BAYESIAN STATISTICS 7, pp. 000–000
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)

The Variational Bayesian EM Algorithm for Incomplete Data: with Application to Scoring Graphical Model Structures

MATTHEW J. BEAL and ZOUBIN GHAHRAMANI
Gatsby Computational Neuroscience Unit, UCL, UK
m.beal@gatsby.ucl.ac.uk zoubin@gatsby.ucl.ac.uk

SUMMARY

We present an efficient procedure for estimating the marginal likelihood of probabilistic models with latent variables or incomplete data. This method constructs and optimises a lower bound on the marginal likelihood using variational calculus, resulting in an iterative algorithm which generalises the EM algorithm by maintaining posterior distributions over both latent variables *and parameters*. We define the family of conjugate-exponential models—which includes finite mixtures of exponential family models, factor analysis, hidden Markov models, linear state-space models, and other models of interest—for which this bound on the marginal likelihood can be computed very simply through a modification of the standard EM algorithm. In particular, we focus on applying these bounds to the problem of scoring discrete directed graphical model structures (Bayesian networks). Extensive simulations comparing the variational bounds to the usual approach based on the Bayesian Information Criterion (BIC) and to a sampling-based gold standard method known as Annealed Importance Sampling (AIS) show that variational bounds substantially outperform BIC in finding the correct model structure at relatively little computational cost, while approaching the performance of the much more costly AIS procedure. Using AIS allows us to provide the first serious case study of the tightness of variational bounds. We also analyse the perfomance of AIS through a variety of criteria, and outline directions in which this work can be extended.

Keywords: MARGINAL LIKELIHOOD; LATENT VARIABLES; VARIATIONAL METHODS; GRAPHICAL MODELS; ANNEALED IMPORTANCE SAMPLING; STRUCTURE SCORING; BAYES FACTORS.

1. INTRODUCTION

Statistical modelling problems often involve a large number of random variables and it is often convenient to express the conditional independence relations between these variables graphically. Such graphical models are an intuitive tool for visualising dependencies between the variables. Moreover, by exploiting the conditional independence relationships, they provide a backbone upon which it has been possible to derive efficient message-propagating algorithms for conditioning and marginalising variables in the model given new evidence (Pearl, 1988; Lauritzen and Spiegelhalter, 1988; Heckerman, 1996; Cowell *et al.* , 1999). Many standard statistical models, especially Bayesian models with hierarchical priors, can be expressed naturally using probabilistic graphical models. This representation can be helpful developing both sampling methods (e.g. Gibbs sampling) and exact inference methods (e.g. junction-tree algorithm) for these models.

An important and difficult problem in Bayesian inference is computing the marginal likelihood of a model. This problem appears under several guises: as computing the Bayes factor (the ratio of two marginal likelihoods; Kass and Raftery, 1995), or computing the normalising constant of a posterior distribution (known in statistical physics as the "partition function" and in machine learning as the "evidence"). The marginal likelihood is an important quantity because it allows us to select between several model structures. It is a difficult quantity to compute because it involves integrating over all parameters and latent variables, which is usually such a high dimensional and complicated integral that most simple approximations fail catastrophically.

In this paper we describe the use of variational methods to approximate the marginal likelihood and posterior distributions of complex models. Variational methods, which have been used extensively in Bayesian machine learning for several years, provide a lower bound on the marginal likelihood which can be computed efficiently. In the next subsections we review Bayesian approaches to learning model structure. In section 2 we turn to describing variational methods applied to Bayesian learning, deriving the variational Bayesian EM algorithm and comparing it to the EM algorithm for maximum a posteriori (MAP) estimation. In section 3, we focus on models in the conjugate-exponential family and derive the basic results. Section 4 introduces the specific problem of learning the conditional independence structure of directed acyclic graphical models with latent variables. We compare variational methods to BIC and annealed importance sampling (AIS). We conclude with a discussion of other variational methods in section 5, and areas for future work in the general area of learning model structure.

