

# Classical and Bayesian inference

Karl J Friston and Will Penny

The Wellcome Dept. of Cognitive Neurology,  
University College London  
Queen Square, London, UK WC1N 3BG  
Tel (44) 020 7833 7456  
Fax (44) 020 7813 1445  
email [k.friston@fil.ion.ucl.ac.uk](mailto:k.friston@fil.ion.ucl.ac.uk)

## Contents

---

<b>I</b>	<b>INTRODUCTION</b>
<b>II</b>	<b>THEORY</b>
<b>III</b>	<b>EM AND VARIANCE COMPONENT ESTIMATION</b>
<b>IV</b>	<b>POSTERIOR PROBABILITY MAPPING AND PPMs</b>
<b>V</b>	<b>BAYESIAN IDENTIFICATION OF DYNAMIC SYSTEMS</b>
<b>VI</b>	<b>APPENDIX</b>
	<b>REFERENCES</b>

---

## I INTRODUCTION

This chapter revisits hierarchical observation models (see **Chapter 13: Hierarchical models**), used in functional neuroimaging, in a Bayesian light. It emphasises 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 *expectation maximisation* or EM. 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 characterise 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).

The first section of this chapter is a theoretical prelude to subsequent sections that deal with applications of the theory to a range of important issues in neuroimaging. These issues include; (i) Estimating non-sphericity 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) Bayesian models for imaging data, in which effects at one voxel are constrained by responses in others and (iii) Bayesian estimation of nonlinear models of hemodynamic responses. Although diverse, all these estimation problems are accommodated by the EM framework described in this chapter.

### A Classical and Bayesian inference

Since its inception, about ten years ago, *statistical parametric mapping* (SPM) has proved useful for characterising neuroimaging data sequences. However, SPM is limited because it is based on classical inference procedures. In this chapter we introduce a more general framework, which places SPM in a broader context and

points to alternative ways of characterising 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 the *expectation maximisation* 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 (**see Chapters 14 and 15: [Introduction to] 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 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 colour *but not motion* and another responds to motion *but not colour*. Secondly, 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 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 chapter describes one such approach. In contradistinction to other proposals, this approach is not a novel way of analysing neuroimaging data. The use of a Bayesian formalism in special models for fMRI data has been usefully explored elsewhere *e.g.* spatio-temporal 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 Højten-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 here 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, **Chapter 12: Random effects analysis**) 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 characterisation of single-subject responses, both in terms of the estimated effects and the nature of the inference.

## **B Overview**

The aim of the first section below 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 non-sphericity that has to be estimated before any inferences about an effect can be made. Important sources of non-sphericity 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 the second section to emphasise both the covariance component estimation and Bayesian perspectives, in terms of the difference between response estimates based on classical maximum likelihood estimators and the conditional means from a Bayesian approach.

In the third section we use the same theory to elaborate hierarchical models that allow the construction of Posterior Probability Maps (PPMs). Again this employs two-level models but focuses on Bayesian inference at the first level. It complements the preceding fMRI application by showing how priors can be estimated using observations *over voxels* at the second level. The final section addresses the Bayesian identification of dynamic systems where empirical Bayesian priors are replaced by knowledge about the biophysics that underlies hemodynamic responses (see **Chapter 11: Hemodynamic Modelling**). This approach will be can be used to characterise hemodynamic responses at a single voxel or, indeed, the response of a network of coupled brain regions (see **Chapter 22: Dynamic Causal Modelling**).

## II. THEORY

In this section we focus on theory and procedures. The key points are reprised in subsequent sections where they are illustrated using real and simulated data. This section 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 behaviour 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  $\sigma^2$ ). 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 chapter 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 section 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.

First we reprise hierarchical linear observation models that form the cornerstone of the ensuing 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 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 that rests on expectation maximisation (EM). This is described briefly 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 part of this 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 behaviour 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 generalised linear models, see Fahrmeir and Tutz (1994). This text provides an exposition of EM algorithm and PEB in linear models, usefully relating EM to classical methods (*e.g.* ReML p225). For an introduction to Bayesian statistics see Lee (1997). This text adopts a more explicit Bayesian perspective and again usefully connects empirical Bayes with classical approaches, *e.g.* the Stein “Shrinkage” estimator and empirical Bayes estimators used below (p232). In most standard texts the hierarchical models considered in the next section are referred to as random effects models.

### **A Hierarchical Linear observation models**

We will deal with hierarchical linear observation models of the form

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

under Gaussian assumptions about the errors  $\varepsilon^{(i)} \sim N\{0, C_\varepsilon^{(i)}\}$ .  $y$  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 (see **Chapter 7: The General Linear Model**). 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 activation 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 Figure 1 (these examples are used in next Section). 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 modelled for each session, effects expressed in each session modelled 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 generalised to non-Gaussian densities) much more transparent.

Figure 1 about here



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 chapter 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

$$y = \varepsilon^{(1)} + X^{(1)}\varepsilon^{(2)} + \dots + X^{(1)} \dots X^{(n-1)}\varepsilon^{(n)} + X^{(1)} \dots X^{(n)}\theta^{(n)} \quad 3$$

In this non-hierarchical 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

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

where  $C_{\varepsilon}^{(i)} = Cov\{\varepsilon^{(i)}\}$ . 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

$$\begin{aligned} y &= \tilde{X}\theta^{(n)} + \tilde{\varepsilon} \\ \tilde{X} &= X^{(1)}X^{(2)} \dots X^{(n)} \\ \tilde{\varepsilon} &= \varepsilon^{(1)} + X^{(1)}\varepsilon^{(2)} + \dots + X^{(1)}X^{(2)} \dots X^{(n-1)}\varepsilon^{(n)} \end{aligned} \tag{5}$$

and Bayesian estimation

$$\begin{aligned} y &= X\theta + \varepsilon^{(1)} \\ X &= [X^{(1)}, \dots, X^{(1)}X^{(2)} \dots X^{(n-1)}, X^{(1)}X^{(2)} \dots X^{(n)}] \\ \theta &= \begin{bmatrix} \varepsilon^{(2)} \\ \vdots \\ \varepsilon^{(n)} \\ \theta^{(n)} \end{bmatrix} \end{aligned} \tag{6}$$

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  $n$ th 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  $\theta$  comprises the errors at all levels 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 Figure 1 for a graphical depiction).

## B A Classical perspective

From a classical perspective (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

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

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

$$T = c^T \eta_{ML} / \sqrt{c^T \text{Cov}\{\eta_{ML}\} c}\tag{8}$$

where, from (5) and (7)

$$\begin{aligned}\text{Cov}\{\eta_{ML}\} &= M C_{\tilde{\varepsilon}} M^T = (\tilde{X}^T C_{\tilde{\varepsilon}}^{-1} \tilde{X})^{-1} \\ C_{\tilde{\varepsilon}} &= C_{\varepsilon}^{(1)} + X^{(1)} C_{\varepsilon}^{(2)} X^{(1)T} \dots + X^{(1)} \dots X^{(n-1)} C_{\varepsilon}^{(n)} X^{(n-1)T} \dots X^{(1)T}\end{aligned}\tag{9}$$

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  $C_{\tilde{\varepsilon}} = \text{Cov}\{\tilde{\varepsilon}\}$ , from all levels, projected down the hierarchy onto the response variable or observation space. 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 multi-stage implementations of random-effect analyses

(Holmes and Friston 1998, **Chapter 12: Random effects analysis**). Although the ensuing variance estimator is not the same as equation (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.

### C 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 to finding its first two moments, the conditional mean  $\eta_{\theta|y}^{(i)}$  and conditional covariance  $C_{\theta|y}^{(i)}$ . 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  $\eta_{\theta|y}^{(i)}$  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

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

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. 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<sup>1</sup> (*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  $\theta$  from (6). From Bayes rule the posterior probability is proportional to the likelihood of obtaining the data, conditional on  $\theta$ , times the prior probability of  $\theta$ ,

$$p(\theta | y) \propto p(y | \theta)p(\theta) \quad 11$$

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

$$\eta_{\theta} = E\{\theta\} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \eta_{\theta}^{(n)} \end{bmatrix}, \quad C_{\theta} = Cov\{\theta\} = \begin{bmatrix} C_{\varepsilon}^{(2)} & \dots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & C_{\varepsilon}^{(n)} & 0 \\ 0 & \dots & 0 & C_{\theta}^{(n)} \end{bmatrix}, \quad \begin{cases} C_{\theta}^{(n)} = \infty & \text{unknown} \\ C_{\theta}^{(n)} = 0 & \text{known} \end{cases}$$

12

Under Gaussian assumptions the likelihood and priors are given by

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

$$p(\theta) \propto \exp\left\{-\frac{1}{2}(\theta - \eta_{\theta})^T C_{\theta}^{-1}(\theta - \eta_{\theta})\right\} \quad 13$$

---

<sup>1</sup> 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

Substituting (12) into (10) gives a posterior density with a Gaussian form

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

where

$$\begin{aligned} C_{\theta|y} &= (X^T C_{\varepsilon}^{(1)-1} X + C_{\theta}^{-1})^{-1} \\ \boldsymbol{\eta}_{\theta|y} &= C_{\theta|y} (X^T C_{\varepsilon}^{(1)-1} y + C_{\theta}^{-1} \boldsymbol{\eta}_{\theta}) \end{aligned} \tag{14}$$

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

The solution (14) is ubiquitous in the estimation literature and is presented under various guises in 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 pre-whitening. If the errors are assumed to be independently and identically distributed, *i.e.*  $C_{\varepsilon}^{(1)} = I$ , then (14) reduces to the ordinary least square estimator. With non-flat priors the form of (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

$$\begin{aligned} C_{\theta|y} &= (\bar{X}^T C_{\varepsilon}^{-1} \bar{X})^{-1} \\ \boldsymbol{\eta}_{\theta|y} &= C_{\theta|y} (\bar{X}^T C_{\varepsilon}^{-1} \bar{y}) \end{aligned} \tag{15}$$

---

becomes flat.

where

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

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

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

See Figure 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  $C_\varepsilon$  that determine the weighting. This is exactly where we ended up in the classical approach, namely reduction to a covariance component estimation problem.

Figure 2 about here

#### **D Covariance component estimation**

The classical approach was portrayed above, 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  $C_\theta^{(i-1)} = C_\varepsilon^{(i)}$ . Both approaches rest on estimating the covariance components. This estimation depends upon some parameterisation of these components; in this chapter we use  $C_\varepsilon^{(i)} = \sum \lambda_j^{(i)} Q_j^{(i)}$  where  $\lambda_j^{(i)}$  are some hyperparameters and  $Q_j^{(i)}$  represent a 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  $j$ th covariance component at the  $i$ th level and can 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 non-sphericity anticipated. For example, they could specify serial correlations within-subject or correlations among the errors induced hierarchically, by repeated measures over subjects (Figure 3 illustrates both these examples). We will illustrate a number of forms for  $Q_j$  in the subsequent sections.

Figure 3 about here.

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

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

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 (16) are generally zero. In this context a linear decomposition of  $C_\epsilon^{(i)}$  is a natural parameterisation 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 parameterisation may be employed. For example, if the constraints were implementing several independent priors in a non-hierarchical model a more natural expansion might be in terms of the precision  $C_\theta^{-1} = \sum \lambda_j Q_j$ . The precision is simply the inverse of the covariance matrix. Here  $Q_j$  correspond to precisions specifying the form of independent prior densities. However, in this chapter, we deal only with priors that are engendered by the observation model that induces hierarchically organised, linearly mixed, variance



components. See Harville (1977, p322) 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

$$C_\varepsilon = Q_\theta + \sum \lambda_k Q_k$$

$$Q_k = \frac{\partial C_\varepsilon}{\partial \lambda_k} \quad 17$$

$$Q_\theta = \begin{bmatrix} 0 & \dots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & 0 \\ 0 & \dots & 0 & C_\theta^{(n)} \end{bmatrix}, \quad Q_k = \begin{bmatrix} 0 & \dots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & 0 \\ 0 & \dots & 0 & 0 \end{bmatrix}$$

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.

### E Expectation-Maximisation

EM or expectation-maximisation is a generic, iterative parameter re-estimation procedure that encompasses many iterative schemes devised to estimate the parameters and hyperparameters of a model (Dempster *et al* 1977, 1981). It was originally introduced as an iterative method to obtain maximum likelihood estimators in incomplete data situations (Hartley 1958) and was generalised 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 **E**xpectation of the parameters, holding the hyperparameters fixed and an **M**-step, which updates the **M**aximum likelihood estimate of the hyperparameters, keeping the parameters fixed.

In brief, EM 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) Restricted ML estimates of the hyperparameters. The EM algorithm described in the appendix (A.1) is depicted schematically in Figure 4. In the context of the linear observation models discussed in this chapter, 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, p226). 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 *Restricted* ML 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.

Figure 4 about here

### F Conditional and classical estimators

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

$$\eta_{\theta|y} = E\{\theta | y\} = \begin{bmatrix} \eta_{\varepsilon|y}^{(2)} \\ \vdots \\ \eta_{\varepsilon|y}^{(n)} \\ \eta_{\theta|y}^{(n)} \end{bmatrix}, \quad C_{\theta|y} = Cov\{\theta | y\} = \begin{bmatrix} C_{\varepsilon|y}^{(2)} & \cdots & & \\ & \ddots & & \\ & & C_{\varepsilon|y}^{(n)} & \\ & & & C_{\theta|y}^{(n)} \end{bmatrix} \quad 18$$

The conditional means for each level obtain recursively with  $\eta_{\theta|y}^{(i-1)} = X^{(i)}\eta_{\theta|y}^{(i)} + \eta_{\varepsilon|y}^{(i)}$ .

The conditional covariances are simply  $C_{\theta|y}^{(i-1)} = C_{\varepsilon|y}^{(i)}$  up to the penultimate level and  $C_{\theta|y}^{(n)}$  at the final level.

The conditional means represent a better ‘collective’ characterisation of the model parameters than the equivalent ML estimates because they are constrained by prior information from higher levels (see discussion below). 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 standardised conditional mean.

The standardised conditional mean, or a contrast of means, is the mean normalised 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).

$$T^{(i)} = c^T \eta_{\theta|y}^{(i)} / \sqrt{c^T C_{\theta|y}^{(i)} c} \quad 19$$

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 emphasise 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 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.

### **G Classical and Bayesian inference compared**

In this subsection 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 standardised 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  $\gamma$ . 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  $\gamma$ . 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 characterised 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  $\alpha$  is a small false positive rate. The sensitivity  $\beta$  or power is the probability of correctly rejecting the null hypothesis. Classically, one rejects the null hypothesis whenever the standardised ML estimator exceeds some specified statistical threshold  $v$ . The probability of this happening is based on its distribution whose standard deviation is given by (9).

$$\begin{aligned}\alpha &= 1 - \Phi(v) \\ \beta &= 1 - \Phi\left(v - \frac{A}{\sqrt{(X^T C_\varepsilon^{-1} X)^{-1}}}\right)\end{aligned}\tag{20}$$

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.  $\alpha$  and  $\beta$  are the probabilities that the ML estimator divided by its standard deviation would exceed  $v$ , 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) 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, 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

$$\langle \eta_{\theta|y} \rangle = \begin{cases} 0 & \text{null} \\ C_{\theta|y} X^T C_\varepsilon^{-1} X A & \text{alternate} \end{cases}$$

$$\text{Cov}\{\eta_{\theta|y}\} = C_\eta = C_{\theta|y} X^T C_\varepsilon^{-1} X C_{\theta|y}$$

from which it follows

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

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 inconstant over voxels. In a classical approach we would choose some fixed threshold  $v$ , say for all voxels in an SPM. This ensures that the resulting inference has the same specificity everywhere because specificity depends on, and only on,  $v$ . 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  $\gamma$  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

$\gamma$  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  $\gamma$  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 a fallacy 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 non-trivial 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 behaviour 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  $\gamma$ ) becomes a useful characterisation. This characterisation 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 non-trivial 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 problems because the ensuing inference has an intrinsically high specificity.

## **H Conceptual issues**

This section has introduced three key components that play a role in the estimation of the linear models; Bayesian estimation, hierarchical models and EM. 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 parameterised in terms of some hyperparameters 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 Figure 5).

Figure 5 about here

- 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 model 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 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 Figures 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 Figure 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 section 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 EM. The use of EM, for variance component estimation, is not limited to hierarchical models but finds a useful application whenever non-sphericity of the errors is specified with more than one hyperparameter (*e.g. serial correlations in fMRI*). This application will be illustrated next. 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 non-hierarchical 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, p232) 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? 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 summarised in a discussion chapter 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 it 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 future  $x$ ’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 chapter 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 chapter 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 chapter, pre-shrunk 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 chapter, are only ‘better’ in collective sense. One example of this collective context is presented below, 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 section 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.

### III EM AND VARIANCE COMPONENT ESTIMATION

In this section we present a series of models that exemplify the diversity of problems that can be addressed with EM. In hierarchical linear observation models, both classical and empirical Bayesian approaches can be framed in terms of *covariance component estimation* (e.g. variance partitioning). To illustrate the use of Expectation-Maximisation (EM) in covariance component estimation we focus on two important problems in fMRI: non-sphericity induced by (i) serial or temporal correlations among errors and (ii) variance components caused by the hierarchical nature of multi-subject studies. In hierarchical observation models, variance components at higher levels can be used as constraints on the parameter estimates of lower levels. This enables the use of parametric empirical Bayesian (PEB) estimators, as distinct from classical maximum likelihood (ML) estimates. We develop this distinction to address the difference between response estimates based on ML and the conditional means.

Empirical Bayes enables the joint estimation of an observation model's parameters (e.g. activations) and its hyperparameters that specify the observation's variance components (e.g. within- and between subject-variability). The estimation procedures conform to EM, which, considering just the hyperparameters in linear observation models, is formally identical to restricted maximum likelihood (ReML). If there is only one variance component these iterative schemes simplify to conventional, non-iterative sum of squares variance estimates. However, there are many situations when a number of hyperparameters have to be estimated. For example, when the correlations among errors are unknown but can be parameterised with a small number of hyperparameters (c.f. serial correlations in fMRI time-series). Another important example, in fMRI, is the multi-subject design in which the hierarchical nature of the observation induces different variance components at each level. The aim of this section is to illustrate how variance component estimation, with EM, can proceed in both single-level and hierarchical contexts. In particular, the examples emphasise that although the mechanisms inducing non-sphericity can be very different, the variance component estimation problems they represent, and the analytic approaches called for, are identical.

We will use two fMRI examples. In the first we deal with the issue of variance component estimation using serial correlations in single-subject fMRI studies.

Because there is no hierarchical structure to this problem there is no Bayesian aspect. However, in the second example we add a second level to the observation model for the first to address inter-subject variability. Endowing the model with a second level affords the opportunity to use empirical Bayes. This enables a quantitative comparison of classical and conditional single-subject response estimates.

### **A Variance component estimation in fMRI: A single-level model**

In this section we review serial correlations in fMRI and use simulated data to compare ReML estimates, obtained with EM, to estimates of correlations based simply on the model residuals. The importance of modelling temporal correlations, for classical inference based on the T statistic, is discussed in terms of correcting for non-sphericity in fMRI time-series. This section concludes with a quantitative assessment of serial correlations within and between subjects.

#### *1 Serial correlations in fMRI*

In this section we restrict ourselves to a single-level model and focus on the covariance component estimation afforded by EM. We have elected to use a simple but important covariance estimation problem to illustrate one of the potential uses of the scheme described in the appendix. Namely, serial correlations in fMRI embodied in the error covariance matrix for the first (and only) level of this observation model  $C_{\varepsilon}^{(1)}$ . Serial correlations have a long history in the analysis of fMRI time-series. fMRI time-series can be viewed as a linear admixture of signal and noise. Noise has many contributions that render it rather complicated in relation to other neurophysiological measurements. These include neuronal and non-neuronal sources. Neuronal noise refers to neurogenic signal not modelled by the explanatory variables and has the same frequency structure as the signal itself. Non-neuronal components have both white (*e.g.* R.F. noise) and coloured components (*e.g.* pulsatile motion of the brain caused by cardiac cycles and local modulation of the static magnetic field  $B_0$  by respiratory movement). These effects are typically low frequency or wide-band and induce long range correlations in the errors over time. These serial correlations can either be used to whiten the data (Bullmore *et al* 1996, Purdon and Weisskoff 1998) or are entered into the non-sphericity corrections described in previous chapters (Worsley and Friston 1995). Both approaches depend upon an accurate estimation of

the serial correlations. In order to estimate correlations among the errors  $C(\lambda)_\varepsilon$ , in terms of some hyperparameters  $\lambda$ , one needs both the residuals of the model  $r$  and the conditional covariance of the parameter estimates that produced those residuals. These combine to give the required error covariance (*c.f.* Equation A.4 in Appendix A.1).

$$C(\lambda)_\varepsilon = rr^T + XC_{\theta|y}X^T \quad 22$$

$XC_{\theta|y}X^T$  represents the conditional covariance of the parameter estimates  $C_{\theta|y}$  ‘projected’ onto the measurement space, by the design matrix  $X$ . The problem is that the covariance of the parameter estimates *is itself a function of the error covariance*. This circular problem is solved by the recursive parameter re-estimation implicit in EM. It is worth noting that estimators of serial correlations based solely on the residuals (produced by any estimator) will be biased. This bias results from ignoring the second term in (22), which accounts for the component of error covariance due to uncertainty about the parameter estimates themselves. It is likely that any valid recursive scheme for estimating serial correlations in fMRI time-series conforms to EM (or ReML) even if the connection is not made explicit. See Worsley *et al* (2002) for a non-iterative approach to AR( $p$ ) models.

In summary, the covariance estimation afforded by EM can be harnessed to estimate serial correlations in fMRI time series that coincidentally provide the most efficient (*i.e.* Gauss-Markov) estimators of the effect one is interested in. In this section we apply the EM algorithm described in Friston *et al* (2002a) to simulated fMRI data sequences and take the opportunity to establish the connections among some commonly employed inference procedures based upon the T statistic. This example concludes with an application of EM to empirical data to demonstrate quantitatively the relative variability in serial correlations over voxels and subjects.

