Chapter 24: Variational Bayes

W. Penny, S. Kiebel and K. Friston

April 13, 2006

Introduction

Bayesian inference can be implemented for arbitrary probabilistic models using Markov Chain Monte Carlo (MCMC) [7]. But MCMC is computationally intensive and so not practical for most brain imaging applications. This chapter describes an alternative framework called 'Variational Bayes (VB)' which is computationally efficient and can be applied to a large class of probabilistic models [27].

The VB approach, also known as 'Ensemble Learning', takes its name from Feynmann's variational free energy method developed in statistical physics. VB is a development from the machine learning community [23, 10] and has been applied in a variety of statistical and signal processing domains [16, 3, 9, 27]. It is now also widely used in the analysis of neuroimaging data [28, 21, 22, 25, 26, 6].

This chapter is structured as follows. We describe the fundamental relationship between model evidence, free energy and Kullback-Liebler (KL) divergence that lies at the heart of VB. Before this we review the salient properties of the KL-divergence. We then describe how VB learning delivers a factorised, minimum KL-divergence approximation to the true posterior density in which learning is driven by an explicit minimisation of the free energy. The theoretical section is completed by relating VB to Laplace approximations and describing how the free energy can also be used as a surrogate for the model evidence, allowing for Bayesian model comparison. The results section then describes simulation studies using models of fMRI data [21]. These are based on a General Linear Model with Auto-Regressive errors, or GLM-AR model.

Theory

In what follows we use upper-case letters to denote matrices and lower-case to denote vectors. $\mathsf{N}(m, \Sigma)$ denotes a uni/multivariate Gaussian with mean m and variance/covariance Σ. X^T denotes the matrix transpose and $\log x$ denotes the natural logarithm.

Kullback-Liebler Divergence

For densities $q(\theta)$ and $p(\theta)$ the Relative Entropy or Kullback-Liebler (KL) divergence from q to p is [5]

$$KL[q||p] = \int q(\theta) \log \frac{q(\theta)}{p(\theta)} d\theta \tag{1}$$

The KL-divergence satisfies the Gibb's inequality [14]

$$KL[q||p] \geq 0 \tag{2}$$

with equality only if $q = p$. In general $KL[q||p] \neq KL[p||q]$, so KL is not a distance measure. Formulae for computing KL, for both Gaussian and Gamma densities, are given in the appendix.

Model evidence and free energy

Given a probabilistic model of some data, the log of the 'evidence' or 'marginal likelihood' can be written as

$$\begin{aligned} \log p(Y) &= \int q(\theta) \log p(Y) d\theta \\ &= \int q(\theta) \log \frac{p(Y,\theta)}{p(\theta|Y)} d\theta \\ &= \int q(\theta) \log \left[\frac{p(Y,\theta) q(\theta)}{q(\theta) p(\theta|Y)} \right] d\theta \\ &= F + KL(q(\theta)||p(\theta|Y)) \end{aligned} \tag{3}$$

where $q(\theta)$ is considered, for the moment, as an arbitrary density. We have

$$F = \int q(\theta) \log \frac{p(Y,\theta)}{q(\theta)} d\theta, \tag{4}$$

which in statistical physics is known as the *negative* free energy. The second term in equation 3 is the KL-divergence between the density $q(\theta)$ and the true posterior $p(\theta|Y)$. Equation 3 is the fundamental equation of the VB-framework and is shown graphically in Figure 5.

Because KL is always positive, due to the Gibbs inequality, F provides a lower bound on the model evidence. Moreover, because KL is zero when two densities are the same, F will become equal to the model evidence when $q(\theta)$ is equal to the true posterior. For this reason $q(\theta)$ can be viewed as an *approximate posterior*.

The aim of VB-learning is to maximise F and so make the approximate posterior as close as possible to the true posterior. This approximate posterior will be the one that best approximates the true posterior in the sense of minimising KL-divergence. We should point out that this divergence cannot be minimised explicitly because $p(\theta|y)$ is only known up to a constant. Instead, it is minimised implicitly by maximising F and by virtue of equation 3. Of course, maximising F, the negative free energy, is the same as minimising $-F$, the free energy.

Factorised Approximations

To obtain a practical learning algorithm we must also ensure that the integrals in F are tractable. One generic procedure for attaining this goal is to assume that the approximating density factorizes over groups of parameters. In physics, this is known as the mean field approximation. Thus, we consider:

$$q(\theta) = \prod_i q(\theta_i) \tag{5}$$

where θ_i is the ith group of parameters. We can also write this as

$$q(\theta) = q(\theta_i)q(\theta_{\setminus i}) \tag{6}$$

where $\theta_{\setminus i}$ denotes all parameters *not* in the ith group. The distributions $q(\theta_i)$ which maximise F can then be derived as follows.

$$\begin{aligned} F &= \int q(\theta) \log \left[\frac{p(Y,\theta)}{q(\theta)} \right] d\theta \qquad (7) \\ &= \int\int q(\theta_i)q(\theta_{\setminus i}) \log \left[\frac{p(Y,\theta)}{q(\theta_i)q(\theta_{\setminus i})} \right] d\theta_{\setminus i} d\theta_i \\ &= \int q(\theta_i) \left[\int q(\theta_{\setminus i}) \log p(Y,\theta) d\theta_{\setminus i} \right] d\theta_i - \int q(\theta_i) \log q(\theta_i) d\theta_i + C \\ &= \int q(\theta_i) I(\theta_i) d\theta_i - \int q(\theta_i) \log q(\theta_i) d\theta_i + C \end{aligned}$$

where the constant C contains terms not dependent on $q(\theta_i)$ and

$$I(\theta_i) = \int q(\theta_{\setminus i}) \log p(Y,\theta) d\theta_{\setminus i} \tag{8}$$

Writing $I(\theta_i) = \log \exp I(\theta_i)$ gives

$$\begin{aligned} F &= \int q(\theta_i) \log \left[\frac{\exp(I(\theta_i))}{q(\theta_i)} \right] d\theta_i + C \qquad (9) \\ &= KL\left[q(\theta_i) || \exp(I(\theta_i))\right] + C \end{aligned}$$

This is minimised when

$$q(\theta_i) = \frac{\exp[I(\theta_i)]}{Z} \tag{10}$$

where Z is the normalisation factor needed to make $q(\theta_i)$ a valid probability distribution. Importantly, this means we are often able to determine the optimal analytic *form* of the component posteriors. This results in what is known as a 'free-form' approximation.

