NeuroImage **16,** 484–512 (2002)
doi:10.1006/nimg.2002.1091, available online at http://www.idealibrary.com on IDEAL®

Classical and Bayesian Inference in Neuroimaging: Applications

K. J. Friston, D. E. Glaser, R. N. A. Henson, S. Kiebel, C. Phillips, and J. Ashburner

The Wellcome Department of Cognitive Neurology and The Institute of Cognitive Neuroscience, University College London, Queen Square, London WC1N 3BG United Kingdom

Received January 11, 2001

In Friston *et al.* ((2002) *Neuroimage* 16: 465–483) we introduced empirical Bayes as a potentially useful way to estimate and make inferences about effects in hierarchical models. In this paper we present a series of models that exemplify the diversity of problems that can be addressed within this framework. 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 the expectation–maximization (EM) algorithm in covariance component estimation we focus first on two important problems in fMRI: nonsphericity induced by (i) serial or temporal correlations among errors and (ii) variance components caused by the hierarchical nature of multisubject 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: (i) The difference between response estimates based on ML and the conditional means from a Bayesian approach and the implications for estimates of intersubject variability. (ii) The relationship between fixed- and random-effect analyses. (iii) The specificity and sensitivity of Bayesian inference and, finally, (iv) the relative importance of the number of scans and subjects. The forgoing is concerned with within- and between-subject variability in multisubject hierarchical fMRI studies. In the second half of this paper we turn to Bayesian inference at the first (within-voxel) level, using PET data to show how priors can be derived from the (between-voxel) distribution of activations over the brain. This application uses exactly the same ideas and formalism but, in this instance, the second level is provided by observations over voxels as opposed to subjects. The ensuing posterior probability maps (PPMs) have enhanced anatomical precision and greater face validity, in relation to underlying anatomy. Furthermore, in comparison to conventional SPMs they are not confounded by the multiple comparison problem that, in a classical context, dictates high thresholds and low sensitivity. We conclude with some general comments on Bayesian approaches to image analysis and on some unresolved issues. © 2002 Elsevier Science (USA)

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

INTRODUCTION

In Friston *et al.* (2002) we reviewed empirical Bayesian approaches that might find a role in neuroimaging. 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 generally 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, noniterative 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 parameterized with a small number of hyperparameters (c.f. serial correlations in fMRI time-series). Another important example, in fMRI, is the multisubject design in which the hierarchical nature of the observation induces different variance components at each level. The aims of the first sections in this paper are to illustrate how variance component estimation, with EM, can proceed in both single-level and hierarchical contexts. Second, we wanted to show how the variance components in supraordinate levels can be used to give Bayesian estimators of effects at lower levels. In particular, the examples emphasize 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.

The fMRI examples are presented in two sections. In the first we deal with the issue of variance component estimation using serial correlations in single-subject

1053-8119/02 $35.00

fMRI studies. In the second section we use a multisubject fMRI study to address intersubject variability by adding a second level to the observation model presented in the first section. 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. The notation and terms used in this paper follow Friston *et al.* (2002).

1. 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 nonsphericity in fMRI time-series. This section concludes with a quantitative assessment of serial correlations within and between subjects.

1.1 Serial Correlations in fMRI

In this section we restrict ourselves to a single-level model and focus on the covariance component estimation offered by the EM algorithm. We have elected to use an important covariance estimation problem to illustrate one of the potential uses of the scheme described in Friston *et al.* (2002). Namely serial correlations in fMRI embodied in the error covariance matrix for the first (and only) level of this observation model $C_\epsilon^{(1)}$ (as in the previous paper the superscript signifies the hierarchical level in question). Serial correlations have a long history in the analysis of fMRI time-series and are still the subject of current work: fMRI time-series can be viewed as a linear admixture of signal and noise. Signal corresponds to neuronally mediated hemodynamic changes that can be modeled as a [non-] linear convolution of some underlying neuronal or synaptic process, responding to changes in experimental factors, by a hemodynamic response functions (HRF). Noise has many contributions that render it rather complicated in relation to other neurophysiological measurements. These include neuronal and nonneuronal sources. Neuronal noise refers to neurogenic signal not modeled by the explanatory variables and has the same frequency structure as the signal itself. Nonneuronal components have both white (e.g., R.F. noise) and colored 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 (e.g., aliased cardiac-locked pulsatile motion) and induce long range correlations in the errors over time. Currently there are two approaches to serial correlations of this sort: (i) The data are filtered with a specified filter to impose a known correlation structure on the errors and are then entered into a generalized least squares scheme as described in Worsley and Friston (1995). (ii) The correlations are estimated, in some fashion, and these estimates are used to give minimum variance or Gauss-Markov estimators (see Eq. (15) in Friston *et al.*, 2002) (e.g., Purdon and Weisskoff, 1998). These are equivalent to ordinary least squares (OLS) estimators based on pre-whitened data (Bullmore *et al.*, 1996). The second approach, in principle, is more efficient but depends upon an accurate estimation of the serial correlations and inversion of the estimated correlation matrix. The first approach eschews this inversion and, more importantly, bias in the estimate of the standard error that ensues from a mismatch between the true and estimated correlations. However, it does so at the expense of efficiency (see Friston *et al.*, 2000, for details). It would be nice to be estimate the serial correlations directly from the data and use the Gauss–Markov estimators but there is a fundamental problem. In order to estimate correlations among the errors $C(\lambda)_\epsilon$, in terms of some hyperparameters λ, one needs both the residuals of the model r and the covariance of the parameter estimates that produced those residuals. These combine to give the required error covariance (c.f. Eq. (A.4) in Friston *et al.*, 2002).

$$C(\lambda)_\epsilon = rr^T + XC_{\theta|y}X^T \tag{1}$$

$XC_{\theta|y}X^T$ 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 reestimation implicit in the EM algorithm. 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 (1), which accounts for the component of error covariance due to the inherent variability of the parameter estimates themselves. It is likely that any valid recursive scheme for estimating serial correlations in fMRI time-series will conform to EM (or ReML) even if the connection is not made explicit. See Worsley *et al.* (2002) for a clever noniterative 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.* (2002) 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 section concludes with an application of EM to empirical data to demonstrate quantitatively the relative variability in serial correlations over voxels and subjects.

1.2 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.

$$\begin{aligned} y &= X^{(1)}\theta^{(1)} + \epsilon^{(1)} \\ Q_1^{(1)} &= I \\ Q_2^{(1)} &= KK^T, \quad k_{ij} = \begin{cases} e^{j-i} & i > j \\ 0 & i \le j \end{cases} \end{aligned} \tag{2}$$

y is the measured response with errors $\epsilon^{(1)} \sim N\{0, C_\epsilon^{(1)}\}$. I is the identity matrix.[1] Here $Q_1^{(1)}$ and $Q_2^{(1)}$ represent covariance components of $C_\epsilon^{(1)}$ 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 modeled as an exponential decay of correlations over nonzero 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. Nonstationary autocorrelations are modeled 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 temporal frequency structure of errors may change with peristimulus time.

In the examples below the covariance constraints were scaled to a maximum of one. This means that the second hyperparameter can be interpreted directly as the covariance between one scan and the next. The basis set enters, along with the data, into the EM algorithm 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 Fig. 1. 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 Fig. 1 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 solid 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) autocorrelation function based on $C_\epsilon^{(1)} = \lambda_1^{(1)}Q_1^{(1)} + \lambda_2^{(1)}Q_2^{(1)}$. 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 variability of 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 $C_\epsilon^{(1)}$ but $RC_\epsilon^{(1)}R^T$, where R is the residual-forming matrix (see Eq. (5)). 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 Fig. 1 legend.

1.3 Inference in the Context of Nonsphericity[2]

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. Because this is a single-level model, classical and empirical Bayes are the same and the standardized conditional mean reduces to a classical T statistic based on the Gauss-Markov estimator (GM/scm in Fig. 2) (see Section 2.6 in Friston *et al.*, 2002). This T statistic can be compared to that which would have been obtained if we had ignored the serial correlations (i.e., had not included the second covariance constraint modeling the AR process above). In this instance the conditional covariance reduces to the standard error of the OLS estimate because there are no priors and the errors are assumed to be i.i.d. In the context of serial

[1] In this and the subsequent paper I_m will represent the identity matrix of size $m \times m$. Similarly we will denote a $m \times 1$ vector of zeros by 0_m and a vector of ones by 1_m.

[2] A Gaussian i.i.d. process is identically and independently distributed and has a probability distribution whose isocontours conform to a sphere. Any departure from this is referred to as nonsphericity.

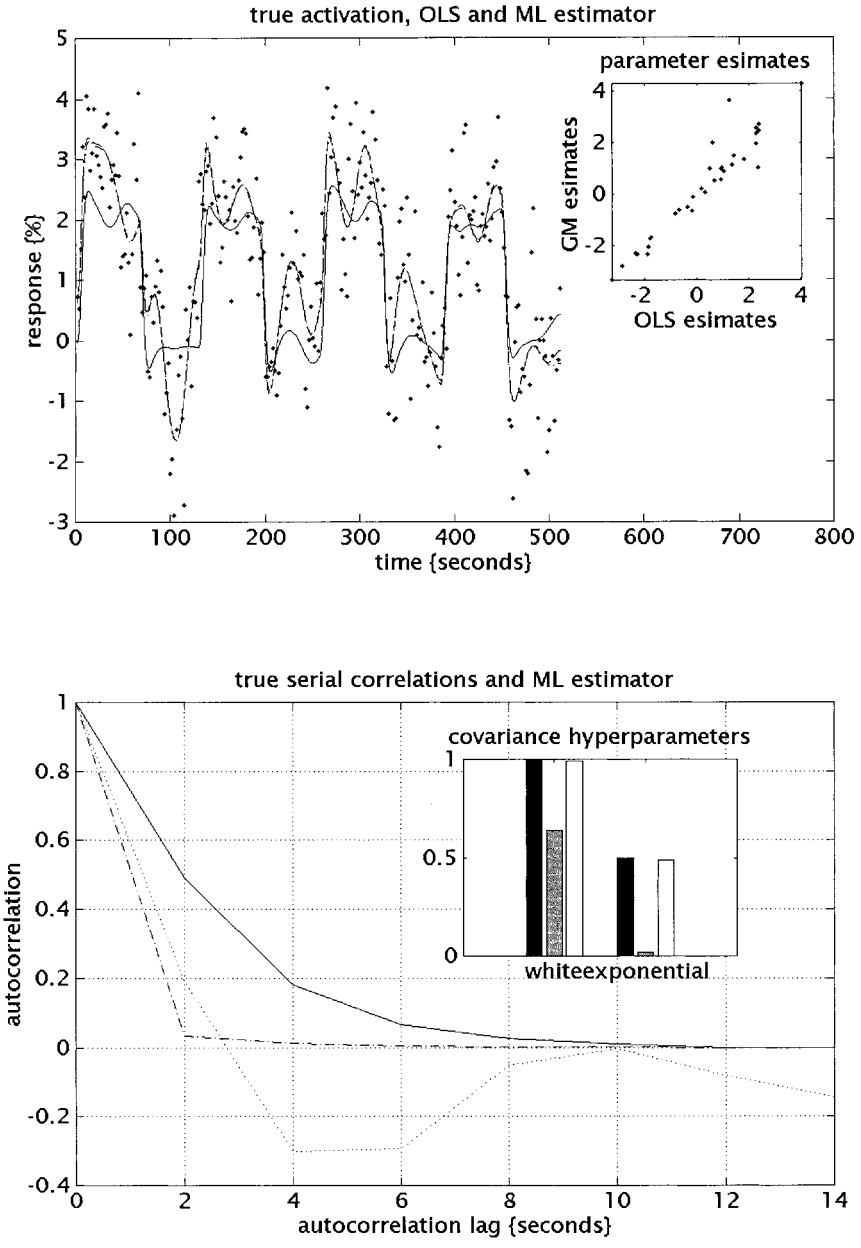


FIG. 1. Top panel: True response (an activation plus random low frequency components) and that based on the OLS and ML estimators for a simulated fMRI experiment. The insert shows the similarity between the OLS and ML predictions. Lower panel: True (dashed) and estimated (solid) autocorrelation functions. 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 interscan interval of 2 s. The activations were modeled with a box-car (duty cycle 64 s) 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].

correlation's the OLS standard error is biased, leading to an underestimate and an inflated T value (i.i.d. in Fig. 2). The difference between the two T values is due to a departure from i.i.d. assumptions that is not accommodated by the OLS scheme. This departure is referred to as nonsphericity in analysis of variance and usually calls for some nonsphericity correction.

The impact of serial correlations on inference was noted early in the fMRI analysis literature (Friston *et al.*, 1994) and led to the generalized least squares

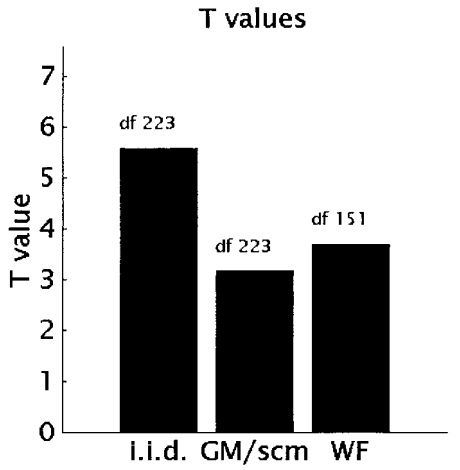


FIG. 2. *T* statistics based on the simulated data of Fig. 1. These corresponding to; (i) the standardized ordinary least square estimator under i.i.d. assumptions about the error terms (i.i.d.). The standardized conditional mean which is equivalent to the standardized Gauss–Markov estimator obtained using the correlation structure estimated by the EM algorithm (GM/scm) and (iii) the standardized generalized least squares estimator of Worsley and Friston (1995) using the same correlation estimate (WF). All are valid apart from the naive i.i.d. scheme that ignores serial correlations or non-sphericity among the errors. The degrees of freedom (d.f.) are provided for each statistic and were calculated according to (5).

