



ACADEMIC  
PRESS

Available online at [www.sciencedirect.com](http://www.sciencedirect.com)

SCIENCE @ DIRECT®

NeuroImage

NeuroImage 19 (2003) 1240–1249

[www.elsevier.com/locate/ynimg](http://www.elsevier.com/locate/ynimg)

Technical Note

## Posterior probability maps and SPMs

K.J. Friston\* and W. Penny

*The Wellcome Department of Imaging Neuroscience, London, Queen Square, London WC1N 3BG, UK*

Received 15 July 2002; revised 5 February 2003; accepted 14 February 2003

### Abstract

This technical note 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 note we present a brief 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.

© 2003 Elsevier Science (USA). All rights reserved.

*Keywords:* Bayesian inference; Posterior probability maps; EM algorithm; Hierarchical models; Neuroimaging

### Introduction

To date, inference in neuroimaging has been restricted largely to classical inferences based on statistical parametric maps (SPMs). The statistics that comprise these SPMs are essentially functions of the data (Friston et al., 1995). The probability distribution of the chosen statistic, under the null hypothesis (i.e., the null distribution), is used to compute a *P* value. This *P* value is the probability of obtaining the statistic, or the data, given that the null hypothesis is true. If sufficiently small, the null hypothesis can be rejected and an inference is made. The alternative approach is to use Bayesian or conditional inference based upon the posterior distribution of the activation given the data (Holmes and 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

on a posterior density analysis. A useful way to summarize 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. In this technical note we describe an approach to computing posterior probability maps for activation effects, or more generally treatment effects, in imaging data sequences. A more thorough account of this approach can be found in Friston et al. (2002a, 2002b). We focus here on a specific procedure that has been incorporated into the SPM software. This approach represents, probably, the most simple and computationally expedient way of constructing posterior probability maps (PPMs).

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 from zero. The problem for classical inference is that trivial departures from the null hypothesis can be declared significant, with sufficient data or sensitiv-

\* Corresponding author. The Wellcome Department of Imaging Neuroscience, Institute of Neurology, 12 Queen Square, London, WC1N 3BG, UK. Fax: +44-020-7813-1445.

*E-mail address:* [k.friston@fil.ion.ucl.ac.uk](mailto:k.friston@fil.ion.ucl.ac.uk) (K.J. Friston).

ity. From the point of view of neuroimaging, posterior inference is especially useful because it eschews the multiple-comparison problem. In classical inference one tries to ensure that the probability of rejecting the null hypothesis incorrectly is maintained at a small rate, despite making inferences over large volumes of the brain. This induces a multiple-comparison problem that, for continuous spatially extended data, requires an adjustment or correction to the  $P$  values using Gaussian random field theory. This Gaussian field correction means that classical inference becomes less sensitive or powerful with large search volumes. In contrast, posterior inference does not have to contend with the multiple-comparison problem because there are no false positives. The probability that an activation has occurred, given the data, at any particular voxel is the same, irrespective of whether one has analyzed 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).

#### *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 rule. 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 hyperparameters at any particular level can be treated as priors on the level below. There are numerous examples of hierarchical observations models. For example, the distinction between fixed- and mixed-effects analyses of multisubject 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 gray matter). The first level of the hierarchy corresponds to the experimental effects at any particular voxel and the second level of the hierarchy com-

prises 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.

This technical note describes the computation of PPMs that is implemented in our software (SPM2, <http://www.fil.ion.ucl.ac.uk/spm>). The theoretical background, on which this approach is based, was presented in Friston et al. (2002a, 2002b) and the reader is referred to these articles for a full description. The model used here is a special case of the spatiotemporal models described in Section 3 of Friston et al. (2002a). This special case 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 Friston et al., 2002b, 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 [restricted maximum likelihood] ([ReML]) estimates of the variance components required for Bayesian inference. The main advance described in this article 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.