## 2 Estimating serial correlations

For each fMRI session we have a single-level observation model that is specified by the design matrix  $X^{(1)}$  and constraints on the observation’s covariance structure  $Q_i^{(1)}$ , in this case serial correlations among the errors.

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

$$Q_1^{(1)} = I \tag{23}$$

$$Q_2^{(1)} = KK^T, \quad k_{ij} = \begin{cases} e^{j-i} & i > j \\ 0 & i \leq j \end{cases}$$

$y$  is the measured response with errors  $\varepsilon^{(1)} \sim N\{0, C_\varepsilon^{(1)}\}$ .  $I$  is the identity matrix. Here  $Q_1^{(1)}$  and  $Q_2^{(1)}$  represent covariance components of  $C_\varepsilon^{(1)}$  that model a white noise and an autoregressive AR(1) process with an AR coefficient of  $1/e = 0.3679$ . Notice that this is a very simple model of autocorrelations; by fixing the AR coefficient there are just two hyperparameters that allow for different mixtures of an AR(1) process and white noise (*c.f.* the 3 hyperparameters needed for a full AR(1) plus white noise model). The AR(1) component is modelled as an exponential decay of correlations over non-zero lag.

These bases were chosen given the popularity of AR plus white noise models in fMRI (Purdon and Weisskoff 1998). Clearly this basis set can be extended in any fashion using Taylor expansions to model deviations of the AR coefficient from  $1/e$  or indeed model any other form of serial correlations. Non-stationary autocorrelations can be modelled by using non-Toeplitz forms for the bases that allow the elements in the diagonals of  $Q_i^{(1)}$  to vary over observations. This might be useful, for example, in the analysis of event-related potentials, where the structure of errors may change with peri-stimulus time.

In the examples below the covariance constraints were scaled to a maximum of one. This means that the second hyperparameter can be interpreted as the covariance between one scan and the next. The basis set enters, along with the data, into the EM algorithm (see appendix A.1) to provide ML estimates of the parameters  $\theta^{(1)}$  and ReML estimates of the hyperparameters  $\lambda^{(1)}$ .

An example, based on simulated data, is shown in Figure 6. In this example the design matrix comprised a boxcar regressor and the first 16 components of a discrete cosine set. The simulated data corresponded to a compound of this design matrix (see



figure legend) plus noise, coloured using hyperparameters of 1 and 0.5 for the white and AR(1) components respectively. The top panel shows the data (dots), the true and fitted effects (broken and sold lines). For comparison, fitted responses based on both ML and OLS (ordinary least squares) are provided. The insert in the upper panel shows these estimators are very similar but not identical. The lower panel shows the true (dashed) and estimated (solid) auto-correlation function based on  $C_{\varepsilon}^{(1)} = \lambda_1^{(1)} Q_1^{(1)} + \lambda_2^{(1)} Q_2^{(1)}$ . They are nearly identical. For comparison the sample autocorrelation function (dotted line) and an estimate based directly on the residuals [*i.e.* ignoring the second term of (1)] (dot-dash line) are provided. The underestimation, that ensues using the residuals, is evident in the insert that shows the true hyperparameters (black), those estimated properly using ReML (white) and those based on the residuals alone (grey). By failing to account for the uncertainty about the parameter estimates, the hyperparameters based only on the residuals are severe underestimates. The sample autocorrelation function even shows negative correlations. This is a result of fitting the low frequency components of the design matrix. One way of understanding this is to note that the autocorrelations among the residuals are not unbiased estimators of  $C_{\varepsilon}^{(1)}$  but  $RC_{\varepsilon}^{(1)}R^T$ , where  $R$  is the residual-forming matrix. In other words, the residuals are not the true errors but what is left after projecting them onto the null space of the design matrix.

The full details of this simulated single-session, boxcar design fMRI study are provided in the figure legend.

Figure 6 about here

### 3 Inference in the context of non-sphericity<sup>2</sup>

This subsection explains why covariance component estimation is so important for inference. In short, although the parameter estimates may not depend on sphericity, the standard error, and ensuing statistics do. The impact of serial correlations on inference was noted early in the fMRI analysis literature (Friston *et al* 1994) and led to the generalised least squares (GLS) scheme described in Worsley and Friston

---

<sup>2</sup> An *i.i.d.* process is identically and independently distributed and has a probability distribution whose iso-contours conform to a *sphere*. Any departure from this is referred to as non-sphericity.

(1995). In this scheme one starts with any observation model that is pre-multiplied by some weighting or convolution matrix  $S$  to give

$$Sy = SX^{(1)}\theta^{(1)} + S\epsilon^{(1)} \quad 24$$

The GLS parameter estimates and their covariance are

$$\begin{aligned} \eta_{LS} &= Ly \\ Cov\{\eta_{LS}\} &= LC_{\epsilon}^{(1)}L^T \\ L &= (SX^{(1)})^+ Sy \end{aligned} \quad 25$$

These estimators minimise the generalised least square index  $(y - X^{(1)}\eta_{LS})^T SS^T (y - X^{(1)}\eta_{LS})$ . This family of estimators are unbiased but not necessarily ML estimates. The Gauss-Markov estimator is the minimum variance and ML estimator that obtains as a special case when  $S = C_{\epsilon}^{(1)-1/2}$ . The T statistic corresponding to the GLS estimator is distributed with  $\nu$  degrees of freedom where (Worsley and Friston 1995)

$$\begin{aligned} T &= \frac{c^T \eta_{LS}}{\sqrt{c^T Cov\{\eta_{LS}\}c}} \\ \nu &= \frac{tr\{RSC_{\epsilon}^{(1)}S\}^2}{tr\{RSC_{\epsilon}^{(1)}SRSC_{\epsilon}^{(1)}S\}} \\ R &= 1 - X^{(1)}L \end{aligned} \quad 26$$

The effective degrees of freedom are based on an approximation due to Satterthwaite (1941). This formulation is formally identical to the non-sphericity correction elaborated by Box (1954) which is commonly known as the Geisser-Greenhouse correction in classical analysis of variance, ANOVA (Geisser and Greenhouse 1958).

The key point here is that EM can be employed to give ReML estimates of correlations among the errors that enter into (26) to enable classical inference, properly adjusted for non-sphericity, *about any GLS estimator*. EM finds a special role in enabling inferences about GLS estimators in statistical parametric mapping. When the relative values of hyperparameters can be assumed to be stationary over

voxels, ReML estimates can be obtained using the sample covariance of the data over voxels, in a single EM (see equation A.7 appendix A.2). After re-normalisation, the ensuing estimate of the non-sphericity  $\Sigma = Q^{(1)} = \sum \lambda_k Q_k^{(1)}$  specifies the serial correlations in terms of a single basis. Voxel-specific hyperparameters can now be estimated in a non-iterative fashion in the usual way, because there is only one hyperparameter to estimate.

#### 4 Application to empirical data

In this subsection we address the variability of serial correlations over voxels within subject and over subjects within the same voxel. Here we are concerned only with the form of the correlations. The next subsection addresses between-subject error variance *per se*.

Using the model specification in (23) serial correlations were estimated using EM in 12 randomly selected voxels from the same slice from a single subject. The results are shown in Figure 7 (left panel) and show that the correlations from one scan to the next can vary between about 0.1 and 0.4. The data sequences and experimental paradigm are described in the figure legend. Briefly these data came from an event-related study of visual word processing in which *new* and *old* words (*i.e.* encoded during a pre-scanning session) were presented in a random order with a stimulus onset asynchrony (SOA) of about 4 seconds. Although the serial correlations within subject vary somewhat there is an even greater variability from subject to subject at the same voxel. The right hand panel of Figure 7 shows the autocorrelation functions estimated separately for 12 subjects at a single voxel. In this instance, the correlations between one scan and the next range from about -0.1 to 0.3 with a greater dispersion relative to the within-subject autocorrelations.

Figure 7 about here

#### 5 Summary

These results are provided to illustrate one potential application of covariance component estimation, not to provide an exhaustive characterisation of serial correlations. This sort of application may be important when it comes to making assumptions about models for serial correlations at different voxels or among

subjects. We have chosen to focus on a covariance estimation problem that requires an iterative parameter re-estimation procedure in which the hyperparameters controlling the covariances depend on the variance of the parameter estimates and *vice versa*. There are other important applications of covariance component estimation we could have considered (although not all require an iterative scheme). One example is the estimation of condition-specific error variances in PET and fMRI. In conventional SPM analyses one generally assumes that the error variance expressed in one condition is the same as that in another. This represents a sphericity assumption over conditions and allows one to pool several conditions when estimating the error variance. Assumptions of this sort, and related sphericity assumptions in multi-subject studies, can be easily addressed in unbalanced designs, or even in the context of missing data, using EM.

## **B Variance component estimation in fMRI: Two-level models**

In this subsection we augment the model above with a second level. This engenders a number of important issues, including the distinction between fixed- and random-effect inferences about the subjects' responses and the opportunity to make Bayesian inferences about single-subject responses. As above, we start with model specification, proceed to simulated data and conclude with an empirical example. In this example the second level represents observations over subjects. Analyses of simulated data are used to illustrate the distinction between fixed- and random-effect inferences by looking at how their respective T values depend on the variance components and design factors. The fMRI data are the same as used above and comprise event-related time-series from 12 subjects. We chose a data set that would be difficult to analyse rigorously using software available routinely. These data not only evidence serial correlations but also the number of trial-specific events varied from subject to subject, giving an unbalanced design.

### *1 Model specification*

The observation model here comprises two levels with the opportunity for subject-specific differences in error variance and serial correlations at the first level and parameter-specific variance at the second. The estimation model here is simply an

extension of that used in the previous subsection to estimate serial correlations. Here it embodies a second level that accommodates observations over subjects.

**level one**

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

$$\begin{bmatrix} y_1 \\ \vdots \\ y_s \end{bmatrix} = \begin{bmatrix} X_1^{(1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & X_s^{(1)} \end{bmatrix} \begin{bmatrix} \theta_1^{(1)} \\ \vdots \\ \theta_s^{(1)} \end{bmatrix} + \varepsilon^{(1)}$$

$$Q_1^{(1)} = \begin{bmatrix} I_t & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}, \dots, Q_s^{(1)} = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & I_t \end{bmatrix}$$

$$Q_{s+1}^{(1)} = \begin{bmatrix} KK^T & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}, \dots, Q_{2s}^{(1)} = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & KK^T \end{bmatrix}$$

**level two**

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

$$X^{(2)} = \mathbf{1}_s \otimes I_p$$

$$Q_1^{(2)} = I_s \otimes \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}, \dots, Q_p^{(2)} = I_s \otimes \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} \quad 27$$

for  $s$  subjects each scanned on  $t$  occasions and  $p$  parameters. The Kronecker tensor product  $A \otimes B$  simply replaces the element of  $A$  with  $A_j B$ . An example of these design matrices and covariance constraints were shown, respectively, in Figures 1 and 3. Note that there are  $2s$  error covariance constraints, one set for the white noise components and one for AR(1) components. Similarly, there are as many prior covariance constraints as there are parameters at the second level.

## 2 Simulations

In the simulations we used 128 scans for each of 12 subjects. The design matrix comprised three effects, modelling an event-related hemodynamic response to frequent but sporadic trials (in fact the instances of correctly identified 'old' words from the empirical example below) and a constant term. Activations were modelled with two regressors, constructed by convolving a series of delta functions with a

canonical hemodynamic response function (HRF)<sup>3</sup> and the same function delayed by three seconds. The delta functions indexed the occurrence of each event. These regressors model event-related responses with two temporal components, which we will refer to as 'early' and 'late' (*c.f.* Henson *et al* 2000). Each subject-specific design matrix therefore comprised three columns giving a total of 36 parameters at the first level and three at the second (the third being a constant term). The HRF basis functions were scaled so that a parameter estimate of one corresponds to a peak response of unity. After division by the grand mean, and multiplication by 100, the units of the response variable and parameter estimates were rendered adimensional and correspond to percent whole brain mean over all scans. The simulated data were generated using (27) with unit Gaussian noise coloured using a temporal, convolution matrix  $(\sum \lambda_k^{(1)} Q_k^{(1)})^{1/2}$  with first-level hyperparameters  $\lambda_j^{(1)} = 0.5$  and  $-0.1$  for each subject's white and AR(1) error covariance components respectively. The second level parameters and hyperparameters were  $\theta^{(2)} = [0.5, 0, 0]^T$ ,  $\lambda^{(2)} = [0.02, 0.006, 0]^T$ . These model substantial early responses with an expected value of 0.5% and a standard deviation over subjects of 0.14% (*i.e.* square root of 0.02). The late component was trivial with zero expectation and a standard deviation of 0.077%. The third or constant terms were discounted with zero mean and variance. These values were chosen because they are typical of real data (see below).

Figures 8 and 9 about here

Figures 8 and 9 show the results after subjecting the simulated data to EM to estimate the conditional mean and covariances of the subject-specific evoked responses. Figure 8 shows the estimated hyperparameters and parameters (black) alongside the true values (white). The first-level hyperparameters controlling within subject error (*i.e.* scan to scan variability) are estimated in a reasonably reliable fashion but note that these estimates show a degree of variation about the veridical values (see Conclusion). In this example the second-level hyperparameters are over-estimated but remarkably good, given only 12 subjects. The parameter estimates at the first and second levels are again very reasonable, correctly attributing the majority of the

---

<sup>3</sup> The canonical HRF was the same as that employed by SPM. It comprises a mixture of two gamma variates modelling peak and undershoot components and is based on a principal component analysis of

experimental variance to an early effect. Figure 8 should be compared with Figure 10 that shows the equivalent estimates for real data.

The top panel in Figure 9 shows the ML estimates that would have been obtained if we had used a single-level model. These correspond to response estimates from a conventional fixed-effects analysis. The insert shows the classical fixed-effect T values, for each subject, for contrasts testing early and late response components. Although these T values properly reflect the prominence of early effects, their variability precludes any threshold that could render the early components significant and yet exclude false positives pertaining to the late component. The lower panel highlights the potential of revisiting the first level, in the context of a hierarchical model. It shows the equivalent responses based on the conditional mean and the posterior inference (insert) based on the conditional covariance. This allows us to reiterate some points made in the previous section. Firstly, the parameter estimates and ensuing response estimates are informed by information abstracted from higher levels. Second, this prior information enables Bayesian inference about the probability of an activation that is specified in neurobiologically meaningful terms.

In Figure 9 the estimated responses are shown (solid lines) with the actual responses (broken lines). Note how the conditional estimates show a regression or ‘shrinkage’ to the conditional mean. In other words, their variance shrinks to reflect, more accurately, the variability in real responses. In particular the spurious variability in the apparent latency of the peak response in the ML estimates disappears when using the conditional estimates. This is because the contribution of the late component, that induces latency differences, is suppressed in the conditional estimates. This, in turn, reflects the fact that the variability in its expression over subjects is small relative to that induced by the observation error. Simulations like these suggest that characterisations of inter-subject variability using ML approaches can severely overestimate the true variability. This is because the ML estimates are unconstrained and simply minimise observation error without considering how likely the ensuing inter-subject variability is.

The posterior probabilities (insert) are a function of the conditional mean  $\eta_{\theta|y}^{(1)}$  and covariance  $C_{\theta|y}^{(1)}$  and a size threshold  $\gamma = 0.1$  that specifies an ‘activation’.

$$1 - \Phi \left( \frac{\gamma - c_j^T \eta_{\theta|y}^{(1)}}{\sqrt{c_j^T C_{\theta|y}^{(1)} c_j}} \right) \quad 28$$

The contrast weight vectors were  $c_{early} = [1, 0, 0]^T$  and  $c_{late} = [0, 1, 0]^T$ . As expected, the probability of the early response being greater than  $\gamma$  was uniformly high for all 12 subjects, whereas the equivalent probability for the late component was negligible. Note that, in contradistinction to the classical inference, there is now a clear indication that each subject expressed an early response but no late response.

### 3 Empirical analyses

Here the analysis is repeated using real data and the results are compared to those obtained using simulated data. The empirical data are described in Henson *et al* (2000). Briefly, they comprised 128+ scans in 12 subjects. Only the first 128 scans were used below. The experimental design was stochastic and event-related, looking for differential responses evoked by *new* relative to *old* (studied prior to the scanning session) words. Either a new or old word was presented every 4 seconds or so (SOA varied between 2.5 and 5.5 seconds). In this design one is interested only in the differences between evoked responses to the two stimulus types. This is because the efficiency of the design to detect the effect of stimuli *per se* is negligible with such a short SOA. Subjects were required to make an old *vs.* new judgement for each word. Drift (the first 8 components of a discrete cosine set) and the effects of incorrect trials were treated as confounds and removed using linear regression<sup>4</sup>. The first-level subject-specific design matrix partitions comprised four regressors with early and late effects for both old and new words.

The analyses proceeded in exactly the same way as for the simulated data. The only difference was that the contrast tested for *differences* between the two word types (*i.e.*  $c = [1, 0, -1, 0]^T$  for an old minus new early effect). The hyperparameter and parameter estimates, for a voxel in the cingulate gyrus (BA 31; -3, -33, 39mm), are shown in Figure 10, adopting the same format as in Figure 8. Here we see that the within-

---

<sup>4</sup> Strictly speaking the projection matrix implementing this adjustment should also be applied to the covariance constraints but this would (i) render the constraints singular and (ii) ruin their sparsity structure. We therefore omitted this and ensured, in simulations, that the adjustment had a negligible effect on the hyperparameter estimates.



subject error varies much more in the empirical data with the last subject showing almost twice the error variance of the first subject. As above, the serial correlations vary considerably from subject to subject and are not consistently positive or negative. The second-level hyperparameters showed the early component of the differential response to be more reliable over subjects than the late component (0.007 and 0.19 respectively). All but two subjects had a greater early response, relative to late, which on average was about 0.28%. In other words, activation differentials, in the order of 0.3%, occurred in the context of an observation error with a standard deviation of 0.5% (see Figure 10). The inter-subject variability was about 30% of the mean response amplitude. A component of the variability in within-subject error is due to uncertainty in the ReML estimates of the hyperparameters (see below) but this degree of inhomogeneity is substantially more than in the simulated data (where subjects had equal error variances). It is interesting to note that, despite the fact that the regressors for the early and late components had exactly the same form, the between-subject error for one was less than half that of the other. Results of this sort speak to the prevalence of non-sphericity (in this instance heteroscedasticity or unequal variances) and a role for the analyses illustrated here.

Figures 10 and 11 about here

The response estimation and inference are shown in Figure 11. Again we see the characteristic 'shrinkage' when comparing the ML to the conditional estimates. It can be seen that all subjects, apart from the first and third, had over a 95% chance of expressing an early differential of 0.1% or more. The late differential response was much less consistent, although one subject expressed a difference with about 84% confidence.

### **C Summary**

The examples presented above allow us to reprise a number of important points made in the previous section (see also Friston *et al* 2002a). In conclusion the main points are:

- There are many instances when an iterative parameter re-estimation scheme is required (*e.g.* dealing with serial correlations or missing data). These schemes are generally variants of EM.
- Even before considering the central role of covariance component estimation in hierarchical or empirical Bayes models it is an important aspect of model estimation in its own right, particularly in estimating non-sphericity among observation errors. Parameter estimates can either be obtained directly from an EM algorithm, in which case they correspond to the ML or Gauss-Markov estimates, or the hyperparameters can be used to determine the error correlations which re-enter a generalised least square scheme, as a non-sphericity correction.
- Hierarchical models enable a collective improvement in response estimates by using conditional, as opposed to maximum-likelihood, estimators. This improvement ensues from the constraints derived from higher levels that enter as priors at lower levels.

In the next section we revisit two-level models but consider hierarchical observations over voxels as opposed to subjects.

## **IV POSTERIOR PROBABILITY MAPPING AND PPMs**

### **A Introduction**

This section describes the construction of *posterior probability maps* that enable conditional or Bayesian inferences about regionally-specific effects in neuroimaging. Posterior probability maps are images of the probability or confidence that an activation exceeds some specified threshold, given the data. Posterior probability maps (PPMs) represent a complementary alternative to statistical parametric maps (SPMs) that are used to make classical inferences. However, a key problem in Bayesian inference is the specification of appropriate priors. This problem can be finessed using *empirical Bayes* in which prior variances are estimated from the data, under some simple assumptions about their form. Empirical Bayes requires a

hierarchical observation model, in which higher levels can be regarded as providing prior constraints on lower levels. In neuroimaging, observations of the same effect over voxels provide a natural, two-level hierarchy that enables an empirical Bayesian approach. In this section we present the motivation and the operational details of a simple empirical Bayesian method for computing posterior probability maps. We then compare Bayesian and classical inference through the equivalent PPMs and SPMs testing for the same effect in the same data.

To date, inference in neuroimaging has been restricted largely to classical inferences based upon statistical parametric maps (SPMs). The alternative approach is to use Bayesian or conditional inference based upon the posterior distribution of the activation given the data (Holmes & Ford 1993). This necessitates the specification of priors (*i.e.* the probability distribution of the activation). Bayesian inference requires the posterior distribution and therefore rests upon a posterior density analysis. A useful way to summarise this posterior density is to compute the probability that the activation exceeds some threshold. This computation represents a Bayesian inference about the effect, in relation to the specified threshold. We now describe an approach to computing posterior probability maps for activation effects, or more generally treatment effects in imaging data sequences. This approach represents, probably, the most simple and computationally expedient way of constructing PPMs.

As established in the previous sections, the motivation for using conditional or Bayesian inference is that it has high face validity. This is because the inference is about an effect, or activation, being greater than some specified size that has some meaning in relation to underlying neurophysiology. This contrasts with classical inference, in which the inference is about the effect being significantly different than zero. The problem for classical inference is that trivial departures from the null hypothesis can be declared significant, with sufficient data or sensitivity. Furthermore, from the point of view of neuroimaging, posterior inference is especially useful because it eschews the multiple-comparison problem. Posterior inference does not have to contend with the multiple-comparison problem because there are no false-positives. The probability that activation has occurred, given the data, at any particular voxel is the same, irrespective of whether one has analysed that voxel or the entire brain. For this reason, posterior inference using PPMs may represent a relatively more powerful approach than classical inference in neuroimaging. The reason that there is no need to adjust the p-values is that we assume independent

prior distributions for the activations over voxels. In this simple Bayesian model the Bayesian perspective is similar to that of the frequentist who makes inferences on a per-comparison basis (see Berry and Hochberg 1999 for a detailed discussion).