(GLS) scheme described in Worsley and Friston (1995). In this scheme one starts with any observation model that is premultiplied by some weighting or convolution matrix S to give

$$Sy = SX^{(1)}\theta^{(1)} + S\epsilon^{(1)} \tag{3}$$

The GLS parameter estimates and their covariance are

$$\begin{aligned} \eta_{LS} &= Ly \\ \mathrm{Cov}\{\eta_{LS}\} &= LC_{\epsilon}^{(1)}L^{T} \\ L &= (X^{(1)T}VX^{(1)})^{-1}X^{(1)}V, \end{aligned} \tag{4}$$

where $V = S^{T}S$ represents the correlations induced by S. These estimators minimize the generalized least square index $(y - X^{(1)}\eta_{LS})^{T}V(y - X^{(1)}\eta_{LS})$. This family of estimators are unbiased but not necessarily ML estimates. The Gauss–Markov estimator is the minimum variance and ML estimator that obtains when $V = C_{\epsilon}^{(1)-1}$. From Eq. (3) it can be seen that this special case is the same as whitening the data with a decorrelating convolution matrix and then using an OLS estimator. Usually one would use the ML estimator. However, there are some situations where a GLS estimator is more practical. For example, when using the same design matrix and filtering for every voxel, the ensuing estimators are GLS if the serial correlations at each voxel are slightly different precluding an exact ML estimate at any single voxel.[3]

The T statistic corresponding to the GLS estimator is distributed with v degrees of freedom where (Worsley and Friston, 1995)

$$\begin{aligned} T &= \frac{c^{T}\eta_{LS}}{\sqrt{c^{T}\mathrm{Cov}\{\eta_{LS}\}c}} \\ v &= \frac{\mathrm{tr}\{RSC_{\epsilon}^{(1)}S\}^{2}}{\mathrm{tr}\{RSC_{\epsilon}^{(1)}SRSC_{\epsilon}^{(1)}S\}} \\ R &= 1 - X^{(1)}L. \end{aligned} \tag{5}$$

The effective degrees of freedom are based on an approximation due to Satterthwaite (1941). This formulation is formally identical to the nonsphericity correction elaborated by Box (1954), which is commonly known as the Geisser–Greenhouse correction in classical analysis of variance, ANOVA (Geisser and Greenhouse, 1958). Note that in Fig. 2 the T statistic based on the GLS estimator (WF) is slightly bigger than the standardized conditional mean, but has a null distribution with fewer degrees of freedom.

The key point here is that EM can be employed to give ReML estimates of correlations among the errors that enter into (5) to enable classical inference, properly adjusted for nonsphericity, about any GLS estimator. The inference is exact, to the extent that the Satterthwaite approximation holds, and is the basis of the Geisser–Greenhouse correction in ANOVA and v the effective degrees of freedom in Worsley and Friston (1995). The ensuing T statistic for the simulated fMRI data, for $S = 1$ (i.e., no temporal filtering), is shown in Fig. 2.

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 Eq. (A.7) in

[3] As discussed in Friston *et al.* (2000) the GLS scheme was originally introduced to ensure robust estimates of variance in the face of unknown serial correlations. In that paper, the hyperparameter was estimated using the filtered residuals

$$\lambda = \frac{\mathrm{tr}\{SRyy^{T}R^{T}S^{T}\}}{\mathrm{tr}\{SRQ^{(1)}R^{T}S^{T}\}} = \frac{y^{T}R^{T}S^{T}SRy}{\mathrm{tr}\{RSQ^{(1)}S\}}$$

This estimate is generally more robust to misspecification of the form of serial correlations $Q^{(1)}$ if a suitable form for S is adopted. One mild but obvious misspecification is when the same correlation structure is assumed for every voxel. When $S = C_{\epsilon}^{-1/2}$ this is the ReML estimator (see footnote 3 in Friston *et al.*, 2002). In Friston *et al.* (2000) the residual forming matrix $\mathbf{R}$ is related to the current notation by $SR = \mathbf{R}S$.

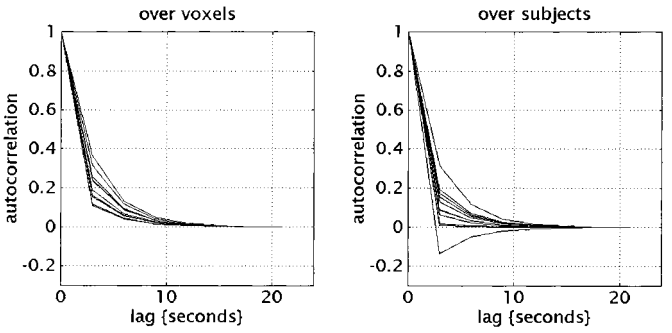


FIG. 3. Estimates of serial correlations expressed as autocorrelation functions based on empirical data. Left panel: Estimates from 12 randomly selected voxels from a single subject. Right panel: Estimates from the same voxel over 12 different subjects. The voxel was in the cingulate gyrus and came from same slice reported in Section 2. The empirical data are described in Henson *et al.* (2000). They comprised 300 vol, acquired with EPI at two Tesla and a TR of 3 s. 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 s (SOA varied randomly between 2.5 and 5.5 s). Subjects were required to make an old vs new judgment 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).

Friston *et al.*, 2002). After renormalization, the ensuing estimate of the nonsphericity $Q^{(1)} = \sum \lambda_k Q_k^{(1)}$ 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 (see footnote 3 of Friston *et al.*, 2002, and this paper). This is very convenient from a computational perspective. This device is not limited to serial correlations in fMRI but can be applied in any context where nonsphericity is an issue. We will pursue this in a subsequent paper dealing with nonsphericity correction in multistage analyses of fMRI data.

1.4 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 (see Section 2.3 for a discussion of between-subject error variance per se).

Using the model specification in (2) serial correlations were estimated using EM in 12 randomly selected voxels from the same slice from a single subject. The results are shown in Fig. 3 (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 prescanning session) were presented in a random order with a stimulus onset asynchrony (SOA) of about 4 s. These data will be used again in the next section. 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 Fig. 3 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.

1.5 Summary

These results are provided to illustrate one potential application of covariance component estimation, not to provide an exhaustive characterization 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 reestimation 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.

2. VARIANCE COMPONENT ESTIMATION IN fMRI: TWO-LEVEL MODELS

In this section we augment the model of the previous section with a second level. This engenders a number of important issues, including (i) the distinction between fixed- and random-effect inferences about the subjects' responses, (ii) the opportunity to make Bayesian inferences about single-subject responses and (iii) the role of variance component estimation in power analyses of classical inference at the second level. As in previous sections we start with model specification, proceed to simulated data and conclude with an empirical example. In this section 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 empirical analyses are used to assess, quantitatively, the sensitivity and specificity of Bayesian inference at the first level and classical inference at the second. The fMRI data are the same as used in section 1 and comprise event-related time-series from 12 subjects. We chose a data set that would be difficult to analyze 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.

2.1 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 section to estimate serial correlations. Here it embodies a second level that accommodates observations over subjects.

$$\text{level one} \qquad y = X^{(1)}\theta^{(1)} + \epsilon^{(1)}$$

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

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

$$Q_{s+1}^{(1)} = \begin{bmatrix} KK^T & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}, \ldots, \qquad (6)$$

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

$$\text{level two} \qquad \theta^{(1)} = X^{(2)}\theta^{(2)} + \epsilon^{(2)}$$

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

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

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

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 $A_{ij}B$. An example of these design matrices and covariance constraints were shown, respectively, in Figs. 1 and 3 of Friston *et al.* (2002). 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.

2.2 Simulations

In the simulations we used 128 scans for each of 12 subjects. The design matrix comprised three effects, modeling 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 modeled with two regressors, constructed by convolving a series of delta functions with a canonical hemodynamic response function (HRF)[4] and the same function delayed by 3 s. 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 constituted three columns giving a total of 36 parameters at the first level and three at the second. 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 (6) 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 4 and 5 show the results after entering the simulated data into the EM algorithm to estimate

[4] The canonical HRF was the same as that employed by SPM. It comprises a mixture of two gamma variates modeling peak and undershoot components and is based on a principal component analysis of empirically determined hemodynamic responses, over voxels, as described in Friston *et al.* (1998).

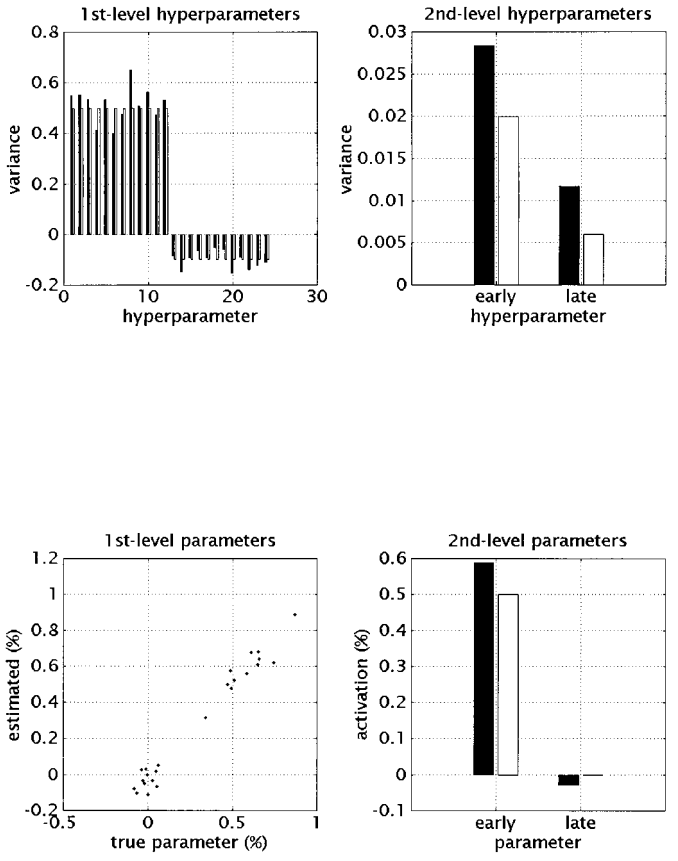


FIG. 4. The results of an analysis of simulated event-related responses in a single voxel. Parameter and hyperparameter estimates based on a simulated fMRI study are shown in relation to the true values. The simulated data comprised 128 scans for each of 12 subjects with a mean peak response over subjects 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 modeled as in Section 1. 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.

the conditional mean and covariances of the subject-specific evoked responses. Figure 4 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 experimental variance to an early effect. Figure 4 should be compared with Fig. 7 that shows the equivalent estimates for real data.

The top panel in Fig. 5 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 Friston *et al.* (2002). First, the parameter estimates and ensuing response estimates are informed by information abstracted from higher levels. Secondly this prior information enables Bayesian inference about the probability of an activation that is specified in neurobiologically meaningful terms.

In Fig. 5 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 characterizations of intersubject variability using ML approaches can severely overestimate the true variability. This is because the ML estimates are unconstrained and simply minimize observation error without considering how likely the ensuing intersubject variability is.

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

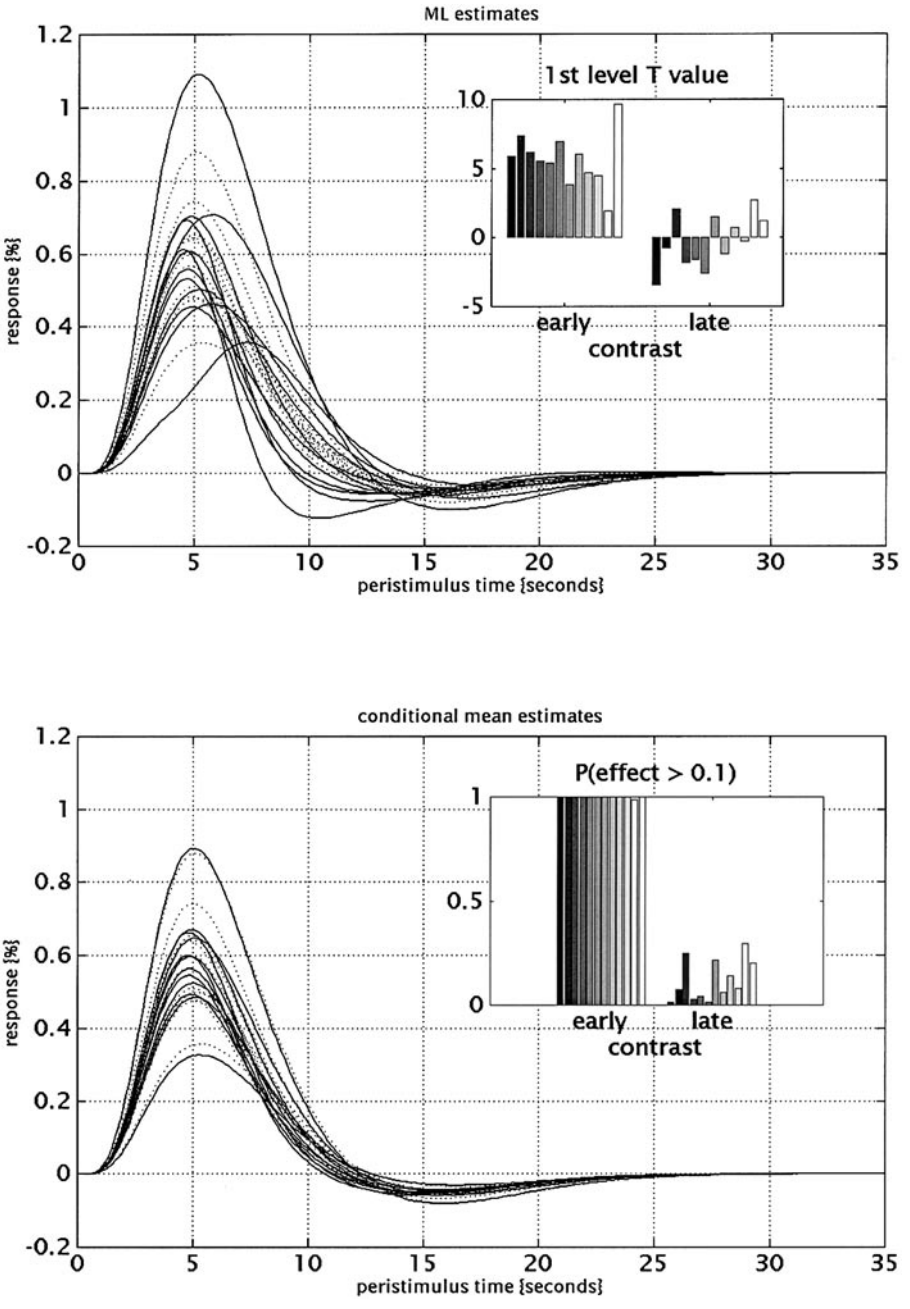


FIG. 5. Response estimates and inferences about the estimates described in the legend of Fig. 4: Upper panel: True (dotted) and ML (solid) estimates of event-related responses to a stimulus over 12 subjects. The units of activation are adimensional and correspond to percent of whole brain mean. The insert 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 intersubject variability in responses. The insert shows the posterior probability that the activation was greater than 0.1%. Because the responses were modeled with early and late components (basis functions corresponding to canonical hemodynamic response functions, separated by 3 s) separate posterior probabilities could be computed for each. The simulated data comprised only early responses as reflected in the posterior probabilities.

$$1 - \Phi\left(\frac{\gamma - c_j^T \eta_{\theta|y}^{(1)}}{\sqrt{c_j^T C_{\theta|y}^{(1)} c_j}}\right) \tag{7}$$

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 component being greater than γ 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.