For example, Mackay [13] considers the case of linear regression models with Gamma priors over error precisions, λ, and Gaussian priors over regression coefficients β, with a factorised approximation $q(\beta, \lambda) = q(\beta)q(\lambda)$. Application of equation 10 then leads to an expression in which $I(\lambda)$ has terms in λ and $\log \lambda$ only. From this we can surmise that the optimal form for $q(\lambda)$ is a Gamma density (see appendix).

More generally, free-form approximations can be derived for models from the 'conjugate-exponential' family [9, 27, 1]. Exponential family distributions include Gaussians and discrete multinomials and conjugacy requires the posterior (over a factor) to have the same functional form as the prior.

This allows free-form VB to be applied to arbitrary directed acyclic graphs comprising discrete multinomial variabes with arbitrary subgraphs of univariate and multivariate Gaussian variables. Special cases include Hidden Markov Models, Linear Dynamical Systems, Principal Component Analysers, as well as mixtures and hierarchical mixtures of these. Moreover, by introducing additional variational parameters free-form VB can be applied to models containing

non-conjugate distributions. This includes eg. independent component analysis [1] and logistic regression [11].

Application of equation 10 also leads to a set of update equations for the *parameters* of the component posteriors. This is implemented for the linear regression example by equating the coefficients of λ and $\log \lambda$ with the relevant terms in the Gamma density (see appendix). In the general case, these update equations are coupled as the solution for each $q(\theta_i)$ depends on expectations with respect to the other factors $q(\theta_{\setminus i})$. Optimisation proceeds by initialising each factor and then cycling through each factor in turn and replacing the current distribution with the estimate from equation 10. Examples of these update equations are provided in the following chapter, which applies VB to spatio-temporal models of fMRI data.

Laplace approximations

Laplace's method approximates the integral of a function $\int f(\theta)d\theta$ by fitting a Gaussian at the maximum $\hat{\theta}$ of $f(\theta)$, and computing the volume of the Gaussian. The covariance of the Gaussian is determined by the Hessian matrix of $\log f(\theta)$ at the maximum point $\hat{\theta}$ [12].

The term 'Laplace approximation' is used for the method of approximating a posterior distribution with a Gaussian centered at the Maximum a Posterior (MAP) estimate. This is the application of Laplace's method with $f(\theta) = p(Y|\theta)p(\theta)$. This can be justified by the fact that under certain regularity conditions, the posterior distribution approaches a Gaussian as the number of samples grows [7]. This approximation is derived in detail in chapter 35.

Despite using a full distribution to approximate the posterior, instead of a point estimate, the Laplace approximation still suffers from most of the problems of MAP estimation. Estimating the variances at the end of iterated learning does not help if the procedure has already lead to an area of low probability mass. This point will be illustrated in the results section.

This motivates a different approach where, for nonlinear models, the Laplace approximation is used at each step of an iterative approximation process. This is described in chapters 22 and 35. In fact, this method is an Expectation-Maximisation (EM) algorithm, which is known to be a special case of VB [15]. This is clear from the fact that, at each step of the approximation, we have an ensemble instead of a point estimate.

The relations between VB, EM, iterative Laplace approximations, and an algorithm from classical statistics called Restricted Maximum Likelihood (ReML) are discussed in a recent publication [6]. This algorithm uses a 'fixed-form' for the approximating ensemble, in this case being a full-covariance Gaussian. This is to be contrasted with the 'free-form' VB algorithms described in the previous section, where the optimal form for $q(\theta)$ is derived from $p(Y, \theta)$ and the assumed factorisation.

Model Inference

As we have seen earlier, the negative free energy, F, is a lower bound on the model evidence. If this bound is tight then F can be used as a surrogate for the

model evidence and so allow for Bayesian model selection and averaging[1]. This provides a mechanism for fine-tuning models. In neuroimaging, F has been used to optimise the choice of hemodynamic basis set [20], the order of autoregressive models [21] (see also chapter 40), and the spatial diffusivity of EEG sources (see chapter 26).

Earlier, the negative free energy was written

$$F = \int q(\theta) \log \frac{p(Y,\theta)}{q(\theta)} d\theta \tag{11}$$

By using $p(Y,\theta) = p(Y|\theta)p(\theta)$ we can express it as the sum of two terms

$$F(\theta) = \int q(\theta) \log p(Y|\theta) d\theta - KL[q(\theta)||p(\theta)] \tag{12}$$

where the first term is the average likelihood of the data and the second term is the KL between the approximating posterior and the *prior*. This is not to be confused with the KL in equation 3 which was between the approximate posterior and the true posterior. In equation 12 the KL term grows with the number of model parameters and so penalizes more complex models. Thus, F contains both accuracy and complexity terms, reflecting the two conflicting requirements of a good model, that it fit the data yet be as simple as possible. Model selection principles are also discussed in chapter 35.

In the very general context of probabilistic graphical models, Beal and Ghahramani [2] have shown that the above VB approximation of model evidence is considerably more accurate than the Bayesian Information Criterion (BIC) whilst incurring little extra computational cost. Chapter 35 shows that BIC is a special case of the Laplace approximation. Moreover, it is of comparable accuracy to a much more computationally demanding method based on Annealed Importance Sampling (AIS) [2].

Results

This section first provides an idealised example which illustrates the difference between Laplace and VB approximations. We then present some simulation results showing VB applied to a model of fMRI data.

Univariate densities

Figures 1 and 2 provide an example showing what it means to minimise KL for univariate densities. The solid lines in Figure 1 show a posterior distribution p which is a Gaussian mixture density comprising two modes. The first contains the Maximimum A Posteriori (MAP) value and the second contains the majority of the probability mass.

The Laplace approximation to p is therefore given by a Gaussian centred around the first, MAP mode. This is shown in Figure 1(a). This approximation does not have high probability mass, so the model evidence will be underestimated.

[1]Throughout this chapter our notation has, for brevity, omitted explicit dependence on the choice of model, m. But strictly eg. $p(Y)$, F, $p(\theta|Y)$ and $q(\theta)$ should be written as $p(Y|m)$, $F(m)$, $p(\theta|Y,m)$ and $q(\theta|m)$.

Figure 1(b) shows a Laplace approximation to the second mode, which could arise if MAP estimation found a local, rather than a global, maximum. Finally, Figure 1(c) shows the minimum KL-divergence approximation, assuming that q is a Gaussian. This is the VB solution and corresponds to a density q which is moment matched to p.

