ELSEVIER

NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage xx (2006) xxx–xxx

Technical Note

Variational free energy and the Laplace approximation

Karl Friston,[a,*] Jérémie Mattout,[a] Nelson Trujillo-Barreto,[b] John Ashburner,[a] and Will Penny[a]

[a]*The Wellcome Department of Imaging Neuroscience, Institute of Neurology, UCL, 12 Queen Square, London, WC1N 3BG, UK*
[b]*Cuban Neuroscience Centre, Havana, Cuba*

Received 30 January 2006; revised 19 July 2006; accepted 16 August 2006

This note derives the variational free energy under the Laplace approximation, with a focus on accounting for additional model complexity induced by increasing the number of model parameters. This is relevant when using the free energy as an approximation to the log-evidence in Bayesian model averaging and selection. By setting restricted maximum likelihood (ReML) in the larger context of variational learning and expectation maximisation (EM), we show how the ReML objective function can be adjusted to provide an approximation to the log-evidence for a particular model. This means ReML can be used for model selection, specifically to select or compare models with different covariance components. This is useful in the context of hierarchical models because it enables a principled selection of priors that, under simple hyperpriors, can be used for automatic model selection and relevance determination (ARD). Deriving the ReML objective function, from basic variational principles, discloses the simple relationships among Variational Bayes, EM and ReML. Furthermore, we show that EM is formally identical to a full variational treatment when the precisions are linear in the hyperparameters. Finally, we also consider, briefly, dynamic models and how these inform the regularisation of free energy ascent schemes, like EM and ReML.

Keywords: Variational Bayes; Free energy; Expectation maximisation; Restricted maximum likelihood; Model selection; Automatic relevance determination; Relevance vector machines

Introduction

The purpose of this note is to describe an adjustment to the objective function used in restricted maximum likelihood (ReML) that renders it equivalent to the free energy in variational learning and expectation maximisation. This is important because the variational free energy provides a bound on the log-evidence for any model, which is exact for linear models. The log-evidence plays a central role in model selection, comparison and averaging (for examples in neuroimaging, see Penny et al., 2004; Trujillo-Barreto et al., 2004).

Previously, we have described the use of ReML in the Bayesian inversion of electromagnetic models to localise distributed sources in EEG and MEG (e.g., Phillips et al., 2002). ReML provides a principled way of quantifying the relative importance of priors that replaces alternative heuristics like L-curve analysis. Furthermore, ReML accommodates multiple priors and provides more accurate and efficient source reconstruction than its precedents (Phillips et al., 2002). More recently, we have explored the use of ReML to identify the most likely combination of priors using model selection, where each model comprises a different set of priors (Mattout et al., in press). This was based on the fact that the ReML objective function is the free energy used in expectation maximisation and is equivalent to the log-evidence $F^{\lambda} = \ln p(y|\lambda,m)$, conditioned on λ, the unknown covariance component parameters (i.e., hyperparameters) and the model m. The noise covariance components encoded by λ include the prior covariances of each model of the data y.

However, this free energy is not a function of the conditional uncertainty about λ and is therefore insensitive to additional model complexity induced by adding covariance components. In this note we finesse this problem and show how F^{λ} can be adjusted to provide the variational free energy, which, in the context of linear models, is exactly the log-evidence $\ln p(y|m)$. This rests on deriving the variational free energy for a general variational scheme and treating expectation maximisation (EM) as a special case, in which one set of parameters assumes a point mass. We then treat ReML as the special case of EM, applied to linear models.

Although this note focuses on the various forms for the free energy, we also take the opportunity to link variational Bayes (VB), EM and ReML by deriving them from basic principles. Indeed, this derivation is necessary to show how the ReML objective function can be generalised for use in model selection. The material in this note is quite technical but is presented here because it underpins many of the specialist applications we have described in the neuroimaging literature over the past years. This didactic treatment may be especially useful for software developers or readers with a particular mathematical interest. For other readers, the main message is that a variational treatment of imaging data can unite a large number of special cases within a relatively simple framework.

Variational Bayes, under the Laplace approximation, assumes a fixed Gaussian form for the conditional density of the parameters

* Corresponding author. Fax: +44 207 813 1445.
E-mail address: k.friston@fil.ion.ucl.ac.uk (K. Friston).
Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2006.08.035

of a model and is used implicitly in ReML and many applications of EM. Bayesian inversion using VB is ubiquitous in neuroimaging (e.g., Penny et al., 2005). Its use ranges from spatial segmentation and normalisation of images during pre-processing (e.g., Ashburner and Friston, 2005) to the inversion of complicated dynamical casual models of functional integration in the brain (Friston et al., 2003). Many of the intervening steps in classical and Bayesian analysis of neuroimaging data call on ReML or EM under the Laplace approximation. This note provides an overview of how these schemes are related and illustrates their applications with reference to specific algorithms and routines we have described in the past (and are currently developing; e.g., dynamic expectation maximisation; DEM). One interesting issue that emerges from this treatment is that VB reduces exactly to EM, under the Laplace approximation, when the precision of stochastic terms is linear in the hyperparameters. This reveals a close relationship between EM and full variational approaches.

This note is divided into seven sections. In the first we summarise the basic theory of variational Bayes and apply it in the context of the Laplace approximation. The Laplace approximation imposes a fixed Gaussian form on the conditional density, which simplifies the ensuing variational steps. In this section we look at the easy problem of approximating the conditional covariance of model parameters and the more difficult problem of approximating their conditional expectation or mode using gradient ascent. We consider a dynamic formulation of gradient ascent, which generalises nicely to cover dynamic models and provides the basis for a temporal regularisation of the ascent. In the second section we apply the theory to nonlinear models with additive noise. We use the VB scheme that emerges as the reference for subsequent sections looking at special cases. The third section considers EM, which can be seen as a special case of VB in which uncertainty about one set of parameters is ignored. In the fourth section we look at the special case of linear models where EM reduces to ReML. The fifth section considers ReML and hierarchical models. Hierarchical models are important because they underpin parametric empirical Bayes (PEB) and other special cases, like relevance vector machines. Furthermore, they provide a link with classical covariance component estimation. In the sixth section we present some toy examples to show how the ReML and EM objective functions can be used to evaluate the log-evidence and facilitate model selection. This section concludes with an evaluation of the Laplace approximation to the model evidence, in relation to Monte Carlo–Markov chain (MCMC) sampling estimates. The final section revisits model selection using automatic model selection (AMS) and relevance determination (ARD). We show how suitable hyperpriors enable EM and ReML to select the best model automatically, by switching off redundant parameters and hyperparameters. The Appendices include some notes on parameterising covariances and the sampling scheme used for validation of the Laplace approximations.

Variational Bayes

Empirical enquiry in science usually rests upon estimating the parameters of some model of how observed data were generated and making inferences about the parameters (or model). Estimation and inference are based on the posterior density of the parameters (or model), conditional on the observations. Variational Bayes is used to evaluate these posterior densities.

The variational approach

Variational Bayes is a generic approach to posterior density (as opposed to posterior mode) analysis that approximates the conditional density $p(\vartheta|y,m)$ of some model parameters ϑ, given a model m and data y. Furthermore, it provides the evidence (also called the marginal or integrated likelihood) of the model $p(y|m)$ which, under prior assumptions about the model, furnishes the posterior density $p(m|y)$ of the model itself (see Penny et al., 2004 for an example in neuroimaging).

Variational approaches rest on minimising the Feynman variational bound (Feynman, 1972). In variational Bayes the free energy represents a bound on the log-evidence. Variational methods are well established in the approximation of densities in statistical physics (e.g., Weissbach et al., 2002) and were introduced by Feynman within the path integral formulation (Titantah et al., 2001). The variational framework was introduced into statistics though ensemble learning, where the ensemble or variational density $q(\vartheta)$ (i.e., approximating posterior density) is optimised to minimise the free energy. Initially (Hinton and von Cramp, 1993; MacKay, 1995a,b) the free energy was described in terms of description lengths and coding. Later, established methods like EM were considered in the light of variational free energy (Neal and Hinton, 1998; see also Bishop, 1999). Variational learning can be regarded as subsuming most other learning schemes as special cases. This is the theme pursued here, with special references to fixed-form approximations and classical methods like ReML (Harville, 1977).

The derivations in this paper involve a fair amount of differentiation. To simplify things we will use the notation $f_x=\partial f/\partial x$ to denote the partial derivative of the function f, with respect to the variable x. For time derivatives we will also use $\dot{x}=x_t$.

The log-evidence can be expressed in terms of the free energy and a divergence term

$$\begin{aligned}\ln p(y|m) &= F + D(q(\vartheta)||p(\vartheta|y,m))\\ F &= \langle L(\vartheta)\rangle_q - \langle \ln q(\vartheta)\rangle_q\\ L &= \ln p(y,\vartheta).\end{aligned} \tag{1}$$

Here $-\langle \ln q(\vartheta)\rangle_q$ is the entropy and $\langle L(\vartheta)\rangle_q$ the expected energy. Both quantities are expectations under the variational density. Eq. (1) indicates that F is a lower-bound approximation to the log-evidence because the divergence or cross-entropy $D(q(\vartheta)||p(\vartheta|y,m))$ is always positive. In this note, all the energies are the negative of energies considered in statistical physics. The objective is to compute $q(\vartheta)$ for each model by maximising F, and then compute F itself, for Bayesian inference and model comparison, respectively. Maximising the free energy minimises the divergence, rendering the variational density $q(\vartheta)\approx p(\vartheta|y,m)$ an approximate posterior, which is exact for linear systems. To make the maximisation easier one usually assumes $q(\vartheta)$ factorises over sets of parameters ϑ^i.

$$q(\vartheta) = \prod_i q^i. \tag{2}$$

In statistical physics this is called a mean field approximation. Under this approximation, the Fundamental Lemma of variational calculus means that F is maximised with respect to $q^i=q(\vartheta^i)$ when, and only when

$$\begin{aligned}\delta F^i = 0 &\Leftrightarrow \frac{\partial f^i}{\partial q^i} = f^i_{q^i} = 0\\ f^i &= F_{\vartheta^i}\end{aligned} \tag{3}$$

Please cite this article as: Karl Friston et al., Variational free energy and the Laplace approximation, NeuroImage (2006), doi:10.1016/j.neuroimage.2006.08.035

δF^i is the variation of the free energy with respect to q^i. From Eq. (1)

$$f^i = \int q^i q^{\setminus i} \ln L(\vartheta) d\vartheta^{\setminus i} - \int q^i q^{\setminus i} \ln q(\vartheta) d\vartheta^{\setminus i}$$
$$f^i_{q^i} = I(\vartheta^i) - \ln q^i - \ln Z^i$$
$$I(\vartheta^i) = \langle L(\vartheta) \rangle_{q^{\setminus i}} \tag{4}$$

Where $\vartheta^{\setminus i}$ denotes the parameters not in the ith set. We have lumped terms that do not depend on ϑ^i into $\ln Z^i$, where Z^i is a normalisation constant (i.e., partition function). We will call $I(\vartheta^i)$ the variational energy, noting its expectation under q^i is the expected energy. Note that when all the parameters are considered in a single set, the energy and variational energy become the same thing; i.e., $I(\vartheta^i)=L(\vartheta)$. The extremal condition in Eq. (3) is met when

$$\ln q^i = I(\vartheta^i) - \ln Z^i \Leftrightarrow$$
$$q(\vartheta^i) = \frac{1}{z^i} \exp(I(\vartheta^i)). \tag{5}$$

If this analytic form were tractable (e.g., through the use of conjugate priors) it could be used directly. See Beal and Ghahramani (2003) for an excellent treatment of conjugate-exponential models. However, we will assume a Gaussian fixed-form for the variational density to provide a generic scheme that can be applied to a wide range of models. Note that assuming a Gaussian form for the conditional density is equivalent to assuming a quadratic form for the variational energy (cf. a second order Taylor approximation).

