Variational Learning : From exponential families to multilinear systems

Ananth Ranganathan

11th February 2005

Abstract

This note aims to give a general overview of variational inference on graphical models. Starting with the need for the variational approach, we proceed to the derivation of the Variational Bayes EM algorithm that creates distributions on the hidden variables in a graphical model. This leads us to the Variational message Passing algorithm for conjugate exponential families, which is shown to result in a set of updates for the parameters of the distributions involved. The updates form an iterative solution to a multilinear system involving the parameters of the exponential distributions.

1 Introduction

The purpose of variational methods is to provide an approximation to the evidence of a model that contains some hidden variables. Given a model with a set of visible (evidence) variables V and a set of hidden variables H, the evidence of the model is obtained by integrating out the hidden variables from the joint posterior over all the variables

$$P(V) \quad = \quad \int_H P(V,H) \tag{1}$$

However, for most realistic models the joint posterior $P(V,H)$ is intractable and consequently, the integration in (1) cannot be performed analytically. A common approach to overcome this intractability is to use sampling techniques such as Markov Chain Monte Carlo (MCMC) to evaluate the integral. Such techniques, while providing any level of accuracy desired, tend to be slow in general.

Variational methods evaluate the integral (1) by computing an approximating distribution to the actual joint posterior. The approximating distribution is chosen so that the integration becomes easy to perform and can be handled analytically. In the process, a tractable approximating distribution to the posterior over hidden variables $P(H|V)$ is also obtained. In this case it is said that a posterior on hidden variables has been learnt. Hence, variational learning can be seen as almost a side effect of the use of variational techniques to compute the evidence of a model.

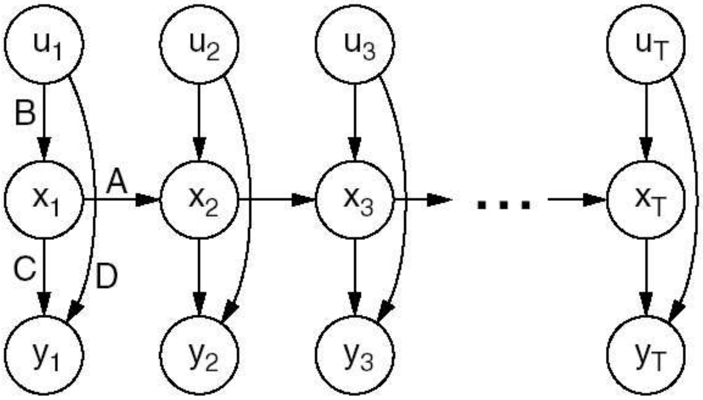

Figure 1: A graphical model for linear dynamical systems with inputs (from [1])

1.1 An example - origin of intractability in Bayesian learning

Consider the Linear Dynamical System (LDS) model shown in Figure 1 where x refers to hidden state variables, y to the observed output measurements, and u to the system inputs. The state evolution and output generation equations are given as

$$\begin{aligned} x_t &= Ax_{t-1} + Bu_t + w_t \\ y_t &= Cx_t + Du_t + v_t \end{aligned}$$

where A and B are matrices determining the state transitions, and C and D are matrices determining the output. v_t is a Gaussian distributed measurement error with mean 0 and covariance R while w_t is a zero mean, unit covariance Gaussian distributed state transition error.

A completely Bayesian model would have priors on all the parameters in the model, namely A,B,C,D and R. In this case, the visible variables are the observations $y_{1:t}$ and the rest of the variables are hidden. The evidence of the model is given as $\int_{\sim y_{1:t}} P(x_{1:t}, y_{1:t}, A, B, C, D, R)$, where $\sim y_{1:t}$ indicates all variables except $y_{1:t}$. Note that the joint posterior (the integrand) contains interaction terms between the variables of upto the *fifth* degree. This can be seen by considering the exponent on the measurement distribution $P(y_t|x_t)$ (ignoring u_t for the moment), which contains the fifth order term $-\frac{1}{2}x_t^T C^T R^{-1} C x_t$. Clearly, it is not feasible to integrate analytically over such highly coupled variables even for this simple and frequently encountered model. Hence, the need for variational approximations.

