FMRIB
Oxford Centre for Functional MRI of the Brain

The FMRIB Variational Bayes Tutorial:

Variational Bayesian inference for a non-linear forward model

Michael A. Chappell, Adrian Groves, Mark W. Woolrich

michael.chappell@clenuro.ox.ac.uk

FMRIB Centre,
University of Oxford,
John Radcliffe Hospital,
Headington,
Oxford,
OX3 9DU.

1. Introduction

Bayesian methods have proved powerful in many applications, including MRI, for the inference of model, e.g. physiological, parameters from data. These methods are based on Bayes' theorem, which itself is deceptively simple. However, in practice the computations required are intractable even for simple cases. Hence methods for Bayesian inference are either significantly approximate, e.g. Laplace approximation, or achieve samples from the exact solution at significant computational expense, e.g. Markov Chain Monte Carlo methods. However, more recently the Variational Bayes (VB) method has been proposed (Attias 2000) that facilitates analytical calculations of the posterior distributions over a model. The method makes use of the mean field approximation, making a factorised approximation to the true posterior, although unlike the Laplace approximation does not need to restrict these factorised posteriors to a Gaussian form. Practical implementations of VB typically make use of factorised approximate posteriors and priors that belong to the conjugate-exponential family, making the required integral tractable. The procedure takes an iterative approach resembling an Expectation Maximisation method and whose convergence is guaranteed. Since the method is approximate the computational expense is significantly less than MCMC approaches and is also less than a Laplace approximation since no Hessian need be evaluated.

Attias (2000) provides the original derivation of the 'Variational Bayes Framework for Graphical Models' (although is not the first person to take such an approach). He introduces the concept of 'free-form' selection of the posterior given the chosen model and priors, although this is ultimately limited by the need for the priors and factorised posteriors to belong to the conjugate exponential family (Beal 2003). A comprehensive example of the application of VB to a one-dimensional Gaussian mixture model has been presented by Penny et al. (2000). Beal (2003) has provided a thorough description of variation Bayes and its relationship to MAP and MLE, as well as its application to a number of standard inference problems. He has shown that Expectation Maximisation algorithm is a special case of VB. Friston et al. (2007) additionally has considered the VB approach and variational free energy in the context of the Laplace approximation and ReML. In this context they use a fixed multi-variate Gaussian form for the approximate posterior, in contrast to the 'free-form' approach. To ensure tractablility of the VB approach the models to which it can be applied are limited (Beal 2003). However, Woolrich & Behrens (2006) have avoided this problem, in the context of spatial mixture models, by using a Taylor expansion of the second order. Friston *et al.* (2007) have also applied their variational Laplace method to non-linear models by way of a Taylor expansion, this time assuming that the model is weakly non-linear and hence ignoring second-order and higher terms. Mackay (2003) has provided a brief non-technical introduction to the VB approach and Penny et al. (2003; 2006) have provided a more mathematical introduction specifically for fMRI data and the GLM, with a comparison to the Laplace approximation approach in the later work.

In this report we present a Variational Bayes solution to problems involving non-linear forward models. This takes a similar approach to (Attias 2000), although unlike (Attias 2000; Penny *et al.* 2000; Beal 2003) the factorisation will be over the parameters alone, like for example (Penny *et al.* 2003), since we do not have any hidden nodes in our model. Motivated by the approach of (Woolrich *et al.* 2006) we will extend VB to non-linear models using a Taylor expansion, primarily restricting ourselves, like Friston et al. (2007), by an expansion of the first order. Since the Variational method is iterative, convergence is an important issues and it is found that the guarantees that hold for pure VB do not hold for our non-linear variant, hence convergence is discussed further and the application of a Levenburg-Marquat approach is proposed.

2. Variational Bayes

2.1. Bayesian Inference

The basic Bayesian inference problem is one where we have a series of measurements, $\mathbf{y}$, and we wish to use them to determine the parameters, $\mathbf{w}$, of our chosen model $\mathfrak{M}$. The method is based on Bayes' theorem:

$$P(\mathbf{w} \mid \mathbf{y},\mathfrak{M}) = \frac{P(\mathbf{y},\mathbf{w} \mid \mathfrak{M})}{P(\mathbf{y} \mid \mathfrak{M})} = \frac{P(\mathbf{y} \mid \mathbf{w},\mathfrak{M})P(\mathbf{w} \mid \mathfrak{M})}{P(\mathbf{y} \mid \mathfrak{M})}, \tag{2.1}$$

which gives the *posterior* probability of the parameters given the data and the model, $P(\mathbf{w} \mid \mathbf{y}, \mathfrak{M})$, in terms of: the *likelihood* of the data given the model with parameters **w**, $P(\mathbf{y} \mid \mathbf{w}, \mathfrak{M})$, the *prior* probability of the parameters for this model, $P(\mathbf{w} \mid \mathfrak{M})$, and the *evidence* for the measurements given the chosen model, $P(\mathbf{y} \mid \mathfrak{M})$. We are not too concerned with the correct normalisation of the posterior probability distribution, hence we can neglect the evidence term to give:

$$P(\mathbf{w} \mid \mathbf{y}) \propto P(\mathbf{y} \mid \mathbf{w}) P(\mathbf{w}), \tag{2.2}$$

where the dependence upon the model is implicitly assumed. $P(\mathbf{y} \mid \mathbf{w})$ is calculated from the model and $P(\mathbf{w})$ incorporates prior knowledge of the parameter values and their variability.

For a general model it may not be possible (let alone easy) to evaluate the posterior probability distribution analytically. In which case we might approximate the posterior with a simpler form: $q(\mathbf{w})$, which itself will parameterised by a series of 'hyper-parameters'. We can measure the fit of this approximate distribution to the true on via the free energy:

$$F = \int q(\mathbf{w}) \log \left[\frac{P(\mathbf{y} \mid \mathbf{w}) P(\mathbf{w})}{q(\mathbf{w})} \right] d\mathbf{w} . \tag{2.3}$$

Inferring the posterior distribution $P(\mathbf{w} \mid \mathbf{y})$ is now a matter of estimation of the correct $q(\mathbf{w})$, which is achieved by maximising the free energy over $q(\mathbf{w})$: "Optimising [*F*] produces the best approximation to the true posterior …, as well as the tightest lower bound on the true marginal likelihood" (Attias 2000).

PROOF

Consider the log evidence :

$$\begin{aligned} \log P(\mathbf{y}) &= \log \int P(\mathbf{y} \mid \mathbf{w}) P(\mathbf{w}) d\mathbf{w}, \\ &= \log \int q(\mathbf{w}) \frac{P(\mathbf{y} \mid \mathbf{w}) P(\mathbf{w})}{q(\mathbf{w})} d\mathbf{w}, \\ &\geq \int q(\mathbf{w}) \log \frac{P(\mathbf{y} \mid \mathbf{w}) P(\mathbf{w})}{q(\mathbf{w})} d\mathbf{w}, \end{aligned} \tag{2.4}$$

using Jensen's inequality. This latter quantity is identified from physics as the free energy and the equality holds when $q(\mathbf{w}) = P(\mathbf{w} \mid \mathbf{y})$. Thus the process of seeking the best approximation $q(\mathbf{w})$ becomes a process of maximization of the free energy.