Figure 2 plots $KL[q||p]$ as a function of the mean and standard deviation of q, showing a global minimum around the moment-matched values. These KL values were computed by discretising p and q and approximating equation 1 by a discrete sum. The MAP mode, maximum mass mode and moment-matched solutions have $KL[q||p]$ values of 11.7, 0.93 and 0.71 respectively. This shows that low KL is achieved when q captures most of the probability mass of p and, minimum KL when q is moment-matched to p.

Capturing probability mass is particularly important if one is interested in nonlinear functions of parameter values, such as model predictions. This is the case for the Dynamic Causal Models described in later chapters. Figures 3 and 4 show histograms of model predictions for squared and logistic-map functions indicating that VB predictions are qualitatively better than those from the Laplace approximation.

Often in Bayesian inference, one quotes posterior exceedance probabilities. Examples of this are the Posterior Probability Maps described in chapter 23 and Dynamic Causal Models in chapter 41. For the squared function, Laplace says 5% of samples are above $g = 12.2$. But in the true density, 71% of samples are. For the logisitic function 62% are above Laplace's 5% point. The percentage of samples above VB's 5% points are 5.1% for the squared function and 4.2% for the logistic-map function. So for this example, Laplace can tell you the posterior exceedance probability is 5% when, in reality it is an order of magnitude greater. This is not the case for VB.

As we shall see later on, the VB solution depends crucially on our assumptions about q. Either, in terms of the factorisation assumed (this is of course, irrelevant for univariate densities) or the family of approximating densities assumed for q. For example, if q were a mixture density, as in [3], then VB would provide an exact approximation of p. It is also important to note that the differences between VB and Laplace depend on the nature of p. For unimodal p, these differences may are likely to be less significant than those in the above example.

Factorised approximation

We now presents results of a simulation study using a General Linear Model with Auto-Regressive errors, or GLM-AR model. The GLM-AR model can describe both the signal and noise characteristics of fMRI data. This model is used in the rest of the results section. For simplicity, we describe application to data at a single voxel. But the next chapter augments the model with a spatial prior and shows it can be applied to to whole slices of data.

We first illustrate VB's factorised approximation to the posterior and compare the marginal distributions obtained with VB to those from exact evalution. We generated data from a known GLM-AR model

$$y_t = x_t w + e_t \tag{13}$$

$$e_t = a e_{t-1} + z_t \tag{14}$$

where $x_t = 1$ for all t, $w = 2.7$, $a = 0.3$ and $1/\lambda = \mathsf{Var}(z) = \sigma^2 = 4$. We generated $N = 128$ samples. Given any particular values of parameters $\theta = \{w, a, \lambda\}$ it is possible to compute the exact posterior distribution up to a normalisation factor, as

$$p(w, a, \lambda|Y) \propto p(Y|w, a, \lambda)p(w|\alpha)p(a|\beta)p(\lambda) \tag{15}$$

where α is the prior precision of regression coefficients and β is the prior precision of AR coefficients (see next chapter for more details). If we evaluate the above quantity over a grid of values w, a, λ we can then normalise it so it sums to one and so make plots of the exact posterior density. We then assumed an approximate posterior $q(w, a, \lambda) = q(w)q(a)q(\lambda)$ and used VB to fit it to the data. Update equations are available in [21].

Figure 6 compares the exact and approximate posterior joint densities for w, a. In the true posterior it is clear that there is a dependence between w and a but VB's approximate posterior ignores this dependence. Figure 7 compares the exact and approximate posterior marginal densities for w,a and σ^2. In this example, VB has accurately estimated the marginal distributions.

Model inference

We generated data from a larger GLM-AR model having two regression coefficients and three autoregressive coefficients

$$y_t = x_t w + e_t \tag{16}$$

$$e_t = \sum_{j=1}^{m} a_j e_{t-j} + z_t \tag{17}$$

where x_t is a two-element row vector, the first element flipping between a '-1' and '1' with a period of 40 scans (ie. 20 -1's followed by 20 1's) and the second element being '1' for all t. The two corresponding entries in w reflect the size of the activation, $w_1 = 2$, and the mean signal level, $w_2 = 3$. We used an AR(3) model for the errors with parameters $a_1 = 0.8$, $a_2 = -0.6$ and $a_3 = 0.4$. The noise precision was set to $1/\lambda = \mathsf{Var}(z) = \sigma^2 = 1$ and we initially generated $N = 400$ samples. This is a larger model than in the previous example as we have more AR and regression coefficients. An example time series produced by this process is shown in Figure 8(a).

We then generated 10 such time series and fitted GLM-AR(p) models to each using the VB algorithm. In each case the putative model order was varied between $m = 0$ and $m = 5$ and we estimated the model evidence for each. Formulae for the model evidence approximation are available in [21]. Figure 8(b) shows a plot of the average value of the negative free energy, $F(m)$ as a function of m, indicating that the maximum occurs at the true model order.

Gibbs Sampling

Whilst it is possible, in principle, to plot the exact posteriors using the method described previously, this would require a prohibitive amount of computer time for this larger model. We therefore validated VB by comparing it with Gibbs sampling [7, 21].

We generated a number of data sets containing either $N = 40$, $N = 160$ or $N = 400$ scans. At each data set size we compared Gibbs and VB posteriors

for each of the regression coefficients. For the purpose of these comparisons the model order was kept fixed at $m = 3$. Figure 9 shows representative results indicating a better agreement with increasing number of scans. We also note that VB requires more iterations for fewer scans (typically 4 iterations for $N = 400$, 5 iterations for $N = 160$ and 7 iterations for $N = 40$). This is because the algorithm was initialised with an Ordinary Least Squares (OLS) solution which is closer to the VB estimate if there are a large number of scans.

Estimation of effect size

Finally, we generated a number of data sets of various sizes to compare VB and OLS estimates of activation size with the true value of $w_1 = 2$. This comparison was made using a matched-pairs t-test on the absolute estimation error. For $N > 100$ the VB estimation error was significantly smaller for VB than for OLS ($p < 0.05$). For $N = 160$, for example, the VB estimation error was 15% smaller than the OLS error ($p < 0.02$).

Discussion

Variational Bayes delivers a factorised, minimum KL-divergence approximation to the true posterior density and model evidence. This provides a computationally efficient implementation of Bayesian inference for a large class of probabilistic models [27]. It allows for parameter inference, based on the approximating density $q(\theta|m)$ and model inference based on a free energy approximation, $F(m)$ to the model evidence, $p(y|m)$.

