Structured Variational Distributions in VIBES

Christopher M. Bishop
Microsoft Research
Cambridge, CB1 2HN, U.K.
cmbishop@microsoft.com
http://research.microsoft.com/∼cmbishop

John Winn
Department of Physics
University of Cambridge, U.K.
jmw39@cam.ac.uk
http://www.inference.phy.cam.ac.uk/jmw39

Abstract

Variational methods are becoming increasingly popular for the approximate solution of complex probabilistic models in machine learning, computer vision, information retrieval and many other fields. Unfortunately, for every new application it is necessary first to derive the specific forms of the variational update equations for the particular probabilistic model being used, and then to implement these equations in application-specific software. Each of these steps is both time consuming and error prone. We have therefore recently developed a general purpose inference engine called VIBES [1] ('Variational Inference for Bayesian Networks') which allows a wide variety of probabilistic models to be implemented and solved variationally without recourse to coding. New models are specified as a directed acyclic graph using an interface analogous to a drawing package, and VIBES then automatically generates and solves the variational equations. The original version of VIBES assumed a fully factorized variational posterior distribution. In this paper we present an extension of VIBES in which the variational posterior distribution corresponds to a sub-graph of the full probabilistic model. Such structured distributions can produce much closer approximations to the true posterior distribution. We illustrate this approach using an example based on Bayesian hidden Markov models.

1 Introduction

Variational methods [2] have been used successfully for a wide range of models, and new applications are constantly being explored. In many ways the variational framework can be seen as a complementary approach to that of Markov chain Monte Carlo (MCMC), with different strengths and weaknesses. The variational approach finds a deterministic approximation to the posterior distribution by optimization over an analytical family of distributions.

For many years there has existed a powerful tool for tackling new problems using MCMC, called BUGS ('Bayesian inference Using Gibbs Sampling') [3]. In BUGS a new probabilistic model, expressed as a directed acyclic graph, can be encoded using a simple scripting notation, and then samples can be drawn from the posterior distribution (given some data set of observed values) using Gibbs sampling in a way that is largely automatic. Furthermore, an extension called WinBUGS provides a graphical front end to BUGS in which the user draws a pictorial representation of the directed graph, and this automatically generates the required script.

We have been inspired by the success of BUGS to produce an analogous tool for the solution of problems using variational methods. The challenge is to build a system which can handle a wide range of graph structures, a broad variety of common conditional probability distributions at the nodes, and a range of variational approximating distributions. All of this must be achieved whilst also remaining computationally efficient.

The VIBES software uses a graphical interface analogous to that used in WinBUGS in order to specify the probabilistic model in terms of a directed acyclic graph. A subset of the nodes of this graph represent observed variables (the 'data') and the remainder represent hidden variables. We are interested primarily in models for which exact inference is intractable.

In the original version of VIBES [1] the variational posterior distribution was assumed to be fully factorized with respect to the nodes of the graph. This already represents a powerful and practical framework for approximate inference that has been widely applied. However, by allowing more structure in the variational posterior we can access a much richer class of approximations and hence obtain better approximations to the true posterior distributions [4, 5, 6]. In this paper we describe an extended version of

VIBES which uses variational distributions given by subgraphs of discrete and/or continuous (Gaussian) nodes obtained by deleting links from the original graph.

In Section 2 we discuss the framework of variational inference in some generality, and review the fully factorized approximation used in the original form of VIBES. Tractability in VIBES is achieved by considering the conditional probability distributions which form the graph to be drawn from the exponential family, as reviewed in Section 3, with conjugacy imposed between all parent-child pairs in the graph. The VIBES software implementation is described in Section 4. In Section 5 we then discuss the general structured variational distribution and derive the corresponding variational update equations. This is illustrated by applying VIBES to a Bayesian hidden Markov model in Section 6. Some directions for future development are discussed in Section 7.

2 Variational Inference

In this section we briefly review the general variational framework, and then we derive the variational update equations for the case of a fully factorized variational posterior distribution.

We denote the set of all variables in the model by $W = (V, X)$ where V are the visible (observed) variables and X are the latent (hidden) variables. Throughout this paper we focus on models which are specified in terms of an acyclic directed graph, although the treatment of undirected graphical models is equally possible and indeed is somewhat more straightforward. The joint distribution $P(V, X)$ is then expressed in terms of conditional distributions $P(W_i|\mathrm{pa_i})$ at each node i, where W_i denotes the variable, or group of variables, associated with node i, and $\mathrm{pa_i}$ denotes the set of variables corresponding to the parents of node i. The joint distribution of all variables is then given by the product of the conditionals

$$P(V, X) = \prod_i P(W_i|\mathrm{pa_i}). \tag{1}$$

Our goal is to find a variational distribution $Q(X|V)$ which approximates the true posterior distribution $P(X|V)$. To do this we note the following decomposition of the log marginal probability of the observed data, which holds for any choice of distribution $Q(X|V)$

