A P P E N D I X

4

Variational Bayes under the Laplace approximation

K. Friston, J. Mattout, N. Trujillo-Barreto, J. Ashburner and W. Penny

INTRODUCTION

This is a rather technical appendix, but usefully connects most of the estimation and inference schemes used in previous chapters, by showing they are all special cases of a single variational approach. This synthesis is underpinned by reference to the various applications we have considered in detail, so that their motivation and interrelationships are more evident.

We will derive 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 maximization, 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 difference covariance components. This is useful in the context of hierarchical models, because it enables a principled selection of priors. 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.

Background

This chapter starts with a very general formulation of inference using approaches developed in statistical physics. It ends with a treatment of a specific objective function used in restricted maximum likelihood (ReML) that renders it equivalent to the free energy in variational learning. 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 (see Penny *et al.*, 2004 and Trujillo-Barreto *et al.*, 2004, for examples in neuroimaging).

Although this appendix focuses on the various forms for the free energy, we use it to link variational Bayes (VB), EM and ReML using the Laplace approximation. This approximation assumes a fixed Gaussian form for the conditional density of the parameters 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 and Chapter 24). Its use ranges from spatial segmentation and normalization of images during pre-processing (e.g. Ashburner and Friston, 2005) to the inversion of complicated dynamical causal models of functional integration in the brain (Friston *et al.*, 2003 and Chapter 34). Many of the intervening steps in classical and Bayesian analysis of neuroimaging data call on ReML or EM under the Laplace approximation. This appendix provides an overview of how these schemes are related and illustrates their applications with reference to specific algorithms and routines we have referred to in this book. 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.

Conditional uncertainty

In previous chapters, we have described the use of ReML in the Bayesian inversion of electromagnetic models to

Statistical Parametric Mapping, by Karl Friston et al.
ISBN–13: 978-0-12-372560-8 ISBN–10: 0-12-372560-7

localize distributed sources in electroencephalography (EEG) and magnetoencephalography (MEG) (e.g. Phillips *et al.*, 2002; Chapters 29 and 30). 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). We have also used ReML to identify the most likely combination of priors using model selection, where each model comprises a different set of priors (Mattout *et al.*, 2006). This was based on the fact that the ReML objective function is the free energy used in expectation maximization 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 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 (i.e. priors). In what follows we 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 maximization (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.

Overview

This appendix is divided into six sections. In the first, we summarize the basic theory of variational Bayes and apply it in the context of the Laplace approximation (see also Chapter 24). 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 generalizes nicely to cover dynamic models and provides the basis for a temporal regularization of the ascent. In the second section, we apply the theory to non-linear 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 final 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.

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 (or marginal 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.

Variational approaches rest on minimizing 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 through ensemble learning, where the ensemble or variational density $q(\theta)$ (i.e. approximating posterior density) is optimized to minimize the free energy. Initially (Hinton and von Camp, 1993; MacKay, 1995), 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 appendix 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:

$$\begin{aligned} \ln p(y|m) &= F + D(q(\vartheta) \| p(\vartheta|y, m)) \\ F &= \langle L(\vartheta) \rangle_q - \langle \ln q(\vartheta) \rangle_q \\ L &= \ln p(y, \vartheta) \end{aligned} \qquad \textbf{A4.1}$$

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. Eqn. **A4.1** indicates that F is a lower-bound approximation to the log-evidence because the divergence $D(q(\vartheta) \| p(\vartheta|y, m))$ 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 maximizing F, and then compute F itself, for Bayesian inference and model comparison respectively. Maximizing the free energy minimizes 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 maximization easier, one usually assumes $q(\vartheta)$ factorizes over sets of parameters ϑ^i:

$$q(\vartheta) = \prod_i q^i \qquad \textbf{A4.2}$$

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

$$\begin{aligned} \delta F^i = 0 &\Leftrightarrow \frac{\partial f^i}{\partial q^i} = f^i_{q^i} = 0 \\ f^i &= F_{\vartheta^i} \end{aligned} \qquad \textbf{A4.3}$$

δF^i is the variation of the free energy with respect to q^i. From Eqn. **A4.1**:

$$\begin{aligned} f^i &= \int q^i q^{\backslash i} \ln L(\vartheta) d\vartheta^{\backslash i} - \int q^i q^{\backslash i} \ln q(\vartheta) d\vartheta^{\backslash i} \\ f^i_{q^i} &= I(\vartheta^i) - \ln q^i - \ln Z^i \\ I(\vartheta^i) &= \langle L(\vartheta) \rangle_{q^{\backslash i}} \end{aligned} \qquad \textbf{A4.4}$$

where $\vartheta^{\backslash i}$ denotes the parameters not in the i-th set. We have lumped terms that do not depend on ϑ^i into $\ln Z^i$, where Z^i is a normalization constant (i.e. partition function). We will call $I(\vartheta^i)$ the variational energy, noting its expectation under q^i is the expected energy. The extremal condition in Eqn. **A4.2** is met when:

$$\begin{aligned} \ln q^i &= I(\vartheta^i) - \ln Z^i \Leftrightarrow \\ q(\vartheta^i) &= \frac{1}{Z^i} \exp(I(\vartheta^i)) \end{aligned} \qquad \textbf{A4.5}$$

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.

