Artif Intell Rev (2012) 38:85–95
DOI 10.1007/s10462-011-9236-8

A tutorial on variational Bayesian inference

Charles W. Fox · Stephen J. Roberts

Published online: 15 June 2011

Abstract This tutorial describes the mean-field variational Bayesian approximation to inference in graphical models, using modern machine learning terminology rather than statistical physics concepts. It begins by seeking to find an approximate mean-field distribution close to the target joint in the KL-divergence sense. It then derives local node updates and reviews the recent Variational Message Passing framework.

Keywords Variational Bayes · Mean-field · Tutorial

1 Introduction

Variational methods have recently become popular in the context of inference problems, (Attias 2000; Winn and Bishop 2005). *Variational Bayes* is a particular variational method which aims to find some approximate joint distribution $Q(x;\theta)$ over hidden variables x to approximate the true joint $P(x)$, and defines 'closeness' as the KL divergence $KL[Q(x;\theta)||P(x)]$. The mean-field form of VB assumes that Q factorises into single-variable factors, $Q(x) = \prod_i Q_i(x_i|\theta_i)$. The asymetric $KL[Q||P]$ is chosen in VB principally to yield useful computational simplifications, but can be viewed as preferring approximation where areas of high Q are accurate, rather than areas of high P. This is often useful because if we were to draw samples from, or integrate over, Q, then the areas used will be largely accurate (though they may well miss out areas of high P).

C. W. Fox (✉)
Adaptive Behaviour Research Group, University of Sheffield,
Sheffield, UK
e-mail: charles.fox@sheffield.ac.uk

S. J. Roberts
Pattern Analysis and Machine Learning Research Group, Department of Engineering Science,
University of Oxford, Oxford, UK

1.1 Problem statement

We wish to find a set of distributions $\{Q_i(x_i;\theta_i)\}$ to minimise the KL-divergence:

$$KL[Q(x)||P(x|D)] = \int dx.Q(x)\ln\frac{Q(x)}{P(x|D)}$$

where D is the observed data and x are the unobserved variables and

$$Q(x) = \prod_i Q_i(x_i|\theta_i)$$

We sometimes omit the notational dependence of Q_i on θ_i for clarity. As the Q_i are approximate beliefs, they are subject to the normalisation constraints:

$$\forall i. \int dx_i Q_i(x_i) = 1.$$

1.2 VB approximates joints, not marginals

Q approximates the joint, but the individual $Q_i(x_i)$ can be poor approximations to the true marginals $P_i(x_i)$. The $Q_i(x_i)$ components should not be expected to resemble—even remotely— the true marginals, for example when inspecting the computational states of network nodes. This makes VB considerably harder to debug than algorithms whose node states do have some local interpretation: the optimal VB components can be highly counter-intuitive and make sense only in the global context.

Figure 1a shows a graphical model of an extreme example of mean-field VB breaking down, as a warning for the algorithms that follow. Suppose a factory produces screws of unknown diameter μ. We know the machines are very accurate so the precision $\gamma = 1/\sigma^2$ of the diameters is high. However no-one has told us what the mean diameter is, except for a drunken engineer in a pub who said it was 5 mm. We might have a prior belief π that μ is say 5 ± 4 mm (the 4 mm deviation reflecting our low confidence in the report.) Suppose we are given a sealed box containing one screw. What is our belief in its diameter x? The exact *marginal* belief in x should be almost identical to our belief in μ, i.e. 5 ± 4 mm. Figure 1b shows one standard deviation of the true Gaussian joint $P(\mu, x)$ and the best mean-field Gaussian approximate joint $Q(\mu, x)$. As discussed above, this Q is chosen to minimise error *in its own domain* so it appears as a tight Gaussian. Importantly, the marginal $Q_x(x)$ is now very small, and not at all equal to the marginal $P(x)$. We emphasise such cases because we found them to be the major cause of debugging time during development.

We also want to find the model likelihood, $P(D|M)$ for model comparison. This quantity will appear naturally as we try to minimise the KL divergence. There are many ways to think about the following derivations, which we will see involves a balance of three terms called lower bound, energy and entropy. The derivation presented here begins by aiming to minimise the above KL divergence. For example, other equivalent derivations may begin by aiming to maximise the lower bound.