The Laplace approximation

Laplace's method (also known as the saddle-point approximation) approximates an integral using a Taylor expansion of the integrands logarithm around its peak. Traditionally, in the statistics and machine leaning literature, the Laplace approximation refers to the evaluation of the marginal likelihood or free energy using Laplace's method. This is equivalent to a local Gaussian approximation of $p(\vartheta|y)$ around a maximum *a posteriori* (MAP) estimate (Kass and Raftery, 1995). A Gaussian approximation is motivated by the fact that, in the large data limit and given some regularity conditions, the posterior approaches a Gaussian around the MAP (Beal and Ghahramani, 2003). However, the Laplace approximation can be inaccurate with non-Gaussian posteriors, especially when the mode is not near the majority of the probability mass. By applying Laplace's method, in a variational context, we can avoid this problem: In what follows, we use a Gaussian approximation to each $p(\vartheta^i|y)$, induced by the mean field approximation. This finesses the evaluation of the variational energy $I(\vartheta^i)$ which is then optimised to find its mode. This contrasts with the conventional Laplace approximation; which is applied post hoc, after the mode has been identified. We will refer to this as the post hoc Laplace approximation.

Under the Laplace approximation, the variational density assumes a Gaussian form $q^i=N(\mu^i,\Sigma^i)$ with variational parameters μ^i and Σ^i corresponding to the conditional mode and covariance of the ith set of parameters. The advantage of this is that the conditional covariance can be evaluated very simply. Under the Laplace assumption

$$F = L(\mu) + \frac{1}{2}\sum_i \left(U^i + \ln|\Sigma^i| + p^i \ln 2\pi e\right)$$
$$I(\vartheta^i) = L(\vartheta^i, \mu^{\setminus i}) + \frac{1}{2}\sum_{j \neq i} U^j$$
$$U^i = tr(\Sigma^i L_{\vartheta^i\vartheta^i}) \tag{6}$$

$p^i=\dim(\vartheta^i)$ is the number of parameters in the ith set. The approximate conditional covariances obtain as an analytic function of the modes by differentiating $I(\vartheta^i)$ in Eq. (6) and solving for zero

$$F_{\Sigma^i} = \frac{1}{2} L_{\vartheta^i\vartheta^i} + \frac{1}{2}\Sigma^{i^{-1}} = 0 \Rightarrow$$
$$\Sigma^i = -L(\mu)^{-1}_{\vartheta^i\vartheta^i}. \tag{7}$$

Note that this solution for the conditional covariances does not depend on the mean-field approximation but only on the Laplace approximation. Eq. (7) recapitulates the conventional Laplace approximation; in which the conditional covariance is determined from the Hessian $L(\mu)_{\vartheta^i\vartheta^i}$, evaluated at the variational mode or maximum *aposteriori* (MAP). Substitution into Eq. (6) means $U^i=p^i$ and

$$F = L(\mu) + \sum_i \frac{1}{2}\left(\ln|\Sigma^i| + p^i \ln 2\pi\right). \tag{8}$$

The only remaining quantities required are the variational modes, which, from Eq. (5), maximise $I(\vartheta^i)$. This leads to the following compact variational scheme.

until convergence
  for all i
    $\mu^i = \max_{\vartheta^i} I(\vartheta^i)$
    $\Sigma^i = -L(\mu)^{-1}_{\vartheta^i\vartheta^i}$
  end
end
(9)

The variational modes

The modes can be found using a gradient ascent based on

$$\dot{\mu}^i = \frac{\partial I(\mu^i)}{\partial \vartheta^i} = I(\mu^i)_{\vartheta^i} \tag{10}$$

It may seem odd to formulate an ascent in terms of the motion of the mode in time. However, this is useful when generalising to dynamic models (see below). The updates for the mode obtain by integrating Eq. (10) to give

$$\Delta\mu^i = (\exp(tJ) - I)J^{-1}\dot{\mu}^i$$
$$J = \frac{\partial \dot{\mu}^i}{\partial \vartheta^i} = I(\mu^i)_{\vartheta^i\vartheta^i}. \tag{11}$$

Please cite this article as: Karl Friston et al., Variational free energy and the Laplace approximation, NeuroImage (2006), doi:10.1016/j.neuroimage.2006.08.035

When t gets large, the matrix exponential disappears; because the curvature is negative definite and we get a conventional Newton scheme

$$\Delta\mu^i = -I(\mu^i)^{-1}_{\vartheta^i\vartheta^i} I(\mu^i)_{\vartheta^i}. \tag{12}$$

Together with the expression for the conditional covariance in Eq. (7), this update furnishes a variational scheme under the Laplace approximation

$$\begin{aligned}
&\text{until convergence}\\
&\quad \text{for all } i\\
&\qquad \text{until convergence}\\
&\qquad\quad I(\mu^i)_{\vartheta^i_k} = L(\mu)_{\vartheta^i_k} + \tfrac{1}{2}\sum_{j\neq i} tr(\Sigma^j L_{\vartheta^j\vartheta^j\vartheta^i_k})\\
&\qquad\quad I(\mu^i)_{\vartheta^i_k\vartheta^i_l} = L(\mu)_{\vartheta^i_k\vartheta^i_l} + \tfrac{1}{2}\sum_{j\neq i} tr(\Sigma^j L_{\vartheta^j\vartheta^j\vartheta^i_k\vartheta^i_l})\\
&\qquad\quad \Delta\mu^i = -I(\mu^i)^{-1}_{\vartheta^i\vartheta^i} I(\mu^i)_{\vartheta^i}\\
&\qquad \text{end}\\
&\qquad \Sigma^i = -L(\mu)^{-1}_{\vartheta^i\vartheta^i}\\
&\quad \text{end}\\
&\text{end}
\end{aligned} \tag{13}$$

Note that this scheme rests on, and only on, the specification of the energy function $L(\vartheta)$ implied by a generative model.

Regularising variational updates

In some instances deviations from the quadratic form assumed for the variational energy $I(\vartheta^i)$ under the Laplace approximation can confound a simple Newton ascent. This can happen when the curvature of the objective function is badly behaved (e.g., when the objective function becomes convex, the curvatures can become positive and the ascent turns into a descent). In these situations some form of regularisation is required to ensure a robust ascent. This can be implemented by augmenting Eq. (10) with a decay term

$$\dot{\mu}^i = I(\mu^i)_{\vartheta^i} - v\Delta\mu^i. \tag{14}$$

This effectively pulls the search back towards the expansion point provided by the previous iteration and enforces a local exploration. Integration to the fixed point gives a classical Levenburg–Marquardt scheme (cf. Eq. (11))

$$\begin{aligned}
\Delta\mu^i &= -J^{-1}\dot{\mu}^i\\
&= (vI - I(\mu^i)_{\vartheta^i\vartheta^i})^{-1} I(\mu^i)_{\vartheta^i}\\
J &= I(\mu^i)_{\vartheta^i\vartheta^i} - vI
\end{aligned} \tag{15}$$

where v is the Levenburg–Marquardt regularisation. However, the dynamic formulation affords a simpler alternative, namely temporal regularisation. Here, instead of constraining the search with a decay term, one can abbreviate it by terminating the ascent after some suitable period $t=v$; from Eq. (11)

$$\begin{aligned}
\Delta\mu^i &= (\exp(vJ) - I)J^{-1}\dot{\mu}^i\\
&= (\exp(vI(\mu^i)_{\vartheta^i\vartheta^i}) - I) I(\mu^i)^{-1}_{\vartheta^i\vartheta^i} I(\mu^i)_{\vartheta^i}\\
J &= I(\mu^i)_{\vartheta^i\vartheta^i}
\end{aligned} \tag{16}$$

This has the advantage of using the local gradients and curvatures while precluding large excursions from the expansion point. In our implementations $v=1/\eta$ is based on the 2-norm of the curvature η for both regularisation schemes. The 2-norm is the largest singular value and, in the present context, represents an upper bound on rate of convergence of the ascent (cf. a Lyapunov exponent).[1] Terminating the ascent prematurely is reminiscent of "early stopping" in the training of neural networks in which the number of weights far exceeds the sample size (e.g., Nelson and Illingworth, 1991, p. 165). It is interesting to note that "early stopping" is closely related to ridge regression, which is another perspective on Levenburg–Marquardt regularisation.

A comparative example using Levenburg–Marquardt and temporal regularisation is provided in Fig. 1 and suggests, in this example, temporal regularisation is better. Either approach can be implemented in the VB scheme by simply regularising the Newton update if the variational energy $I(\vartheta^i)$ fails to increase after each iteration. We prefer temporal regularisation because it is based on a simpler heuristic and, more importantly, is straightforward to implement in dynamic schemes using high-order temporal derivatives.

A note on dynamic models

The second reason we have formulated the ascent as a time-dependent process is that it can be used to invert dynamic models. In this instance, the integration time in Eq. (16) is determined by the interval between observations. This is the approach taken in our variational treatment of dynamic systems; namely, dynamic expectation maximisation or DEM (introduced briefly in Friston, 2005 and implemented in spm_DEM.m). DEM produces conditional densities that are a continuous function of time and avoids many of the limitations of discrete schemes based on incremental Bayes (e.g., extended Kalman filtering). In dynamic models the energy is a function of the parameters and their high-order motion; i.e., $I(\vartheta^i)\rightarrow I(\vartheta^i, \dot{\vartheta}^i,\ldots,t)$. This entails the extension of the variational density to cover this motion, using generalised coordinates $q(\vartheta^i)\rightarrow q(\vartheta^i, \dot{\vartheta}^i,\ldots,t)$. This approach will be described fully in a subsequent paper. Here we focus on static models.

Having established the operational equations for VB under the Laplace approximation we now look at their application to some specific models.

Variational Bayes for nonlinear models

Consider the generative model with additive error $y=G(\theta)+\varepsilon(\lambda)$. Gaussian assumptions about the errors or innovations $p(\varepsilon)=N(0,\Sigma(\lambda))$ furnish a likelihood $p(y|\theta,\lambda)=N(G(\theta),\Sigma(\lambda))$. In this example, we can consider the parameters as falling into two sets $\vartheta=\{\theta,\lambda\}$ such that $q(\vartheta)=q(\theta)q(\lambda)$, where $q(\theta)=N(\mu^\theta,\Sigma^\theta)$ and $q(\lambda)=N(\mu^\lambda,\Sigma^\lambda)$. We will also assume Gaussian priors $p(\theta)=N(\eta^\theta,\Pi^{\theta-1})$ and $p(\lambda)=N(\eta^\lambda,\Pi^{\lambda-1})$. We will refer to the two sets as the parameters and hyperparameters. These likelihood and priors define the energy $L(\vartheta)=\ln p(y|\theta,\lambda)+\ln p(\theta)+\ln p(\lambda)$. Note that Gaussian priors are not too restrictive because both $G(\theta)$ and $\Sigma(\lambda)$ can be nonlinear functions that embody a probability integral transform (i.e., can implement a re-parameterisation in terms of non-Gaussian priors).

[1] Note that the largest singular value is the largest negative eigenvalue of the curvature and represents the largest rate of change of the gradient locally.