The 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 i-th set of parameters. The advantage of this is that the conditional covariance can be evaluated very simply. Under the Laplace assumption:

$$\begin{aligned} F &= L(\mu) + \frac{1}{2} \sum_i (U^i + \ln |\Sigma^i| + p^i \ln 2\pi e) \\ I(\vartheta^i) &= L(\vartheta^i, \mu^{\backslash i}) + \frac{1}{2} \sum_{j \neq i} U^j \\ U^i &= tr(\Sigma^i L_{\vartheta^i \vartheta^i}) \end{aligned} \qquad \textbf{A4.6}$$

$p^i = \dim(\vartheta^i)$ is the number of parameters in the i-th set. The approximate conditional covariances obtain as an analytic function of the modes by differentiating Eqn. **A4.6** and solving for zero:

$$\begin{aligned} F_{\Sigma^i} &= \frac{1}{2} L_{\vartheta^i \vartheta^i} + \frac{1}{2} \Sigma^{i^{-1}} = 0 \Rightarrow \\ \Sigma^i &= -L(\mu)^{-1}_{\vartheta^i \vartheta^i} \end{aligned} \qquad \textbf{A4.7}$$

Note that this solution for the conditional covariances does not depend on the mean-field approximation, but only on the Laplace approximation. Substitution into Eqn. **A4.6** means $U^i = p^i$ and:

$$F = L(\mu) + \sum_i \frac{1}{2} (\ln |\Sigma^i| + p^i \ln 2\pi) \qquad \textbf{A4.8}$$

The only remaining quantities required are the variational modes which, from Eqn. **A4.5** maximize $I(\vartheta^i)$. The

leads to the following compact variational scheme, under the Laplace approximation:

until convergence

for all i

$$\mu^i = \max_{\vartheta^i} I(\vartheta^i)$$

$$\Sigma^i = -L(\mu)^{-1}_{\vartheta^i\vartheta^i}$$

end

end **A4.9**

The variational modes

The modes can be found using a gradient ascent based on:

$$\dot{\mu}^i = \frac{\partial I(\mu^i)}{\partial \vartheta^i} = I(\mu^i)_{\vartheta^i} \qquad \textbf{A4.10}$$

It may seem odd to formulate an ascent in terms of the motion of the mode in time. However, this is useful when generalizing to dynamic models (see below). The updates for the mode obtain by integrating Eqn. **A4.10** to give:

$$\begin{aligned} \Delta\mu^i &= (\exp(tJ) - I)J^{-1}\dot{\mu}^i \\ J &= \frac{\partial \dot{\mu}^i}{\partial \vartheta^i} = I(\mu^i)_{\vartheta^i\vartheta^i} \end{aligned} \qquad \textbf{A4.11}$$

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

$$\Delta\mu^i = -I(\mu^i)^{-1}_{\vartheta^i\vartheta^i} I(\mu^i)_{\vartheta^i} \qquad \textbf{A4.12}$$

Together with the expression for the conditional covariance in Eqn. **A4.7**, this update furnishes a variational scheme under the Laplace approximation:

until convergence

for all i

until convergence

$$\begin{aligned} I(\mu^i)_{\vartheta^i_k} &= L(\mu)_{\vartheta^i_k} + \frac{1}{2}\sum_{j\neq i} tr(\Sigma^j L_{\vartheta^j\vartheta^j\vartheta^i_k}) \\ I(\mu^i)_{\vartheta^i_k\vartheta^i_l} &= L(\mu)_{\vartheta^i_k\vartheta^i_l} + \frac{1}{2}\sum_{j\neq i} tr(\Sigma^j L_{\vartheta^j\vartheta^j\vartheta^i_k\vartheta^i_l}) \\ \Delta\mu^i &= -I(\mu^i)^{-1}_{\vartheta^i\vartheta^i} I(\mu^i)_{\vartheta^i} \end{aligned} \qquad \textbf{A4.13}$$

end

$$\Sigma^i = -L(\mu)^{-1}_{\vartheta^i\vartheta^i}$$

end

end

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

Regularizing 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 Gauss-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 regularization is required to ensure a robust ascent. This can be implemented by augmenting Eqn. **A4.10** with a decay term:

$$\dot{\mu}^i = I(\mu^i)_{\vartheta^i} - \nu\Delta\mu^i \qquad \textbf{A4.14}$$

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. Eqn. **A4.11**):

$$\begin{aligned} \Delta\mu^i &= -J^{-1}\dot{\mu}^i \\ &= (\nu I - I(\mu^i)_{\vartheta^i\vartheta^i})^{-1} I(\mu^i)_{\vartheta^i} \\ J &= I(\mu^i)_{\vartheta^i\vartheta^i} - \nu I \end{aligned} \qquad \textbf{A4.15}$$

where ν is the Levenburg-Marquardt regularization. However, the dynamic formulation affords a simpler alternative, namely temporal regularization. Here, instead of constraining the search with a decay term, one can abbreviate it by terminating the ascent after some suitable period $t = \nu$; from Eqn. **A4.11**:

$$\begin{aligned} \Delta\mu^i &= (\exp(\nu J) - I)J^{-1}\dot{\mu}^i \\ &= (\exp(\nu I(\mu^i)_{\vartheta^i\vartheta^i}) - I) I(\mu^i)^{-1}_{\vartheta^i\vartheta^i} I(\mu^i)_{\vartheta^i} \\ J &= I(\mu^i)_{\vartheta^i\vartheta^i} \end{aligned} \qquad \textbf{A4.16}$$

This has the advantage of using the local gradients and curvatures while precluding large excursions from the expansion point. In our implementations $\nu = 1/\eta$ is based on the 2-norm of the curvature η for both regularization 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

[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.