ASIDE

The maximisation of F is equivalent to minimising the Kullback-Liebler (KL distance), also known as the Relative Entropy (Penny *et al.* 2006), between $q(\mathbf{w})$ and the true posterior. Start with the log evidence:

$$\log P(\mathbf{y}) = \log \frac{P(\mathbf{y},\mathbf{w})}{P(\mathbf{w} \mid \mathbf{y})}, \tag{2.5}$$

take the expectation with respect to the (arbitrary) density $q(\mathbf{w})$:

$$\begin{aligned}
&= \int q(\mathbf{w}) \log \frac{P(\mathbf{y},\mathbf{w})}{P(\mathbf{w} \mid \mathbf{y})} d\mathbf{w}, \\
&= \int q(\mathbf{w}) \log \left[\frac{P(\mathbf{y},\mathbf{w})}{P(\mathbf{w} \mid \mathbf{y})} \cdot \frac{q(\mathbf{w})}{q(\mathbf{w})} \right] d\mathbf{w}, \\
&= \int q(\mathbf{w}) \log \frac{P(\mathbf{y},\mathbf{w})}{q(\mathbf{w})} d\mathbf{w} + \int q(\mathbf{w}) \log \frac{q(\mathbf{w})}{P(\mathbf{w} \mid \mathbf{y})} d\mathbf{w}, \\
&= F + KL,
\end{aligned} \tag{2.6}$$

where KL is the KL divergence between $q(\mathbf{w})$ and $P(\mathbf{w} \mid \mathbf{y})$. Since KL satisfies the Gibb's inequality (Mackay 2003) it is always positive, hence F is a lower bound for the log evidence. Thus to achieve a good approximation we either maximise F or minimise KL, only the former being possible in this case.

2.2. *Variational approach*

To make the integrals tractable the variational method chooses mean field approximation for $q(\mathbf{w})$:

$$q(\mathbf{w}) = \prod_i q_{\mathbf{w}_i}(\mathbf{w_i}), \tag{2.7}$$

where we have collected the parameters in $\mathbf{w}$ into separate groups $\mathbf{w}_i$, each with their own approximate posterior distribution $q(\mathbf{w}_i)$. This is the key restriction in the Variational Bayes method, making q approximate. It assumes that the parameters in the separate groups are independent, although we do not require complete factorisation of all the individual parameters (Attias 2000). The computation of $q(\mathbf{w}_i)$ proceeds by the maximisation of $q(\mathbf{w}_i)$ over *F*, by application of the calculus of variations this gives:

$$\log q_{\mathbf{w}_i}(\mathbf{w}_i) \propto \int q_{\mathbf{w}_{\backslash i}}(\mathbf{w}_{\backslash i}) \log P(\mathbf{y} \mid \mathbf{w}) P(\mathbf{w}) d\mathbf{w}_{\backslash i} \tag{2.8}$$

where $\mathbf{w}_{\backslash i}$ refer to the parameters not in the *i*th group. We can write (2.8) in terms of an expectation as:

$$\log q_{\mathbf{w}_i}(\mathbf{w}_i) \propto \left\langle \log P(\mathbf{y} \mid \mathbf{w}) P(\mathbf{w}) \right\rangle_{q_{\mathbf{w}_{\backslash i}}} , \tag{2.9}$$

where $\langle \cdot \rangle_X$ is the expectation of the expression taken with respect to *X*.

PROOF

We wish to maximise the free energy:

$$F = \int q(\mathbf{w}) \log \frac{P(\mathbf{y} \mid \mathbf{w}) P(\mathbf{w})}{q(\mathbf{w})} d\mathbf{w} , \tag{2.10}$$

with respect to each factorised posterior distribution in turn. *F* is a functional (a function of a function), i.e. $F = \int f(\mathbf{w}, q(\mathbf{w})) d\mathbf{w}$, hence to maximise *F* we need to turn to the calculus of variations. We require the maximum of F with respect to a subset of the parameters, $\mathbf{w}_i$, thus we write the functional in terms of these parameters alone as:

$$F = \int g(\mathbf{w}_i, q_{\mathbf{w}_i}(\mathbf{w}_i)) d\mathbf{w}_i ,$$

where:

$$g(\mathbf{w}_i, q_{\mathbf{w}_i}(\mathbf{w}_i)) = \int f(\mathbf{w}, q(\mathbf{w})) d\mathbf{w}_{\backslash i} . \tag{2.11}$$

From variational calculus the maximum of *F* is the solution of the Euler differential equation:

$$\begin{aligned} &\frac{\partial}{\partial q_{\mathbf{w}_i}(\mathbf{w}_i)} \left[g(\mathbf{w}_i, q(\mathbf{w}_i), q'(\mathbf{w}_i)) \right] \\ &- \frac{d}{d\mathbf{w}_i} \left\{ \frac{\partial}{\partial q'_{\mathbf{w}_i}(\mathbf{w}_i)} \left[g(\mathbf{w}_i, q(\mathbf{w}_i), q'(\mathbf{w}_i)) \right] \right\} = 0 , \end{aligned} \tag{2.12}$$

where the second term is zero, in this case, as g is not dependant upon $q'_{\mathbf{w}_i}(\mathbf{w}_i)$. Using equation (2.11) this can be written as[1]:

$$\frac{\partial}{\partial q_{\mathbf{w}_i}(\mathbf{w}_i)}\int q(\mathbf{w})\log\frac{P(\mathbf{y}\,|\,\mathbf{w})P(\mathbf{w})}{q(\mathbf{w})}d\mathbf{w}_{\backslash i}=0. \qquad (2.13)$$

$$=\int q_{\mathbf{w}_{\backslash i}}(\mathbf{w}_{\backslash i})\log P(\mathbf{y}\,|\,\mathbf{w})P(\mathbf{w})d\mathbf{w}_{\backslash i}-\int q_{\mathbf{w}_{\backslash i}}(\mathbf{w}_{\backslash i})\log q_{\mathbf{w}_{\backslash i}}(\mathbf{w}_{\backslash i})d\mathbf{w}_{\backslash i}$$
$$-\int q_{\mathbf{w}_{\backslash i}}(\mathbf{w}_{\backslash i})\log q_{\mathbf{w}_i}(\mathbf{w}_i)d\mathbf{w}_{\backslash i}=0. \ (2.14)$$

Hence:

$$\log q_{\mathbf{w}_i}=\int q_{\mathbf{w}_{\backslash i}}(\mathbf{w}_{\backslash i})\log P(\mathbf{y}\,|\,\mathbf{w})P(\mathbf{w})d\mathbf{w}_{\backslash i}+\text{constant}, \qquad (2.15)$$

which is the result in equation (2.8). Since $q_{\mathbf{w}_i}(\mathbf{w}_i)$ is a probability distribution it should be normalised:

$$q_{\mathbf{w}_i}(\mathbf{w}_i)=\frac{e^{\int I}}{\int e^{\int I}d\mathbf{w}_i}, \qquad (2.16)$$

with $I=\int q_{\mathbf{w}_{\backslash i}}(\mathbf{w}_{\backslash i})\ln P(\mathbf{y}\,|\,\mathbf{w})P(\mathbf{w})d\mathbf{w}_{\backslash i}$. Although often the form of q is chosen (e.g. use of factorised posteriors conjugate to the priors) such that the normalisation is unnecessary. A derivation that incorporates the normalisation, using Lagrange multipliers, is given by (Beal 2003).