1.1. *Bayesian Learning of Model Structure*

We use the term "model structure" to denote a variety of things. (1) In probabilistic graphical models, each graph implies a set of conditional independence statements between the variables in the graph. For example, in directed acyclic graphs (DAGs) if two variables X and Y are d-separated by a third Z (see Pearl, 1988, for a definition), then X and Y are conditionally independent given Z. The model structure learning problem is inferring the conditional independence relationships that hold given a set of (complete or incomplete) observations of the variables. (2) A special case of this problem is input variable selection in regression. Selecting which input (i.e. explanatory) variables are needed to predict the output (i.e. response) variable in the regression can be equivalently cast as deciding whether each input variable is a parent of the output variable in the corresponding directed graph. (3) Many statistical models of interest contain discrete nominal latent variables. A model structure learning problem of interest is choosing the cardinality of the discrete latent variable. Examples of this problem include deciding how many mixture components in a finite mixture model or how many hidden states in a hidden Markov model. (4) Other statistical models contain real-valued vectors of latent variables, making it necessary to do inference on the dimensionality of the latent vector. Examples of this include choosing the instrinsic dimensionality in a probabilistic principal components analysis (PCA) or factor analysis (FA) model or in a linear-Gaussian state-space model.

An obvious problem with maximum likelihood methods is that the likelihood function will generally be higher for more complex model structures, leading to overfitting. Bayesian approaches overcome overfitting by treating the parameters $\boldsymbol{\theta}_i$ from a model m_i (out of a set of models) as unknown random variables and averaging over the likelihood one would obtain from different settings of $\boldsymbol{\theta}_i$:

$$p(\mathbf{y} \,|\, m_i) = \int p(\mathbf{y} \,|\, \boldsymbol{\theta}_i, m_i) p(\boldsymbol{\theta}_i \,|\, m_i) \; d\boldsymbol{\theta}_i \; . \tag{1}$$

$p(\mathbf{y} \,|\, m_i)$ is the *marginal likelihood*[1] for a data set $\mathbf{y}$ assuming model m_i, where $p(\boldsymbol{\theta} \,|\, m_i)$ is the prior distribution over parameters. Integrating out parameters penalises models with more degrees of freedom since these models can *a priori* model a larger range of data sets. This property of Bayesian integration has been called Ockham's razor, since it favors simpler explanations (models) for the data over complex ones (Jefferys and Berger, 1992; MacKay, 1995).[2] The overfitting problem is avoided simply because no parameter in the pure Bayesian approach is actually *fit* to the data.

Given a prior distribution over model structures $p(m_i)$ and a prior distribution over parameters for each model structure $p(\boldsymbol{\theta} \,|\, m_i)$, observing the data set $\mathbf{y}$ induces a posterior distribution over models given by Bayes rule:

$$p(m_i \,|\, \mathbf{y}) = \frac{p(m_i) p(\mathbf{y} \,|\, m_i)}{p(\mathbf{y})}, \qquad p(\boldsymbol{\theta} \,|\, \mathbf{y}, m_i) = \frac{p(\mathbf{y} \,|\, \boldsymbol{\theta}, m_i) p(\boldsymbol{\theta} \,|\, m_i)}{p(\mathbf{y} \,|\, m_i)}. \tag{2}$$

Our uncertainty over parameters is quantified by the posterior $p(\boldsymbol{\theta} \,|\, \mathbf{y}, m_i)$. The density at a new data point y is obtained by averaging over both the uncertainty in the model structure and in the parameters, $p(y \,|\, \mathbf{y}) = \sum_{m_i} \int p(y \,|\, \boldsymbol{\theta}, m_i, \mathbf{y}) p(\boldsymbol{\theta} \,|\, m_i, \mathbf{y}) p(m_i \,|\, \mathbf{y}) \; d\boldsymbol{\theta}$, which is the *predictive distribution.* Although in theory we should average over all possible structures, in practice, constraints on storage and computation or ease of interpretability may lead us to select a most probable model structure by maximising $p(m_i \,|\, \mathbf{y})$. We focus on methods for computing probabilities over structures.

1.2. *Practical Bayesian Approaches*

For most models of interest it is computationally and analytically intractable to perform the integrals required for (1) and (2) exactly. Not only do these involve very high dimensional integrals but for models with parameter symmetries (such as mixture models) the integrand can have exponentially many well-separated modes. Focusing on (1), we briefly outline several methods that have been used to approximate this integral before turning to variational methods.

In the Bayesian statistics community *Markov chain Monte Carlo* (MCMC) is the method of choice for approximating difficult high dimensional expectations and integrals. While there are many MCMC methods which result asymptotically in samples from the posterior distribution over parameters, for models with symmetries it is hard to get mixing between modes, which can be crucial in obtaining valid estimates of the marginal likelihood. Several methods that have been proposed for estimating marginal likelihoods are the candidate method (Chib, 1995; and see Neal, 1998), bridge sampling and path sampling (Gelman and Meng, 1998) and the closely related Annealed Importance Sampling (Neal, 2001). For large-scale problems, sampling methods are often not the method of choice because they can be slow and the posterior distribution

[1] In the machine learning community this is sometimes referred to as the *evidence* for model m_i.