$$\ln P(V) = \mathcal{L}(Q) + \mathrm{KL}(Q\|P) \tag{2}$$

where

$$\mathcal{L}(Q) = \sum_X Q(X|V) \ln \frac{P(V, X)}{Q(X|V)} \tag{3}$$

$$\mathrm{KL}(Q\|P) = -\sum_X Q(X|V) \ln \frac{P(X|V)}{Q(X|V)} \tag{4}$$

and the sums are replaced by integrals in the case of continuous variables. Here $\mathrm{KL}(Q\|P)$ is the Kullback-Lieber divergence between the variational approximation $Q(X|V)$ and the true posterior $P(X|V)$. Since this satisfies $\mathrm{KL}(Q\|P) \geq 0$ it follows from (2) that the quantity $\mathcal{L}(Q)$ forms a rigorous lower bound on $\ln P(V)$.

We now choose some family of distributions to represent $Q(X|V)$ and then seek a member of that family which maximizes the lower bound $\mathcal{L}(Q)$. If we allow $Q(X|V)$ to have complete flexibility then we see that the maximum of the lower bound occurs for $Q(X|V) = P(X|V)$ so that the variational posterior distribution equals the true posterior. In this case the Kullback-Leibler divergence vanishes and $\mathcal{L}(Q) = \ln P(V)$. However, working with the true posterior distribution is computationally intractable (otherwise we wouldn't be resorting to variational methods). We must therefore consider a more restricted family of Q distributions which has the property that the lower bound (3) can be evaluated and optimized efficiently and yet which is still sufficiently flexible as to give a good approximation to the true posterior distribution.

2.1 Factorized Distributions

In the original version of VIBES [1] we focussed on distributions which factorize with respect to disjoint groups X_i of variables

$$Q(X|V) = \prod_i Q_i(X_i). \tag{5}$$

This is already a powerful approximation which has been successfully used in many applications of variational methods [7, 8, 9]. Substituting (5) into (3) we can maximize variationally with respect to $Q_i(X_i)$ keeping all Q_j for $j \neq i$ fixed, which leads to the solution [1]

$$\ln Q_i^\star(X_i) = \langle \ln P(V, X) \rangle_{\{j \neq i\}} + \mathrm{const.} \tag{6}$$

where $\langle \cdot \rangle_k$ denotes an expectation with respect to the distribution $Q_k(X_k)$. Taking exponentials of both sides and normalizing we obtain

$$Q_i^\star(X_i) = \frac{\exp\langle \ln P(V, X) \rangle_{\{j \neq i\}}}{\sum_{X_i} \exp\langle \ln P(V, X) \rangle_{\{j \neq i\}}}. \tag{7}$$

Note that these are coupled equations since the solution for each $Q_i(X_i)$ depends on expectations with respect to the other factors $Q_{j \neq i}$. The variational optimization proceeds by initializing each of the $Q_i(X_i)$ and then cycling through each factor in turn replacing the current distribution with a revised estimate given by (7). The original version of VIBES was based on a factorization of the form (5) in which each factor $Q_i(X_i)$ corresponds to one of the nodes of the graph.

An important property of the variational update equations, from the point of view of VIBES, is that the right hand

side of (7) does not depend on all of the conditional distributions $P(W_i|\mathrm{pa_i})$ which define the joint distribution but only on those which have a functional dependence on X_i, namely the conditional $P(X_i|\mathrm{pa_i})$, together with the conditional distributions for any children of node i since these have X_i in their parent set. Thus the expectations which must be performed on the right hand side of (7) involve only those variables lying in the Markov blanket of node i, in other words the parents, children and co-parents of i, as illustrated in Figure 1(a). This is a key concept in VIBES since it allows the variational update equations to be formulated in terms of local operations which can therefore be expressed in terms of generic code which is independent of the global structure of the graph.

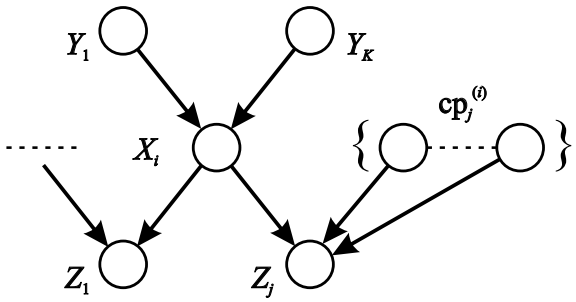

Figure 1: A central observation is that the variational update equations for node X_i depend only on expectations over variables appearing in the *Markov blanket* of X_i, namely the set of parents, children and co-parents.