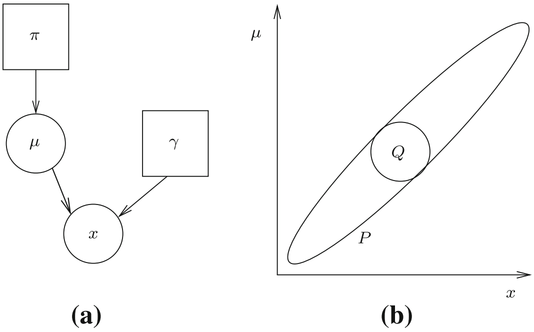


Fig. 1 **a** Graphical model for a population mean problem. Square nodes indicate observed variables. **b** True joint P and VB approximation Q

1.3 Rewriting KL optimisation as an easier problem

We will rewrite the KL equation in terms that are more tractable. First we flip the numerator and denominator, and flip the sign:

$$KL[Q(x)||P(x|D)] = \int dx.Q(x) \ln \frac{Q(x)}{P(x|D)} = -\int dx.Q(x) \ln \frac{P(x|D)}{Q(x)}$$

Next, replace the conditional $P(x|D)$ with a joint $P(x, D)$ and a prior $P(D)$. The reason for making this rewrite is that for Bayesian networks with exponential family nodes, the log $P(x, D)$ term will be a be a very simple sum of node energy terms, whereas log $P(x|D)$ is more complicated. This will simplify later computations.

$$\begin{aligned} P(x, D) &= P(x|D)P(D) \\ &= \ln P(x|D) = \ln P(x, D) - \ln P(D) \\ KL[Q(x)||P(x|D)] &= -\int dx \left(Q(x) \ln \frac{P(x, D)}{Q(x)} - \ln P(D) \right) \end{aligned}$$

The log $P(D)$ term does not involve Q so we can ignore it for the purposes of our minimisation.

Finally, define

$$L[Q(x)] = \int dx \left(Q(x) \ln \frac{P(x, D)}{Q(x)} \right)$$

So that

$$KL[Q(x)||P(x|D)] = -L + \ln P(D)$$

So to minimize the KL divergence, we must *maximize* L.

The maximisation is still subject to the normalisation constraints:

$$\forall i. \int dx_i Q_i(x_i) = 1$$

$L[Q(x)]$ are *lower bounds* on the model log-likelihood, $P(D) = P(D|M)$ (where we generally drop the M notation as we are working with a single model only). The best bound is thus achieved when $L[Q(x)]$ is maximised over Q. The reason for L being a lower bound is seen by rearranging as:

$$\ln P(D) = L[Q(x)] + KL[Q(x|D)||P(x)]$$

Thus when the KL-divergence is zero (a perfect fit), L is equal to the model log-likelihood. When the fit is not perfect, the KL-divergence is always positive and so $L[Q(x)] < \ln P(D)$.

Another rearrangement gives

$$L[Q(x)] = \ln P(D) - KL[Q(x|D)||P(x)]$$

showing that the KL divergence is the error between L and $\ln P(D)$.

1.4 Solution of free energy optimisation

We now wish to find Q to maximise the lower bound, subject to normalisation constraints,

$$\begin{aligned} L[Q(x)] &= \int dx Q(x) \ln \frac{P(x, D)}{Q(x)} \\ &= \int dx Q(x) \log P(x, D) - \int dx Q(x) \ln Q(x) \\ &= \langle E(x, D)\rangle_{Q(x)} + H[Q(x)] \end{aligned}$$

where we define *energy* as $E = \ln P$ and *entropy*[1] as $H[Q(x)] = -\int dx Q(x) \ln Q(x)$. For exponential models, this will become a convenient sum of linear functions. By the mean field assumption:

$$L[Q(x)] = \int dx \left(\prod_i Q_i(x_i)\right) E(x, D) - \int dx \left(\prod_k Q_k(x_k)\right) \sum_i \ln Q_i(x_i) \qquad (1)$$

Consider the entropy (rightmost) term. We can bring out the sum:

$$\sum_i \int dx \left(\prod_k Q_k(x_k)\right) \ln Q_i(x_i)$$

Consider the partitions $x = \{x_i, \bar{x}_i\}$ where $\bar{x}_i = x \backslash x_i$.