2 Lower bounding the evidence

Variational methods use Jensen's inequality to construct a lower bound to the log evidence of the model. This is done as follows -

$$\begin{aligned}
\log P(V) &= \log \int_H P(V,H) \\
&= \log \int_H Q(H) \frac{P(V,H)}{Q(H)} \\
&\geq \int_H Q(H) \log \frac{P(V,H)}{Q(H)} \qquad (2)
\end{aligned}$$

where $Q(H)$ is an arbitrary distribution on the hidden variables used to construct the lower bound, and the last relation is obtained using Jensen's inequality. This bound can also be alternately obtained as

$$\begin{aligned}
\log P(V) &= \log \frac{P(V,H)}{P(H|V)} \\
&= \int_H Q(H) \log \frac{Q(H)}{Q(H)} \frac{P(V,H)}{P(H|V)} \\
&= \int_H Q(H) \log \frac{P(V,H)}{Q(H)} + \int_H Q(H) \log \frac{Q(H)}{P(H|V)} \\
&= \mathcal{L}(Q) + \mathrm{KL}(Q||P) \qquad (3)
\end{aligned}$$

where we have made use of the fact that $\int_H Q(H) = 1$. $\mathrm{KL}(Q||P)$ is the KL-divergence between $Q(H)$ and the posterior over hidden variables $P(H|V)$. Since the KL-divergence is always positive (or, of course, zero), $\mathcal{L}(Q)$ is a lower bound on the evidence, which confirms our result from (2).

Note that the sum in the right hand side of (3) is constant since it is the log evidence of the model. Hence, we can proceed to approximate the evidence by minimizing the KL-divergence $\mathrm{KL}(Q||P)$ or by maximizing the lower bound $\mathcal{L}(Q)$, both of which are equivalent and yield the same result. In either case the optimization is performed wrt the $Q(H)$ distribution. The approximate posterior on the hidden variables that is obtained using variational learning is then given by

$$P^{\star}(H|V) = \underset{Q(H)}{\operatorname{argmax}} \; \mathcal{L}(Q) = \underset{Q(H)}{\operatorname{argmin}} \; \mathrm{KL}(Q||P) \qquad (4)$$

In this document, we will take the approach of maximizing the lower bound $\mathcal{L}(Q)$ since this is more popular.

3 Deriving the variational updates

From the preceding discussion, it can be seen that learning the posterior on the hidden variables (and evaluating the evidence) through a variational approximation can be done by maximizing the lower bound on the log evidence. However, maximizing $\mathcal{L}(Q)$

without any constraints on $Q(H)$ is fruitless. To see this, we differentiate $\mathcal{L}(Q)$ wrt $Q(H)$ after adding a lagrangian term that ensures that $Q(H)$ is a proper distrbution, and set the derivative to zero

$$\frac{d}{dQ(H)}\left(\mathcal{L}(Q)+\lambda\left(1-\int_H Q(H)\right)\right) \quad = \quad \log P(V,H)-\log Q(H)-1-\lambda=0 \tag{5}$$

and hence,

$$Q(H) \quad \propto \quad P(V,H)$$

Obviously, the true joint distribution maximizes the bound but this is exactly the distribution that we were trying to avoid having to work with in the first place. Also note that we have not introduced any approximations in the above derivation.

To make $Q(H)$ tractable, we make the assumption, called the *mean field assumption*, that all the variables in the set H are independent, so that $Q(H)$ can be factored into the distributions on the individual variables in H, i.e. $Q(H)=\prod_i Q_i(H_i)\ \forall i\,\text{s.t. } H_i \in H$. In making this assumption, we constrain $Q(H)$ to those distributions that adhere to the independence assumptions.

Given the mean field assumption, we can write the lower bound on the log evidence as

$$\begin{aligned}\mathcal{L}(Q) &= \int_H \left(\prod_i Q_i(H_i)\right)\log\frac{P(V,H)}{\prod_i Q_i(H_i)} \\ &= \int_H \left(\prod_i Q_i(H_i)\right)\log P(V,H)-\sum_i\int_{H_i} Q_i(H_i)\log Q_i(H_i) \qquad (6)\end{aligned}$$

which in turn can be written by separating out the terms in $Q_i(H_i)$ as

$$\mathcal{L}(Q) \quad = \quad \int_H Q_i(H_i)\,\langle\log P(V,H)\rangle_{\sim Q_i(H_i)}-\int_{H_i} Q_i(H_i)\log Q_i(H_i)+\text{other terms} \tag{7}$$

where $\langle\log P(V,H)\rangle_{\sim Q_i(H_i)}$ is the expectation of the log joint posterior wrt to the product of all the Qs except Q_i. We can maximize the bound by performing the maximization wrt to each Q_i individually. Performing the maximization wrt Q_i in exactly the same fashion as (5), we get from (7)

$$Q_i(H_i) \quad = \quad \frac{1}{Z}\exp\langle\log P(V,H)\rangle_{\sim Q_i(H_i)} \tag{8}$$