### *1 Priors and Bayesian inference*

PPMs require the posterior distribution or conditional distribution of the activation (a contrast of conditional parameter estimates) given the data. This posterior density can be computed, under Gaussian assumptions, using Bayes rules. Bayes rule requires the specification of a likelihood function and the prior density of the model's parameters. The models used to form PPMs, and the likelihood functions, are exactly the same as in classical SPM analyses. The only extra bit of information that is required is the prior probability distribution of the parameters of the general linear model employed. Although it would be possible to specify these in terms of their means and variances using independent data, or some plausible physiological constraints, there is an alternative to this fully Bayesian approach. The alternative is empirical Bayes in which the variances of the prior distributions are estimated directly from the data. Empirical Bayes requires a *hierarchical observation model* where the parameters and hyper-parameters at any particular level can be treated as priors on the level below. There are numerous examples of hierarchical observation models. For example, the distinction between fixed- and mixed-effects analyses of multi-subject studies relies upon a two-level hierarchical model. However, in neuroimaging there is a natural hierarchical observation model that is common to all brain mapping experiments. This is the hierarchy induced by looking for the same effects at every voxel within the brain (or grey matter). The first level of the hierarchy corresponds to the experimental effects at any particular voxel and the second level of the hierarchy comprises the effects over voxels. Put simply, the variation in a particular contrast, over voxels, can be used as the prior variance of that contrast at any particular voxel.

. The model used here is one in which the spatial relationship among voxels is discounted. The advantage of treating an image like a 'gas' of unconnected voxels is that the estimation of between-voxel variance in activation can be finessed to a considerable degree (see Eq. A.7 in the Appendix and following discussion). This renders the estimation of posterior densities tractable because the between-voxel variance can then be used as a prior variance at each voxel. We therefore focus on this simple and special case and on the 'pooling' of voxels to give precise [ReML]

estimates of the variance components required for Bayesian inference. The main focus of this section is the pooling procedure that affords a computational saving necessary to produce PPMs of the whole brain. In what follows we describe how this approach is implemented and provide some examples of its application.

## B Theory

### 1 Conditional estimators and the posterior density

In this subsection we describe how the posterior distribution of the parameters of any general linear model can be estimated at each voxel from imaging data sequences. Under Gaussian assumptions about the errors  $\varepsilon \sim N\{0, C_\varepsilon\}$  of a general linear model with design matrix  $X$  the responses are modelled as

$$y = X\theta + \varepsilon \quad 29$$

The conditional or posterior covariances and mean of the parameters  $\theta$  are given by (Friston *et al* 2002a).

$$\begin{aligned} C_{\theta|y} &= (X^T C_\varepsilon^{-1} X + C_\theta^{-1})^{-1} \\ \eta_{\theta|y} &= C_{\theta|y} X^T C_\varepsilon^{-1} y \end{aligned} \quad 30$$

where  $C_\theta$  is the prior covariance and assuming a prior expectation of zero. Once these moments are known, the posterior probability that a particular effect or contrast specified by a contrast weight vector  $c$  exceeds some threshold  $\gamma$  is easily computed

$$p = 1 - \Phi \left( \frac{\gamma - c^T \eta_{\theta|y}}{\sqrt{c^T C_{\theta|y} c}} \right) \quad 31$$

$\Phi(\cdot)$  is the cumulative density function of the unit normal distribution. An image of these posterior probabilities constitutes a PPM.

## 2 Estimating the error covariance with ReML

Clearly, to compute the conditional moments in (30) one needs to know the error and prior covariances  $C_\varepsilon$  and  $C_\theta$ . In the next section we will describe how the prior covariance  $C_\theta$  can be estimated. For the moment, assume the prior covariance is known. In this case the error covariance can be estimated in terms of a hyperparameter  $\lambda_\varepsilon$  where  $C_\varepsilon = \lambda_\varepsilon V$ , and  $V$  is the correlation or non-sphericity matrix of the errors (see below). This hyperparameter is estimated simply using Restricted Maximum Likelihood (ReML) as described in the appendix<sup>5</sup>.

Until convergence { **E-Step**

$$C_\varepsilon = \lambda_\varepsilon V$$

$$C_{\theta|y} = (X^T C_\varepsilon^{-1} X + C_\theta^{-1})^{-1}$$

**M-Step**

$$P = C_\varepsilon^{-1} - C_\varepsilon^{-1} X C_{\theta|y} X^T C_\varepsilon^{-1}$$

$$g = -\frac{1}{2} \text{tr}\{PV\} + \frac{1}{2} \text{tr}\{P^T V P y y^T\}$$

$$H = \frac{1}{2} \text{tr}\{PVPV\}$$

$$\lambda_\varepsilon \leftarrow \lambda_\varepsilon + H^{-1} g$$

}

32

In brief,  $P$  represents the residual forming matrix, pre-multiplied by the inverse of the error covariance. It is this projector matrix that ‘restricts’ the estimation of variance components to the null space of the design matrix.  $g$  and  $H$  are the first- and expected second-order derivatives (*i.e.* gradients and expected negative curvature) of the ReML objective function. The **M-Step** can be regarded as a Fisher Scoring scheme that maximises the ReML objective function. Given that there is only one hyperparameter to estimate this scheme converges very quickly (2 to 3 iterations for a tolerance of  $10^{-6}$ ).

### 3 Estimating the prior density with empirical Bayes

Simply computing the conditional moments using (30) corresponds to a fully Bayesian analysis at each and every voxel. However, there is an outstanding problem in the sense that we do not know the prior covariances of the parameters. It is at this point that we introduce the hierarchical perspective that enables an empirical Bayesian approach. If we now consider (29) as the first level of the two-level hierarchy, where the second level corresponds to observations over voxels, we have a hierarchical observation model for all voxels that treats some parameters as random effects and others as fixed. The random effects  $\theta_1$  are those that we are interested in and the fixed effects  $\theta_0$  are nuisance variables or confounds (*e.g.* drifts or the constant term) modelled by the regressors in  $X_0$  where  $X = [X_1, X_0]$  and

$$y = [X_1, X_0] \begin{bmatrix} \theta_1 \\ \theta_0 \end{bmatrix} + \varepsilon^{(1)} \quad 33$$

$$\theta_1 = 0 + \varepsilon^{(2)}$$

This model posits that there is a voxel-wide prior distribution for the parameters  $\theta_1$  with zero mean and unknown covariance  $E\{\varepsilon^{(2)} \varepsilon^{(2)T}\} = \sum_i \lambda_i Q_i$ . The bases  $Q_i$  specify the prior covariance structure of the interesting effects and would usually comprise a basis for each parameter whose  $i$ th leading diagonal element was one and zero elsewhere. This implies that if we selected a voxel at random from the search volume, the  $i$ th parameter at that voxel would conform to a sample from a Gaussian distribution of zero expectation and variance  $\lambda_i$ . The reason this distribution can be assumed to have zero mean is that parameters of interest reflect region-specific effects that, by definition sum to zero over the search volume<sup>6</sup>. By concatenating the data from all voxels and using Kronecker tensor products of the design matrices and covariance bases, it is possible to create a very large hierarchical observation model

---

<sup>5</sup> Note that the augmentation step shown in Figure 4 is unnecessary because the prior covariance enters explicitly into the conditional covariance.

<sup>6</sup> In the SPM2 implementation we allow for any mean of the parameters at the second level by subtracting the mean over voxels from the data. This mean represents an estimate of the prior expectation projected onto the observation space by the design matrix.

that could be subject to EM (see for example Friston *et al* 2002b, Section 3.2). However, given the enormous number of voxels in neuroimaging this is, computationally, prohibitive. A mathematically equivalent but more tractable approach is to consider the estimation of the prior hyperparameters as a variance component estimation problem after collapsing (33) to a single-level model

$$\begin{aligned} y &= X_0 \theta_0 + \xi \\ \xi &= X_1 \boldsymbol{\varepsilon}^{(2)} + \boldsymbol{\varepsilon}^{(1)} \end{aligned} \quad 34$$

This is simply a rearrangement of (33) to give a linear model with a compound error covariance that includes the observation error covariance and  $m$  components for each parameter in  $\theta_1$ . These components are induced by variation of the parameters over voxels.

$$\begin{aligned} C_\xi &= E\{\xi\xi^T\} = \sum \lambda_k Q_k \\ Q &= \{X_1 Q_1 X_1^T, \dots, X_1 Q_m X_1^T, V\} \\ \lambda &= [\lambda_1, \dots, \lambda_m, \lambda_\varepsilon]^T \end{aligned} \quad 35$$

This equation says that the covariance of the compound error can be linearly decomposed into  $m$  components (usually one for each parameter) and the error variance. The form of the observed covariances, due to variation in the parameters, is determined by the design matrix  $X$  and  $Q_i$  that model variance components in parameter space.

Equation (35) affords a computationally expedient way to estimate the prior covariances for the parameters that then enter into (30) to provide for voxel-specific error hyperparameter estimates and conditional moments. In brief, the hyperparameters are estimated by pooling the data from all voxels to provide ReML estimates of the variance components of  $C_\xi$  according to (35). The nice thing about this pooling is that the hyperparameters of the parameter covariances are, of course, the same for all voxels. This is not the case for the error covariance hyperparameters that may change from voxel to voxel. The pooled estimate of  $\lambda_\varepsilon$  can be treated as an



estimate of the average  $\lambda_\varepsilon$  over voxels. The hyperparameters are estimated by iterating

*Until convergence* { **E-Step**

$$C_\xi = \sum \lambda_k Q_k$$

$$C_{\theta_0|y} = (X_0^T C_\varepsilon^{-1} X_0)^{-1}$$

**M-Step**

36

$$P = C_\xi^{-1} - C_\xi^{-1} X_0 C_{\theta_0|y} X_0^T C_\xi^{-1}$$

$$g_i = -\frac{1}{2} \text{tr}\{PQ_i\} + \frac{1}{2} \text{tr}\{P^T Q_i P Y Y^T / n\}$$

$$H_{ij} = \frac{1}{2} \text{tr}\{PQ_i P Q_j\}$$

$$\lambda \leftarrow \lambda + H^{-1} g$$

}

It can be seen that this has exactly the form as (32) used for the analysis at each voxel. The differences are (i)  $yy^T$  has been replaced by its sample mean over voxels  $YY^T/n$  and (ii) there are no priors because the parameters controlling the expression of confounding effects or nuisance variables are treated as fixed effects. This is equivalent to setting their prior variance to infinity (*i.e.* flat priors) so that  $C_{\theta_0}^{-1} \rightarrow 0$ . Finally, (iii) the regressors in  $X_1$  have disappeared from the design matrix because these effects are embodied in the covariance components of the compound error. As above, the inclusion of confounds restricts the hyperparameter estimation to the null space of  $X_0$ , hence *restricted* maximum likelihood (ReML). In the absence of confounds the hyperparameters would simply be maximum likelihood (ML) estimates that minimise the difference between the estimated and observed covariance of the data, averaged over voxels. The ensuing ReML estimates are very high precision estimators. Their precision increases linearly with the number of voxels  $n$  and is in fact equal to  $nH$ . These hyperparameters now enter as priors into the voxel-specific estimation along with the flat priors for the nuisance variables

$$\mathbf{C}_\theta = \begin{bmatrix} \sum \lambda_i \mathcal{Q}_i & \cdots & 0 \\ \vdots & \infty & \\ 0 & & \ddots \\ & & & \infty \end{bmatrix} \quad 37$$

We now have a very precise estimate of the prior covariance that can be used to re-visit each voxel to compute the conditional or posterior density using equations (30) and (32). Finally, the conditional moments enter Eq. (31) to give the posterior probability for each voxel. See Figure 12 for a schematic illustration of this scheme.

#### 4 Summary

A natural hierarchy characterises all neuroimaging experiments, where the second level is provided by variation over voxels. Although it would be possible to form a very large two-level observation model and estimate the conditional means and covariances of the parameters at the first level this would involve dealing with matrices of size  $(ns) \times (ns)$  (number of voxels  $n$  times the number of scans  $s$ ). The same conditional estimators can be computed using the two-step approach described above. First, the data covariance components induced by parameter variation over voxels and observation error are computed using ReML estimates of the associated covariance hyperparameters. Second, each voxel is revisited to compute voxel-specific error variance hyperparameters and the conditional moments of the parameters, using the empirical priors from the first step (see Figure 12). Both these steps deal only with matrices of size  $n \times n$ . The voxel-specific estimation sacrifices the simplicity of a single large iterative scheme for lots of quicker iterative schemes at each voxel. This exploits the fact that the same first-level design matrix is employed for all voxels.

### C Empirical demonstration

In this section we compare and contrast Bayesian and classical inference using PPMs and SPMs based on real data. The first data is the PET verbal fluency data that has been used to illustrate methodological advances in SPM over the years. In brief, these data were required from five subjects each scanned 12 times during the performance of one of two word generation tasks. The subjects were asked to either repeat a heard

letter or to respond with a word that began with the heard letter. These tasks were performed in alternation over the 12 scans and the order randomised over subjects. The second data set comprised data from a study of attention to visual motion (Büchel & Friston 1997). The data used in this note came from the first subject studied. This subject was scanned at 2T to give a time series of 360 images comprising 10 block epochs of different visual motion conditions. These conditions included a fixation condition, visual presentation of static dots, visual presentation of radially moving dots under attention and no-attention conditions. In the attention condition subjects were asked to attend to changes in speed (which did not actually occur). These data were re-analysed using a conventional SPM procedure and using the empirical Bayesian approach described in the previous section. The ensuing SPMs and PPMs are presented below for the PET and fMRI data respectively. The contrast for the PET data compared the word generation with the word shadowing condition and the contrast for the fMRI data tested for the effect of visual motion above and beyond that due to photic stimulation with stationary dots.

### *1 Inference for the PET data*

The upper panel of Figure 13 shows the PPM for a deactivating effect of verbal fluency. There are two thresholds for the PPM. The first and more important is  $\gamma$  in equation 3. This defines what we mean by “activation” and, by default, is set at one deviation of the prior variance of the contrast, in this instance 2.2. This corresponds to a change in rCBF of 2.2 adimensional units (equivalent to ml/dl/min). The second threshold is more trivial and simply enables the use of maximum intensity projections. This is the probability the voxel has to exceed in order to be displayed. In the PPM shown this was set at 95%. This means that all voxels shown have greater than 95% probability of being deactivated by 2.2 or more. The PPM can be regarded as a way of summarising ones confidence that an effect is present (*c.f.* the use of confidence intervals where the lower bound on the interval is set at  $\gamma$ ). It should be noted that posterior inference would normally require the reporting of the conditional probability whether it exceeded some arbitrary threshold or not. However, for the visual display of posterior probability maps it is useful to remove voxels that fall below some threshold.

Figure 14 provides a quantitative representation of Bayesian inference afforded by PPMs. In the upper panel the posterior expectation for the twelve condition-specific effects are shown, encompassed by the 95% confidence intervals (bars) based on the posterior covariance. It can be seen that in the fifth condition (the third word shadowing condition) one could be almost certain the activation is greater than zero. The prior and posterior densities for this activation are shown in the lower panel. These are the probability distributions before and after observing the data. Note that the posterior variance is always smaller than the prior variance, depending on how noisy the data is.

The corresponding SPM is shown in the lower panel (Figure 13b). The SPM has been thresholded at 0.05 adjusted for the search volume using a Gaussian field correction. There is a remarkable correspondence between the activation profiles inferred by the PPM and the SPM. The similarity between the PPM and the SPM for these data should not be taken as characteristic. The key difference between Bayesian inference, based on the confidence we have about activation, and classical inference, based on rejecting the null hypothesis, is that the latter depends on the search volume. The classical approach, when applied in a mass univariate setting (*i.e.* over a family of voxels) induces a multiple comparison problem that calls for a procedure to control for family-wise false positives. In the context of imaging data this procedure is a Gaussian field adjustment to the threshold. This adjustment depends on the search volume. The consequence is that if we increased the search volume the threshold would rise and some of the voxels seen in the SPM would disappear. Because the PPM does not label any voxel as ‘activated’, there is no multiple comparison problem and the 95% confidence threshold is the same irrespective of search volume. This difference between PPMs and SPMs is highlighted in the analysis of the fMRI data. Here, the search volume is increased by reducing the smoothness of the data. We do this by switching from PET to fMRI. Smoothness controls the ‘statistical’ search volume, which is generally much greater for fMRI than for PET.

## *2 Inference for the fMRI data*

The difference between the PPM and SPM for the fMRI analysis is immediately apparent on inspection of Figures 15 and 16. Here the default threshold for the PPM was 0.7% (equivalent to percentage whole brain mean signal). Again only voxels that exceed 95% confidence are shown. These are restricted to visual and extrastriate

cortex involved in motion processing. The critical thing to note here is that the corresponding SPM identifies a smaller number of voxels than the PPM. Indeed the SPM appears to have missed a critical and bilaterally represented part of the V5 complex (circled cluster on the PPM in the lower panel of Figure 15). The SPM is more conservative because the correction for multiple comparisons in these data is very severe, rendering classical inference relatively insensitive. It is interesting to note that dynamic motion in the visual field has such widespread (if small) effects at a hemodynamic level.

### *3 PPMs and FDR*

There is an interesting connection between false discovery rate (FDR) control and thresholded PPMs. Subjecting PPMs to a 95% threshold means that surviving voxels have, at most, a 5% probability of not exceeding the default threshold  $\gamma$ . In other words, if we declared these voxels as “activated”, 5% of the voxels could be false activations. This is exactly the same as FDR in the sense that the FDR is the proportion of voxels that are declared significant but are not. It should be noted that many voxels will have a posterior probability that is more than 95%. Therefore, the 5% is an upper bound on the FDR. This interpretation rests explicitly on thresholding the PPM and labelling the excursion set as “activated”. It is reiterated that this declaration is unnecessary and only has any meaning in relation to classical inference. However, thresholded PPMs do have this interesting connection to SPMs in which false discovery rate has been controlled.

## **D Conclusion**

In this section we looked at a simple way to construct posterior probability maps using empirical Bayes. Empirical Bayes can be used because of the natural hierarchy in neuroimaging engendered by looking for the same thing over multiple voxels. The approach provides simple shrinkage priors based on between-voxel variation in parameters controlling effects of interest. A computationally expedient way of computing these priors using ReML has been presented that pools over voxels. This pooling device offers an enormous computational saving through simplifying the matrix algebra and enabling the construction of whole-brain PPMs. The same device has found an interesting application in the ReML estimation of prior variance

components in space, by pooling over time bins, in the EEG source reconstruction problem (Phillips *et al* 2003 - submitted).

A key consideration, in the use of empirical Bayes in this setting is “which voxels to include in the hierarchy?” There is no right or wrong answer here (*c.f.* the search volume in classical inference with SPMs). The most important thing to bear in mind is that the conditional estimators of an activation or effect are those which minimise some cost function. This cost function can be regarded as the ability to predict the observed response with minimum error, on average, over the voxels included in the hierarchical model. In other words, the voxels over which the priors are computed define the space one wants, on average, the best estimates for. In this work we have simply used potentially responsive voxels within the brain as defined by thresholding the original images (to exclude extra-cranial regions).

In the next section we turn to Bayesian inferences based on Full Bayes where the priors come, not from empirical estimates based on hierarchical observations over voxels, but from biophysical parameters mediating the response at a single voxel.

## V BAYESIAN IDENTIFICATION OF DYNAMIC SYSTEMS

### A Introduction

This section presents a method for estimating the conditional or posterior distribution of the parameters of deterministic dynamical systems. The procedure conforms to an EM search for the maximum of the conditional or posterior density. The inclusion of priors in the estimation procedure ensures robust and rapid convergence and the resulting conditional densities enable Bayesian inference about the model parameters. The method is demonstrated using an input-state-output model of the hemodynamic coupling between experimentally designed causes or factors in fMRI studies and the ensuing BOLD response (see **Chapter 11: Hemodynamic modelling**). This example represents a generalisation of current fMRI analysis models that accommodates nonlinearities and in which the parameters have an explicit physical interpretation.

This section is about the identification of deterministic nonlinear dynamical models. Deterministic here refers to models where the dynamics are completely determined by the state of the system. Random or stochastic effects enter only at the point that the

system's outputs or responses are observed<sup>7</sup>. We will focus on a particular model of how changes in neuronal activity translate into hemodynamic responses. By considering a voxel as an *input-state-output* system one can model the effects of an input (*i.e.* stimulus function) on some state variables (*e.g.* flow, volume, deoxyhemoglobin content *etc.*) and the ensuing output (*i.e.* BOLD response). The scheme adopted here uses Bayesian estimation, where the aim is to identify the posterior or conditional distribution of the parameters, given the data. Knowing the posterior distribution allows one to characterise an observed system in terms of the parameters that maximise their posterior probability (*i.e.* those parameters that are most likely given the data) or indeed, make inferences about whether the parameters are bigger or smaller than some specified value.

By demonstrating the approach using hemodynamic models, we can establish the notion that biophysical and physiological models of evoked brain responses can be used to make Bayesian inferences about experimentally induced regionally specific activations. Including parameters that couple experimentally changing stimulus or task conditions to the system's states enables this inference. The posterior or conditional distribution of these parameters can then be used to make inferences about the efficacy of experimental inputs in eliciting measured responses. Because the parameters we want to make an inference about have an explicit physical interpretation, in the context of the hemodynamic model used, the face validity of the ensuing inference is more grounded in physiology. Furthermore, because the 'activation' is parameterised in terms of processes that have natural biological constraints, these constraints can be used as priors in a Bayesian scheme.