$$\begin{aligned} &= \sum_i \int dx_i d\bar{x}_i Q_i(\bar{x}_i) Q_i(x_i) \ln Q_i(x_i) \\ &= \sum_i \left\langle \int dx_i Q_i(x_i) \ln Q_i(x_i) \right\rangle_{Q(\bar{x}_i)} \\ &= \sum_i \int dx_i Q_i(x_i) \ln Q_i(x_i) \end{aligned}$$

Substituting this into the right term of Eq. 1:

$$L[Q(x)] = \int dx \left(\prod_i Q(x_i)\right) E(x, D) - \sum_i \int dx_i . Q(x_i) \ln Q(x_i) \qquad (2)$$

[1] This is Shannon entropy, used by convention.

Now look at and rearrange the *energy* (left) term of Eq. 2, again separating out one variable:

$$\begin{aligned}\int dx\left(\prod_i Q_i(x_i)\right)E(x,D) &= \int dx_i\, Q_i(x_i)\int d\bar{x}_i.Q(\bar{x}_i)E(x,D)\\ &= \int dx_i\, Q_i(x_i)\langle E(x,D)\rangle_{Q(\bar{x}_i)}\\ &= \int dx_i\, Q_i(x_i)\ln\exp\langle E(x,D)\rangle_{Q(\bar{x}_i)}\\ &= \int dx_i\, Q_i(x_i)\ln Q_i^*(x_i)+\ln Z\end{aligned}$$

where we have defined $Q_i^*(x_i)=\frac{1}{Z}\exp\langle E(x,D)\rangle_{Q(\bar{x}_i)}$ and Z normalises $Q_i^*(x_i)$. Substituting this new form of the energy back into Eq. 2 yields

$$L[Q(x)]=\int dx_i\, Q_i(x_i)\ln Q_i^*(x_i)-\sum_i\int dx_i\, Q_i(x_i)\ln Q_i(x_i)+\ln Z$$

Separate out the entropy for $H_i=H[Q_i(x_i)]$ from the rest of the entropy sum:

$$L[Q(x)]=\left\{\int dx_i\, Q_i(x_i)\ln Q_i^*(x_i)-\int dx_i\, Q_i(x_i)\ln Q_i(x_i)\right\}+H[(Q(\bar{x}_i)]+\ln Z$$

Consider the terms in the brackets:

$$\begin{aligned}\int dx_i\, Q_i(x_i)\ln Q_i^*(x_i)-\int dx_i\, Q_i(x_i)\ln Q_i(x_i) &= \int dx_i\, Q_i(x_i)\ln\frac{Q_i^*(x_i)}{Q_i(x_i)}\\ &= -KL[Q_i(x_i)||Q_i^*(x_i)]\end{aligned}$$

What a lucky co-incidence! Though we started by trying to minimise the KL-divergence between large joint distributions (which is hard), we have converted the problem to that of minimising KL-divergences between *individual 1D* distributions (which is easier). Write:

$$L[Q(x)]=-KL[Q_i(x_i)||Q^*(x_i)]+H[Q_i(\bar{x}_i)]+\ln Z$$

Thus L depends on each individual Q_i only through the KL term. We wish to maximise L with respect to each Q_i, subject to the constraint that all Q_i are normalized to unity. This could be achieved by Lagrange multipliers and functional differentiation:

$$\frac{\delta}{\delta Q_i(x_i)}\left\{-KL[Q_i(x_i)||Q^*(x_i)]-\lambda_i\left(\int_{s_i} Q_i(x_i)dx_i-1\right)\right\}:=0$$

A long algebraic derivation would then eventually lead to a Gibbs distribution. However, thanks to the KL form rearrangement we do not need to perform any of this, because we can see immediately that L will be maximised when the KL divergence is zero, hence when

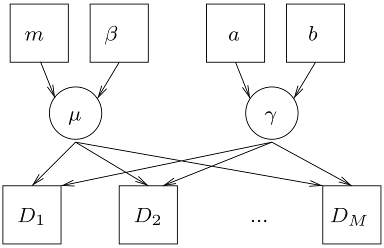
$$Q(x_i)=Q^*(x_i)$$

(The normalisation constraint on Q_i is satisfied thanks to the inclusion of Z in the previous definition of Q_i^*). Expanding back the definition gives the optimal Q_i to be

$$Q(x_i)=\frac{1}{Z}\exp\langle E(x_i,\bar{x}_i,D)\rangle_{Q(\bar{x}_i)}$$

where $E(x_i,\bar{x}_i,D)=\log P(x_i,\bar{x}_i,D)$ is the energy.

