NeuroImage **16,** 465–483 (2002)
doi:10.1006/nimg.2002.1090, available online at http://www.idealibrary.com on IDEAL®

Classical and Bayesian Inference in Neuroimaging: Theory

K. J. Friston, W. Penny, C. Phillips, S. Kiebel, G. Hinton, and J. Ashburner

The Wellcome Department of Imaging Neuroscience, and The Gatsby Computational Neuroscience Unit, University College London, Queen Square, London WC1N 3BG, United Kingdom

Received January 11, 2001

This paper reviews hierarchical observation models, used in functional neuroimaging, in a Bayesian light. It emphasizes the common ground shared by classical and Bayesian methods to show that conventional analyses of neuroimaging data can be usefully extended within an empirical Bayesian framework. In particular we formulate the procedures used in conventional data analysis in terms of hierarchical linear models and establish a connection between classical inference and parametric empirical Bayes (PEB) through covariance component estimation. This estimation is based on an expectation maximization or EM algorithm. The key point is that hierarchical models not only provide for appropriate inference at the highest level but that one can revisit lower levels suitably equipped to make Bayesian inferences. Bayesian inferences eschew many of the difficulties encountered with classical inference and characterize brain responses in a way that is more directly predicated on what one is interested in. The motivation for Bayesian approaches is reviewed and the theoretical background is presented in a way that relates to conventional methods, in particular restricted maximum likelihood (ReML). This paper is a technical and theoretical prelude to subsequent papers that deal with applications of the theory to a range of important issues in neuroimaging. These issues include; (i) Estimating nonsphericity or variance components in fMRI time-series that can arise from serial correlations within subject, or are induced by multisubject (i.e., hierarchical) studies. (ii) Spatiotemporal Bayesian models for imaging data, in which voxels-specific effects are constrained by responses in other voxels. (iii) Bayesian estimation of nonlinear models of hemodynamic responses and (iv) principled ways of mixing structural and functional priors in EEG source reconstruction. Although diverse, all these estimation problems are accommodated by the PEB framework described in this paper.

Key Words: **fMRI; PET; random effects; EM algorithm; ReML; Bayesian inference; hierarchical models.**

1. INTRODUCTION

Since its inception, about ten years ago, statistical parametric mapping (SPM) has proved useful for characterizing neuroimaging data sequences. However, SPM is limited because it is based on classical inference procedures. In this paper we introduce a more general framework, that places SPM in a broader context and points to alternative ways of characterizing and making inferences about regionally specific effects in neuroimaging. In particular we formulate the procedures used in conventional data analysis in terms of hierarchical linear models and establish the connection between classical inference and empirical Bayesian inference through covariance component estimation. This estimation is based on an expectation maximization or EM algorithm.

Statistical parametric mapping entails the use of the general linear model and classical statistics, under parametric assumptions, to create a statistic (usually the *T* statistic) at each voxel. Inferences about regionally specific effects are based on the ensuing image of *T* statistics, the SPM{*T*}. The requisite distributional approximations for the peak height, or spatial extent, of voxel clusters, surviving a specified threshold, are derived using Gaussian random field theory. Random field theory enables the use of classical inference procedures, and the latitude afforded by the general linear model, to give a powerful and flexible approach to continuous spatially extended data. It does so by protecting against family-wise false positives over all the voxels that constitute a search volume; i.e., it provides a way of adjusting the *P* values, in the same way that a Bonferroni correction does for discrete data (Worsley, 1994; Friston *et al.*, 1995).

Despite its success statistical parametric mapping has a number of fundamental limitations. In SPM the *P* value, ascribed to a particular effect, does not reflect the likelihood that the effect is present but simply the probability of getting the observed data in the effect's absence. If sufficiently small, this *P* value can be used to reject the null hypothesis that the effect is negligible. There are several shortcomings of this classical

1053-8119/02 $35.00

approach. Firstly, one can never reject the alternate hypothesis (i.e., say that an activation has not occurred) because the probability that an effect is exactly zero is itself zero. This is problematic, for example, in trying to establish double dissociations or indeed functional segregation; one can never say one area responds to color but not motion and another responds to motion but not color. Second, because the probability of an effect being zero is vanishingly small, given enough scans or subjects one can always demonstrate a significant effect at every voxel. This fallacy of classical inference is becoming relevant practically, with the thousands of scans entering into some fixed-effect analyses of fMRI data. The issue here is that a trivially small activation can be declared significant if there are sufficient degrees of freedom to render the variability of the activation's estimate small enough. A third problem, that is specific to SPM, is the correction or adjustment applied to the *P* values to resolve the multiple comparison problem. This has the somewhat nonsensical effect of changing the inference about one part of the brain in a way that is contingent on whether another part is examined. Put simply, the threshold increases with the search volume, rendering inference very sensitive to what that inference encompasses. Clearly the probability that any voxel has activated does not change with the search volume and yet the classical *P* value does.

All these problems would be eschewed by using the probability that a voxel had activated, or indeed its activation was greater than some threshold. This sort of inference is precluded by classical approaches, which simply give the likelihood of getting the data, given no activation. What one would really like is the probability distribution of the activation given the data. This is the posterior probability used in Bayesian inference. The posterior distribution requires both the likelihood, afforded by assumptions about the distribution of errors, and the prior probability of activation. These priors can enter as known values or can be estimated from the data, provided we have observed multiple instances of the effect we are interested in. The latter is referred to as empirical Bayes. A key point here is that in many situations we do assess repeatedly the same effect over different subjects, or indeed different voxels, and are in a position to adopt an empirical Bayesian approach. This paper describes one such approach.

In contradistinction to other proposals, we are not suggesting a novel way of analyzing neuroimaging data. The use of a Bayesian formalism in special models for fMRI data has been usefully explored elsewhere, e.g., spatiotemporal Markov field models, Descombes *et al.*, 1998; and mixture models, Everitt and Bullmore, 1999. See also the compelling work of Hartvig and Jensen (2000) that combines both these approaches and Højen-Sørensen *et al.* (2000) who focus on temporal aspects with hidden Markov models. Generally these approaches assume that voxels are either active or not and use the data to infer their status. Because of this underlying assumption, there is little connection with conventional models that allow for continuous or graded hemodynamic responses. The aim of this paper is to highlight the fact that the conventional models, we use routinely, conform to hierarchical observation models that can be treated in a Bayesian fashion. The importance of this rests on: (i) the connection between classical and Bayesian inference that ensues and (ii) the potential to apply Bayesian procedures that are overlooked from a classical perspective. For example, random-effect analyses of fMRI data (Holmes and Friston, 1998) adopt two-level hierarchical models. In this context, people generally focus on classical inference at the second level, unaware that the same model can support Bayesian inference at the first. Revisiting the first level, within a Bayesian framework, provides for a much better characterization of single-subject responses, both in terms of the estimated effects and the nature of the inference. This example is developed in Friston *et al.* (2002).

The aim of this paper is to describe hierarchical observation models and establish the relationship between classical maximum likelihood (ML) and empirical Bayes estimators. Parametric empirical Bayes can be formulated classically in terms of covariance component estimation (e.g., within subject vs between subject contributions to error). The covariance component formulation is important because it is ubiquitous in fMRI. Different sources of variability in the data induce nonsphericity that has to be estimated before any inferences about an effect can be made. Important sources of nonsphericity in fMRI include serial or temporal correlations among the errors in single-subject studies, or in multisubject studies, the differences between within and between-subject variability. These issues are used in a companion paper (Friston *et al.*, 2002) to emphasize both the covariance component estimation and Bayesian perspectives, in terms of: (i) The difference between response estimates based on classical maximum likelihood estimators and the conditional means from a Bayesian approach. (ii) The relationship between fixed- and random-effect analyses. (iii) The specificity and sensitivity of Bayesian inferences at the first level and, finally, (iii) the relative importance of the number of scans and subjects for the sensitivity of second-level inferences.

In Friston *et al.* (2002) we use the same theory to elaborate spatiotemporal models for PET. Again this employs two-level models but focuses on Bayesian inference at the first level. It complements the previous fMRI application by looking at spatial correlations in data, using PET data to show how priors can be estimated using observations over voxels at the second level. The examples presented in the companion paper (Friston *et al.*, 2002) illustrate how posterior probability maps (PPMs) can be endowed with greater spatial resolution than the equivalent SPMs and demonstrate their relative immunity from the multiple comparison problem.

1.1 Overview

In this paper we focus on theory and procedures. The key points are reprised in a series of subsequent papers where they are illustrated using real and simulated data. This paper describes how the parameters and hyperparameters of a hierarchical model can be estimated jointly given some data. The distinction between a parameter and a hyperparameter depends on the context established by the estimation or inference in question. Here parameters are quantities that determine the expected response, that is observed. Hyperparameters pertain to the probabilistic behavior of the parameters. Perhaps the simplest example is provided by a single-sample t test. The parameter of interest is the true effect causing the observations to differ from zero. The hyperparameter corresponds to the variance of the observation error (usually denoted by σ^2). Note that one can estimate the parameter, with the sample mean, without knowing the hyperparameter. However, if one wanted to make an inference about that estimate it is necessary to know (or estimate using the residual sum of squares) the hyperparameter. In this paper all the hyperparameters are simply variances of different quantities that cause the measured response (e.g., within-subject variance and between-subject variance). The estimation procedure described below is Bayesian in nature. Because the hyperparameters are estimated from the data it represents an empirical Bayesian approach. However, the aim of this paper is to show the close relationship between Bayesian and maximum likelihood estimation implicit in conventional analyses of imaging data, using the general linear model. Furthermore, we want to place classical and Bayesian inference within the same framework. In this way we show that conventional analyses are special cases of the more general PEB approach.

The first section of this paper introduces hierarchical linear observation models that form the cornerstone of the ensuring estimation procedures. These models are then reviewed from the classical perspective of estimating the model parameters using maximum likelihood and statistical inference using the T statistic. The same model is then considered in a Bayesian light to a make an important point: The estimated error variances, at any level, play the role of priors on the variability of the parameters in the level below. At the highest level, the ML and Bayes estimators are the same, as are their standard error and conditional standard deviation. Both classical and Bayesian approaches rest upon covariance component estimation for which we use an EM algorithm. This is described briefly in the first section and presented in detail in the appendix. The EM algorithm is related to that described in Dempster *et al.* (1981) but extended to cover hierarchical models with any number of levels. The final section addresses Bayesian inference in classical terms of sensitivity and specificity. To do this we "convert" Bayesian inference into a classical one by thresholding the posterior probability to label a region as "activated" or not. This device opens up some interesting questions that are especially relevant to neuroimaging: in classical approaches the same threshold is applied to all voxels in a SPM, to ensure uniform specificity over the brain. Thresholded PPMs, on the other hand, adapt their specificity according to the behavior of local error terms, engendering a uniform confidence in activations of a given size. This complementary aspect of SPMs and PPMs highlights the relative utility of both approaches in making inferences about regional responses.