2.3. *Conjugate-exponential restriction*

We will take the approach referred to by Attias (2000) as 'free form' optimization, whereby "rather than assuming a specific parametric form for the posteriors, we let them fall out of free-form optimisation of the VB objective function." We will, however, restrict ourselves to priors that are conjugate with the complete data likelihood. The prior is said to be conjugate to the likelihood if and only if (Beal 2003) the posterior (in this case we are interested in the approximate factorised posterior):

$$q_{\mathbf{w}_i}(\mathbf{w}_i)\overset{\sim}{\propto}P(\mathbf{Y}\,|\,\mathbf{w}_i)P(\mathbf{w}_i) \qquad (2.17)$$

[1] Note that this is equivalent to the form used in (Friston *et al.* 2007): $\frac{\partial}{\partial q(\mathbf{w}_i)}\left(\frac{\partial F}{\partial \mathbf{w}_i}\right)=0$

is the same parametric form as the prior. This naturally simplifies the computation of the factorised posteriors, as the VB update becomes a process of updating the posteriors hyper parameters. In general we are restricted by this choice to requiring that our complete data likelihood comes from the exponential family: "In general the exponential families are the only classes of distributions that have natural conjugate prior distributions because they are the only distributions with a fixed number of sufficient statistics apart from some irregular cases" (Beal 2003). Additionally, the advantage of requiring an exponential distribution for the complete data likelihood can be see by examining equation (2.8), where this choice naturally leads to an exponential form for the factorised posterior allowing a tractable VB solution. Hence VB methods typically deal with models which are conjugate-exponential, where setting the requirement that the likelihood come from the exponential family usually allows the conjugacy of the prior to be satisfied. In general the restriction to models whose likelihood is from the exponential family is not restrictive as many models of interest satisfy this requirement (Beal 2003). Neither does this severely limit our choice of priors (which by conjugacy will also need to be from the exponential family), since this still leaves a large family including non-informative distributions as limiting cases (Attias 2000).

We now have a series of equations for the hyper-parameters of each $q_{\mathbf{w}_i}\left(\mathbf{w}_i\right)$ in terms of the parameters of the priors and potentially of the other factorised posteriors. Since the equation for $q_{\mathbf{w}_i}\left(\mathbf{w}_i\right)$ is typically dependent upon the other $\mathbf{w}_i$ the resultant Variational Bayes algorithm follows an EM update procedure: the values for the hyper-parameters are calculated based on the current values, these values are then used for the next iteration and so on until convergence. Since VB is essentially an EM update it is guaranteed to converge (Attias 2000).

3. A simple example: inferring a single Gaussian

The procedure of arriving at a VB algorithm from equation (2.8) is best illustrated by a trivial example. Penny & Roberts (2000) provide the VB update equations for a Gaussian mixture model including inference on the structure of the model, this is a little beyond what we wish to consider here. However, they also provide the results for inferring on a single Gaussian, which we will derive here. We draw measurements from a Gaussian distribution with mean μ and precision β:

$$P(y_n \mid \mu,\beta) = \frac{\sqrt{\beta}}{\sqrt{2\pi}} e^{-\frac{\beta}{2}(y_n-\mu)^2} . \tag{3.1}$$

If we draw N samples that are identically independently distributed (i.i.d.) we have:

$$. \tag{3.2}$$

We wish to infer over the two Gaussian parameters, hence we may factorise our approximate posterior as:

$$q(\mu,\beta) = q(\mu)q(\beta) . \tag{3.3}$$

Thus we need to choose factorised posteriors for both parameters. We restrict ourselves to priors that belong to the conjugate-exponential family; hence we choose prior distributions as normal for μ and Gamma for β. The optimal form for the approximate factorised posteriors is determined by our choice of priors and the requirement of conjugacy, thus we have a Normal distribution over μ and Gamma distribution over β:

$$q(\mu \mid m,v) = \mathrm{N}(\mu;m,v) = \frac{1}{\sqrt{2\pi v}} e^{-\frac{1}{2v}(\mu-m)^2} , \tag{3.4}$$

$$q(\beta \mid b,c) = \mathrm{Ga}(\beta;b,c) = \frac{1}{\Gamma(c)} \frac{\beta^{c-1}}{b^c} e^{-\frac{\beta}{b}} . \tag{3.5}$$

Thus we have four 'hyper-parameters' (m,v,b,c) over the parameters of our posterior distribution. The log factorised-posteriors (which we will need later) are given by:

$$\log q(\mu) = -\frac{(\mu-m)^2}{2v} + \mathrm{const}\{\mu\} , \tag{3.6}$$

$$\log q(\beta) = -\frac{\beta}{b} + (c-1)\log\beta + \mathrm{const}\{\beta\} , \tag{3.7}$$

where const$\{X\}$ contains all terms constant with respect to X. Likewise the log priors are given by:

$$\log P(\mu) = -\frac{(\mu-m_0)^2}{2v_0} + \mathrm{const}\{\mu\} , \tag{3.8}$$

$$\log P(\beta) = -\frac{\beta}{b_0} + (c_0-1)\log\beta + \mathrm{const}\{\beta\} , \tag{3.9}$$

where we have prior values for each of our hyper-parameters denoted by a ‘0’ subscript.

Bayes theorem gives:

$$P(\mu,\beta \mid \mathbf{Y}) \propto P(\mathbf{Y} \mid \mu,\beta)P(\mu)P(\beta), \tag{3.10}$$

which allows us to write down the log posterior up to proportion, which we will need for equation (2.8)

$$\begin{aligned} L &= \log P(\mathbf{Y} \mid \mu, v) + \log P(\mu) + \log P(\beta) + \text{const}\{\mu,\beta\}, \\ &= \frac{N}{2}\log\beta - \frac{\beta}{2}\sum_n (y_n - \mu)^2 - \frac{\beta}{b_0} + (c_0 - 1)\log\beta - \frac{(\mu - m_0)^2}{2v_0} + \text{const}\{\mu,\beta\}. \end{aligned}$$

(3.11)

We are now in a position to derive the updates for μ and β.