## Theory

### *Conditional estimators and the posterior density*

In this section 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 modeled as

$$y = X\theta + \varepsilon. \quad (1)$$

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

$$C_{\theta|y} = (X^T C_\varepsilon^{-1} X + C_\theta^{-1})^{-1} \quad (2)$$

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

where  $C_\theta$  is the prior covariance and assuming a prior expectation of 0. 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

$$p = 1 - \Phi\left(\frac{\gamma - c^T \eta_{\theta|y}}{\sqrt{c^T C_{\theta|y} c}}\right). \quad (3)$$

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

### Estimating the error covariance with ReML

Clearly, to compute the conditional moments in Eq. (2), 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 that 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 nonsphericity matrix of the errors (see below). This hyperparameter is estimated simply using ReML. ReML estimation involves recursion of the following equations that can be formulated in terms of expectation maximization (EM, Dempster et al., 1977), as described in Friston et al. (2002b)

*Until convergence*

**{E-Step}**

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

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

**M-Step**

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

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

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

$$\lambda_\varepsilon \leftarrow \lambda_\varepsilon + H^{-1}g \quad (4)$$

These equations are presented in the form used in Friston et al. (2002b) and below, to make the connection with EM explicit. See the Appendix of Friston et al. (2002b) for a detailed explanation of this scheme.<sup>1</sup> In brief,  $P$  represents the residual forming matrix, premultiplied 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 thus be regarded as a Fisher scoring scheme that maximizes the ReML objective function. Given that there is only one hyperparameter to estimate this scheme converges very quickly (two to three iterations for a tolerance of  $10^{-6}$ ).

### Estimating the prior density with empirical Bayes

Simply computing the conditional moments using Eq. (2) 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 Eq. (1) 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) modeled by the regressors in  $X_0$ , where  $X = [X_1, X_0]$  and

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

$$\theta_1 = 0 + \varepsilon^{(2)} \quad (5)$$

This model posits that there is a voxelwise 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 1 and zeros 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 0 over the search volume.<sup>2</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 that could be subject to EM (see, for example, Friston et al., 2002a, 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 Eq. (5) to a single-level model

$$y = X_0 \theta_0 + \xi$$

$$\xi = X_1 \varepsilon^{(2)} + \varepsilon^{(1)}. \quad (6)$$

This is simply a rearrangement of Eq. (5) 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.

<sup>1</sup> Note that the augmentation step shown in Fig. 4 of Friston et al. (2002b) is unnecessary because the prior covariance enters explicitly into the conditional covariance.

<sup>2</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.

$$\begin{aligned}
C_\xi &= E\{\xi\xi^T\} = \sum \lambda_k Q_k \\
Q &= \{X_1 Q_1 X_1^T, \dots, X_m Q_m X_m^T, V\} \\
\lambda &= [\lambda_1, \dots, \lambda_m, \lambda_e]^T.
\end{aligned} \tag{7}$$

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, owing to variation in the parameters, is determined by the design matrix  $X$  and  $Q_i$  that model variance components in parameter space.

Eq. (7) affords a computationally expedient way to estimate the prior covariances for the parameters that then enter into Eq. (2) 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 Eq. (7). 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  $\lambda_e$  can be treated as an estimate of the average  $\lambda_e$  over voxels. The hyperparameters are estimated by iterating

*Until convergence*

**{E-Step}**

$$\begin{aligned}
C_\xi &= \sum \lambda_k Q_k \\
C_{\theta_{0y}} &= (X_0^T C_\xi^{-1} X_0)^{-1}
\end{aligned}$$

**M-Step**

$$\begin{aligned}
P &= C_\xi^{-1} - C_\xi^{-1} X_0 C_{\theta_{0y}} X_0^T C_\xi^{-1} \\
g_i &= -\frac{1}{2} \text{tr}\{P Q_i\} + \frac{1}{2} \text{tr}\{P^T Q_i P Y Y^T / n\} \\
H_{ij} &= \frac{1}{2} \text{tr}\{P Q_i P Q_j\} \\
\lambda &\leftarrow \lambda + H^{-1} g
\end{aligned} \tag{8}$$

It can be seen that this has exactly the form as Eq. (4) used for the analysis at each voxel. The differences are (1)  $yy^T$  has been replaced by its sample mean over voxels  $YY^T/n$  and (2) 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  $C_{\theta_0}^{-1} \rightarrow 0$ . (3) Finally, 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. In the absence of confounds the hyperparameters would simply be maximum likelihood (ML) estimates that minimize 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

$$C_\theta = \begin{bmatrix} \sum \lambda_i Q_i & \dots & 0 \\ \vdots & \infty & \vdots \\ 0 & \vdots & \infty \end{bmatrix} \tag{9}$$

We now have a very precise estimate of the prior covariance that can be used to revisit each voxel to compute the conditional or posterior density using Eqs. (2) and Eq. (4). Finally, the conditional moments enter Eq. (3) to give the posterior probability for each voxel (see Fig. 1 for a schematic illustration of this scheme).