Note that these equations are coupled since each Q_i depends on all the others. The Q distributions can now be obtained through fixed point iteration on the equations (8).

The optimization performed above is unconstrained regarding the form of the Q distributions. The form of the Q distributions is given by the form of the true joint posterior $P(V,H)$ from (8). However, it is also possible to assume some parametric distribution for the Q_i and constrain the optimization to this family. If the parametric form of the distributions is written as $Q_i(H_i|\lambda_i)$, the optimization performed above can be modified to directly obtain the values of the λ_i, referred to as the *variational parameters,* instead of the Q_i. Such a direct optimization is only tractable if the lower bound in (6) can be computed relatively efficiently. This is turn is possible if the log joint posterior $\log P(V,H)$ is a polynomial and the parametric distributions $Q_i(H_i|\lambda_i)$ have only a few moments. If these conditions are satisfied, variational learning can be reduced to a simultaneous, (possibly) non-linear minimization problem on the variational parameters.

4 Variational message passing

Performing variational learning on a model with joint posterior $P(V,H)$ using (8) according to the above exposition requires the calculation of the expectation of the log posterior wrt to the Q distributions. This is an onerous, time consuming task that has to be performed manually. In addition, if a slight change to the model is required, all the expectations and update equations need to be re-calculated.

Variational message passing seeks to automate the task of performing the variational update for graphical models. The requirement for this to be feasible is that the model should be representable in the form of a Bayes network, i.e we now have additional factorizations available in the true posterior $P(V,H) = \prod_i P(X_i|\text{pa}(i))$, where pa($i$) represents the parents of the X_i variable in the bayes network. Substituting this form of the factored true posterior into the variational update equations (8), we see that the update for Q_i involves only the parents and co-parents of H_i, i.e the Markov blanket of H_i, since all the expectations not involving this variables result in constant terms. (Co-parents of H_i are the variables that share one or more children with H_i). Hence, we get the variational update equations as

$$\begin{aligned} Q_i(H_i) &= \frac{1}{Z}\exp\left\langle \log P(H_i|\text{pa}(i)) + \sum_{k\in\text{ch}(i)} \log P(X_k|H_i,\text{cp}(i,k))\right\rangle_{\sim Q_i(H_i)} + \text{const} \\ &= \frac{1}{Z}\exp\langle \log P(H_i|\text{pa}(i))\rangle_{\sim Q_i(H_i)} + \frac{1}{Z}\sum_{k\in\text{ch}(i)} \exp\langle \log P(X_k|H_i,\text{cp}(i,k))\rangle_{\sim Q_i(H_i)} + \text{const} \end{aligned} \tag{9}$$

where ch(i) refers to the children of H_i and cp(i,k) refers to the co-parents of H_i wrt the child X_k.

From the form of the update equations in (9), it can be seen that the updates admit a message passing algorithm. The first term on the right hand side of (9) is a term involving the parent nodes of H_i while the other terms involve a single child node each. These terms can be seen as messages. The term involving the parents can also be viewed as a prior on H_i while all the other terms from the children together can be said to constitute a likelihood. Hence, the overall optimization problem can be decomposed into a set of local updates on a Bayes network implementable by way of message passing.

5 Conjugate exponential families

We now present a special case of the above exposition for the case where all the distributions on the Bayes net are in the family of exponential distributions and, in addition, distributions on parent nodes are conjugate to distributions on child nodes. It will be seen that this particular special case yields particularly simple forms for the message equations derived in the previous section.