[2] However, it is important to keep in mind that a realistic model of the data might need to be complex. It is therefore often advisable to use the most "complex" model for which it is possible to do inference, ideally setting up priors that allow the limit of infinitely many parameters to be taken, rather than to artificially limit the number of parameters in the model (Neal, 1996; Rasmussen and Ghahramani, 2001).

over parameters is stored as a set of samples, which can be inefficient from a memory standpoint. In this paper we chose Annealed Importance Sampling (AIS) as the gold standard with respect to which we compare faster non-sampling based approximations.

Another approach to Bayesian integration is the *Laplace approximation* which makes a local Gaussian approximation around a maximum *a posteriori* (MAP) parameter estimate (Kass and Raftery, 1995; MacKay, 1995). This is based on the fact that in the large data limit, given some regularity conditions, the posterior approaches a Gaussian around the MAP estimate. However, for models with symmetries these regularity conditions do not hold and the posterior can approach a mixture of exponentially many Gaussians. The Gaussianity assumption can be inaccurate for small data sets (for which, in principle, the advantages of Bayesian integration over MAP are largest). Local Gaussian approximations are also poorly suited to bounded, constrained or positive parameters such as the mixing proportions of a mixture model, so it is often advisable to reparameterise to make Gaussianity reasonable. Finally, the Gaussian approximation requires computing or approximating the determinant of the Hessian matrix at the MAP estimate, which can be computationally costly, i.e. $O(d^3)$ for d parameters.

An even more drastic but much less costly approximation to the marginal likelihood is given by the Bayesian Information Criterion (BIC; Schwarz, 1978) which can be derived from the Laplace approximation by dropping all terms that do not scale with the number of data points, n. For a model with d well-determined parameters, using the MAP parameter estimate $\hat{\boldsymbol{\theta}}$ the BIC approximation to the log marginal likelihood is: $\ln p(\mathbf{y} \,|\, m_i) \approx \ln p(\mathbf{y} \,|\, \hat{\boldsymbol{\theta}}, m_i) - \frac{d}{2} \ln n$.

2. MARGINAL LIKELIHOOD VIA A VARIATIONAL PRINCIPLE

We review how variational methods can be used to approximate the integrals required for Bayesian learning. While examples of these methods have been used for several years in the machine learning community they are less known to the Bayesian statistics community. Moreover, here we present the general framework for a large family of models, we investigate a novel application of the framework to scoring the structures of discrete graphical models, and provide the first serious assessments of how tight the variational bounds are and how well they compare to sampling methods. We focus on models with latent or hidden variables, considering a general incomplete data setting.

2.1 *Lower Bounding the Marginal Likelihood*

Let $\mathbf{y}$ denote the observed variables, $\mathbf{x}$ denote the latent variables, and $\boldsymbol{\theta}$ denote the parameters. The log marginal likelihood of a data set $\mathbf{y}$ 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, and then appealing to Jensen's inequality (due to the concavity of the logarithm function):

$$\ln p(\mathbf{y} \,|\, m) = \ln \int p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta} \,|\, m)\, d\mathbf{x}\, d\boldsymbol{\theta} = \ln \int q(\mathbf{x}, \boldsymbol{\theta}) \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta} \,|\, m)}{q(\mathbf{x}, \boldsymbol{\theta})}\, d\mathbf{x}\, d\boldsymbol{\theta} \tag{3}$$

$$\geq \int q(\mathbf{x}, \boldsymbol{\theta}) \ln \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta} \,|\, m)}{q(\mathbf{x}, \boldsymbol{\theta})}\, d\mathbf{x}\, d\boldsymbol{\theta}. \tag{4}$$

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 true posterior distribution

$p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y}, m)$ requires knowing its normalising constant, the marginal likelihood. Instead we use a simpler, factorised approximation to $q(\mathbf{x}, \boldsymbol{\theta}) \approx q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$:

$$\ln p(\mathbf{y} \,|\, m) \;\geq\; \int q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta} \,|\, m)}{q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \, d\mathbf{x} \, d\boldsymbol{\theta} \;=\; \mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathbf{y}). \qquad (5)$$

The quantity $\mathcal{F}$ is a functional of the free distributions $q_{\mathbf{x}}(\mathbf{x})$ and $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$.