3.1. Update on μ

From equation (2.8):

$$\log q(\mu) = \int Lq(\beta)d\beta. \tag{3.12}$$

Performing the integral on the right-hand side:

$$\begin{aligned} &\int Lq(\beta)d\beta, \\ &= \int \left\{ \frac{N}{2}\log\beta - \frac{\beta}{2}\sum_n (y_n - \mu)^2 - \frac{\beta}{b_0} + (c_0 - 1)\log\beta - \frac{(\mu - m_0)^2}{2v_0} \right\} \text{Ga}(\beta;\text{b},\text{c})d\beta, \\ &= -\frac{(\mu - m_0)^2}{2v_0}\int \text{Ga}(\beta;\text{b},\text{c})d\beta + \left(\frac{N}{2} + c_0 - 1\right)\int \log\beta\,\text{Ga}(\beta;b,c)d\beta \\ &-\frac{1}{b_0}\int \beta\text{Ga}(\beta;b,c)d\beta - \frac{1}{2}\sum_n (y_n - \mu)^2 \int \beta\text{Ga}(\beta;\text{b},\text{c})d\beta. \end{aligned}$$

(3.13)

This simplifies by noting that the second and third terms are constant with respect to μ, that the integral of a probability distribution is unity, and that the integral in the final term is simply the first moment of the Gamma distribution. Hence:

$$\int Lq(\beta)d\beta = -\frac{(\mu - m_0)^2}{2v_0} - \frac{bc}{2}\sum_n (y_n - \mu)^2 + \text{const}\{\mu\}. \tag{3.14}$$

Now:

$$\begin{aligned}\sum_n (y_n-\mu)^2 &= N\mu^2 - 2\mu\sum_n y_n + \sum_n y_n^2,\\ &= N\mu^2 - 2\mu s_1 + s_2,\end{aligned} \tag{3.15}$$

hence, using this result and completing the square:

$$\int Lq(\beta)d\beta = -\frac{1+Nv_0bc}{2v_0}\left\{\mu - \frac{m_0+v_0bcs_1}{1+Nv_0bc}\right\}^2 + \text{const}\{\mu\}. \tag{3.16}$$

Comparing coefficients with the expression for the log factorised-posterior finally gives:

$$m = \frac{m_0+v_0bcs_1}{1+Nv_0bc}, \tag{3.17}$$

$$v = \frac{v_0}{1+Nv_0bc}. \tag{3.18}$$

Note that having ignored the terms which are constant in μ we can only define $q(\mu)$ up to scale. If we need the correctly scaled version we can fully account for all the terms in our derivation, alternatively we can normalise our un-scaled q at this stage, as in equation (2.16). Typically finding the update over the hyper-parameters is sufficient, i.e. in this case we are only interested in what the parameters of our distributions become, we don't care about having a correctly scaled distribution.

3.2. Update on β

Likewise we can arrive at the update for β, again starting from (2.8):

$$\begin{aligned}\log q(\beta) &= \int Lq(\mu)d\mu,\\ &= \int\left\{\frac{N}{2}\log\beta - \frac{\beta}{2}\sum_n (y_n-\mu)^2 - \frac{\beta}{b_0} + (c_0-1)\log\beta - \frac{(\mu-m_0)^2}{2v_0}\right\}\mathrm{N}(\mu;\mathrm{m},v)d\mu,\\ &= \frac{N}{2}\log\beta - \frac{\beta}{b_0} + (c_0-1)\log\beta - \frac{\beta}{2}\int\sum_n (y_n-\mu)^2\mathrm{N}(\mu;\mu,v)d\mu + \text{const}\{\beta\},\\ &= \left(\frac{N}{2}+c_0-1\right)\log\beta - \left(\frac{1}{b_0}+\frac{X}{2}\right)\beta,\end{aligned}$$

(3.19)

where X is a function of μ only:

$$\begin{aligned} X &= \int \left(s_2 - 2\mu s_1 + N\mu^2\right) \mathrm{N}(\mu;\mathrm{m},v)\, d\mu, \\ &= s_2 - 2s_1 \int \mu \mathrm{N}(\mu;m,v)\, d\mu + N\int \mu^2 \mathrm{N}(\mu;\mathrm{m},v)\, d\mu, \\ &= s_2 - 2s_1 m + N\left(m^2 + v\right). \end{aligned} \tag{3.20}$$

Comparing coefficients with the log factorised-posterior, equation (3.9), gives the updates for β:

$$\frac{1}{b} = \frac{1}{b_0} + \frac{X}{2}, \tag{3.21}$$

$$c = \frac{N}{2} + c_0 . \tag{3.22}$$

Thus we now have the updates, informed by the data, for the hyper parameters. Hence we can arrive at an estimate for the parameters of our Gaussian distribution. Since the update equations for the hyper-parameters for μ depend on the hyper-parameter values for β and vice versa, these update have to proceed as an iterative process.

3.3. Numerical example

Since this example is sufficiently simple it is possible to plot the factorised approximation to the posterior against the true posterior, as is done in Figure 1. Where 100 samples were drawn from a normal distribution with zero mean and unity variance, and where the following relatively uninformative prior values were used: $m_0 = 0,\ v_0 = 1000,\ b_0 = 1000,\ c_0 = 0.001$. The VB updates were run over 1000 iterations (more than sufficient for convergence) giving estimates for the mean of the distribution as 0.0918 and variance as 1.1990. Figure 2 compares the approximate posterior for μ to the true marginal posterior, showing that as the size of the data increases the approximation improves.

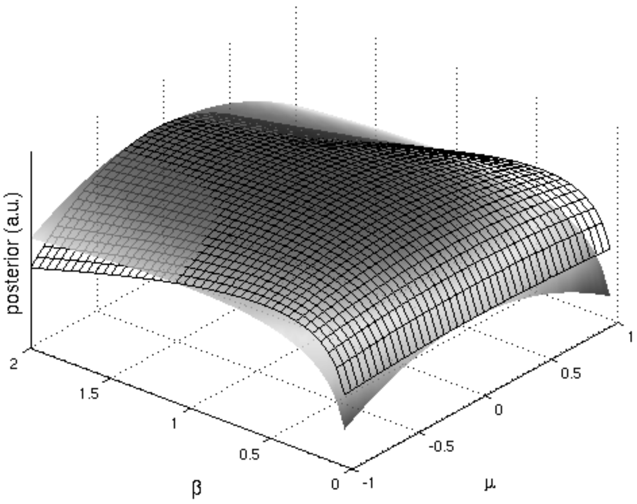

Figure 1: Comparison of (log) true posterior (wireframe) to the factorised approximation (shaded) for VB inference of the parameters of a single Gaussian.