We start with the definition of an exponential family distribution. A distribution is said to be in the exponential family if it can be written in the form

$$p(x|\theta) \quad = \quad h(x)\exp\left(\phi(\theta)^T u(x) + f(x) + g(\phi)\right) \tag{10}$$

where θ are the parameters of the distribution, and ϕ and u are vectors, called the *natural parameters* and *natural statistics* respectively. Almost all commonly encountered distributions belong to the exponential family. For example, the Gaussian distribution, which has the parametric form $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp -\frac{(x-\mu)^2}{2\sigma^2}$, is in the exponential family with natural statistics $\begin{bmatrix} x \\ x^2 \end{bmatrix}$, as can be seen by re-writing the function as $\exp\left(\begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}^T \begin{bmatrix} x \\ x^2 \end{bmatrix} - \frac{1}{2}(\log\sigma^2 + \frac{\mu^2}{\sigma^2} + \log 2\pi)\right)$.

The natural statistics of an exponential family distribution encode its sufficient statistics. For example, the expectation of the natural statistics of the Gaussian wrt the Gaussian give the mean and second central moment - the sufficient statistics of a Gaussian. Also, the natural parameters of any exponential family distrbution can be expressed as a function of the expectation of the natural statistics wrt the distribution.

Conjugacy is an important property found in some members of the exponential family. Two distributions P_X and P_Y are said to be conjugate if they have the same functional form. Since P_X and P_Y are in the exponential family, they can be written as

$$\begin{aligned} P_X(x) &= h(x)\exp\left(\phi_X(\theta_x)^T u_X(x) + f(x) + g(\phi_X)\right) \\ P_Y(y) &= h'(y)\exp\left(\phi_Y(\theta_y)^T u_Y(y) + f'(y) + g'(\phi_Y)\right) \end{aligned}$$

Then conjugacy implies that P_Y can be written in terms of the natural statistics $u_X(x)$ as

$$P_Y(y) \quad = \quad h'(y)\exp\left(\phi_{XY}(x)^T u_X(x) + f'(x,y)\right)$$

where $\phi_{XY}(x)$ can be determined from the form of P_Y. In a Bayes network, conjugacy is imposed on the model by ensuring that distributions on the children of a particular node are always conjugate to the distribution on the node itself.

As an example, consider a Gaussian distribution on a variable X with mean Y and precision β. In addition, let the distribution on Y also be a Gaussian with mean m and precision γ, so that the distributions can be written in exponential family form as

$$\begin{aligned} \log P(X|Y) &= \begin{bmatrix} \beta Y \\ -\frac{\beta}{2} \end{bmatrix}^T \begin{bmatrix} X \\ X^2 \end{bmatrix} + \frac{1}{2}(\log\beta - \beta Y^2 - \log 2\pi) \\ \log P(Y) &= \begin{bmatrix} \gamma m \\ -\frac{\gamma}{2} \end{bmatrix}^T \begin{bmatrix} Y \\ Y^2 \end{bmatrix} + \frac{1}{2}(\log\gamma - \gamma m^2 - \log 2\pi) \end{aligned}$$

where we have used the log distributions to simplify the equations. This is a conjugate model since the Gaussian on X can be re-written in terms of the natural statistics of Y

$$\log P(X|Y) \quad = \quad \begin{bmatrix} \beta X \\ -\frac{\beta}{2} \end{bmatrix}^T \begin{bmatrix} Y \\ Y^2 \end{bmatrix} + \frac{1}{2}(\log \beta - \beta X^2 - \log 2\pi)$$

We are now ready to use conjugacy to derive simpler updates for variational learning. We first provide the general form of the distribution on a variable Y in the Bayes net and the conjugate distribution on one of its children X

$$\begin{aligned} \log P(Y|\mathrm{pa}(Y)) &= \phi_Y(\mathrm{pa}(Y))^T u_Y(Y) + f_Y(Y) + g_Y(\mathrm{pa}(Y)) \\ \log P(X|Y,\mathrm{cp}(Y,X)) &= \phi_{XY}(X,\mathrm{cp}(Y,X))^T u_Y(Y) + \lambda(X,\mathrm{cp}(Y,X)) \end{aligned} \qquad (11)$$

where the form of $\phi_{XY}(X,\mathrm{cp}(Y,X))$ can be determined from the form of $P(X|Y,\mathrm{cp}(Y,X))$.