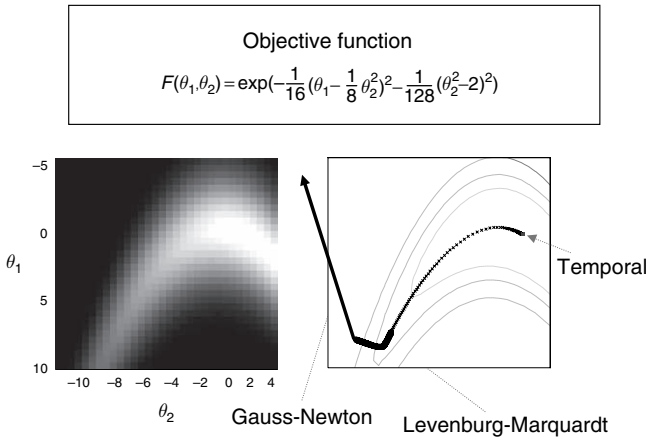

FIGURE A4.1 Examples of Levenburg-Marquardt and temporal regularization. The left panel shows an image of the landscape defined by the objective function $F(\theta_1, \theta_2)$ of two parameters (upper panel). This landscape was chosen because it is difficult for conventional schemes exhibiting curvilinear 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 regularization parameter was the 2-norm of the curvature evaluated at each update. Note how the ascent goes off in the wrong direction with no regularization (Gauss-Newton). The regularization adopted by Levenburg-Marquardt makes its progress slow, in relation to the temporal regularization, so that it fails to attain the maximum after 256 iterations.

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). It is interesting to note that 'early stopping' is closely related to ridge regression, which is another perspective on Levenburg-Marquardt regularization.

A comparative example using Levenburg-Marquardt and temporal regularization is provided in Figure A4.1 and suggests temporal regularization is better, in this example. Either approach can be implemented in the VB scheme by simply regularizing the Gauss-Newton update if the variational energy $I(\vartheta^i)$ fails to increase after each iteration. We prefer temporal regularization 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 Eqn. **A4.16** is determined by the interval between observations. This is the approach taken in our variational treatment of dynamic systems, namely, dynamic expectation maximization or DEM (introduced briefly in Friston *et al.*, 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. $I(\vartheta^i) \to I(\vartheta^i, \dot{\vartheta}^i, \ldots, t)$. This entails the extension of the variational density to cover this motion, using generalized coordinates $q(\vartheta^i) \to q(\vartheta^i, \dot{\vartheta}^i, \ldots, t)$. Dynamic schemes are important for the identification of stochastic dynamic casual models. However, the applications considered in this book are restricted to deterministic systems, without random fluctuations in the hidden states, and so we will 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 NON-LINEAR MODELS

Consider the non-linear generative model with additive error $y = G(\theta) + \varepsilon$. 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 $\Sigma(\lambda)$ are non-linear functions that can embody a probability integral transform (i.e. can implement a re-parameterization in terms of non-Gaussian priors).

Given n samples, p parameters and h hyperparameters:

$$
\begin{aligned}
L(\theta) = & \\
& -\frac{1}{2}\varepsilon^T\Sigma^{-1}\varepsilon + \frac{1}{2}\ln\left|\Sigma^{-1}\right| - \frac{n}{2}\ln 2\pi \\
& -\frac{1}{2}\varepsilon^{\theta T}\Pi^\theta\varepsilon^\theta + \frac{1}{2}\ln\left|\Pi^\theta\right| - \frac{p}{2}\ln 2\pi \\
& -\frac{1}{2}\varepsilon^{\lambda T}\Pi^\lambda\varepsilon^\lambda + \frac{1}{2}\ln\left|\Pi^\lambda\right| - \frac{h}{2}\ln 2\pi \\
\varepsilon = & \; G(\mu^\theta) - y \\
\varepsilon^\theta = & \; \mu^\theta - \eta^\theta \\
\varepsilon^\lambda = & \; \mu^\lambda - \eta^\lambda
\end{aligned}
\qquad \textbf{A4.17}
$$

and

$$
\begin{aligned}
L_\theta &= -G_\theta^T \Sigma^{-1} \varepsilon - \Pi^\theta \varepsilon^\theta \\
L_{\theta\theta} &= -G_\theta^T \Sigma^{-1} G_\theta - \Pi^\theta \\
L_{\lambda i} &= -\frac{1}{2} tr(P_i(\varepsilon\varepsilon^T - \Sigma)) - \Pi^\lambda_{i\bullet} \varepsilon^\lambda \\
L_{\lambda\lambda ij} &= -\frac{1}{2} tr(P_{ij}(\varepsilon\varepsilon^T - \Sigma)) - \frac{1}{2} tr(P_i \Sigma P_j \Sigma) - \Pi^\lambda_{ij} \qquad \textbf{A4.18} \\
P_i &= \frac{\partial \Sigma^{-1}}{\partial \lambda_i} \qquad P_{ij} = \frac{\partial^2 \Sigma^{-1}}{\partial \lambda_i \partial \lambda_j}
\end{aligned}
$$

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

$$
\begin{aligned}
I(\theta)_{\theta k} &= L(\theta, \mu^\lambda)_{\theta k} + \frac{1}{2} tr(\Sigma^\lambda A^k) & I(\lambda)_{\lambda i} &= L(\mu^\theta, \lambda)_{\lambda i} + \frac{1}{2} tr(\Sigma^\theta C^i) \\
I(\theta)_{\theta\theta kl} &= L(\theta, \mu^\lambda)_{\theta\theta kl} + \frac{1}{2} tr(\Sigma^\lambda B^{kl}) & I(\lambda)_{\lambda\lambda ij} &= L(\mu^\theta, \lambda)_{\lambda\lambda ij} + \frac{1}{2} tr(\Sigma^\theta D^{ij}) \\
A^k_{ij} &= -G^T_{\theta\bullet k} P_{ij} \varepsilon & C^i &= -G_\theta^T P_i G_\theta \\
B^{kl}_{ij} &= -G^T_{\theta\bullet k} P_{ij} G_{\theta\bullet l} & D^{ij} &= -G_\theta^T P_{ij} G_\theta
\end{aligned}
\qquad \textbf{A4.19}
$$