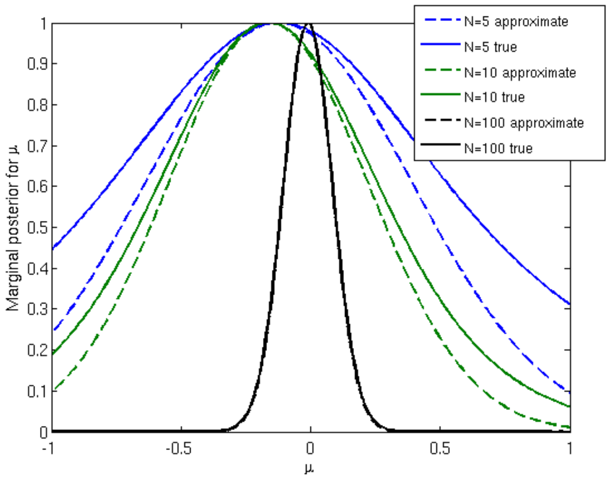

Figure 2: Accuracy of the marginal posterior for μ as the size of the data increases.

3.4. Free energy

The expression for the free energy for this problem is given by (Penny *et al.* 2000):

$$F = L_{av} - KL\left(q(\mu) \parallel p(\mu)\right) - KL\left(q(\beta) \parallel p(\beta)\right), \tag{3.23}$$

where the average likelihood is:

$$L_{av} = 0.5N\left(\psi(c) + \log b\right) - 0.5bc\left(s_2 + N\left(m^2 + v\right) - 2ms_1\right), \tag{3.24}$$

the KL divergence between the factorised posteriors and priors is given by:

$$
\begin{aligned}
KL\left(q(\mu)\,\|\,p(\mu)\right) &= 0.5\log\frac{v_0}{v}+\frac{m^2+m_0^2+v-2mm_0}{2v_0}-0.5,\\
KL\left(q(\beta)\,\|\,p(\beta)\right) &= (c-1)\psi(c)-\log b-c-\log\Gamma(c)+\log\Gamma(c_0)\\
&\quad +c_0\log b_0-(c_0-1)\left(\psi(c_0)+\log b_0\right)+\frac{bc}{b_0},
\end{aligned}
\tag{3.25}
$$

and $\psi(x)$ is the digamma function evaluated at x (see the appendix).

4. Variational Bayes updates for non-linear forward models

Now we can turn to a more useful VB derivation that of inferring the parameters for a non-linear forward model with additive noise. The model for the measurements, y, is

$$\mathbf{y}=\mathbf{g}(\theta)+\mathbf{e}, \tag{4.1}$$

where $\mathbf{g}(\theta)$ is the non-linear forward model for the measurements and $\mathbf{e}$ is additive Gaussian noise with precision ϕ:

$$\mathbf{e}\sim N\left(0,\phi^{-1}\right), \tag{4.2}$$

Hence:

$$P(y_n\,|\,\phi)=\left(\frac{\phi}{2\pi}\right)^{-\frac{1}{2}}e^{-\frac{1}{2}\mathbf{e}^{\mathrm{T}}\phi\mathbf{e}}. \tag{4.3}$$

Thus for N observations we have a log likelihood of:

$$\log P(\mathbf{y}\,|\,\Theta)=\frac{N}{2}\log\phi-\frac{1}{2}\left(\mathbf{y}-\mathbf{g}(\theta)\right)^{\mathrm{T}}\phi\left(\mathbf{y}-\mathbf{g}(\theta)\right), \tag{4.4}$$

where $\Theta=\{\theta,\phi\}$ is the set of all the parameters we wish to infer: those of the model and the noise.

For VB we factorise the approximate posterior separately over the model parameters θ and the noise parameter ϕ:

$$q(\Theta\,|\,\mathbf{y})=q_\theta(\theta\,|\,\mathbf{y})q_\phi(\phi\,|\,\mathbf{y}), \tag{4.5}$$

From here on the subscripts on q will be dropped as the function should be clear from the domain of the function. The following distributions are chosen for the priors:

$$P(\theta)\sim\mathrm{MVN}\left(\theta;\mathbf{m}_0,\Lambda_0^{-1}\right), \tag{4.6}$$

$$P(\phi) \sim \mathrm{Ga}(\phi; s_0, c_0). \tag{4.7}$$

The factorised posteriors are chosen conjugate with the factorised posteriors as:

$$q(\theta \,|\, \mathbf{y}) \sim \mathrm{MVN}(\theta; \mathbf{m}, \Lambda^{-1}), \tag{4.8}$$

$$q(\phi \,|\, \mathbf{y}) \sim \mathrm{Ga}(\phi; s, c). \tag{4.9}$$

Now we can use Bayes' theorem equation (2.1) to get the log-posterior, which we will need to derive the update equations:

$$L = \log P(\Theta \,|\, \mathbf{y}) = \log P(\mathbf{y} \,|\, \Theta) + \log P(\theta) + \log P(\phi) + \mathrm{const}\{\theta, \phi\}, \tag{4.10}$$

where we place any terms that are constant in θ, ϕ in the final term. Hence:

$$\begin{aligned} L = -\frac{1}{2}(\mathbf{y} - \mathbf{g}(\theta))^{\mathrm{T}}(\mathbf{y} - \mathbf{g}(\theta)) + \frac{N}{2}\log\phi \\ -\frac{1}{2}(\theta - \mathbf{m_0})^{\mathrm{T}}\Lambda_0(\theta - \mathbf{m_0}) \\ (c_0 - 1)\log\phi - \frac{1}{s_0}\phi \\ +\mathrm{const}\{\theta, \phi\}. \end{aligned} \tag{4.11}$$

We are now almost ready to use equation (2.8) to derive the updates for the parameters of each factorised posterior distribution. However, L (equation (4.11)) may not produce tractable VB updates for any general non-linear model. In this case we will ensure the tractability by considering a linear approximation of the model. In practice it may not be necessary to restrict ourselves to a purely linear approximation as long as we ensure that the likelihood still belongs to the conjugate-exponential family, we will return to this point later. We approximate $\mathbf{g}(\theta)$ by a first-order Taylor expansion about the mode of the posterior distribution (which for a MVN is also the mean):

$$\mathbf{g}(\theta) \approx \mathbf{g}(\mathbf{m}) + \mathbf{J}(\theta - \mathbf{m}), \tag{4.12}$$

where $\mathbf{J}$ is the Jacobian (matrix of partial derivates):

$$(\mathbf{J})_{x,y} = \left.\frac{\mathrm{d}\left(\mathbf{g}(\theta)_x\right)}{\mathrm{d}\theta_y}\right|_{\theta=\mathbf{m}}. \tag{4.13}$$

This linearization means that we no longer have 'pure' VB. The main consequence of this that the guarantee of convergence for VB no longer applies. The problems associate with convergence will be pursued further in later.

4.1. Parameter update equations

This section summarises the resulting equations for the updates of the parameters that will be derived in detail in the following sections.

Forward model parameters:

$$\Lambda = sc\mathbf{J}^{\mathrm{T}}\mathbf{J} + \Lambda_{\mathbf{0}},$$

$$\Lambda\mathbf{m}_{\text{new}} = sc\mathbf{J}^{\mathrm{T}}\left(\mathbf{k}+\mathbf{J}\mathbf{m}_{\text{old}}\right) + \Lambda_0\mathbf{m}_{\mathbf{0}},$$

Noise precision parameters:

$$c = \frac{N}{2} + c_o,$$

$$\frac{1}{s} = \frac{1}{s_0} + \frac{1}{2}\mathbf{k}^{\mathrm{T}}\mathbf{k}_i + \frac{1}{2}\mathrm{Tr}\left(\Lambda^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{J}\right),$$

where $\mathbf{k} = \mathbf{y} - \mathbf{g}(\mathbf{m})$.