3 Conjugate Exponential Models

It has already been noted [1, 7, 8] that important simplifications to the variational update equations occur when the distributions of variables, conditioned on their parameters, are drawn from the exponential family and are conjugate with respect to the prior distributions of the parameters. Here we adopt a somewhat different viewpoint in that we make no distinction between latent variables and model parameters. In a Bayesian setting these both correspond to unobserved stochastic variables and can be treated on an equal footing. This allows us to consider conjugacy not just between variables and their parameters, but hierarchically between all parent-child pairs in the graph.

Thus we consider models in which each conditional distribution takes the standard exponential family form

$$\ln P(X_i|Y) = \phi_i(Y)^{\mathrm{T}} u_i(X_i) + f_i(X_i) + g_i(Y) \quad (8)$$

where $Y = \{Y_1, \ldots, Y_K\}$, and the vector $\phi(Y)$ is called the *natural parameter* of the distribution. Now consider a node Z_j with parent X_i and co-parents $\mathrm{cp}_j^{(i)}$, as indicated in Figure 1(a). As far as the pair of nodes X_i and Z_j are concerned, we can think of $P(X_i|Y)$ as a prior over X_i and the conditional $P(Z_j|X_i, \mathrm{cp}_j^{(i)})$ as a (contribution to) the likelihood function. Conjugacy requires that, as a function of X_i, the product of these two conditionals must take the same form as (8). Since the conditional $P(Z_j|X_i, \mathrm{cp}_j^{(i)})$ is also in the exponential family it can be expressed as

$$\begin{aligned} \ln P(Z_j|X_i, \mathrm{cp}_j^{(i)}) &= \phi_j(X_i, \mathrm{cp}_j^{(i)})^{\mathrm{T}} u_j(Z_j) \\ &\quad + f_j(Z_j) + g_j(X_i, \mathrm{cp}_j^{(i)}). \quad (9) \end{aligned}$$

Conjugacy then requires that this be expressible in the form

$$\begin{aligned} \ln P(Z_j|X_i, \mathrm{cp}_j^{(i)}) &= \widetilde{\phi}_{j\to i}(Z_j, \mathrm{cp}_j^{(i)})^{\mathrm{T}} u_i(X_i) \\ &\quad + \lambda(Z_j, \mathrm{cp}_j^{(i)}). \quad (10) \end{aligned}$$

Since this must hold for each of the parents of Z_j it follows that $\ln P(Z_j|X_i, \mathrm{cp}_j^{(i)})$ must be a multi-linear function of the $u_k(X_k)$ for each of the parents X_k of node X_i. Similarly, we observe from (9) that the dependence of $\ln P(Z_j|X_i, \mathrm{cp}_j^{(i)})$ on Z_j is again linear in the function $u_j(Z_j)$. We can apply a similar argument to the conjugate relationship between node X_j and each of its parents, showing that the contribution from the conditional $P(X_i|Y)$ can again be expressed in terms of expectations of the natural parameters for the parent node distributions. Hence the right hand side of the variational update equation (6) for a particular node X_i will be a multi-linear function of the expectations $\langle u_k \rangle$ for each node in the Markov blanket of X_i.

The variational update equation then takes the form

$$\ln Q_j^\star(X_j) = \widehat{\phi}^{\mathrm{T}} u_i(X_i) + \mathrm{const.} \quad (11)$$

where we have defined

$$\widehat{\phi} \equiv \langle \phi_i(Y) \rangle_Y + \sum_{j=1}^{M} \langle \widetilde{\phi}_{j\to i}(Z_j, \mathrm{cp}_j^{(i)}) \rangle_{Z_j, \mathrm{cp}_j^{(i)}}. \quad (12)$$

This involves summation of bottom up 'messages' $\langle \widetilde{\phi}_{j\to i} \rangle_{Z_j, \mathrm{cp}_j^{(i)}}$ from the children together with a top-down message $\langle \phi_i(Y) \rangle_Y$ from the parents. Note that since all of these messages are expressed in terms of the same basis $u_i(X_i)$, we can write compact, generic code for updating any type of node, instead of having to take account explicitly of the many possible combinations of node types in each Markov blanket.