The quality of inference provided by VB depends on the nature of the approximating distribution. There are two distinct approaches here. Fixed-form approximations fix the form of q to be eg. a diagonal [10] or full-covariance Gaussian ensemble [6]. Free-form approximations choose a factorisation that depends on $p(Y, \theta)$. These range from fully-factorised approximations, where there are no dependencies in q, to structured approximations. These identify substructures in $p(Y, \theta)$, such as trees or mixtures of trees, in which exact inference is possible. Variational methods are then used to handle interactions between them [8].

VB also delivers an approximation to the model evidence, allowing for Bayesian model comparison. However, it turns out that model selections based on VB are systematically biased towards simpler models [2]. Nevertheless, they have been shown empirically to be more accurate than BIC approximations and faster than sampling approximations [2]. Bayesian model selection is discussed further in chapter 35.

Chapter 24 described a Parametric Empirical Bayes (PEB) algorithm for inference in hierarchical linear Gaussian models. This algorithm may be viewed as a special case of VB with a fixed-form full-covariance Gaussian ensemble [6]. More generally, however, VB can be applied to models with discrete as well as continuous variables.

A classic example here is the Gaussian mixture model. This has been applied to an analysis of intersubject variability in fMRI data. Model comparisons based on VB identified two overlapping degenerate neuronal systems in subjects performing a crossmodal priming task [18].

In the dynamic realm, VB has been used to fit and select Hidden Markov Models (HMMs) for the analysis of EEG data [4]. These HMMs use discrete variables to enumerate the hidden states and continuous variables to parameterise the activity in each. Here, VB identifies the number of stationary dynamic regimes, when they occur, and describes activity in each with a Multivariate Autoregressive (MAR) model. The application of VB to MAR models is described further in chapter 40.

The following chapter uses a spatio-temporal model for the analysis of fMRI. This includes spatial regularisation of the autoregressive processes which characterise fMRI noise. This regularistion requires a prior over error terms which is precluded in chapter 22's PEB framework but is readily accomodated using free-form VB.

Appendix

For univariate Normal densities $q(x) = \mathsf{N}(\mu_q, \sigma_q^2)$ and $p(x) = \mathsf{N}(\mu_p, \sigma_p^2)$ the KL-divergence is

$$KL_{N_1}(\mu_q, \sigma_q; \mu_p, \sigma_p) = 0.5 \log \frac{\sigma_p^2}{\sigma_q^2} + \frac{\mu_q^2 + \mu_p^2 + \sigma_q^2 - 2\mu_q\mu_p}{2\sigma_p^2} - 0.5 \tag{18}$$

The multivariate Normal density is given by

$$\mathsf{N}(\mu, \Sigma) = (2\pi)^{-d/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \tag{19}$$

The KL divergence for Normal densities $q(x) = \mathsf{N}(\mu_q, \Sigma_q)$ and $p(x) = \mathsf{N}(\mu_p, \Sigma_p)$ is

$$\begin{aligned} KL_N(\mu_q, \Sigma_q; \mu_p, \Sigma_p) &= 0.5 \log \frac{|\Sigma_p|}{|\Sigma_q|} + 0.5Tr(\Sigma_p^{-1}\Sigma_q) \\ &+ 0.5(\mu_q - \mu_p)^T\Sigma_p^{-1}(\mu_q - \mu_p) - \frac{d}{2} \end{aligned} \tag{20}$$

where $|\Sigma_p|$ denotes the determinant of the matrix Σ_p.

The Gamma density is defined as

$$Ga(b, c) = \frac{1}{\Gamma(c)} \frac{x^{c-1}}{b^c} \exp\left(\frac{-x}{b}\right) \tag{21}$$

The log of the gamma density

$$\log Ga(b, c) = -\log\Gamma(c) - c\log b + (c-1)\log x - \frac{x}{b} \tag{22}$$

In [13], application of equation 10 for the approximate posterior over the error precision $q(\lambda)$ leads to an expression containing terms in λ and $\log\lambda$ only. This identifies $q(\lambda)$ as a Gamma density. The coefficients of these terms are then equated with those in the above equation to identify the parameters of $q(\lambda)$.

For Gamma densities $q(x) = Ga(\boldsymbol{x}; b_q, c_q)$ and $p(x) = Ga(\boldsymbol{x}; b_p, c_p)$ the KL-divergence is

$$\begin{aligned} KL_{Ga}(b_q, c_q; b_p, c_p) &= (c_q - 1)\Psi(c_q) - \log b_q - c_q - \log\Gamma(c_q) \\ &+ \log\Gamma(c_p) + c_p \log b_p - (c_p - 1)(\Psi(c_q) + \log b_q) + \frac{b_q c_q}{b_p} \end{aligned} \tag{23}$$

where $\Gamma()$ is the gamma function and $\Psi()$ the digamma function [24]. Similar equations for multinomial and Wishart densities are given in [19].

References

[1] H. Attias. Inferring Parameters and Structure of Latent Variable Models by Variational Bayes. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 1999.

[2] M. Beal and Z. Ghahramani. The Variational Bayesian EM algorithms for incomplete data: with application to scoring graphical model structures. In J. Bernardo, M. Bayarri, J. Berger, and A. Dawid, editors, *Bayesian Statistics 7*. Cambridge University Press, 2003.

[3] C.M. Bishop, N. Lawrence, T.S. Jaakkola, and M.I. Jordan. Approximating posterior distributions in belief networks using mixtures. In M.J. Kearns M.I. Jordan and S.A. Solla, editors, *Advances in Neural Information Processing Systems 10*, 1998.

[4] M.J. Cassidy and P. Brown. Hidden Markov based autoregressive analysis of stationary and non-stationary electrophysiological signals for functional coupling studies. *Journal of Neuroscience Methods*, 116(35), 2002.

[5] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley, 1991.

[6] K. Friston, J. Mattout, N. Trujillo-Barreto, J. Ashburner, and W. Penny. Variational free energy and the Laplace approximation. 2006. Submitted.

[7] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Chapman and Hall, Boca Raton, 1995.

[8] Z. Ghahramani. On Structured Variational Approximations. Technical report, Gatsby Computational Neuroscience Unit, UCL, UK, 2002.

[9] Z. Ghahramani and M.J. Beal. Propagation algorithms for Variational Bayesian learning. In T. Leen et al, editor, *NIPS 13*, Cambridge, MA, 2001. MIT Press.