2.3 Classical Fixed- and Random-Effect Analyses

In this subsection we focus on the importance of covariance component estimation from a classical perspective, in particular the differences in classical inference using fixed (single-level) and random (two-level) effect analyses. In this two-level model the variance partitioning is

$$E\{yy^T\} = \underbrace{C_\epsilon^{(1)}}_{error} + \underbrace{X^{(1)}C_\epsilon^{(2)}X^{(1)T}}_{2nd\text{-}level\ random\ effects} + \underbrace{X^{(1)}X^{(2)}\theta^{(2)}\theta^{(2)T}X^{(2)T}X^{(1)T}}_{fixed\ effects} \quad (8)$$

The first term on the right is simply observation error. The second term corresponds to variance in the response variable due to between-subject variability in the parameters that is projected down to the observation space by the first-level design matrix. The final term corresponds to the sum of squares due the fixed effects at the second level (i.e., the mean effects over subjects) projected down by the design matrices at both levels. Observation error corresponds to within-subject error, second-level random effects correspond to between-subject error and the fixed effects to the variance attributable to activations per se.

What implications does this variance partitioning have for the classical inference? Recall from equation (9) in Friston *et al.* (2002) that the T statistic can be expressed in terms of error variances at all levels specified. For this two-level model the random effects T statistic is

$$\begin{aligned} T^{(2)} &= c^T M^{(2)} y / \sqrt{c^T M^{(2)} C_{\tilde{\epsilon}} M^{(2)T} c} \\ C_{\tilde{\epsilon}} &= C_\epsilon^{(1)} + X^{(1)} C_\epsilon^{(2)} X^{(1)T}, \end{aligned} \quad (9)$$

where $M^{(2)}$ is the ML projector as defined in (7) in Friston *et al.* (2002). The second expression is the combined error in the observation that contributes to the standard error of the contrast. It comprises within-subject error and between-subject error projected by the first-level design matrix onto the observation space (i.e., random effects). Their relative contributions to the standard error of the T statistic can be understood in terms of the difference between random and fixed-effect inference in classical analyses:

Pretend that we had only specified our model to the first level. To test for the mean activation we would have to augment c to average over subjects giving $c^{(1)T} = c^T X^{(2)+}$ (+ denotes pseudoinverse). Here the second-level design matrix enters, not as a constraint on the parameter expectations but directly into the contrast at the first level. The corresponding fixed effects T statistic is

$$\begin{aligned} T^{(1)} &= c^T X^{(2)+} M^{(1)} y / \sqrt{c^T X^{(2)+} M^{(1)} C_{\tilde{\epsilon}} M^{(1)T} X^{(2)+T} c} \\ C_{\tilde{\epsilon}} &= C_\epsilon^{(1)}. \end{aligned} \quad (10)$$

For balanced designs with equal variances, $X^{(2)+}M^{(1)} = M^{(2)}$. In this case the only difference between the two T statistics is the contribution of the random effects $X^{(1)}C_\epsilon^{(2)}X^{(1)T}$ to the standard error. This contribution will be large when (i) the second-level error is big or (ii) when the first-level design matrix amplifies its projection onto the observation space, i.e., many observations at the first level, relative to the second. This accounts for the well-known fact that the distinction between the fixed and random-effect T statistics is greater when the within-subject error is small relative to between-subject error and when there are many repeated measures per subject. This is why random effect analyses are more critical in fMRI than in PET. In fMRI the scan to scan variability is much smaller than in PET and typically there are many more scans per session in fMRI.

These points are illustrated in Fig. 6 where the T statistics were computed according to (9) and (10) using the estimated parameters and hyperparameters from the simulation. In the upper panel the first-level design matrix was decimated to reduce the number of scans per subject. It can be seen that at around 16 scans the fixed (broken line) and random-effect (solid line) T statistics converge, whereas there is a substantial difference by 32 scans. In the lower panel the error covariance was scaled while keeping the design matrices constant. In this illustration, as the error covariance falls to zero the fixed-effect T statistic tends to infinity whereas the random-effect T statistic properly reflects the intrinsic between-subject variability.

In summary, covariance component estimation is critical for inference in hierarchical models because mixtures of variance components are required to compute the standard error of contrasts. In random-effect analyses intersubject differences are treated as a variance component, rendering the inference about a contrast of effects relative to that contrast's inherent variability. In fixed-effect analyses this variance component is discounted and the inference is in relation to the precision with which the effect can be measured.

2.4 Empirical Analyses

Here the analysis is repeated using real data and the results compared to those obtained using simulated data. We focus first on the parameter and hyperparameter estimates and how these are used to form posterior probability maps or PPMs. We then use the covariance component estimates to quantify the specificity and sensitivity of Bayesian inference at the first level, and classical inference at the second. 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 s or so (SOA varied between 2.5 and 5.5 s). 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 judgment for each word. Drift (the first 8 components of a discrete cosine set) and the effects of incorrect trials were treated as confounds and were removed using linear regression.[5] The first-level subject-specific design matrix partitions comprised four regressors with early and late effects for both old and new words.

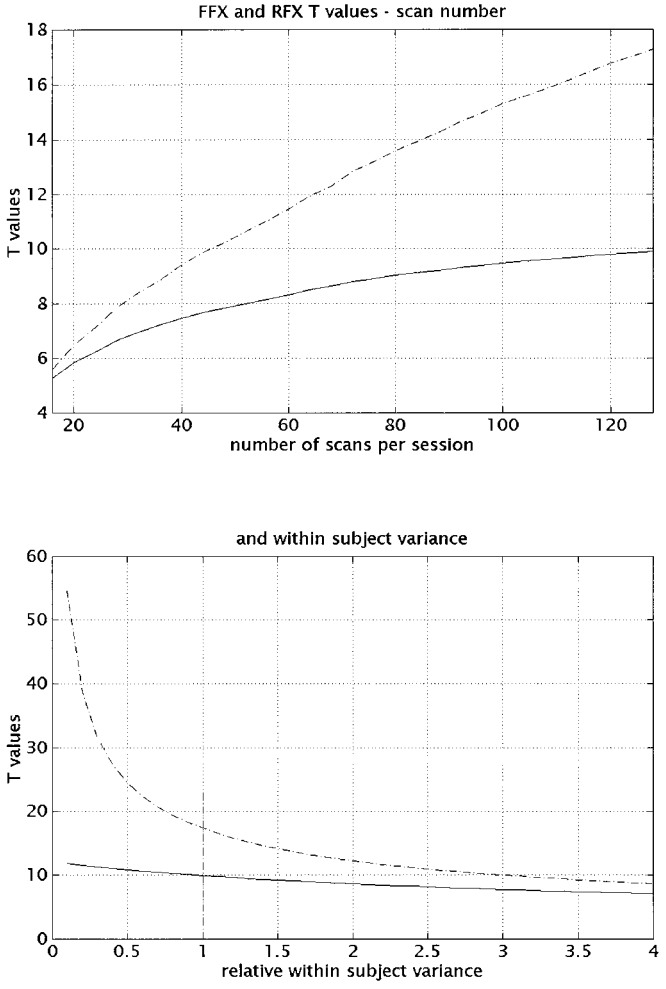


FIG. 6. Comparison of fixed and random-effect *T* values. Upper panel: The relationship between the *T* values based on a single-level (FFX—broken line) and two-level hierarchical model (RFX—solid line) as a function of the number of scans for each subject's session. Lower panel: Equivalent relationship as a function of relative within-subject error (the dashed vertical line corresponds to the error estimated empirically). In the limit of high within-subject error, or variability in its estimate due to a small number of scans the two *T* statistics converge. These results used the hyperparameter estimates from the analysis of the simulated data described in the legend of Fig. 4.

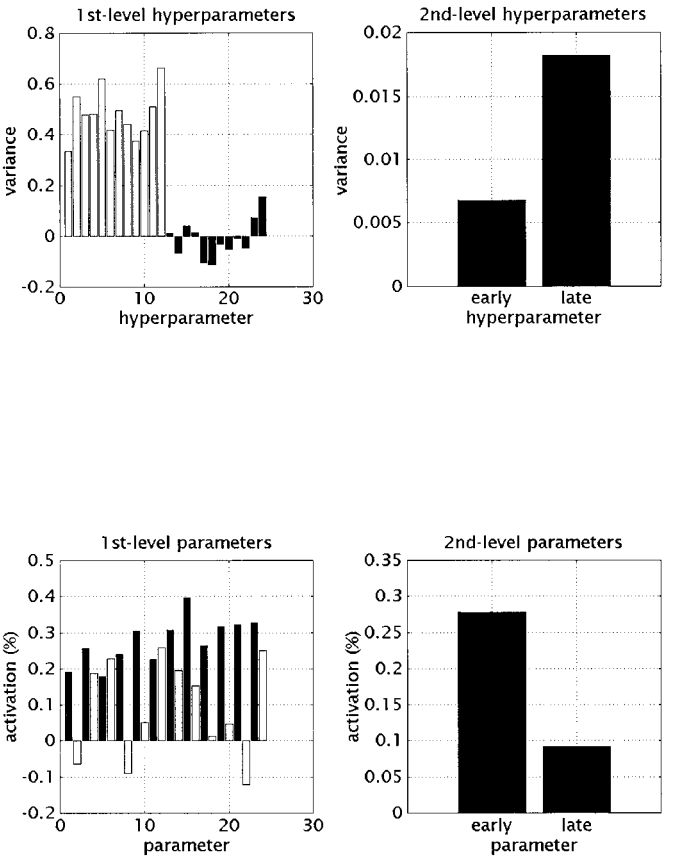


FIG. 7. Estimation of differential event-related responses in real data. The format of this figure is identical to that of Fig. 4. The only differences are that these results are based on real data where the response is due to the difference between studied or familiar (old) words and novel (new) words. In this example we used the first 128 scans from 12 subjects. Clearly in this figure we cannot include true effects.

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, 39 mm), are shown in Fig. 7, adopting the same

[5] 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.

format as in Fig. 4. Here we see that the within-subject error varies much more in the empirical data with the last subject showing almost twice the error variance of the first subject. As we found in Section 1 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 Fig. 7). The intersubject 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 Section 6.1) 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 nonsphericity (in this instance heteroscedasticity or unequal variances) and a role for the analyses illustrated here.

The response estimation and inference are shown in Fig. 8. Again we see the characteristic "shrinkage" when comparing the ML to the conditional estimates. It can be seen that all subjects, apart from 1 and 3, 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.

2.5 Posterior Probability Maps (PPMs)

The analysis was repeated for all voxels, in the slice containing the voxel reported above, surviving a conventional fixed-effect analysis for any condition-related effect using a capricious F ratio ($P < 0.001$ uncorrected). The distributions, over voxels, of the contrast testing for an early differential response are shown in Fig. 9. The differential activations from a single subject (subject 5) range from about -0.28% to 0.75%. It is interesting to note that even with the pre-selection of both activated and deactivated voxels, the distribution is unimodal and shows positive skew (upper left panel). In contradistinction the activations at the second level, over subjects, are roughly Gaussian (upper right panel). This distribution is interesting for two reasons. First, its central nature suggests voxels show an equal tendency to activate and deactivate even within a slice. Global normalization would require the masses under the positive and negative sides of the distribution to be equal, over the entire brain, but not for a subset of voxels from a single slice. More importantly the unimodal distribution suggests that activations are continuously distributed and speak against two distinct distributions for activated and nonactivated voxels. This touches on the assumption made in the next section, in which observations over voxels constitute the second level. The distributions of within and between-subject error, over voxels, are presented in the lower panels of Fig. 9. These are the conditional variances of the contrasts in the upper panels. Both correspond to scaled chi-squared distributions with a mean within- and between-subject error of 0.72% (standard error) and 0.13% (standard error), respectively. This means that the within-subject error is about the same as the maximum activations expressed by subjects. Similarly the between-subject variability is only slightly less than the maximum activation averaged over subjects.