### Dealing with multiple error hyperparameters

Above it was assumed that the error covariance could be modeled with a single hyperparameter that scaled a known nonsphericity matrix  $V$ . In many instances the exact form of  $V$  may not be known (e.g., serial correlations in fMRI or heteroscedasticity when using different sorts of data). In this case, nonsphericity among the errors is modeled by more than one covariance basis. That is, the bases and hyperparameters in Eq. (7) are augmented to

$$\begin{aligned}
Q &= \{X_1 Q_1 X_1^T, \dots, X_m Q_m X_m^T, Q_1^e, \dots, Q_l^e\} \\
\lambda &= [\lambda_1, \dots, \lambda_m, \lambda_1^e, \dots, \lambda_l^e]^T.
\end{aligned}$$

Additional bases can model serial correlations or a block diagonal set could accommodate different forms of heteroscedasticity. The nonsphericity matrix can then be computed by renormalizing  $V = \sum \lambda_i^e Q_i^e$  such that  $\text{tr}\{V\} = n$ . This ensures that the voxel-specific error hyperparameters are estimated with high precision by reducing the number of error hyperparameters to 1. However, it does entail the assumption that nonsphericity among the errors has the same structure at all voxels.

### A computational saving

Although it would be possible to save the conditional covariances of the parameters for every voxel, this is quite burdensome, in terms of both memory and time (because one would have to save a matrix for each voxel). A simple approximation to the conditional covariance of the parameters obtains through a first-order Taylor expansion in terms of the voxel-specific error hyperparameter that obtains after iterating Eq. (4).

$$C_{\theta_{0y}}(\lambda_e) \approx C_{\theta_{0y}}(E\{\lambda_e\}) + \frac{\partial C_{\theta_{0y}}}{\partial \lambda_e} (\lambda_e - E\{\lambda_e\}), \tag{10}$$