Substituting the forms of the distributions in (11) into the message passing updates (9), we get the update for the variational distribution Q_Y corresponding to the variable Y as

$$\begin{aligned} \log Q_Y(Y) \quad = \quad & \left\langle \phi_Y(\mathrm{pa}(Y))^T u_Y(Y) + f_Y(Y) + g_Y(\mathrm{pa}(Y)) \right\rangle_{\sim Q_Y(Y)} \\ & + \sum_{k \in \mathrm{ch(Y)}} \left\langle \phi_{XY}(X,\mathrm{cp}(Y,X))^T u_Y(Y) + \lambda(X,\mathrm{cp}(Y,X)) \right\rangle_{\sim Q_Y(Y)} + \mathrm{const} \end{aligned}$$

which can be rearranged to give

$$\begin{aligned} \log Q_Y(Y) \quad = \quad & \left[\left\langle \phi_Y(\mathrm{pa}(Y))^T u_Y(Y) \right\rangle_{\sim Q_Y(Y)} + \sum_{k \in \mathrm{ch(Y)}} \left\langle \phi_{XY}(X,\mathrm{cp}(Y,X))^T u_Y(Y) \right\rangle_{\sim Q_Y(Y)} \right]^T u_Y(Y) \\ & + f_Y(Y) + \mathrm{const} \end{aligned} \qquad (12)$$

and hence, $Q_Y(Y)$ is also in the exponential family but with an updated natural parameter vector given by

$$\phi^{\star}_{Q(Y)} \quad = \quad \langle \phi_Y(\mathrm{pa}(Y)) \rangle_{\sim Q_Y(Y)} + \sum_{k \in \mathrm{ch(Y)}} \langle \phi_{XY}(X,\mathrm{cp}(Y,X)) \rangle_{\sim Q_Y(Y)} \qquad (13)$$

Hence, the messages on the Bayes network can be implemented as expectations of the natural parameter vectors wrt the Q distribution.

6 An example

Consider the simple Gaussian model shown in Figure 2. The data x_n are generated from a Gaussian distribution with mean μ and precision γ. Further, we assume conjugate priors on μ and γ, so that the prior on μ is a Gaussian and the prior on γ is a Gamma distrbution. Hence, we can write the complete model as

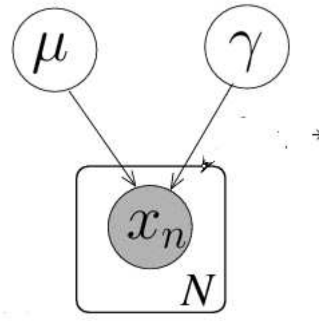

Figure 2: A simple Bayes network

$$
\begin{aligned}
P(x|\mu,\gamma) &= \sqrt{\frac{\gamma}{2\pi}}\exp-\frac{\gamma}{2}(x-\mu)^2 \\
P(\mu|m,\tau) &= \sqrt{\frac{\tau}{2\pi}}\exp-\frac{\tau}{2}(\mu-m)^2 \\
P(\gamma|a,b) &= \frac{b^a}{\Gamma(a)}e^{-\gamma b}\gamma^{a-1}
\end{aligned}
\tag{14}
$$

Now consider the log probability of the Gaussian distribution $P(x|\mu,\gamma)$. This can be written in the standard form of a distribution in the exponential family as

$$
\log P(x|\mu,\gamma) = \begin{bmatrix} \gamma\mu \\ -\frac{\gamma}{2} \end{bmatrix}^T \begin{bmatrix} x \\ x^2 \end{bmatrix} + \frac{1}{2}\left(\log\gamma - \gamma\mu^2 - \log 2\pi\right) \tag{15}
$$

where $\begin{bmatrix} \gamma\mu \\ -\frac{\gamma}{2} \end{bmatrix}$ are the natural parameters and $\begin{bmatrix} x \\ x^2 \end{bmatrix}$ are the natural statistics. We can also re-write this Gaussian distribution in two other ways

$$
\log P(x|\mu,\gamma) = \begin{bmatrix} \gamma x \\ -\frac{\gamma}{2} \end{bmatrix}^T \begin{bmatrix} \mu \\ \mu^2 \end{bmatrix} + \frac{1}{2}\left(\log\gamma - \gamma x^2 - \log 2\pi\right) \tag{16}
$$

$$
\log P(x|\mu,\gamma) = \begin{bmatrix} -\frac{1}{2}(x-\mu)^2 \\ \frac{1}{2} \end{bmatrix}^T \begin{bmatrix} \gamma \\ \log\gamma \end{bmatrix} - \frac{1}{2}\log 2\pi \tag{17}
$$