Previous sections have focussed on *empirical* Bayesian approaches in which the priors were derived from the data being analysed. In this section we use a *fully* Bayesian approach, where the priors are assumed to be known and apply it to the hemodynamic model described in Friston *et al* (2000) and **Chapter 11 (Hemodynamic Modelling)**. In Friston *et al* (2000) we presented a hemodynamic model that embedded the Balloon/Windkessel (Buxton *et al* 1998, Mandeville *et al* 1999) model of flow to BOLD coupling to give a complete dynamical model of how neurally mediated signals cause a BOLD response. In this work we restricted ourselves to single input-single output (SISO) systems by considering only one input.

---

<sup>7</sup> There is another important class of models where stochastic processes enter at the level of the state variables themselves (*i.e.* deterministic noise). These are referred to as stochastic dynamical models.

Here we demonstrate a general approach to nonlinear system identification using an extension of these SISO models to multiple input-single output (MISO) systems. This allows for a response to be caused by multiple experimental effects and we can assign a causal efficacy to any number of explanatory variables (*i.e.* stimulus functions). Later (**Chapter 22: Dynamic Causal Modelling**) we will generalise to multiple input-multiple output systems (MIMO) such that interactions among brain regions, at a neuronal level can be addressed.

An important aspect of the proposed model is that it can be reduced, exactly, to the model used in classical SPM-like analyses, where one uses the stimulus functions, convolved with a canonical hemodynamic response function, as explanatory variables in a general linear model. This classical analysis is a special case, that obtains when the model parameters of interest (the efficacy of a stimulus) are treated as fixed effects with flat priors and the remaining biophysical parameters enter as known canonical values with infinitely small prior variance (*i.e.* high precision). In this sense the current approach can be viewed as a Bayesian generalisation of that normally employed. The advantages of this generalisation rest upon (i) the use of a nonlinear observation model and (ii) Bayesian estimation of that model's parameters. The fundamental advantage, of a nonlinear MISO model over linear models, is that only the parameters linking the various inputs to hemodynamics are input- or trial-specific. The remaining parameters, pertaining to the hemodynamics *per se*, are the same for each voxel. In conventional analyses the hemodynamic response function, for each input, is estimated in a linearly separable fashion (usually in terms of a small set of temporal basis functions) despite the fact that the form of the impulse response function to each input is likely to be the same. In other words, a nonlinear model properly accommodates the fact that many of the parameters shaping input-specific hemodynamic responses are shared by all inputs. For example, the components of a compound trial (*e.g.* cue and target stimuli) might not interact at a neuronal level but may show sub-additive effects in the measured response, due to nonlinear hemodynamic saturation. In contradistinction to conventional linear analyses the analysis proposed in this section could, in principle, disambiguate between interactions at the neuronal and hemodynamic levels. The second advantage is that Bayesian inferences about input-specific parameters can be framed in terms of whether the efficacy for a particular cause exceeded some specified threshold or, indeed the probability that it was less than some threshold (*i.e.* infer that a voxel did



*not* respond). The latter is precluded in classical inference. These advantages should be weighed against the difficulties of establishing a valid model and the computational expense of identification.

### *1 Overview*

This section is divided into four parts. In the first we reprise briefly the hemodynamic model and motivate the four differential equations that it comprises. We will touch on the Volterra formulation of nonlinear systems to show the output can always be represented as a nonlinear function of the input and the model parameters. This nonlinear function is used as the basis of the observation model that is subject to Bayesian identification. This identification requires priors which, here, come from the distribution, over voxels, of parameters estimated in Friston *et al* (2000). The second part describes these priors and how they were determined. Having specified the form of the nonlinear observation model and the prior densities on the model's parameters, the third section describes the estimation of their posterior densities. The ensuing scheme can be regarded as a Gauss-Newton search for the maximum posterior probability (as opposed to the maximum likelihood as in conventional applications) that embeds the EM scheme in Appendix A.1. This description concludes with a note on integration, required to evaluate the local gradients of the objective function. This, effectively generalises the EM algorithm for linear systems so that it can be applied to nonlinear models.

Finally we demonstrate the approach using empirical data. First, we revisit the same data used to construct the priors using a single input. We then apply the technique to the same study of visual attention used in the previous section, to make inferences about the relative efficacy of multiple experimental effects in eliciting a BOLD response.

## **B The Hemodynamic Model**

The hemodynamic model considered here was presented in detail in Friston *et al* (2000). Although relatively simple it is predicated on a substantial amount of previous careful theoretical work and empirical validation [*e.g.* Buxton *et al* (1998), Mandeville *et al* 1999, Hoge *et al* 1999, Mayhew *et al* 1998]. The model is a SISO system with a stimulus function as input (that is supposed to elicit a neuronally mediated flow-inducing signal) and BOLD response as output. The model has six

parameters and four state variables each with its corresponding differential equation. The differential or state equations express how each state variable changes over time as a function of the others. These state equations and the output nonlinearly (a static nonlinear function of the state variables that gives the output) specify the form of the model. The parameters determine any specific realisation of the model. In what follows we review the state equations, the output nonlinearity, extension to a MISO system and the Volterra representation.

### 1 The state equations

Assuming that the dynamical system linking synaptic activity and rCBF is linear (Miller *et al* 2000) we start with

$$\dot{f}_{in} = s \quad 38$$

where  $f_{in}$  is inflow and  $s$  is some flow inducing signal. The signal is assumed to subsume many neurogenic and diffusive signal sub-components and is generated by neuronal responses to the input (the stimulus function)  $u(t)$

$$\dot{s} = \varepsilon u(t) - \kappa_s s - \kappa_f (f_{in} - 1) \quad 39$$

$\varepsilon$ ,  $\kappa_s$  and  $\kappa_f$  are parameters that represent the efficacy with which input causes an increase in signal, the rate-constant for signal decay or elimination and the rate-constant for auto-regulatory feedback from blood flow. The existence of this feedback term can be inferred from; (i) post-stimulus undershoots in rCBF and (ii) the well-characterised vasomotor signal in optical imaging (Mayhew *et al* 1998). Inflow determines the rate of change of volume through

$$\begin{aligned} \dot{v} &= f_{in} - f_{out}(v) \\ f_{out}(v) &= v^{1/\alpha} \end{aligned} \quad 40$$

This says that normalised venous volume changes reflect the difference between inflow  $f_{in}$  and outflow  $f_{out}$  from the venous compartment with a time constant

(transit time)  $\tau$ . Outflow is a function of volume that models the balloon-like capacity of the venous compartment to expel blood at a greater rate when distended (Buxton *et al* 1998). It can be modelled with a single parameter (Grubb *et al* 1974)  $\alpha$  based on the Windkessel model (Mandeville *et al* 1999). The change in normalised total deoxyhemoglobin voxel content  $\dot{q}$  reflects the delivery of deoxyhemoglobin into the venous compartment minus that expelled (outflow times concentration)

$$\tau\dot{q} = f_{in} \frac{E(f_{in}, E_0)}{E_0} - f_{out}(v)q/v \quad 41$$

$$E(f_{in}, E_0) = 1 - (1 - E_0)^{1/f_{in}}$$

where  $E(f_{in}, E_0)$  is the fraction of oxygen extracted from inflowing blood. This is assumed to depend on oxygen delivery and is consequently flow-dependent. This concludes the state equations, where there are six unknown parameters namely efficacy  $\varepsilon$ , signal decay  $\kappa_s$ , auto-regulation  $\kappa_f$ , transit time  $\tau$ , Grubb's exponent  $\alpha$  and resting net oxygen extraction by the capillary bed  $E_0$ .

## 2 The Output nonlinearity

The BOLD signal  $y(t) = \lambda(v, q, E_0)$  is taken to be a static nonlinear function of volume ( $v$ ), and deoxyhemoglobin content ( $q$ )

$$y(t) = \lambda(v, q) = V_0(k_1(1 - q) + k_2(1 - q/v) + k_3(1 - v))$$

$$k_1 = 7E_0 \quad 42$$

$$k_2 = 2$$

$$k_3 = 2E_0 - 0.2$$

where  $V_0$  is resting blood volume fraction. This signal comprises a volume-weighted sum of extra- and intra-vascular signals that are functions of volume and deoxyhemoglobin content. A critical term in (42) is the concentration term  $k_2(1 - q/v)$ , which accounts for most of the nonlinear behaviour of the hemodynamic model.. The architecture of this model is summarised in Figure 17.

### 3 Extension to a MISO

The extension to a multiple input system is trivial and involves extending Eq(39) to cover  $n$  inputs

$$\dot{s} = \varepsilon_1 u(t)_1 + \dots + \varepsilon_n u(t)_n - \kappa_s s - \kappa_f (f_{in} - 1) \quad 43$$

The model now has  $5 + n$  parameters; five biophysical parameters  $\kappa_s, \kappa_f, \tau, \alpha$  and  $E_0$  and  $n$  efficacies  $\varepsilon_1, \dots, \varepsilon_n$ . Although all these parameters have to be estimated we are only interested in making inferences about the efficacies. Note that the biophysical parameters are the same for all inputs.

### 4 The Volterra formulation

In our hemodynamic model the state variables are  $X = \{x_1, \dots, x_4\}^T = \{s, f_{in}, v, q\}^T$  and the parameters are  $\theta = \{\theta_1, \dots, \theta_{5+n}\}^T = \{\kappa_s, \kappa_f, \tau, \alpha, E_0, \varepsilon_1, \dots, \varepsilon_n\}^T$ . The state equations and output nonlinearity specify a multiple input-single output (MISO) model

$$\begin{aligned} \dot{X}(t) &= f(X, u(t)) \\ y(t) &= \lambda(X(t)) \\ \dot{x}_1 &= f_1(X, u(t)) = \varepsilon_1 u(t)_1 + \dots + \varepsilon_n u(t)_n - \kappa_s x_1 - \kappa_f (x_2 - 1) \\ \dot{x}_2 &= f_2(X, u(t)) = x_1 \\ \dot{x}_3 &= f_3(X, u(t)) = \frac{1}{\tau} (x_2 - f_{out}(x_3, \alpha)) \\ \dot{x}_4 &= f_4(X, u(t)) = \frac{1}{\tau} \left( x_2 \frac{E(x_2, E_0)}{E_0} - f_{out}(x_3, \alpha) \frac{x_4}{x_3} \right) \\ y(t) &= \lambda(x_1, \dots, x_4) = V_0 (k_1 (1 - x_4) + k_2 (1 - x_4 / x_3) + k_3 (1 - x_3)) \end{aligned} \quad 44$$

This is the state-space representation. The alternative Volterra formulation represents the output  $y(t)$  as a nonlinear convolution of the input  $u(t)$ , critically without reference to the state variables  $X(t)$  (see Bendat 1990). This series can be considered a nonlinear convolution that obtains from a functional Taylor expansion of  $y(t)$  about  $X(0)$  and  $u(t) = 0$ . For a single input this can be expressed as

$$\begin{aligned}
y(t) &= h(\theta, u) \\
&= \kappa_0 + \sum_{i=1}^{\infty} \int_0^{\infty} \dots \int_0^{\infty} \kappa_i(\sigma_1, \dots, \sigma_i) u(t - \sigma_1) \dots u(t - \sigma_i) d\sigma_1 \dots d\sigma_i \\
\kappa_i(\sigma_1, \dots, \sigma_i) &= \frac{\partial^i y(t)}{\partial u(t - \sigma_1) \dots \partial u(t - \sigma_i)}
\end{aligned} \tag{45}$$

where  $\kappa_i$  is the  $i$ th generalised convolution kernel (Fliess *et al* 1983). (45) now expresses the output as a function of the input and the parameters whose posterior distribution we require. The Volterra kernels are a time-invariant characterisation of the input-output behaviour of the system and can be thought of as generalised high order convolution kernels that are applied to a stimulus function to emulate the observed BOLD response. Integrating (45) and applying the output nonlinearity to the state variables is the same as convolving the inputs with the kernels. Both give the system's response in terms of the output. In what follows, the response is evaluated by integrating (45). This means the kernels are not required. However, the Volterra formulation is introduced for several reasons. First, it demonstrates that the output is a nonlinear function of the inputs  $y(t) = h(\theta, u)$ . This is critical for the generality of the estimation scheme below. Secondly, it provides an important connection with conventional analyses using the general linear model (see below). Finally, we use the kernels to characterise evoked responses.

### C The Priors

Bayesian estimation requires informative priors on the parameters. Under Gaussian assumptions these prior densities can be specified in terms of their expectation and covariance. These moments are taken here to be the sample mean and covariance, over voxels, of the parameter estimates reported in Friston *et al* (2000). Normally priors play a critical role in inference; indeed the traditional criticism levelled at Bayesian inference reduces to reservations about the validity of the priors employed. However, in the application considered here, this criticism can be discounted. This is because the priors, on those parameters about which inferences are made, are relatively flat. Only the five biophysical parameters have informative priors.

In Friston *et al* (2000) the parameters were identified as those that minimised the sum of squared differences between the Volterra kernels implied by the parameters

and those derived directly from the data. This derivation used ordinary least square estimators, exploiting the fact that Volterra formulation is linear in the unknowns, namely the kernel coefficients. The kernels can be thought of as a re-parameterisation of the model that does not refer to the underlying state representation. In other words, for every set of parameters there is a corresponding set of kernels (see Friston *et al* 2000 for the derivation of the kernels as a function of the parameters). The data and Volterra kernel estimation are described in detail in Friston *et al* (1998). In brief, we obtained fMRI time-series from a single subject at 2 Tesla using a Magnetom VISION (Siemens, Erlangen) whole body MRI system, equipped with a head volume coil. Multi-slice T2\*-weighted fMRI images were obtained with a gradient echo-planar sequence using an axial slice orientation (TE = 40ms, TR = 1.7 seconds, 64x64x16 voxels). After discarding initial scans (to allow for magnetic saturation effects) each time-series comprised 1200 volume images with 3mm isotropic voxels. The subject listened to monosyllabic or bi-syllabic concrete nouns (i.e. 'dog', 'radio', 'mountain', 'gate') presented at 5 different rates (10 15 30 60 and 90 words per minute) for epochs of 34 seconds, intercalated with periods of rest. The presentation rates were repeated according to a Latin Square design.

The distribution of the five biophysical parameters, over 128 voxels, was computed to give our prior expectation  $\eta_\theta$  and covariance  $C_\theta$ . Signal decay  $\kappa_s$  had a mean of about 0.65 per sec. giving a half-life  $t_{1/2} = \ln 2 / \kappa_s \approx 1\text{sec}$ . Mean feedback rate  $\kappa_f$  was about 0.4 per sec. Mean Transit time  $\tau$  was 0.98 seconds. Under steady state conditions Grubb's parameter  $\alpha$  is about 0.38. The mean over voxels was 0.326. Mean resting oxygen extraction  $E_0$  was about 34% and the range observed conformed exactly with known values for resting oxygen extraction fraction (between 20% and 55%). Figure 18 shows the covariances among the biophysical parameters along with the correlation matrix (left-hand panel). The correlations suggest a high correlation between transit time and the rate constants for signal elimination and auto-regulation.

The priors for the efficacies were taken to be relatively flat with an expectation of zero and a variance of 16 per sec. The efficacies were assumed to be independent of the biophysical parameters with zero covariance. A variance of 16, or standard deviation of 4, corresponds to time constants in the range of 250ms. In other words, inputs can elicit flow-inducing signal over wide range of time constants from

infinitely slowly to very fast (250ms) with about the same probability. A 'strong' activation usually has an efficacy in the range of 0.5 to 0.6 per sec. Notice that from a dynamical perspective 'activation' depends upon the speed of the response not the percentage change. Equipped with these priors we can now pursue a fully Bayesian approach to estimating the parameters using new data sets and multiple input models:

#### D System identification

This subsection describes Bayesian inference procedures for nonlinear observation models, with additive noise, of the form

$$y = h(\theta, u) + e \quad 46$$

under Gaussian assumptions about the parameters  $\theta$  and errors  $e \sim N\{0, C_\varepsilon\}$ . These models can be adopted for any analytic dynamical system due to the existence of the equivalent Volterra series expansion above. Assuming the posterior density of the parameters is approximately Gaussian the problem reduces to finding its first two moments, the conditional mean  $\eta_{\theta|y}$  and covariance  $C_{\theta|y}$ .

The observation model can be made linear by expanding (46) about a working estimate  $\eta_{\theta|y}$  of the conditional mean.

$$\begin{aligned} h(\theta, u) &\approx h(\eta_{\theta|y}) + J(\theta - \eta_{\theta|y}) \\ J &= \frac{\partial h(\eta_{\theta|y})}{\partial \theta} \end{aligned} \quad 47$$

such that  $y - h(\eta_{\theta|y}) \approx J(\theta - \eta_{\theta|y}) + \varepsilon$ . This linear model can now be placed in the EM scheme described in Appendix A.1 to give

*Until convergence* {

**E-step**

$$\begin{aligned}
J &= \frac{\partial h(\eta_{\theta|y})}{\partial \theta} \\
\bar{y} &= \begin{bmatrix} y - h(\eta_{\theta|y}) \\ \eta_{\theta} - \eta_{\theta|y} \end{bmatrix}, \quad \bar{J} = \begin{bmatrix} J \\ I \end{bmatrix}, \quad \bar{C}_{\epsilon} = \begin{bmatrix} \sum \lambda_i Q_i & 0 \\ 0 & C_{\theta} \end{bmatrix} \\
C_{\theta|y} &= (\bar{J}^T \bar{C}_{\epsilon}^{-1} \bar{J})^{-1} \\
\eta_{\theta|y} &\leftarrow \eta_{\theta|y} + C_{\theta|y} \bar{J}^T \bar{C}_{\epsilon}^{-1} \bar{y}
\end{aligned}$$

48

**M-Step**

$$\begin{aligned}
P &= \bar{C}_{\epsilon}^{-1} - \bar{C}_{\epsilon}^{-1} \bar{J} C_{\theta|y} \bar{J}^T \bar{C}_{\epsilon}^{-1} \\
\frac{\partial F}{\partial \lambda_i} &= -\frac{1}{2} \text{tr}\{P Q_i\} + \frac{1}{2} \bar{y}^T P^T Q_i P \bar{y} \\
\left\langle \frac{\partial^2 F}{\partial \lambda_{ij}^2} \right\rangle &= -\frac{1}{2} \text{tr}\{P Q_i P Q_j\} \\
\lambda &\leftarrow \lambda - \left\langle \frac{\partial^2 F}{\partial \lambda^2} \right\rangle^{-1} \frac{\partial F}{\partial \lambda}
\end{aligned}$$

)

This EM scheme is effectively a *Gauss-Newton* search for the posterior mode or MAP estimate of the parameters. The relationship between the **E**-step and a conventional Gauss-Newton ascent can be seen easily in terms of the derivatives of their respective objective functions. For conventional Gauss-Newton this function is the *log likelihood*

$$\begin{aligned}
l &= \ln p(y|\theta) = -\frac{1}{2} (y - h(\theta))^T C_{\epsilon}^{-1} (y - h(\theta)) + \text{const.} \\
\frac{\partial l}{\partial \theta}(\eta_{ML}) &= J^T C_{\epsilon}^{-1} (y - h(\eta_{ML})) \\
-\frac{\partial^2 l}{\partial \theta^2}(\eta_{ML}) &\approx J^T C_{\epsilon}^{-1} J \\
\eta_{ML} &\leftarrow \eta_{ML} + (J^T C_{\epsilon}^{-1} J)^{-1} J^T C_{\epsilon}^{-1} (y - h(\eta_{ML}))
\end{aligned}$$

49

This is a conventional Gauss-Newton scheme. By simply augmenting the log likelihood with the log prior we get the *log posterior*



$$\begin{aligned}
l &= \ln p(\boldsymbol{\theta}|y) = \ln p(y|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) \\
&= -\frac{1}{2}(y - h(\boldsymbol{\theta}))^T C_\varepsilon^{-1} y - h(\boldsymbol{\theta}) - \frac{1}{2}(\boldsymbol{\eta}_\theta - \boldsymbol{\theta})^T C_\theta^{-1} (\boldsymbol{\eta}_\theta - \boldsymbol{\theta}) + \text{const.} \\
\frac{\partial l}{\partial \boldsymbol{\theta}}(\boldsymbol{\eta}_{\theta|y}) &= J^T C_\varepsilon^{-1} (y - h(\boldsymbol{\eta}_{\theta|y})) + C_\theta^{-1} (\boldsymbol{\eta}_\theta - \boldsymbol{\eta}_{\theta|y}) \\
-\frac{\partial^2 l}{\partial \boldsymbol{\theta}^2}(\boldsymbol{\eta}_{\theta|y}) &\approx J^T C_\varepsilon^{-1} J + C_\theta^{-1} \\
\boldsymbol{\eta}_{\theta|y} &\leftarrow \boldsymbol{\eta}_{\theta|y} + (J^T C_\varepsilon^{-1} J + C_\theta^{-1})^{-1} (J^T C_\varepsilon^{-1} (y - h(\boldsymbol{\eta}_{\theta|y})) + C_\theta^{-1} (\boldsymbol{\eta}_\theta - \boldsymbol{\eta}_{\theta|y}))
\end{aligned} \tag{50}$$

which is identical to the expression for the conditional expectation in the **E-Step**.

In summary, the only difference between the **E-step** and a conventional Gauss-Newton search is that priors are included in the objective log probability function converting it from a log likelihood into a log posterior. The use of an EM algorithm rests upon the need to find not only the conditional mean but also the hyperparameters of unknown variance components. The **E-step** finds (i) the current MAP estimate that provides the next expansion point for the Gauss-Newton search and (ii) the conditional covariance required by the **M-Step**. The **M-step** then updates the ReML estimates of the covariance hyperparameters that are required to compute the conditional moments in the **E-step**. Technically (48) is a *generalised* EM (GEM) because the **M-step** increases the log likelihood of the hyperparameter estimates, as opposed to maximising it.

### *1 Relationship to established procedures*

The procedure presented above represents a fairly obvious extension to conventional Gauss-Newton searches for the parameters of nonlinear observation models. The extension has two components: First, maximisation of the *posterior* density that embodies priors, as opposed to the likelihood. This allows for the incorporation of prior information into the solution and ensures uniqueness and convergence. Second the estimation of unknown covariance components. This is important because it accommodates non-sphericity in the error terms. The overall approach engenders a relatively simple way of obtaining Bayes estimators for nonlinear systems with unknown additive observation error. Technically, the algorithm represents a *posterior mode estimation* for nonlinear observation models using EM. It can be regarded as approximating the posterior density of the parameters by replacing the conditional mean with the mode and the conditional precision with the curvature (at the current

expansion point). Covariance hyperparameters are then estimated, which maximise the expectation of the log likelihood of the data over this approximate posterior density.