For an introduction to EM algorithms in generalized linear models, see Fahrmeir and Tutz (1994). This text provides an exposition of EM algorithms and PEB in linear models, usefully relating EM to classical methods (e.g., ReML, p. 225). For an introduction to Bayesian statistics see Lee (1997). This text adopts a more explicit Bayesian perspective and again usually connects empirical Bayes with classical approaches, e.g., the Stein "Shrinkage" estimator and empirical Bayes estimators used below (p. 232). In most standard texts the hierarchical models considered in the next section are referred to as random effects models.

2. THEORY

2.1 Hierarchical Linear Observation Models

In this paper we deal with hierarchical linear observation models of the form

$$\begin{aligned} y &= X^{(1)}\theta^{(1)} + \epsilon^{(1)} \\ \theta^{(1)} &= X^{(2)}\theta^{(2)} + \epsilon^{(2)} \\ &\vdots \\ \theta^{(n-1)} &= X^{(n)}\theta^{(n)} + \epsilon^{(n)} \end{aligned} \tag{1}$$

under Gaussian assumptions about the errors $\epsilon^{(i)} \sim N\{0, C_\epsilon^{(i)}\}$. y is the response variable, usually observed both within units over time and over several units (e.g., subject or voxels). $X^{(i)}$ are specified [design] matrices containing explanatory variables or constraints on the parameters $\theta^{(i-1)}$ of the level below. if the hierarchical model has only one level it reduces to the familiar general linear model employed in conventional data analysis. Two-level models will be familiar to readers who use mixed or random-effect analyses. In this instance the first-level design matrix models the activa-

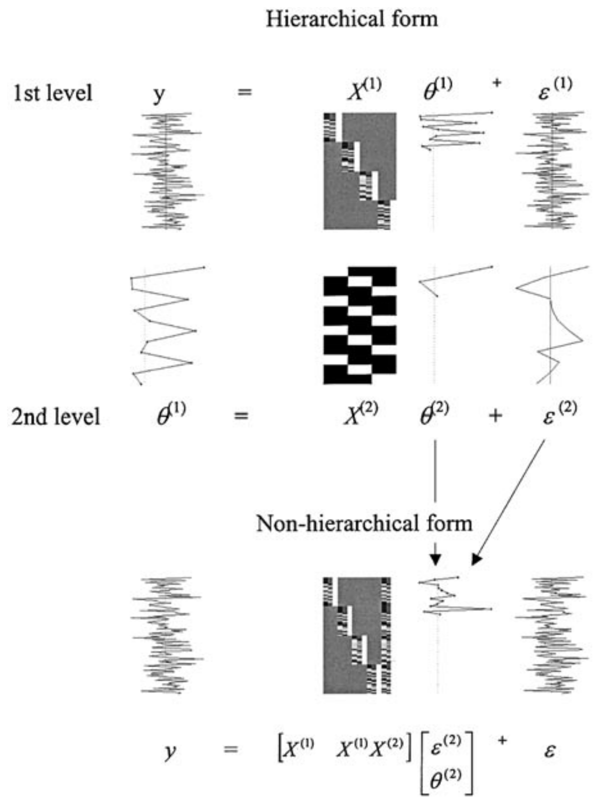

FIG. 1. Schematic showing the form of the design matrices in a two-level model and how the hierarchical form (upper panel) can be reduced to a nonhierarchical form (lower panel). The design matrices are shown in image format with an arbitrary color scale. The response variable, parameters, and error terms are depicted as plots. In this example there are four subjects or units observed at the first level. Each subject's response is modeled with the same three effects, one of these being a constant term. These design matrices are part of those used in Friston *et al.* (2002) to generate simulated fMRI data and are based on the design matrices used in the subsequent empirical event-related fMRI analyses.

tion effects, over scans within subjects, in a subject-separable fashion (i.e., in partitions constituting the blocks of a block diagonal matrix). The second-level design matrix models the subject-specific effects over subjects. Usually, but not necessarily, design matrices at all levels are block diagonal matrices with each partition modelling the observations in each unit at that level (e.g., session, subject, or group).

$$X^{(i)} = \begin{bmatrix} X_1^{(i)} & 0 & \cdots & 0 \\ 0 & X_2^{(i)} & & \\ \vdots & & \ddots & \\ 0 & & & X_j^{(i)} \end{bmatrix} \quad (2)$$

Some examples are shown in Fig. 1 (these examples are used in the empirical analyses of Friston *et al.*, 2002, Section 2). The design matrix at any level has as many rows as the number of columns in the design matrix of the level below. One can envisage three-level models, which embody activation effects in scans modeled for each session, effects expressed in each session modeled for each subject and finally effects over subjects.

The Gaussian or parametric assumptions implicit in these models imply that all the random sources of variability, in the observed response variable, have a Gaussian distribution. This is appropriate for most models in neuroimaging and makes the relationship between classical approaches and Bayesian treatments (that can be generalized to non-Gaussian densities) much more transparent. To ensure valid inference this assumption must not be violated substantially, especially for high-level random effects (e.g., by bimodal distribution of responses over subjects in two-level models).

Technically, models that conform to (1) fall into the class of conditionally independent hierarchical models when the response variables and parameters are independent across units, conditionally on the hyperparameters controlling the error terms (Kass and Steffey, 1989). These models are also called parametric empirical Bayes (PEB) models because the obvious interpretation of the higher-level densities as priors led to the development of PEB methodology (Efron and Morris, 1973). Although the procedures considered in this paper accommodate general models, that are not conditionally independent, we refer to the Bayesian procedures below as PEB because the motivation is identical and most of the examples assume conditional independence. Having posited a model with a hierarchical form, the aim is to estimate its parameters and make some inferences about these estimates using their estimated variability, or more generally their probability distribution. In classical inference one is, usually, only interested in inference about the parameters at the highest level to which the model is specified. In a Bayesian context the highest level is regarded as providing constraints or empirical priors that enable posterior inferences about the parameters in lower levels. Identifying the system of equations in (1) can proceed under two perspectives that are formally identical; a classical statistical perspective and a Bayesian one.

After recursive substitution, to eliminate all but the final level parameters, (1) can be written in an alternative form

$$\begin{aligned} y = \epsilon^{(1)} + X^{(1)}\epsilon^{(2)} + \ldots + X^{(1)} \ldots X^{(n-1)}\epsilon^{(n)} \\ + X^{(1)} \ldots X^{(n)}\theta^{(n)} \end{aligned} \quad (3)$$

In this nonhierarchical form the components of the response variable comprise linearly separable contributions from all levels. Those components that embody error terms are referred to as random effects where the last-level parameters enter as fixed effects. The covariance partitioning implied by (3) is

$$E\{yy^T\} = \underbrace{C_\epsilon^{(1)}}_{error} + \ldots + \underbrace{X^{(1)} \ldots X^{(i-1)} C_\epsilon^{(i)} X^{(i-1)T} \ldots X^{(1)T}}_{ith\text{-}level\ random\ effects}$$
$$+ \ldots + \underbrace{X^{(1)} \ldots X^{(n)} \theta^{(n)} \theta^{(n)T} X^{(n)T} \ldots X^{(1)T}}_{fixed\ effects}, \quad (4)$$

where $C_\epsilon^{(i)} = \text{Cov}\{\epsilon^{(i)}\}$. If only one level is specified the random effects vanish and a fixed-effect analysis ensues. If n is greater than one, the analysis corresponds to a random-effect analysis (or more exactly a mixed-effect analysis that includes random terms). (3) can be interpreted in two ways that form, respectively, the basis for a classical

$$\begin{aligned} y &= \tilde{X}\theta^{(n)} + \tilde{\epsilon} \\ \tilde{X} &= X^{(1)}X^{(2)} \ldots X^{(n)} \\ \tilde{\epsilon} &= \epsilon^{(1)} + X^{(1)}\epsilon^{(2)} + \ldots + X^{(1)}X^{(2)} \ldots X^{(n-1)}\epsilon^{(n)} \end{aligned} \quad (5)$$

and Bayesian estimation

$$\begin{aligned} y &= X\theta + \epsilon^{(1)} \\ X &= [X^{(1)}, \ldots, X^{(1)}X^{(2)} \ldots X^{(n-1)}, X^{(1)}X^{(2)} \ldots X^{(n)}] \\ \theta &= \begin{bmatrix} \epsilon^{(2)} \\ \vdots \\ \epsilon^{(n)} \\ \theta^{(n)} \end{bmatrix}. \end{aligned} \quad (6)$$

In the first, classical formulation (5) the random effects are lumped together and treated as a composite error, rendering the last-level parameters the only ones to appear explicitly. Inferences about nth level parameters are obtained by simply specifying the model to the order required. In contradistinction, the second formulation (6) treats the error terms as parameters, so that θ comprises the errors at all but the first-level and the final-level parameters. Here we have effectively collapsed the hierarchical model into a single level by treating the error terms as parameters (see Fig. 1 for a graphical depiction).

2.2 A Classical Perspective

From a classical perceptive (5) represents an observation model with response variable y, design matrix $\tilde{X}$ and parameters $\theta^{(n)}$. The objective is to estimate these parameters and make some inference about how large they are based upon an estimate of their standard error. Classically, estimation proceeds using the maximum likelihood (ML) estimator of the final-level parameters. Under our model assumptions this is the Gauss–Markov estimator (see Section 2.3).

$$\begin{aligned} \eta_{ML} &= My \\ M &= (\tilde{X}^T C_{\tilde{\epsilon}}^{-1} \tilde{X})^{-1} \tilde{X}^T C_{\tilde{\epsilon}}^{-1}, \end{aligned} \quad (7)$$

where M is an estimator-forming matrix that projects the data onto the estimate. Inferences about this estimate are based upon its covariance, against which any contrast (i.e., linear compound specified by the contrast weight vector c) of the estimates can be compared using the T statistic

$$T = c^T \eta_{ML} / \sqrt{c^T \text{Cov}\{\eta_{ML}\} c}, \quad (8)$$

where, from Eqs. (5) and (7),

$$\begin{aligned} \text{Cov}\{\eta_{ML}\} &= MC_{\tilde{\epsilon}}M^T = (\tilde{X}^T C_{\tilde{\epsilon}}^{-1} \tilde{X})^{-1} \\ C_{\tilde{\epsilon}} &= C_\epsilon^{(1)} + X^{(1)} C_\epsilon^{(2)} X^{(1)T} \ldots \\ &\quad + X^{(1)} \ldots X^{(n-1)} C_\epsilon^{(n)} X^{(n-1)T} \ldots X^{(1)T}. \end{aligned} \quad (9)$$

The covariance of the ML estimator represents a mixture of covariances offered up to the highest level by the error at all previous levels. To implement this classical procedure we need the covariance of the composite errors, from all levels, projected down the hierarchy onto the response variable or observation space $C_{\tilde{\epsilon}} = \text{Cov}\{\tilde{\epsilon}\}$. In other words we need the error covariance components of the model. In fact to proceed, in the general case, one has to turn to the second formulation (6) and some iterative procedure to estimate these covariance components, in our case an EM algorithm. This dependence, on the same procedures used by PEB methods, reflects the underlying equivalence between classical and empirical Bayes methods.