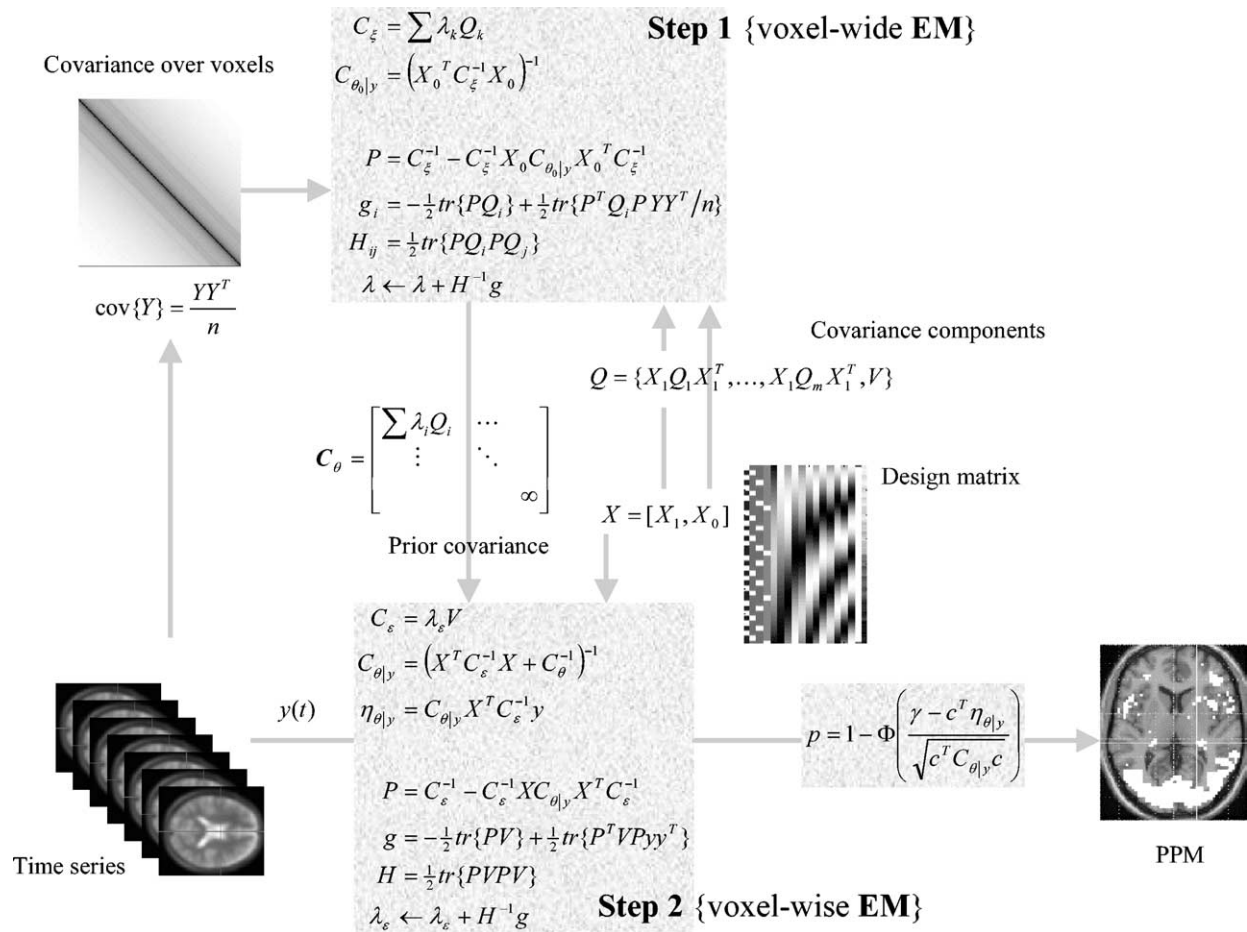


Fig. 1. Schematic summarizing the two-step procedure for (1) ReML estimation of the prior covariance based on the data covariance, pooled over voxels and (2) a voxel-by-voxel estimation of the posterior expectation and covariance of the parameters, required for inference. See the main text for a detailed explanation of the equations.

where the expectations are over voxels. Adopting this first-order approximation means that we only have to save one hyperparameter for each voxel. This speeds up the computation of voxel-specific conditional variances of contrasts (as in Eq. (3)). This first-order approximation provides an upper bound on the variance owing to the concavity of the function relating the conditional covariance to the hyperparameters (see Fig. 2). This upper bound on the variance translates into a conservative lower bound on the posterior probability.

In summary, a natural hierarchy characterizes 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 Fig. 1). 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. For those interested in computing architectures in neuroimaging, this approach lends itself nicely to parallelizing.

The approach conforms to empirical Bayes under parametric assumptions (PEB). PEB is simple [restricted] maximum likelihood estimation in the context of hierarchical models, where the ReML estimators can be viewed as conditional estimators by treating any level a source of priors for its subordinate (cf. Efron and Morris, 1973).

### Applications

In this section we compare and contrast Bayesian and classical inference using PPMs and SPMs based on real