Please cite this article as: Karl Friston et al., Variational free energy and the Laplace approximation, NeuroImage (2006), doi:10.1016/j.neuroimage.2006.08.035

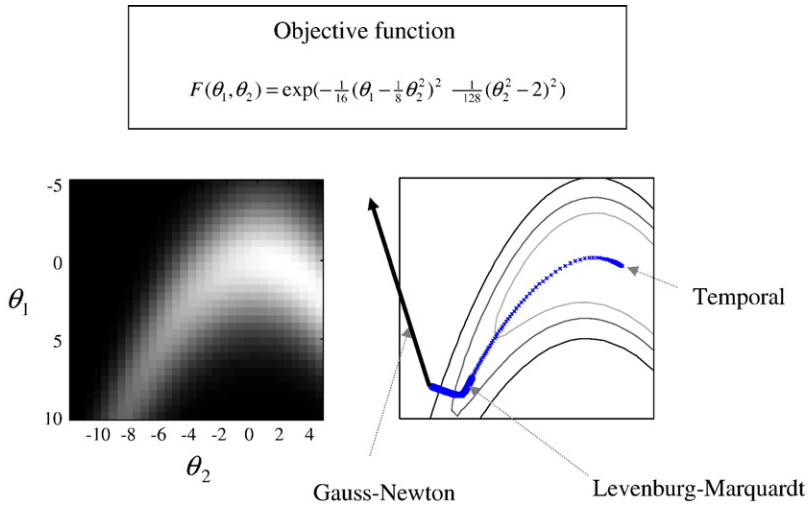

Fig. 1. Examples of Levenburg–Marquardt and temporal regularisation. The left panel shows an image of the landscape defined by the objective function $F(\theta_1,\theta_2)$ of two parameters (upper panel). This was chosen to be difficult for conventional schemes; exhibiting curvilinear valleys and convex regions. The right panel shows the ascent trajectories, over 256 iterations (starting at 8, −10), superimposed on a contour plot of the landscape. In these examples the regularisation parameter was the 2-norm of the curvature evaluated at each update. Note how the ascent goes off in the wrong direction with no regularisation (Newton). The regularisation adopted by Levenburg–Marquardt makes its progress slow, in relation to the temporal regularisation, so that it fails to attain the maximum after 256 iterations.

Given n samples, p parameters and h hyperparameters the energy and its derivatives are

$$\begin{aligned} L(\vartheta) = &-\frac{1}{2}\varepsilon^T\Sigma^{-1}\varepsilon + \frac{1}{2}\ln|\Sigma^{-1}| - \frac{n}{2}\ln 2\pi \\ &-\frac{1}{2}\varepsilon^{\theta T}\Pi^\theta\varepsilon^\theta + \frac{1}{2}\ln|\Pi^\theta| - \frac{p}{2}\ln 2\pi \\ &-\frac{1}{2}\varepsilon^{\lambda T}\Pi^\lambda\varepsilon^\lambda + \frac{1}{2}\ln|\Pi^\lambda| - \frac{h}{2}\ln 2\pi \end{aligned} \tag{17}$$

$$\varepsilon = G(\mu^\theta) - y$$
$$\varepsilon^\theta = \mu^\theta - \eta^\theta$$
$$\varepsilon^\lambda = \mu^\lambda - \eta^\lambda$$

and

$$\begin{aligned} L(\mu)_\theta &= -G_\theta^T\Sigma^{-1}\varepsilon - \Pi^\theta\varepsilon^\theta \\ L(\mu)_{\theta\theta} &= -G_\theta^T\Sigma^{-1}G_\theta - \Pi^\theta \\ L(\mu)_{\lambda i} &= -\frac{1}{2}tr\left(P_i\left(\varepsilon\varepsilon^T - \Sigma\right)\right) - \Pi^\lambda_{i\cdot}\varepsilon^\lambda \\ L(\mu)_{\lambda\lambda ij} &= -\frac{1}{2}tr\left(P_{ij}\left(\varepsilon\varepsilon^T - \Sigma\right)\right) - \frac{1}{2}tr\left(P_i\Sigma P_j\Sigma\right) - \Pi^\lambda_{ij} \end{aligned} \tag{18}$$

$$P_i = \frac{\partial\Sigma^{-1}}{\partial\lambda_i} \qquad P_{ij} = \frac{\partial^2\Sigma^{-1}}{\partial\lambda_i\partial\lambda_j}$$

Note that we have ignored second-order terms that depend on $G_{\theta\theta}$, under the assumption that the generative model is only weakly nonlinear. The requisite gradients and curvatures are

$$\begin{aligned} I(\theta)_{\theta k} &= L(\theta,\mu^\lambda)_{\theta k} + \frac{1}{2}tr\left(\Sigma^\lambda A^k\right) & I(\lambda)_{\lambda i} &= L(\mu^\theta,\lambda)_{\lambda i} + \frac{1}{2}tr\left(\Sigma^\theta C^i\right) \\ I(\theta)_{\theta\theta kl} &= L(\theta,\mu^\lambda)_{\theta\theta kl} + \frac{1}{2}tr\left(\Sigma^\lambda B^{kl}\right) & I(\lambda)_{\lambda\lambda ij} &= L(\mu^\theta,\lambda)_{\lambda\lambda ij} + \frac{1}{2}tr\left(\Sigma^\theta D^{ij}\right) \\ A^k_{ij} &= -G^T_{\theta\cdot k}P_{ij}\varepsilon & C^i &= -G_\theta^T P_i G_\theta \\ B^{kl}_{ij} &= -G^T_{\theta\cdot k}P_{ij}G_{\theta\cdot l} & D^{ij} &= -G_\theta^T P_{ij}G_\theta \end{aligned} \tag{19}$$

where $G_{\theta\cdot k}$ denotes the kth column of G_θ. These enter the VB scheme in Eq. (13), giving the two-step scheme

until convergence

  until convergence

$$\begin{aligned} \Sigma^{\theta^{-1}} &= G_\theta^T\Sigma^{-1}G_\theta + \Pi^\theta \\ L(\mu) &= -G^T\Sigma_\theta^{-1}\varepsilon - \Pi^\theta\varepsilon^\theta \\ I(\mu)_{\theta k} &= L(\mu)_{\theta k} + \tfrac{1}{2}tr(\Sigma^\lambda A^k) \\ I(\mu)_{\theta\theta kl} &= -\Sigma^{\theta^{-1}}_{kl} + \tfrac{1}{2}tr(\Sigma^\lambda B^{kl}) \\ \Delta\mu^\theta &= -I(\mu)^{-1}_{\theta\theta}\, I(\mu)_\theta \end{aligned}$$

  end
  until convergence

$$\begin{aligned} \Sigma^{\lambda^{-1}}_{ij} &= \tfrac{1}{2}tr(P_{ij}(\varepsilon\varepsilon^T - \Sigma) + P_i\Sigma P_j\Sigma) + \Pi^\lambda_{ij} \\ I(\mu)_{\lambda i} &= -\tfrac{1}{2}tr(P_i(\varepsilon\varepsilon^T - \Sigma + G_\theta\Sigma^\theta G_\theta^T)) + \Pi^\lambda_{i\cdot}\varepsilon_\lambda \\ I(\mu)_{\lambda\lambda ij} &= -\Sigma^{\lambda^{-1}}_{ij} - \tfrac{1}{2}tr(\Sigma^\theta G_\theta^T P_{ij}G_\theta) \\ \Delta\mu^\lambda &= -I(\mu)^{-1}_{\lambda\lambda}\, I(\mu)_\lambda \end{aligned}$$

  end
end

(20)

The negative free energy for these models is

$$F = -\frac{1}{2}\varepsilon^T\Sigma^{-1}\varepsilon + \frac{1}{2}\ln|\Sigma^{-1}| - \frac{n}{2}\ln 2\pi$$

Please cite this article as: Karl Friston et al., Variational free energy and the Laplace approximation, NeuroImage (2006), doi:10.1016/j.neuroimage.2006.08.035

$$-\frac{1}{2}\varepsilon^{\theta T}\Pi^{\theta}\varepsilon^{\theta}+\frac{1}{2}\ln|\Pi^{\theta}|+\frac{1}{2}\ln|\Sigma^{\theta}|$$

$$-\frac{1}{2}\varepsilon^{\lambda T}\Pi^{\lambda}\varepsilon^{\lambda}+\frac{1}{2}\ln|\Pi^{\lambda}|+\frac{1}{2}\ln|\Sigma^{\lambda}| \quad (21)$$

In principle, these equations cover a large range of models and will work provided the true posterior is unimodal (and roughly Gaussian). The latter requirement can usually be met by a suitable transformation of parameters. In the next section, we consider a further simplification of our assumptions about the variational density and how this leads to expectation maximisation.

Expectation maximisation for nonlinear models

There is a key distinction between θ and λ in the generative model above: The parameters λ are hyperparameters in the sense, like the variational parameters, they parameterise a density. In many instances their conditional density per se is uninteresting. In variational expectation maximisation EM, we ignore uncertainty about the hyperparameters. In this case, the free energy is effectively conditioned on λ and reduces to

$$\begin{aligned}F^{\lambda} &= \ln p(y|\lambda) - D(q(\theta)||p(\theta|y,\lambda))\\ &= -\frac{1}{2}\varepsilon^{T}\Sigma^{-1}\varepsilon+\frac{1}{2}\ln|\Sigma^{-1}|-\frac{n}{2}\ln 2\pi\\ &\quad -\frac{1}{2}\varepsilon^{\theta T}\Pi^{\theta}\varepsilon^{\theta}+\frac{1}{2}\ln|\Pi^{\theta}|+\frac{1}{2}\ln|\Sigma^{\theta}|. \quad (22)\end{aligned}$$

Here, $F^{\lambda}\leq\ln p(y|\lambda)$ becomes a lower bound on the log likelihood of the hyperparameters. This means the variational step updating the hyperparameters maximises the likelihood of the hyperparameters $\ln p(y|\lambda)$ and becomes an M-step. In this context, Eq. (20) simplifies because we can ignore the terms that involve Σ^{λ} and Π^{λ} to give

until convergence
  until convergence: E-step

$$\Sigma^{\theta^{-1}} = G_{\theta}^{T}\Sigma^{-1}G_{\theta} + \Pi^{\theta}$$

$$\Delta\mu^{\theta} = -\Sigma^{\theta^{-1}}(G_{\theta}^{T}\Sigma^{-1}\varepsilon + \Pi^{\theta}\varepsilon^{\theta})$$

  end
  until convergence: M-step

$$I(\mu)_{\lambda i} = -\tfrac{1}{2}tr(P_i(\varepsilon\varepsilon^{T}-\Sigma+G_{\theta}\Sigma^{\theta}G_{\theta}^{T}))$$

$$I(\mu)_{\lambda\lambda ij} = -\tfrac{1}{2}tr(P_{ij}(\varepsilon\varepsilon^{T}-\Sigma+G_{\theta}\Sigma^{\theta}G_{\theta}^{T}))+P_i\Sigma P_j\Sigma)$$

$$\Delta\mu_{\lambda}^{\lambda} = -I(\mu)_{\lambda\lambda}^{-1}I(\mu)_{\lambda}$$

  end
end

(23)

Expectation–maximisation or EM is an iterative parameter re-estimation procedure devised to estimate the parameters and hyperparameters of a model. It was introduced as an iterative method to obtain maximum likelihood estimators with incomplete data (Hartley, 1958) and was generalised by Dempster et al. (1977). Strictly speaking, EM refers to schemes in which the conditional density of the E-step is known exactly, obviating the need for fixed-form assumptions. This is why we used the term 'variational EM' above.