The posterior probabilities of differential activations (early component) for the first subject were calculated using (7) and assembled into a PPM (Fig. 10). The PPM is shown alongside the corresponding conventional fixed-effect SPM$\{T\}$ for this subject. The upper panels show the relationship between the posterior probabilities and conditional means (left) and the fixed-effect T values (right). Notice that when the conditional mean approaches the size threshold used for posterior inference (in this case 0.1%) the posterior probabilities approach 50%. The PPM identifies enhanced activation for studied words in the right prefrontal, bilateral posterior parietal, cingulate cortices and the precuneus. A similar profile, although less complete, obtains from the SPM$\{T\}$ thresholded at $P = 0.001$ uncorrected. The critical thing to note is that there is no simple one to one relationship between the fixed-effect T value and the posterior probability (see the upper right panel). The discrepancy between the SPM and PPM reflects the essential differences in the nature of the inference and the implied specificity and sensitivity. As discussed in Section 3 of the previous paper (Friston *et al.*, 2002) the PPM has the latitude to adjust its specificity and sensitivity according to the inherent variability of the activations over subjects and the local error variance. It does this to maintain focus on the object of the inference; namely whether the activation is greater than the specified threshold. In contradistinction the SPM$\{T\}$ has no notion of how variable the response is or how big the response has to be before it is meaningful. It simply adopts the same specificity for all brain regions. The sensitivity and specificity of Bayesian inferences in this example are now dealt with in more detail.

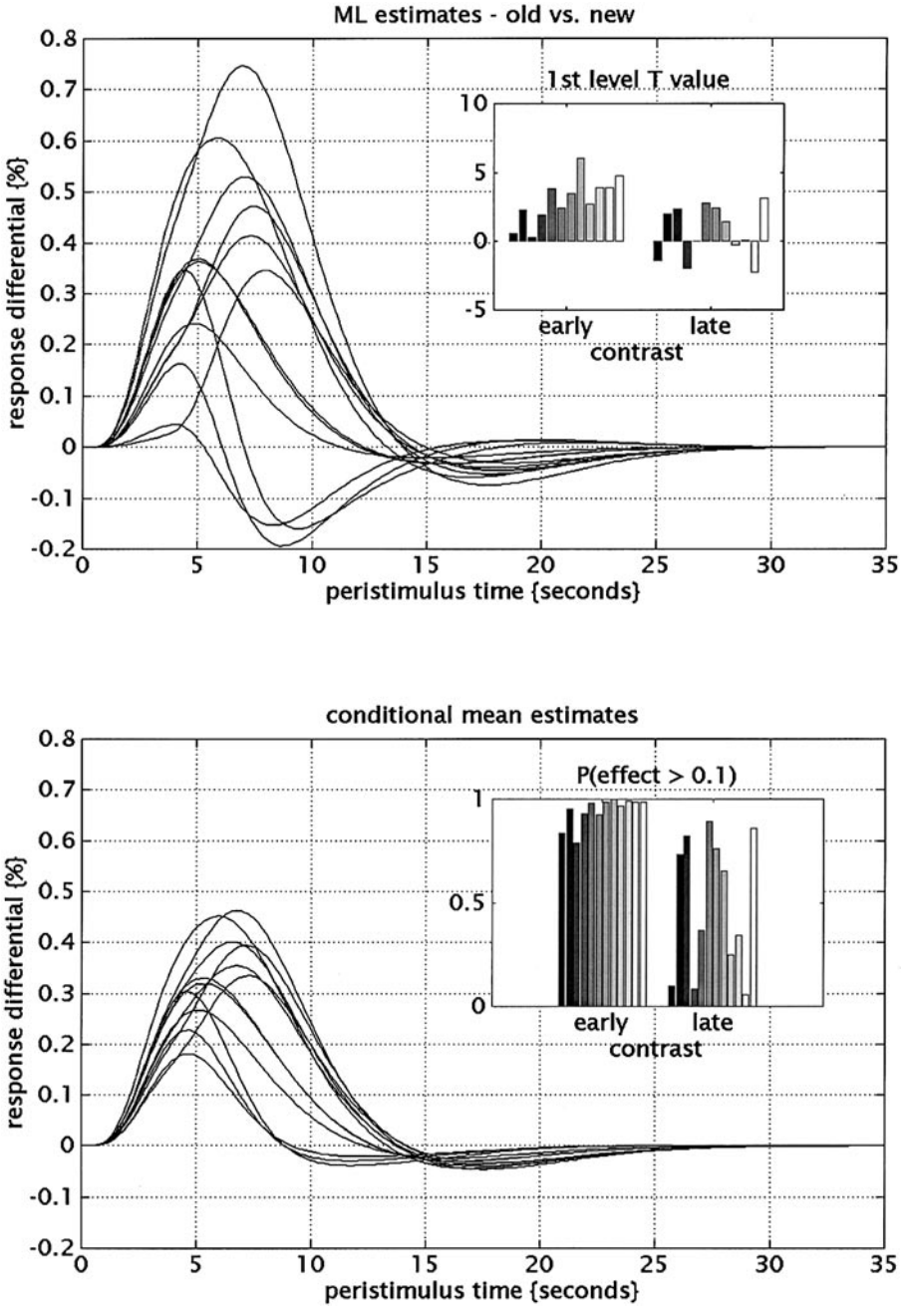


FIG. 8. The format of this figure is identical to that of Fig. 5. The only differences are that these results are based on real data were the response is due to the difference between studied or familiar (old) words and novel (new) words. The same regression of conditional responses to the conditional 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.

2.6 Sensitivity and Specificity of Bayesian Inference

In section 2.5 we introduced some heuristics concerning the sensitivity and specificity of Bayesian inference in relation to classical inference. In this subsection we consider this issue in quantitative terms using the empirical estimates of error and prior covariances from the previous analysis. By extending the expressions in Section 3 in Friston *et al.* (2002) to cover design matrices with multiple columns, we can compute the sensitivity and specificity of the thresholding PPMs (at a confidence level specified by the Z-variate u).

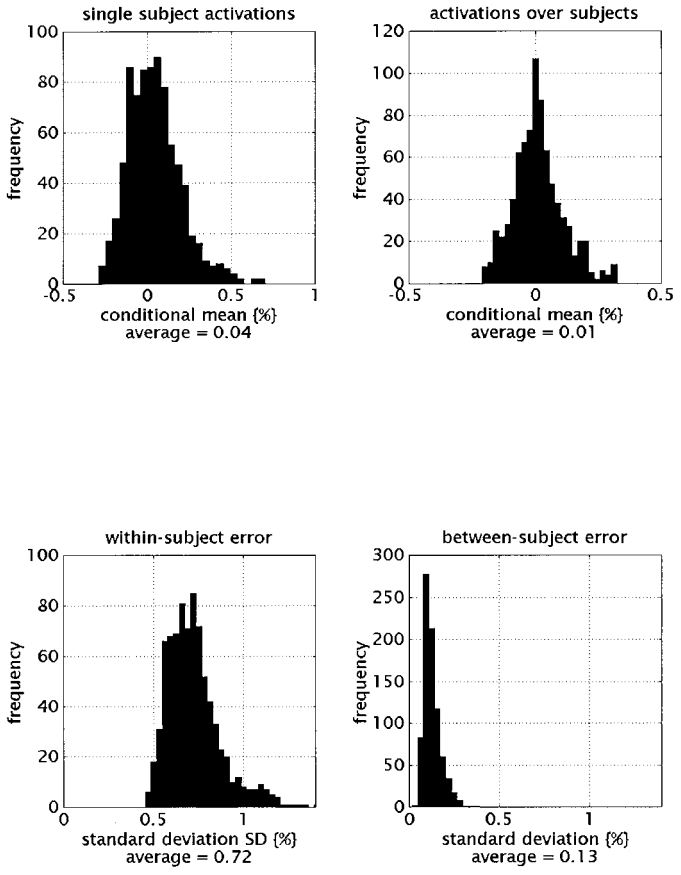


FIG. 9. Results from an analyses over multiple voxels selected on the basis of a conventional SPM analysis (F test for all event-related responses, $P < 0.001$, uncorrected) at $z = 39$ mm. Upper left panel: Distribution of a contrast of conditional mean responses, here the differential peak response evoked by old words relative to new, in terms of the early component. This distribution is for subject 5. The noncentral nature and skew of this distribution suggests that regional responses in this part of the brain tend to be greater for "old" relative to "new" words. However, relative deactivations are nearly as prevalent. A reasonable differential activation, in this context would be about 0.2%. Upper right panel: Equivalent contrast at the second level reflecting conditional responses averaged over subjects. Lower left panel: Distribution of within-subject error over the same voxels. This corresponds to the conditional variance of this subject-specific contrast. The average standard deviation of this error is 0.72%. Therefore responses of about 0.3% are occurring in the context of errors that are over twice their magnitude. Lower panel: Between-subject error over the same voxels. This represents the variability in responses over subjects. It can be seen that the average standard deviation (0.13%) is a little less than the magnitude of the larger responses themselves. These results are interesting because they frame, quantitatively the size of responses in relation to the within-session noise and their variability over subjects.

$$\alpha = 1 - \Phi(w)$$
$$\beta = 1 - \Phi\left(w - \frac{c^T C_{\theta|y} X^T C_\epsilon^{-1} X A}{\sqrt{c^T C_\eta c}}\right) \qquad (11)$$
$$w = \frac{\gamma + u\sqrt{c^T C_{\theta|y} c}}{\sqrt{c^T C_\eta c}}$$

Here the contrast weights c were chosen to test for early activation differentials in the first subject, at the cingulate voxel reported above. Superscripts have been dropped from (11) because these expressions hold for any level considered. Here we are dealing with Bayesian inference at the first level. Both sensitivity and specificity are functions of the size threshold γ specified, whereas only sensitivity is a function of A, the true effect. Figure 11 shows the sensitivity or power β (solid line) and false positive rate $\alpha = 1 -$ specificity (broken line) as functions of the size threshold. These results are for 90% confidence, given a true activation of 0.5% (vertical line). As one might expect both power and false positive rate increase as the threshold is reduced. The lower panel shows the same relationship but on a semilog scale. It can be seen that false positive rate approaches very small levels as the threshold approaches the true activation. In the example shown, a voxel would be correctly declared as activating by 0.1% or more on about 50% of occasions while retaining considerable specificity ($\alpha \approx 10^{-4}$). Conversely we can examine sensitivity and specificity as functions of the true activation for a fixed size threshold. Figure 12 shows the results for a threshold that maintains a low false positive rate of $\alpha < 10^{-4}$ (vertical line). Specificity is not a function of the true activation but sensitivity increases dramatically with activations above 0.4%.

Finally, we demonstrate the dependence on error variance. This is interesting because different error variances in different brain regions imply that Bayesian inference has a self-adjusting sensitivity and specificity depending on the local noise. Figure 13 shows power and false positive rate as functions of relative error variance, modeled by scaling the error covariance between 0 and 4 while holding the true activation and threshold fixed. For relatively large errors, sensitivity and false positive rate fall in tandem with increasing error. However, the critical thing to note is that the proportional difference between the power and false positive rate in the semilog plot (lower panel) decreases with relative error variance. This means that the power, relative to false positive rate falls more slowly with increasing error. In other words, the balance between specificity and sensitivity is implicitly adjusted depending on the reliability of the measured response.

2.7 Sensitivity and Specificity of Classical Inference

In this final subsection we turn to inference at the final level using the ML estimator and the T statistic. The aim here is to demonstrate how the covariance component estimation can be useful from a purely classical perspective. We do this by using the variance partitioning to evaluate how the sensitivity of a classical second-level inference depends on the relative number of subjects and scans. We created a series of synthetic balanced designs in which

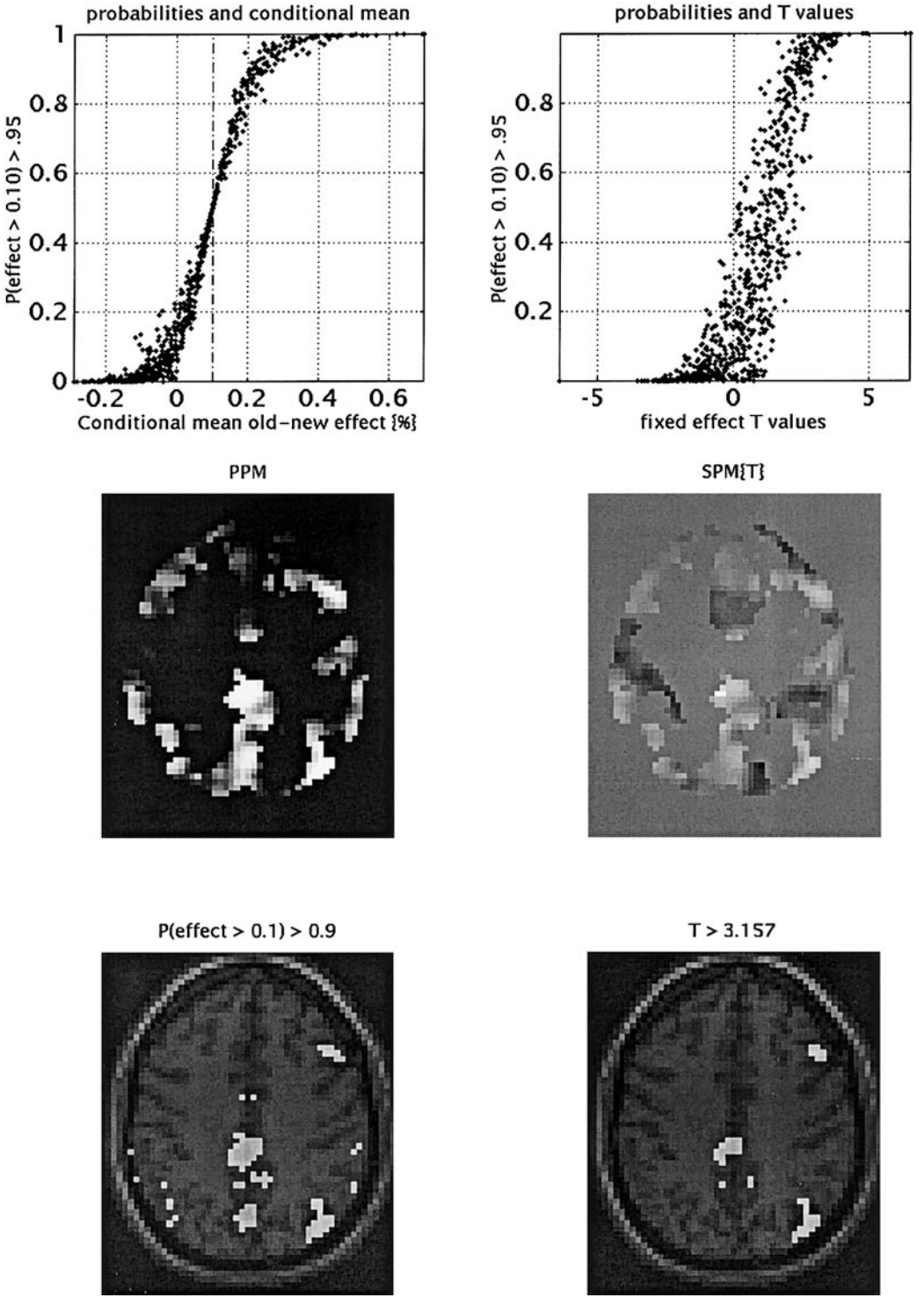


FIG. 10. Posterior probability map (PPM) and SPM{*T*} pertaining to first-level effects. The posterior probability of an early differential response in subject 5 was computed for the voxels described in the legend of Fig. 9. Upper panels: Posterior probability that the differential effects were greater than 0.1% expressed as a function of the conditional means (left) and fixed-effect *T* values (right). Lower panels (left). Posterior probability map (PPM) (above) and thresholded (below) to show regions that evidenced a differential peak response of 0.1% or more with 90% confidence, or more. Lower panels (right). Equivalent SPM{*T*} based on the single-level model, fixed-effects *T* values. The threshold adopted here (lower right panel) was 0.001 uncorrected. It can be seen that this particular subject shows quite an extensive differential response involving cingulate, right prefrontal, and bilateral parietal cortices that is similar to, but more complete than, the activation profile inferred on the basis of the SPM{*T*} (even with this liberal threshold).