Posterior mode estimation is an alternative to full posterior density analysis, which avoids numerical integration (Fahrmeir and Tutz 1994, p58) and has been discussed extensively in the context of *generalised linear models* (e.g. Leonard 1972, Santner and Duffy 1989). The departure from Gaussian assumptions in generalised linear models comes from non-Gaussian likelihoods, as opposed to nonlinearities in the observation model considered here, but the issues are similar. Posterior mode estimation usually assumes the error covariances and priors are known. If the priors are unknown constants then empirical Bayes can be employed to estimate the required hyperparameters.

It is important not to confuse this application of EM with Kalman filtering. Although Kalman filtering can be formulated in terms of EM and, indeed, posterior mode estimation, Kalman filtering is used with completely different observation models - *state-space models*. State space or dynamic models comprise a *transition* equation and an *observation* equation (c.f. the state equation and output nonlinearity above) and cover systems in which the underlying state is hidden and is treated as a stochastic variable. This is not the sort of model considered here, in which the inputs (experimental design) and the ensuing states are known. This means that the conditional densities can be computed for the entire time-series simultaneously (Kalman filtering updates the conditional density recursively, by stepping through the time-series). If we treated the inputs as unknown and random then the state equation could be re-written as a stochastic differential equation (SDE) and a transition equation derived from it, using local linearity assumptions. This would form the basis of a state-space model. This approach may be useful for accommodating deterministic noise in the hemodynamic model but, in this treatment, we consider the inputs to be fixed. This means that the only random effects enter at the level of the observation or output nonlinearity. In other words, we are assuming that the measurement error in fMRI is the principal source of randomness in our measurements and that hemodynamic responses *per se* are determined by known inputs. This is the same assumption used in conventional analyses of fMRI data.

## 2 A note on Integration

To iterate Eq (48) the local gradients  $J = \partial h / \partial \theta$  have to be evaluated. This involves evaluating  $h(\theta, u)$  around the current expansion point with the generalised convolution of the inputs for the current conditional parameter estimates according to (45) or, equivalently, the integration of (44). The latter can be accomplished efficiently by capitalising on the fact that stimulus functions are usually sparse. In other words inputs arrive as infrequent events (*e.g.* event-related paradigms) or changes in input occur sporadically (*e.g.* boxcar designs). We can use this to evaluate  $y(t) = h(\eta_{\theta|y}, u)$  at the times the data were sampled using a bilinear approximation to (44). The Taylor expansion of  $\dot{X}(t)$  about  $X(0) = X_0 = [0, 1, 1, 1]^T$

$$\dot{X}(t) \approx f(X_0, 0) + \frac{\partial f(X_0, 0)}{\partial X} (X - X_0) + \sum_i u(t)_i \left( \frac{\partial^2 f(X_0, 0)}{\partial X \partial u_i} (X - X_0) + \frac{\partial f(X_0, 0)}{\partial u_i} \right)$$

has a bilinear form, following a change of variables (equivalent to adding an extra state variable  $x_0(t) = 1$ )

$$\begin{aligned} \dot{\tilde{X}}(t) &\approx A\tilde{X} + \sum_i u(t)_i B_i \tilde{X} \\ \tilde{X} &= \begin{bmatrix} 1 \\ X \end{bmatrix} \\ A &= \begin{bmatrix} 0 & 0 \\ f(X_0, 0) - \frac{\partial f(X_0, 0)}{\partial X} X_0 & \frac{\partial f(X_0, 0)}{\partial X} \end{bmatrix} \\ B_i &= \begin{bmatrix} 0 & 0 \\ \frac{\partial f(X_0, 0)}{\partial u_i} - \frac{\partial^2 f(X_0, 0)}{\partial X \partial u_i} X_0 & \frac{\partial^2 f(X_0, 0)}{\partial X \partial u_i} \end{bmatrix} \end{aligned} \tag{51}$$

This bilinear approximation is important because the Volterra kernels of bilinear systems have closed-form expressions. This means that the kernels can be derived analytically, and quickly, to provide a characterisation of the impulse response properties of the system. The integration of (51) is predicated on its solution over periods  $\Delta t_k = t_{k+1} - t_k$  within which the inputs are constant.

$$\begin{aligned}
\tilde{X}(t_{k+1}) &\approx e^{J\Delta t_k} \tilde{X}(t_k) \\
y(t_{k+1}) &\approx \lambda(X(t_{k+1})) \\
J &= A + \sum_i u(t_k)_i B_i
\end{aligned}
\tag{52}$$

This quasi-analytical integration scheme can be an order of magnitude quicker than straightforward numerical integration, depending on the sparsity of inputs.

### E Relation to conventional fMRI analyses

Note that if we treated the five biophysical parameters as known canonical values and discounted all but the first order terms in the Volterra expansion (45) the following linear model would result

$$\begin{aligned}
h(u, \theta) &= \kappa_0 + \sum_{i=1}^n \int_0^t \kappa_1(\sigma) u(t-\sigma)_i d\sigma = \sum_{i=1}^n \kappa_1 * u(t)_i \\
&\approx \kappa_0 + \sum_{i=1}^n \left( \frac{\partial \kappa_1}{\partial \varepsilon_i} * u(t)_i \right) \varepsilon_i
\end{aligned}
\tag{53}$$

where \* denotes convolution and the second expression is a first order Taylor expansion around the expected values of the parameters<sup>8</sup>. This is exactly the same as the general linear model adopted in conventional analysis of fMRI time series, if we elect to use just one (canonical) hemodynamic response function *HRF* to convolve our stimulus functions with. In this context the *HRF* plays the role of  $\partial \kappa_1 / \partial \varepsilon_i$  in (53). This partial derivative is shown in Figure 19 (upper panel) using the prior expectations of the parameters and conforms closely to the sort of *HRF* used in practice. Now, by treating the efficacies as fixed effects (*i.e.* with flat priors) the MAP and ML estimators reduce to the same thing and the conditional expectation reduces to the Gauss-Markov estimator

$$\eta_{ML} = (J^T C_\varepsilon^{-1} J)^{-1} J^T C_\varepsilon^{-1} y$$

where  $J$  is the design matrix. This is precisely the estimator used in conventional analyses when whitening strategies are employed.

Consider now the second order Taylor approximation to (53) that obtains when we do not know the exact values of the biophysical parameters and they are treated as unknown

$$h(\theta, u) \approx \kappa_0 + \sum_{i=1}^n \left[ \left( \frac{\partial \kappa_1}{\partial \varepsilon_i} * u(t)_i \varepsilon_i + \frac{1}{2} \sum_{j=1}^5 \left( \frac{\partial^2 \kappa_1}{\partial \varepsilon_i \partial \theta_j} * u(t)_i \right) \varepsilon_i \theta_j \right) \right] \quad 54$$

This expression<sup>9</sup> is precisely the general linear model proposed in Friston *et al* (1998) and implemented in our software. In this instance the explanatory variables comprise the stimulus functions, each convolved with a small temporal basis set corresponding to the canonical  $\partial \kappa_1 / \partial \varepsilon_i$  and its partial derivatives with respect to the biophysical parameters. Examples of these second order partial derivatives are provided in the lower panel of Figure 19. The unknowns in this general linear model are the efficacies  $\varepsilon_i$  and the interaction between the efficacies and the biophysical parameters  $\varepsilon_i \theta_j$ . Of course, the problem with this linear approximation is that generalised least squares estimates of the unknown coefficients  $\beta = [\varepsilon_1, \dots, \varepsilon_n, \varepsilon_1 \theta_1, \dots, \varepsilon_n \theta_1, \varepsilon_1 \theta_2, \dots]^T$  are not constrained to factorise into stimulus-specific efficacies  $\varepsilon_i$  and biophysical parameters  $\theta_j$  that are the same for all inputs. Only a nonlinear estimation procedure can do this.

In the usual case of using a temporal basis set (*e.g.* a canonical form and various derivatives) one obtains a ML or generalised least squares estimate of [functions of]

---

<sup>8</sup> Note that in this first order Taylor approximation  $\kappa_1 = 0$  when expanding around the prior expectations of the efficacies = 0. Furthermore, all first order partial derivatives  $\partial \kappa_1 / \partial \theta_i = 0$  unless they are with respect to an efficacy.

<sup>9</sup> Note that in this second order Taylor approximation all the second order partial derivatives  $\partial^2 \kappa_1 / \partial \theta_i \partial \theta_j = 0$  unless they are with respect to an efficacy and one of the biophysical parameters.

the parameters in some subspace defined by the basis set. Operationally this is like specifying priors but of a very particular form. This form can be thought of as uniform priors over the support of the basis set and zero elsewhere. In this sense basis functions implement hard constraints that may not be very realistic but provide for efficient estimation. The soft constraints implied by the Gaussian priors in the EM approach are more plausible but are computationally more expensive to implement.

### *1 Summary*

This subsection has described a nonlinear EM algorithm that can be viewed as a Gauss-Newton search for the conditional distribution of the parameters of deterministic dynamical system, with additive Gaussian error. It was shown that classical approaches to fMRI data analysis are special cases that ensue when considering only first order kernels and adopting flat or uninformative priors. Put another way, the scheme can be regarded as a generalisation of existing procedures that is extended in two important ways. First, the model encompasses nonlinearities and second; it moves the estimation from a classical into a Bayesian frame.

## **G An empirical illustration**

### *1 Single input example*

In this, the first of the two examples, we revisit the original data set on which the priors were based. This constitutes a single-input study where the input corresponds to the aural presentation of single words, at different rates, over epochs. The data were subject to a conventional event-related analysis where the stimulus function comprised trains of spikes indexing the presentation of each word. The stimulus function was convolved with a canonical *HRF* and its temporal derivative. The data were high pass filtered by removing low frequency components modelled by a discrete cosine set. The resulting  $SPM\{T\}$ , testing for activations due to words, is shown in Figure 20 (left hand panel) thresholded at  $p = 0.05$  (corrected).

A single region in the left superior temporal gyrus was selected for analysis. The input comprised the same stimulus function used in the conventional analysis and the output was the first eigenvariate of high-pass filtered time-series, of all voxels, within a 4mm sphere, centred on the most significant voxel in the  $SPM\{T\}$  (marked by an arrow in Figure 20). The error covariance basis set  $Q$  comprised two bases; an

identity matrix modelling white or an *i.i.d.* component and a second with exponentially decaying off-diagonal elements modelling an AR(1) component (see Friston *et al* 2002b and equation 23). This models serial correlations among the errors. The results of the estimation procedure are shown in the right hand panel in terms of (i) the conditional distribution of the parameters and (ii) the conditional expectation of the first and second order kernels. The kernels are a function of the parameters and their derivation using a bilinear approximation is described in Friston *et al* (2000). The upper right panel shows the first order kernels for the state variables (signal, inflow, deoxyhemoglobin content and volume). These can be regarded as impulse response functions detailing the response to a transient input. The first and second order output kernels for the BOLD response are shown in the lower right panels. They concur with those derived empirically in Friston *et al* (2000). Note the characteristic undershoot in the first order kernel and the pronounced negativity in the upper left of the second order kernel, flanked by two off-diagonal positivities at around 8 seconds. These lend the hemodynamics a degree of refractoriness when presenting paired stimuli less than a few seconds apart and a super-additive response with about 8 seconds separation. The left-hand panels show the conditional or posterior distributions. The density for the efficacy is presented in the upper panel and those for the five biophysical parameters are shown in the lower panel using the same format. The shading corresponds to the probability density and the bars to 90% confidence intervals. The values of the biophysical parameters are all within a very acceptable range. In this example the signal elimination and decay appears to be slower than normally encountered, with the rate constants being significantly larger than their prior expectations. Grubb's exponent here is closer to the steady state value of 0.38 than the prior expectation of 0.32. Of greater interest is the efficacy. It can be seen that the efficacy lies between 0.4 and 0.6 and is clearly greater than 0. This would be expected given we chose the most significant voxel from the conventional analysis. Notice there is no null hypothesis here and we do not even need a  $p$  value to make the inference that words evoke a response in this region. An important facility, with inferences based on the conditional distribution and precluded in classical analyses, is that one can infer a cause did not elicit a response. This is demonstrated in the second example.

## 2 Multiple input example

In this example we turn to a data set used in previous sections, in which there are three experimental causes or inputs. This was a study of attention to visual motion. Subjects were studied with fMRI under identical stimulus conditions (visual motion subtended by radially moving dots) whilst manipulating the attentional component of the task (detection of velocity changes). The data were acquired from normal subjects at 2 Tesla using a Magnetom VISION (Siemens, Erlangen) whole body MRI system, equipped with a head volume coil. Here we analyse data from the first subject. Contiguous multi-slice T2\*-weighted fMRI images were obtained with a gradient echo-planar sequence (TE = 40ms, TR = 3.22 seconds, matrix size = 64x64x32, voxel size 3x3x3mm). Each subject had 4 consecutive 100-scan sessions comprising a series of 10-scan blocks under 5 different conditions D F A F N F A F N S. The first condition (D) was a dummy condition to allow for magnetic saturation effects. F (Fixation) corresponds to a low-level baseline where the subjects viewed a fixation point at the centre of a screen. In condition A (Attention) subjects viewed 250 dots moving radially from the centre at 4.7 degrees per second and were asked to detect changes in radial velocity. In condition N (No attention) the subjects were asked simply to view the moving dots. In condition S (Stationary) subjects viewed stationary dots. The order of A and N was swapped for the last two sessions. In all conditions subjects fixated the centre of the screen. In a pre-scanning session the subjects were given 5 trials with 5 speed changes (reducing to 1%). During scanning there were no speed changes. No overt response was required in any condition.

This design can be reformulated in terms of three potential causes, photic stimulation, visual motion and directed attention. The F epochs have no associated cause and represent a baseline. The S epochs have just photic stimulation. The N epochs have both photic stimulation and motion whereas the A epochs encompass all three causes. We performed a conventional analysis using boxcar stimulus functions encoding the presence or absence of each of the three causes during each epoch. These functions were convolved with a canonical *HRF* and its temporal derivative to give two repressors for each cause. The corresponding design matrix is shown in the left panel of Figure 21. We selected a region that showed a significant attentional effect in the lingual gyrus for Bayesian inference. The stimulus functions modelling the three inputs were the box functions used in the conventional analysis. The output corresponded to the first eigenvariate of high-pass filtered time-series from all voxels in a 4mm sphere centred on 0, -66, -3mm (Talairach and Tournoux 1998). The error



covariance basis set was simply the identity matrix<sup>10</sup>. The results are shown in the right-hand panel of Figure 21 using the same format as Figure 20. The critical thing here is that there are three conditional densities, one for each of the input efficacies. Attention has a clear activating effect with more than a 90% probability of being greater than 0.25 per sec. However, in this region neither photic stimulation *per se* or motion in the visual field evokes any real response. The efficacies of both are less than 0.1 and are centred on 0. This means that the time constants of the response to visual stimulation would range from about ten seconds to never. Consequently these causes can be discounted from a dynamical perspective. In short, this visually unresponsive area responds substantially to attentional manipulation *showing a true functional selectivity*. This is a crucial statement because classical inference does not allow one to infer any region does not respond and therefore precludes a formal inference about the selectivity of regional responses. The only reason one can say “this region responds *selectively* to attention” is because Bayesian inference allows one to say “it does *not* respond to photic stimulation with random dots or motion”.

## H Conclusion

In this section we have looked a method, that conforms to an EM implementation of the Gauss-Newton method, for estimating the conditional or posterior distribution of the parameters of a deterministic dynamical system. The inclusion of priors in the estimation procedure ensures robust and rapid convergence and the resulting conditional densities enable Bayesian inference about the model's parameters. We have examined the coupling between experimentally designed causes or factors in fMRI studies and the ensuing BOLD response. This application represents a generalisation of existing linear models to accommodate nonlinearities in the transduction of experimental causes to measured output in fMRI. Because the model is predicated on biophysical processes the parameters have a physical interpretation. Furthermore the approach extends classical inference about the likelihood of the data, to more plausible inferences about the parameters of the model given the data. This inference provides confidence intervals based on the conditional density.

---

<sup>10</sup> We could motivate this by noting the TR is considerably longer in these data than in the previous example. However, in reality, serial correlations were ignored because the loss of sparsity in the

Perhaps the most important extension of the scheme described in this section is to MIMO systems where we deal with multiple regions or voxels at the same time. The fundamental importance of this extension is that one can incorporate interactions among brain regions at the neuronal level. This provides a promising framework for the dynamic causal modelling of functional integration in the brain (**see Chapter 22: Dynamic causal modelling**)

## VI APPENDIX

### A.1 The EM algorithm

This appendix describes EM 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 maximised 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 maximise the likelihood of the observed data  $p(y|\lambda)$ , conditional on some hyperparameters, in the presence of unobserved variables or parameters  $\theta$ . This is equivalent to maximising 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

---

associated inverse covariance matrices considerably increases computation time and we wanted to repeat the analysis many times (see next subsection).

algorithm alternates between maximising  $F$ , and implicitly the likelihood of the data, with respect to the distribution  $q(\theta)$  and the hyperparameters  $\lambda$ , holding the other fixed

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

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

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  $\lambda$  that maximise  $p(y|\lambda)$  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 optimise the hyperparameters, in a maximum likelihood sense, using this distribution. These new hyperparameters re-enter 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 maximise the first term in the expression for  $F$  above. From (15)

$$\begin{aligned}
\log p(\theta, y|\lambda) &= -\frac{1}{2} \ln |C_\epsilon| - \frac{1}{2} (\bar{y} - \bar{X}\theta)^T C_\epsilon^{-1} (\bar{y} - \bar{X}\theta) + \text{const.} \\
\int q(\theta) \ln p(\theta, y|\lambda) d\theta &= -\frac{1}{2} \ln |C_\epsilon| - \frac{1}{2} r^T C_\epsilon^{-1} r - \frac{1}{2} \left\langle (\theta - \eta_{\theta|y})^T \bar{X}^T C_\epsilon^{-1} \bar{X} (\theta - \eta_{\theta|y}) \right\rangle_q + \text{const.} \\
&= -\frac{1}{2} \ln |C_\epsilon| - \frac{1}{2} r^T C_\epsilon^{-1} r - \frac{1}{2} \text{tr} \{ C_{\theta|y} \bar{X}^T C_\epsilon^{-1} \bar{X} \} + \text{const.} \\
\int q(\theta) \log q(\theta) &= -\frac{1}{2} \ln |C_{\theta|y}| + \text{const.}
\end{aligned}$$

$$F = \frac{1}{2} \ln |C_\epsilon^{-1}| - \frac{1}{2} r^T C_\epsilon^{-1} r - \frac{1}{2} \text{tr} \{ C_{\theta|y} \bar{X}^T C_\epsilon^{-1} \bar{X} \} + \frac{1}{2} \ln |C_{\theta|y}| + \text{const.} \quad \text{A.2}$$

where the residuals  $r = \bar{y} - \bar{X}\eta_{\theta|y}$ . 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)

$$\frac{\partial F}{\partial C_\epsilon^{-1}} = \frac{1}{2} C_\epsilon - \frac{1}{2} r r^T - \frac{1}{2} \bar{X} C_{\theta|y} \bar{X}^T \quad \text{A.3}$$

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

$$C(\lambda)_\epsilon = r r^T + \bar{X} C_{\theta|y} \bar{X}^T \quad \text{A.4}$$

(*c.f.* Dempster *et al* (1981) p350). 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.

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

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 realisations of the data. The ensuing matrix  $H$  is referred to as Fisher’s Information matrix<sup>11</sup>. The computation of the gradient vector  $g$  can be made computationally efficient by capitalising 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 expansion of  $C\{\lambda\}_\varepsilon$ . In many instances the bases can be constructed so that they do not ‘overlap’ or interact through the design matrix *i.e.*  $PQ_i PQ_j = 0$  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  $C(\lambda)_\varepsilon$  the hyperparameters can be calculated very simply<sup>12</sup>. However, we work with the general solution above that encompasses all these special cases.

---

<sup>11</sup> 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

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

and (ii) taking the expectation, using  $\left\langle \text{tr} \{ P Q_i P \bar{y} \bar{y}^T P Q_j \} \right\rangle_q = \text{tr} \{ P Q_i P C_\varepsilon P Q_j \} = \text{tr} \{ P Q_i P Q_j \}$

<sup>12</sup> Note that if there is only one hyperparameter then  $g = 0$  can be solved directly

$$\begin{aligned}
\text{tr} \{ P Q \} &= \bar{y}^T P Q P \bar{y} \Rightarrow \\
\lambda &= \frac{r^T Q^{-1} r}{\text{tr} \{ R \}}
\end{aligned}$$

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 Figure 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 generalised 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 parameterisation of  $C_\varepsilon$  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-parameterisation described above is that does not guarantee that  $C_\varepsilon$  is positive definite. This is because the hyperparameters can take negative values with extreme degrees of non-sphericity. 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) maximise  $F$  (or minimise 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

---

where  $C_\varepsilon = \lambda Q$  and  $R = I - \bar{X}(\bar{X}^T Q^{-1} \bar{X})^{-1} \bar{X}^T Q^{-1}$  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.

function: Substituting the  $C_{\theta|y} = (\bar{X}^T C_\varepsilon^{-1} \bar{X})^{-1}$  from (15) into the expression for the negative free energy (A.2) gives

$$F = -\frac{1}{2} \ln |C_\varepsilon| - \frac{1}{2} r^T C_\varepsilon^{-1} r - \frac{1}{2} \ln |\bar{X}^T C_\varepsilon^{-1} \bar{X}| + \text{const.} \quad \text{A.6}$$

which is the ReML objective function (see Harville 1977, p325). Critically the derivatives of A.6, with respect to the hyperparameters, are exactly the same as those given in (A.5)<sup>13</sup>. Operationally, (A.5) can be rearranged to give a ReML scheme by removing any explicit reference to the conditional covariance.

$$\begin{aligned} g_i &= -\frac{1}{2} \text{tr}\{PQ_i\} + \frac{1}{2} \text{tr}\{P\bar{y}\bar{y}^T P^T Q_i\} \\ H_{ij} &= \frac{1}{2} \text{tr}\{PQ_i P Q_j\} \end{aligned} \quad \text{A.7}$$

$$P = C_\varepsilon^{-1} - C_\varepsilon^{-1} \bar{X} (\bar{X}^T C_\varepsilon^{-1} \bar{X})^{-1} \bar{X}^T C_\varepsilon^{-1}$$

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-parameterisations 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 covariance matrix  $yy^T$ . This is possible because the conditional parameter estimates are not required in the ReML formulation. This is used in the current version of the SPM software to estimate voxel-wide non-sphericity.