There are special cases where one does not need to resort to iterative covariance component estimation. For example, single-level models. With balanced designs, where $X_1^{(i)} = X_j^{(i)}$ for all i and j, one can replace the response variable with the ML estimates at the penultimate level and proceed as if one had a single-level model. This is the trick harnessed by multistage implementations of random-effect analyses (Holmes and Friston, 1998). Although the ensuing variance estimator is not the same as Eq. (9), its expectation is.

In summary, parameter estimation and inference, in hierarchical models, can proceed given estimates of the appropriate covariance components. The reason for introducing inference based on the ML estimate is to motivate the importance of covariance component estimation. In the next section we take a Bayesian approach to the same issue.

2.3 A Bayesian Perspective

Bayesian inference is based on the conditional probability of the parameters given the data $p(\theta^{(i)}|y)$. Under the assumptions above, this posterior density is Gaussian and the problem reduces the finding its first two moments, the conditional mean $\eta^{(i)}_{\theta|y}$ and conditional covariance $C^{(i)}_{\theta|y}$. These posterior or conditional distributions can be determined for all levels enabling, in contradistinction to classical approaches, inferences at any level using the same hierarchical model. Given the posterior density we can work out the maximum a posteriori (MAP) estimate of the parameters (a point estimator equivalent to $\eta^{(i)}_{\theta|y}$ for the linear systems considered here) or the probability that the parameters exceed some specified value. Consider (1) from a Bayesian point of view. Here level i can be thought of as providing prior constraints on the expectation and covariances of the parameters below

$$\begin{aligned} E\{\theta^{(i-1)}\} &= \eta_\theta^{(i-1)} = X^{(i)}\theta^{(i)} \\ \mathrm{Cov}\{\theta^{(i-1)}\} &= C_\theta^{(i-1)} = C_\epsilon^{(i)}. \end{aligned} \qquad (10)$$

In other words the parameters at level i play the role of supraordinate parameters for level $i - 1$ that control the prior expectation under the constraints specified by $X^{(i)}$. Similarly the prior covariances are simply specified by the error covariances of the level above. For example, given several subjects we can use information about the distribution of activations, over subjects, to inform an estimate pertaining to any single subject. In this case the between-subject variability, from the second level, enters as a prior on the parameters of the first level. The general idea is that in many instances we measure the same effect repeatedly in different contexts. The fact that we have some handle on this effect's inherent variability means that the estimate for a single instance can be constrained by knowledge about others. At the final level we can treat the parameters as; (i) unknown, in which case their priors are flat[1] (c.f. fixed effects) giving an empirical Bayesian approach, or (ii) known. In the latter case the connection with the classical formulation is lost because there is nothing to make an inference about, at the final level.

The objective is to estimate the conditional means and covariances such that the parameters at lower levels can be estimated in a way that harnesses the information available from higher levels. All the information we require is contained in the conditional mean and covariance of θ from (6). From Bayes rule the posterior probability is proportional to the likelihood of obtaining the data, conditional on θ, times the prior probability of θ,

$$p(\theta|y) \propto p(y|\theta)p(\theta), \qquad (11)$$

where the Gaussian priors $p(\theta)$ are specified in terms of their expectation and covariance

$$\eta_\theta = E\{\theta\} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \eta_\theta^{(n)} \end{bmatrix},$$

$$C_\theta = \mathrm{Cov}\{\theta\} = \begin{bmatrix} C_\epsilon^{(2)} & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & C_\epsilon^{(n)} & 0 \\ 0 & \cdots & 0 & C_\theta^{(n)} \end{bmatrix}, \qquad (12)$$

$$\begin{cases} C_\theta^{(n)} = \infty & unknown \\ C_\theta^{(n)} = 0 & known \end{cases}.$$

Under Gaussian assumptions the likelihood and priors are given by

$$\begin{aligned} p(y|\theta) &\propto \exp\left\{-\frac{1}{2}(X\theta - y)^T C_\epsilon^{(1)-1}(X\theta - y)\right\} \\ p(\theta) &\propto \exp\left\{-\frac{1}{2}(\theta - \eta_\theta)^T C_\theta^{-1}(\theta - \eta_\eta)\right\}. \end{aligned} \qquad (13)$$

Substituting Eq. (13) into Eq. (11) gives a posterior density with a Gaussian form

$$p(\theta|y) \propto \exp\left\{-\frac{1}{2}(\theta - \eta_{\theta|y})^T C_{\theta|y}^{-1}(\theta - \eta_{\theta|y})\right\},$$

where

$$\begin{aligned} C_{\theta|y} &= (X^T C_\epsilon^{(1)-1} X + C_\theta^{-1})^{-1} \\ \eta_{\theta|y} &= C_{\theta|y}(X^T C_\epsilon^{(1)-1} y + C_\theta^{-1}\eta_\theta). \end{aligned} \qquad (14)$$

Note that when we adopt an empirical Bayesian scheme $C_\theta^{(n)} = \infty$ and $C_\theta^{-1}\eta_\theta = 0$ (see Eq. (12)). This means we never have to specify the prior expectation at the last level because it never appears explicitly in Eq. (14).

The solution Eq. (14) is ubiquitous in the estimation literature and is presented under various guises in

[1] Flat or uniform priors denote a probability distribution that is the same everywhere, reflecting a lack of any predilection for specific values. In the limit of very high variance a Gaussian distribution becomes flat.

different contexts. If the priors are flat, i.e., $C_\theta^{-1} = 0$, the expression for the conditional mean reduces to the minimum variance linear estimator, referred to as the Gauss–Markov estimator. The Gauss–Markov estimator is identical to the ordinary least square (OLS) estimator that obtains after prewhitening. If the errors are assumed to be independently and identically distributed, i.e., $C_\epsilon^{(1)} = I$, then Eq. (14) reduces to the ordinary least square estimator. With nonflat priors the form of Eq. (14) is identical to that employed by ridge regression and [weighted] minimum norm solutions (e.g., Tikhonov and Arsenin, 1977) commonly found in the inverse problem literature. The Bayesian perspective is useful for minimum norm formulations because it motivates plausible forms for the constraints that can be interpreted in terms of priors.

Equation (14) can be expressed in an exactly equivalent but more compact [Gauss–Markov] form by augmenting the design matrix with an identity matrix and augmenting the data matrix with the prior expectations such that

$$\begin{aligned} C_{\theta|y} &= (\overline{X}^T C_\epsilon^{-1} \overline{X})^{-1} \\ \eta_{\theta|y} &= C_{\theta|y}(\overline{X}^T C_\epsilon^{-1} \bar{y}), \end{aligned} \tag{15}$$

where

$$\bar{y} = \begin{bmatrix} y \\ \eta_\theta \end{bmatrix}$$

$$\overline{X} = \begin{bmatrix} X \\ I \end{bmatrix}$$

$$C_\epsilon = \begin{bmatrix} C_\epsilon^{(1)} & 0 \\ 0 & C_\theta \end{bmatrix}.$$

See Fig. 2 for schematic illustration of the linear model implied by this augmentation. If the priors at the last level are flat, the last-level prior expectation can be set to zero. Note from (12) the remaining prior expectations are zero. This augmented form is computationally more efficient to deal with and simplifies the exposition of the EM algorithm. Furthermore, it highlights the fact that a Bayesian scheme of this sort can be reformulated as the simple weighted least square or ML problem that (15) represents. The problem now reduces to estimating the error covariances C_ϵ that determine the weighting. This is exactly where we ended up in the classical approach, namely reduction to a covariance component estimation problem.

2.4 Covariance Component Estimation

In the previous sections the classical approach was portrayed as using the error covariances to construct an appropriate statistic. The PEB approach was described as using the error covariances as priors to estimate the conditional means and covariances, recall from (10) that $C_\theta^{(i-1)} = C_\epsilon^{(i)}$. Both approaches rest on estimating the covariance components. This estimation depends upon some parameterization of these components; in this paper we use $C_\epsilon^{(i)} = \Sigma\, \lambda_j^{(i)} Q_j^{(i)}$ where $\lambda_j^{(i)}$ are some hyperparameters and $Q_j^{(i)}$ represent some basis set for the covariance matrices. The bases can be construed as constraints on the prior covariance structures in the same way as the design matrices $X^{(i)}$ specify constraints on the prior expectations. $Q_j^{(i)}$ embodies the form of the jth covariance component at the ith level and model different variances for different levels and different forms of correlations within levels. The bases or constraints Q_j are chosen to model the sort of nonsphericity anticipated. For example, they could specify serial correlations within-subject (see Friston *et al.*, 2002, Section 1.1) or correlations among the errors induced hierarchically by repeated measures over subjects (Fig. 3 illustrates both these examples). We will illustrate a number of forms for Q_j in subsequent papers.

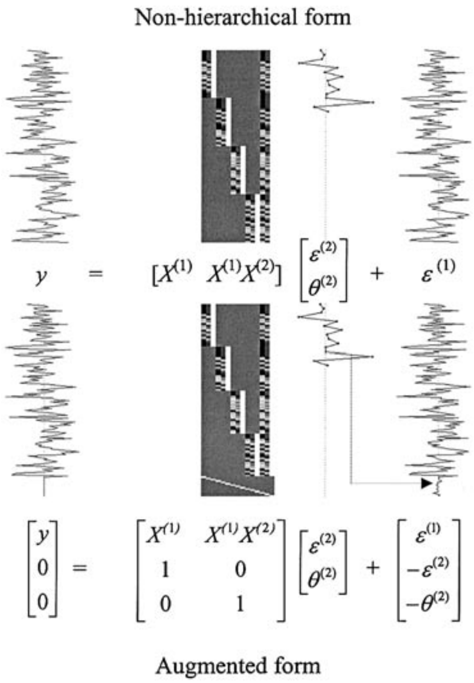
Non-hierarchical form
$y = [X^{(1)}\ X^{(1)}X^{(2)}] \begin{bmatrix} \varepsilon^{(2)} \\ \theta^{(2)} \end{bmatrix} + \varepsilon^{(1)}$
$\begin{bmatrix} y \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} X^{(1)} & X^{(1)}X^{(2)} \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \varepsilon^{(2)} \\ \theta^{(2)} \end{bmatrix} + \begin{bmatrix} \varepsilon^{(1)} \\ -\varepsilon^{(2)} \\ -\theta^{(2)} \end{bmatrix}$
Augmented form

FIG. 2. As for Fig. 1 but here showing how the nonhierarchical form is augmented so that the parameter estimates (that include the error terms from all but the first level and the final level parameters) now appear in the model's residuals. A Gauss–Markov estimator will minimize these residuals in inverse proportion to their prior variance.