In terms of the VB scheme, the M-step for $\mu^{\lambda}=\max I(\lambda)$ is unchanged because $I(\lambda)$ does not depend on Σ^{λ}. The remaining variational steps (i.e., E-steps) are simplified because one does not have to average over the conditional density $q(\lambda)$. This ensuing scheme is that described in Friston (2002) for nonlinear system identification and is implemented in spm_nlsi.m. Although this scheme is applied to time series it actually treats the underlying model as static, generating finite-length data sequences. This routine is used to identify hemodynamic models in terms of biophysical parameters for regional responses and dynamic causal models (DCMs) of distributed responses in a variety of applications; e.g., fMRI (Friston et al., 2003), EEG (David et al., 2006), MEG (Kiebel et al., 2006) and mean-field models of neuronal activity (Harrison et al., 2005).

A formal equivalence

A key point here is that VB and EM are exactly the same when $P_{ij}=0$. In this instance the matrices A, B and D in Eq. (19) disappear. This means the VB-step for the parameters does not depend on Σ^{λ} and becomes formally identical to the E-step. Because the VB-step for the hyperparameters is already the same as the M-step (apart from the loss of hyperpriors) the two schemes converge. One can ensure $P_{ij}=0$ by adopting a hyper-parameterisation, which renders the precision linear in the hyperparameters; for example, a linear mixture of precision components Q_i (see Appendix 1). This resulting variational scheme is used by the SPM5 version of spm_nlsi.m for nonlinear system identification.

The second key point that follows from this analysis is that one can adjust the EM free energy to approximate the log-evidence, as described next.

Accounting for uncertainty about the hyperparameters

The EM free energy in Eq. (22) discounts uncertainty about the hyperparameters because it is conditioned upon them. This is a well-recognised problem, sometimes referred to as the over-confidence problem, for which a number of approximate solutions have been suggested (e.g., Kass and Steffey, 1989). Here we describe a solution that appeals to the variational framework within which EM can be treated.

If we treat EM as an approximate variational scheme, we can adjust the EM free energy to give the variational free energy required for model comparison and averaging. By comparing Eqs. (21) and (22) we can express the variational free energy in terms of F^{λ} and an extra term[2]

$$F = F^{\lambda}+\frac{1}{2}\ln|\Sigma^{\lambda}|$$

$$\Sigma_{ij}^{\lambda} = -L(\mu)_{\lambda\lambda}^{-1} \quad (24)$$

where the expression for $L(\mu)_{\lambda\lambda}$ comes from Eq. (18). Intuitively, the extra term encodes the conditional information (i.e., entropy) about the models covariance components. The log-evidence will only increase if there is conditional information about the extra com-

[2] We are assuming that there are no hyperpriors on the hyperparameters so that terms involving Π^{λ} can be ignored.

Please cite this article as: Karl Friston et al., Variational free energy and the Laplace approximation, NeuroImage (2006), doi:10.1016/j.neuroimage.2006.08.035

ponent. Adding redundant components will have no effect on F (see the section below on automatic model selection). This term can be regarded as additional Occam factor (Mackay and Takeuchi, 1996).

Note that even when conditional uncertainty about the hyperparameters has no effect on the conditional density of the parameters (e.g., when the precisions are linear in the hyperparameters—see above) this uncertainty can still have a profound effect on model selection because it is an important component of the free energy and therefore the log-evidence for a particular model.

Adjusting the EM free energy to approximate the log-evidence is important because of the well-known connections between EM for linear models and restricted maximum likelihood. This connection suggests that ReML could also be used to evaluate the log-evidence and therefore be used for model selection. We now consider ReML as a special case of EM.

Restricted maximum likelihood for linear models

In the case of general linear models $G(\theta)=G\theta$ with additive Gaussian noise and no priors on the parameters (i.e., $\Pi^{\theta}=0$) the free energy reduces to

$$\begin{aligned} F^{\theta} &= \ln p(y|\lambda) - D(q(\theta)||p(\theta|y,\lambda)) \\ &= -\frac{1}{2}\varepsilon^T\Sigma^{-1}\varepsilon + \frac{1}{2}\ln|\Sigma^{-1}| - \frac{n}{2}\ln 2\pi + \frac{1}{2}\ln|\Sigma^{\theta}|. \end{aligned} \tag{25}$$

Critically, the dependence on $q(\theta)$ can be eliminated using the closed form solutions for the conditional moments

$$\mu^{\theta} = \Sigma^{\theta}G^T\Sigma^{-1}y$$

$$\Sigma^{\theta} = (G^T\Sigma^{-1}G)^{-1}$$

to eliminate the divergence term and give

$$\begin{aligned} F^{\theta} &= \ln p(y|\lambda) \\ &= -\frac{1}{2}tr\left(\Sigma^{-1}Ryy^TR^T\right) + \frac{1}{2}\ln|\Sigma^{-1}| - \frac{n}{2}\ln 2\pi - \frac{1}{2}\ln|G^T\Sigma^{-1}G| \end{aligned}$$

$$\varepsilon = Ry$$

$$R = I - G(G^T\Sigma^{-1}G)^{-1}G^T\Sigma^{-1} \tag{26}$$

This free energy is also known as the ReML objective function (Harville, 1977). ReML or restricted maximum likelihood was introduced by Patterson and Thompson in 1971 as a technique for estimating variance components, which accounts for the loss in degrees of freedom that result from estimating fixed effects (Harville, 1977). The elimination makes the free energy a simple function of the hyperparameters and, effectively, the EM scheme reduces to a single M-step or ReML-step

$$\begin{aligned} &\text{until convergence: ReML-step} \\ &\quad L(\mu)_{\lambda i} = -\tfrac{1}{2}tr(P_iR(yy^T-\Sigma)R^T) \\ &\quad \left\langle L(\mu)_{\lambda\lambda ij}\right\rangle = -\tfrac{1}{2}tr(P_iR\Sigma P_jR\Sigma) \\ &\quad \Delta\mu^{\theta} = -\left\langle L(\mu)_{\lambda\lambda}\right\rangle^{-1}L(\mu)_{\lambda} \\ &\text{end} \end{aligned} \tag{27}$$

Notice that the energy has replaced the variational energy because they are the same: from Eq. (6) $I(\vartheta)=L(\lambda)$. This is a result of eliminating $q(\theta)$ from the variational density. Furthermore, the curvature has been replaced by its expectation to render the Newton decent a Fisher–Scoring scheme using

$$\left\langle Ryy^TR^T\right\rangle = R\Sigma R^T = \Sigma - G\Sigma_q^{\theta}G^T = R\Sigma. \tag{28}$$

To approximate the log-evidence we can adjust the ReML free energy, after convergence, as with the EM free energy

$$\begin{aligned} F &= F^{\theta} + \frac{1}{2}\ln|\Sigma^{\lambda}| \\ \Sigma^{\lambda} &= -\left\langle L(\mu)_{\lambda\lambda}\right\rangle^{-1}. \end{aligned} \tag{29}$$

The conditional covariance of the hyperparameters uses the same curvature as the ascent in Eq. (27). Being able to compute the log-evidence from ReML is useful because ReML is used widely in an important class of models, namely hierarchical models reviewed next.

Restricted maximum likelihood for hierarchical linear models

Parametric empirical Bayes

The application of ReML to the linear models of the previous section did not accommodate priors on the parameters. However, one can generally absorb these priors into the error covariance components using a hierarchical formulation. This enables the use of ReML to identify models with full or empirical priors. Hierarchical linear models are equivalent to parametric empirical Bayes models (Efron and Morris, 1973) in which empirical priors emerge from conditional independence of the errors $\varepsilon^{(i)}\sim N(0,\Sigma^{(i)})$:

$$\begin{aligned} y^{(1)} &= & y^{(1)} &= \varepsilon^{(1)} \\ \theta^{(1)} &= G^{(1)}\theta^{(2)} + \varepsilon^{(1)} & &+ G^{(1)}\varepsilon^{(2)} \\ \theta^{(2)} &= G^{(2)}\theta^{(3)} + \varepsilon^{(2)} \quad\equiv & &+ G^{(1)}G^{(2)}\varepsilon^{(3)} \\ &\vdots & &\vdots \\ \theta^{(n)} &= \varepsilon^{(n)} & &+ G^{(1)}\ldots G^{(n-1)}\theta^{(n)} \end{aligned} \tag{30}$$

In hierarchical models, the random terms model uncertainly about the parameters at each level and $\Sigma(\lambda)^{(i)}$ are treated as prior covariance constraints on $\theta^{(i)}$. Hierarchical models of this sort are very common and underlie all classical mixed effects analyses of variance.[3] ReML identification of simple two-level models like

$$y^{(1)} = G^{(1)}\theta^{(2)} + \varepsilon^{(1)}$$

$$\theta^{(2)} = \varepsilon^{(2)} \tag{31}$$

is a useful way to impose shrinkage priors on the parameters and covers early approaches (e.g., Stein shrinkage estimators) to recent developments, such as relevance vector machines (e.g.,

[3] For an introduction to EM algorithms in generalised linear models, see Fahrmeir and Tutz (1994). This text provides an exposition of EM and PEB in linear models, usefully relating EM to classical methods (e.g., ReML p. 225).

Please cite this article as: Karl Friston et al., Variational free energy and the Laplace approximation, NeuroImage (2006), doi:10.1016/j.neuroimage.2006.08.035

Tipping, 2001). Relevance vector machines represent a Bayesian treatment of support vector machines, in which the second-level covariance $\Sigma(\lambda)^{(2)}$ has a component for each parameter. Most of the ReML estimates of these components shrink to zero. This means the columns of $G^{(1)}$ whose parameters have zero mean and variance can be eliminated, providing a new model with sparse support. This is also known as automatic relevance determination (ARD; MacKay, 1995a,b) and will be illustrated below.

Estimating these models through their covariances $\Sigma^{(i)}$ with ReML corresponds to empirical Bayes. This estimation can proceed in one of two ways: First, we can augment the model and treat the random terms as parameters to give

$$y = J\theta + \varepsilon$$

$$y = \begin{bmatrix} y^{(1)} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad J = \begin{bmatrix} K^{(2)} & \cdots & K^{(n)} \\ -I & & \\ & \ddots & \\ & & -I \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon^{(1)} \\ \varepsilon^{(2)} \\ \vdots \\ \varepsilon^{(n)} \end{bmatrix} \quad \theta = \begin{bmatrix} \varepsilon^{(2)} \\ \vdots \\ \theta^{(n)} \end{bmatrix}$$

$$K^{(i)} = \prod_{j=1}^{i} G^{(j-1)}$$

$$\Sigma = \begin{bmatrix} \Sigma^{(1)} & & \\ & \ddots & \\ & & \Sigma^{(n)} \end{bmatrix} \tag{32}$$

with $G^{(0)}=I$. This reformulation is a nonhierarchical model with no explicit priors on the parameters. However, the ReML estimates of $\Sigma(\lambda)^{(i)}$ are still the empirical prior covariances of the parameters $\theta^{(i)}$ at each level. If $\Sigma^{(i)}$ is known a priori, it simply enters the scheme as a known covariance component. This corresponds to a full Bayesian analysis with known or full priors for the level in question.

spm_peb.m uses this reformulation and Eq. (27) for estimation. The conditional expectations of the parameters are recovered by recursive substitution of the conditional expectations of the errors into Eq. (30) (cf. Friston, 2002). spm_peb.m uses a computationally efficient substitution

$$\frac{1}{2}tr\left(P_iR\left(yy^T - \Sigma\right)R^T\right) = \frac{1}{2}y^TR^TP_iRy - \frac{1}{2}tr\left(P_iR\Sigma R^T\right) \tag{33}$$

to avoid computing the potentially large matrix yy^T. We have used this scheme extensively in the construction of posterior probability maps or PPMs (Friston and Penny, 2003) and mixed-effect analysis of multi-subject studies in neuroimaging (Friston et al., 2005). Both these examples rest on hierarchical models, using hierarchical structure over voxels and subjects, respectively.