where $G_{\theta\bullet k}$ denotes the k-th column of G_θ. These enter the VB scheme in Eqn. **A4.13**, giving the two-step scheme:

until convergence

  until convergence

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

  end

  until convergence

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

  end

end **A4.20**

The negative free energy for these models is:

$$
\begin{aligned}
F = &-\frac{1}{2} \varepsilon^T \Sigma^{-1} \varepsilon + \frac{1}{2} \ln |\Sigma^{-1}| - \frac{n}{2} \ln 2\pi \\
&-\frac{1}{2} \varepsilon^{\theta T} \Pi^\theta \varepsilon^\theta + \frac{1}{2} \ln |\Pi^\theta| + \frac{1}{2} \ln |\Sigma^\theta| \qquad \textbf{A4.21} \\
&-\frac{1}{2} \varepsilon^{\lambda T} \Pi^\lambda \varepsilon^\lambda + \frac{1}{2} \ln |\Pi^\lambda| + \frac{1}{2} \ln |\Sigma^\lambda|
\end{aligned}
$$

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 maximization.

EXPECTATION MAXIMIZATION FOR NON-LINEAR 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 parameterize a density. In many instances, their conditional density *per se* is uninteresting. In variational expectation maximization, we ignore uncertainty about the hyperparameters and assume $q(\lambda)$ is a point mass (i.e. $\Sigma^\lambda = 0$). In this case, the free energy is effectively conditioned on λ and reduces to:

$$
\begin{aligned}
F^\lambda &= \ln p(y|\lambda) - D(q(\theta)||p(\theta|y, \lambda)) \\
&= \\
&-\frac{1}{2} \varepsilon^T \Sigma^{-1} \varepsilon + \frac{1}{2} \ln |\Sigma^{-1}| - \frac{n}{2} \ln 2\pi \qquad \textbf{A4.22} \\
&-\frac{1}{2} \varepsilon^{\theta T} \Pi^\theta \varepsilon^\theta + \frac{1}{2} \ln |\Pi^\theta| + \frac{1}{2} \ln |\Sigma^\theta|
\end{aligned}
$$

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 maximizes the likelihood of the hyperparameters $\ln p(y|\lambda)$ and becomes an **M**-step. In this context, Eqn. **A4.20** simplifies because we can ignore the terms that involve Σ^λ and Π^λ to give:

until convergence

  until convergence: **E-step**

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

  end

until convergence: **M-step**

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

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

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

end

end **A4.23**

Expectation maximization 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 generalized by Dempster *et al.* (1977) (see Appendix 3 for more details). 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 non-linear system identification (see Chapter 34) 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 haemodynamic models in terms of biophysical parameters for regional responses and dynamic causal models (DCMs) of distributed responses in a variety of applications, e.g. functional magnetic resonance imaging (fMRI) (Friston *et al.*, 2003 and Chapter 41), EEG (David *et al.*, 2005 and Chapter 42), MEG (Kiebel *et al.*, 2006), and mean-field models of neuronal activity (Harrison *et al.*, 2005 and Chapter 31).

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 Eqn. **A4.19** 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 hyperparameterization, which *renders the precision linear in the hyperparameters*, e.g. a linear mixture of precision components Q_i (see below). This resulting variational scheme is used by the SPM5 version of spm_nlsi.m for non-linear system identification.

Hyperparameterizing precisions

One can ensure $P_{ij} = 0$ by adopting a hyperparameterization, where the precision is linear in the hyperparameters, e.g. a linear mixture of precision components Q_i. Consider the more general parameterization of precisions:

$$\begin{aligned} \Sigma^{-1} &= \sum_i f(\lambda_i) Q_i \\ P_i &= f'(\lambda_i) Q_i \\ P_{ij} &= \begin{cases} 0 & i \neq j \\ f''(\lambda_i) Q_i & i = j \end{cases} \end{aligned} \qquad \textbf{A4.24}$$

where $f(\lambda_i)$ is any analytic function. The simplest is $f(\lambda_i) = \lambda_i \Rightarrow f' = 1 \Rightarrow f'' = 0$. 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 hyperparameterization 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 non-informative. 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 it is necessary to hyperparameterize in terms of linear mixtures of covariance components:

$$\begin{aligned} \Sigma &= \sum_i f(\lambda_i) Q_i \\ P_i &= -f'(\lambda_i) \Sigma^{-1} Q_i \Sigma^{-1} \\ P_{ij} &= \begin{cases} 2P_i \Sigma P_j & i \neq j \\ 2P_i \Sigma P_i + \frac{f''(\lambda_i)}{f'(\lambda_i)} P_i & i = j \end{cases} \end{aligned} \qquad \textbf{A4.25}$$

This is necessary when hierarchical generative models induce multiple covariance components. These are important models because they are central to empirical Bayes (see Chapter 22). See Harville (1977) 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 hyperparameterizations 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 hyperparameterizations and an analysis of their convexity. Interested readers may find the material in Leonard and Hsu (1992) useful further reading.