One way of thinking about these covariance constraints is in terms of the Taylor expansion of any function of hyperparameters that produced the actual covariance structure

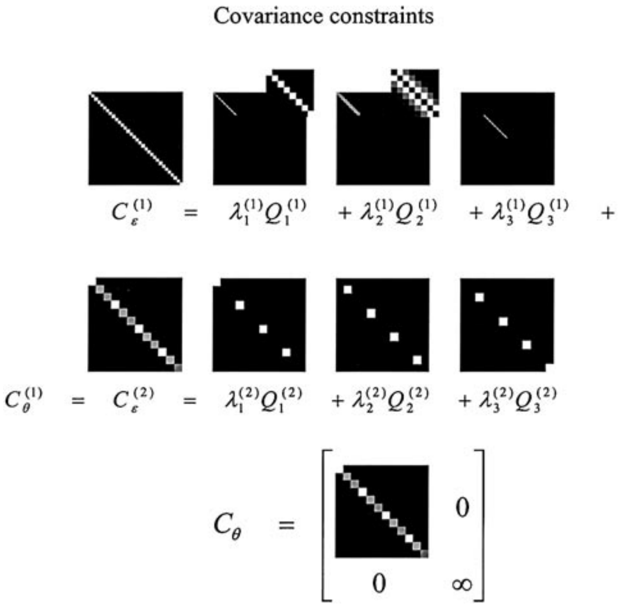
Covariance constraints
$C_\varepsilon^{(1)} = \lambda_1^{(1)}Q_1^{(1)} + \lambda_2^{(1)}Q_2^{(1)} + \lambda_3^{(1)}Q_3^{(1)} +$
$C_\theta^{(1)} = C_\varepsilon^{(2)} = \lambda_1^{(2)}Q_1^{(2)} + \lambda_2^{(2)}Q_2^{(2)} + \lambda_3^{(2)}Q_3^{(2)}$
$C_\theta =$
0
0
∞

FIG. 3. Schematic illustrating the form of the covariance constraints. These can be thought of as "design matrices" for the second-order behavior of the response variable and form a basis set for estimating the error covariance and implicitly the prior covariances. The hyperparameters scale the contribution of each constraint to the error and prior covariances. These covariance constraints correspond to the model described in the legend of Fig. 1. The top row depicts the constraints on the errors. For each subject there are two constraints, one modelling white (i.e., independent) errors and another serial correlation with an AR(1) form. The second level constraints simply reflect the fact that each of the three parameters estimated on the basis of repeated measures at the first level has its own variance. The estimated priors at each level are assembled with the prior for the last level (here a flat prior) to completely specify the models priors (lower panel). Constraints of this form are used in Friston *et al.* (2002) during the simulation of serially correlated fMRI data-sequences and covariance component estimation using real data.

$$C(\lambda)_\epsilon^{(i)} = \sum \lambda_j^{(i)} \frac{\partial C(0)_\epsilon^{(i)}}{\partial \lambda_j^{(i)}} + \ldots, \qquad (16)$$

where the basis set corresponds to the partial derivatives of the covariances with respect to the hyperparameters. In variance component estimation the high-order terms in Eq. (16) are generally zero. In this context a linear decomposition of $C_\epsilon^{(i)}$ is a natural parameterization because the different sources of conditionally independent variance add linearly and the constraints can be specified directly in terms of these components. There are other situations where a different parameterization may be employed. For example, if the constraints were implementing several independent priors in a nonhierarchical model a more natural expansion might be in terms of the precision $C_\theta^{-1} = \Sigma\, \lambda_j Q_j$. The precision is simply the inverse of the covariance matrix. Here Q_j correspond to precisions specifying the form of independent prior densities (see Appendix A.3). However, in this paper, we deal only with priors that are engendered by the observation model that induces hierarchically organized, linearly mixed, variance components. See Harville (1977, p. 322) for comments on the usefulness of making the covariances linear in the hyperparameters.

The augmented form of the covariance constraints obtains by placing them in the appropriate partition in relation to the augmented error covariance matrix

$$C_\epsilon = C_\theta + \sum \lambda_k Q_k$$

$$Q_k = \frac{\partial C_\epsilon}{\partial \lambda_k}$$

$$C_\theta = \begin{bmatrix} 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \\ 0 & \cdots & 0 & C_\theta^{(n)} \end{bmatrix}, \qquad (17)$$

$$Q_k = \begin{bmatrix} 0 & & \cdots & & 0 & 0 \\ & \ddots & & & & \\ \vdots & & Q_j^{(i)} & & \vdots & \vdots \\ & & & \ddots & & \\ 0 & & \cdots & & 0 & 0 \\ 0 & & \cdots & & 0 & 0 \end{bmatrix},$$

where the subscript k runs over both levels and the constraints within each level. Having framed the covariance estimation in terms of estimating hyperparameters, we can now use an EM algorithm to estimate them.

2.5 *The EM Algorithm*

EM or expectation-maximization is a generic, iterative parameter reestimation procedure that encompasses many iterative schemes devised to estimate jointly the parameters and hyperparameters of a model (Dempster *et al.*, 1977, 1981). It was original introduced as an iterative method to obtain maximum likelihood estimators in incomplete data situations (Hartley, 1958) and was generalized by Dempster *et al.* (1977). More recently, it has been formulated (e.g., Neal and Hinton, 1998) in a way that highlights its elegant nature using a statistical mechanical interpretation. This formulation considers the EM algorithm as a coordinate descent on the free energy of a system. The descent comprises an **E**-step, that finds the conditional expectation of the parameters, holding the hyperparameters fixed and a **M**-step, which updates the maximum likelihood estimate of the hyperparameters, keeping the parameters fixed.

In brief, EM algorithms provide a way to estimate both the parameters and hyperparameters from the data. In other words, it estimates the model parameters when the exact densities of the observation error and priors are unknown. For linear models under Gaussian assumptions the EM algorithm returns: (i) the posterior density of the parameters, in terms of their expectation and covariance and (ii) the ML estimates of the hyperparameters. The EM algorithm described in the appendix (A.1) is depicted schematically in Fig. 4. In the context of the linear observation models discussed in this paper, the EM scheme is the same as using restricted maximum likelihood (ReML) estimates of the hyperparameters, that properly account for the loss of degrees of freedom, incurred by parameter estimation. The operational equivalence between ReML and EM has been established for many years (see Fahrmeir and Tutz, 1994, p. 226). However, it is useful to understand their equivalence because EM algorithms are usually employed to estimate the conditional densities of model parameters when the hyperparameters of the likelihood and prior densities are not known. In contradistinction, ReML is generally used to estimate unknown variance components without explicit reference to the parameters. In the hierarchical linear observation model considered here, the unknown hyperparameters become variance components which means they can be estimated using ReML. It should be noted that EM algorithms are not restricted to linear observation models or Gaussian priors, and have found diverse applications in the machine learning community. On the other hand ReML was developed explicitly for linear observation models under Gaussian assumptions.

In the appendix we have made an effort to reconcile the free energy formulation based on statistical mechanics (Neal and Hinton, 1998) with classical ReML (Harville, 1977). This might be relevant for understanding ReML in the context of extensions to the free energy formulation, afforded by the use of hyperpriors (priors on the hyperparameters). One key insight into the EM approach is that the **M**-step returns, not simply the ML estimate of the hyperparameters, but the ReML that is properly restricted from a classical perspective.

Having computed the conditional mean and covariances of the parameters we are now in a position to make inferences about the effects at any level using their posterior density.

2.6 Conditional and Classical Estimators

Given an estimate of the error covariance of the augmented form C_ϵ and implicitly the priors that are embedded in it, one can compute the conditional mean and covariance at each level, where

$$\eta_{\theta|y} = E\{\theta|y\} = \begin{bmatrix} \eta^{(2)}_{\epsilon|y} \\ \vdots \\ \eta^{(n)}_{\epsilon|y} \\ \eta^{(n)}_{\theta|y} \end{bmatrix},$$

$$C_{\theta|y} = \mathrm{Cov}\{\theta|y\} = \begin{bmatrix} C^{(2)}_{\epsilon|y} & \cdots & & \\ \vdots & \ddots & & \\ & & C^{(n)}_{\epsilon|y} & \\ & & & C^{(n)}_{\theta|y} \end{bmatrix}. \tag{18}$$

The conditional means for each level obtain recursively with $\eta^{(i-1)}_{\theta|y} = X^{(i)}\eta^{(i)}_{\theta|y} + \eta^{(i)}_{\epsilon|y}$. The conditional covariances are simply $C^{(i-1)}_{\theta|y} = C^{(i)}_{\epsilon|y}$ up to the penultimate level and $C^{(n)}_{\theta|y}$ at the final level. The conditional means represent a better "collective" characterization of the model parameters than the equivalent ML estimates because they are constrained by prior information from higher levels (see discussion). At the last level the conditional mean and ML estimators are the same. In PEB, inferences about the parameters at subordinate levels are enabled through having an estimate of their posterior density. At the last level the posterior density reduces to the likelihood distribution and inference reverts to a classical one based on the standardized conditional mean.

The standardized conditional mean, or a contrast of means, is normalized by its conditional error. This conditional error is larger than the standard error of the conditional mean with equivalence when the priors are flat (i.e., the conditional variability of a parameter is greater than the estimate of its mean, except at the last level where they are the same).

$$T^{(i)} = c^T\eta^{(i)}_{\theta|y}/\sqrt{c^T C^{(i)}_{\theta|y} c} \tag{19}$$

This statistic indicates the number of standard deviations by which the mean of the conditional distribution of the contrast deviates from zero. The critical thing, we want to emphasize here, is that this statistic is identical to the classical T statistic at the last level. This means that the ML estimate and the conditional mean are the same and the conditional covariance is exactly the same as the covariance of the ML estimate. The convergence of classical and Bayesian inference at the last level rests on this identity and depends on adopting an empirical Bayesian approach. This establishes a close connection between classical random effect analyses and hierarchical Bayesian models. However, the two approaches diverge if we consider that

Augment to embody priors in error covariance

$$\overline{X}=\begin{bmatrix}\prod_{i=1}^{1}X^{(i)} & \cdots & \prod_{i=1}^{n}X^{(i)}\\ I & & 0\\ \vdots & \ddots & \\ 0 & \cdots & I\end{bmatrix},\quad \overline{y}=\begin{bmatrix}y\\ 0\\ \vdots\\ \eta_\theta^{(n)}\end{bmatrix},\quad \mathrm{C}_\theta=\begin{bmatrix}0 & & 0 & 0\\ & \ddots & & \vdots\\ 0 & & 0 & 0\\ 0 & \cdots & 0 & \mathrm{C}_\theta^{(\mathrm{n})}\end{bmatrix},\quad Q_1=\begin{bmatrix}\mathrm{Q}_1^{(1)} & & 0 & 0\\ & \ddots & & \vdots\\ 0 & & 0 & 0\\ 0 & \cdots & 0 & 0\end{bmatrix},\quad Q_2=\ldots$$

```
Until convergence {     E-Step
                            $C_\varepsilon = C_\theta + \sum \lambda_k Q_k$
                            $C_{\theta|y} = (\overline{X}^T C_\varepsilon^{-1} \overline{X})^{-1}$
                            $\eta_{\theta|y} = C_{\theta|y} \overline{X}^T C_\varepsilon^{-1} \overline{y}$

                        M-Step
                            $P = C_\varepsilon^{-1} - C_\varepsilon^{-1} \overline{X} C_{\theta|y} \overline{X}^T C_\varepsilon^{-1}$
                            $g_i = -\frac{1}{2} tr\{PQ_i\} + \frac{1}{2} \overline{y}^T P^T Q_i P \overline{y}$
                            $H_{ij} = \frac{1}{2} tr\{PQ_i P Q_j\}$
                            $\lambda = \lambda + H^{-1} g$
                  }
```

assemble estimates of error covariance, priors, conditional covariances and means

$$C_\varepsilon=\begin{bmatrix}C_\varepsilon^{(1)} & & 0 & 0\\ & \ddots & & \vdots\\ 0 & & C_\varepsilon^{(n)} & 0\\ 0 & \cdots & 0 & \mathrm{C}_\theta^{(\mathrm{n})}\end{bmatrix},\qquad \mathrm{C}_\theta^{(\mathrm{i})}=C_\varepsilon^{(i-1)}$$

$$C_{\theta|y}=\begin{bmatrix}C_{\varepsilon|y}^{(2)} & \cdots & & \\ \vdots & \ddots & & \\ & & C_{\varepsilon|y}^{(n)} & \\ & & & C_{\theta|y}^{(n)}\end{bmatrix},\qquad C_{\theta|y}^{(i)}=C_{\varepsilon|y}^{(i-1)}$$

$$\eta_{\theta|y}=\begin{bmatrix}\eta_{\varepsilon|y}^{(2)}\\ \vdots\\ \eta_{\varepsilon|y}^{(n)}\\ \eta_{\theta|y}^{(n)}\end{bmatrix},\qquad \eta_{\theta|y}^{(i-1)}=X^{(i)}\eta_{\theta|y}^{(i)}+\eta_{\varepsilon|y}^{(i)}$$