Classical covariance component estimation

An equivalent identification of hierarchical models rests on an alternative and simpler reformulation of Eq. (30) in which all the hierarchically induced covariance components $K^{(i)T}\Sigma^{(i)}K^{(i)T}$ are treated as components of a compound error

$$y = \varepsilon$$

$$y = y^{(1)}$$

$$\varepsilon = \sum_{i=1}^{n} K^{(i)}\varepsilon^{(i)}$$

$$\Sigma = \sum_{i=1}^{n} K^{(i)T}\Sigma^{(i)}K^{(i)T}. \tag{34}$$

The ensuing ReML estimates of $\Sigma(\lambda)^{(i)}$ can be used to compute the conditional density of the parameters in the usual way. For example, the conditional expectation and covariance of the ith level parameters $\theta^{(i)}$ are

$$\mu^{\theta(i)} = \Sigma^{\theta(i)}K^{(i)^T}\tilde{\Sigma}^{-1}y$$

$$\Sigma^{\theta(i)} = \left(K^{(i)^T}\tilde{\Sigma}^{-1}K^{(i)} + \Sigma^{(i)^{-1}}\right)^{-1}$$

$$\tilde{\Sigma} = \sum_{j\neq i} K^{(j)T}\Sigma^{(j)}K^{(j)T} \tag{35}$$

where $\tilde{\Sigma}$ represents the ReML estimate of error covariance, excluding the level of interest. This component $\Sigma^{(i)}=\Sigma(\lambda)^{(i)}$ is treated as an empirical prior on $\theta^{(i)}$. spm_ReML.m uses Eq. (27) to estimate the requisite hyperparameters. Critically, it takes as an argument the matrix yy^T. This may seem computationally inefficient. However, there is a special but very common case where dealing with yy^T is more appropriate than dealing with y (cf. the implementation using Eq. (33) in spm_peb.m):

This is when there are r multiple observations that can be arranged as a matrix $Y=[y_1,\ldots,y_r]$. If these observations are independent then we can express the covariance components of the vectorised response in terms of Kronecker tensor products

$$y = vec\{Y\} = \varepsilon$$

$$\varepsilon = \sum_{i=1}^{n} I \otimes K^{(i)}\varepsilon^{(i)}$$

$$\mathrm{cov}\left\{\varepsilon^{(i)}\right\} = I \otimes \Sigma^{(i)}. \tag{36}$$

This leads to a computationally efficient scheme employed by spm_ReML.m, which uses the compact forms[4]

$$\begin{aligned} L_{\lambda i} &= -\frac{1}{2}tr\left((I \otimes P_iR)\left(yy^T - I \otimes \Sigma\right)\left(I \otimes R^T\right)\right) \\ &= -\frac{r}{2}tr\left(P_iR\left(\frac{1}{r}YY^T - \Sigma\right)R^T\right) \\ \langle L_{\lambda\lambda ij}\rangle &= -\frac{1}{2}tr\left(I \otimes P_iR\Sigma P_jR\Sigma\right) \\ &= -\frac{r}{2}tr\left(P_iR\Sigma P_jR\Sigma\right). \end{aligned} \tag{37}$$

Critically, the update scheme is a function of the sample covariance of the data $(1/r)YY^T$ and can be regarded as a

[4] Note that we have retained the residual forming matrix R, despite the fact that there are no parameters. This is because in practice one usually models confounds as fixed effects at the first level. The residual forming matrix projects the data onto the null space of these confounds.

Please cite this article as: Karl Friston et al., Variational free energy and the Laplace approximation, NeuroImage (2006), doi:10.1016/j.neuroimage.2006.08.035

covariance component estimation scheme. This can be useful in two situations:

First, if the augmented form in Eq. (32) produces prohibitively long vectors. This can happen when the number of parameters is much greater than the number of responses. This is a common situation in underdetermined problems. An important example is source reconstruction in electroencephalography, where the number of sources is much greater than the number of measurement channels (see Phillips et al., 2005, for an application that uses spm_ReML.m in this context). In these cases one can form conditional estimates of the parameters using the matrix inversion lemma and again avoid inverting large ($p \times p$) matrices.

$$\begin{aligned}\mu^{\theta(i)} &= \Sigma^{(i)} K^{(i)^T} \tilde{\Sigma}^{-1} Y \\ \Sigma^{\theta(i)} &= \Sigma^{(i)} - \Sigma^{(i)} K^{(i)T} \tilde{\Sigma}^{-1} K^{(i)} \Sigma^{(i)} \\ \tilde{\Sigma} &= \sum_{i=1}^{n} K^{(i)T} \Sigma^{(i)} K^{(i)T}. \end{aligned} \tag{38}$$

The second situation is where there are a large number of realisations. In these cases it is much easier to handle the second-order matrices of the data YY^T than the data Y itself. An important application here is the estimation of nonsphericity over voxels in the analysis of fMRI time-series (see Friston et al., 2002, for this use of spm_ReML.m). Here, there are many more voxels than scans and it would not be possible to vectorise the data. However, it is easy to collect the sample covariance over voxels and partition it into nonspherical covariance components using ReML.

In the case of sequential correlations among the errors $\mathrm{cov}\{\varepsilon^{(i)}\} = V \otimes \Sigma^{(i)}$ one simply replaces YY^T with $YV^{-1}Y^T$. Heuristically, this corresponds to sequentially whitening the observations before computing their second order statistics. We have used this device in the Bayesian inversion of models of evoked and induced responses in EEG/MEG (Friston et al., 2006).

In summary, hierarchical models can be identified through ReML estimates of covariance components. If the response vector is relatively small it is generally more expedient to reduce the hierarchical form by augmentation, as in Eq. (32), and use Eq. (33) to compute the gradients. When the augmented form becomes too large, because there are too many parameters, reformulation in terms of covariance components is computationally more efficient because the gradients can be computed from the sample covariance of the data. The latter formulation is also useful when there are multiple realisations of the data because the sample covariance, over realisations, does not change in size. This leads to very fast Bayesian inversion. Both approaches rest on estimating covariance components that are induced by the observation hierarchy. This enforces a hyper-parameterisation of the covariances, as opposed to precisions (see Appendix 1).

Model selection with ReML

This section contains a brief demonstration of model selection using ReML and its adjusted free energy. In these examples, we use the covariance component formulation (spm_ReML.m) as in Eq. (34), noting exactly the same results would be obtained with augmentation (spm_peb.m). We use a simple hierarchical two-level linear model, implementing shrinkage priors, because this sort of model is common in neuroimaging data analysis and represents the simplest form of empirical Bayes. The model is described in Fig. 2.

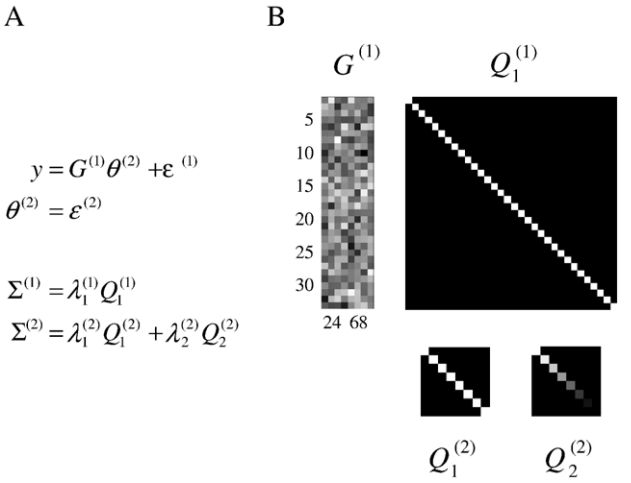

Fig. 2. A hierarchical linear model. (A) The form of the model with two levels. The first level has a single error covariance component, whereas the second has two. The second level places constraints on the parameters of the first, through the second-level covariance components. Conditional estimation of the hyperparameters, controlling these components, corresponds to an empirical estimate of their prior covariance (i.e., empirical Bayes). Because there is no second level design matrix the priors shrink the conditional estimates towards zero. These are known as shrinkage priors. (B) The design matrix and covariance components used to generate 128 realisations of the response variable y, using hyperparameters of unity for all components. The design matrix comprised random Gaussian variables.

Briefly it has eight parameters that cause a 32-variiate response. The parameters are drawn from a multivariate Gaussian that was a mixture of two known covariance components. Data were generated repeatedly (128 samples) using different parameters for each realization. This model can be regarded as generating fMRI data over 32 scans, each with 128 voxels; or EEG data from 32 channels over 128 time bins. These simulations are provided as a proof of concept and illustrate how one might approach numerical validation in the context of other models.

The free energy can, of course, be used for model selection when models differ in the number and deployment of parameters. This is because both F and F^θ are functions of the number of parameters and their conditional uncertainty. This can be shown by evaluating the free energy as a function of the number of model parameters, for the same data. The results of this sort of evaluation are seen in Fig. 3 and demonstrate that model selection correctly identifies a model with eight parameters. This was the model used to generate the data (Fig. 2). In this example, we used a simple shrinkage prior on all parameters (i.e., $\Sigma^{(2)} = \lambda^{(2)} I$) during the inversions.

The critical issue is whether model selection will work when the models differ in their hyperparameterisation. To address this, we analysed the same data, produced by two covariance components at the second level, with models that comprised an increasing number of second-level covariance components (Fig. 4). These components can be regarded as specifying the form of empirical priors over solution space (e.g., spatial constraints in an EEG source reconstruction problem). The results of these simulations show that the adjusted free energy F correctly identified the model with two components. Conversely, the unadjusted free energy F^θ rose progressively as the number of components and accuracy increased. See Fig. 5.

Please cite this article as: Karl Friston et al., Variational free energy and the Laplace approximation, NeuroImage (2006), doi:10.1016/j.neuroimage.2006.08.035

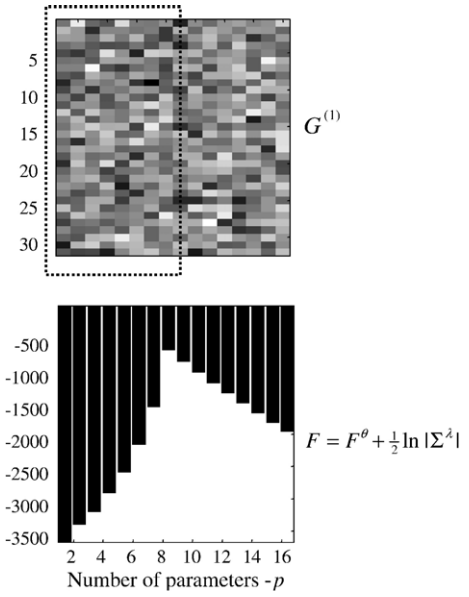

Fig. 3. Model selection in terms of parameters using ReML. The data generated by the eight-parameter model in Fig. 2 were analysed with ReML using a series of models with an increasing numbers of parameters. These models were based on the first p columns of the design matrix above. The profile of free energy clearly favours the model with eight parameters, corresponding to the design matrix (dotted line in upper panel) used to generate the data.