Fig. 2 Graphical model for variational example. Square nodes are observed



1.5 Solution as iterative update equations

Converting the equation above into an iterative update equation gives:

$$Q(x_i) \leftarrow \frac{1}{Z} \exp\langle E(x_i, \bar{x}_i, D)\rangle_{Q(\bar{x}_i)}$$

where x_j is a hidden node to be updated; D are the observed data, and $\bar{x_i} = (x \backslash x_i)$ are the other hidden nodes. These updates give us an EM-like algorithm, optimising one node at a time in the hope of converging to a local minimum. Graphical models define conditional independence relationships between nodes, so for models having such structure there exists a *Markov blanket* set of nodes $mb(x_i)$ for each node x_i such that $P(x_i|\bar{x_i}) = P(x_i|mb(x_i))$. For undirected graphical model links, the Markov blanket is the net of neighbouring nodes of x_i; for directed Bayesian networks it is the neighbours of x_i augmented with the set of co-parents of x_i. By the definition of Markov blankets, we may write

$$\ln Q(x_i) \leftarrow \langle \ln P(x_i, mb(x_i)), D\rangle_{Q(mb(x_i))}$$

where $mb(x_i)$ is the Markov blanket of the node x_i of interest.

1.6 Example: variational Bayesian mean update

Until recently, variational implementations have consisted of calculating the iterative update equations by hand for each specific network model. These calculations required much algebra and were error-prone. We consider the simple graphical model shown in Fig. 2: the task is to infer the mean μ component of the posterior joint of a Gaussian population of unknown precision γ given a set of M observations $D = \{D_i\}_{i=1}^{M}$ drawn from this population.

Making the mean-field approximation and assuming conjugate exponential priors we have:

$$Q(x) = Q(\mu)Q(\gamma) = N(\mu; m, \beta^{-1})\Gamma(\gamma; a, b)$$

where m, β, a and b are constants specifying conjugate priors on the population parameters.

The update equation is

$$\ln Q(x_i) \leftarrow \langle \ln P(x_i, mb(x_i)), D\rangle_{Q(mb(x_i))}$$

Substituting the variables for the mean update, $x_i = \mu$:

$$\begin{aligned}\ln Q(\mu) &\leftarrow \int d\gamma \Gamma(\gamma; a, b) \ln P(\mu, \gamma, \{D_i\})\\ &= \int d\gamma \Gamma(\gamma; a, b) \ln N(\mu; m, \beta^{-1}) \Gamma(\gamma; a, b) \prod_i N(D_i|\mu, \gamma)\\ &= \int d\gamma \Gamma(\gamma; a, b) \left\{ \ln N(\mu; m, \beta^{-1}) + \ln \Gamma(\gamma; a, b) + \sum_i \ln N(D_i|\mu, \gamma) \right\}\\ &= \int d\gamma \Gamma(\gamma; a, b) \ln N(\mu; m, \beta^{-1}) + \int d\gamma \Gamma(\gamma; a, b) \ln \Gamma(\gamma; a, b)]\\ &\quad + \sum_i \int d\gamma \Gamma(\gamma; a, b) \ln N(D_i|\mu, \gamma)\end{aligned}$$

This simplifies greatly. First, integrands not dependent on μ can be discarded and replaced by a normalising Z, since we are only interested in the PDF for $Q(\mu)$:

$$\ln Q(\mu) = \int d\gamma \Gamma(\gamma; a, b) \ln N(\mu; m, \beta^{-1}) + \sum_i \int d\gamma \Gamma(\gamma; a, b) \ln N(D_i|\mu, \gamma) + \ln \frac{1}{Z}$$