2.2. *Variational Bayesian EM*

The variational Bayesian algorithm iteratively maximises $\mathcal{F}$ in equation (5) with respect to the free distributions, $q_{\mathbf{x}}(\mathbf{x})$ and $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$. We use elementary calculus of variations to take functional derivatives of the lower bound with respect $q_{\mathbf{x}}(\mathbf{x})$ and $q_{\boldsymbol{\theta}}(\theta)$, each while holding the other fixed. This results in the following update equations where the superscript (t) denotes the iteration number.

$$q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \;\propto\; \exp\left[\int \ln p(\mathbf{x},\mathbf{y} \,|\, \boldsymbol{\theta}, m) \, q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta}) \, d\boldsymbol{\theta}\right] \qquad (6)$$

$$q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}) \;\propto\; p(\boldsymbol{\theta} \,|\, m) \, \exp\left[\int \ln p(\mathbf{x},\mathbf{y} \,|\, \boldsymbol{\theta}, m) \, q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \, d\mathbf{x}\right] . \qquad (7)$$

Clearly $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and $q_{\mathbf{x}_i}(\mathbf{x}_i)$ are coupled, so we iterate these equations until convergence. Readers familiar with the EM algorithm (Dempster *et al.* , 1977) may note the similarity between this iterative algorithm and EM. We call this procedure the *Variational Bayesian EM Algorithm* for reasons which will become clearer in the following sections; see also Attias (2000) and Ghahramani and Beal (2001).

Re-writing (5), it is easy to see that maximising $\mathcal{F}$ is equivalent to minimising the KL divergence between $q_{\mathbf{x}}(\mathbf{x}) \, q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and the joint posterior $p(\mathbf{x}, \boldsymbol{\theta} \,|\, \mathbf{y}, m)$:

$$\ln p(\mathbf{y} \,|\, m) - \mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathbf{y}) = \int q_{\mathbf{x}}(\mathbf{x}) \, q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \frac{q_{\mathbf{x}}(\mathbf{x}) \, q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{p(\boldsymbol{\theta}, \mathbf{x} \,|\, \mathbf{y}, m)} \, d\mathbf{x} \, d\boldsymbol{\theta} = \mathrm{KL}(q \| p) \; . \quad (8)$$

Note that whilst this factorisation of the posterior distribution over latent variables and parameters may seem drastic, one can think of it as replacing stochastic dependencies between $\mathbf{x}$ and $\boldsymbol{\theta}$ with deterministic dependencies between relevant moments of the two sets of variables.

Variational methods for lower bounding probabilities have been explored by several researchers in the past decade. Hinton and van Camp (1993) proposed an early approach for Bayesian learning of one-hidden layer neural networks using the restriction that $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is Gaussian. Neal and Hinton (1998) presented a generalisation of EM which made use of Jensen's inequality to allow partial E-steps. Jordan *et al.* (1998) review variational methods in a general context. Variational Bayesian methods have been applied to various models with latent variables (Waterhouse *et al.* , 1995; MacKay, 1997; Bishop, 1999; Attias, 2000; Ghahramani and Beal, 2000). The structural EM algorithm for scoring discrete graphical models (Friedman, 1998) is closely related to the variational method described here except that in (6) the distribution over $\boldsymbol{\theta}$ is replaced by the MAP estimate.

3. CONJUGATE-EXPONENTIAL MODELS

We consider a particular class of graphical models with latent variables, which we call *conjugate-exponential* (CE) models. We explicitly apply the variational method to these parametric families, resulting in a simple generalisation of EM[3]. Conjugate-exponential models satisfy two conditions:

Condition (1). *The complete data likelihood is that of an exponential family:* $p(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\theta}) = f(\mathbf{x}, \mathbf{y})\ g(\boldsymbol{\theta}) \exp\left\{\boldsymbol{\phi}(\boldsymbol{\theta})^T \mathbf{u}(\mathbf{x}, \mathbf{y})\right\}$*, where* $\boldsymbol{\phi}(\boldsymbol{\theta})$ *is the vector of natural parameters, and* $\mathbf{u}$ *and* f *and* g *are the functions that define the exponential family.*

Condition (2). *The parameter prior is conjugate to the complete data likelihood:* $p(\boldsymbol{\theta} \mid \eta, \boldsymbol{\nu}) = h(\eta, \boldsymbol{\nu})\ g(\boldsymbol{\theta})^\eta \exp\left\{\boldsymbol{\phi}(\boldsymbol{\theta})^T \boldsymbol{\nu}\right\}$*, where* η *and* $\boldsymbol{\nu}$ *are hyperparameters.*