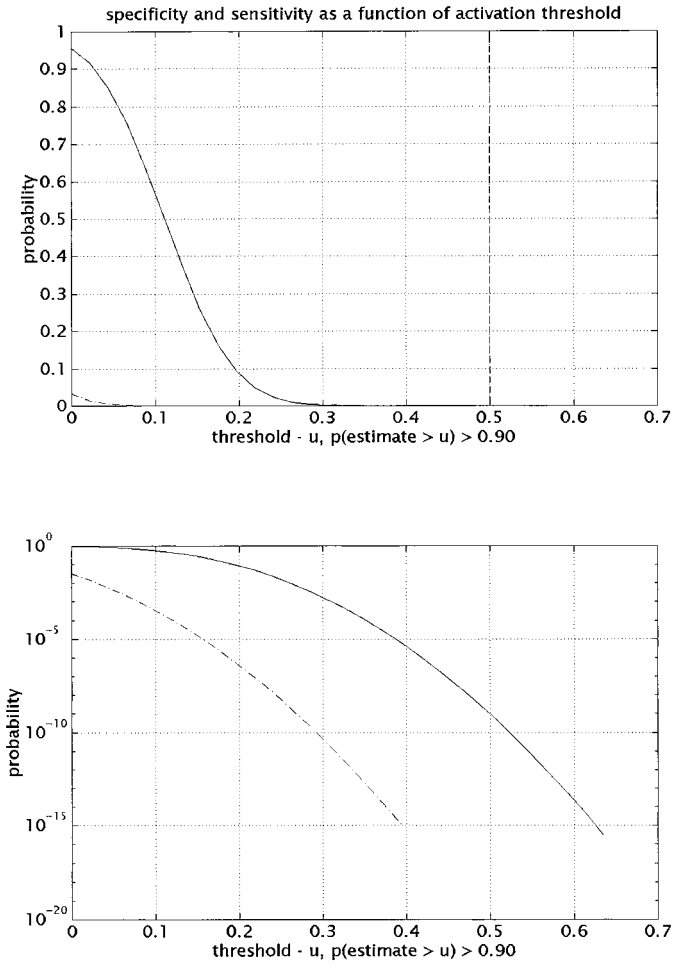


FIG. 11. Specificity and sensitivity of first-level inferences as a function of activation threshold: Upper panel: The probability of declaring a voxel "activated" to degree u or more, with 90% confidence in the context of no activation (broken curve) and with a true activation of 0.5% (solid curve and vertical dashed line). These probabilities are based on the hyperparameters from the analysis of section 2.4. In this instance the posterior probabilities pertain to the activations in the cingulate gyrus voxel described in the legend of Fig. 7, in the first subject. Lower panel: The same as in the upper panel but plotted on a semilog scale. A 50% sensitivity is achieved with an activation threshold of about 0.1% whilst retaining a high specificity.

subject-specific responses were modeled by the same design matrix (that of the first subject). Different numbers of scans per subject and numbers of subjects were modeled by decimating the first- and second-level design matrices, respectively. These design matrices and the covariance hyperparameters, for the voxel of the previous subsection, were entered into (9) to give the standard error of the second-level ML contrast testing for an early activation differential. The sensitivity of the corresponding T test is simply

$$\beta = 1 - \Phi_T\left(w - \frac{A}{\sqrt{c^T M^{(2)} C_{\hat{\varepsilon}} M^{(2)T} c}}\right), \qquad (12)$$

where w is some T value threshold and A is the true activation, here set to 0.001 and 0.5% respectively. The results are shown in Fig. 14. As might be anticipated there is a trade-off between the number of subjects and number of scans per subject. The minimum number of subjects, required to attain 90% sensitivity for voxels such as the one chosen, appears to be about 8 and this requires about 100 scans per subject. Scanning 16 subjects with about 24 scans each can approximate the same sensitivity. This power analysis is presented as an illustration of how covariance component estimation can be used in a classical power analysis. The quantitative conclusions pertain to, and only to the voxel reported.

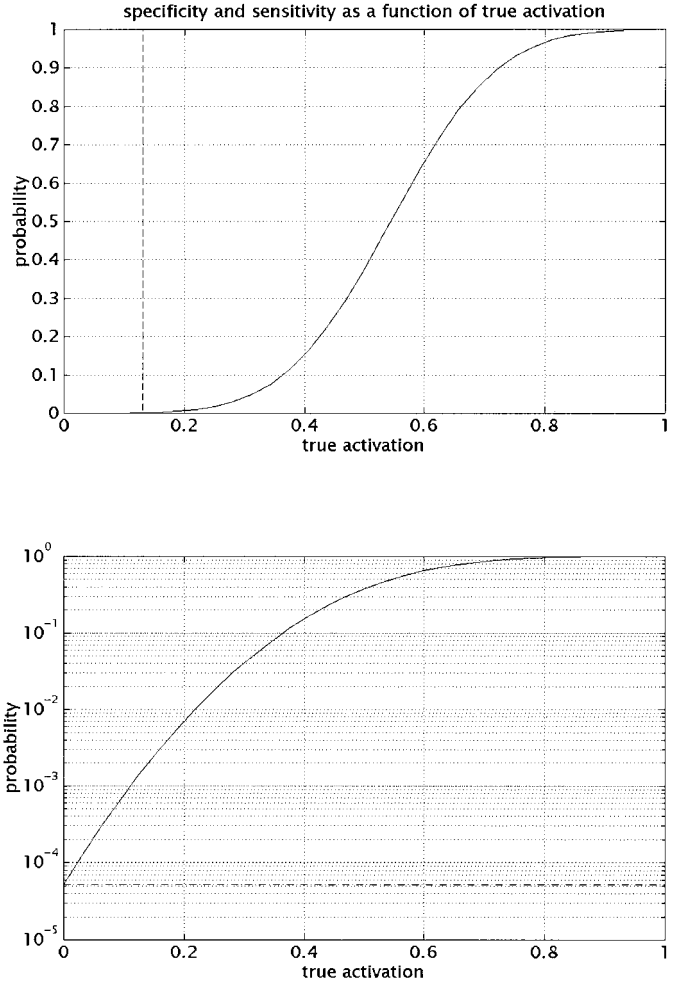
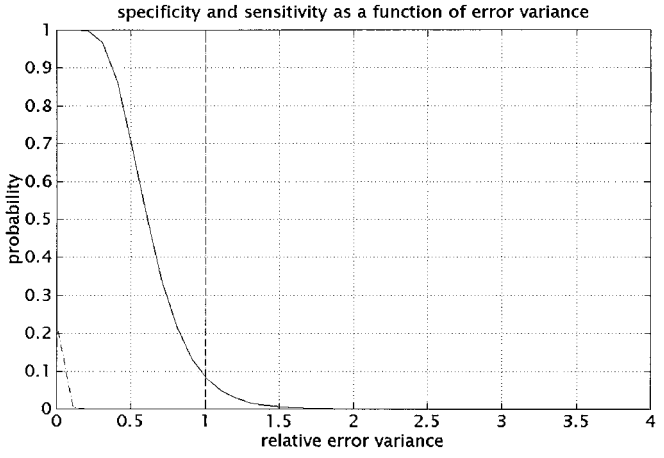


FIG. 12. As for Fig. 11 but holding the threshold constant at about 0.14% (such that the specificity was at least $1 - 10^{-4}$) and varying the true activation. In this case the specificity is constant but sensitivity increases as the true activation exceeds the specified threshold.



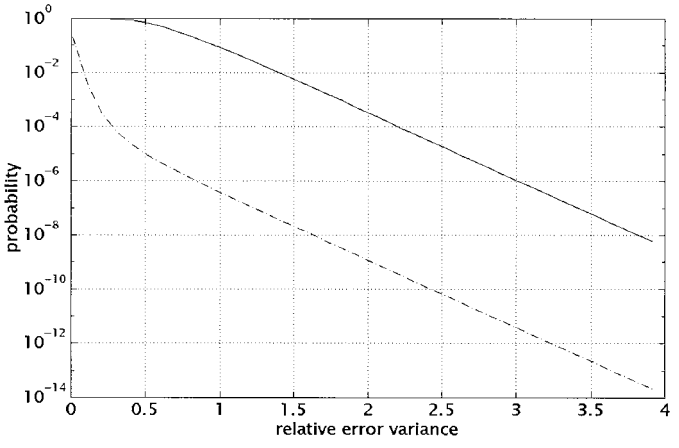
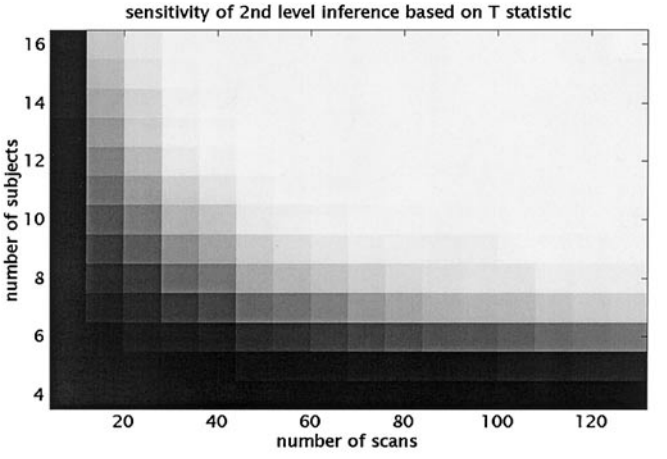


FIG. 13. As for Figs. 11 and 12 but holding the threshold and true activation constant (at 0.14% and 0.5% respectively) and varying the relative amount of error variance. As might be expected, increasing the error variance reduces the probability of declaring a voxel to be activated irrespective of the true activation. Critically, it does so in proportion at high levels of error. However, at very low levels of within-subject error the specificity falls, relative to sensitivity to a lower limit that is determined by the between-subject variability and the threshold chosen.

2.8 Summary

The examples presented above allow us to reprise a number of important points made in the previous paper (Friston *et al.*, 2002). 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 an EM algorithm.
- 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 nonsphericity 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 reenter a generalized least square scheme, as a nonsphericity 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.
- The sensitivity and specificity of Bayesian inference differs from that of classical approaches. In particular, Bayesian inference maintains a high specificity while adjusting its sensitivity according to the prevail-



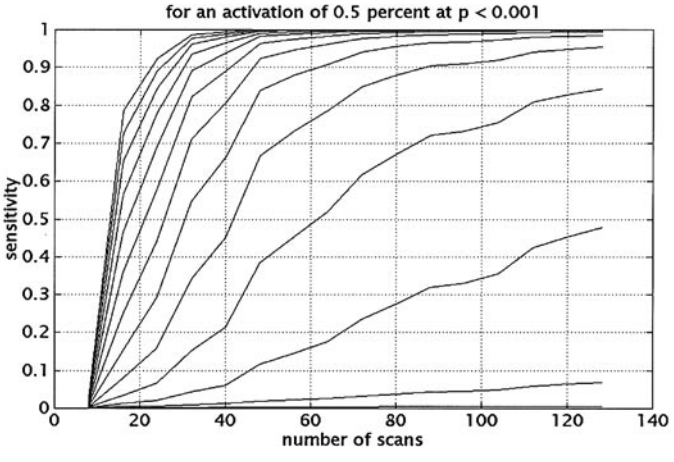


FIG. 14. Sensitivity of second-level inferences. Upper panel: Sensitivity or the probability of declaring a voxel (in the cingulate gyrus) significant at $P < 0.001$ based on the second-level T statistic. The sensitivity is shown in image format (white = 1 and black = 0) as a function of the number of subjects and scans per subject. Lower panel: The same data but plotted graphically as a function of scan number. Here the activation was assumed to be 0.5%. Note the trade-off between the number of scans and subjects, lending a hyperbolic-like form to the sensitivity function.

ing error variance and inherent variability of the response. Specificity can be ensured by making an inference about an "unlikely" (i.e., reasonably large) effect. This is precluded in classical inferences because the inference is about the data, not the activation.