FIG. 4. Pseudo-code schematic showing the recursive structure of the EM algorithm (described in the appendix) as applied in the context of conditionally independent hierarchical models. See main text for a full explanation. This formulation follows Harville (1977).

the real power of Bayesian inference lies in (i) coping with incomplete data or unbalanced designs and (ii) looking at the conditional or posterior distributions at lower levels. The relationship between classical and empirical Bayesian inference is developed in the next section.

3. CLASSICAL AND BAYESIAN INFERENCE COMPARED

In this section we establish a relationship between classical and Bayesian inference by applying Bayes in a classical fashion. As noted above, at the last level, PEB inference based on the standardized conditional mean is identical to classical inference based on the T statistic. In this context the ML estimators and the conditional means are the same, as are the conditional covariance and the covariance of the ML estimator. What about inference at intermediate levels? Bayesian inference is based on the conditional or posterior densities (means and covariances) to give the posterior probability that a compound of parameters (i.e., contrast) is greater than some value say γ. How does this relate to the equivalent classical inference? Clearly the essence of both inferences are quite distinct. The P value in classical inference pertains to the probability of getting the data under the null hypothesis, whereas in Bayesian inference it is the probability that, given the data, the contrast exceeds γ. However, we can demonstrate the connection between Bayesian and classical inference by taking a classical approach to the former:

Consider the following heuristic argument. Take an observation model with a single parameter and assume that the error and prior covariance of the parameter are known. Classical inference is characterized in terms of specificity and sensitivity given the null $\theta = 0$ and alternate $\theta = A$ hypotheses. Specificity is the probability of correctly accepting the null hypothesis and is $1 - \alpha$, where α is a small false positive rate. The sensitivity β or power is the probability of correctly rejecting the null hypothesis. Classically, one rejects the null hypothesis whenever the standardized ML estimator exceeds some specified statistical threshold ν. The probability of this happening is based on its distribution whose standard deviation is given by Eq. (9).

$$\begin{aligned}\alpha &= 1 - \Phi(\nu)\\ \beta &= 1 - \Phi\left(\nu - \frac{A}{\sqrt{(X^TC_\epsilon^{-1}X)^{-1}}}\right),\end{aligned} \tag{20}$$

where $\Phi(\cdot)$ is the cumulative density function of the unit normal distribution. Note that one would use the Student's T distribution if the error covariance had to be estimated but here we are treating the error variance as known. α and β are the probabilities that the ML estimator divided by its standard deviation would exceed ν, under the null and alternative hypotheses, respectively. Note that this classical inference disregards any priors on the parameter's variance, assuming them to be infinite. We can now pursue an identical analysis for Bayesian inference. By thresholding the posterior probability (or PPM) at a specified confidence (say 95%) one could declare the surviving voxels as showing a significant effect. This corresponds to thresholding the conditional mean at $\gamma + u\sqrt{C_{\theta|y}}$, where u is a standard Gaussian deviate specifying the level of confidence required. For example $u = 1.64$ for 95% confidence. One can regard u as a Bayesian threshold. Although thresholding the posterior probability to declare a voxel "activated" is, of course, unnecessary (see discussion), it is used here as a device to connect Bayesian and classical inference.

Under the null and alternate hypotheses the expectation and variance of the conditional mean are

$$\langle \eta_{\theta|y} \rangle = \begin{cases} 0 & null \\ C_{\theta|y}X^TC_\epsilon^{-1}XA & alternate \end{cases}$$

$$\mathrm{Cov}\{\eta_{\theta|y}\} = C_\eta = C_{\theta|y}X^TC_\epsilon^{-1}XC_{\theta|y},$$

from which it follows

$$\begin{aligned}\alpha &= 1 - \Phi(w)\\ \beta &= 1 - \Phi\left(w - \frac{C_{\theta|y}X^TC_\epsilon^{-1}XA}{\sqrt{C_\eta}}\right)\\ &= 1 - \Phi\left(w - \frac{A}{\sqrt{(X^TC_\epsilon^{-1}X)^{-1}}}\right)\\ w &= \frac{\gamma}{\sqrt{C_\eta}} + \frac{u\sqrt{C_{\theta|y}}}{\sqrt{C_\eta}},\end{aligned} \tag{21}$$

where $C_{\theta|y} \geq C_\eta$, with equality when the priors are flat. Comparing (20) and (21) reveals a fundamental difference and equivalence between classical and Bayesian inference. The first thing to note is that the expressions for power and sensitivity have exactly the same form, such that if we chose a threshold u that gave the same specificity as a classical test, then the same sensitivity would ensue. In other words there is no magical increase in power afforded by a Bayesian approach. The classical approach is equally as sensitive given the same specificity.

The essential difference emerges when we consider that the relationship between the posterior probability threshold u and the implied classical threshold w depends on quantities (i.e., error and prior variance) that vary over voxels. In a classical approach we would choose some fixed threshold ν, say for all voxels in a classical SPM. This ensures that the resulting inference has the same specificity everywhere because specificity depends on, and only on, ν. To emulate this uniform specificity, when thresholding a PPM, we would have to keep w constant. The critical thing here is that if the prior covariance or observation error changes from voxel to voxel then either γ or u must change to maintain the same specificity. This means that the nature of the inference changes fundamentally, either in terms of the size of the inferred activation γ or the confidence about that effect u. In short, one can either have a test with uniform specificity (the classical approach) or one can infer an effect of uniform size with uniform confidence (the Bayesian approach) but not both at the same time. For example, given a confidence level determined by u, as the prior variance gets smaller γ must also decrease to maintain the same specificity. Consequently, in some regions a classical inference corresponds to a Bayesian inference about a big effect and in other regions, where the estimate is intrinsically less variable, the inference is about a small effect. In the limit of estimates that are very reliable the classical inference pertains to trivially small effects. This is one of the fallacies of classical inference alluded to in the introduction. There is nothing statistically invalid about this: One might argue that a very reliable activation that is exceedingly small is interesting. However, in many contexts, including neuroimaging, we are generally interested in activations of a nontrivial magnitude and this speaks to the usefulness of Bayesian inference.

In summary, classical inference uses a criterion that renders the specificity fixed. However, this is at the price that the size of the effect, subtending the inferred activation, will change from voxel to voxel or brain region to brain region. By explicitly framing the inference in terms of the posterior probability, Bayesian inference sacrifices a constant specificity to ensure the inference is about the same thing at every voxel. Intuitively one can regard Bayesian inference as adjusting the classical threshold according to the inherent variability of the effect one is interested in. In regions with high prior variability the classical threshold is relaxed to ensure type II errors are avoided. In this context the classical specificity represents the lower bound for Bayesian inference. In other words Bayesian inference is generally much more specific than classical inference (by several orders of magnitude in the empirical examples presented later) with equivalence when the prior variance becomes very large.

In concluding it should be noted one does not usually consider issues like specificity from a Bayesian point of view (the null hypothesis plays no role because the real world behavior is already specified by the priors). From a purely Bayesian perspective the specificity and sensitivity of an inference are meaningless because at no point is an activation declared significant (correctly or falsely). It is only when we impose a categorical classification (activated vs not activated) by thresholding on the posterior probability that specificity and sensitivity become an issue. Ideally, one would report ones inferences in terms of the conditional density of the activation at every voxel. This is generally impractical in neuroimaging and the posterior probability (that is a function of the conditional density and γ) becomes a useful characterization. This characterization is, and should be, the same irrespective of whether we have analysed just one voxel or the entire brain. To threshold the posterior probabilities is certainly tenable for summary or display purposes, but to declare the surviving voxels as "activated" represents a category error. This is because the inherent nature of the inference already specifies that the voxel is probably active with a nontrivial probability of not being activated. However, it is comforting to note that, by enforcing a classical take on Bayesian inference, we do not have to worry too much about the multiple comparison problem because the ensuing inference has an intrinsically high specificity.

4. SUMMARY

This paper has introduced three key components that play a role in the estimation of the linear models considered, Bayesian estimation, hierarchical models and the EM algorithm. The summary points below attempt to clarify the relationships among these components. It is worth while keeping in mind there are essentially three sorts of estimation. (i) Fully Bayesian, when the priors are known. (ii) Empirical Bayesian, when the priors are unknown but they can be parameterized in terms of some hyperparameters that are estimated from the data and (iii) maximum likelihood estimation, when the priors are assumed to be flat. In the final instance the ML estimators correspond to weighted least square or minimum norm solutions. All these procedures can be implemented with an EM algorithm (see Fig. 5).