[10] G.E. Hinton and D. van Camp. Keeping neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, pages 5–13, 1993.

[11] T. S. Jaakkola and M. I. Jordan. A variational approach to Bayesian logistic regression models and their extensions. Technical Report 9702, MIT Computational Cognitive Science, January 1997.

[12] D. J. C. MacKay. Choice of basis for Laplace approximations. *Machine Learning*, 33:77–86, 1998.

[13] D.J.C. Mackay. Ensemble Learning and Evidence Maximization. Technical report, Cavendish Laboratory, University of Cambridge, 1995.

[14] D.J.C Mackay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, Cambridge, 2003.

[15] M.Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[16] T.S. Jaakola M.I. Jordan, Z. Ghahramani and L.K. Saul. An Introduction to Variational Methods for Graphical Models. In M.I. Jordan, editor, *Learning in Graphical Models*. Kluwer Academic Press, 1998.

[17] T. Mullin. *The Nature of Chaos*. Oxford Science Publications, 1993.

[18] U. Noppeney, W.D. Penny, C.J. Price, G. Flandin, and K.J. Friston. Identification of degenerate neuronal systems based on intersubject variability. *NeuroImage*, 2006. In Press.

[19] W.D. Penny. Kullback-Liebler Divergences of Normal, Gamma, Dirichlet and Wishart densities. Technical report, Wellcome Department of Imaging Neuroscience, 2001.

[20] W.D. Penny, G. Flandin, and N. Trujillo-Bareto. Bayesian Comparison of Spatially Regularised General Linear Models. *Human Brain Mapping*, 2006. Accepted for publication.

[21] W.D. Penny, S.J. Kiebel, and K.J. Friston. Variational Bayesian Inference for fMRI time series. *NeuroImage*, 19(3):727–741, 2003.

[22] W.D. Penny, N. Trujillo-Barreto, and K.J. Friston. Bayesian fMRI time series analysis with spatial priors. *NeuroImage*, 24(2):350–362, 2005.

[23] C. Peterson and J. Anderson. A mean field theory learning algorithm for neural networks. *Complex Systems*, 1:995–1019, 1987.

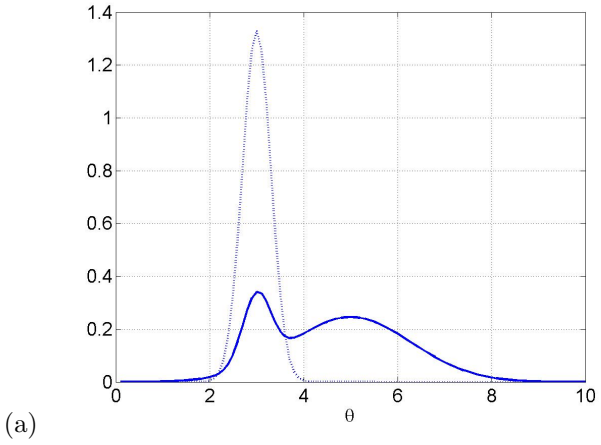
[24] W. H. Press, S.A. Teukolsky, W.T. Vetterling, and B.V.P. Flannery. *Numerical Recipes in C*. Cambridge, 1992.

[25] M. Sahani and S. S. Nagarajan. Reconstructing MEG sources with unknown correlations. In L. Saul S. Thrun and B. Schoelkopf, editors, *Advances in Neural Information Processing Systems*, volume 16. MIT, Cambridge, MA, 2004.

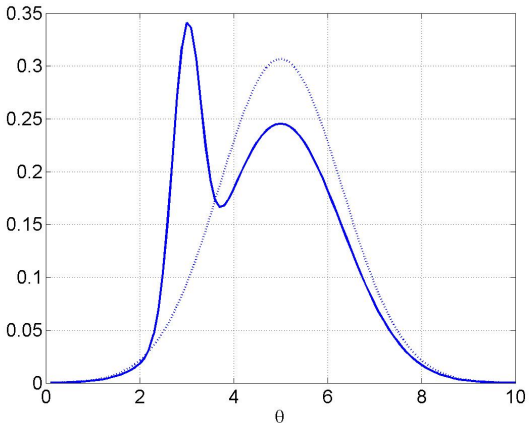
[26] M. Sato, T. Yoshioka, S. Kajihara, K. Toyama, N. Goda, K. Doya, and M. Kawato. Hierarchical Bayesian estimation for MEG inverse problem. *NeuroImage*, 23:806–826, 2004.

[27] J. Winn and C. Bishop. Variational message passing. *Journal of Machine Learning Research*, 6:661–694, 2005.

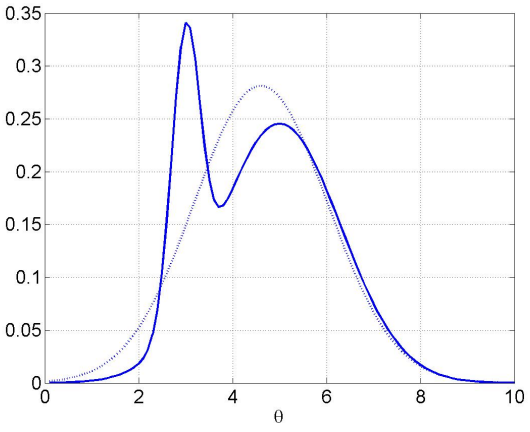
[28] M.W. Woolrich, T.E. Behrens, and S.M. Smith. Constrained linear basis sets for HRF modelling using Variational Bayes. *NeuroImage*, 21:1748–1761, 2004.