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

3. SPATIOTEMPORAL MODELS WITH EMPIRICAL BAYES

3.1 Introduction

There has been a growing interest in spatiotemporal Bayes models for imaging data-sequences as exemplified by some recent and engaging proposals. For example Descombes *et al.* (1998) have explored the use of spatiotemporal Markov field models to characterize activations in fMRI, while Everitt and Bullmore (1999) have looked at mixture models to assign conditional activation probabilities. The compelling work of Hartvig and Jensen (2000) combines both these approaches. The dynamics of fMRI time-series has been addressed by Højen-Sørensen *et al.* (2000) who use Hidden Markov Models to making inferences about which state the brain is in.

These proposals are exciting and will probably be a focus of research for many years. However, the purpose of this paper is not to propose a new model but to show that existing models can be treated in a Bayesian fashion. To enable this we have to make one assumption, about the distribution of activations, in addition to those made in conventional analyses. Namely that regionally specific responses have an expectation of zero, over the whole brain (this is true by definition, otherwise they would not be regionally specific) and a Gaussian distribution (this can be motivated using the empirical results of the previous section).[6] With this assumption conventional models (e.g., anatomically informed basis functions AIBF; Kiebel *et al.*, 2000; Phillips *et al.*, 2000) can be simply extended to facilitate inference through empirical Bayes.

In this section we focus on priors that derive from making multiple observations over voxels and how these observations can be harnessed in an empirical Bayesian framework to make more informed inferences about any single voxel. In Bayesian inference the posterior probability of an effect is proportional to its likelihood and prior probability. If the latter is known this allows a full Bayes treatment. If the priors are not know then they can be [hyper]parameterized in terms of some hyperparameters that are estimated from the data using empirical Bayes. The basic idea, here, is that the prior probability of a particular voxel activating can be estimated from the distribution of estimated activations over all remaining voxels. This rests upon a hierarchical observation model where the first level is exactly the same as in a conventional voxel-based general linear model and the second level comprises observations over voxels. The ensuing variance can be partitioned into within- and between-voxel components that are estimated, jointly with voxel-specific activations per se, using the EM algorithm described in Friston *et al.* (2002).

This section revisits Bayesian inference in a practical sense and illustrates the latitude afforded by being able to incorporate prior knowledge into estimation and inference schemes. We have chosen PET data to illustrate some key points because the benefits are more apparent given PETs relatively poor spatial resolution and the smaller number of scans. In what follows we present a spatiotemporal model and discuss how prior information of different sorts can be incorporated. The underlying form and motivation for the model is described in this section. In the next section, simulated PET data is subject to Bayesian analysis to show the advantages, in relation to known underlying activations. The final section applies exactly the same procedures to real PET data. This data is the verbal fluency data set used in many of our previous theoretical publications and is the SPM training PET data, available from http://www.fil.ion.ucl.ac.uk/spm. We conclude with a general discussion of empirical Bayes in neuroimaging.

3.2 Spatio-Temporal Hierarchical Models

As in the previous sections we start with model specification and develop the implied constraints under which its parameters are estimated. In this example we treat each voxel as a replication of the same observation at the first level

$$
\begin{aligned}
&\text{level one} \quad y = X^{(1)}\theta^{(1)} + \epsilon^{(1)} \\
&\begin{bmatrix} y_1 \\ \vdots \\ y_s \end{bmatrix} = P \begin{bmatrix} X_1^{(1)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & X_s^{(1)} \end{bmatrix} \begin{bmatrix} \theta_1^{(1)} \\ \vdots \\ \theta_s^{(1)} \end{bmatrix} + \epsilon^{(1)} \\
&Q_1^{(1)} = PP^T \\
&\text{level two} \quad \theta^{(1)} = X^{(2)}\theta^{(2)} + \epsilon^{(2)} \\
&X^{(2)} = 0_s \\
&Q_1^{(2)} = GDG^T
\end{aligned}
\tag{13}
$$

At the first level, all the voxel time-series are stacked on top of each other to create a large response vector y

[6] This assumption differs from those made by the approaches mentioned in the first paragraph, in which a separate distribution is assumed for activated and nonactivated voxels.

with $t \times s$ elements (t scans for each of s voxels). The design matrix $X^{(1)}$ is a leading block diagonal matrix with the voxel-specific design matrix (in this case the same design matrix for each voxel) along the leading diagonal. Vectorizing the response variable in this way essentially converts a s-variate multivariate problem into a univariate model with repeated measures over s voxels. The explanatory variables are pre-multiplied by P, which models the spatial blurring due to the point spread function psf and is given by

$$P = psf \otimes I_t \tag{14}$$

where $\otimes$ is the Kronecker tensor product and psf_{ij} is the value of the point spread function at the distance between voxels i and j. Because we are dealing with PET, temporal autocorrelations are not considered.[7] By modeling spatial correlations in this fashion, parameter estimation effects an implicit least squares de-convolution. This is because the spatial correlations, induced by the point spread function, are embodied in the explanatory variables (i.e., the forward model from the point of view of source estimation) and not in the parameter estimates. This is only possible because of the spatiotemporal form adopted in (13). The constraints $Q^{(1)}$ on the error variance are scaled by a hyperparameter to give the error covariances at the first level $\lambda_1 Q^{(1)} = C_\epsilon^{(1)}$. These correspond here to stationary error terms over voxels whose spatial correlation structure conforms to the point spread function. Although this is appropriate for the simulated data in the next section, this basis set should obviously be extended to allow for voxel-specific variations in error variance (see the Section 4). The parameters $\theta^{(1)}$ are a large vector with p parameters for each of the s voxels. For simplicity we have used only one regressor in the design matrices in the examples below.

At the second level the voxel-specific parameter $\theta^{(2)}$ estimates are modeled as zero mean variates with a spatially structured covariance. The first-level parameters are assigned an expectation of zero because we are only interested in regionally specific effects.[8] This means that the average of any effect over voxels is zero and presupposes that global effects have been removed from the data before estimation (i.e., global normalization). Adopting a second level allows one to model the spatial dependencies of the signal in nearby voxels and use the variability of responses, over voxels, as priors on the estimate of any particular voxel's response.

[7] If we wanted to incorporate some known temporal convolution we would simply multiply the design matrix by $I_s \otimes hrf$ where hrf is a temporal convolution matrix for each time-series, for example, a hemodynamic response function.

[8] More generally $X^{(2)} = 1_s \otimes I_p$ giving a second level design matrix with $s \times p$ rows and p columns where each column is associated with a second level parameter.

The second-level spatial dependencies among the responses are based on spatial priors that are constructed according to the following arguments. First, in the absence of any information about the relative tissue composition of each voxel (e.g., grey matter, white matter, CSF etc) we know that hemodynamic signals evidence spatial correlations. These short-range correlations are due to the mediation of increases in rCBF by diffusive signals and the local architecture of cerebral vasculature. Optical imaging experiments suggest an intrinsic smoothness of between 2 and 5 mm. We can build this information into the model by specifying these intrinsic correlations in terms of a covariance constraint at the second level. Here modeled by D, a Gaussian correlation matrix of 4 mm full width at half maximum (FWHM). Second, we can impose neuroanatomical constraints by modulating these stationary correlations using grey matter priors G. These priors enter as a leading diagonal matrix whose elements reflect the probability that the corresponding voxel is a grey matter voxel (Ashburner and Friston, 1997). Alternatively, these priors can be construed as the proportion of the voxel that is grey matter and, consequently, capable of engendering a measurable response. It follows that, in the absence of any functional information, the prior covariance of the signal has the form $Q_1^{(2)} = GDG^T$. An intuitive way to motive this form for the spatial priors is to think about biophysical signals that induce blood flow as being smoothed and dispersed by some intrinsic convolution matrix $D^{1/2}$. The hemodynamic response, induced in a voxel, will be proportional to the interaction between, or product of, the amount of dispersed signal and the proportion of that voxel that can respond (i.e., the grey matter probability). The resulting convolution with $GD^{1/2}$ of i.i.d. sources would give a signal with covariance proportional to GDG^T. We could of course incorporate the fact that grey matter is the most likely origin of these flow-inducing signals to give $Q_1^{(2)} = GD^{1/2}GG^TD^{1/2T}G^T$ but the simpler form in (13) is sufficient for current purposes. Before proceeding to estimation we have to consider the size of the matrices in (13). If we wanted to include a hundred thousand voxels, they would be prohibitively large. In order to make the estimation computationally tractable these matrices have to be reduced. This is not a generic aspect of the Bayesian approach but something one has to consider in spatiotemporal models with large numbers of scans and voxels.

3.3 Model Reduction and "Hard" Priors

In this subsection we suggest a form of model reduction that uses the priors to motivate a suitable basis set, onto which the data at various levels can be projected. This effectively reduces the problem of dealing with s voxels to dealing with a much smaller number of

m spatial modes. To do this we borrow a device developed previously for the inverse problem in EEG source estimation (Phillips *et al.*, 2002) using AIBF (anatomically informed basis functions) (Kiebel *et al.*, 2000). Namely, use the spatial basis set that preserves the most information about sources, that conform to the prior covariance, after projection into the subspace in which they are estimated.[9] These bases are simply the eigenvectors U of the prior covariance matrix with the largest m eigenvalues

$$Q_1^{(2)}U = US \tag{15}$$

where S is a $m \times m$ leading diagonal matrix of eigenvalues.

The basic idea behind anatomically informed basis functions (AIBF) is to establish a small number of spatial patterns or modes that can be linearly mixed to approximate the profile of voxel-specific responses. Because the number of modes or AIBF is much smaller than the number of voxels, iterative schemes like EM can be used with relative computational ease. The basis set is chosen to maximize the amount of information (entropy) in the responses, under the prior distribution, after the responses are projected onto this basis (i.e., expressed in terms of the coefficients of the basis set). Under Gaussian assumptions, these basis functions are the eigenvectors of the prior covariance matrix with the largest eigenvalues. The use of AIBF can be construed as setting the prior variances of activation patterns conforming to the "unused" minor eigenvectors to zero. This precludes the estimates from lying in the subspace spanned by these minor modes, because the resulting priors enforce zero response with infinite precision. One could simply modify the prior covariance matrix and proceed in voxel space. However, using AIBF is mathematically the same, but is much more efficient and represents a useful way to implement "hard" priors.[10] An example of a hard constraint would be setting the prior variance of hemodynamic responses to be zero in brain ventricles or white matter. These are the sorts of constraints embodied in AIBFs.

The eigenvectors or AIBF now enter into (13) to give a reduced form that is computationally tractable

level one

$$(U^T \otimes I_t)y = (U^T \otimes I_t)X^{(1)}UU^T\theta^{(1)} + (U^T \otimes I_t)\epsilon^{(1)}$$

$$\text{giving} \quad y_u = X_u^{(1)}\theta_u^{(1)} + \epsilon_u^{(1)}$$

where $X_u^{(1)} = (U^T \otimes I_t)X^{(1)}U$

$$\theta_u^{(1)} = U^T\theta^{(1)}$$

$$Q_u^{(1)} = (U^T \otimes I_t)PP^T(U^T \otimes I_t)^T$$

level two

$$U^T\theta^{(1)} = U^TX^{(2)}\theta^{(2)} + U^T\epsilon^{(2)}$$

$$\text{giving} \quad \theta_u^{(1)} = X_u^{(2)}\theta^{(2)} + \epsilon_u^{(2)}$$

$$\text{where} \quad X_u^{(2)} = U^TX^{(2)} = 0_u$$

$$Q_u^{(2)} = U^TGDG^TU = S. \tag{16}$$

Note that when using these spatial bases the covariance constraints at the second level reduce to the leading diagonal matrix of eigenvalues in (15). In this form we are solving, not for voxel-specific estimates, but mode-specific estimates of the conditional means and covariances. From these we can reconstruct the relevant statistics in voxel space as shown below. In the simulations below and in the empirical example of the subsequent section we used 64 spatial modes. The expressions in (16) have been written out in full to show the relationship between the reduced and non-reduced forms. In practice the operations involving Kronecker tensor products can be implemented in a simpler and computationally more efficient fashion.

4. SPATIOTEMPORAL MODELS: A SIMULATION STUDY

The aim of these simulations is to compare and contrast Bayesian and classical inference at the first level to highlight the potential usefulness of the former. We simulated data according to (13) using an activation, centred in the left dorsolateral prefrontal cortex, of 16 mm Gaussian width, modulated by the grey matter priors described above (see Fig. 15). The modulated activation was scaled to a peak height of 0.5. The error process at the first level was Gaussian, with unit variance, convolved with a Gaussian point spread function (psf) of 8 mm FWHM. The activation, encoded by the level-one design matrix, was a simple alternating baseline-activation sequence of 16 scans. The simulated data, design matrices and constraints according to (16) were entered into the EM algorithm to provide estimates of the conditional mean $\eta_{\theta|y}^{(i)}$ and covariances $C_{\theta|y}^{(i)}$ at each level. The conditional probability that the activation exceeded a size threshold $\gamma = 0.1$ was computed for each voxel j at the first level with

[9] Note that this device can only be used if the constraints on the priors comprise just a single matrix.

[10] "Hard" (as opposed to "soft") constraints are priors that are specified with infinite precision or zero variance.