- Model estimation and inference are greatly enhanced by being able to make probabilistic statements about the model parameters given the data, as opposed to probabilistic statements about the data, under some arbitrary assumptions about the parameters (e.g., the null hypothesis), as afforded by classical statistics. The former is predicated on the posterior or conditional distribution of the parameters that is derived using Bayes rule.
- Bayesian estimation and inference require priors. If the priors are known then a fully Bayesian estimation can proceed. In the absence of known priors there may be constraints on the form of the priors that can be harnessed using empirical Bayes estimates of the associated hyperparameters.
- A model with a hierarchical form embodies implicit constraints on the form of the prior distributions. Hyperparameters that, in conjunction with these constraints, specify the priors can then be estimated with

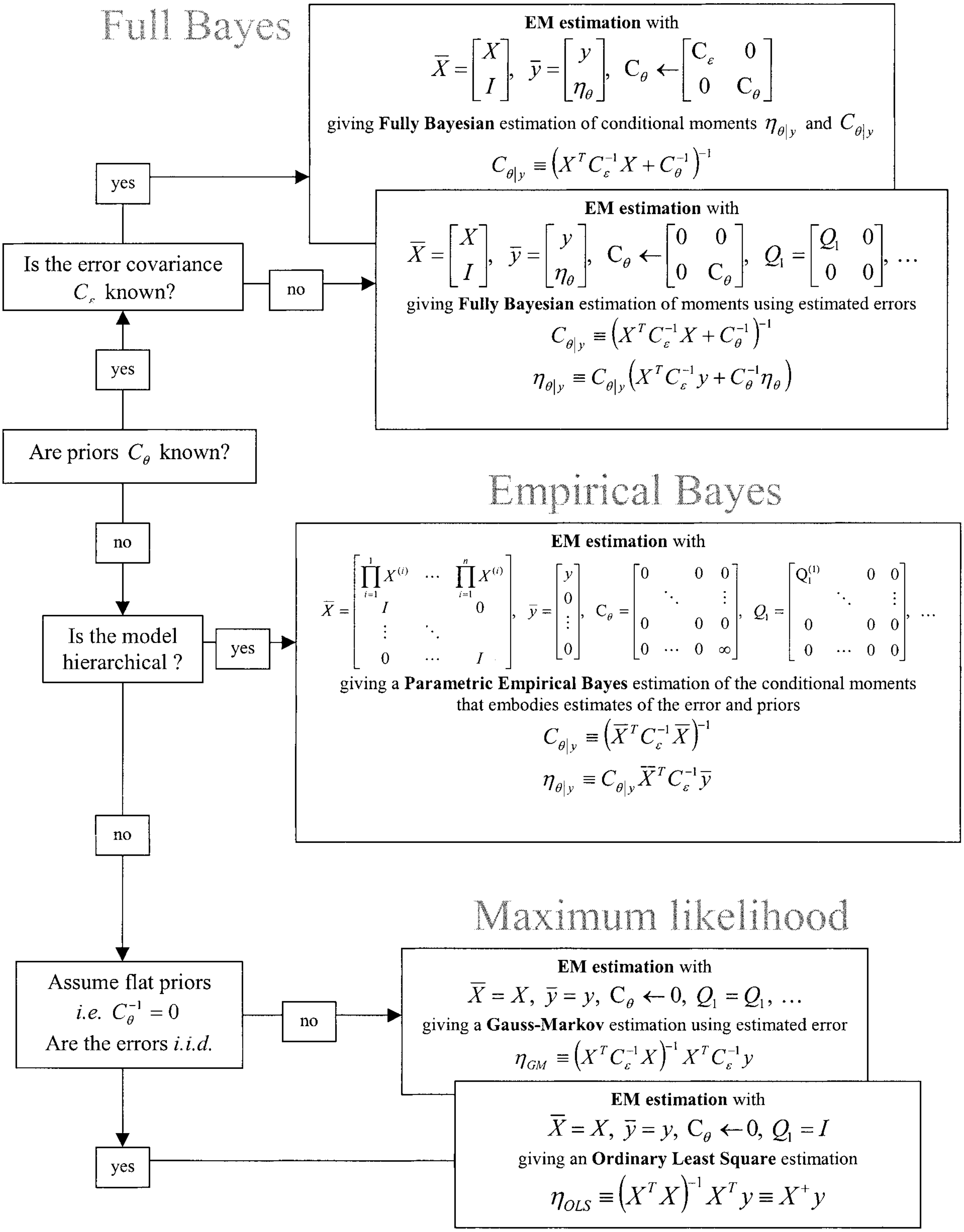
Full Bayes
EM estimation with
$\bar{X}=\begin{bmatrix}X\\I\end{bmatrix},\ \bar{y}=\begin{bmatrix}y\\\eta_\theta\end{bmatrix},\ C_\theta\leftarrow\begin{bmatrix}C_\varepsilon&0\\0&C_\theta\end{bmatrix}$
giving Fully Bayesian estimation of conditional moments $\eta_{\theta|y}$ and $C_{\theta|y}$
$C_{\theta|y}\equiv\left(X^TC_\varepsilon^{-1}X+C_\theta^{-1}\right)^{-1}$
yes
Is the error covariance C_ε known?
no
EM estimation with
$\bar{X}=\begin{bmatrix}X\\I\end{bmatrix},\ \bar{y}=\begin{bmatrix}y\\\eta_\theta\end{bmatrix},\ C_\theta\leftarrow\begin{bmatrix}0&0\\0&C_\theta\end{bmatrix},\ Q_1=\begin{bmatrix}Q_1&0\\0&0\end{bmatrix},\ \dots$
giving Fully Bayesian estimation of moments using estimated errors
$C_{\theta|y}\equiv\left(X^TC_\varepsilon^{-1}X+C_\theta^{-1}\right)^{-1}$
$\eta_{\theta|y}\equiv C_{\theta|y}\left(X^TC_\varepsilon^{-1}y+C_\theta^{-1}\eta_\theta\right)$
yes
Are priors C_θ known?
Empirical Bayes
no
EM estimation with
$\bar{X}=\begin{bmatrix}\prod_{i=1}^{1}X^{(i)}&\cdots&\prod_{i=1}^{n}X^{(i)}\\I&&0\\\vdots&\ddots&\\0&\cdots&I\end{bmatrix},\ \bar{y}=\begin{bmatrix}y\\0\\\vdots\\0\end{bmatrix},\ C_\theta=\begin{bmatrix}0&&0&0\\&\ddots&&\vdots\\0&&0&0\\0&\cdots&0&\infty\end{bmatrix},\ Q_1=\begin{bmatrix}Q_1^{(1)}&&0&0\\&\ddots&&\vdots\\0&&0&0\\0&\cdots&0&0\end{bmatrix},\ \dots$
Is the model hierarchical ?
yes
giving a Parametric Empirical Bayes estimation of the conditional moments that embodies estimates of the error and priors
$C_{\theta|y}\equiv\left(\bar{X}^TC_\varepsilon^{-1}\bar{X}\right)^{-1}$
$\eta_{\theta|y}\equiv C_{\theta|y}\bar{X}^TC_\varepsilon^{-1}\bar{y}$
no
Maximum likelihood
EM estimation with
$\bar{X}=X,\ \bar{y}=y,\ C_\theta\leftarrow0,\ Q_1=Q_1,\ \dots$
Assume flat priors
i.e. $C_\theta^{-1}=0$
Are the errors i.i.d.
no
giving a Gauss-Markov estimation using estimated error
$\eta_{GM}\equiv\left(X^TC_\varepsilon^{-1}X\right)^{-1}X^TC_\varepsilon^{-1}y$
EM estimation with
$\bar{X}=X,\ \bar{y}=y,\ C_\theta\leftarrow0,\ Q_1=I$
yes
giving an Ordinary Least Square estimation
$\eta_{OLS}\equiv\left(X^TX\right)^{-1}X^Ty\equiv X^+y$

FIG. 5. Schematic showing the relationship among estimation schemes for linear observation models under parametric assumptions. This figure highlights the universal role of the EM algorithm, showing that all conventional estimators can be cast in terms of, or implemented with, the EM algorithm described in the legend of Fig. 4.

PEB. In short, a hierarchical form for the observation model enables an empirical Bayesian approach.

- If the observation model does not have a hierarchical structure then one knows nothing about the form of the priors, and they are assumed to be flat. Bayesian estimation with flat priors reduces to maximum likelihood estimation.
- In the context of an empirical Bayesian approach the priors at the last level are generally unknown and enter as flat priors. This is equivalent to treating the parameters at the last level as fixed effects (i.e., effects with no intrinsic or random variability). One consequence of this is that the conditional mean and the ML estimate, at the last level, are identical.
- In terms of inference, at the last level, PEB and classical approaches are formally identical. At subordinate levels PEB can use the posterior densities to provide for Bayesian inference about the effects of interest. This is precluded from a classical perspective because there are no priors.
- EM provides a generic framework in which fully Bayes, PEB or ML estimation can proceed. Its critical utility is the estimation of covariance components, given some data, through the ReML estimation of hyperparameters mixing these covariance components. An EM algorithm can be used to estimate the error covariance in the context of known priors or to estimate both the error and priors by embedding the latter in the former. This embedding is achieved by augmenting the design matrix and data (see Figs. 2 and 4).
- In the absence of priors, or hierarchical constraints on their form, EM can be used in a ML setting to estimate the error covariance to enable Gauss-Markov estimates (see Fig. 5). These estimators are the optimum weighted least square estimates in the sense they have the minimum variance of all unbiased linear estimators. In the limiting case that the covariance constraints reduce to a single basis (synonymous with known correlations or a single hyperparameter) the EM algorithm converges in a single iteration and emulates a classical sum of square estimation of error variance. When this single basis is the identity matrix (i.e., i.i.d. errors), an EM algorithm simply implements an ordinary least square estimation.

In this paper we have reviewed hierarchical observation models of the sort commonly encountered in neuroimaging. Their hierarchical nature induces different sources of variability in the observations at different levels (i.e., variance components) that can be estimated using an EM algorithm. The use of an EM algorithm, for variance component estimation, is not limited to hierarchical models but finds a useful application whenever nonsphericity of the errors is specified with more than one hyperparameter (e.g., serial correlations in fMRI). This application will be illustrated in Friston *et al.* (2002). The critical thing, about hierarchical models, is that they conform to a Bayesian scheme where variance estimates at higher levels can be used as constraints on the estimation of effects at lower levels. This perspective rests upon exactly the same mathematics that pertains to variance component estimation in nonhierarchical models but allows one to frame the estimators in conditional or Bayesian terms. An intuitive understanding of the conditional estimators, at a given level, is that they "shrink" towards their average, in proportion to the error variance at that level, relative to their intrinsic variability (error variance at the supraordinate level). See Lee (1997, p. 232) for a discussion of PEB and Stein "Shrinkage" estimators. In what sense are these Bayes predictors a better characterization of the model parameters than the equivalent ML estimates? In other words, what are the gains in using a shrinkage estimator? The following, prepared by Keith Worsley (personal communication), addresses this question.

This is a topic that has been debated at great length in the statistics literature and even in the popular press. See the Scientific American article "Stein's paradox in statistics" (Efron and Morris, 1977). The answer depends on ones definition of "better," or in technical terms, the loss function. If the aim is to find the best predictor for a specific subject, then one can do no better than the ML estimator for that subject. Here the loss function is simply the squared difference between the estimated and real effects for the subject in question. Conversely, if the loss function is averaged over subjects then the shrinkage estimator is best. This has been neatly summarized in a discussion paper read before the Royal Statistical Society entitled "Regression, prediction, and shrinkage" by Copas (1983). The vote of thanks was given by Dunsmore, who said:

> "Suppose I go to the doctor with some complaint and ask him to predict the time *y* to remission. He will take some explanatory measurements **x** and provide some prediction for *y*. What I am interested in is a prediction for my **x**, not for any other **x** that I might have had—but did not. Nor am I really interested in his necessarily using a predictor which is "best" over all possible **x**'s. Perhaps rather selfishly, but I believe justifiably, I want the best predictor for my **x**. Does is necessarily follow that the best predictor for my **x** should take the same form as for some other **x**? Of course this can cause problems for the esteem of the doctor or his friendly statistician. Because we are concerned with actual observations the goodness or otherwise of the prediction will eventually become apparent. In this case the statistician will not be able to hide behind the screen provided by averaging over all possible futures **x**'s."