As an example, consider the Gaussian $\mathcal{N}(X|\mu, \tau^{-1})$ for a single variable X with mean μ and precision (inverse variance) τ. The natural coordinates are $u_X = [X, X^2]^{\mathrm{T}}$ and the natural parameterization is $\phi = [\mu\tau, -\tau/2]^{\mathrm{T}}$. Then $\langle u \rangle = [\mu, \mu^2 + \tau^{-1}]^{\mathrm{T}}$, and the function $f_i(X_i)$ is simply zero in this case. Conjugacy allows us to choose a distribution for the parent μ which is Gaussian and a prior for τ which is a Gamma distribution. The corresponding natural

parameterizations and update messages are given by

$$u_\mu = \begin{bmatrix} \mu \\ \mu^2 \end{bmatrix}, \quad \widetilde{\phi}_{X\to\mu} = \begin{bmatrix} \langle\tau\rangle\langle X\rangle \\ -\langle\tau\rangle/2 \end{bmatrix}$$
$$u_\tau = \begin{bmatrix} \tau \\ \ln\tau \end{bmatrix}, \quad \widetilde{\phi}_{X\to\tau} = \begin{bmatrix} -\langle(X-\mu)^2\rangle \\ 1/2 \end{bmatrix}.$$

We can similarly consider multi-dimensional Gaussian distributions, with a Gaussian prior for the mean and a Wishart prior for the inverse covariance matrix.

A generalization of the Gaussian is the rectified Gaussian which is defined as $P(X|\mu,\tau) \propto \mathcal{N}(X|\mu,\tau)$ for $X \geq 0$ and $P(X|\mu,\tau) = 0$ for $X < 0$, for which moments can be expressed in terms of the 'erf' function. This corresponds to the introduction of a step function for $f_i(X_i)$ in (8), and so is carried through the variational update equations unchanged. Similarly, we can consider doubly truncated Gaussians which are non-zero only over some finite interval.

Another example is the discrete distribution for categorical variables. These are most conveniently represented using the 1-of-M coding scheme in which $S = \{S_k\}$ with $k = 1,\ldots,K$, $S_k \in \{0,1\}$ and $\sum_k S_k = 1$. This has a distribution

$$P(S|\pi) = \prod_{k=1}^{K} \pi_k^{S_k} \tag{13}$$

and we can place a conjugate Dirichlet distribution over the parameters $\{\pi_k\}$.

3.1 Allowable Distributions

We now characterize the class of models which can be solved by VIBES using the factorized variational distribution given by (5). This will also be the class of distributions which will be considered in the context of structured Q distributions in Section 5.

First of all we note that, since a Gaussian variable can have a Gaussian parent for its mean, we can extend this hierarchically to any number of levels to give a sub-graph which is a DAG of Gaussian nodes of arbitrary topology. Each Gaussian can have Gamma (or Wishart) prior over its precision.

Next, we observe that discrete variables $S = \{S_k\}$ can be used to construct 'pick' functions which choose a particular parent node $\widehat{Y}$ from amongst several conjugate parents $\{Y_k\}$, so that $\widehat{Y} = Y_k$ when $s_k = 1$, which can be written $\widehat{Y} = \prod_{k=1}^{K} Y_k^{S_k}$. Under any non-linear function $h(\cdot)$ we have $h(Y) = \prod_{k=1}^{K} h(Y_k)^{S_k}$. Furthermore the expectation under S takes the form $\langle h(Y)\rangle_S = \sum_k \langle S_k\rangle h(Y_k)$. Variational inference will therefore be tractable for this model provided it is tractable for each of the parents Y_k individually.

Thus we can handle the following very general architecture: an arbitrary DAG of multi-nomial discrete variables (each having Dirichlet priors) together with an arbitrary DAG of linear Gaussian nodes (each having Wishart priors) and with arbitrary pick links from the discrete nodes to the Gaussian nodes. This graph represents a generalization of the Gaussian mixture model, and includes as special cases the hidden Markov model, Kalman filters, factor analysers and principal component analysers, as well as mixtures and hierarchical mixtures of all of these.

There are other classes of models which are tractable under this scheme, for example Poisson variables having Gamma priors, although these may be of more limited interest.

We can further extend the class of tractable models by considering nodes whose natural parameters are formed from deterministic *functions* of the states of several parents. This is a key property of the VIBES approach which, as with BUGS, greatly extends its applicability. Suppose we have some conditional distribution $P(X|Y,\ldots)$ and we want to make Y some deterministic function of the states of some other nodes so that $Y = \psi(Z_1,\ldots,Z_M)$. In effect we have a pseudo-parent which is a deterministic function of other nodes, and indeed is represented explicitly through additional deterministic nodes in the graphical interface both in WinBUGS and in VIBES. This will be tractable under VIBES provided the expectation of $u_\psi(\psi)$ can be expressed in terms of the expectations of the corresponding functions $u_j(Z_j)$ of the parents. The pick functions discussed earlier are a special case of these deterministic functions.

Thus for a Gaussian node the mean can be formed from products and sums of the states of other Gaussian nodes provided the function is linear with respect to each of the nodes individually. Similarly, the precision of the Gaussian can comprise the products (but not sums) of any number of Gamma distributed variables.

We also wish to be able to evaluate the lower bound (3), both to confirm the correctness of the variational updates (since the value of the bound should never decrease), as well as to monitor convergence and set termination criteria. This can be done efficiently, largely using quantities which have already been calculated during the variational updates.