---

<sup>13</sup> Note that  $\frac{\partial \ln |\bar{X}^T C_\varepsilon^{-1} \bar{X}|}{\partial \lambda_i} = \text{tr} \left\{ (\bar{X}^T C_\varepsilon^{-1} \bar{X})^{-1} \frac{\partial \bar{X}^T C_\varepsilon^{-1} \bar{X}}{\partial \lambda_i} \right\} = -\text{tr} \{ C_{\theta|y} \bar{X}^T C_\varepsilon^{-1} Q_i C_\varepsilon^{-1} \bar{X} \}$

## References

- Bendat JS. (1990) *Nonlinear System Analysis and Identification from Random Data*. John Wiley and Sons, New York USA
- Berry DA and Hochberg Y (1999) Bayesian perspectives on multiple comparisons. *J. Statistical Planning and Inference*. 82: 215-227
- Büchel C and Friston KJ. (1997) Modulation of connectivity in visual pathways by attention: Cortical interactions evaluated with structural equation modelling and fMRI. *Cerebral Cortex* 7,768-778
- Box GEP (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *Ann. Math. Stats.* 25:290-302
- Bullmore ET Brammer MJ Williams SCR Rabe-Hesketh S Janot N David A Mellers J Howard R and Sham P. (1996) Statistical methods of estimation and inference for functional MR images. *Mag. Res. Med.* 35:261-277
- RB Buxton, EC Wong, and LR Frank. Dynamics of blood flow and oxygenation changes during brain activation: The Balloon model. (1998) *MRM* 39, 855-864
- Copas JB (1983) Regression prediction and shrinkage. *J. Roy. Statistical. Soc. Series B* 45;311-354
- Dempster AP, Laird NM and Rubin (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. Series B* 39;1-38
- Dempster AP, Rubin DB and Tsutakawa R.K. (1981) Estimation in covariance component models. *J. Am. Statistical Assoc.* 76;341-353
- Descombes X, Kruggel F and von Cramon DY (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. Stats. Assoc.* 68:117-130
- Efron B and Morris C (1977) Stein's paradox in statistics. *Scientific American*. May:119-127
- Everitt BS and Bullmore ET (1999) Mixture model mapping of brain activation in functional magnetic resonance images. *Human Brain Mapp.* 7:1-14
- Fahrmeir L and Tutz G (1994) *Multivariate statistical modelling based on generalised linear Models*, Pringer-Verlag Inc. New York. p355-356
- M Fliess, M Lamnabhi and F Lamnabhi-Lagarrigue (1983) An algebraic approach to nonlinear functional expansions. *IEEE Trans. Circuits Syst.* 30, 554-570
- Friston KJ Jezzard PJ and Turner R. (1994) Analysis of functional MRI time-series *Hum. Brain Mapp.* 1:153-171



- Friston, KJ Holmes, AP Worsley, KJ Poline, J-B Frith, CD and Frackowiak RSJ. (1995) Statistical parametric maps in functional imaging: A general linear approach. *Hum. Brain Map.* **2**,189-210
- Friston KJ Josephs O Rees G Turner R (1998) Non-linear event-related responses in fMRI. *Magnetic Resonance in Medicine* **39**:41-52
- Friston KJ, Mechelli A, Turner R and Price CJ (2000) Nonlinear responses in fMRI: The Balloon model, Volterra kernels and other hemodynamics. *NeuroImage* **12**, 466-477
- Friston KJ, Penny W, Phillips C, Kiebel S, Hinton G and Ashburner J. (2002a) Classical and Bayesian inference in neuroimaging: Theory. *NeuroImage.* 16:465-483
- Friston KJ, Glaser DE, Henson RNA Kiebel S and Phillips C and Ashburner J. (2002b) Classical and Bayesian inference in neuroimaging: Applications. *NeuroImage.* 16:484-512
- Friston KJ (2002) Bayesian estimation of dynamical systems: An application to fMRI. *NeuroImage* – 16:465-483
- Geisser S and Greenhouse SW (1958). An extension of Box's results on the use of the F distribution in multivariate analysis. *Ann. Math. Stats.* **29**:885-891
- Grubb RL, Rachael ME, Euchring JO, and Ter-Pogossian MM. (1974) The effects of changes in PCO<sub>2</sub> on cerebral blood volume, blood flow and vascular mean transit time. *Stroke* **5**, 630-639
- Hartley H (1958). Maximum likelihood estimation from incomplete data. *Biometrics.* **14**;174-194
- Hartvig NV and Jensen JL (2000) Spatial mixture modelling of fMRI data. *Hum. Brain Mapp.* in press
- Harville DA (1977) Maximum likelihood approaches to variance component estimation and to related problems. *J. Am. Stat. Assoc.* **72**:320-338
- Henson RNA, Rugg MD, Shallice T and Dolan RJ (2000) Confidence in recognition memory for words: Dissociating right prefrontal roles in episodic retrieval. *J. Cog. Neurosci.* in press
- Hoge RD, Atkinson J, Gill B, Crelier GR, Marrett S and Pike GB (1999) Linear coupling between cerebral blood flow and oxygen consumption in activated human cortex. *Proc. Natl. Acad. Sci.* **96**, 9403-9408
- Højten-Sørensen P, Hansen LK and Rasmussen CE (2000) Bayesian modelling of fMRI time-series. In *Advances in Neural Information Processing Systems 12*. SA Solla, TK Leen and KR Muller (Eds.) MIT Press pp 754-760.
- Holmes A. and Ford I. (1993) A Bayesian approach to significance testing for statistic images from PET. In *Quantification of Brain function, Tracer kinetics and Image analysis in brain PET*. Eds. K. Uemura, N.A. Lassen, T. Jones and I. Kanno. Excerpta Medica, Int. Cong. Series No. 1030: 521-534

- Holmes AP and Friston KJ (1998) Generalisability, Random Effects and Population Inference. *NeuroImage* 7:754
- Kass RE and Steffey D (1989) Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J. Am. Stat. Assoc.* **407**:717-726
- Laird NM and Ware JH (1982) Random effects models for longitudinal data. *Biometrics.* **38**:963-974
- Lee PM (1997) *Bayesian Statistics an Introduction*. John Wiley and Son Inc. 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*, MI Jordan Ed. Kluwer Academic Press, pp 355-368
- Mandeville JB, Marota JJ, Ayata C, Zararchuk G, Moskowitz MA, Rosen B and Weisskoff RM (1999) Evidence of a cerebrovascular postarteriole windkessel with delayed compliance. *J. Cereb. Blood Flow Metab.* **19**, 679-689
- Mayhew J, Hu D, Zheng Y, Askew S, Hou Y, Berwick J, Coffey PJ, and Brown N (1998) An evaluation of linear models analysis techniques for processing images of microcirculation activity *NeuroImage* 7, 49-71
- Miller KL, Luh WM, Liu, TT, Martinez A, Obata T, Wong EC, Frank LR and Buxton RB (2000) Characterizing the dynamic perfusion response to stimuli of short duration. *Proc. ISRM* **8**, 580
- Phillips C, Mattout J, Rugg MD, Pierre Maquet P, and Friston KJ (2003) Restricted maximum likelihood solution of the source localization problem in EEG. *NeuroImage* (submitted abstract)
- Purdon PL and Weisskoff R (1998) Effect of temporal autocorrelations due to physiological noise stimulus paradigm on voxel-level false positive rates in fMRI. *Hum. Brain Mapp.* **6**:239-249
- Santner TJ and Duffy DE (1989). *The statistical analysis of discrete data*. New York Springer.
- Satterthwaite EF (1941). Synthesis of variance. *Psychometrika* **6**:309-316
- Talairach J and Tournoux P (1988) *A Co-planar stereotaxic atlas of a human brain*. Thieme, Stuttgart
- Tikhonov AN and Arsenin VY (1977) *Solution of ill posed problems*. Winston and Sons.
- Worsley KJ and Friston. KJ (1995) Analysis of fMRI time-series revisited - again *NeuroImage* **2**,173-181
- Worsley. KJ (1994) Local Maxima and the expected Euler characteristic of excursion sets of chi squared, F and t fields. *Advances Appl. Prob.* **26**,13-42
- Worsley, K.J., Liao, C., Aston, J., Petre, V., Duncan, G.H., Evans, AC (2002). A general statistical analysis for fMRI data. *NeuroImage*, - in press



## Figure legends

### *Figure 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 non-hierarchical form (lower panel). The design matrices are shown in image format with an arbitrary colour 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 modelled with the same three effects, one of these being a constant term. These design matrices are part of those used in Friston *et al* (2002b) to generate simulated fMRI data and are based on the design matrices used in the subsequent empirical event-related fMRI analyses.

### *Figure 2*

As for Figure 1 but here showing how the non-hierarchical form is augmented so that the parameter estimates (that include the error terms from all levels and the final level parameters) now appear in the model's residuals. A Gauss-Markov estimator will minimise these residuals in inverse proportion to their prior variance.

### *Figure 3*

Schematic illustrating the form of the covariance constraints. These can be thought of as 'design matrices' for the second-order behaviour 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 in Figure 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* (2002b) during the simulation of serially correlated fMRI data-sequences and covariance component estimation using real data.

*Figure 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).

*Figure 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 in Figure 4.

*Figure 6*

Top panel: True response (activation plus random low frequency components) and that based on the OLS and ML estimators for a simulated fMRI experiment. The insert shows the similarity between the OLS and ML predictions. Lower panel: True (dashed) and estimated (solid) autocorrelation functions. The sample autocorrelation function of the residuals (dotted line) and the best fit in terms of the covariance constraints (dot-dashed) are also shown. The insert shows the true covariance hyperparameters (black), those obtained just using the residuals (grey) and those estimated by the EM algorithm (white). Note, in relation to the EM estimates, those based directly on the residuals severely underestimate the actual correlations. The simulated data comprised 128 observations with an inter-scan interval of 2 seconds. The activations were modelled with a box-car (duty cycle 64 seconds) convolved with a canonical hemodynamic response function and scaled to a peak height of 2. The constant terms and low frequency components were simulated with a linear combination of the first 16 components of a discrete cosine set, each scaled by a random unit Gaussian variate. Serially correlated noise was formed by filtering unit Gaussian noise with a convolution kernel based on covariance hyperparameters of 1.0 [uncorrelated or white component] and 0.5 [AR(1) component].

*Figure 7*

Estimates of serial correlations expressed as autocorrelation functions based on empirical data. Left panel: Estimates from 12 randomly selected voxels from a single subject. Right panel: Estimates from the same voxel over 12 different subjects. The voxel was in the cingulate gyrus. The empirical data are described in Henson *et al* (2000). They comprised 300 volumes, acquired with EPI at two Tesla and a TR of three seconds. The experimental design was stochastic and event-related looking for differential response evoked by *new* relative to *old* (studied prior to the scanning session) words. Either a new or old word was presented visually with a mean stimulus onset asynchrony (SOA) of 4 seconds (SOA varied randomly between 2.5 and 5.5 seconds). Subjects were required to make an old *vs.* new judgement for each word. The design matrix for these data comprised two regressors (early and late) for each of the four trial types (old *vs.* new and correct *vs.* incorrect) and the first 16 components of a discrete cosine set (as in the simulations).

*Figure 8*

The results of an analysis of simulated event-related responses in a single voxel. Parameter and hyperparameter estimates based on a simulated fMRI study are shown in relation to the true values. The simulated data comprised 128 scans for each of 12 subjects with a mean peak response over subjects of 0.5%. The construction of these data is described in the main text. Stimulus presentation conformed to the presentation of 'old' words in the empirical analysis described in the main text. Serial correlations were modelled as in the main text. Upper left: first-level hyperparameters. The estimated subject-specific values (black) are shown alongside the true values (white). The first 12 correspond to the 'white' term or variance. The second 12 control the degree of autocorrelation and can be interpreted as the covariance between one scan and the next. Upper right: Hyperparameters for the early and late components of the evoked response. Lower left: The estimated subject-specific parameters pertaining to the early and late response components are plotted against their true values. Lower right: The estimated and true parameters at the second level representing the conditional mean of the distribution from which the subject-specific effects are drawn.

*Figure 9*

Response estimates and inferences about the estimates presented in Figure 8: Upper panel: True (dotted) and ML (solid) estimates of event-related responses to a stimulus over 12 subjects. The units of activation are adimensional and correspond to percent of whole brain mean. The insert shows the corresponding subject-specific T values for contrasts testing for early and late responses. Lower panel: The equivalent estimates based on the conditional means. It can be seen that the conditional estimates are much 'tighter' and reflect better the inter-subject variability in responses. The insert shows the posterior probability that the activation was greater than 0.1%. Because the responses were modelled with early and late components (basis functions corresponding to canonical hemodynamic response functions, separated by 3 seconds) separate posterior probabilities could be computed for each. The simulated data comprised only early responses as reflected in the posterior probabilities.

*Figure 10*

Estimation of differential event-related responses in real data. The format of this figure is identical to that of Figure 8. The only differences are that these results are based on real data where the response is due to the difference between studied or familiar (old) words and novel (new) words. In this example we used the first 128 scans from 12 subjects. Clearly in this figure we cannot include true effects.

*Figure 11*

The format of this figure is identical to that of Figure 9. The only differences are that these results are based on real data where the response is due to the difference between studied or familiar (old) words and novel (new) words. The same regression of conditional responses to the conditional mean is seen on comparing the ML and conditional estimates. In relation to the simulated data, there is more evidence for a late component but no late activation could be inferred for any subject with any degree of confidence. The voxel from which these data were taken was in the cingulate gyrus (BA 31) at  $-3, -33, 39$ mm.

*Figure 12*

Schematic summarising the two-step procedure for (1) ReML estimation of the prior covariance based on the data covariance, pooled over voxels and (2) a voxel-by-voxel

estimation of the posterior expectation and covariance of the parameters, required for inference. See the main text for a detailed explanation of the equations.

*Figure 13*

Bayesian and classical and inference for the PET study of word generation. **3a)** PPM for a contrast reflecting the difference between word-shadowing and word-generation, using an activation threshold of 2.2 and a confidence of 95%. The design matrix and contrast for this model are shown (right) in image format. We have modelled each scan as a specific effect that has been replicated over subjects. **3b)** Classical SPM of the  $t$  statistic for the same contrast. This SPM has been thresholded at  $p=0.05$ , corrected using a Gaussian field adjustment.

*Figure 14*

Illustrative results for a single voxel - the maximum in the left temporal region of the PPM in the previous figure (-54, -4, -2mm). Upper panel: These are the conditional or posterior expectations and 95% confidence intervals for the activation effect associated with each of the 12 conditions. Note that the odd conditions (word shadowing) are generally higher. In condition 5 one would be more than 95% certain the activation exceeded 2.2. Lower panel: The prior and posterior densities for the parameter estimate for condition 5.

*Figure 15*

PPM for the fMRI study of attention to visual motion. The display format in the lower panel uses an axial slice through extrastriate regions but the thresholds are the same as employed in maximum intensity projections (upper panels). The activation threshold for the PPM was 0.7. As can be imputed from the design matrix, the statistical model of evoked responses comprised box-car regressors convolved with a canonical hemodynamic response function.

*Figure 16*

As for Figure 15, but this time showing the corresponding SPM using a corrected threshold at  $p = 0.05$ .



*Figure 17*

Schematic illustrating the architecture of the hemodynamic model. This is a fully nonlinear single-input  $u(t)$ , single-output  $y(t)$  state model with four state variables  $s, f_{in}, v$  and  $q$ . The form and motivation for the changes in each state variable, as functions of the others, is described in the main text.

*Figure 18*

Prior covariances for the five biophysical parameters of the hemodynamic model in Figure 17. Left panel: Correlation matrix showing the correlations among the parameters in image format (white = 1). Right panel: Corresponding covariance matrix in tabular format. These priors represent the sample covariances of the parameters estimated by minimising the difference between the Volterra kernels implied by the parameters and those estimated, empirically using ordinary least squares as described in Friston *et al* (2000).

*Figure 19*

Partial derivatives of the kernels with respect to parameters of the model evaluated at their prior expectation. Upper panel: First-order partial derivative with respect to efficacy. Lower panels: Second-order partial derivatives with respect to efficacy and the biophysical parameters. When expanding around the prior expectations of the efficacies = 0 the remaining first- and second-order partial derivatives with respect to the parameters are zero.

*Figure 20*

A SISO example: Left panel: Conventional SPM{T} testing for an activating effect of word presentation. The arrow shows the centre of the region (a sphere of 4mm radius) whose response was entered into the Bayesian estimation procedure. The results for this region are shown in the right hand panel in terms of (i) the conditional distribution of the parameters and (ii) the conditional expectation of the first- and second-order kernels. The upper right panel shows the first-order kernels for the state variables (signal, inflow, deoxyhemoglobin content and volume). The first- and second-order output kernels for the BOLD response are shown in the lower right panels. The left-hand panels show the conditional or posterior distributions. That for

efficacy is presented in the upper panel and those for the five biophysical parameters in the lower panel. The shading corresponds to the probability density and the bars to 90% confidence intervals.

*Figure 21*

A MISO example using visual attention to motion. The left panel shows the design matrix used in the conventional analysis and the right panel shows the results of the Bayesian analysis of a lingual extrastriate region. This panel has the same format as Figure 20.



## Hierarchical form

1st level

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

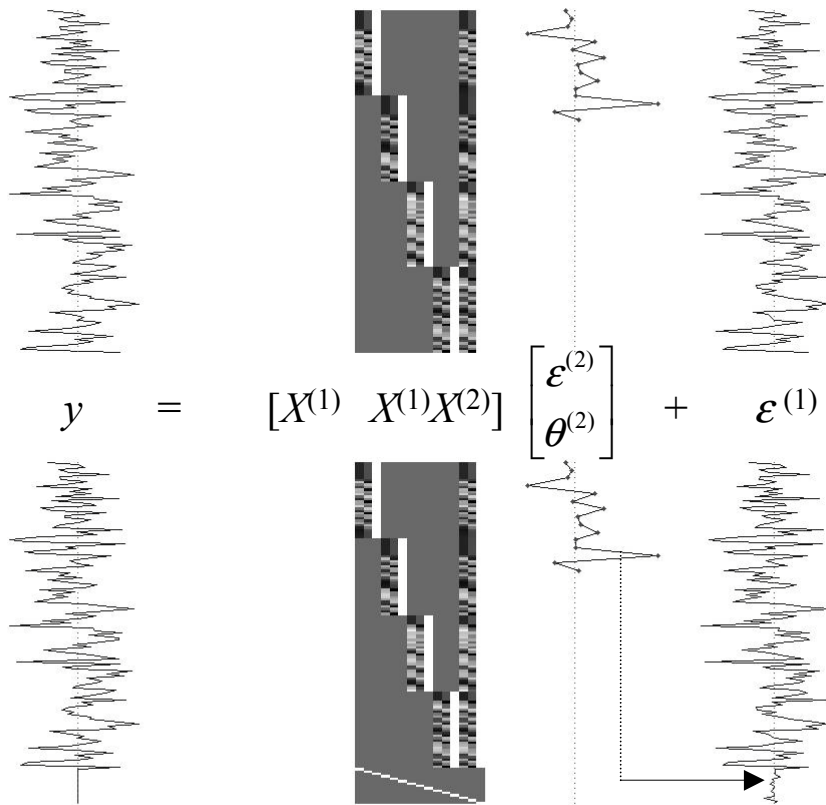
2nd level

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

## Non-hierarchical form

$$y = X^{(1)} \begin{bmatrix} \theta^{(2)} \\ \theta^{(1)} \end{bmatrix} + \varepsilon^{(1)}$$

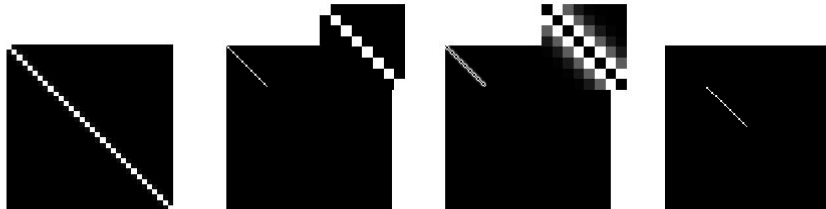
### Non-hierarchical form



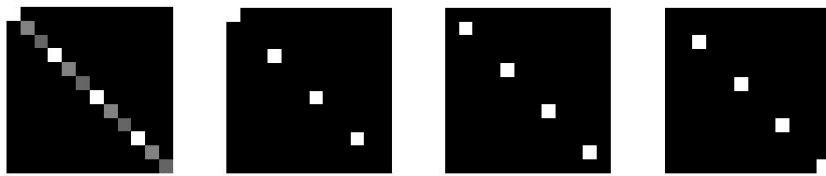
$$\begin{bmatrix} y \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} X^{(1)} & X^{(1)}X^{(2)} \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \epsilon^{(2)} \\ \theta^{(2)} \end{bmatrix} + \begin{bmatrix} \epsilon^{(1)} \\ -\epsilon^{(2)} \\ -\theta^{(2)} \end{bmatrix}$$