Copas then replied:

> "Dr. Dunsmore raises two general points that repay careful thought. Firstly, he questions the assumption made at the very start of the paper that predictions are to be judged in the context of a population of future **x**'s and not just at some specific **x**. To pursue the analogy of the doctor and the patient, all I can say is that the paper is written from the doctor's point of view and not from the patients! No doubt the doctor will feel he is doing a better job if he cures 95% of patients rather than only

> 90%, even though a particular patient (Dr. Dunsmore) might do better in the latter situation than the former. As explained in the paper, preshrunk predictors do better than least squares for most **x**'s at the expense of doing worse at a minority of **x**'s. Perhaps if we think our symptoms are unusual we should seek a consultant who is prepared to view our complaint as an individual research problem rather than rely on the blunt instrument of conventional wisdom."

The implication for Bayesian estimators, in the context of neuroimaging, is that they are the best for each subject [or voxel] on average over subjects [or voxels]. In this sense Bayesian or conditional estimates of individual effects are only better on average, over the individual effects estimated. The issues, framed by Keith Worsley above, speak to the important consideration that Bayesian estimates, of the sort discussed in this paper, are only "better" in collective sense. One example of this collective context is presented in Friston *et al.* (2002), where between-voxel effects are used to "shrink" within-voxel estimates that are then reported together in a PPM.

The estimators and inference from a PEB approach do not inherently increase the sensitivity or specificity of the analysis. The most appropriate way to do this would be to simply increase sample size. PEB methodology can be better regarded as providing a set of estimates or predictors that are internally consistent within and over hierarchies of the observation model. Furthermore, they enable Bayesian inference (comments about the likelihood of an effect given the data) that complement classical inference (comments about the likelihood of the data). Bayesian inference does not necessarily decide whether an activation is present or not, it simply estimates the probability of an activation, specified in terms of the size of the effect. Conversely, classical inference is predicated on a decision (is the null hypothesis true or is the size of the effect different from zero?). The product of classical inference is a decision or declaration, which induces a sensitivity and specificity of the inference. In this paper we have used classical notions of sensitivity and specificity to link the two sorts of inference by thresholding the posterior probability. However, one is not compelled to threshold maps of posterior probability. Indeed, one of the motivations, behind Bayesian treatments, is to eschew the difficult compromise between sensitivity and specificity engendered by classical inference in neuroimaging.

APPENDIX

A.1 The EM Algorithm

This appendix describes the EM algorithm using a statistical mechanics perspective adopted by the machine learning community (Neal and Hinton, 1998). The second section of the appendix connects this formulation with classical ReML methods. We show that, in the context of linear observation models, the negative free energy is the same as the objective function maximized in classical schemes like restricted maximum likelihood (ReML).

The EM algorithm is ubiquitous in the sense that many estimation procedures can be formulated as such, from mixture models through to factor analysis. Its objective is to maximize the likelihood of the observed data $p(y|\lambda)$, conditional on some hyperparameters, in the presence of unobserved variables or parameters θ. This is equivalent to maximizing the log likelihood

$$\begin{aligned}\ln p(y|\lambda) &= \ln \int p(\theta, y|\lambda)d\theta \geq F(q, \lambda) \\ &= \int q(\theta)\ln p(\theta, y|\lambda)d\theta - \int q(\theta)\ln q(\theta)d\theta\end{aligned} \tag{A.1}$$

where $q(\theta)$ is any distribution over the model parameters (Neal and Hinton, 1998). Equation (A.1) rests on Jensen's inequality that follows from the concavity of the log function, which renders the log of an integral greater than the integral of the log. F corresponds to the negative free energy in statistical thermodynamics and comprises two terms, related to the energy (first term) and entropy (second term). The EM algorithm alternates between maximizing F, and implicitly the likelihood of the data, with respect to the distribution $q(\theta)$ and the hyperparameters λ, holding the other fixed

$$\text{E-step:} \quad q(\theta) \leftarrow \arg\max_q F(q(\theta), \lambda)$$

$$\text{M-step:} \quad \lambda \leftarrow \arg\max_\lambda F(q(\theta), \lambda)$$

This iterative alternation performs a co-ordinate ascent on F. It is easy to show that the maximum in the E-step obtains when $q(\theta) = p(\theta|y, \lambda)$, at which point (A.1) becomes an equality. The M-step finds the ML estimate of the hyperparameters, i.e., the values of λ that maximize $p(y|\lambda)$ by integrating $p(\theta, y|\lambda)$ over the parameters using the current estimate of their conditional distribution. In short the E-step computes sufficient statistics (in our case the conditional mean and covariance) relating to the distribution of the unobserved parameters to enable the M-step to optimize the hyperparameters, in a maximum likelihood sense, using this distribution. These new hyperparameters reenter into the estimation of the conditional distribution and so on until convergence.

The E-Step

In our hierarchical model, with Gaussian (i.e., parametric) assumptions, the E-step is trivial and corresponds to taking the conditional mean and covariance according to (15). These are then used, with the data, to estimate the hyperparameters of the covariance components in the M-step.

The M-Step

Given that we can reduce the problem to estimating the error covariances with the augmented expressions for the conditional mean and covariance (15) we only need to estimate the hyperparameters of the error covariances (which contain the prior covariances). Specifically, we require the hyperparameters that maximize the first term in the expression for F above. From (15)

$$\begin{aligned}\log p(\theta, y|\lambda) &= -\frac{1}{2}\ln|C_\epsilon| \\ &\quad -\frac{1}{2}(\bar{y} - \overline{X}\theta)^T C_\epsilon^{-1}(\bar{y} - \overline{X}\theta) \\ &\quad + const.\end{aligned}$$

$$\begin{aligned}\int q(\theta)\ln p(\theta, y|\lambda)d\theta &= -\frac{1}{2}\ln|C_\epsilon| - \frac{1}{2}r^T C_\epsilon^{-1} r \\ &\quad -\frac{1}{2}\langle(\theta - \eta_{\theta|y})^T \overline{X}^T \\ &\quad \times C_\epsilon^{-1}\overline{X}(\theta - \eta_{\theta|y})\rangle_q + const. \\ &= -\frac{1}{2}\ln|C_\epsilon| - \frac{1}{2}r^T C_\epsilon^{-1} r \qquad \text{(A.2)} \\ &\quad -\frac{1}{2}\operatorname{tr}\{C_{\theta|y}\overline{X}^T C_\epsilon^{-1}\overline{X}\} + const.\end{aligned}$$

$$\int q(\theta)\log q(\theta) = -\frac{1}{2}\ln|C_{\theta|y}| + const.$$

$$\begin{aligned}F &= \frac{1}{2}\ln|C_\epsilon^{-1}| - \frac{1}{2}r^T C_\epsilon^{-1} r \\ &\quad -\frac{1}{2}\operatorname{tr}\{C_{\theta|y}\overline{X}^T C_\epsilon^{-1}\overline{X}\} \\ &\quad +\frac{1}{2}\ln|C_{\theta|y}| + const.\end{aligned}$$

where the residuals $r = \bar{y} - \overline{X}\eta_{\theta|y}$. We now simply take the derivatives of F with respect to the hyperparameters and use some nonlinear search to find the maximum. Note that the second [entropy] term does not depend on the hyperparameters. There is an interesting intermediate derivative. From (A.2)

$$\frac{\partial F}{\partial C_\epsilon^{-1}} = \frac{1}{2}C_\epsilon - \frac{1}{2}rr^T - \frac{1}{2}\overline{X}C_{\theta|y}\overline{X}^T \qquad \text{(A.3)}$$

Setting this derivative to zero (at the maximum of F) requires

$$C(\lambda)_\epsilon = rr^T + \overline{X}C_{\theta|y}\overline{X}^T \qquad \text{(A.4)}$$

(c.f. Dempster *et al.* (1981) p. 350). Equation (A.4) says that the error covariance estimate has two components: that due to differences between the data observed and predicted by the conditional expectation of the parameters and another component due to the variation of the parameters about their conditional mean. More generally one can adopt a Fischer scoring algorithm and update the hyperparameters $\lambda \leftarrow \lambda + \Delta\lambda$ using the first and expected second partial derivatives of the negative free energy.

$$\begin{aligned}\Delta\lambda &= H^{-1}g \\ g_i &= \frac{\partial F}{\partial \lambda_i} = \operatorname{tr}\left\{-\frac{\partial F}{\partial C_\epsilon^{-1}}C_\epsilon^{-1}Q_i C_\epsilon^{-1}\right\} \\ &= -\frac{1}{2}\operatorname{tr}\{PQ_i\} + \frac{1}{2}\bar{y}^T P^T Q_i P\bar{y} \\ \frac{\partial^2 F}{\partial \lambda_{ij}^2} &= \frac{\partial g_i}{\partial \lambda_j} = \frac{1}{2}\operatorname{tr}\{PQ_i PQ_j\} - \bar{y}^T PQ_i PQ_j P\bar{y} \qquad \text{(A.5)} \\ H_{ij} &= E\left\{-\frac{\partial^2 F}{\partial \lambda_{ij}^2}\right\} = \frac{1}{2}\operatorname{tr}\{PQ_i PQ_j\} \\ P &= C_\epsilon^{-1} - C_\epsilon^{-1}\bar{X}C_{\theta|y}\bar{X}^T C_\epsilon^{-1}\end{aligned}$$

Fisher scoring corresponds to augmenting a simple Newton-Raphson scheme by replacing the second derivatives or "curvature" observed at the particular response y with its expectation over realizations of the data. The ensuing matrix H is referred to as Fisher's Information matrix.[2] The computation of the gradient vector g can be made computationally efficient by capitalizing on any sparsity structure in the constraints and by bracketing the multiplications appropriately. (A.5) is general in that it accommodates almost any form for the covariance constraints through a Taylor

[2] The derivation of the expression for the Information matrix uses standard linear algebra results and is most easily seen by: (i) differentiating the form for g in (A.7) by noting

$$\frac{\partial P}{\partial \lambda_j} = -PQ_j P$$

and (ii) taking the expectation, using $\langle\operatorname{tr}\{PQ_i P\overline{yy}^T PQ_j\}\rangle_q = \operatorname{tr}\{PQ_i PC_\epsilon PQ_j\} = \operatorname{tr}\{PQ_i PQ_j\}$.

expansion of $C\{\lambda\}_\epsilon$. In many instances the bases can be constructed so that they do not "overlap" or interact through the design matrix, i.e., $Q_iPQ_j = 0$ and estimates of the hyperparameters can be based directly on the first partial derivatives in (A.5) by solving for $g = 0$. For certain forms of $C(\lambda)_\epsilon$ the hyperparameters can be calculated very simply.[3] However, we work with the general solution above that encompasses all these special cases.