The second key point that follows from the variational treatment 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 Eqn. **A4.22** discounts uncertainty about the hyperparameters because it is conditioned upon them. This is a well-recognized problem, sometimes referred to as the overconfidence 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 Eqn. **A4.21** and Eqn. **A4.22**, we can express the variational free energy in terms of F^λ and an extra term from Eqn. **A4.18**:

$$\begin{aligned} F &= F^\lambda + \frac{1}{2}\ln|\Sigma^\lambda| \\ \Sigma^\lambda_{ij} &= -L(\mu)^{-1}_{\lambda\lambda} \end{aligned} \qquad \textbf{A4.26}$$

Intuitively, the extra term encodes the conditional information (i.e. entropy) about the model's covariance components. The log-evidence will only increase if an extra component adds information. Adding redundant components will have no effect on F. This term can be regarded as additional Occam factor (Mackay and Takeuchi, 1996). Adjusting the EM free energy to approximate the log-evidence is important because of the well-know connections between EM for linear models and restricted maximum likelihood. This connection suggests that the ReML objective function 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:

$$\begin{aligned} F^\theta &= \ln p(y|\lambda) - D(q(\theta)||p(\theta|y,\lambda)) \\ &= -\frac{1}{2}\varepsilon^T\Sigma^{-1}\varepsilon + \frac{1}{2}\ln\left|\Sigma^{-1}\right| - \frac{n}{2}\ln 2\pi + \frac{1}{2}\ln\left|\Sigma^\theta\right| \end{aligned} \qquad \textbf{A4.27}$$

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

$$\begin{aligned} \mu^\theta &= \Sigma^\theta G^T\Sigma^{-1}y \\ \Sigma^\theta &= (G^T\Sigma^{-1}G)^{-1} \end{aligned}$$

to eliminate the divergence term and give:

$$\begin{aligned} F^\theta &= \ln p(y|\lambda) \\ &= -\frac{1}{2}tr(\Sigma^{-1}Ryy^TR^T) + \frac{1}{2}\ln\left|\Sigma^{-1}\right| - \frac{n}{2}\ln 2\pi \\ &\quad - \frac{1}{2}\ln\left|G^T\Sigma^{-1}G\right| \\ \varepsilon &= Ry \\ R &= I - G(G^T\Sigma^{-1}G)^{-1}G^T\Sigma^{-1} \end{aligned} \qquad \textbf{A4.28}$$

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:

until convergence: ReML-step

$$\begin{aligned} L(\mu)_{\lambda i} &= -\frac{1}{2}tr(P_iR(yy^T - \Sigma)R^T) \\ \left\langle L(\mu)_{\lambda\lambda ij}\right\rangle &= -\frac{1}{2}tr(P_iR\Sigma P_jR\Sigma) \\ \Delta\mu^\lambda &= -\left\langle L(\mu)_{\lambda\lambda}\right\rangle^{-1} L(\mu)_\lambda \end{aligned} \qquad \textbf{A4.29}$$

end

Notice that the energy has replaced the variational energy because they are the same: from Eqn. **A4.6** $I(\vartheta) = L(\lambda)$. This is a result of eliminating $q(\theta)$ from the variational density. Furthermore, the curvature has been replaced by its expectation to render the Gauss-Newton descent a Fisher-Scoring scheme using:

$$\left\langle Ryy^TR^T\right\rangle = R\Sigma R^T = R\Sigma \qquad \textbf{A4.30}$$

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

$$\begin{aligned} F &= F^\theta + \frac{1}{2}\ln|\Sigma^\lambda| \\ \Sigma^\lambda_{ij} &= -\left\langle L(\mu)_{\lambda\lambda}\right\rangle^{-1} \end{aligned} \qquad \textbf{A4.31}$$

The conditional covariance of the hyperparameters uses the same curvature as the ascent in Eqn. **A4.29**. 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 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 (see Chapters 11 and 22) are equivalent to parametric empirical Bayes models (Efron and Morris, 1973) in which empirical priors emerge from conditional independence of the errors $\varepsilon^{(i)} \sim N(0, \Sigma^{(i)})$:

$$\begin{aligned} y^{(1)} &= & y^{(1)} &= \varepsilon^{(1)} \\ \theta^{(1)} &= G^{(1)}\theta^{(2)} + \varepsilon^{(1)} & &+ G^{(1)}\varepsilon^{(2)} \\ \theta^{(2)} &= G^{(2)}\theta^{(3)} + \varepsilon^{(2)} \equiv & &+ G^{(1)}G^{(2)}\varepsilon^{(3)} \\ &\vdots & &\vdots \\ \theta^{(n)} &= \varepsilon^{(n)} & &+ G^{(1)}\cdots G^{(n-1)}\theta^{(n)} \end{aligned} \qquad \textbf{A4.32}$$

In hierarchical models, the random terms model uncertainty 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.[2] ReML identification of simple two-level models like:

$$\begin{aligned} y^{(1)} &= G^{(1)}\theta^{(2)} + \varepsilon^{(1)} \\ \theta^{(2)} &= \varepsilon^{(2)} \end{aligned} \qquad \textbf{A4.33}$$

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. 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.

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:

$$\begin{aligned} y &= J\theta + \varepsilon \\ y &= \begin{bmatrix} y^{(1)} \\ 0 \\ \vdots \\ 0 \end{bmatrix} J = \begin{bmatrix} K^{(2)} & \cdots & K^{(n)} \\ -I & & \\ & \ddots & \\ & & -I \end{bmatrix} \varepsilon = \begin{bmatrix} \varepsilon^{(1)} \\ \varepsilon^{(2)} \\ \vdots \\ \varepsilon^{(n)} \end{bmatrix} \theta = \begin{bmatrix} \varepsilon^{(2)} \\ \vdots \\ \theta^{(n)} \end{bmatrix} \\ K^{(i)} &= \prod_{j=1}^{i} G^{(j-1)} \\ \Sigma &= \begin{bmatrix} \Sigma^{(1)} & & \\ & \ddots & \\ & & \Sigma^{(n)} \end{bmatrix} \end{aligned} \qquad \textbf{A4.34}$$

with $G^{(0)} = I$. This reformulation is a non-hierarchical 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 Eqn. **A4.29** for estimation. The conditional expectations of the parameters are recovered by recursive substitution of the conditional expectations of the errors into Eqn. **A4.33** (cf. Friston, 2002). spm_peb.m uses a computationally efficient substitution:

$$\frac{1}{2}tr(P_iR(yy^T - \Sigma)R^T) = \frac{1}{2}y^TR^TP_iRy - \frac{1}{2}tr(P_iR\Sigma R^T) \qquad \textbf{A4.35}$$