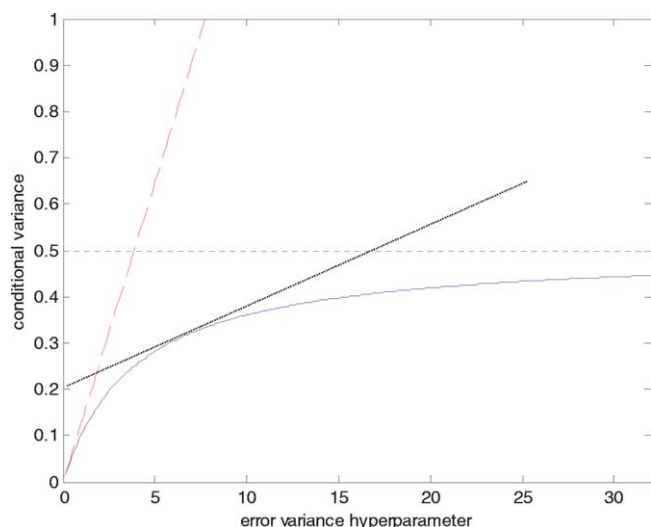


Fig. 2. An example of the relationship between the conditional variance of a parameter estimate and the value of a single hyperparameter governing error variance. It can be seen that the conditional variance of the parameter (or contrast) is always less than it would have been in the absence of priors (dashed line). Critically, when the error variance is high, the conditional variance asymptotes to the prior covariance (here 0.5). Conversely, when the error variance is very low, so is the conditional variance. The important thing is that this relationship is concave, such that a first-order Taylor expansion around any expected value for the hyperparameter will provide an upper bound on the conditional variance (dotted line).

data. The first data are the PET verbal fluency data that have 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 either to 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 was randomized over subjects. The second data set comprised data from a study of attention to visual motion (Büchel and Friston 1997). The data used in this note came from the first subject studied. This subject was scanned at 2 T 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 reanalyzed 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.

### Inference for the PET data

The top panel of Fig. 3 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 Eq. (3). This defines what we mean by “activation” and, by default, is set at one standard 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 must exceed 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 summarizing ones confidence that an effect is present (cf. 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

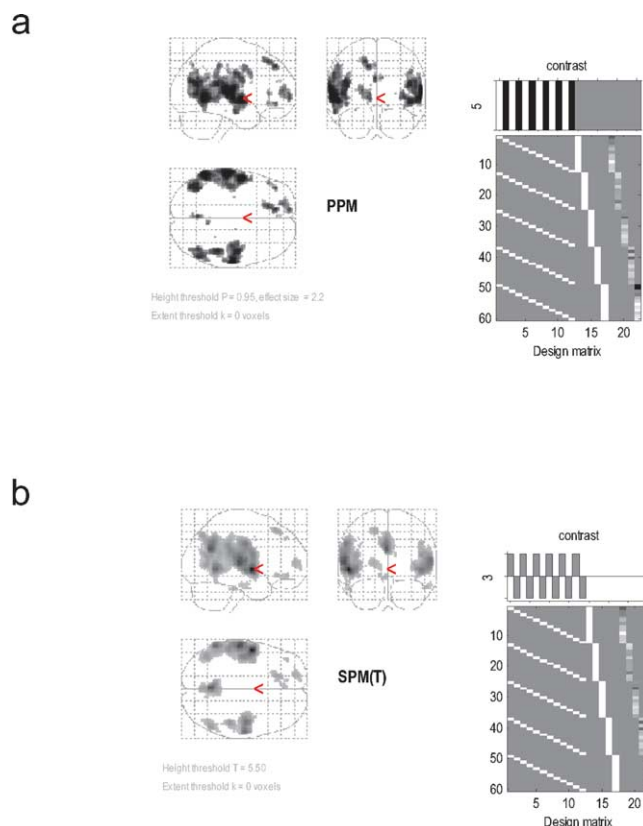


Fig. 3. Bayesian and classical and inference for the PET study of word generation. (a) PPM for a contrast reflecting the difference between word shadowing and word generation, using an activation threshold of 2.2 and a confidence of 95%. The design matrix and contrast for this model are shown (right) in image format. We have modeled each scan as a specific effect that has been replicated over subjects. (b) 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.