4.2. Updates for forward model parameters

From equation (2.8):

$$\log q(\theta \mid \mathbf{y}) \propto \int Lq(\phi \mid \mathbf{y})\,d\phi, \tag{4.14}$$

The factorised log-posterior is (from (4.8)):

$$\log q(\theta \mid \mathbf{y}) = -\frac{1}{2}\theta^{\mathrm{T}}\Lambda\theta + \frac{1}{2}\theta^{\mathrm{T}}\Lambda\mathbf{m} + \frac{1}{2}\mathbf{m}^{\mathrm{T}}\Lambda\theta + \mathrm{const}\{\theta\}, \tag{4.15}$$

The right-hand side of equation (4.14):

$$\begin{aligned}
&\iint Lq(\phi \mid \mathbf{y})\,d\phi,\\
&= \iint\left(-\frac{1}{2}\phi\left(\mathbf{y}-\mathbf{g}(\theta)\right)^{\mathrm{T}}\left(\mathbf{y}-\mathbf{g}(\theta)\right) - \frac{1}{2}\left(\theta-\mathbf{m}_{\mathbf{0}}\right)^{\mathrm{T}}\Lambda_0\left(\theta-\mathbf{m}_{\mathbf{0}}\right) + \mathrm{const}\{\theta\}\right)\mathrm{Ga}(\phi; s, c)\,d\phi,\\
&= -\frac{1}{2}\left(\theta-\mathbf{m}_{\mathbf{0}}\right)^{\mathrm{T}}\Lambda_0\left(\theta-\mathbf{m}_{\mathbf{0}}\right) - \frac{1}{2}\left(\mathbf{y}-\mathbf{g}(\theta)\right)^{\mathrm{T}}\left(\mathbf{y}-\mathbf{g}(\theta)\right)\int \phi\mathrm{Ga}(\phi; s, c)\,d\phi + \mathrm{const}\{\theta\},\\
&= -\frac{1}{2}\left(\theta-\mathbf{m}_{\mathbf{0}}\right)^{\mathrm{T}}\Lambda_0\left(\theta-\mathbf{m}_{\mathbf{0}}\right) - \frac{1}{2}sc\left(\mathbf{y}-\mathbf{g}(\theta)\right)^{\mathrm{T}}\left(\mathbf{y}-\mathbf{g}(\theta)\right) + \mathrm{const}\{\theta\}.
\end{aligned} \tag{4.16}$$

Now, using the linearization of $\mathbf{g}(\theta)$ from equation (4.12):

$$\begin{aligned}
\mathbf{y}-\mathbf{g}(\theta) &= \mathbf{y}-\mathbf{g}(\mathbf{m}) + \mathbf{J}(\theta-\mathbf{m}),\\
&= \mathbf{k} + \mathbf{J}(\theta-\mathbf{m}),
\end{aligned} \tag{4.17}$$

We can write equation (4.16) as:

$$=-\frac{1}{2}\left\{\theta^{\mathrm{T}}\left(\Lambda_0+sc\mathbf{J}^{\mathrm{T}}\mathbf{J}\right)\theta-\theta^{\mathrm{T}}\left(\Lambda_0\mathbf{m}_0+sc\mathbf{J}\left(\mathbf{k}+\mathbf{Jm}\right)\right)-\left(\Lambda_0\mathbf{m}_0+sc\mathbf{J}\left(\mathbf{k}+\mathbf{Jm}\right)\right)^{\mathrm{T}}\theta\right\},$$

(4.18)

Comparing coefficients with equation (4.15), gives the updates for $\mathbf{m}$ and Λ:

$$\Lambda = sc\mathbf{J}^{\mathrm{T}}\mathbf{J}+\Lambda_{\mathbf{0}}, \quad (4.19)$$

$$\Lambda\mathbf{m}_{\mathrm{new}} = sc\mathbf{J}^{\mathrm{T}}\left(\mathbf{k}+\mathbf{Jm}_{\mathrm{old}}\right)+\Lambda_0\mathbf{m}_{\mathbf{0}}. \quad (4.20)$$

Note that in equation (4.20) the new value of **m** is dependant upon its previous value. This is unlike VB for linear forward models (and all the other updates for this formulation), where the new value for each hyper-parameter is only dependent upon the other hyper-parameters and hyper-parameter priors.

4.3. Updates for the noise precision

For the noise precision posterior distribution we have from equation (2.8):

$$\log q(\phi \mid \mathbf{y}) = \iint Lq(\theta \mid \mathbf{y})d\theta. \quad (4.21)$$

The log-posterior is given by:

$$\log q(\phi \mid \mathbf{y}) = (c-1)\log\phi - \frac{\phi}{s} + \mathrm{const}\{\phi\}, \quad (4.22)$$

and the right-hand side of equation (4.21) as:

$$=\int\left\{-\frac{1}{2}\phi\left(\mathbf{y}-\mathbf{g}(\theta)\right)^{\mathrm{T}}\left(\mathbf{y}-\mathbf{g}(\theta)\right)+\frac{N}{2}\log\phi+\left(c_0-1\right)\log\phi-\frac{1}{s_0}\phi+\mathrm{const}\{\phi\}\right\}q(\theta)d\theta,$$
$$=\left(\frac{N}{2}+c_0-1\right)\log\phi-\frac{1}{s_0}\phi-\frac{1}{2}\phi\int\left(\mathbf{y}-\mathbf{g}(\theta)\right)^{\mathrm{T}}\left(\mathbf{y}-\mathbf{g}(\theta)\right)\mathrm{MVN}(\theta;\mathrm{m},\Lambda)d\theta.$$

(4.23)

Using the linearization as in equation (4.17):

$$\int\left(\mathbf{y}-\mathbf{g}(\theta)\right)^{\mathrm{T}}\left(\mathbf{y}-\mathbf{g}(\theta)\right)\mathrm{MVN}(\theta;\mathrm{m},\Lambda)d\theta,$$
$$=\int\left\{\mathbf{k}^{\mathrm{T}}\mathbf{k}-\cancel{(\theta-\mathbf{m})^{\mathrm{T}}\mathbf{J}^{\mathrm{T}}\mathbf{k}}-\cancel{\mathbf{k}^{\mathrm{T}}\mathbf{J}(\theta-\mathbf{m})}+(\theta-\mathbf{m})^{\mathrm{T}}\mathbf{J}^{\mathrm{T}}\mathbf{J}(\theta-\mathbf{m})\right\}\mathrm{MVN}(\theta)d\theta,$$
$$=\mathbf{k}^{\mathrm{T}}\mathbf{k}+\mathrm{Trace}\left(\Lambda^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{J}\right).$$