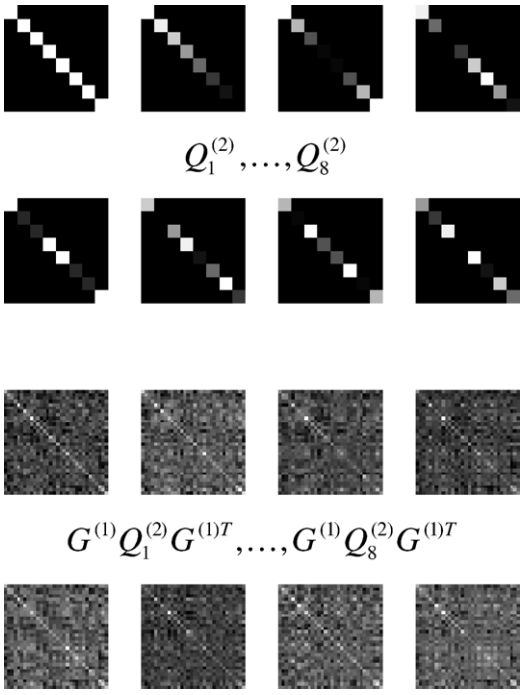

Fig. 4. Covariance components used to analyse the data generated by the model in Fig. 2. The covariance components are shown at the second level (upper panels) and after projection onto response space (lower panel) with the eight-parameter model. Introducing more covariance components creates a series models with an increasing number of hyperparameters, which we examined using model selection in Fig. 5. These covariance components were leading diagonal matrices, whose elements comprised a mean-adjusted discrete cosine set.

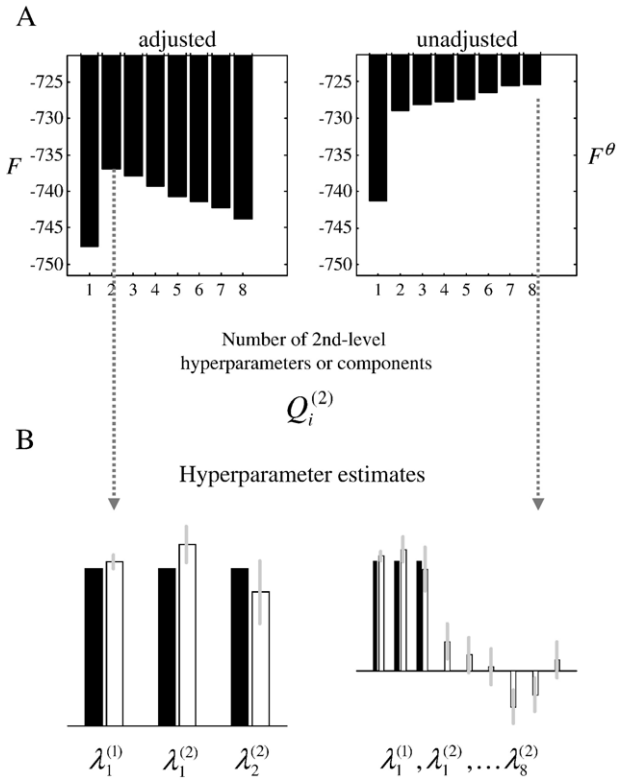

Fig. 5. Model selection in terms of hyperparameters using ReML. (A) The free energy was computed using the data generated by the model in Fig. 2 and a series of models with an increasing number of hyperparameters. The ensuing free energy profiles (adjusted—left; unadjusted—right) are shown as a function of the number of second-level covariance components used (from the previous figure). The adjusted profile clearly identified the correct model with two second-level components. (B) Conditional estimates (white) and true (black) hyperparameter values with 90% confidence intervals for the correct (3-component; left) and redundant (9-component; right) models.

The lower panel in Fig. 5 shows the hyperparameter estimates for two models. With the correctly selected model the true values fall within the 90% confidence interval. However, when the model is over-parameterised, with eight second-level components, this is not the case. Although the general profile of hyperparameters has been captured, this suboptimum model has clearly overestimated some hyperparameters and underestimated others.

Validation using MCMC

Finally, to establish that the variational approximation to the log-evidence is veridical, we computed $\ln p(y|\vartheta)$ using a standard Monte Carlo–Markov chain (MCMC) procedure, described in Appendix 2. MCMC schemes are computationally intensive but allow one to sample from the posterior distribution $p(\vartheta|y)$ without making any assumptions about its form. These samples can then be used to estimate the marginal likelihood using, in this instance, a harmonic mean (see Appendix 2). These resulting estimates are not biased by the mean-field and Laplace approximations implicit in the variational scheme and can be used to asses the impact of these approximations on model comparison. The sampling estimates of free energy are provided in Fig. 6 (upper panels) for the eight models analysed in Fig. 5. The profile of true [sampling] log-evidences concurs with the free-energy profile in a pleasing way and suggests that the approximations entailed by the variational

Please cite this article as: Karl Friston et al., Variational free energy and the Laplace approximation, NeuroImage (2006), doi:10.1016/j.neuroimage.2006.08.035

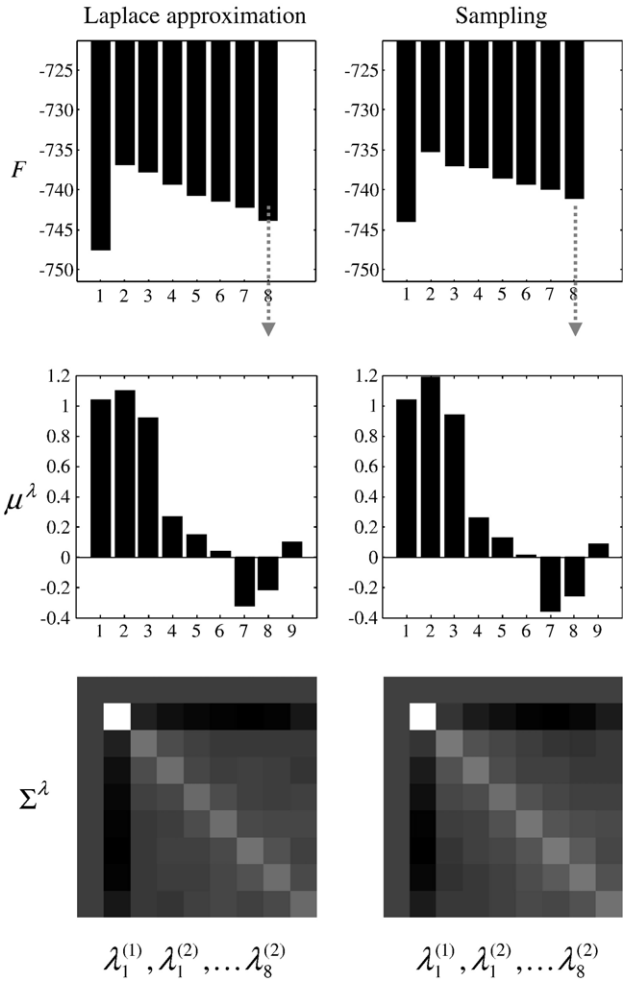

Fig. 6. Log-evidence or marginal likelihoods for the models in Fig. 5 estimated by ReML under the Laplace approximation (left) and a harmonic mean, based on samples from the posterior density using a Metropolis–Hasting sampling algorithm (right). These results should be compared with the free-energy approximations in the first panel of the previous figure (Fig. 5). Details of the sampling scheme can be found Appendix 2. The lower panels compare the two estimates of the conditional density in terms of their expectations and covariances. The agreement is self-evident.

approach do not lead to inaccurate model selection, under these linear models. Furthermore, the sampled posterior $p(\lambda|y)$ of the largest model (with eight second-level covariance components) is very similar to the Laplace approximation $q(\lambda)$ as judged by their first two moments (Fig. 6; lower panels).

Automatic model selection (AMS)

Hitherto, we have looked at model selection in terms of categorical comparisons of model evidence. However, the log-evidence is the same as the objective function used to optimise $q(\vartheta^i)$ for each model. Given that model selection and inversion maximise the same objective function; one might ask if inversion of an over-parameterised model finds the optimal model automatically. In other words, does maximising the free energy switch off redundant parameters and hyperparameters by setting their conditional density $q(\vartheta^i)$ to a point mass at zero; i.e., $\mu^i \to 0$ and $\Sigma^i \to 0$. Most fixed-form variational schemes do this; however, this is precluded in classical[5] ReML because the Laplace approximation admits improper, negative, covariance components, when these components are small. This means classical schemes cannot switch off redundant covariance components. Fortunately, it is easy to finesse this problem by applying the Laplace assumption to $\lambda = \ln\alpha$, where α are scale parameters encoding the expression of covariance components. This renders the form of $q(\alpha)$ log-normal and places a positivity constraint on α.

Log-normal hyperpriors

Consider the hyper-parameterisation

$$\Sigma = \sum_i \exp(\lambda_i) Q_i \tag{39}$$

with priors $p(\lambda) = N(\eta^\lambda, \Pi^{\lambda^{-1}})$. This corresponds to a transformation in which scale parameters $\alpha_i = \exp(\lambda_i)$ control the expression of each covariance component, Q_i. To first order, the conditional variance of each scale parameter is

$$\Sigma_i^\alpha = \frac{\partial \alpha_i}{\partial \lambda_i} \Sigma_i^\lambda \frac{\partial \alpha_i}{\partial \lambda_i}\bigg|_{\lambda = \mu_i^\lambda} = \mu_i^\lambda \Sigma_i^\alpha \mu_i^\alpha. \tag{40}$$

This means that when $\mu_i^\alpha = \exp(\mu_i^\lambda) \to 0$ we get $\Sigma_i^\alpha \to 0$, which is necessary for automatic model selection. In this limit, the conditional covariance Σ_i^λ is given by Eqs. (7) and (18)

$$\Sigma^\lambda = -L_{\lambda\lambda}^{-1} = \Pi^{\lambda^{-1}} \tag{41}$$

because $P_i = P_{ij} \to 0$ when $\exp(\mu_i^\lambda) \to 0$ (see Appendix 1). In short, by placing log-normal hyperpriors on α_i, we can use a conventional EM or ReML scheme for AMS. In this case, the conditional uncertainty about covariance components shrinks with their expectation, so that we can be certain they do not contribute to the model. It is simple to augment the ReML scheme and free energy to include hyperpriors (cf. Eq. (27))

until convergence: ReML-step

$$\begin{aligned} \varepsilon^\lambda &= \mu^\lambda - \eta^\lambda \\ \Delta\mu^\lambda &= -\left\langle L_{\lambda\lambda} - \Pi^\lambda \right\rangle^{-1} (L_{\lambda\lambda} - \Pi^\lambda \varepsilon^\lambda) \end{aligned}$$

end (42)

$$F = F^\theta + \frac{1}{2}\ln|\Sigma^\lambda| + \frac{1}{2}\ln|\Pi^\lambda| - \frac{1}{2}\varepsilon^{\lambda T}\Pi^\lambda\varepsilon^\lambda$$

Note, that adding redundant covariance components to the model does not change the free energy because the entropy associated with conditional uncertainty is offset exactly by the prior uncertainty they induce; due to the equality in Eq. (41).[6] This means that conventional model selection will show that all over-parameterised models are equally optimal. This is intuitive because the inversion has already identified the optimal model. Fig. 7 shows the hyperparameter estimates for the model

[5] In which covariances are hyper-parameterised as a linear mixture of covariance components.

[6] Assuming $\mu^\lambda = \eta^\lambda$, where the prior mean η^λ shrinks the conditional mean towards minus infinity (and the scale parameter to zero).