Once the hyperparameters have been updated they enter into (19) to give the new covariance estimate which, in turn enters (15) to give the new conditional estimates which re-enter into (A.5) to give new updates until convergence. A pseudo-code illustration of the complete algorithm is presented in Fig. 4. Note that in this implementation one is effectively performing a single Fisher scoring iteration for each M-step. One could postpone each E-step until this search converged but a single step is sufficient to perform a co-ordinate ascent on F. Technically this renders (A.5) a generalized EM or GEM algorithm.

It should be noted that the search for the maximum of F does not have to employ a Fisher scoring scheme or indeed the parameterization of C_ϵ used in (18). Other search procedures such as quasi-Newton searches are commonly employed (Fahrmeir and Tutz, 1994). Harville (1977) originally considered Newton-Raphson and scoring algorithms, and Laird and Ware (1982) recommend several versions of the EM algorithm. One limitation of the hyper-parameterization described above is that does not guarantee that C_ϵ is positive definite. This is because the hyperparameters can take negative values with extreme degrees of nonsphericity. The EM algorithm employed by multistat (Worsley *et al.*, 2002), for variance component estimation in multi-subject fMRI studies, uses a slower but more stable EM algorithm that ensures positive definite covariance estimates. The common aspect of all these algorithms is that they (explicitly or implicitly) maximize F (or minimize free energy). As shown next, this is equivalent to the method of restricted maximum likelihood.

A.2. Relationship to ReML

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). It is commonly employed in standard statistical packages (e.g., SPSS). Under the present model assumptions ReML is formally identical to EM. One can regard ReML as embedding the E-step into the M-step to provide a single log-likelihood objective function: Substituting the $C_{\theta|y} = (\overline{X}^TC_\epsilon^{-1}\overline{X})^{-1}$ from (15) into the expression for the negative free energy (A.2) gives

$$F = -\frac{1}{2}\ln|C_\epsilon| - \frac{1}{2}r^TC_\epsilon^{-1}r - \frac{1}{2}\ln|\overline{X}^TC_\epsilon^{-1}\overline{X}| + const. \tag{A.6}$$

which is the ReML objective function (see Harville, 1977, p. 325). Critically the derivatives of (A.6), with respect to the hyperparameters, are exactly the same as those given in (A.5).[4] Operationally, (A.5) can be rearranged to give a ReML scheme by removing any explicit reference to the conditional covariance.

$$\begin{aligned} g_i &= -\frac{1}{2}\operatorname{tr}\{PQ_i\} + \frac{1}{2}\operatorname{tr}\{P\overline{y}\overline{y}^TP^TQ_i\} \\ H_{ij} &= \frac{1}{2}\operatorname{tr}\{PQ_iPQ_j\} \qquad\qquad (A.7) \\ P &= C_\epsilon^{-1} - C_\epsilon^{-1}\overline{X}(\overline{X}^TC_\epsilon^{-1}\overline{X})^{-1}\overline{X}^TC_\epsilon^{-1} \end{aligned}$$

These expressions are formally identical to those described in Section 5 of Harville (1977, p. 326). Because (A.7) does not depend explicitly on the conditional density, one could think of ReML as estimating the hyperparameters in a subspace that is restricted in the sense that the estimates are conditionally independent of the parameters. See Harville (1977) for a discussion of expressions, comparable to the terms in (A.7) that are easier to compute, for particular hyper-parameterizations of the variance components.

The particular form of (A.7) has a very useful application when y is a multivariate data matrix and the hyperparameters are the same for all columns (i.e., voxels). Here, irrespective of the voxel-specific parameters, the voxel-wide hyperparameters can be obtained efficiently by iterating (A.7) using the sample covari-

[3] Note that if there is only one hyperparameter then $g = 0$ can be solved directly

$$\operatorname{tr}\{PQ\} = \bar{y}PQP\bar{y} \Rightarrow \lambda = \frac{r^TQ^{-1}r}{\operatorname{tr}\{R\}}$$

where $C_\epsilon = \lambda Q$ and $R = I - \overline{X}(\overline{X}^TQ^{-1}\overline{X})^{-1}\overline{X}^TQ^{-1}$ is a residual forming matrix. This is the expression used in classical schemes, given the correlation matrix Q, to estimate the error covariance using the sum of squared de-correlated residuals.

[4] Note that

$$\begin{aligned} \frac{\partial\ \ln|\overline{X}^TC_\epsilon^{-1}\overline{X}|}{\partial\lambda_i} &= \operatorname{tr}\left\{(\overline{X}^TC_\epsilon^{-1}\overline{X})^{-1}\frac{\partial\overline{X}^TC_\epsilon^{-1}\overline{X}}{\partial\lambda_i}\right\} \\ &= -\operatorname{tr}\{C_{\theta|y}\overline{X}^TC_\epsilon^{-1}Q_iC_\epsilon^{-1}\overline{X}\} \end{aligned}$$

ance matrix yy^T. This is possible because the conditional parameter estimates are not required in the ReML formulation. This is used in the current development version of the SPM software to estimate voxel-wide nonsphericity.

A.3 Hyper-Parameterizing the Precision

In the forgoing we have parameterized the covariances of the likelihood and prior densities. This is natural when the priors become variance components on augmenting hierarchical models. However, there are other situations when the priors are more naturally specified in terms of precisions, each precision component $U_j^{(i)}$ corresponding to the jth independent prior specified for the ith level of the model. Following augmentation we get

$$C_\epsilon^{-1} = U_\theta + \sum \lambda_k U_k$$

$$U_k = \frac{\partial C_\epsilon^{-1}}{\partial \lambda_k}$$

$$U_\theta = \begin{bmatrix} 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \\ 0 & \cdots & 0 & C_\theta^{(n)-1} \end{bmatrix}, \qquad \text{(A.8)}$$

$$U_k = \begin{bmatrix} 0 & & \cdots & & 0 & 0 \\ & \ddots & & & & \\ \vdots & & U_j^{(i)} & & \vdots & \vdots \\ & & & \ddots & & \\ 0 & & \cdots & & 0 & 0 \\ 0 & & \cdots & & 0 & 0 \end{bmatrix}$$

c.f. Equation (19). Notice that with this hyper-parameterization large values of the hyperparameters correspond to high precision and a small variance component contribution. The Fisher scoring scheme of (A.5) now takes a slightly simpler form,

$$g_i = \frac{1}{2}\,\mathrm{tr}\{OU_i\} - \frac{1}{2}\, r^T U_i r$$

$$H_{ij} = \frac{1}{2}\,\mathrm{tr}\{OU_i O U_j\} \qquad \text{(A.9)}$$

$$O = C_\epsilon P C_\epsilon$$

that is most easily derived by noting $Q = \partial C_\epsilon / \partial \lambda_i = -C_\epsilon \times (\partial C_\epsilon^{-1} / \partial \lambda_i)\, C_\epsilon = -C_\epsilon U_i C_\epsilon$ and substituting in (A.5).

This approach will be illustrated in a subsequent paper that uses a simple version of (A.9) to find the right mixture of structural and functional priors in the EEG source reconstruction problem (Phillips *et al.*, in preparation). This application effectively solves the problem of identifying the most appropriate regularization hyperparameters using an empirical Bayesian scheme.

ACKNOWLEDGMENTS

The Wellcome Trust and MRC (UK) funded this work. We thank members of the Gatsby Computational Neuroscience Unit for tuition and guidance during its development. We also thank Pasi Karjalainen for a clear-minded synthesis of model estimation issues. Finally we thank Keith Worsley for introducing us to ReML and helping with the conceptual issues arising from this work.

REFERENCES

Copas, J. B. 1983. Regression prediction and shrinkage. *J. R. Statistical Soc. Series B* **45:** 311–354.

Dempster, A. P., Laird, N. M., and Rubin 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Series B* **39:** 1–38.

Dempster, A. P., Rubin, D. B., and Tsutakawa, R. K. 1981. Estimation in covariance component models. *J. Am. Stat. Assoc.* **76:** 341–353.

Descombes, X., Kruggel, F., and von Cramon, D. Y. 1998. fMRI signal restoration using a spatio-temporal Markov random field preserving transitions. *NeuroImage* **8:** 340–349.

Efron, B., and Morris, C. 1973. Stein's estimation rule and its competitors—An empirical Bayes approach. *J. Am. Stat. Assoc.* **68:** 117–130.

Efron, B., and Morris, C. 1977. Stein's paradox in statistics. *Sci. Am.* **May:** 119–127.

Everitt, B. S., and Bullmore, E. T. 1999. Mixture model mapping of brain activation in functional magnetic resonance images. *Hum. Brain Mapp.* **7:** 1–14.

Fahrmeir, L., and Tutz, G. 1994. *Multivariate Statistical Modelling Based on Generalized Linear Models,* pp. 355–356. Springer-Verlag, New York.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-B., Frith, C. D., and Frackowiak, R. S. J. 1995. Statistical parametric maps in functional imaging: A general linear approach. *Hum. Brain Mapp.* **2:** 189–210.

Friston, K. J., Glaser, D. E., Henson, R. N. A., Kiebel, S., Phillips, C., and Ashburner, J. 2002. Classical and Bayesian inference in neuroimaging: Applications. *NeuroImage* **16:** 484–512.

Hartley, H. 1958. Maximum likelihood estimation from incomplete data. *Biometrics* **14:** 174–194.

Hartvig, N. V., and Jensen, J. L. 2002. Spatial mixture modelling of fMRI data. *Hum. Brain Mapp.*, in press.

Harville, D. A. 1977. Maximum likelihood approaches to variance component estimation and to related problems. *J. Am. Stat. Assoc.* **72:** 320–338.

Højen-Sørensen, P., Hansen, L. K., and Rasmussen, C. E. 2000. Bayesian modelling of fMRI time-series. In *Advances in Neural Information Processing Systems* (S. A. Solla, T. K. Leen, and K. R. Muller, Eds.), Vol. 12, pp. 754–760. MIT Press.

Holmes, A. P., and Friston, K. J. 1998. Generalizability, random effects and population inference. *NeuroImage* S754.

Kass, R. E., and Steffey, D. 1989. Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J. Am. Stat. Assoc.* **407:** 717–726.

Laird, N. M., and Ware, J. H. 1982. Random effects models for longitudinal data. *Biometrics* **38:** 963–974.

Lee, P. M. 1997. *Bayesian Statistics: An Introduction*. Wiley, New York.

Neal, R. M., and Hinton, G. E. 1998. A view of the EM algorithm that justifies incremental, sparse and other variants. In *Learning in Graphical Models* (M. I. Jordan, Ed.), pp. 355–368. Kluwer Academic Press.

Phillips, C., Rugg, M. D., and Friston, K. J. 2002. Systematic regularisation for linear inverse solutions of the EEG source localization problem. Submitted.

Tikhonov, A. N., and Arsenin, V. Y. 1977. *Solution of Ill Posed Problems.* Winston and Sons.

Worsley, K. J. 1994. Local Maxima and the expected Euler characteristic of excursion sets of chi squared, F and t fields. *Adv. Appl. Prob.* **26:** 13–42.

Worsley, K. J., Liao, C., Aston, J., Petre, V., Duncan, G. H., and Evans, A. C. 2002. A general statistical analysis for fMRI data. *NeuroImage* **15:** 1–15.