Theorem. (Conjugate-Exponential Models). *Given an iid data set* $\mathbf{y} = \{\mathbf{y}_1, \dots \mathbf{y}_n\}$*, if the model satisfies conditions (1) and (2), then at every iteration of the variational Bayesian EM algorithm and at the maxima of* $\mathcal{F}(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathbf{y})$*:*

(a) $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ *is conjugate with parameters* $\tilde{\eta} = \eta + n$, $\tilde{\boldsymbol{\nu}} = \boldsymbol{\nu} + \sum_{i=1}^n \overline{\mathbf{u}}(\mathbf{y}_i)$*:*

$$q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = h(\tilde{\eta}, \tilde{\boldsymbol{\nu}}) g(\boldsymbol{\theta})^{\tilde{\eta}} \exp\left\{\boldsymbol{\phi}(\boldsymbol{\theta})^T \tilde{\boldsymbol{\nu}}\right\} \tag{9}$$

where $\overline{\mathbf{u}}(\mathbf{y}_i) = \mathrm{E}_{q_{\mathbf{x}_i}}(\mathbf{u}(\mathbf{x}_i, \mathbf{y}_i))$*, using* $\mathrm{E}_{q_{\mathbf{x}_i}}$ *to denote expectation under the variational posterior over the latent variable(s) associated with the* i*th datum.*

(b) $q_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^n q_{\mathbf{x}_i}(\mathbf{x}_i)$ *with*

$$q_{\mathbf{x}_i}(\mathbf{x}_i) = p(\mathbf{x}_i \mid \mathbf{y}_i, \overline{\boldsymbol{\phi}}) \propto f(\mathbf{x}_i, \mathbf{y}_i) \exp\left\{\overline{\boldsymbol{\phi}}^T \mathbf{u}(\mathbf{x}_i, \mathbf{y}_i)\right\} \tag{10}$$

where $\overline{\boldsymbol{\phi}} = \mathrm{E}_{q_{\boldsymbol{\theta}}}(\boldsymbol{\phi}(\boldsymbol{\theta}))$*, the expectation of the natural parameter.*

Proof. Substitute the parametric forms from the definition of the CE family into the variational extrema given in (6) and (7), revealing forms for $q_{\mathbf{x}}(\mathbf{x})$ and $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ according to (10) and (9) respectively. For CE models these forms are then closed under iterations of variational Bayesian EM, ensuring the Theorem continues to hold through to convergence to a local maximum of the lower bound on the marginal likelihood.

3.1. *Comparison of Variational Bayesian EM and EM for MAP estimation*

It is instructive to compare (6) and (7) with the EM algorithm for MAP estimation. We use an alternative derivation of EM due to Neal and Hinton (1998):

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

The Variational Bayesian EM algorithm reduces to the ordinary EM algorithm if we restrict the parameter density to a point estimate (i.e. Dirac delta function),

[3]This section follows the exposition in Ghahramani and Beal (2001), which also includes several general results for directed and undirected graphs.

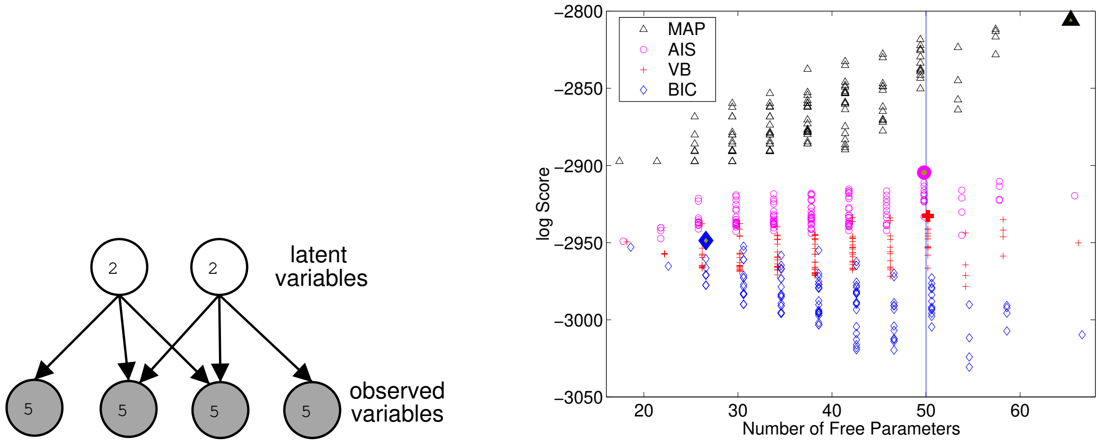

Figure 1. (left) *The true structure used to generate the data—one of 136 possible distinct structures accounting for permutations of latent variables.* **(right)** *Marginal likelihood estimates at* $n = 480$ *for all structures using MAP, AIS, VB and BIC scores, plotted against the number of free parameters in the structure (staggered for each number of parameters for clarity). The true structure has 50 parameters (vertical line); the highest scoring structure for each method is shown with a bold symbol.*