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 Chapter 23) 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 Eqn. **A4.30**

[2] For an introduction to EM algorithms in generalized 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).

in which all the hierarchically induced covariance components $K^{(i)T}\Sigma^{(i)}K^{(i)T}$ are treated as components of a compound error:

$$\begin{aligned} y &= \varepsilon \\ y &= y^{(1)} \\ \varepsilon &= \sum_{i=1}^{n} K^{(i)}\varepsilon^{(i)} \\ \Sigma &= \sum_{i=1}^{n} K^{(i)T}\Sigma^{(i)}K^{(i)T} \end{aligned} \qquad \textbf{A4.36}$$

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 i-th level parameters $\theta^{(i)}$ are:

$$\begin{aligned} \mu^{\theta(i)} &= \Sigma^{\theta(i)}K^{(i)^T}\tilde{\Sigma}^{-1}y \\ \Sigma^{\theta(i)} &= (K^{(i)^T}\tilde{\Sigma}^{-1}K^{(i)} + \Sigma^{(i)^{-1}})^{-1} \\ \tilde{\Sigma} &= \sum_{j\neq i} K^{(j)T}\Sigma^{(j)}K^{(j)T} \end{aligned} \qquad \textbf{A4.37}$$

where $\tilde{\Sigma}$ represents the ReML estimate of error covariance, excluding the component of interest. This component $\Sigma^{(i)} = \Sigma(\lambda)^{(i)}$ is treated as an empirical prior on $\theta^{(i)}$. spm_reml.m uses Eqn. **A4.29** 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 Eqn. **A4.35** 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 vectorized response in terms of Kronecker tensor products:

$$\begin{aligned} y &= vec\{Y\} = \varepsilon \\ \varepsilon &= \sum_{i=1}^{n} I \otimes K^{(i)}\varepsilon^{(i)} \\ \mathrm{cov}\{\varepsilon^{(i)}\} &= I \otimes \Sigma^{(i)} \end{aligned} \qquad \textbf{A4.38}$$

This leads to a computationally efficient scheme employed by spm_reml.m, which uses the compact forms:[3]

$$\begin{aligned} L(\mu)_{\lambda i} &= -\frac{1}{2}tr((I \otimes P_i R)(yy^T - I \otimes \Sigma)(I \otimes R^T)) \\ &= -\frac{r}{2}tr(P_i R(\frac{1}{r}YY^T - \Sigma)R^T) \\ \langle L(\mu)_{\lambda\lambda ij}\rangle &= -\frac{1}{2}tr(I \otimes P_i R\Sigma P_j R\Sigma) \\ &= -\frac{r}{2}tr(P_i R\Sigma P_j R\Sigma) \end{aligned} \qquad \textbf{A4.39}$$

Critically, the update scheme is a function of the sample covariance of the data $\frac{1}{r}YY^T$ and can be regarded as a covariance component estimation scheme. This can be useful in two situations: first, if the augmented form in Eqn. **A4.33** 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 Chapters 29 and 30 and 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:

$$\begin{aligned} \mu^{\theta(i)} &= \Sigma^{(i)}K^{(i)^T}\tilde{\Sigma}^{-1}Y \\ \Sigma^{\theta(i)} &= \Sigma^{(i)} - \Sigma^{(i)}K^{(i)^T}\tilde{\Sigma}^{-1}K^{(i)}\Sigma^{(i)} \\ \tilde{\Sigma} &= \sum_{i=1}^{n} K^{(i)T}\Sigma^{(i)}K^{(i)T} \end{aligned} \qquad \textbf{A4.40}$$

The second situation is where there are a large number of realizations. 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 non-sphericity over voxels in the analysis of fMRI time-series (see Chapter 22 and 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 vectorize the data. However, it is easy to collect the sample covariance over voxels and partition it into non-spherical 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 (Chapter 30 and 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 Eqn. **A4.34**, and use Eqn. **A4.35** to compute the

[3] 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.

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 realizations of the data because the sample covariance, over realizations, 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 hyperparameterization of the covariances, as opposed to precisions.

MODEL SELECTION WITH REML

We conclude with 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), 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 Figure A4.2.

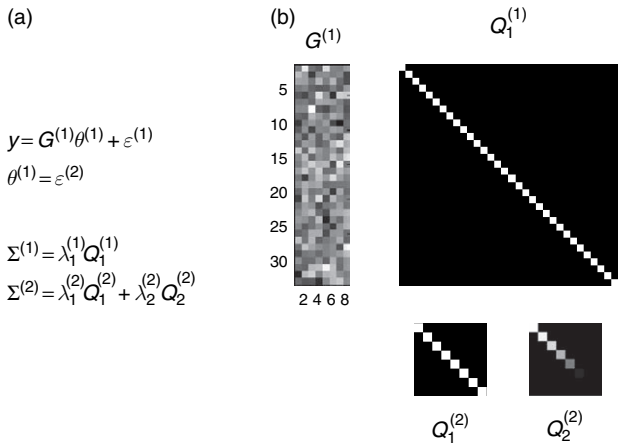

FIGURE A4.2 A hierarchical linear model. (a) The form of the model with two levels. The first level has a single error covariance component, while the second has two. The second level places constraints on the parameters of the first, through the second-level covariance components. Conditional estimation 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 realizations of the response variable y, using hyperparameters of one for all components. The design matrix comprised random Gaussian variables.