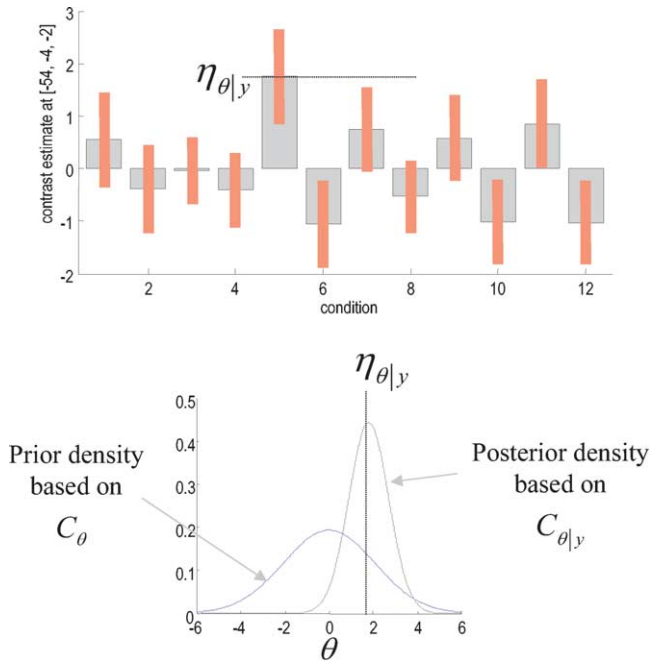


Fig. 4. Illustrative results for a single voxel—the maximum in the left temporal region of the PPM in the previous figure ( $-54, -4, -2\text{mm}$ ). (Top) These are the conditional or posterior expectations and 95% confidence intervals for the activation effect associated with each of the 12 conditions. 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. (Bottom) The prior and posterior densities for the parameter estimate for condition 5.

maps it is useful to remove voxels that fall below some threshold.

Fig. 4 provides a quantitative representation of Bayesian inference afforded by PPMs. In the top panel the posterior expectation for the 12 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 bottom 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 are.

The corresponding SPM is shown in the bottom panel (Fig. 3b). 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 familywise 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 by changing modalities from PET to fMRI. Smoothness controls the “statistical” search volume, which is generally much greater for fMRI than for PET.

#### *Inference for the fMRI data*

The difference between the PPM and SPM for the fMRI analysis is immediately apparent on inspection of Fig. 5 and 6. Here the default threshold for the PPM was 0.7% (equivalent to percentage of 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 bottom panel of Fig. 5). 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.

#### *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 labeling 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.