$q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$. The VB-E 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, $\overline{\boldsymbol{\phi}}$. In particular, we can make use of all relevant propagation algorithms such as junction tree, Kalman smoothing, or belief propagation. The VB-M step computes a *distribution* over parameters (in the conjugate family) rather than a point estimate. Both algorithms monotonically increase an objective function.

4. VARIATIONAL SCORING OF MODEL STRUCTURES

We apply the variational method to the non-trivial problem of learning the conditional independence structure of directed acyclic graphical models with latent variables. We compare our VB bounds for the marginal likelihood to the standard method of scoring graphs based on the Bayesian Information Criterion (BIC; Schwarz, 1978), and also to a sampling-based gold standard: Annealed Importance Sampling (AIS; Neal, 2001).

We consider a small graph consisting of six discrete variables: two binary-valued latent (hidden) variables and four observed variables each of cardinality five. We restrict ourselves to all bipartite structures in which latent variables are parents of the observed variables. We are interested in how successful different scoring methods are at learning from data the true graph structure (i.e. which latent variables are parents of which observed variables). Since all variables are discrete, by placing independent Dirichlet priors on all parameters the model becomes conjugate-exponential.

Data was generated from the graph shown in Figure 1; we call this the "true" structure. We instantiated a setting of this graph's parameters under the prior (once only), and generated incrementally larger data sets from the model.[4] We chose this particular structure because it contains enough links to induce non-trivial correlations amongst the observed variables, whilst it has few enough nodes that we can exhaustively evaluate the marginal likelihood of all possible alternative structures.[5]

[4]Experiments averaging results over multiple true structures and parameter settings would have been prohibitive as our sampling runs took over 300 CPU hours.

[5]Exhaustive enumeration is of academic interest only—in practice one would embed different structure scoring methods in a greedy model search outer loop (Friedman, 1998) to find probable structures.

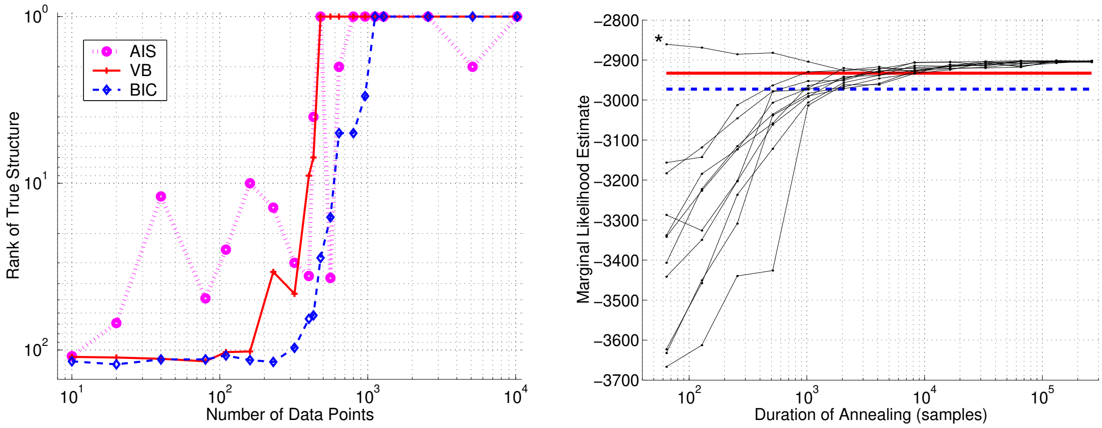

Figure 2. **(left)** *Rank given to the true structure by each scoring method for varying data set sizes (higher in plot is better).* **(right)** *AIS estimates of the marginal likelihood for $n = 480$, for different initial conditions of the sampler against different duration annealing schedules ($*$ indicates setting initial parameters to true values). Shown are also the BIC score (dashed) and the VB lower bound (solid).*