Please cite this article as: Karl Friston et al., Variational free energy and the Laplace approximation, NeuroImage (2006), doi:10.1016/j.neuroimage.2006.08.035

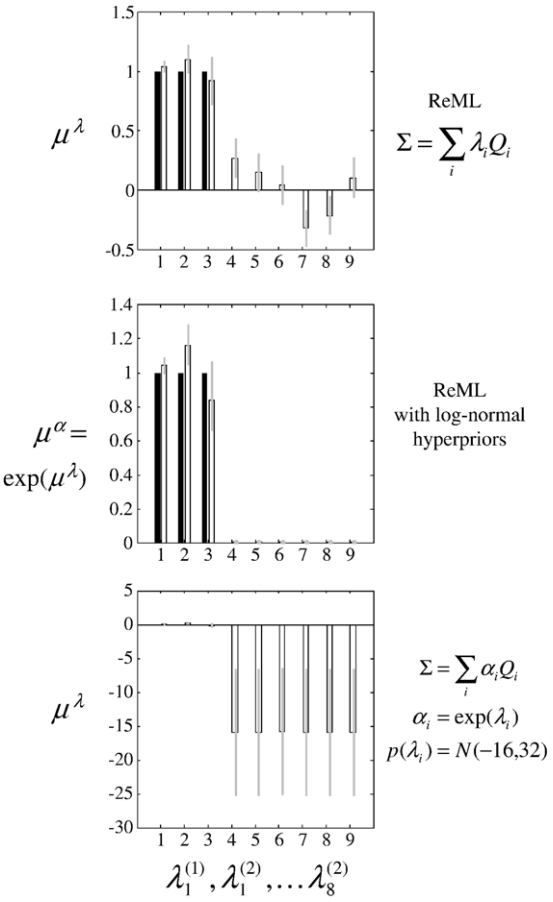

Fig. 7. Hyperparameter estimates for the model described in Fig. 5; with an increasing number of second-level covariance components. The upper panel show the conditional estimates with a classical hyper-parameterisation and the lower panels show the results with log-normal hyperpriors before (lower) and after (middle) log-transformation. True values are shown as filled bars and 90% confidence intervals are shown in light grey.

described in Fig. 5; with eight second-level covariance components. The upper panel shows the results with a classical hyper-parameterisation and the lower panels show the results with log-normal hyperpriors (before and after log-transformation). Note that hyperpriors are necessary to eliminate unnecessary components. In this example (and below) we used relatively flat hyperpriors; $p(\lambda)=N(-16,32)$.

Automatic relevance determination (ARD)

When automatic model selection is used to eliminate redundant parameters, it is known as automatic relevance determination (ARD). In ARD one defines an empirical prior on the parameters that embodies the notion of uncertain relevance. This enables the inversion to infer which parameters are relevant and which are not (MacKay, 1995a,b). This entails giving each parameter its own shrinkage prior and estimating an associated scale parameter. In the context of linear models, this is implemented simply by adding a level to induce empirical shrinkage priors on the parameters. ReML (with hyperpriors) can then be used to switch off redundant parameters by eliminating their covariance components at the first level. We provide an illustration of this in Fig. 8, using the models reported in Fig. 3. The lower panel shows the conditional estimates of the parameters using Eq. (35) for the over-parameterised model with 16 parameters. Note that ARD with ReML correctly shrinks and switches off the last eight redundant parameters and has implicitly performed AMS. The upper panel shows that the free energy of all models, with an increasing number of parameters, reaches a maximum at the correct parameterisation and stays there even when redundant parameters (i.e., components) are added.

Discussion

We have seen that restricted maximum likelihood is a special case of expectation maximisation and that expectation maximisation is a special case of variational Bayes. In fact, nearly every routine used in neuroimaging analysis (certainly in SPM5; http://www.fil.ion.ucl.ac.uk/spm) is a special case of variational Bayes, from ordinary least squares estimation to dynamic causal modelling. We have focussed on adjusting the objective functions used by EM and ReML to approximate the variational free energy under the Laplace approximation. This free energy is a lower bound approximation (exact for linear models) to the log-evidence, which plays a central role in model selection and averaging. This means one can use computationally efficient schemes like ReML for both model selection and Bayesian inversion.

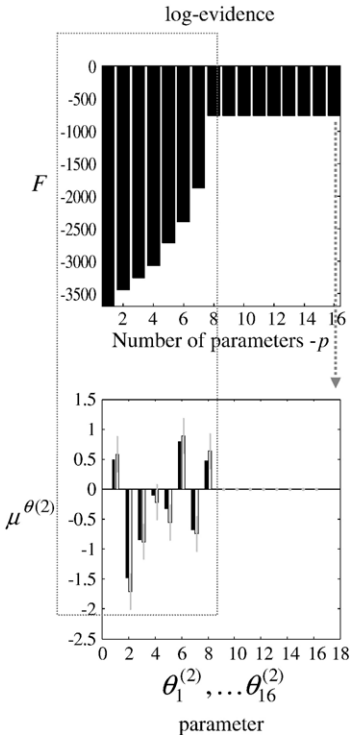

Fig. 8. Automatic relevance determination using ReML and the models reported in Fig. 3. The upper panel shows that the free energy reaches a maximum with the correct number of parameters and remains there, even when redundant parameters are added. The lower panel shows the conditional estimates (white bars) of the parameters (for the first sample) using Eq. (35). 90% confidence intervals are shown in light grey. These estimates are from an over-parameterised model with 16 parameters. The true values are depicted as filled bars. Note that ARD switches off the last eight redundant parameters (outside the box) and has implicitly performed a model selection.

Please cite this article as: Karl Friston et al., Variational free energy and the Laplace approximation, NeuroImage (2006), doi:10.1016/j.neuroimage.2006.08.035

Variational Bayes and the Laplace approximation

Variational inference finds itself between the conventional post hoc Laplace approximation and sampling methods (see Adami, 2003). At one extreme, the post hoc Laplace approximation, although simple, can be inaccurate and unwieldy in a high-dimensional setting (requiring large numbers of second-order derivatives). At the other extreme, we can approximate the evidence using numerical techniques such as MCMC methods (e.g., the Metropolis–Hastings algorithm used above). However, these are computationally intensive. Variational inference attempts to approximate the integrand to make the integral tractable. The basic idea is to bound the integral, reducing the integration problem to an optimisation problem, i.e., making the bound as tight as possible. No parameter estimation is required and the integral is optimised directly. The Kullback–Leibler cross-entropy or divergence measures the disparity between the true and approximate posterior and quantifies the loss of information incurred by the approximation. Variational Bayes under the Laplace approximation entails two approximations to the conditional density. The first is the factorisation implicit in the mean field approximation and the second in the Laplace assumption about the ensuing factors. Both can affect the estimation of the evidence.

Mean-field factorisation

Although a mean field factorisation of the posterior distribution may seem severe, one can regard it as replacing stochastic dependencies among ϑ^i with deterministic dependencies between their relevant moments (see Beal, 1998). The advantage of ignoring how fluctuations in ϑ^i induce fluctuations in ϑ^i (and vice-versa) is that we can obtain analytical free-form or fixed-form approximations to the log-evidence. These ideas underlie mean-field approximations from statistical physics, where lower-bounding variational approximations were conceived (Feynman, 1972). Using the bound for model selection and averaging rests on assumptions about the tightness of that bound: for example, the log-Bayes factor comparing two models m and m' is

$$\ln\frac{p(y|m)}{p(y|m')} = F - F' + D(q(\vartheta)\,||p(\vartheta|y,m)) - D(q'(\vartheta)\,||p(\vartheta|y,m')). \quad (39)$$

When we perform model selection by comparing the free energies, $F-F'$, we are assuming that the tightness or divergence of the two approximations are the same. Unfortunately, it is nontrivial to predict analytically how tight a particular bound is; if this were possible, we could estimate the marginal likelihood more accurately (Beal and Ghahramani, 2003). However, as illustrated above, sampling methods can be used to validate the free-energy estimates of log-evidence for a particular class of model. See Girolami and Rogers (2005) for an example of comparing Laplace and variational approximations to exact Inference via Gibbs sampling in the context of multinomial probit regression with Gaussian process priors.

The Laplace approximation

These arguments also apply to the Laplace approximation for each mean-field partition $q(\vartheta^i)$. However, this approximation is less severe for the models considered here. In the context of linear models, $q(\theta)$ is exactly Gaussian. Even for nonlinear or dynamic models there are several motivations for a Gaussian approximation. First, the large number of observations, encountered typically in neuroimaging, render the posterior nearly Gaussian, around its mode (Beal and Ghahramani, 2003). Second, Gaussian assumptions about errors and empirical priors in hierarchical models are motivated easily by the central limit theorem entailed by the averaging implicit in most imaging applications.

Priors and model selection

The log-evidence, and ensuing model selection, can depend on the choice of priors. This is an important issue because model selection could be dominated by the priors entailed by different models. This becomes acute when the priors change systematically with the model. An example of this is dynamic causal modelling, in which shrinkage priors are used to ensure stable dynamics. These priors become tighter as the number of connections among neuronal sources increases. This example is discussed in Penny et al. (2004), where the use of approximations to the log-evidence (Akaike and Bayesian information criteria; AIC and BIC) are used to provide consistent evidence in favour of one model over another. The AIC and BIC depend less on the priors. Generally, however, sensitivity to prior assumptions can be finessed by adopting noninformative hyperpriors. This involves optimising the priors per se, with respect to the free energy, by introducing hyperparameters that encode the prior density. The use of flat hyperpriors, on these hyperparameters, enables model comparison that is not confounded by prior assumptions: The section on AMS provided an example of this, which speaks to the usefulness of hierarchical models and empirical priors: In the example used above, the models differed only in the number of covariance components, each with flat hyperpriors on their expression.

In a subsequent publication (Henson et al., in preparation), we will illustrate the use of automatic model selection using ReML in the context of distributed source reconstruction. This example uses MEG data to localise responses to face processing and shows that a relatively simple model of both sensor noise and source-space priors supervenes over more elaborate models.

Acknowledgments

The Wellcome Trust and British Council funded this work.

Appendix A

A.1. Hyper-parameterising covariances

This appendix discusses briefly the various hyper-parameterisations one can use for the covariances of random effects. Recall that the variational scheme and EM become the same when $P_{ij}=\partial^2\Sigma/\partial\lambda_i\partial\lambda_j=0$. One can ensure $P_{ij}=0$ by adopting a hyper-parameterisation, where the precision is linear in the hyperparameters; for example, a linear mixture of precision components Q_i. Consider the more general parameterisation of precisions

$$\Sigma^{-1} = \sum_i f(\lambda_i)Q_i$$

$$P_i = f'(\lambda_i)Q_i$$

$$P_{ij} = \begin{cases} 0 & i \neq j \\ f''(\lambda_i)Q_i & i = j \end{cases} \quad \text{(A.1)}$$

Where $f(\lambda_i)$ is any analytic function. The simplest is $f(\lambda_i)=\lambda_i \Rightarrow f' \Rightarrow 1 = f'' = 0$. In this case VB and EM are formally identical. However, this allows negative contributions to the precisions, which can lead to improper covariances. Using $f(\lambda_i)=\exp(\lambda_i) \Rightarrow f''=f'=f$ precludes improper covariances. This hyper-parameterisation effectively implements a log-normal hyperprior, which imposes scale-invariant positivity constraints on the precisions. This is formally related to the use of conjugate [gamma] priors for scale parameters like $f(\lambda_i)$ (cf. Berger, 1985), when they are noninformative. Both imply a flat prior on the log-precision, which means its derivatives with respect to $\ln f(\lambda_i)=\lambda_i$ vanish (because it has no maximum). In short, one can either place a gamma prior on $f(\lambda_i)$ or a normal prior on $\ln f(\lambda_i)=\lambda_i$. These hyperpriors are the same when uninformative.