The left term does depend on μ but not on γ, so the integral over the normalised $Q(\gamma)$ has no effect. Its integral is simply its integrand, so we can write

$$\ln Q(\mu) = \ln N(\mu; m, \beta^{-1}) + \sum_i \int d\gamma \Gamma(\gamma; a, b) \ln N(D_i|\mu, \gamma) + \ln \frac{1}{Z}$$

We next consider the integral containing data D_i. Writing out its log Gaussian energy expression in full, this term becomes

$$\int d\gamma \Gamma(\gamma; a, b) \ln N(D_i|\mu, \gamma) = \int d\gamma \Gamma(\gamma; a, b)[c + (D_i - \mu)\gamma(D_i - \mu)]$$

where c is its normalising constant. Rewriting up to normalisation gives

$$\propto (D - \mu) \left[\int d\gamma \Gamma(\gamma; a, b)\gamma \right] (D - \mu)$$

The required integral is now just the expectation (first moment) of a Gamma distribution, which can be found in standard statistics tables:

$$\langle \gamma \rangle = E[\Gamma(\gamma; a, b)] = ab^{-1}$$

(We notate pre- and post-multiplications by $(D - \mu)$ rather than a single multiplication by $(D - \mu)^2$ so that the multivariate Wishart case follows the same derivation and notation.) The term is now

$$\propto (D - \mu)\langle \gamma \rangle (D - \mu)$$

Substituting this back into the equation for ln $Q(\mu)$ gives

$$\ln Q(\mu) = \ln N(\mu; m, \beta^{-1}) + \sum_i (D_i - \mu)\langle\gamma\rangle(D_i - \mu) + \ln\frac{1}{Z}$$
$$Q(\mu) = \frac{1}{Z}N(\mu; m, \beta^{-1}) \exp\left\{\sum_i (D_i - \mu)\langle\gamma\rangle(D_i - \mu)\right\}$$
$$Q(\mu) = \frac{1}{Z}N(\mu; m, \beta^{-1}) \prod_i N(D_i|\mu, \langle\gamma\rangle)$$

Switching round the dependent variable we obtain the following. (For μ in Gaussian distribution, the flipped result is still Gaussian. For other distributions, it will be conjugate.)

$$Q(\mu) = \frac{1}{Z}N(\mu; m, \beta^{-1}) \prod_i N(\mu|D_i, \langle\gamma\rangle)$$

Finally we can use the standard equation for product of Gaussians to give

$$Q(\mu) = N(\mu|m', \beta'^{-1})$$

with

$$\beta' = \beta + M\langle\gamma\rangle$$
$$m' = \beta'^{-1}\left(\beta m + \langle\gamma\rangle \sum_{i=1}^{M} x_i\right)$$

The above illustrates how the general VB updates can be transformed into particular updates for graphical models. The $Q(\mu)$ component is meaningless by itself as VB aims to approximate the full joint rather than local marginals, so a more useful analysis would repeat the above to obtain the $Q(\gamma)$ updates as well—requiring even more algebra. We see that this process is time-consuming and error prone even for very simple networks such as the Gaussian population used here.

1.7 Variational Bayes with message passing

The previous method of by-hand derivation was time-consuming and tedious, though until recently was state-of-the-art. However the recent variational message passing (VMP) algorithm (Bishop et al. 2002; Winn and Bishop 2005) has shown how to automate these derivations in the case of conjugate-exponential networks. For such networks, the updates all have a standard form, involving the sufficient statistics and natural parameters of the particular node types. For unavoidable non-conjugate-exponential nodes (such as mixture models in particular) it is possible to make further approximations to bring them into the standard form. For conjugate-exponential networks, VMP should now be the standard method for variational Bayesian inference, replacing derivations by hand. For non-conjugate-exponential networks, VMP may still be useful if fast approximations are required at the expense of accuracy.

1.8 VMP details

A standard theorem (Bernardo and Smith 2000) about exponential family distributions shows that the expectation of sufficient statistics are given simply by:

$$\langle u(x_i) \rangle = \bigtriangledown_{\phi} g(\phi)|_{\phi}$$

where the Dell symbol ($\bigtriangledown$) means we form a vector whose rth component is the derivative of g with respect to the rth component of the vector ϕ.