4.1. *Annealed Importance Sampling (AIS)*

We estimate $\mathcal{Z}_1 \equiv \int p(\boldsymbol{\theta} \,|\, m) p(\mathbf{y} \,|\, \boldsymbol{\theta}, m)\, d\boldsymbol{\theta}$ by computing $\frac{\mathcal{Z}_1}{\mathcal{Z}_0} = \frac{\mathcal{Z}_{\tau(1)}}{\mathcal{Z}_{\tau(0)}} \frac{\mathcal{Z}_{\tau(2)}}{\mathcal{Z}_{\tau(1)}} \cdots \frac{\mathcal{Z}_{\tau(T)}}{\mathcal{Z}_{\tau(T-1)}}$ where $0 = \tau(0) \leq \tau(1) \ldots \leq \tau(T) = 1$, and $\mathcal{Z}_\tau \equiv \int p(\boldsymbol{\theta} \,|\, m) p(\mathbf{y} \,|\, \boldsymbol{\theta}, m)^\tau d\boldsymbol{\theta}$. The sequence $\tau(0) \ldots \tau(T)$ is an annealing schedule of inverse temperatures. Each ratio $\frac{\mathcal{Z}_{\tau(t)}}{\mathcal{Z}_{\tau(t-1)}}$ is estimated by running a Metropolis sampler for $\boldsymbol{\theta}^{(i)} \sim p(\boldsymbol{\theta} \,|\, m) p(\mathbf{y} \,|\, \boldsymbol{\theta}, m)^{\tau(t-1)} / \mathcal{Z}_{\tau(t-1)}$, and computing the importance estimate: $\frac{1}{N} \sum_i p(\mathbf{y} \,|\, \boldsymbol{\theta}^{(i)}, m)^{\tau(t) - \tau(t-1)}$. This estimate of $\frac{\mathcal{Z}_{\tau(t)}}{\mathcal{Z}_{\tau(t-1)}}$ is unbiased if each step is sampled from equilibrium; we approximate this by taking one sample at each τ and changing τ very slowly.

4.2. *Experiments*

Scoring all possible structures with MAP, BIC, VB, and AIS. There are 136 distinct structures with the basic architecture described above. For a large range of data set sizes, we ran EM on each structure to compute the MAP estimate of the parameters, and from it computed the BIC score. We also ran the variational Bayesian EM algorithm with the same initial conditions to obtain the lower bound on the marginal likelihood. Finally we estimated the marginal likelihood using AIS, annealing from the prior to the posterior in 16384 steps, with a nonlinear annealing schedule in τ tuned to reduce the variance in the estimate, and a Metropolis proposal tuned to give reasonable acceptance rates. In Figure 1 we have shown the MAP, AIS, VB and BIC scores for each structure, ordered by number of parameters, for 480 data points. This data set size was chosen as it was the smallest in which both VB and AIS gave the highest score to the true structure. We can see the general upward trend for MAP, which prefers more complex structures and the general downward trend for BIC which (over)penalises complexity. AIS lies above the VB lower bound for all structures as we would expect. At 480 data points VB appears to be close to AIS and finds the correct structure.

Rank of the True Model. Figure 2 shows the rank out of 136 given by each scoring method to the true structure for 20 data set sizes. All methods eventually find the correct structure, although the AIS rank is noisy, which may be due to annealing too rapidly (we examine the effect of annealing time on AIS below). BIC finds it at 1120 data points, while VB does so after 480; moreover the true structure is almost always ranked better by VB than BIC. Especially for small amounts of data, AIS outperforms both VB and BIC; nevertheless it finds the true structure only after 480 data points.

Performance of AIS with Sampling Time. We ran AIS on the 480 point data set for varying duration annealing schedules, ranging from 64 samples to over 260,000. Figure 2 (right) shows annealing runs starting at 10 random initial parameters sampled from the prior, and also at one starting from the true parameters that generated the data (marked with a $*$); also shown are the BIC and VB scores. We see that all runs converge for sufficiently long annealing schedules. It takes roughly 1000 samples to approach the BIC score and about 5000 to pass the VB lower bound.

Computation Time. Scoring all 136 structures at 480 data points on a 1GHz Pentium III processor took: 200 seconds with BIC, 575 seconds with VB, and 55000 seconds (15 hours) with AIS (using 16384 samples as in the main experiments). Referring back to Figure 2 (left), we can infer that in this example, not only was VB about 100 times faster than AIS, but it also locked on to the true structure more reliably than AIS.