4 VIBES: A Software Implementation

Creation of a model in VIBES simply involves drawing the graph (using operations similar to those in a simple drawing package) and then assigning properties to each node such as the functional form for the distribution, a list of which other variables it is conditioned on, and the location of the corresponding data file if the node is observed. The menu of distributions available to the user is dynamically adjusted at each stage to ensure that only valid conjugate

models can be constructed.

As in WinBUGS we have adopted the convention of making logical (deterministic) nodes explicit in the graphical representation as this greatly simplifies the specification and interpretation of the model. We also use the 'plate' notation of a box surrounding one or more nodes to denote that those nodes are replicated some number of times as specified by the parameter appearing in the bottom right hand corner of the box.

Once the model is completed (and the file or files containing the observed variables are specified) it is then 'compiled', which involves allocation of memory for the variables and initializing the distributions Q_i (which is done using simple heuristics but which can also be over-ridden by the user). If desired, monitoring of the lower bound (3) can be switched on (at the expense of slightly increased computation) and this can also be used to set a termination criterion. Alternatively the variational optimization can be run for a fixed number of iterations.

We illustrate VIBES for factorized Q distributions using a Bayesian model for independent component analysis [10] shown in Figure 2. Independent component analysis relies

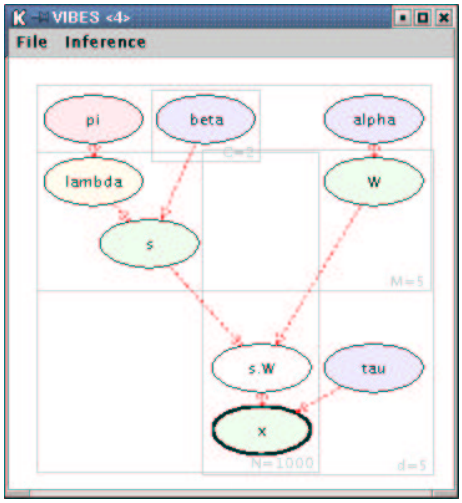

Figure 2: VIBES screen shot for a Bayesian model for independent component analysis. Square boxes correspond to 'plates' in the usual graphical models notation, and indicate multiple replicated copies of the nodes therein. The dashed red links indicates links that will be removed in defining the Q distribution, which will therefore be fully factorized.

on the use of non-Gaussian latent variable distributions, which are defined in this model through a Gaussian mixture representation. The node x has a heavy outline indicating that it represents an observed variable, and the node labelled $s \cdot W$ represents a deterministic function (inner product) of the variables s and W. In this model τ, β and α have Gamma distributions, λ is discrete, π is Dirichlet, while x, W, and s are Gaussians.

5 Structured Variational Distributions

Although the fully factorized variational approximation has been widely used with great success in many applications, it nevertheless represents a somewhat restrictive approximation. It cannot, for instance, capture the posterior correlations between variables. We therefore wish to extend VIBES to allow for a much broader class of variational distributions, which will include the fully factorized distribution as a special case, but which will in general give closer approximations to the true posterior distribution. However, we must ensure that this richer family of distributions remains computationally tractable.

We are particularly interested in Q distributions corresponding to sub-graphs of the original graphical model, since we anticipate that these will capture many of the important dependencies in the P distribution [6, 4]. Our strategy is therefore to take the original graphical model and to delete links as required in order to achieve tractability.

We will see shortly that a particular graphical structure will in general be tractable provided either (1) it comprises only discrete nodes in which the conditional distribution of a node given its parents is given by a pick function over the states of the parents, or (2) it comprises Gaussian nodes each of whose mean is a multi-linear functions of the node's parents. In fact a directed graph of mixed discrete and Gaussian nodes is also tractable provided there are no links representing a Gaussian parent with a discrete child [11], however, we do not consider this more complex case in this paper.

We shall therefore define the Q distribution by removing any links from the original graph other than those which connect two discrete nodes or those which connect two Gaussian nodes in a linear-Gaussian relationship. Note that when a link is removed it is replaced by a corresponding variational parameter. For instance, consider a node represents a Gaussian variable $x \sim \mathcal{N}(\mu, \tau)$ with two parent nodes representing a Gaussian distribution over the mean μ and Gamma distribution over the precision τ. We delete the link from τ to x and in so doing the factor in the Q corresponding to the variable x takes the form $\mathcal{N}(\mu, \lambda)$ where λ is a variational parameter whose value is to be optimized. The link from μ to x is left intact.

Thus in general the Q distribution is described by some number of connected sub-graphs of discrete nodes, along with some number of connected sub-graphs of linear-Gaussian nodes, together with isolated nodes representing Wishart, Dirichlet and other distributions. The user may choose to delete further links within the connected sub-graphs in order to improve the speed of inference, at the expense of some further restriction on the form of the Q distribution. It would be straightforward to assist the user in this process by providing guidance on the expected com-

putational time of each update based on clique sizes.