### Augmented form

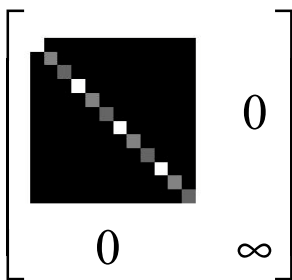
### Covariance constraints



$$C_{\varepsilon}^{(1)} = \lambda_1^{(1)} Q_1^{(1)} + \lambda_2^{(1)} Q_2^{(1)} + \lambda_3^{(1)} Q_3^{(1)} +$$



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



$$C_{\theta} = \begin{bmatrix} \text{[Diagonal Matrix]} & 0 \\ 0 & \infty \end{bmatrix}$$

Augment to embody priors in error covariance

$$\bar{X} = \begin{bmatrix} \prod_{i=1}^1 X^{(i)} & \dots & \prod_{i=1}^n X^{(i)} \\ I & & 0 \\ \vdots & \ddots & \\ 0 & \dots & I \end{bmatrix}, \quad \bar{y} = \begin{bmatrix} y \\ 0 \\ \vdots \\ \eta_{\theta}^{(n)} \end{bmatrix}, \quad C_{\theta} = \begin{bmatrix} 0 & 0 & 0 \\ & \ddots & \vdots \\ 0 & 0 & 0 \\ 0 & \dots & 0 & C_{\theta}^{(n)} \end{bmatrix}, \quad Q_1 = \begin{bmatrix} Q_1^{(1)} & 0 & 0 \\ & \ddots & \vdots \\ 0 & 0 & 0 \\ 0 & \dots & 0 & 0 \end{bmatrix}, \quad Q_2 = \dots$$

Until convergence { **E-Step**

$$\begin{aligned} C_{\varepsilon} &= C_{\theta} + \sum \lambda_k Q_k \\ C_{\theta|y} &= (\bar{X}^T C_{\varepsilon}^{-1} \bar{X})^{-1} \\ \eta_{\theta|y} &= C_{\theta|y} \bar{X}^T C_{\varepsilon}^{-1} \bar{y} \end{aligned}$$

**M-Step**

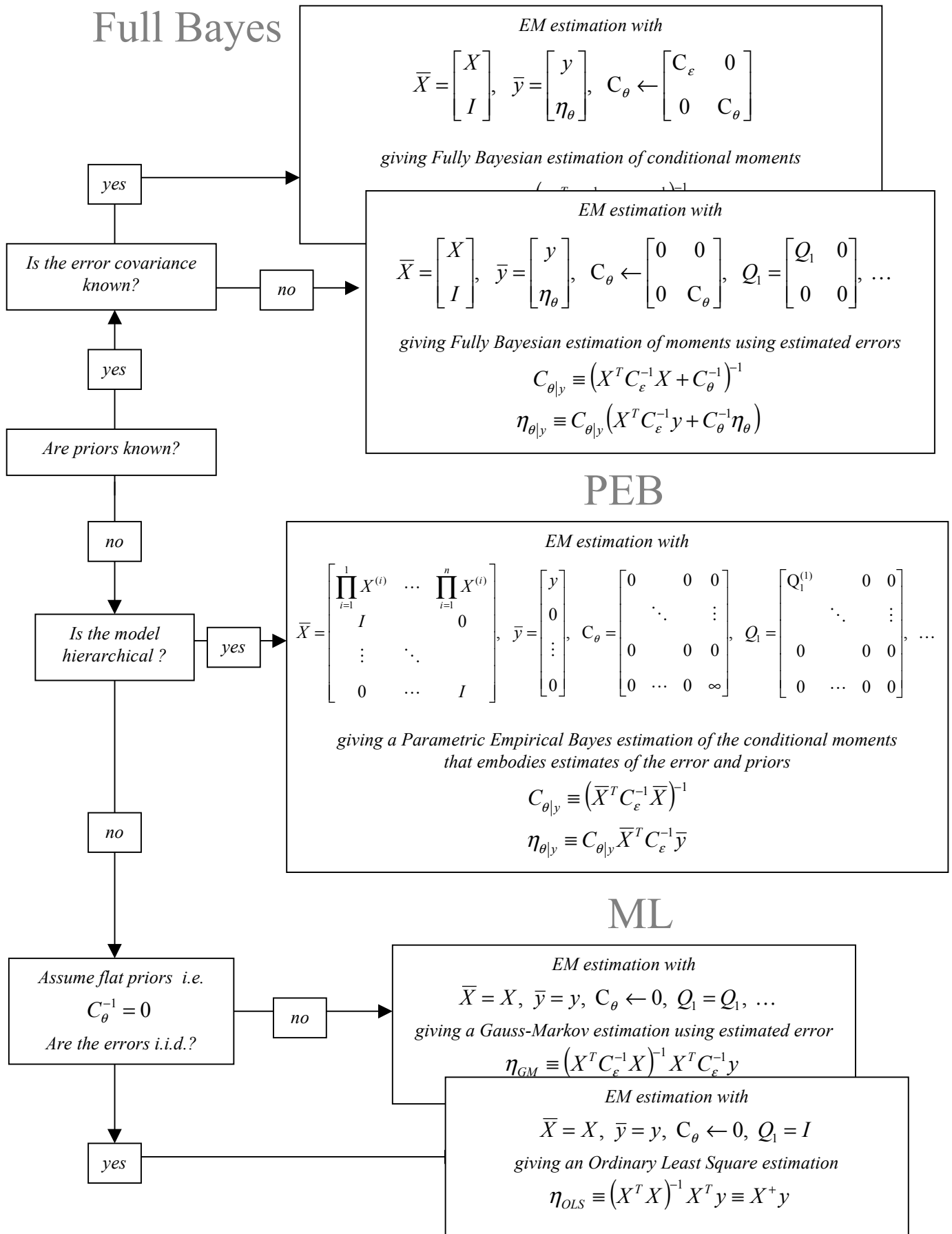
$$\begin{aligned} P &= C_{\varepsilon}^{-1} - C_{\varepsilon}^{-1} \bar{X} C_{\theta|y} \bar{X}^T C_{\varepsilon}^{-1} \\ g_i &= -\frac{1}{2} \text{tr}\{P Q_i\} + \frac{1}{2} \bar{y}^T P^T Q_i P \bar{y} \\ H_{ij} &= \frac{1}{2} \text{tr}\{P Q_i P Q_j\} \\ \lambda &= \lambda + H^{-1} g \end{aligned}$$

}

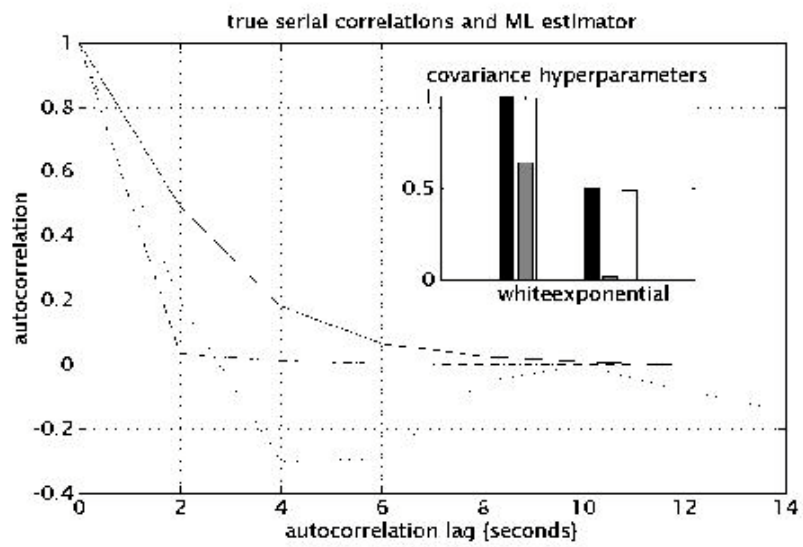
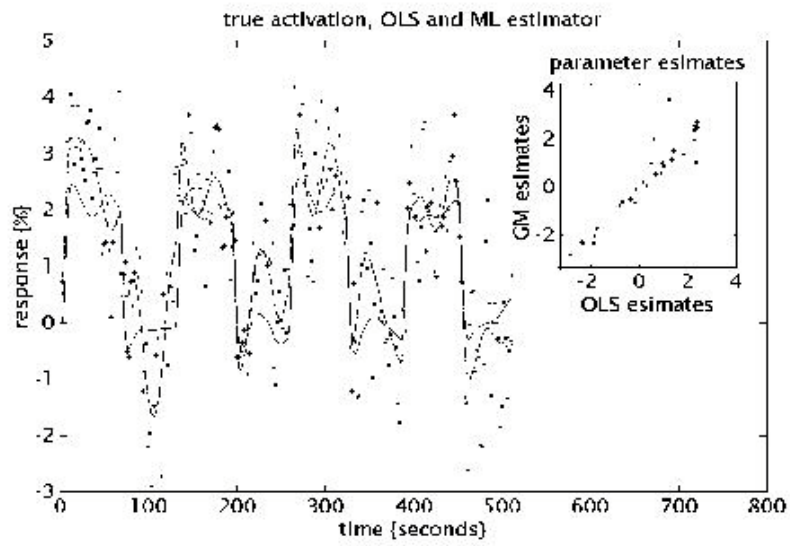
assemble estimates of error covariance, priors, conditional covariances and means

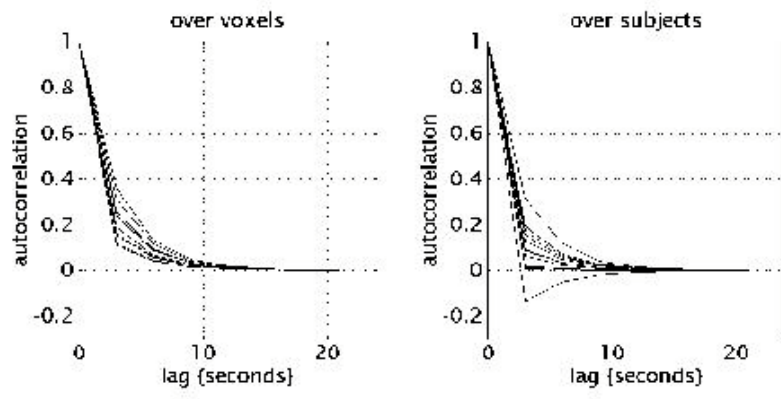
$$\begin{aligned} C_{\varepsilon} &= \begin{bmatrix} C_{\varepsilon}^{(1)} & & 0 & 0 \\ & \ddots & & \vdots \\ 0 & & C_{\varepsilon}^{(n)} & 0 \\ 0 & \dots & 0 & C_{\theta}^{(n)} \end{bmatrix}, \quad C_{\theta}^{(i)} = C_{\varepsilon}^{(i-1)} \\ C_{\theta|y} &= \begin{bmatrix} C_{\varepsilon|y}^{(2)} & \dots & & \\ \vdots & \ddots & & \\ & & C_{\varepsilon|y}^{(n)} & \\ & & & C_{\theta|y}^{(n)} \end{bmatrix}, \quad C_{\theta|y}^{(i)} = C_{\varepsilon|y}^{(i-1)} \\ \eta_{\theta|y} &= \begin{bmatrix} \eta_{\varepsilon|y}^{(2)} \\ \vdots \\ \eta_{\varepsilon|y}^{(n)} \\ \eta_{\theta|y}^{(n)} \end{bmatrix}, \quad \eta_{\theta|y}^{(i-1)} = X^{(i)} \eta_{\theta|y}^{(i)} + \eta_{\varepsilon|y}^{(i)} \end{aligned}$$

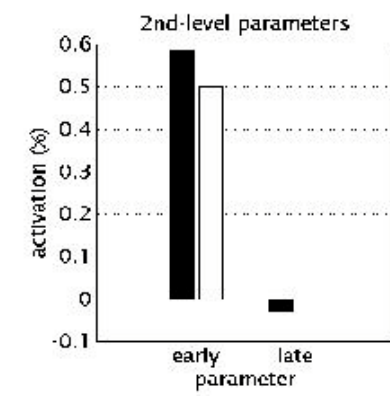
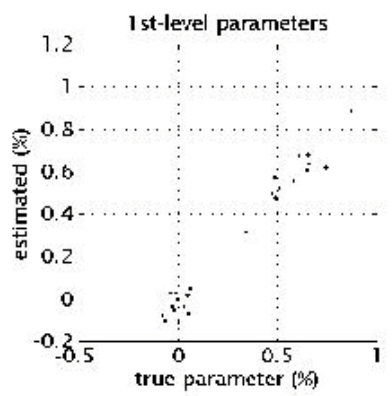
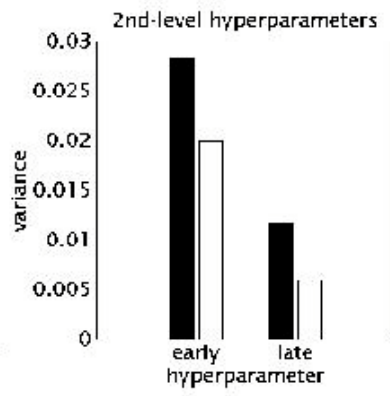
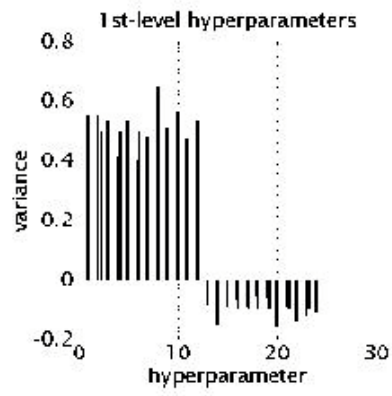
# Full Bayes

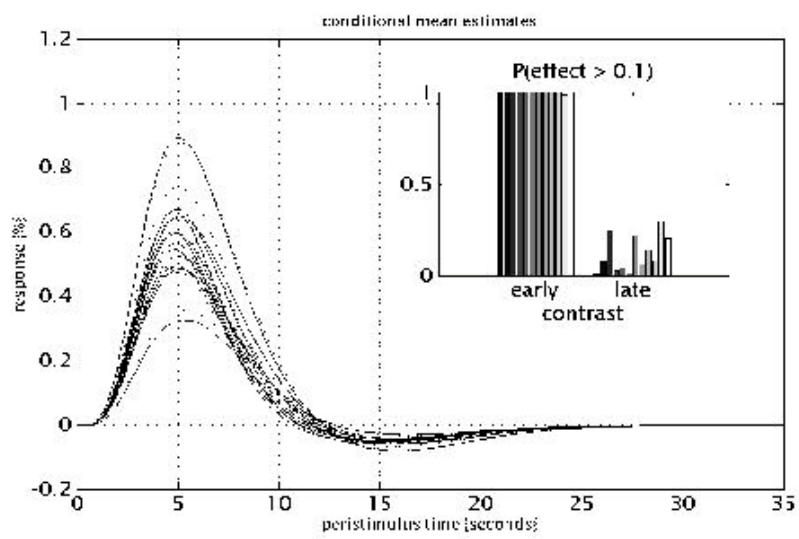
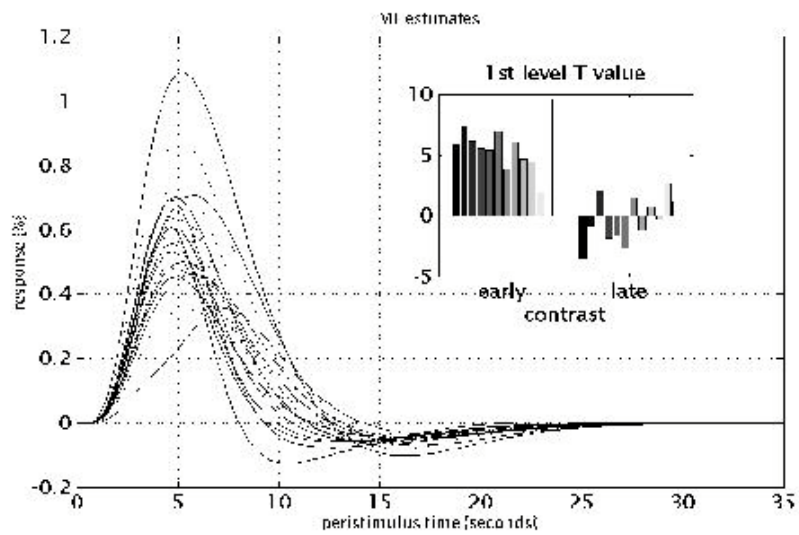


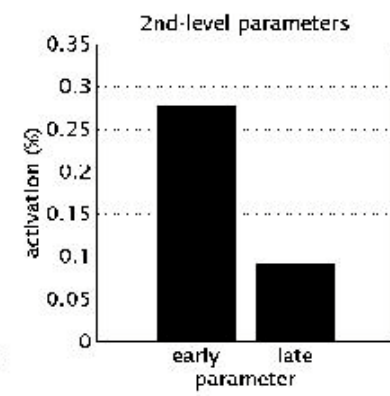
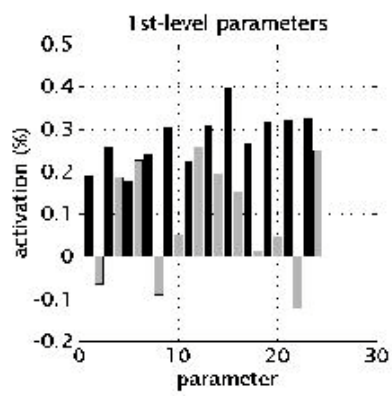
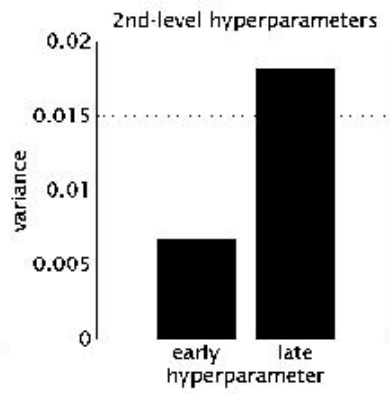
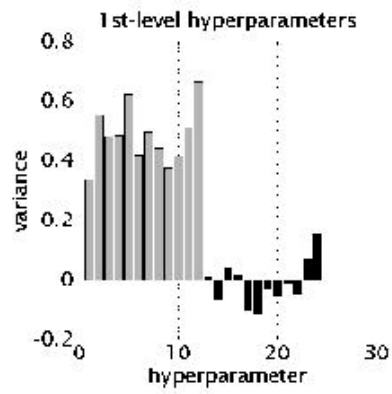


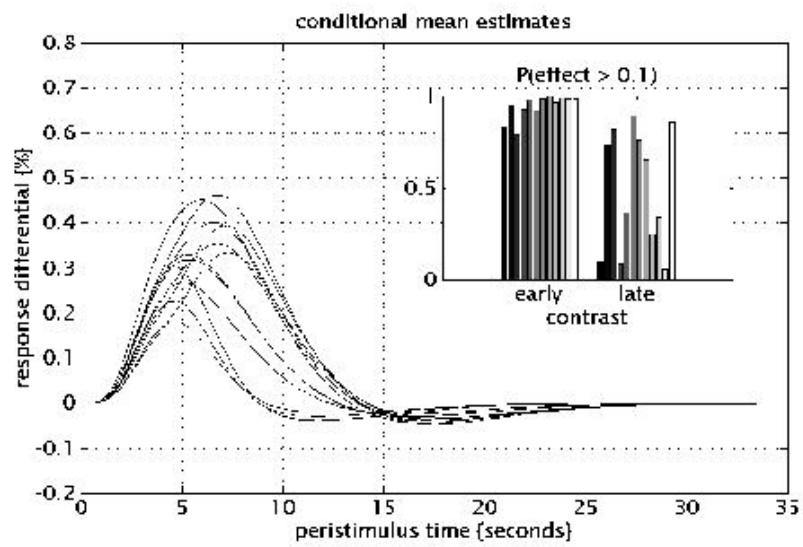
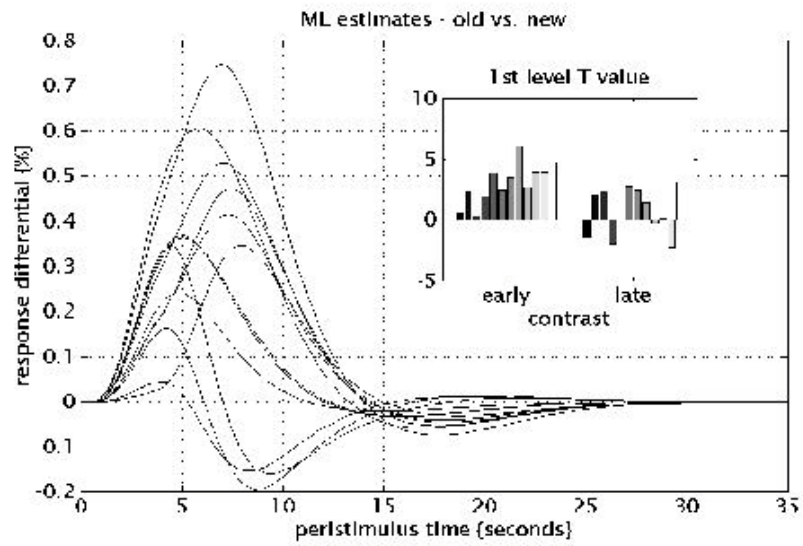