However, there are many models where is necessary to hyper-parameterise in terms of linear mixtures of covariance components

$$\Sigma = \sum_i f(\lambda_i)Q_i$$

$$P_i = -f'(\lambda_i)\Sigma^{-1}Q_i\Sigma^{-1}$$

$$P_{ij} = \begin{cases} 2P_i\Sigma P_j & i \neq j \\ 2P_i\Sigma P_i + \dfrac{f''(\lambda_i)}{f'(\lambda_i)}P_i & i = j \end{cases} \quad \text{(A.2)}$$

This is necessary when hierarchical generative models induce multiple covariance components. These are important models because they are central to empirical Bayes. See Harville (1977, p. 322) for comments on the usefulness of making the covariances linear in the hyperparameters; i.e., $f(\lambda_i)=\lambda_i \Rightarrow f' = 1 \Rightarrow f'' = 0$.

An important difference between these two hyper-parameterisations is that the linear mixture of precisions is conditionally convex (Mackay and Takeuchi, 1996), whereas the mixture of covariances is not. This means there may be multiple optima for the latter. See Mackay and Takeuchi (1996) for further covariance hyper-parameterisations and an analysis of their convexity. Interested readers may find the material in Leonard and Hsu (1992) useful further reading.

Appendix B

B.1. Estimating the marginal likelihood via MCMC sampling and the harmonic mean identity

Following Raftery et al. (2006); consider data y, a likelihood function $p(y|\vartheta,m)$ for a model m and a prior distribution $p(\vartheta|m)$. The integrated or marginal likelihood is

$$p(y|m) = \int p(y|\vartheta,m)p(\vartheta|m)d\vartheta. \quad \text{(A.3)}$$

The integrated or marginal likelihood is the normalising constant for the product of the likelihood and the prior in forming the posterior density $p(\vartheta|y)$. Evaluating the marginal likelihood can present a difficult computational problem, which has been the focus of this note. Newton and Raftery (1994) showed that the marginal likelihood can be expressed as an expectation with respect to the posterior distribution of the parameters, thus motivating an estimate based on a Monte Carlo sample from the posterior. By Bayes theorem

$$\frac{1}{p(y|m)} = \int \frac{p(\vartheta|y,m)}{p(y|\vartheta,m)}d\vartheta = E\left\{\frac{1}{p(y|\vartheta,m)}|y\right\} \quad \text{(A.4)}$$

Eq. (A.4) says that the marginal likelihood is the posterior harmonic mean of the likelihood. This suggests that the integrated likelihood can be approximated by the sample harmonic mean of the likelihoods

$$p(y|m) \approx \left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{p(y|\vartheta_i,m)}\right]^{-1} \quad \text{(A.5)}$$

based on N samples of ϑ_i from the posterior distribution. These samples can come from a standard MCMC implementation, for example a Metropolis–Hastings scheme.

B.2. Metropolis–Hastings (MH) sampling

MH involves the construction of a Markov chain whose equilibrium distribution is the desired posterior distribution. At equilibrium, a sample from the chain is a sample from the posterior. Note that the posterior distribution reconstructed in this way will not be constrained to be Gaussian or factorise over mean-field partitions, thereby circumventing the approximations of the variational scheme described in the main text.

This algorithm has the following recursive form, starting with an initial value ϑ_0 of the parameters (i.e., the prior expectation):

1. Propose ϑ_{i+1} from $\pi(\vartheta_{i+1}|\vartheta_i)$
2. Calculate the ratio $\alpha = \dfrac{L(\vartheta_{i+1})\pi(\vartheta_{i+1}|\vartheta_i)}{L(\vartheta_i)\pi(\vartheta_i|\vartheta_{i+1})}$ (A.6)
3. Accept or reject

$$\vartheta_{i+1} = \begin{cases} \vartheta_{i+1} & \alpha > 1 \\ \vartheta_i & \text{with probability } 1-\alpha \text{ otherwise} \end{cases}$$

Where $L(\vartheta_i)=p(y,\vartheta|m)$ is defined in Eq. (17). We use 256 'burn-in' iterations and 2^{16} samples. The proposal density was $\pi(\vartheta_{i+1}|\vartheta_i)=N(\vartheta_i, (1/32)I)$. This symmetric density is convenient because the proposal densities $\pi(\vartheta_{i+1}|\vartheta_i) = \pi(\vartheta_i|\vartheta_{i+1})$ in step 2 cancel, leading to a very simple algorithm.

References

Adami, K.Z., 2003. Variational Methods in Bayesian Deconvolution. PHYSTAT2003, SLAC, Stanford, California. September 8–11.

Ashburner, J., Friston, K.J., 2005. Unified segmentation. NeuroImage 26 (3), 839–851.

Beal, M.J., 1998. Variational algorithms for approximate Bayesian inference; PhD thesis: http://www.cse.buffalo.edu/faculty/mbeal/thesis/, p. 58.

Beal, M.J., Ghahramani, Z., 2003. The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. In: Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A. P., Heckerman, D., Smith, A.F.M., West, M. (Eds.), Bayesian Statistics. OUP, UK. Chapter 7.

Please cite this article as: Karl Friston et al., Variational free energy and the Laplace approximation, NeuroImage (2006), doi:10.1016/j.neuroimage.2006.08.035

Berger, J.O., 1985. Statistical Decision Theory and Bayesian Analysis, 2nd ed. Springer.
Bishop, C., 1999. Latent variable models. In: Jordan, M. (Ed.), Learning in Graphical Models. MIT Press, London, England.
David, O., Kiebel, S.J., Harrison, L.M., Mattout, J., Kilner, J.M., Friston, K.J., 2006. Dynamic causal modeling of evoked responses in EEG and MEG. NeuroImage 30, 1255–1272.
Dempster, A.P., Laird, N.M., Rubin, 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc., Ser. B 39, 1–38.
Efron, B., Morris, C., 1973. Stein's estimation rule and its competitors—An empirical Bayes approach. J. Am. Stat. Assoc. 68, 117–130.
Fahrmeir, L., Tutz, G., 1994. Multivariate Statistical Modelling Based on Generalised Linear Models. Springer-Verlag Inc, New York, pp. 355–356.
Feynman, R.P., 1972. Statistical Mechanics. Benjamin, Reading MA, USA.
Friston, K.J., 2002. Bayesian estimation of dynamical systems: an application to fMRI. NeuroImage 16, 513–530.
Friston, K., 2005. A theory of cortical responses. Philos. Trans. R. Soc. Lond., B Biol. Sci. 360, 815–836.
Friston, K.J., Penny, W., 2003. Posterior probability maps and SPMs. NeuroImage 19, 1240–1249.
Friston, K.J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., Ashburner, J., 2002. Classical and Bayesian inference in neuroimaging: theory. NeuroImage 16, 465–483.
Friston, K.J., Harrison, L., Penny, W., 2003. Dynamic causal modelling. NeuroImage 19, 1273–1302.
Friston, K.J., Stephan, K.E., Lund, T.E., Morcom, A., Kiebel, S., 2005. Mixed-effects and fMRI studies. NeuroImage 24, 244–252.
Friston, K.J., Henson, R., Phillips, C., Mattout, J., 2006. Bayesian estimation of evoked and induced responses. Human Brain Mapping (Feb 1st; electronic publication ahead of print).
Girolami, M., Rogers, S.,2005. Variational Bayesian Multinomial Probit Regression With Gaussian Process Priors Department of Computing Science; University of Glasgow. Technical Report: TR-2005-205.
Harrison, L.M., David, O., Friston, K.J., 2005. Stochastic models of neuronal dynamics. Philos. Trans. R. Soc. Lond. B, Biol. Sci. 360, 1075–1091.
Hartley, H., 1958. Maximum likelihood estimation from incomplete data. Biometrics 14, 174–194.
Harville, D.A., 1977. Maximum likelihood approaches to variance component estimation and to related problems. J. Am. Stat. Assoc. 72, 320–338.
Henson R.N., Mattout, J., Singh, K.D., Barnes, G.R., Hillebrand, A., Friston, K.J., in preparation. Group-based inferences for distributed source localisation using multiple constraints: application to evoked MEG data on face perception.
Hinton, G.E., von Cramp, D., 1993. Keeping neural networks simple by minimising the description length of weights. Proceedings of COLT-93, pp. 5–13.
Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Am. Stat. Assoc. 90, 773–795.
Kass, R.E., Steffey, D., 1989. Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). J. Am. Stat. Assoc. 407, 717–726.
Kiebel, S.J., David, O., Friston, K.J., 2006. Dynamic causal modelling of evoked responses in EEG/MEG with lead field parameterization. NeuroImage 30, 1273–1284.
Leonard, T., Hsu, J.S.L., 1992. Bayesian inference for a covariance matrix. Ann. Stat. 20, 1669–1696.
MacKay, D.J.C., 1995a. Probable networks and plausible predictions—A review of practical Bayesian methods for supervised neural networks. Netw.: Comput. Neural Syst. 6, 469–505.
MacKay, D.J.C., 1995b. Free energy minimisation algorithm for decoding and cryptoanalysis. Electron. Lett. 31, 445–447.
Mackay, D.J.C., Takeuchi, R., 1996. In: Skilling, J., Sibisi, S. (Eds.), Interpolation Models with Multiple Hyperparameters. Maximum Entropy and Bayesian Methods. Kluwer, pp. 249–257.
Mattout, J., Phillips, C., Rugg, M.D., Friston, K.J., 2005. MEG source localisation under multiple constraints: an extended Bayesian framework. *NeuroImage*, in press.
Neal, R.M., Hinton, G.E., 1998. A view of the EM algorithm that justifies incremental sparse and other variants. In: Jordan, M.I. (Ed.), Learning in Graphical Models. Kulver Academic Press.
Nelson, M.C., Illingworth, W.T., 1991. A Practical Guide to Neural Nets, Reading. Addison-Wesley, MA, p. 165.
Newton, M.A., Raftery, A.E., 1994. Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). J. R. Stat. Soc., Ser. B 56, 3–48.
Penny, W.D., Stephan, K.E., Mechelli, A., Friston, K.J., 2004. Comparing dynamic causal models. NeuroImage 22, 1157–1172.
Penny, W.D., Trujillo-Barreto, N.J., Friston, K.J., 2005. Bayesian fMRI time series analysis with spatial priors. NeuroImage 24, 350–362.
Phillips, C., Rugg, M., Friston, K.J., 2002. Systematic regularisation of linear inverse solutions of the EEG source localisation problem. NeuroImage 17, 287–301.
Phillips, C., Mattout, J., Rugg, M.D., Maquet, P., Friston, K.J., 2005. An empirical Bayesian solution to the source reconstruction problem in EEG. NeuroImage 24, 997–1011.
Raftery, A.E, Newton, M.A., Satagopan, J.M., Krivitsky, P.N., 2006. Estimating the Integrated Likelihood Via Posterior Simulation Using the Harmonic Mean Identity Technical Report No. 499 Department of Statistics University of Washington Seattle, Washington, USA. http://www.bepress.com/mskccbiostat/paper6.
Tipping, M.E., 2001. Sparse Bayesian learning and the relevance vector machine. J. Mach. Learn. Res. 1, 211–244.
Titantah, J.T., Pierlioni, C., Ciuchi, S., 2001. Free energy of the Fröhlich Polaron in two and three dimensions. Phys. Rev. Lett. 87, 206–406.
Trujillo-Barreto, N., Aubert-Vazquez, E., Valdes-Sosa, P., 2004. Bayesian model averaging. NeuroImage 21, 1300–1319.
Weissbach, F., Pelster, A., Hamprecht, 2002. High-order variational perturbation theory for the free energy. Phys. Rev. Lett. 66, 036129.

Please cite this article as: Karl Friston et al., Variational free energy and the Laplace approximation, NeuroImage (2006), doi:10.1016/j.neuroimage.2006.08.035