(4.24)

where the indicated terms are zero[2] and the following result has been used:

$$\int(\theta-\mathbf{m})^{\mathrm{T}}\mathbf{U}(\theta-\mathbf{m})\mathrm{MVN}\left(\theta;\mathbf{m},\Lambda^{-1}\right)d\theta=\mathrm{Tr}\left(\Lambda^{-1}\mathbf{U}\right)\ \forall\mathbf{U}, \tag{4.25}$$

Hence equation (4.23) becomes:

$$=\left(\frac{N}{2}+c_0-1\right)\log\phi-\frac{1}{s_0}\phi-\frac{1}{2}\phi\left\{\mathbf{k}^{\mathrm{T}}\mathbf{k}+\mathrm{Trace}\left(\Lambda^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{J}\right)\right\}. \tag{4.26}$$

Comparing co-efficients with equation (4.22) gives the following update equations:

$$c=\frac{N}{2}+c_o, \tag{4.27}$$

$$\frac{1}{s}=\frac{1}{s_0}+\frac{1}{2}\mathbf{k}^{\mathrm{T}}\mathbf{k}+\frac{1}{2}\mathrm{Tr}\left(\Lambda^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{J}\right). \tag{4.28}$$

In this case the update for c is not dependant upon the hyper-parameters for θ, hence it does not need to be iteratively determined.

4.4. Numerical Example

A simple example of a non-linear model will now be considered to illustrate the performance of this VB algorithm. The forward model takes the form of a decaying exponential:

$$\mathbf{g}(\theta)=Ae^{-\lambda t} \tag{4.29}$$

where $\theta=\{A,\lambda\}$. Figure 3 shows the fit to the data for two values of noise precision representing relatively large and small quantities of additive noise, values for the parameters used were A=1, λ=1 and ϕ=100 or 10, with 50 data points equally spaced in t between 0 and 5. Figure 4 shows group results for a range of noise precision, at each value 10 sets of data were generated and parameters estimated using VB, the mean value of the estimate is shown along with the mean estimate of the variance. The variance is a measure of the confidence in the estimate and as might be expected increases with increasing noise.

[2] Since, for example, $\int(\theta-\mathbf{m})\mathrm{MVN}(\theta;\mathbf{m},\Lambda)d\theta=\mathbf{m}-\mathbf{m}=\mathbf{0}$.

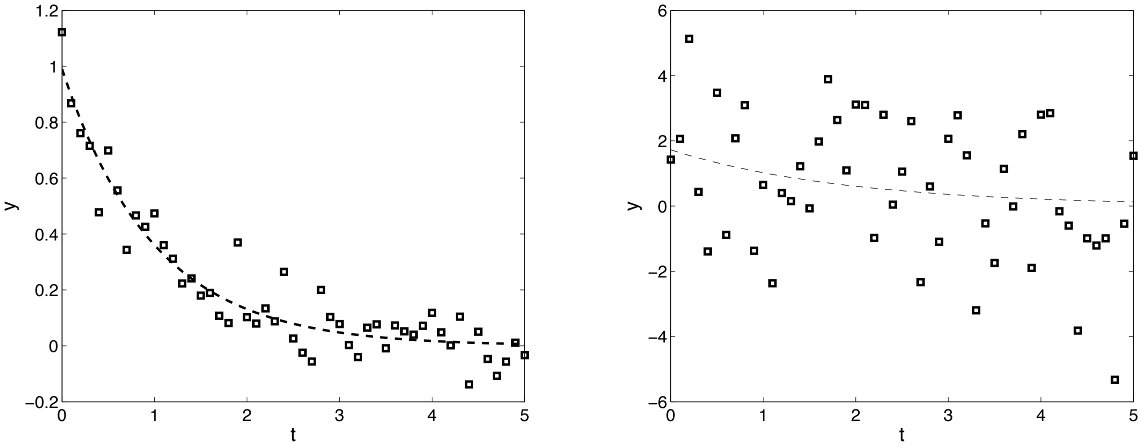

Figure 3: Fit of the VB estimated exponential decay model to simulated data with noise precisions of 100 (left) and 10 (right).

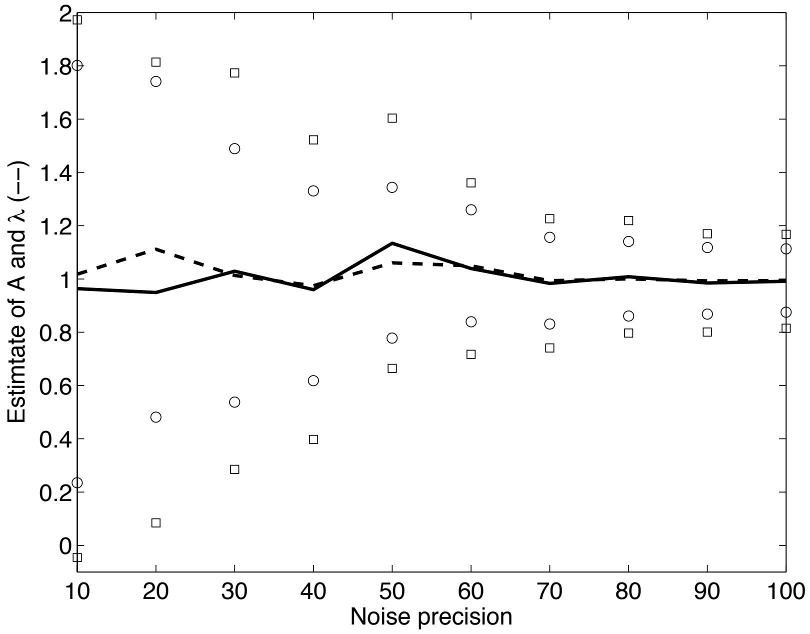

Figure 4: Estimated values of A and λ with varying precision, average of 10 separate data sets at each value. Estimates of the variance are also shown for A (□) and λ (o).

The VB approach can be compared to using a linear regression of the logarithm of the data as a simple estimator. This estimator, however, is ultimately unsatisfactory since through taking the logarithm the Gaussian noise process becomes log-Gaussian and hence sum-of-squares error on the linear regression is no longer optimal. A further problem with this approach is that by taking the log of the data we cannot handle negative values of **g** (which arise as a result of the additive noise).

5. Variational Bayes convergence issues

Convergence of the Variational Bayes iterative updates is guaranteed since it is fundamentally a generalisation of EM (Beal 2003). However, since we here use a Taylor expansion to approximate a non-linear model, such a guarantee of convergence no longer exists, as the model seen by VB is not identical to the true model anymore. If convergence is simply measured by stabilisation of the parameters then it is easy to reach a condition where the iterations alternates between a limited set of solutions without settling to a stable set of values.