We will write x_i's set of parents as $pa(x_i)$; its set of children as $ch(x_i)$; its set of co-parents with respect to particular child ch as $cop(x_i; ch)$; and its set of co-parents over all children as $cop(x_i)$. We wish to compute the variational Bayesian update as in the previous section:

$$Q_i(x_i) \leftarrow \langle \ln P(x_i, mb(x_i)), D \rangle_{Q(mb(x_i))}$$

Assuming the effects of D are already in the Markov blanket nodes, and separating, this becomes

$$\langle \ln P(pa(x_i)) + \ln P(cop(x_i)) + \ln P(x_i|pa(x_i)) + \ln P(ch(x_i)|x_i, cop(x_i)) \rangle_{Q(mb(x_i))}$$

Simplifying and dropping constant parent and co-parent terms:

$$= \langle \ln P(x_i|pa(x_i)) \rangle_{Q(pa(x_i))} + \langle \ln P(ch(x_i)|x_i, cop(x_i)) \rangle_{Q(ch(x_i),cop(x_i))}$$

The children separate:

$$= \langle \ln P(x_i|pa(x_i)) \rangle_{Q(pa(x_i))} + \sum_{ch \in ch(x_i)} \langle \ln P(ch|x_i, cop(x_i; ch)) \rangle_{Q(ch,cop(x_i;ch))}$$

We will consider the two parts one at a time.

1.8.1 Messages from parents

A conjugate-exponential node x_i is parametrised by a natural parameter vector ϕ_i. By the definition of such nodes,

$$\begin{aligned}\langle \ln P(x_i|pa(x_i)) \rangle_{Q(pa(x_i))} &= \langle \phi_i u(x_i) + f_i(x_i) + g_i(\phi_i) \rangle_{Q(pa(x_i))} \\ &= \langle \phi_i \rangle_{Q(pa(x_i))} u_i(x_i) + f_i(x_i) + \langle g(\phi_i) \rangle_{Q(pa(x_i))}\end{aligned}$$

As ϕ and g are multi-linear functions of the parent sufficient statistics (by construction), and using the mean field assumption, we can simply take their formulæ (defined as conditional on single values of the parents) and substitute the expectations for the sufficient statistics, to get the expectation of the whole expression as required. So parents of x_i need only to send their sufficient statistic expectations to x_i as messages. (A concrete example is shown in Sect. 1.9.)

1.8.2 Messages to parents

A key property of the exponential family is that we can multiply (fuse) similar distributions by adding their natural parameter vectors ϕ:

$$\begin{aligned}&\exp\{\phi_1 u(x_i) + f(x_i) + g(\phi_1)\} \exp\{\phi_1 u(x_i) + f(x_i) + g(\phi_1)\} \\ &\quad = \exp\{(\phi_1 + \phi_2) u(x_i) + f(x_i) + g(\phi_1 + \phi_2)\}\end{aligned}$$

A second property is that by conjugacy, ϕ and g are also multi-linear in parental sufficient statistics. So we can always rearrange the formula by finding functions ϕ_{ij}, f_j, g_{ij} to make it look like a function of a parent $x_j \in pa(x_i)$:

$$\langle \ln P(x_i|pa(x_i))\rangle_{Q(pa(x_i))} = \langle \phi_{ij} u_j(x_j) + f_{ij}(x_j) + g_{ij}(\phi_{ij})\rangle_{Q(pa(x_i))}$$

As before, we may handle the expectation by using the multi-linear property to push all the expectations tight around the sufficient statistics. So from the point of view of the parent, this is written in terms of the sufficient statistic expectations of its child and co-parents. We can thus pass a likelihood message consisting of

$$\phi_{ij}\left(\langle u(x_i)\rangle, \{\langle u(cop)\rangle\}_{cop\in cop(x_j;x_i)}\right)$$

The parent may then simply add these to its prior parameters, by the first property.

Observed data nodes D may be treated as Delta distributions which send standard messages to their parents and children.

1.9 Example: mean, precision and data using VMP

To demonstrate the power of the VMP formalism, we consider the same scenario as used to hand-derive the VB updates in Sect. 1.6. Using VMP we can quickly substitute the particular Gaussian and Gamma distributions into the VMP update equations and quickly obtain *all* network messages for mean, precision, and data nodes as follows. (Messages to D are not required in this particular example, but are useful in general for making inferences about unobserved data.) The exponential forms used here are standard (Bernardo and Smith 2000).