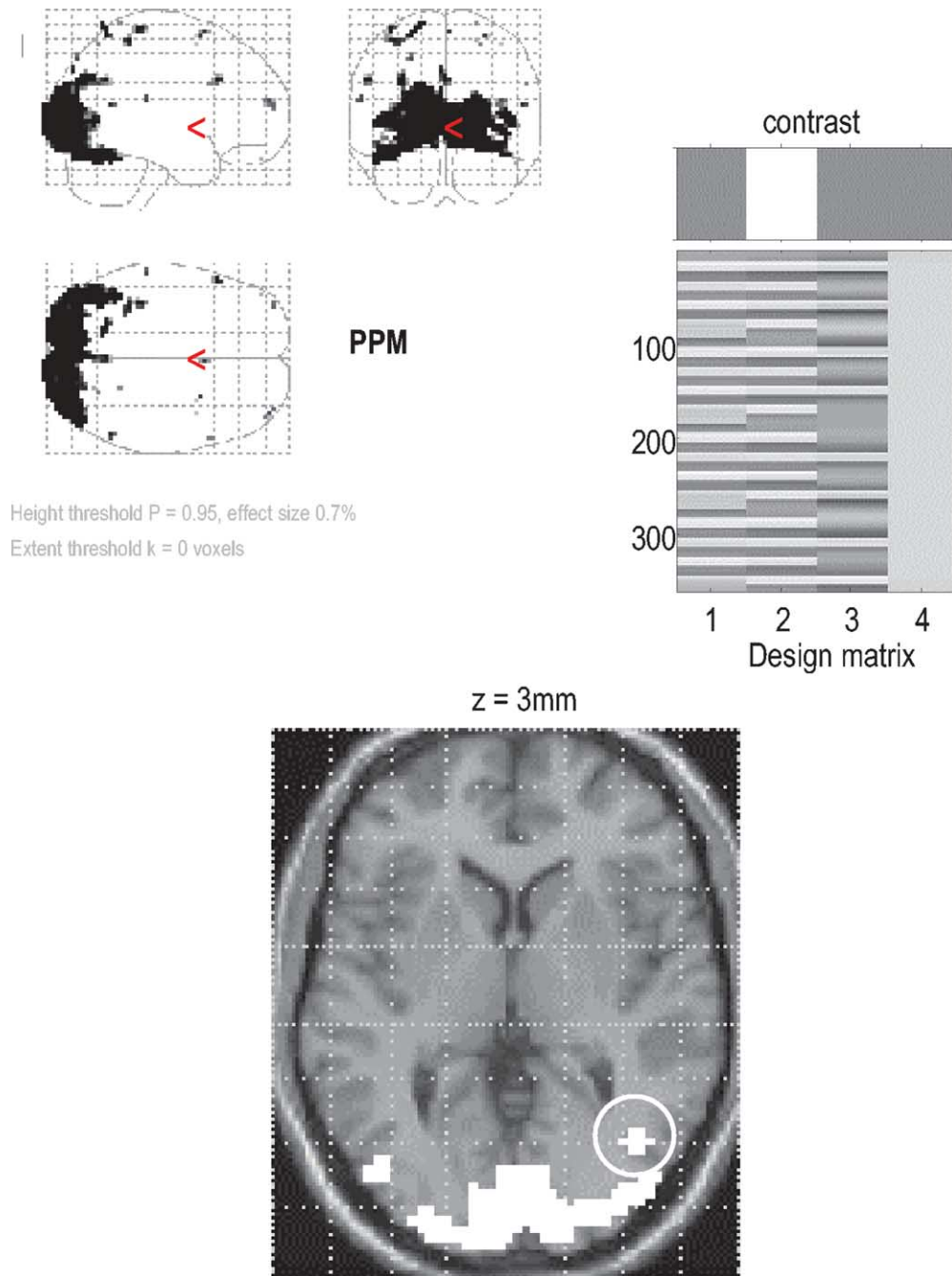


Fig. 5. PPM for the fMRI study of attention to visual motion. The display format in the bottom panel uses an axial slice through extrastrate regions but the thresholds are the same as employed in maximum intensity projections (top panels). The activation threshold for the PPM was 0.7. As can be imputed from the design matrix, the statistical model of evoked responses comprised boxcar regressors convolved with a canonical hemodynamic response function.

## Conclusion

In this note we have presented 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).