(a)



(b)



(c)

Figure 1: *Probability densities $p(\theta)$ (solid lines) and $q(\theta)$ (dashed lines) for a Gaussian mixture $p(\theta) = 0.2 \times \mathsf{N}(m_1, \sigma_1^2) + 0.8 \times \mathsf{N}(m_2, \sigma_2^2)$ with $m_1 = 3, m_2 = 5$,$\sigma_1 = 0.3, \sigma_2 = 1.3$, and a single Gaussian $q(\theta) = \mathsf{N}(\mu, \sigma^2)$ with (a) $\mu = \mu_1, \sigma = \sigma_1$ which fits the first mode, (b) $\mu = \mu_2, \sigma = \sigma_2$ which fits the second mode and (c) $\mu = 4.6, \sigma = 1.4$ which is moment-matched to $p(\theta)$.*

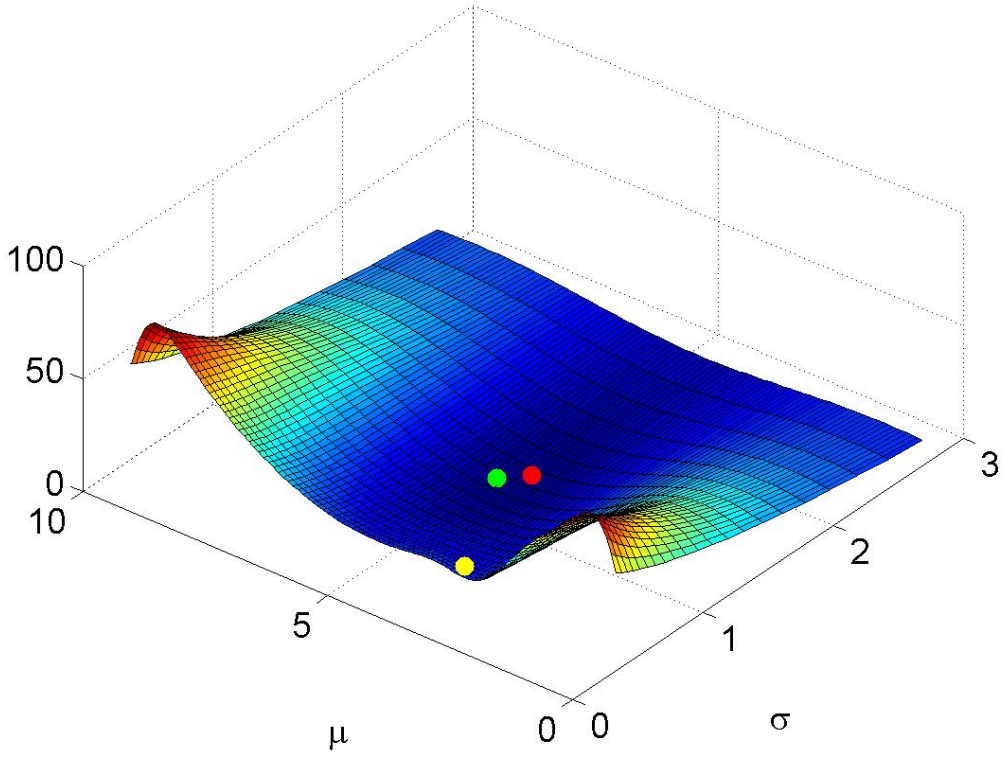


Figure 2: *KL-divergence, $KL(q||p)$ for p as defined in Figure 1 and q being a Gaussian with mean μ and standard deviation σ. The KL-divergences of the approximations in Figure 1 are (a) 11.73 for the first mode (yellow ball), (b) 0.93 for the second mode (green ball) and (c) 0.71 for the moment-matched solution (red ball).*

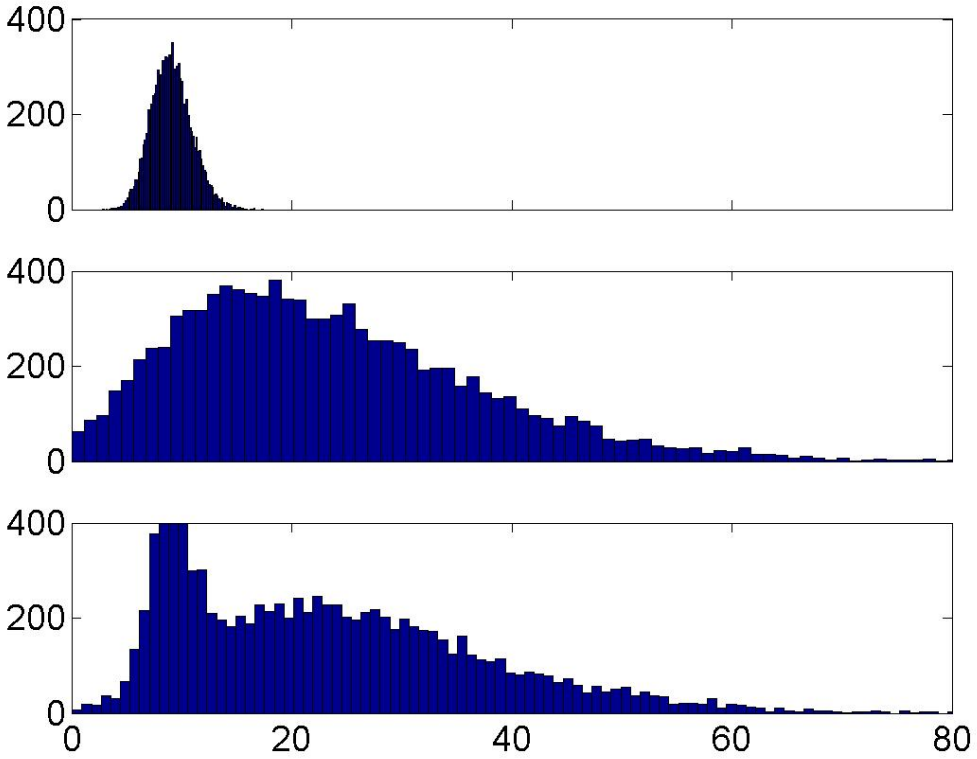


Figure 3: *Histograms of 10,000 samples drawn from $g(\theta)$ where the distribution over θ is from the Laplace approximation (top), VB approximation (middle) and true distribution, p, (bottom) for $g(\theta) = \theta^2$.*