5. CONCLUSIONS AND EVOLVING DIRECTIONS FOR STRUCTURE SCORING

This paper has presented the variational lower bound method for handling intractable marginal likelihoods. A promising future direction is to explore the Bethe and Kikuchi family of variational methods (Yedidia *et al.* , 2001) which may be more accurate but do not provide the assurance of being a bound. These re-express the negative log marginal likelihood as a "free energy" from statistical physics, and then approximate the entropy of the posterior distribution over latent variables and parameters neglecting high order terms. There are several procedures for minimising the Bethe free energy as a functional of the approximate posterior distribution to obtain estimates of the marginal likelihood. Interestingly, the belief propagation algorithm, even when run on multiply-connected graphs (i.e. 'loopy' graphs), has fixed points at the stationary points of the Bethe free energy. While belief propagation on loopy graphs is not guaranteed to converge, it often works well in practice, and has become the standard approach to decoding state-of-the-art error-correcting codes.

We have also explored approximations to the marginal likelihood based on the Cheeseman-Stutz score as outlined in Chickering and Heckerman (1996), and corrections to the BIC criterion as outlined in Geiger *et al.* (1996). Results on these, and suggestions for how they can be combined with variational methods are reported in Beal (2003). A reasonable comparison should take into account the tradeoff between accuracy and computational complexity. We found in this paper that the VB lower bound on the marginal likelihood is a good criterion for model selection which can be computed orders of magnitude more quickly than sampling-based criteria. We have had similar experiences on the usefulness of VB for model selection in a variety of other models, and we note that the VB framework can be readily applied to all conjugate-exponential models with incomplete data, including e.g., graphical Gaussian models.

ACKNOWLEDGEMENTS

We thank Carl Edward Rasmussen for extensive input and helpful suggestions, especially regarding AIS. Part of this work was done while ZG was visiting the Center for Automated Learning and Discovery at Carnegie Mellon University.

REFERENCES

Attias, H. (2000). A variational Bayesian framework for graphical models. In *Advances in Neural Information Processing Systems 12*, MIT Press.

Beal, M. J. (2003). Variational algorithms for approximate Bayesian inference. Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London.

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

Chib, S. (1995). Marginal likelihood from the Gibbs output. *JASA*, 90:1313–1321.

Chickering, D. M. and Heckerman, D. (1996). Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. In *Proc. 12th Conf. on Uncertainty in Artificial Intelligence. (UAI '96)*, pages 158–68. Morgan Kaufmann Publishers.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems*. Springer-Verlag, New York.

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. 14th Conf. on Uncertainty in Artificial Intelligence (UAI '98)*, San Francisco, CA. Morgan Kaufmann.

Geiger, D., Heckerman D., and Meek C. (1996). Asymptotic model selection for directed networks with hidden variables. In *Proc. 12th Conf. on Uncertainty in Artificial Intelligence (UAI '96)*. 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*, MIT Press.

Ghahramani, Z. and Beal, M. J. (2001). Propagation algorithms for variational Bayesian learning. In *Advances in Neural Information Processing Systems 13*, MIT Press.

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.

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.*

Jefferys,W. H. and Berger,J. O. (1992). Ockham's razor and Bayesian analysis. *Amer. Scientist*, 80:64–72.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1998). An Introduction to variational methods in graphical models. In Jordan, M. I., editor, *Learning in Graphical Models*. Kluwer.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.*, 90:773–795.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *J. Roy. Statist. Soc. B*, 50:154–227.

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.

MacKay, D. J. C. (1997). Ensemble learning for hidden Markov models. Technical report, Cavendish Laboratory, University of Cambridge.

Neal, R. M. (1996). *Bayesian Learning in Neural Networks*. Springer Verlag.

Neal, R. M. (1998). Erroneous results in "marginal likelihood from the Gibbs output".

Neal, R. M. (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., editor, *Learning in Graphical Models*, pages 355–369. Kluwer.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA.

Rasmussen, C. E. and Ghahramani, Z. (2001). Occam's razor. In *Advances in Neural Information Processing Systems 13*, MIT Press.

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

Waterhouse, S., MacKay, D. J. C., and Robinson, T. (1995). Bayesian methods for mixtures of experts. In *Advances in Neural Information Processing Systems 7*, MIT Press.

Yedidia, J., Freeman, W. T., and Weiss, Y. (2001). Generalized belief propagation. In *Advances in Neural Information Processing Systems 13*, MIT Press.