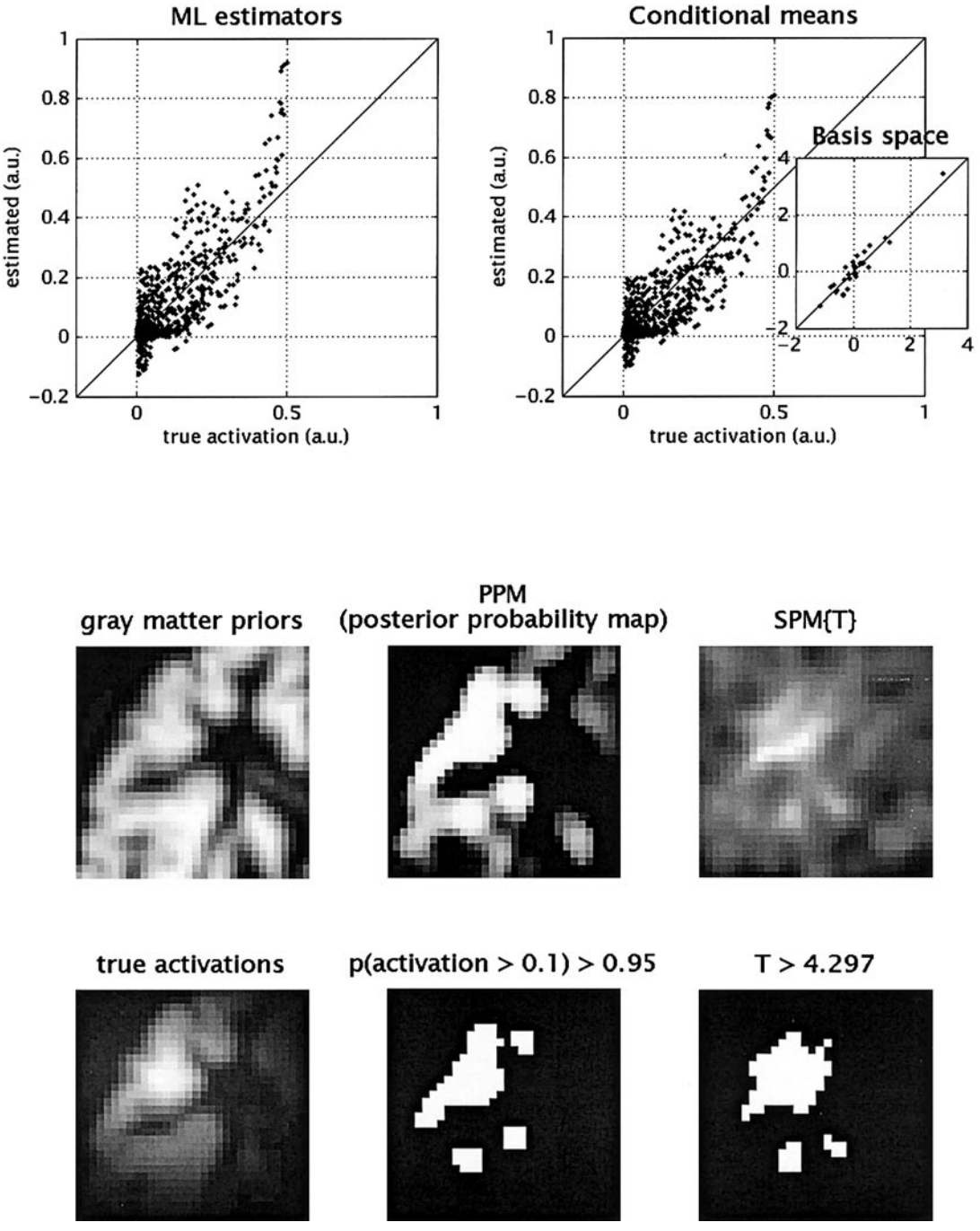


FIG. 15. Analysis of simulated PET data. Upper left panel: The ML estimates plotted against the true effect for each voxel. These ML estimates were obtained by solving (16) using only the first level. Upper right panel: The same, but now for the first-level conditional means using both levels. Insert: Plot of the true and conditional responses in the space of the basis functions. Lower Panels (left): Grey matter priors and the true activation. This activation was centred at -36, 24, 12 mm (Talairach and Tournoux, 1988) and comprised a 16-mm Gaussian function modulated by the grey matter priors. This simulated activation was scaled to a peak height of 0.5 arbitrary units (that can be considered ml/dl/min equivalents). Lower panel (middle): Posterior probability map (PPM) detailing the voxel-wise probability that the activation is greater than 0.1. The lower panel is the PPM threshold at $P > 0.95$ (i.e., the showing regions where one would be at least 95% confident that the activation was 0.1 or more). Lower panels (right): Classical analysis in terms of the SPM$\{T\}$. This SPM has been thresholded in the lower panel at $P < 0.05$ (corrected). These simulated data constitute a time-series of 16 (32×32 voxel) axial slices through the left dorsolateral prefrontal cortex, with the simulated activation appearing in alternate scans. Spatial correlations were modeled by smoothing a unit Gaussian noise field with a stationary Gaussian convolution kernel of 8 mm FWHM. The dimension reduction employed 64 spatial modes (see main text).

$$1 - \Phi\left(\frac{\gamma - U_{\cdot j}\eta^{(1)}_{\theta|y}}{\sqrt{U_{\cdot j}C^{(1)}_{\theta|y}U^{T}_{\cdot j}}}\right), \qquad (17)$$

where $U_{\cdot j}$ corresponds to the row of the eigenvector matrix associated with voxel j. $U_{\cdot j}$ plays the role of a vector of contrast weights that would be specified directly in more conventional settings (see Eq. (7) in the previous paper, Friston *et al.*, 2002). These probabilities form the posterior probability map (PPM), shown in Fig. 15 (middle panels) before and after thresholding.

The size threshold operationally defines the semantic or meaning of "activation," in much the same way that an age threshold of 12 might arbitrarily, but usefully, distinguish a child from a young person. Because this semantic formalism is precluded in classical inference there are few precedents to fall back on here. One could motivate the choice of $\gamma = 0.1$ by noting this is about 1% of a full-blown activation in PET (around 10 ml/dl/min). In other words, we are stipulating that anything less than 0.1 does not conform to an "activation" in the sense of a hemodynamic response that is typically evoked in an experiment.

To compare the PPM with the equivalent SPM, T values were computed for each voxel using conventional least squares and the first-level design matrix. The resulting SPM$\{T\}$ is shown in the left-hand panels of Fig. 15, before and after thresholding at a conventional level of 0.05 (corrected for the volume analyzed). It is clear from Fig. 15 that the PPM is a more anatomically informed characterization of the activations than the SPM. A critical thing here is that the SPM appears to have quite significant values even in areas with very low signal (lower right panels). This is because, although the activation effect may be very low, so is the estimated error variance. Only when we specify explicitly an interest in non-trivial effects (by setting $\gamma = 0.1$ in the Bayesian scheme) is the essential role of the priors evident.

The thresholded PPM identifies voxels whose posterior probability of activating is at least 95%, where an activation is γ, or more. This means that, at most, 5% of the voxels identified could have activations less than γ. In other words, thresholding a PPM establishes an upper bound on the false discovery rate or FDR (Benjamini and Hochberg, 1995). The FDR is the proportion of voxels declared active (i.e., discovered) that are not. This is very different from the false positive rate that is the proportion of all voxels tested that are declared falsely significant. In short, thresholding a SPM controls false positive rate, whereas thresholding a PPM can be regarded as controlling FDR (see Genovese *et al.*, 2002). In this sense thresholding PPMs has a much closer connection to FDR control in classical schemes. However, we reiterate there is no reason to threshold a PPM, other than to enable a classical inference.

The sizes of the activations estimated by the classical and hierarchical Bayes models are actually very similar. The voxel-specific first-level parameter estimates $U_{\cdot j}\eta^{(1)}_{\theta|y}$ are plotted against the true values in the upper panels of Fig. 15. The left-hand panel shows the ML estimates that obtain using the first level only and are slight overestimates in relation to the conditional means. Although the Bayesian estimates approximate the true effects more closely, there is still some discrepancy at high levels of activation. This is due largely to the curvilinear relationship between the true and estimated voxel-specific responses. This nonlinearity reflects the inability of the relatively "coarse" spatial basis set to fit the true activation profile exactly. When plotting the actual and estimated responses in the space of the basis functions U, the anticipated, tight linear relationship is seen (insert in the upper right panel).

Notice how the conditional means in Fig. 15 (upper panel) cluster around the true values more than the ML estimates. This "shrinkage" or regression is due to the effect of the priors. As will be shown next, this effect can be very pronounced when the prior variance is small.

4.1 A Null Analysis

To illustrate a fundamental difference between the Bayesian approach, relative to the classical analysis, the analyses were repeated exactly, while setting the activation to zero (i.e., using the same error terms). Figure 16 shows the results in the same format as Fig. 15. At the threshold used, the classical approach still identifies false positive activations and, with a suitably low threshold, will always do so. We have deliberately chosen a low threshold (0.05 uncorrected) to highlight the arbitrary inferences that ensue from thresholding in a classical framework. On the other hand the PPM correctly shows no activations whatsoever. As intimated in the previous section this is not due to enhanced sensitivity of the Bayesian approach. It is simply a reflection of the fact that the criterion used to assess the conditional activation in the Bayesian scheme has a far greater specificity than the criterion adopted by the classical one, despite retaining sufficient sensitivity to give results with face validity. In this example the FDR is still 5%, or less, but there are no "discoveries" to be false.

The empirical Bayes and ML parameter estimates in the top panel of Fig. 16 evidence a further important difference. Here the classical ML estimates (left-hand panel) are wildly overestimated in relation to the conditional estimates (right panel). In this instance the estimates should all be zero and the shrinkage of the first-level conditional means to zero is clearly evident. This illustrates an essential benefit of the Bayesian approach. By embodying the knowledge that all brain

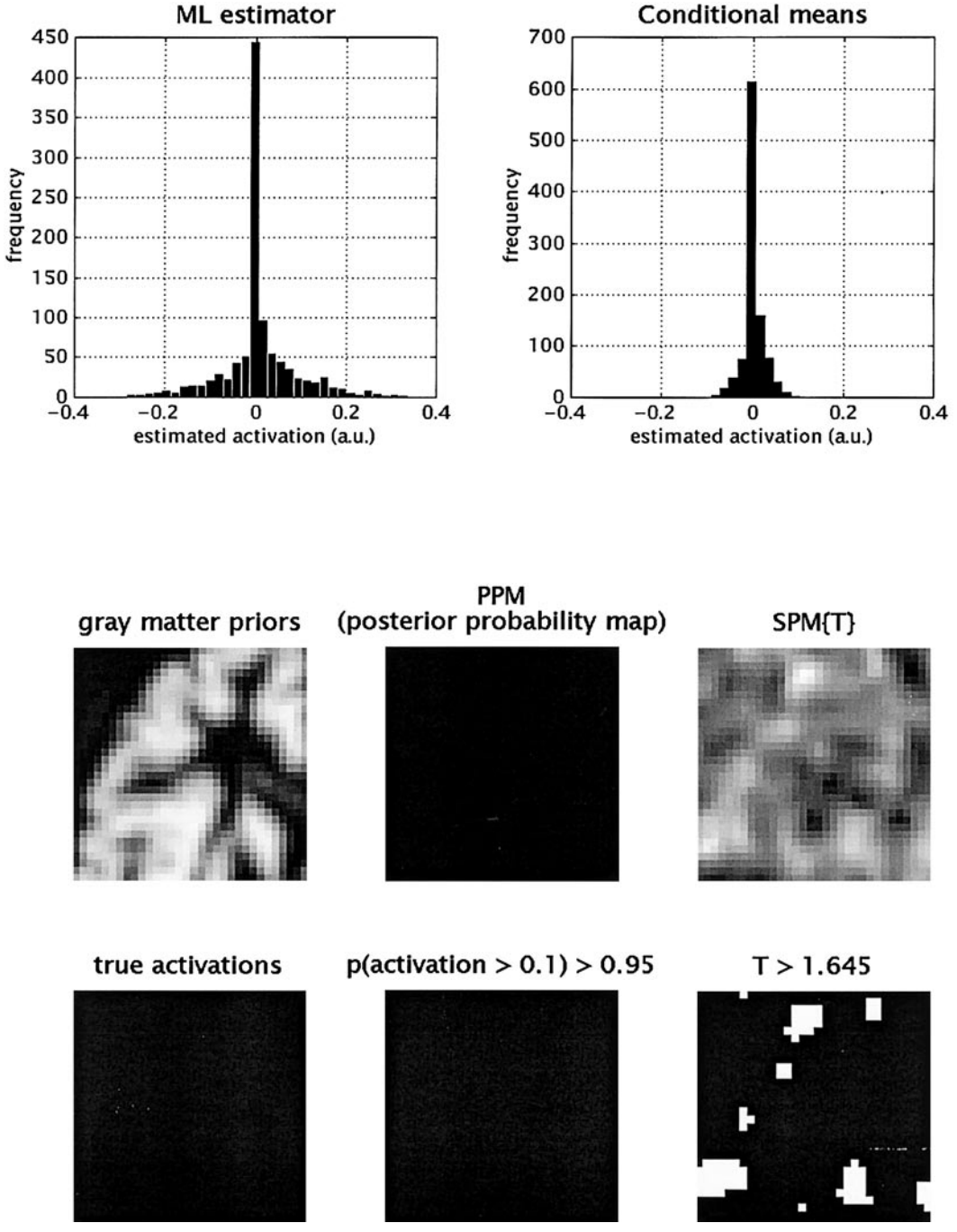


FIG. 16. As for Fig. 1 but in this instance the activation was omitted from the simulated data. Histograms of the ML and conditional estimates replace the plots of Fig. 15. The critical thing to note here is that the conditional estimates of the activation now show a profound regression to the true (zero) activation whereas the ML estimates are not constrained and provide wildly inflated estimates of the effect (upper left panel). As expected the SPM{*T*} shows a few false positives at this uncorrected threshold of 0.05. In contradistinction the PPM indicates, correctly, that no activation can be inferred with 95% confidence.

regions respond to a greater or lesser extent, augmenting the classical single-level model with a second level allows the Bayesian approach to estimate the intrinsic variability in activations from voxel to voxel. When variability is very small the estimates of activation at any voxel are suitably moderated. In contrast, the classical approach is uninformed and simply minimizes error variance at the first level in an unconstrained fashion, overestimating them by one or more orders of magnitude. In short the Bayesian scheme enables a regression to the [conditional] mean based on the relative variability induced in those estimates by the ob-

servation error and the intrinsic variability over voxels. This is the strength of Bayesian estimators.