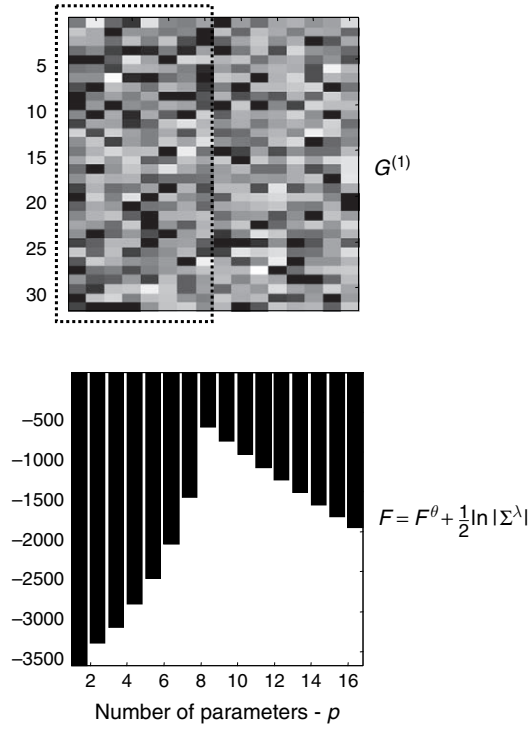

FIGURE A4.3 Model selection in terms of parameters using ReML. The data generated by the eight-parameter model in Figure A4.2 were analysed with ReML using a series of models with an increasing number of parameters. These models were based on the first p columns of the design matrix above. The profile 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.

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 Figure A4.3 and demonstrate that model selection correctly identifies a model with eight parameters. This was the model used to generate the data (Figure A4.2).

The critical issue is whether model selection works when the models differ in their hyperparameterization. To illustrate 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 (Figure A4.4). These components can be regarded as specifying the form of empirical priors over solution space (e.g. spatial constraints in a 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

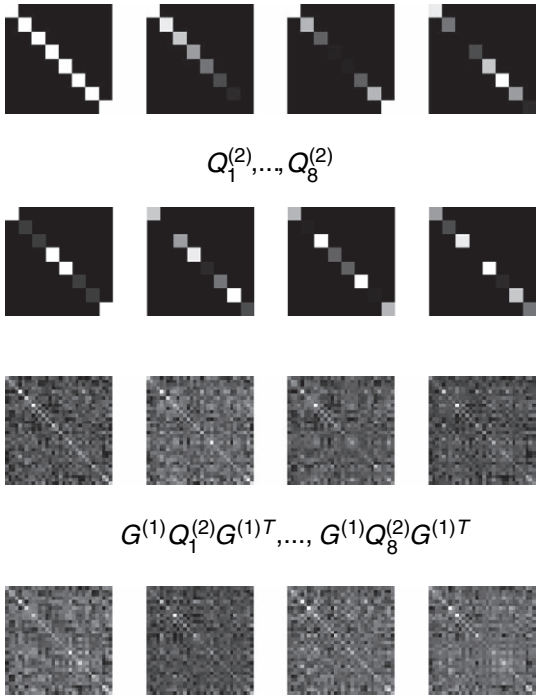

FIGURE A4.4 Covariance components used to analyse the data generated by the model in Figure A4.2. The covariance components are shown at the second level (upper panels) and after projection onto response space (lower panels) with the eight-parameter model. Introducing more covariance components creates a series model with an increasing number of hyperparameters, which we examined using model selection in Figure A4.5. These covariance components were leading diagonal matrices, whose elements comprised a mean-adjusted discrete cosine set.

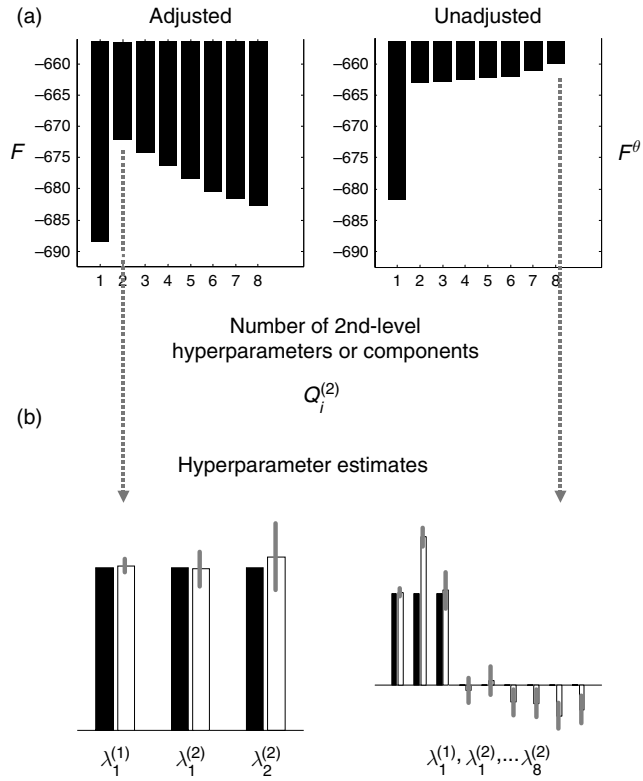

FIGURE A4.5 Model selection in terms of hyperparameters using ReML. (a) The free energy was computed using the data generated by the model in Figure A4.2 and a series of models with an increasing number of hyperparameters. The ensuing free energy profiles (adjusted – left; unadjusted – right) are shown as a function of the number of second-level covariance components used (from Figure A4.4). The adjusted profile clearly identified the correct model with two second-level components. (b) Conditional estimates (white) and true (black) hyperparameter values with 90 per cent confidence intervals for the correct (3-component, left) and redundant (9-component, right) models.

progressively as the number of components and accuracy increased (Figure A4.5).