1.9.1 Mean node (with known prior parameters)

Beginning with the conjugate-exponential form of the Gaussian distribution:

$$\ln P(\mu|m,\beta) = \begin{bmatrix} m\beta \\ -\beta/2 \end{bmatrix} \cdot \begin{bmatrix} \mu \\ \mu^2 \end{bmatrix} - g(m,\beta)$$

with

$$g(m,\beta) = \frac{1}{2}(\ln\beta - \beta m^2 - \ln 2\pi)$$

The message to child D is the expectation of sufficient statistics:

$$\left\langle \begin{bmatrix} \mu \\ \mu^2 \end{bmatrix} \right\rangle = \nabla_\phi g(\phi)|_\phi = \begin{bmatrix} \mu \\ \mu^2 + \beta^{-1} \end{bmatrix}$$

1.10 Computing the log-likelihood bound using VMP

VMP also makes computation of the log-likelihood bound simple. Recall that

$$\begin{aligned} \ln P(D|M) &\geq L[Q(x)] \\ L[Q(x)] &= \int dx Q(x) \log \frac{P(x,D)}{Q(x)} \\ &= \langle \log P(x,D)\rangle_{Q(x)} - \langle \log Q(x)\rangle_{Q(x)} \end{aligned}$$

Writing as the sum of individual node contributions from the universe of all (data and hidden) nodes $\Omega = x \cup D$,

$$= \sum_{\omega \in \Omega} \langle \log P(\omega|pa(\omega)) \rangle_{Q(\omega,pa(\omega))} - \langle \log Q(\omega) \rangle_{Q(\omega)} = \sum_{\omega \in \Omega} L_\omega$$

with

$$\begin{aligned} L_\omega &= \langle \log P(\omega|pa(\omega)) \rangle_{Q(\omega,pa(\omega))} - \langle \log Q(\omega) \rangle_{Q(\omega)} \\ &= \langle \phi^\pi_\omega u(\omega) + f(\omega) + g(\phi^\pi_\omega) \rangle_{Q(\omega,pa(\omega))} - \langle \phi^*_\omega u(\omega) + f(\omega) + g(\phi^*_\omega) \rangle_{Q(\omega)} \end{aligned}$$

where ϕ^π are the *prior* natural parameters conditioned only on the parents, and ϕ^* are the posterior parameters, after child message are fused. $Q(\omega)$ uses the posterior parameters. By multi-linearity we can push in the expectations and simplify to obtain:

$$L_\omega = (\langle \phi^\pi_\omega \rangle_{Q(pa(\omega))} + \phi^*_\omega) \langle u(\omega) \rangle_{Q(\omega)} + \langle g(\phi^\pi_\omega) \rangle_{Q(pa(\omega))} + g(\phi^*_\omega)$$

which is simple to compute locally, from each node's received messages. (The term $\langle \phi^\pi_\omega \rangle_{Q(pa(\omega))}$ is computed by substituting in the received parent sufficient statistics expectations into the conjugate-exponential formula for ϕ_ω. By multi-linearity, the expectation can be pushed into the sufficient statistics.) The global model log-likelihood bound is computed by summing the contributions from all nodes, both hidden and observed.

1.11 Summary

In this tutorial we have seen how variational methods may be used to approximate joint posteriors with mean-field distributions. An long-hand calculation of variational inference was shown, then the more general variational message passing framework introduced, which greatly simplifies calculations for conjugate-exponential networks. Variational methods are not appropriate when the marginals or the correlations structure of the joint are required, but are useful in model comparison tasks when the joint is integrated out.

References

Attias H (2000) A variational Bayesian framework for graphical models. In: Advances in neural information processing systems. MIT Press

Bernardo JM, Smith AFM (2000) Bayesian theory. Wiley, London

Bishop CM, Winn JM, Spiegelhalter D (2002) VIBES: a variational inference engine for Bayesian networks. In: Advances in neural information processing systems

Winn J, Bishop C (2005) Variational message passing. J Mach Learn Res 6:661–694