5.1 Variational Inference

We now derive the generalized variational update equations corresponding to this more complex Q distribution. First of all we note that the distribution is represented by a product of factors, one for each disconnected component in the Q distribution graph. We will denote the disjoint *groups* of nodes associated with these factors by X_α and the corresponding factors in the Q distribution by $Q_\alpha(X_\alpha)$ so that

$$Q(X) = \prod_\alpha Q_\alpha(X_\alpha). \tag{14}$$

For isolated nodes, the analysis of Section 2.1 holds and the corresponding factor in the Q distribution can be updated using (7). This requires of course that the appropriate expectations with respect to other factors in the Q distribution can be evaluated.

Next consider the update of the factor Q_α corresponding to a connected sub-graph. We would like to perform exact inference over this sub-graph, and this can most conveniently be handled by exploiting the junction tree formalism [12, 6]. We therefore take this directed subgraph and moralize it by adding links connecting all pairs of parents for every node, and then dropping the arrows on the links to obtain an undirected graph. The cluster potentials have functional forms governed by the corresponding conditional distributions in the original directed graph, with appropriate variational parameters corresponding to deleted links. We next triangulate the graph and then find a junction tree. This represents a tree-structured cluster graph satisfying the running intersection property.

From (3) and (14) we can dissect out the terms which depend on $Q_\alpha(X_\alpha)$ to give

$$\begin{aligned}\mathcal{L} &= \sum_{X_\alpha} Q_\alpha(X_\alpha) \sum_{\{X_{\beta\neq\alpha}\}} \prod_{\beta\neq\alpha} Q_\beta(X_\beta) \\ &\quad \cdot \left\{ \ln P(X) - \sum_\beta \ln Q_\beta(X_\beta) \right\} \\ &= \sum_\alpha Q_\alpha \left\{ \langle \ln P \rangle_{\beta\neq\alpha} - \ln Q_\alpha \right\} + \text{const.} \end{aligned} \tag{15}$$

where the constant is independent of the variables X_α.

We now write $X_\alpha = \cup_\gamma \mathcal{C}_{\alpha\gamma}$ where $\mathcal{C}_{\alpha\gamma}$ are (in general non-disjoint) clusters of variables. The corresponding factor in the Q distribution is written as a normalized product of cluster potentials

$$Q_\alpha(X_\alpha) = \frac{1}{Z_\alpha} \prod_\gamma \Psi_{\alpha\gamma}(\mathcal{C}_{\alpha\gamma}) \tag{16}$$

where Z_α is the normalization constant. We now substitute (16) into (15) and then pull out the contribution from potential $\Psi_{\alpha\gamma}$ to give

$$\begin{aligned}\mathcal{L} &= \frac{1}{Z_\alpha} \sum_{\mathcal{C}_{\alpha\gamma}} \Psi_{\alpha\gamma} \sum_{X_\alpha \setminus \mathcal{C}_{\alpha\gamma}} \prod_{\rho\neq\gamma} \Psi_{\alpha\rho} \left\{ \langle \ln P(X) \rangle_{\beta\neq\alpha} \right. \\ &\quad \left. - \sum_\rho \ln \Psi_{\alpha\rho} + \ln Z_\alpha \right\} + \text{const.} \end{aligned} \tag{17}$$

We can enforce the normalization constraint on $Q_\alpha(X_\alpha)$ by means of a Lagrange multiplier λ, so we seek to maximize

$$\widetilde{\mathcal{L}} = \mathcal{L} + \lambda \left(\frac{1}{Z_\alpha} \sum_{X_\alpha} \prod_\gamma \Psi_{\alpha\gamma}(\mathcal{C}_{\alpha\gamma}) - 1 \right). \tag{18}$$

In doing so we also note that the distribution $P(X)$ is composed of a product of conditional distributions over the nodes of the original directed graph

$$P(W) = \prod_l P(W_i|\text{pa}_\text{i}). \tag{19}$$

where $W = (X, V)$ as before. We therefore obtain the following stationarity condition

$$\begin{aligned} 0 &= \sum_{X_\alpha \setminus \mathcal{C}_{\alpha\gamma}} \prod_{\rho\neq\gamma} \Psi_{\alpha\rho} \left\{ \sum_i \langle \ln P(W_i|\text{pa}_\text{i}) \rangle_{\beta\neq\alpha} \right. \\ &\quad \left. - \sum_\rho \ln \Psi_{\alpha\rho} + \ln Z_\alpha + 1 \right\}. \end{aligned} \tag{20}$$