A key consideration, in the use of empirical Bayes in this



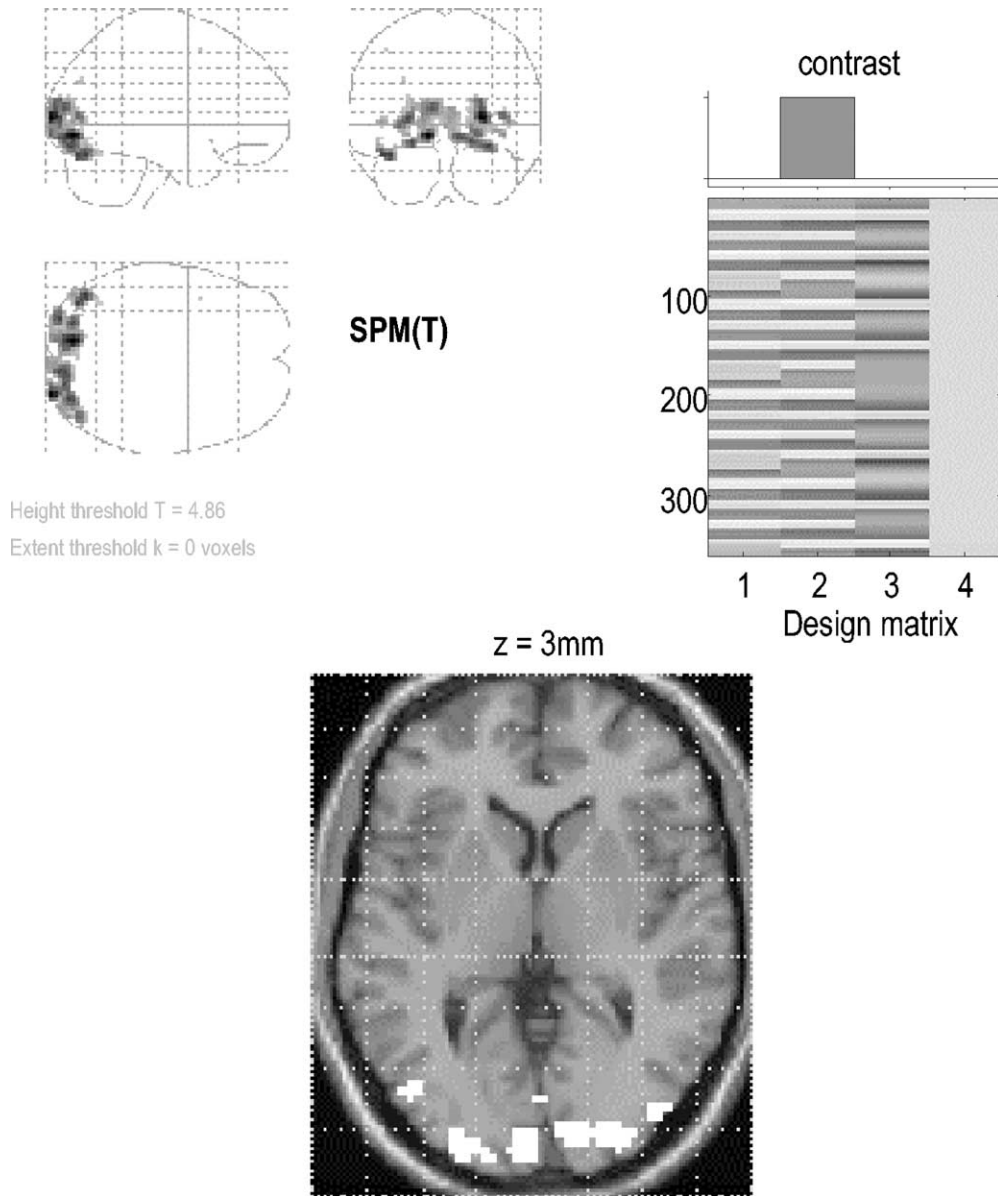


Fig. 6. As for Fig. 5, but this time showing the corresponding SPM using a corrected threshold at  $P = 0.05$ .

setting is “which voxels to include in the hierarchy?” There is no right or wrong answer here (cf. 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 that minimize 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 extracranial regions).

The theory described in this technical note has been implemented in the current version of the SPM software

(SPM2). Simple posterior inferences that are enabled by this approach may find a useful role in characterizing evoked brain responses or differences among cohorts.

**Acknowledgments**

This work was funded by the Wellcome Trust. We thank Marcia Bennett for preparing the manuscript.

**References**

Berry, D.A., Hochberg, Y., 1999. Bayesian perspectives on multiple comparisons. *J. Stat. Plan. Infer.* 82, 215–227.

- Büchel, C., Friston, K.J., 1997. Modulation of connectivity in visual pathways by attention: cortical interactions evaluated with structural equation modeling and fMRI. *Cereb. Cortex* 7, 768–778.
- Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B* 39, 1–38.
- Efron, B., Morris, C., 1973. Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Am. Stat. Assoc.* 68, 117–130.
- Friston, K.J., Glaser, D.E., Henson, R.N.A., Kiebel, S., Phillips, C., Ashburner, J., 2002a. Classical and Bayesian inference in neuroimaging: applications. *NeuroImage* 16, 484–512.
- Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.-B., Frith, C.D., Frackowiak, R.S.J., 1995. Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Map.* 2, 189–210.
- Friston, K.J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., Ashburner, J., 2002b. Classical and Bayesian inference in neuroimaging: theory. *NeuroImage* 16, 465–483.
- Holmes, A., Ford, I., 1993. A Bayesian approach to significance testing for statistic images from PET, in: Uemura, K., Lassen, N.A., Jones, T., Kanno, I. (Eds.), *Quantification of Brain Function: Tracer Kinetics and Image Analysis in Brain PET*, Int. Cong. Series No., 1030. Excerpta Medica, Almere, the Netherlands, pp. 521–534.
- Phillips, C., Mattout, J., Rugg, M.D., Pierre Maquet, P., Friston, K.J., 2003. Restricted maximum likelihood solution of the source localization problem in EEG. *NeuroImage*, in press abstract.