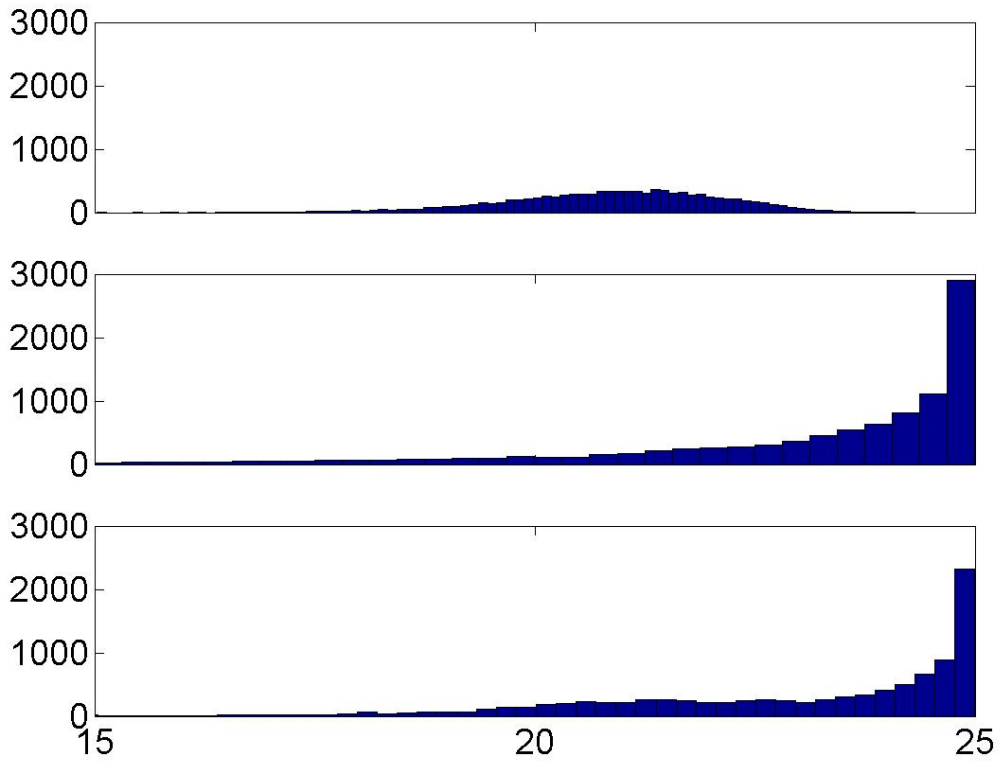


Figure 4: *Histograms of 10,000 samples drawn from $g(\theta)$ where the distribution over θ is from the Laplace approximation (top), VB approximation (middle) and true distribution, p, (bottom) for $g(\theta) = \theta * (10 - \theta)$. This is akin to a logistic map function encountered in dynamical systems [17].*

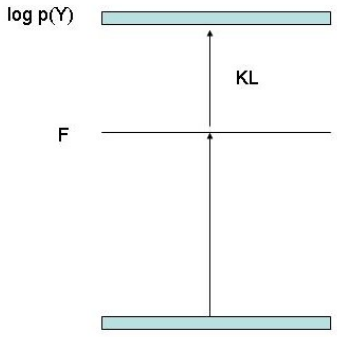
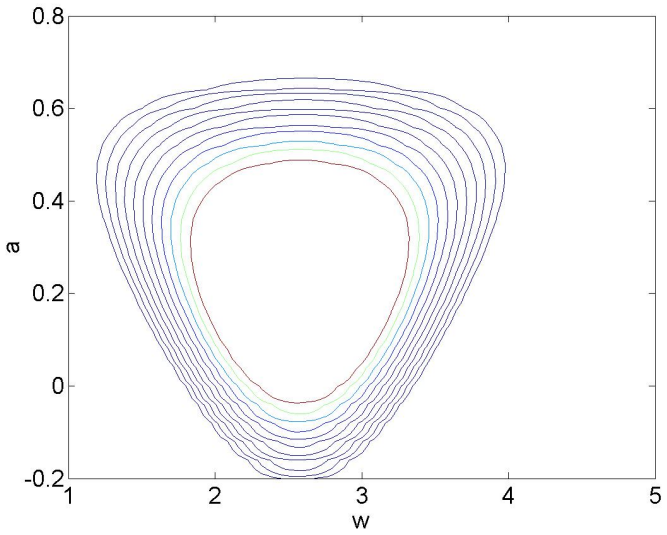
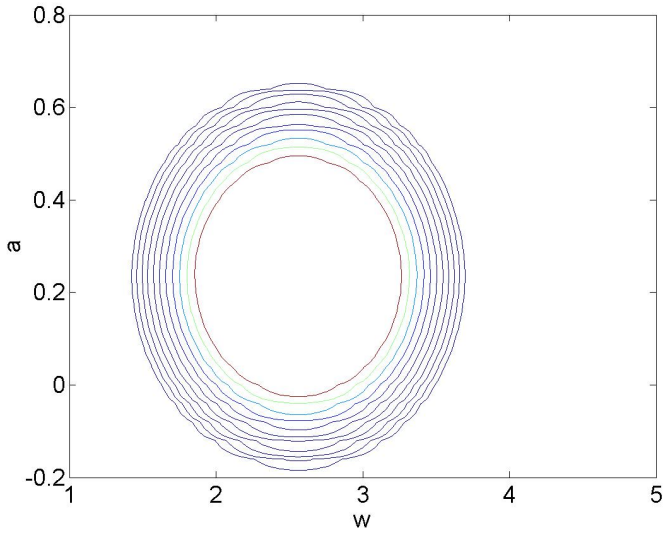


Figure 5: *The negative free energy, F, provides a lower bound on the log-evidence of the model with equality when the approximate posterior equals the true posterior.*



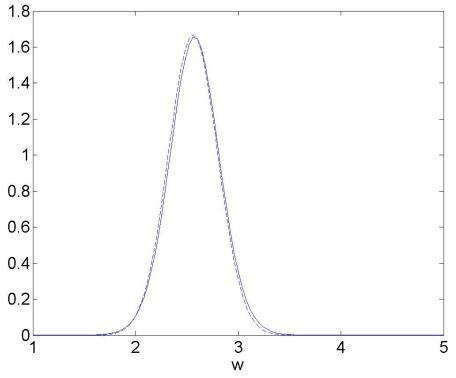
(a)



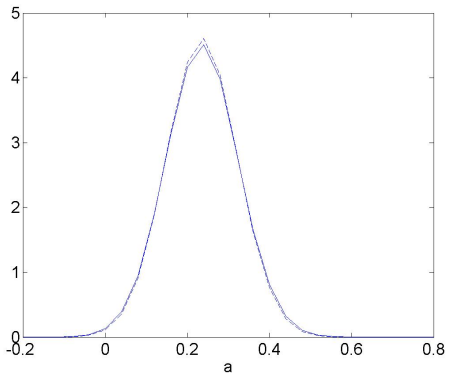
(b)

Figure 6: The figures show contour lines of constant probability density from (a) the exact posterior $p(a, w|Y)$ and (b) the approximate posterior used in VB, $q(a, w)$ for the GLM-AR model. This clearly shows the effect of the factorisation, $q(a, w) = q(a)q(w)$.