We now solve for $\Psi_{\alpha\gamma}$ which appears as one of the terms in the sum over ρ. This gives our final results

$$\begin{aligned} \ln \Psi^\star_{\alpha\gamma} &= \left\langle \sum_{i(\gamma)} \langle \ln P(W_i|\text{pa}_\text{i}) \rangle_{\beta\neq\alpha} - \sum_{\rho_\alpha(\gamma)} \ln \Psi_{\alpha\rho} \right\rangle_{(\alpha\overline{\gamma})} \\ &\quad + \text{const.} \end{aligned} \tag{21}$$

where $i(\gamma)$ denotes the set of all nodes l from the original graph whose conditional distributions $P(W_i|\text{pa}_\text{i})$ have variables which intersect the cluster $\mathcal{C}_{\alpha\gamma}$, and similarly $\rho_\alpha(\gamma)$ denotes the set of clusters $\mathcal{C}_{\alpha\rho}$ which have non-zero intersection with the cluster $\mathcal{C}_{\alpha\gamma}$. Thus again we arrive at a local update scheme, albeit a little more complex than in the fully factorized case.

The expectation in (21) is taken with respect to the conditional distribution defined by

$$Q(\mathcal{C}_{\alpha\gamma}|\{X_{\beta\neq\alpha}\}, X_\alpha \setminus \mathcal{C}_{\alpha\gamma}) = \frac{\prod_{\rho\neq\gamma} \Psi_{\alpha\rho}}{\sum_{X_\alpha \setminus \mathcal{C}_{\alpha\rho}} \prod_{\rho\neq\gamma} \Psi_{\alpha\rho}}. \tag{22}$$

Note that in order to implement these variational update equations we need to be able to compute appropriate expectations and this requires that the cluster potentials be interpretable as (un-normalized) marginals. We can achieve

this using the standard *DistributeEvidence* procedure. For this purpose it is convenient to maintain the separator sets of the junction tree as distinct cluster potentials, and is the essential motivation for using the junction tree representation for the connected sub-graphs. Note that only the cluster potentials, and not the separator potentials, need to be updated using (21). After each of the clusters has been updated once, using an appropriate ordering [6], we will have performed exact inference over the connected sub-graph.

Note that, although inference is performed exactly over the sub-graph, there are situations in which the procedure described so far does not necessarily extract the full marginals. Consider the graph shown in Figure 3. In updating the factor $Q(Y)$ we need to compute the expectation of $P(Y|X_1, X_5)$ with respect to the variational posterior distribution $Q(X_1, X_5)$. As it stands, this will be represented by a product of marginals $Q(X_1)Q(X_5)$. We can extend the formalism to capture correctly the correlation between X_1 and X_5 by adding a link connecting these two nodes thereby ensuring that they are in the same cluster of the junction tree. This can be achieved in general by moralizing nodes such as Y before removing the links.

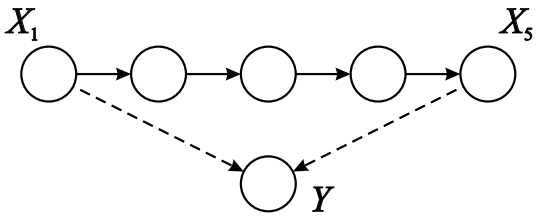

Figure 3: Example showing the need for additional moralization in order to find optimal variational marginals. The Q distribution for this example is defined by the subgraph comprising the Markov chain at the top, with the dashed links removed.

6 Illustration: Bayesian HMM

We have extended the original VIBES software to implement the framework described in the Section 5. Currently we have implemented the case in which any connected subgraph comprises purely discrete nodes. The extension to linear-Gaussian subgraphs, while more complex due to the presence of multi-linear interactions, is analogous to the discrete case. Our implementation is also currently limited to tree-structured sub-graphs, so that the moralization and triangulation steps are not required, although again the required extensions are straightforward.

We illustrate the extended VIBES using a Bayesian hidden Markov model in which we put prior distributions over the probabilities for the initial state of the hidden variables as well as over the transition and emission matrices. This model was described, and also solved variationally, in [13]. In order to highlight the comparison against the structured framework we have allowed all of the variables to be unobserved. The screen shot from VIBES for the directed graph defining the $P(X)$ distribution is shown in Figure 4. As a point of comparison we first solve this model using the fully factorized variational approximation.