Note that these are linear functions of the natural statistics of the distributions on μ and γ respectively. Hence, from (15), (16), and (17), $\log P(x|\mu,\gamma)$ is a multilinear function of the natural statistics of x, μ,and γ. While we have demonstrated this property only for this example, it also holds in general. Thus, for any node x in any Bayes network, $\log P\left(x|\mathrm{Pa}(x)\right)$ is a multilinear function of the natural statistics of x and all the members of $\mathrm{Pa}(x)$, the parents of x in the Bayes network. However, this is only true when all the distributions on the parent nodes are conjugate to the $\log P\left(x|\mathrm{Pa}(x)\right)$.

Additionally, also note that the natural parameters of $P(x|\mu,\gamma)$ are multilinear functions of the expectations of the natural statistics of all the dependent variables (the expectations being taken wrt the joint distribution of all the dependent variables). In general, the natural parameters of $P\left(x|\mathrm{Pa}(x)\right)$ are multilinear functions of the expectations of the natural statistics of all the variables in the markov blanket of x, i.e all parents, children and co-parents of x.

We can now perform variational learning in the Bayes network of Figure 2. Let us assume that the x_n are evidence nodes and it is required to learn the distributions on $\mu$and γ. Assuming the same conjugate form of the distributions on these variables as before in the factored mean-field approximation, we see that the learning equations consist of updates to the natural parameters of each distribution, and are given by

$$\begin{aligned}
\phi_\mu &= \begin{bmatrix} \tau m \\ -\frac{\tau}{2} \end{bmatrix} + \sum_{n=1}^{N} \begin{bmatrix} \langle\gamma\rangle x_n \\ -\frac{\langle\gamma\rangle}{2} \end{bmatrix} \\
\phi_\gamma &= \begin{bmatrix} -b \\ a-1 \end{bmatrix} + \sum_{n=1}^{N} \begin{bmatrix} -\frac{1}{2}\left(x_n^2 - 2x_n\langle\mu\rangle + \langle\mu^2\rangle\right) \\ \frac{1}{2} \end{bmatrix}
\end{aligned} \tag{18}$$

where ϕ_μ and ϕ_γ are natural parameters of the distributions on μ and γ respectively.

It is to be kept in mind that after each iteration of the above equations, the expected values of the natural statistics of μ and γ need to be updated. This can easily be done since these are deterministic functions of the natural paremeters. We can get rid of this step by writing (18) in terms of purely the natural parameters of μ and γ (without referencing the expected values of the natural statistics)

$$\begin{aligned}
\begin{bmatrix} \phi_{\mu0} \\ \phi_{\mu1} \end{bmatrix} &= \begin{bmatrix} \tau m \\ -\frac{\tau}{2} \end{bmatrix} + \sum_{n=1}^{N} \begin{bmatrix} -\frac{\phi_{\gamma1}+1}{\phi_{\gamma0}} x_n \\ \frac{\phi_{\gamma1}+1}{2\phi_{\gamma0}} \end{bmatrix} \\
\begin{bmatrix} \phi_{\gamma0} \\ \phi_{\gamma1} \end{bmatrix} &= \begin{bmatrix} -b \\ a-1 \end{bmatrix} + \sum_{n=1}^{N} \begin{bmatrix} -\frac{1}{2}\left(x_n^2 + x_n\frac{\phi_{\mu0}}{\phi_{\mu1}} - \frac{\phi_{\mu0}^2}{4\phi_{\mu1}^2} + \frac{1}{2\phi_{\mu0}}\right) \\ \frac{1}{2} \end{bmatrix}
\end{aligned}$$

For our example, the equations turn out to be part of a non-linear system. This is also true in general and this non-linear system is solved by an iterative scheme in the Variational Message Passing algorithm [2]. In fact, Variational Message Passing turns out to be nothing but the fixed point iterative scheme for solving a nonlinear system of simultaneous equations.

References

[1] M.J. Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[2] John Winn. *Variational Message Passing and its Applications*. PhD thesis, University of Cambridge, 2003.