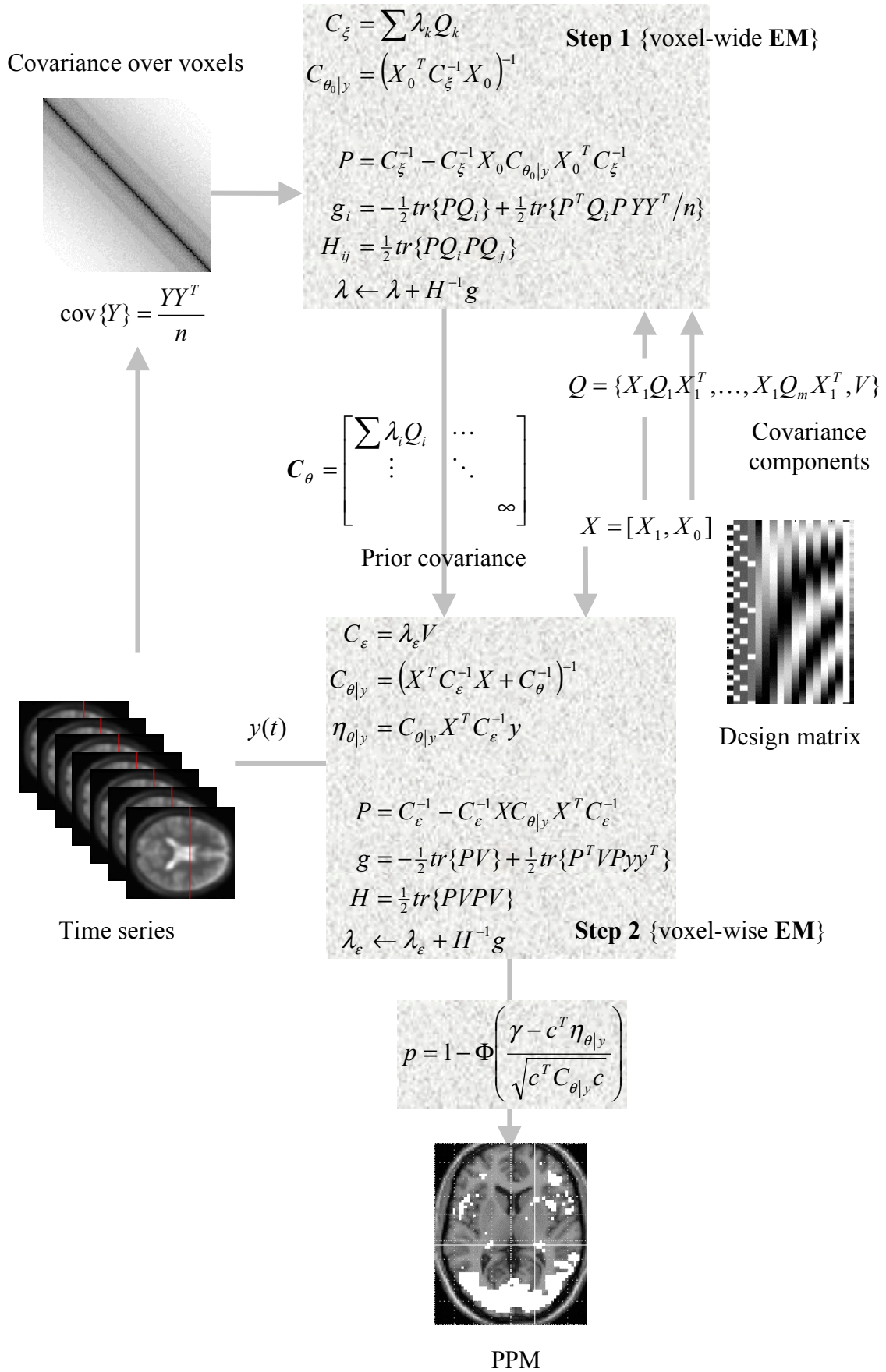




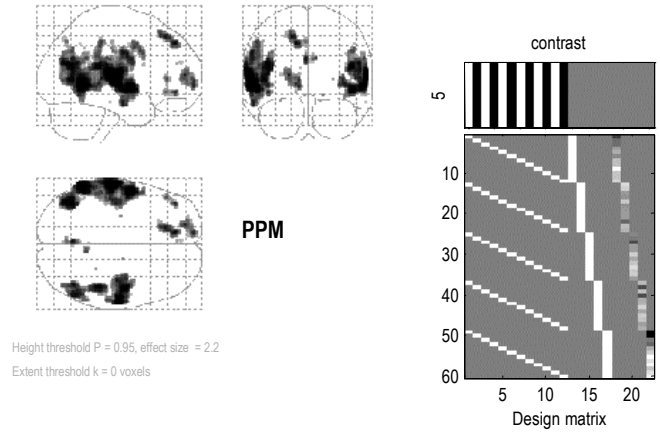




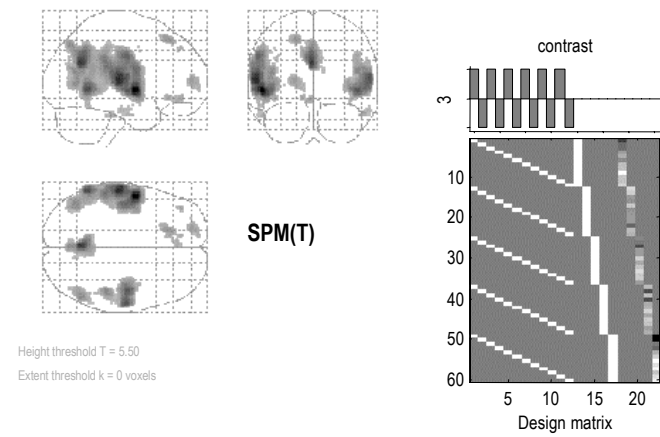




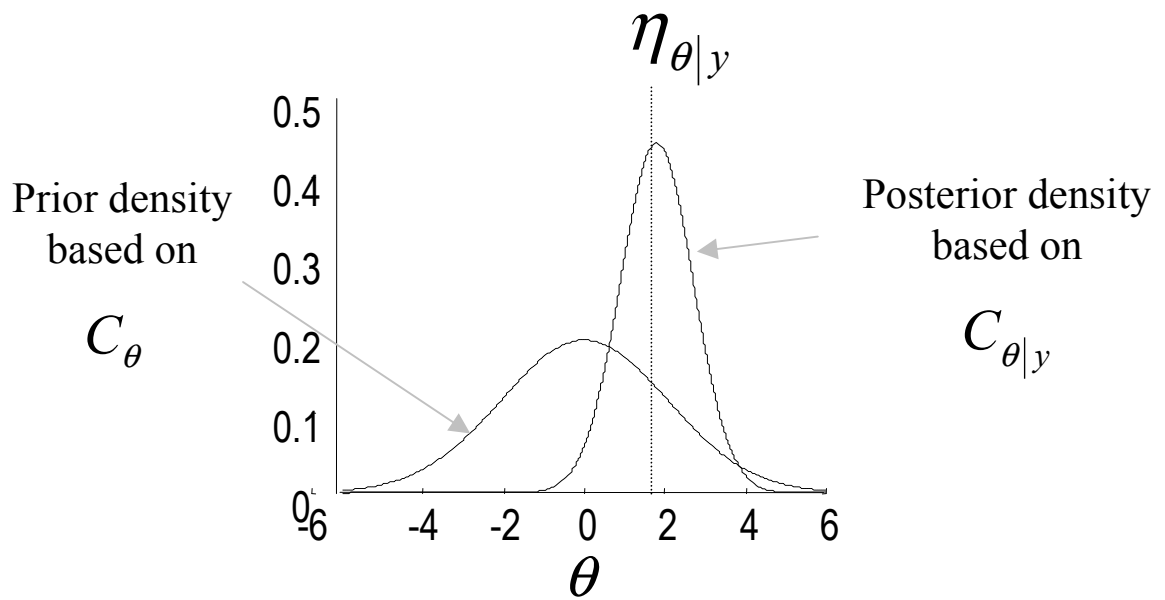
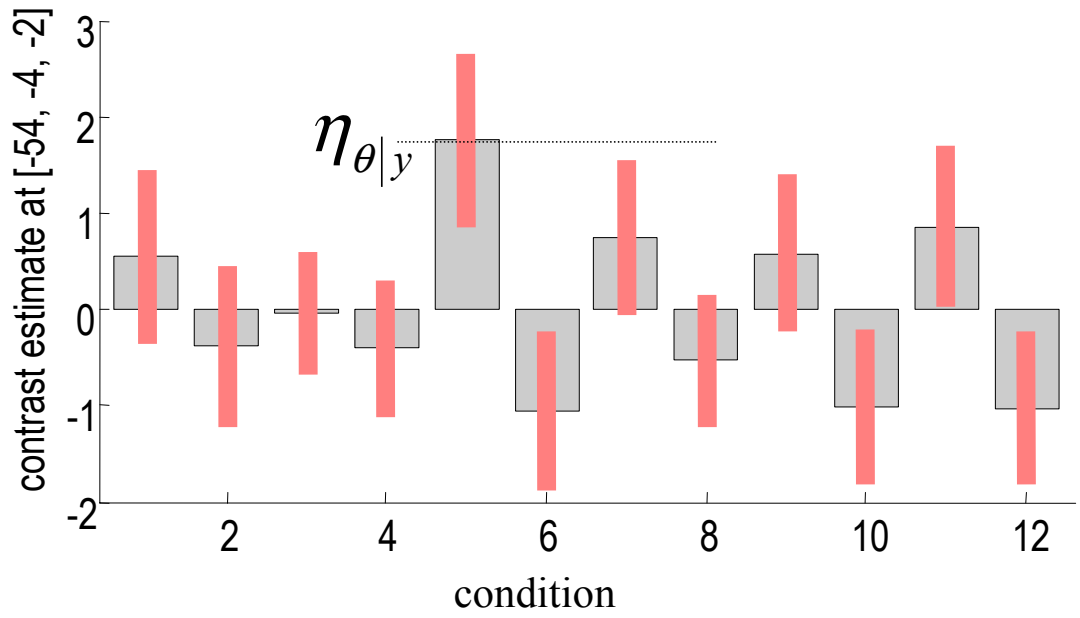
a

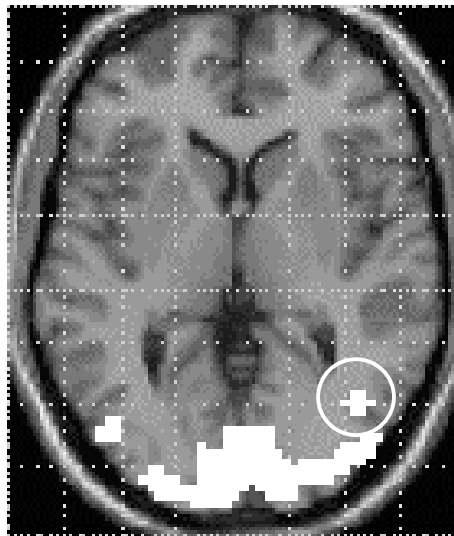
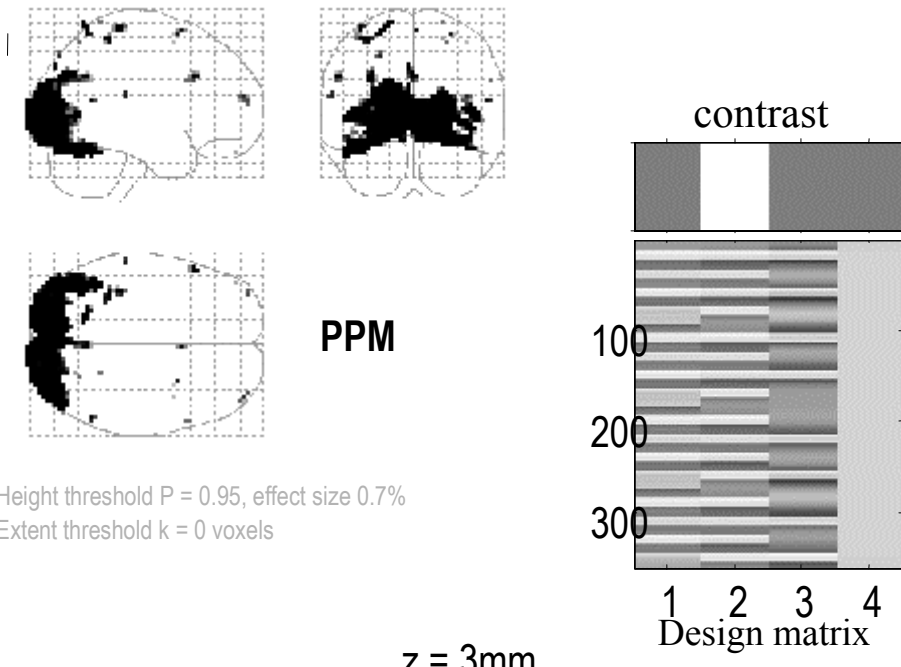


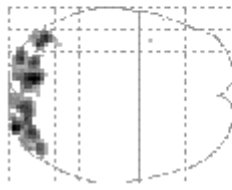
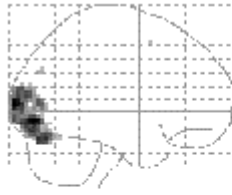
b





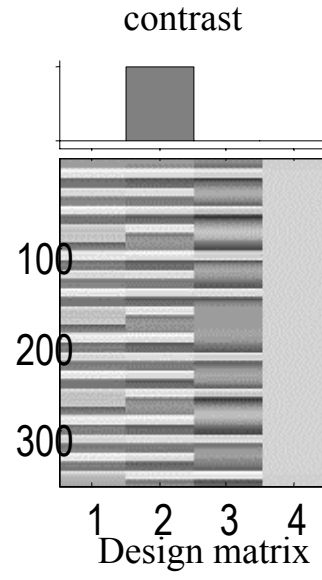




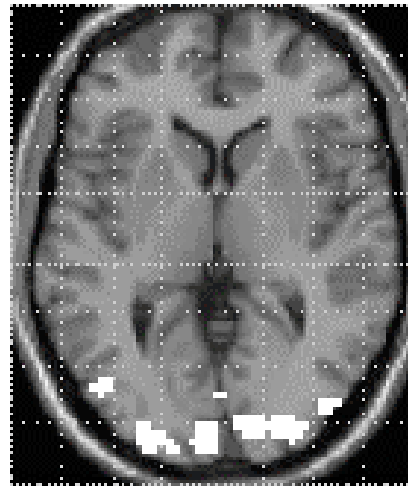


**SPM(T)**

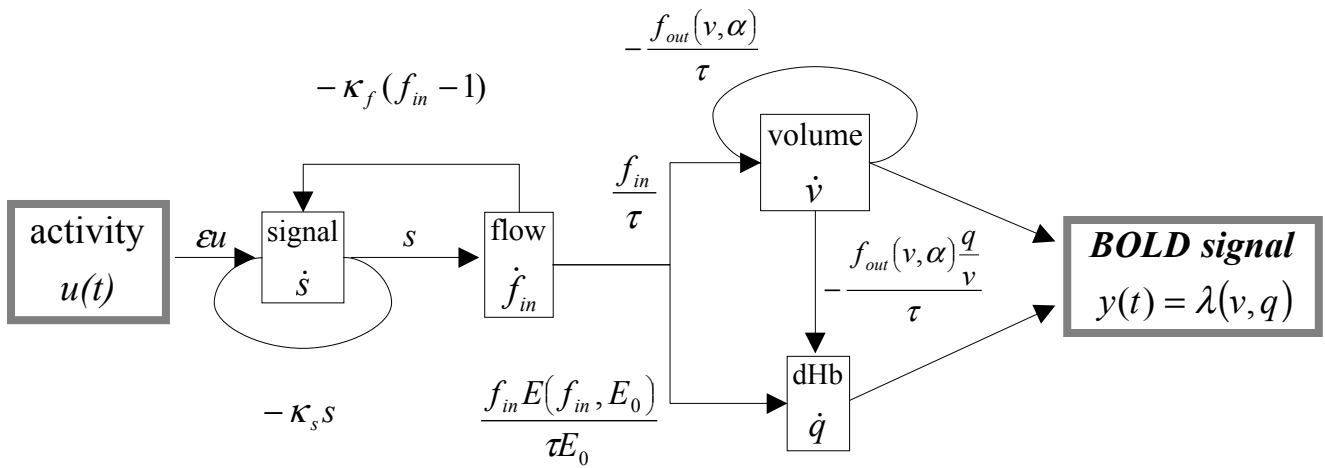
Height threshold  $T = 4.86$   
Extent threshold  $k = 0$  voxels



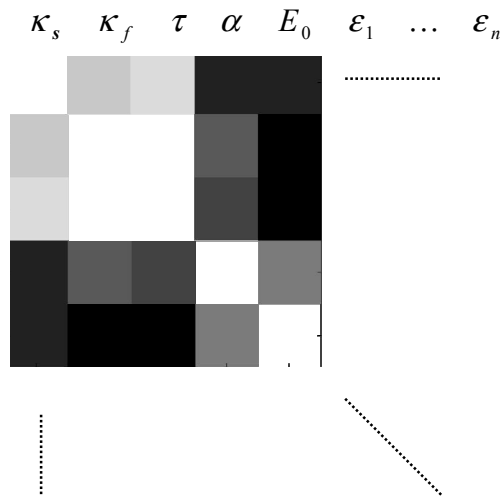
$z = 3\text{mm}$



# The hemodynamic model



### Correlations among parameters

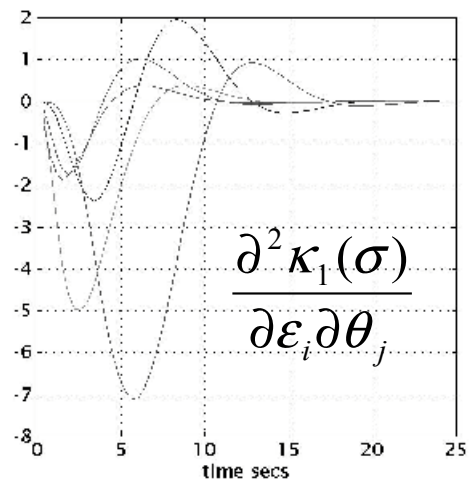
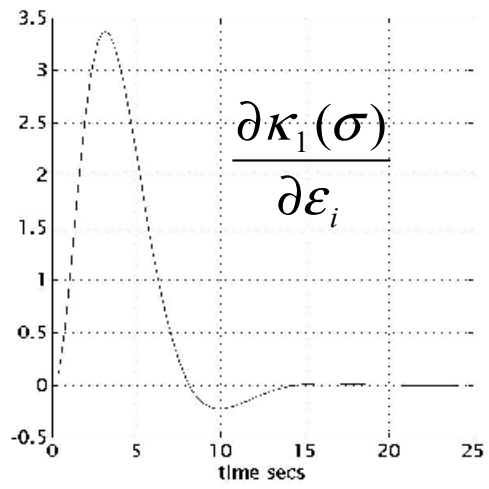


### Covariances

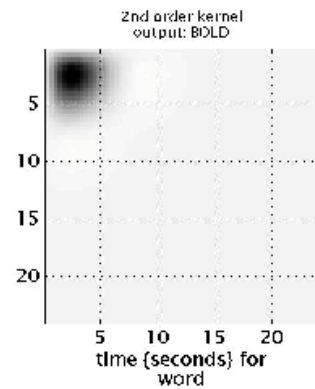
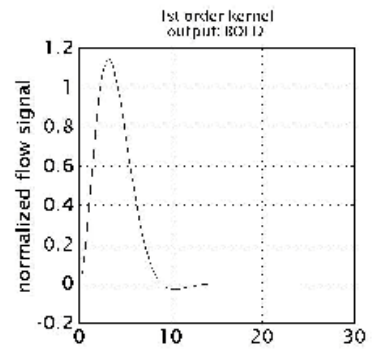
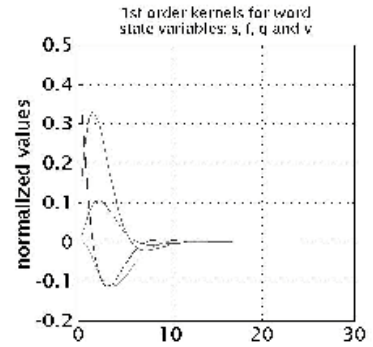
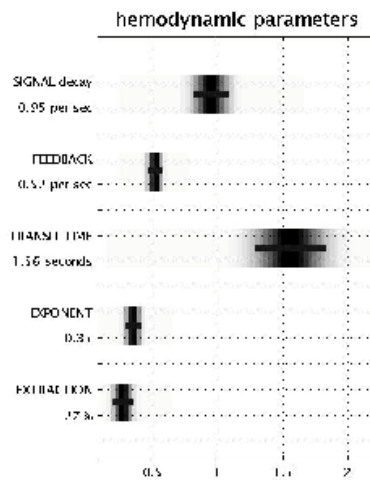
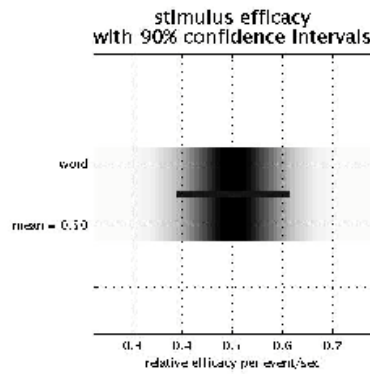
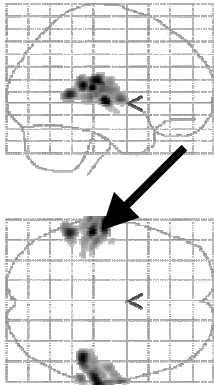
	$\kappa_s$	$\kappa_f$	$\tau$	$\alpha$	$E_0$	$\epsilon_1$	...	$\epsilon_n$
$\kappa_s$	.0150	.0052	.0283	.0002	-.0027	0	...	0
$\kappa_f$	$\vdots$	.0020	.0104	.0004	-.0013			
$\tau$			.0568	.0010	.0069			
$\alpha$				.0013	-.0010			
$E_0$					.0024			
$\epsilon_1$								
...								
$\epsilon_n$								



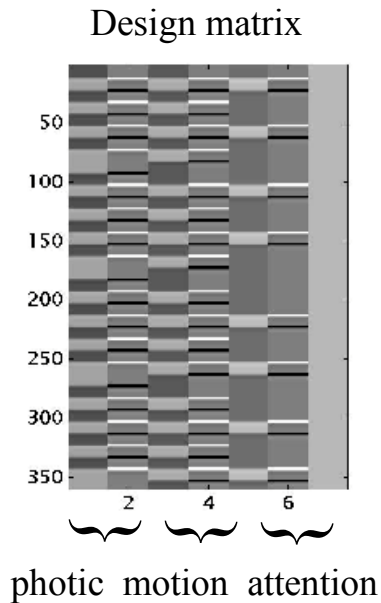
Bayesian estimation



# fMRI study of single word processing at different rates



# Extension to MISO



fMRI study of attention to visual motion