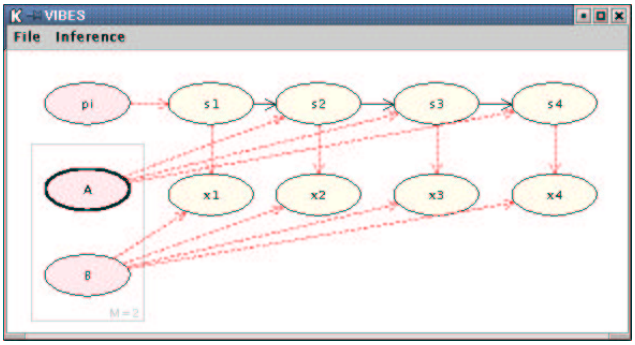

Figure 4: VIBES screen shot showing the graphical model defining the $P(X)$ distribution for a Bayesian hidden Markov model. Links shown in dashed red are those that will be removed in defining the structured Q distribution, while those shown in black will remain.

Next we apply a structured variational approximation as shown in Figure 5. Here the links along the hidden Markov chain are retained, leading to a more flexible class of Q distributions. The converged value of the lower bound $\mathcal{L}$ for the structured distribution of Figure 5 is 3.873 compared to 3.631 for the fully factorized distribution of Figure **??**.

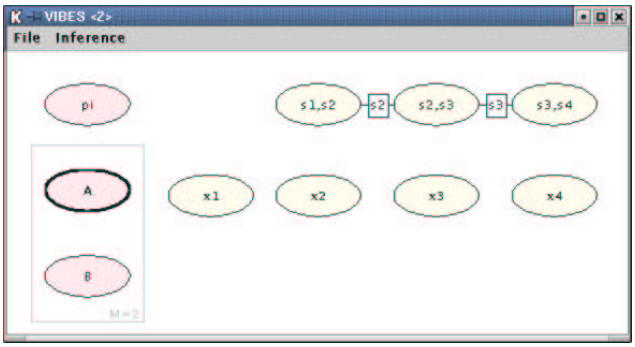

Figure 5: VIBES screen shot of the structured variational distribution, showing the cluster graph (junction tree) .

7 Discussion

Our early experiences with VIBES have shown that it dramatically simplifies the construction and testing of new variational models, and readily allows a range of alternative models to be evaluated on a given problem. We aim to make VIBES freely available to the research community later this year.

Note that this does not encompass all possible tractable substructures. For instance, a Gaussian node having a Gaussian prior for its mean and a Wishart prior for its inverse covariance (precision) matrix is a tractable substructure described by the Normal-Wishart distribution. Also, we could consider Q distributions represented by a tractable graph which is not a sub-graph of the original P distribution graph. We do not consider such possibilities in the present paper.

Finally, there are many possible extensions to the basic VIBES we have described here. For example, in order to broaden the range of models which can be tackled we can combine variational with other methods techniques such as Gibbs sampling or optimization (empirical Bayes) to allow for non-conjugate hyper-priors, for instance.

Acknowledgements

We would like to thank David Spiegelhalter, Zoubin Ghahramani and Matthew Beal for many useful discussions relating to this work.

References

[1] Christopher M. Bishop, David Spiegelhalter, and John Winn. VIBES: A variational inference engine for bayesian networks. In *Advances in Neural Information Processing Systems*, 2002. Accepted for publication.

[2] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 105–162. Kluwer, 1998.

[3] D J Lunn, A Thomas, N G Best, and D J Spiegelhalter. WinBUGS – a Bayesian modelling framework: concepts, structure and extensibility. *Statistics and Computing*, 10:321–333, 2000. http://www.mrc-bsu.cam.ac.uk/bugs/.

[4] L. K. Saul and M. I. Jordan. Exploiting tractable substructures in intractable networks. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 486–492. MIT Press, 1996.

[5] Z. Ghahramani and M. I. Jordan. Factorial hidden Markov models. *Machine Learning*, 29:245–273, 1997.

[6] W. Wiegerinck. Variational approximations between mean field theory and the junction tree algorithm. In *Uncertainty in Artificial Intelligence*. Morgan Kauffmann, 2000.

[7] Z. Ghahramani and M. J. Beal. Propagation algorithms for variational Bayesian learning. In T. K. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13, Cambridge MA, 2001. MIT Press.

[8] H. Attias. A variational Bayesian framework for graphical models. In S. Solla, T. K. Leen, and K-L Muller, editors, *Advances in Neural Information Processing Systems*, volume 12, pages 209–215, Cambridge MA, 2000. MIT Press.

[9] C. M. Bishop. Variational principal components. In *Proceedings Ninth International Conference on Artificial Neural Networks, ICANN'99*, volume 1, pages 509–514. IEE, 1999.

[10] J. W. Miskin and D. J. C. MacKay. Ensemble learning for blind image separation and deconvolution. In M. Girolami, editor, *Advances in Independent Component Analysis*. Springer-Verlag, 2000.

[11] S. L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, 1996.

[12] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Statistics for Engineering and Information Science. Springer-Verlag, 1999.

[13] D. J. C. MacKay. Ensemble learning for hidden Markov models, 1997. Unpublished manuscript, Department of Physics, University of Cambridge.