(a)



(b)



(c)

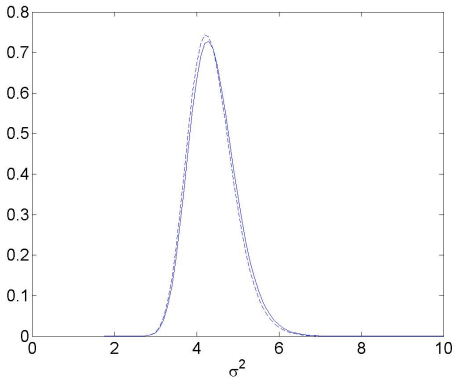
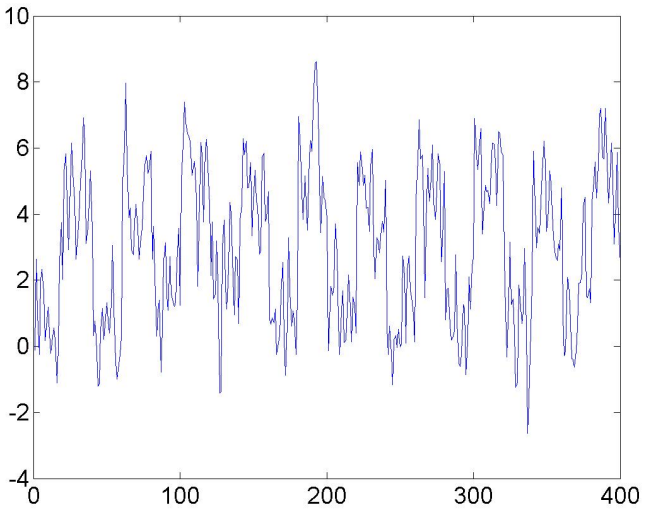
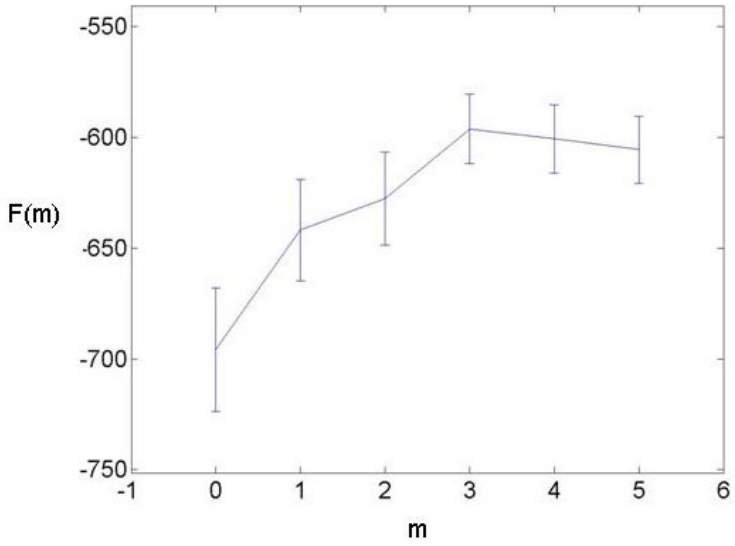


Figure 7: The figures compare the exact (solid lines) and approximate (dashed lines) marginal posteriors (a) $p(w|Y)$ and $q(w)$, (b) $p(a|Y)$ and $q(a)$, (c) $p(\sigma^2|Y)$ and $q(\sigma^2)$ (where $\sigma^2 = 1/\lambda$).



(a)



(b)

Figure 8: The figures show (a) an example time series from a GLM-AR model with AR model order $m = 3$ and (b) a plot of the average negative free energy $F(m)$, with error bars, versus m. This shows that $F(m)$ picks out the correct model order.

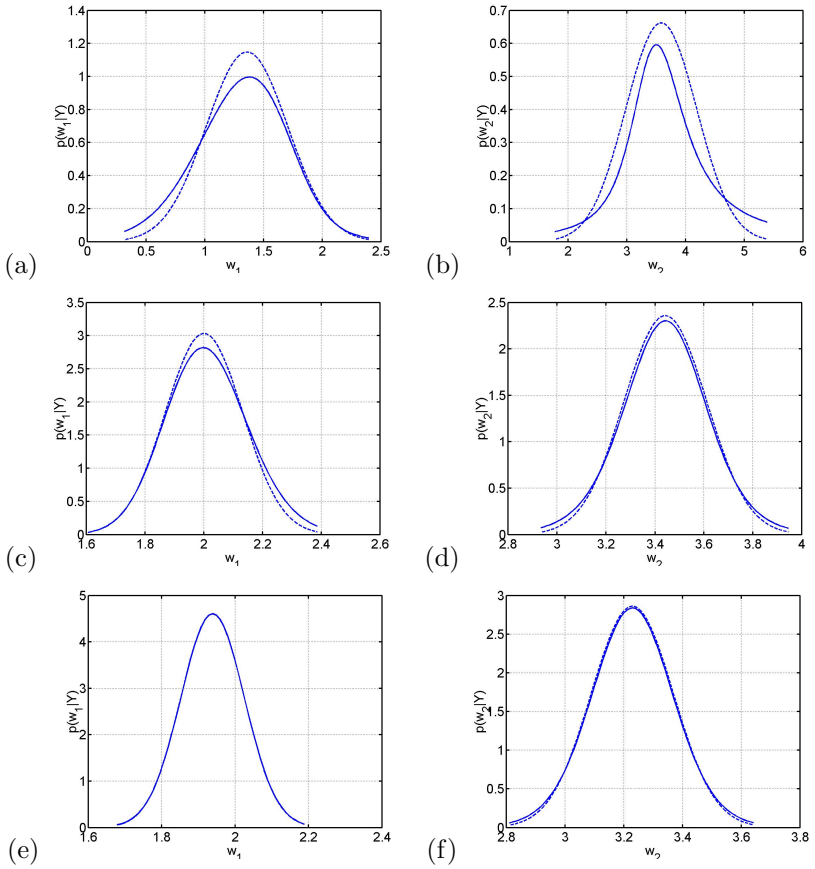


Figure 9: The figures show the posterior distributions from Gibbs sampling (solid lines) and Variational Bayes (dashed lines) for data sets containing 40 scans (top row), 160 scans (middle row) and 400 scans (bottom row). The distributions in the left column are for the first regression coefficient (size of activation) and in the right column for the second regression coefficient (offset). The fidelity of the VB approximation increases with number of scans.