The lower panel in Figure A4.5 shows the hyperparameter estimates for two models. With the correctly selected model, the true values fall within the 90 per cent confidence interval. However, when the model is overparameterized, 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.

Conclusion

We have seen that restricted maximum likelihood is a special case of expectation maximization and that expectation maximization is a special case of variational Bayes. In fact, nearly every routine used in neuroimaging analysis is a special case of variational Bayes, from ordinary least squares estimation to dynamic causal modelling. We have focused 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.

REFERENCES

Ashburner J, Friston KJ (2005) Unified segmentation. *NeuroImage* **26**: 839–51

Beal MJ, Ghahramani Z (2003) The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. In *Bayesian statistics*, Bernardo JM, Bayarri MJ, Berger JO *et al.* (eds). OUP, Milton Keynes, ch 7

Berger JO (1985) *Statistical decision theory and Bayesian analysis*, 2nd edn. Springer, Berlin

Bishop C (1999) Latent variable models. In *Learning in graphical models*, Jordan M (ed.). MIT Press, London

David O, Harrison L, Friston KJ (2005) Modelling event-related responses in the brain. *NeuroImage* **25**: 756–70

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc Series* 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**: 117–30

Fahrmeir L, Tutz G (1994) *Multivariate statistical modelling based on generalised linear models*. Springer-Verlag Inc., New York, pp 355–56

Feynman RP (1972) *Statistical mechanics*. Benjamin, Reading, MA

Friston KJ (2002) Bayesian estimation of dynamical systems: an application to fMRI. *NeuroImage* **16**: 513–30

Friston K (2005) A theory of cortical responses. *Philos Trans R Soc Lond B Biol Sci* **360**: 815–36

Friston KJ, Penny W, Phillips C *et al.* (2002) Classical and Bayesian inference in neuroimaging: theory. *NeuroImage* **16**: 465–83

Friston KJ, Penny W. (2003) Posterior probability maps and SPMs. *NeuroImage* **19**: 1240–49

Friston KJ, Harrison L, Penny W (2003) Dynamic causal modelling. *NeuroImage* **19**: 1273–302

Friston KJ, Stephan KE, Lund TE *et al.* (2005) Mixed-effects and fMRI studies. *NeuroImage* **24**: 244–52

Friston KJ, Henson R, Phillips C *et al.* (2006) Bayesian estimation of evoked and induced responses. *Hum Brain Mapp* in press

Harrison LM, David O, Friston KJ (2005) Stochastic models of neuronal dynamics. *Philos Trans R Soc Lond B Biol Sci* **360**: 1075–91

Hartley H (1958) Maximum likelihood estimation from incomplete data. *Biometrics* **14**: 174–94

Harville DA (1977) Maximum likelihood approaches to variance component estimation and to related problems. *J Am Stat Assoc* **72**: 320–38

Hinton GE, von Camp D (1993) Keeping neural networks simple by minimising the description length of weights. In *Proc COLT*-93 pp 5–13

Kass RE, Steffey D (1989) Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J Am Stat Assoc* **407**: 717–26

Kiebel SJ, David O, Friston KJ (2006) Dynamic causal modelling of evoked responses in EEG/MEG with lead field parameterization. *NeuroImage* **30**: 1273–84

Leonard T, Hsu JSL (1992) Bayesian inference for a covariance matrix. *Ann Stat* **20**: 1669–96

Mackay DJC (1995) Free energy minimisation algorithm for decoding and cryptoanalysis. *Electron Lett* **31**: 445–47

Mackay DJC, Takeuchi R (1996) Interpolation models with multiple hyperparameters. In *Maximum entropy & Bayesian methods*, Skilling J, Sibisi S (eds). Kluwer, Dordrecht, pp 249–57

Mattout J, Phillips C, Rugg MD *et al.* (2006) MEG source localisation under multiple constraints: an extended Bayesian framework. *NeuroImage* **30**: 753–67

Neal RM, Hinton GE (1998) A view of the EM algorithm that justifies incremental sparse and other variants. In *Learning in graphical models*, Jordan MI (ed.). Kulver Academic Press, Dordrecht

Nelson MC, Illingworth WT (1991) *A practical guide to neural nets*. Addison-Wesley, Reading, MA, pp 165

Penny WD, Stephan KE, Mechelli A *et al.* (2004) Comparing dynamic causal models. *NeuroImage* **22**: 1157–72

Penny WD, Trujillo-Barreto NJ, Friston KJ (2005) Bayesian fMRI time series analysis with spatial priors. *NeuroImage* **24**: 350–62

Phillips C, Rugg M, Friston KJ (2002) Systematic regularisation of linear inverse solutions of the EEG source localisation problem. *NeuroImage* **17**: 287–301

Phillips C, Mattout J, Rugg MD *et al.* (2005) An empirical Bayesian solution to the source reconstruction problem in EEG. *NeuroImage* **24**: 997–1011

Tipping ME (2001) Sparse Bayesian learning and the relevance vector machine. *J Machine Learn Res* **1**: 211–44

Titantah JT, Pierlioni C, Ciuchi S (2001) Free energy of the Fröhlich Polaron in two and three dimensions. *Phys Rev Lett* **87**: 206406

Trujillo-Barreto N, Aubert-Vazquez E, Valdes-Sosa P (2004) Bayesian model averaging. *NeuroImage* **21**: 1300–19

Weissbach F, Pelster A, Hamprecht B (2002) High-order variational perturbation theory for the free energy. *Phys Rev Lett* **66**: 036129