A more rigorous method is to monitor the value of *F* and halt the iteration once a maximum has been reached. Alternatively the likelihood multiplied by the priors as an estimate of the posterior may be used since this is easier to calculate. However, the value of F is the correct measure to use since we are aiming to maximise it. The expression for *F* is given by:

$$F = \int q(\mathbf{w}) \ln \frac{P(\mathbf{y},\mathbf{w})}{q(\mathbf{w})} d\mathbf{w}, \tag{5.1}$$

Hence for the non-linear forward model considered here:

$$\begin{aligned} F = &-\frac{sc}{s_0} + \left(\frac{N}{2} + c_0 - 1\right)\left[\log s + \psi(c)\right] \\ &-\frac{1}{2}\left\{(\mathbf{m}-\mathbf{m}_0)^{\mathrm{T}}\Lambda_0(\mathbf{m}-\mathbf{m}_0) + \mathrm{Tr}\left(\Lambda^{-1}\Lambda_0\right)\right\} \\ &-\frac{1}{2}\left\{\mathbf{k}^{\mathrm{T}}\mathbf{k} + \mathrm{Tr}\left(\Lambda^{-1}\mathbf{J}^{\mathrm{T}}\mathbf{J}\right)\right\} - s\log c - \log\Gamma(c) \\ &-c + \left(\frac{N}{2} + c - 1\right)\left[\log s + \psi(c)\right] \\ &+\frac{1}{2}\log\det(\Lambda) + \text{constant}, \end{aligned} \tag{5.2}$$

where $\psi(c)$ is the digamma function as defined in the appendix.

Since the non-linear version of VB is deviates from an EM approach the value of *F* during iteration may pass through a maximum and start to decrease again, this is associated with ill conditioning of the matrix inversion is required in equation (4.20) for the calculation of the means for the model parameters. If the precision matrix is ill conditioned for inversion this can produce spurious solutions that show in an increase in *F*. This may arise because the algorithm makes a step based on the linear approximation to the model, which will be mis-directed in regions where the model is highly non-linear. Often if iteration is allowed to proceed past this point improvement in *F* may recur. Therefore, one approach to reach convergence is to halt only once the value of *F* has decreased at an iteration and has not improved even after a further number of iterations set empirically, i.e. even after a decrease in *F* take a further number of 'trial' steps to test if the algorithm can pass though the problem.

Alternatively the case of an ill condition matrix inversion within an iterative estimation scheme is a well know problem and can be dealt with using Levernburg-Marquat (L-M) approach, e.g. (Andersson 2007; Friston *et al.* 2007). The L-M approach deals with a minimization scheme of the form:

$$\gamma_{\text{new}} = \gamma_{\text{old}} + H^{-1}\delta, \tag{5.3}$$

i.e. an incremental update δ in the parameter γ scaled by an inverted matrix, that typically is a Hessian. If the convergence fails it will typically be because **H** becomes negative definite or poorly conditioned (Andersson 2007). L-M deals with this problem by introducing a scalar, α, initialised to a small value:

$$\gamma_{\text{new}} = \gamma_{\text{old}} + \left(\mathbf{H} + \alpha \cdot \text{diag}[\mathbf{H}]\right)^{-1}\delta. \tag{5.4}$$

If this achieves an improvement in the convergence measure then we accept the new value of γ, if not then we increment α. Ultimately if we do not find an improvement in our convergence measure then we keep incrementing α until the matrix that we are inverting becomes diagonally dominant with large values and hence equation (5.4) reduces to $\gamma_{\text{new}} = \gamma_{\text{old}}$ and we conclude that we cannot find a better solution.

The L-M scheme is implemented in the Variational Bayes on the update for the means of the forward model parameters:

$$\mathbf{m} = \mathbf{m}_{\text{old}} + \left(\Lambda + \alpha \cdot \text{diag}[\Lambda]\right)^{-1}\Delta, \tag{5.5}$$

where:

$$\Delta = \left(\mathbf{J}^{\text{T}}\mathbf{X}\left(\mathbf{y} - \mathbf{g}(\mathbf{m}) + \mathbf{J}\mathbf{m}\right) + \Lambda_0\mathbf{m}_0\right) - \Lambda\mathbf{m}_{\text{old}}. \tag{5.6}$$

If the convergence measure falls, i.e. takes a backward step, then an update according to equation (5.5) is attempted with α = 0.01, during this update all the other (i.e. noise) parameters are not updated. If this results in a reduction in the F then the VB updates proceed with these new values for the forward model parameter means, otherwise α is increased by a factor of 10 and the process repeated until F increases. If no improvement can be found, i.e. α reaches a large value at which no change in $\mathbf{m}$ is called for by equation (5.5), then we halt. During this LM update phase the noise parameters and the value of the model parameter precisions are not updated, only the means alone. We only resume 'normal' VB updates if α returns to its original value.

Essentially by using an LM approach we seek to reduce the size of step made by the algorithm when the Talyor expansion is a poor approximation to the model. However applying LM to the parameter means within VB is an artificial interference into the updates and whilst it will always achieve convergence it seems likely that it will end up in some local minimum. In practice results produced where the LM approach to convergence is applied compares well with those using the 'trial' method described above. However, in a number of cases it is found (by monitoring the value of F) that the 'trial' method produces more optimal solutions and typically with fewer iterations.

6. Acknowledgements

With grateful thanks to Saad and Salima for helpful comments and advice in preparing this tutorial.

7. Appendix – function definitions

The Gamma distribution may be defined as:

$$\mathrm{Ga}(x;b,c)=\frac{1}{\Gamma(c)}\frac{x^{c-1}}{b^c}e^{-\frac{x}{b}}. \qquad (6.1)$$

The di-gamma function is defined as:

$$\psi(x)=\frac{d}{dx}\ln\Gamma(x)=\frac{\Gamma'(x)}{\Gamma(x)}. \qquad (6.2)$$

8. References

Andersson, J. L. R. (2007). Non-Linear Optimisation, Technical Report, TBC, FMRIB Centre,

Attias, H. (2000). A Variational Bayesian Framework for Graphical Models. In proceedings: Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA.

Beal, M. J. (2003). Variational Algorithms for Approximate Bayesian Inference, PhD Thesis, Gatsby Computational Neurosicence Unit, University College London, London.

Friston, K., J. Mattout, et al. (2007). "Variational Free Energy and the Laplace Approximation." NeuroImage **34**(1): 220.

Mackay, D. (2003). Variational Methods. Information Theory, Inference, and Learning Algorithms, CUP**:** 422-436.

Penny, W., S. Kiebel, et al. (2003). "Variational Bayesian Inference for Fmri Time Series." NeuroImage **19**(3): 727.

Penny, W. D., S. Kiebel, et al. (2006). Variational Bayes. ?

Penny, W. D. and S. J. Roberts (2000). Variational Bayes for 1-Dimensional Mixture Models, Technical Report, PARG-2000-01, Department of Engineering Science, Univeristy of Oxford, Oxford.

Woolrich, M. W. and T. E. J. Behrens (2006). "Variational Bayes Inference of Spatial Mixture Models for Segmentation " IEEE Transactions on Medical Imaging **25**(10): 1380-1391.