5. SPATIOTEMPORAL MODELS: AN EMPIRICAL EXAMPLE

5.1 Analysis of Real Data

In this section we apply the same model to PET data to demonstrate the role of spatial priors in an empirical setting. These data are described in the legend to Fig. 17. Briefly they came from 5 subjects each scanned 12 times while performing a verbal fluency task (orthographic word generation) alternating with a word shadowing (word repetition) baseline. The model adopted was exactly the same as above (i.e., Eq. (16)) with one exception. In the simulations the error variance was the same for all voxels. However, in the real brain, error variance may change from brain structure to brain structure. This is accommodated easily by expanding the constraints on the error covariances to cover the nonstationariness of error variance anticipated. Here we adopt a very simple model and assume that error variance has two components, one that is stationary (as used above) and one that conforms to the same spatial profile as the hemodynamic signal itself, where, prior to reduction

$$\begin{aligned} Q_1^{(1)} &= PP^T \\ Q_2^{(1)} &= P(Q_1^{(2)} \otimes I_t)P^T. \end{aligned} \tag{18}$$

As in the simulations the intrinsic correlation matrix D was Gaussian with 4 mm FWHM. The point spread function was Gaussian with 16 mm FWHM, corresponding to the smoothing kernel applied to the data after spatial normalization. The motivation for smoothing the data, and then incorporating the smoothing kernel in the estimation model, is that one can reduce variations in spatial correlations from scan to scan and subject to subject. Furthermore, this allows one to accommodate variation in gyral and functional anatomy among subjects that is smaller than the smoothing kernel. By smoothing the data and then applying the same anatomical constraints we effectively estimate what would have been observed if every subject had exactly the same neuroanatomy, namely that of the standard space used in anatomical normalization and to derive the priors. Subject-specific effects, linear time effects, global effects, and the constant term were all treated as nuisance variables and were removed from the data (and design matrix) prior to estimation.[11] As above, each partition of the block diagonal first level design matrix was a single column of alternating +1 s and −1 s depending on whether the scan was an activation scan or baseline. The adjusted data were entered into the EM algorithm as above.

The results of the analysis are presented in Fig. 17 using a similar format to Figs. 15 and 16. In this instance we show the ML estimators (upper panel) that obtain from a voxel by voxel estimation, i.e., without using spatial basis functions or incorporating the point spread function into the estimation model. We omitted these components to show (i) the impoverished spatial resolution (lower panels) and (ii) partial volume effect that ensues (upper panel). The PPM and SPM are pleasingly congruent particularly in light of the realistic threshold used for the SPM$\{T\}$ (0.05 corrected). However, the enhanced spatial information available in the PPM, relative to the SPM is immediately obvious. The slice shown is at 12 mm above the ACPC line. The regions showing an activation of 0.1 ml/dl/min (Equivalents), or more, comprise Broca's area (BA 44) and contiguous premotor cortex (BA 6), an activation deep in the anterior frontal operculum and a subcortical activation centred on the mediodorsal nucleus of the thalamus. This anatomical precision could not be supported using the conventional SPM (right hand panels in Fig. 17). Equally interesting are the conditional means and ML estimates in the upper panel. The critical thing to note here is that, due to smoothing or partial volume effects, the ML estimates (dots) are too small. Because the conditional means (line) are informed about the spatial configuration of sources, subtending signals observed after convolution, they are approximately twice as large. This suggests (as we already knew) that conventional analyses can substantially underestimate true effects, especially when they are focal.

5.2 Summary

In this section we have illustrated how Bayesian estimates can supervene over classical ML estimates by harnessing constraints at higher levels. We have used spatial constraints in this example and introduced spatiotemporal models to this end. The key points include:

- The notion that observations over voxels represent repeated measures of the same neuronal response. This motivates a two-level hierarchical observation model and enables Bayesian inference at the first [voxel] level. The first-level conditional estimates are generally better than conventional ML estimates be-

[11] This is not necessary because we could have simply included these effects in the first-level design. However, given the computational load of spatiotemporal models, it is sometimes easier to treat nuisance variables and confounds as fixed effects and remove them before applying the EM algorithm. Note that when this is done the covariance constraints are assumed to hold for the adjusted data, as opposed to the original data.

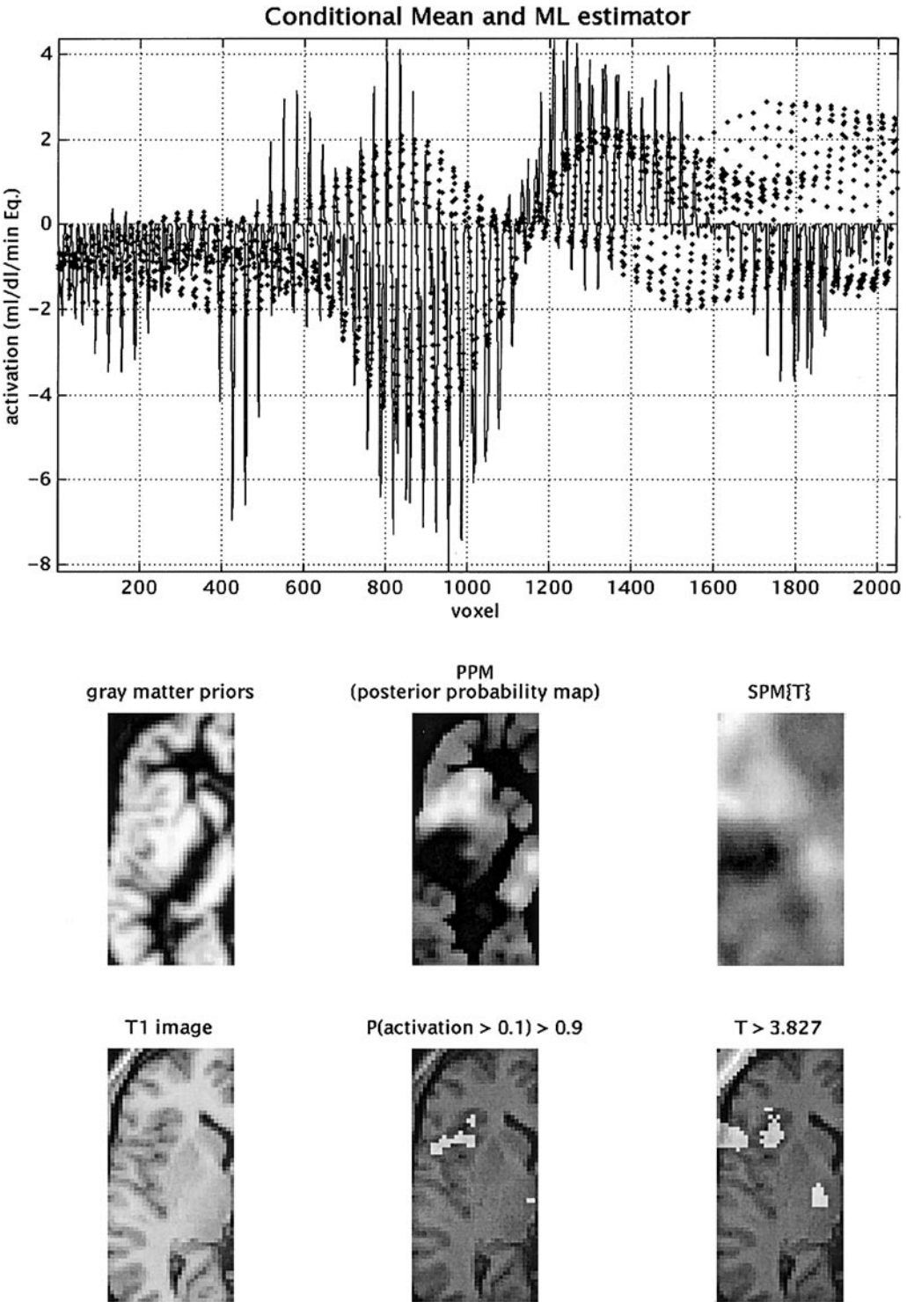


FIG. 17. Analysis of real PET data. The layout of this figure corresponds to Figs. 15 and 16 but in this instance the true activation is unknown (the lower right panel simply shows the structural image on which the priors were based). These data came from a PET study of paced FAS orthographic verbal fluency in which five subjects were asked to produce words beginning with a heard letter (word generation) or simply repeat the letter (word shadowing). These two conditions alternated for 12 scans. In this analysis we treated all 60 scans as if they came from the same subject because we wanted to illustrate second-level constraints in terms of spatial realizations over voxels. The data were smoothed with a 16-mm FWHM Gaussian kernel, which ameliorates differences in functional anatomy among subjects. The results shown are for a 32 × 64 (2 × 2 × 2 mm) voxel slice through the left hemisphere at $z = 12$ mm. The SPM{T} is based on a conventional voxel-wise approach that benefits from neither the dimension reduction (hard constraints afforded by the grey matter priors) or the Bayesian estimation (soft constraints). Consequently the classical analysis has markedly poorer anatomical resolution and is subject to partial volume

cause they are constrained by information from the second level. Furthermore, the form of Bayesian inference means one can avoid the fallacies of classical inference in the context of very small effects.

• Applying a threshold to PPMs sets an upper bound on the false discovery rate (FDR). This provides an interesting connection with FDR schemes in classical inference (Genovese *et al.,* 2002). In short, Bayesian inference offers a principled way of controlling the sensitivity and specificity of an inference by referring explicitly to the size of the activation. However, the secondary thresholding of PPMs is not an intrinsic part of Bayesian inference.

• The two-level spatiotemporal model is predicated on the assumption that the brain shows a high degree of functional integration such that responses to changing task or stimulus conditions are expressed in a distributed fashion everywhere, to a greater or lesser extent. In other words, we anticipate all brain areas respond but the vast majority do so imperceptibly with little or trivially small activations. Once in a while we encounter a brain area that is sufficiently specialized to respond in a substantial and unambiguous fashion (defined operationally by the threshold γ). By assuming the distribution of activations is approximately Gaussian we can simply enter observations over voxels, as a second-level constraint, on the estimates and inference at any one voxel. Note that this assumption is different from other proposals where voxels are assumed to be activated or not (e.g., mixture models). In these models brain regions either show zero activation or some constant mean activation. One could debate which is the more biologically plausible (see below), but our primary motivation here is to demonstrate how extensions of analyses people currently employ can bring them into a Bayesian framework.

• Anatomical priors can enter into the estimation scheme at two levels, (i) in terms of soft constraints embodied in the form of the prior covariances and (ii) in terms of hard constraints implicit in the eigenvectors of these priors that define an estimation (AIBF) subspace. The former requires a Bayesian framework, the latter does not (but can be seen as an implementation of priors with infinite precision). The former is interesting because it is formally identical to minimum norm approaches to source estimation in the EEG literature (e.g., ridge regression). In these applications the hyperparameters controlling the contribution of the spatial priors are usually fixed. The framework presented here allows not only the hyperparameters to be estimated using empirical Bayes but allows a number of different spatial priors to be entered into the model concurrently.

• By incorporating the point-spread function in the first-level design matrix (13) the estimation is effectively performing a least squares de-convolution. While this is, in principle, possible using a classical analysis the results are generally unstable and require some form of regularization, such as a smoothness constraint on the estimators. This regularization is exactly what the hierarchical model provides through the priors. As in the previous paragraph the hyperparameters controlling the degree of regularization are estimated automatically and have a much more natural interpretation, or plausible motivation, when placed within a Bayesian framework.

Some of these ideas are not inherently Bayesian in nature but all are facilitated operationally by a hierarchical framework. In short, adopting a Bayesian approach gives one the latitude to explore and use different devices and sources of information to refine the estimation and inference procedure in ways that are precluded by classical approaches.

6. CONCLUSION

In Friston *et al.* (2002) and in this paper we have provided a fairly technical but didactic introduction to the use of hierarchical observation models in functional neuroimaging. We have emphasized the points of connection between the classical perspective and Bayesian inference in an effort to show that conventional analyses of PET and fMRI data can be usefully extended within an empirical Bayes frame of reference. A critical point is that hierarchical models not only provide for appropriate inference at the last level but that one can revisit lower levels suitably equipped to make Bayesian inferences. Bayesian inferences eschew many of the fallacies of classical inference and characterize brain responses in a way that is more directly predicated on the things one is interested in.

There are a large number of potential applications of the analytic framework presented here which have not been considered. Among the more important is the generalization to dynamic and nonlinear models of neuronal signals. These and further applications, in the context of linear models, will be dealt with in sub-

---

effects in terms of the estimated activations. See the upper panel where the conditional estimates are much larger than the ML estimates in areas of activation. In this instance the ML estimates were computed on a voxel by voxel basis and, like the SPM{*T*} due not benefit from the implicit least squares deconvolution implemented when solving the spatiotemporal model in (16). Activations of 0.1 or more can be inferred with at least 90% confidence in Broca's area/premotor cortex (BA 44 and 6), deep in the frontal operculum and in the dorsomedial thalamus. The latter anatomical attribution would be precluded with the conventional analysis and is interesting given the cortical projections of this subcortical region.

sequent papers (e.g., Friston, 2002). The examples in this paper are used to illustrate ideas and serve as "proof of concept." In particular, conclusions regarding sensitivity and specificity pertain only to the specific data considered.

The drawbacks of empirical Bayes estimation and inference procedures are theoretical and practical. Practical difficulties are related simply to computational load and time of processing (see software implementation note below). The theoretical difficulties include the overconfidence problem.

6.1 The Overconfidence Problem

In PEB methods the hyperparameters pertaining to the covariance components are estimated by maximum likelihood and are then used to compute the posterior means and covariance of the parameters. From a fully Bayesian perspective the PEB approach effectively substitutes the conditional posterior density, $p(\theta|y, \lambda)$ with $\lambda = \hat{\lambda}$ for the correct posterior density given by

$$p(\theta|y) = \int p(\theta|y, \lambda)p(\lambda|y)d\lambda. \tag{19}$$

Thus the PEB approximation fails to take account of the uncertainty about the hyperparameters that enters through $p(\lambda|y)$. It is well known that the posterior mean conditional on the ReML estimate $\hat{\lambda}$ is approximately equal to the true posterior mean. However, the corresponding conditional posterior covariance is too small (Kass and Steffrey, 1989). This results in over confidence when making inferences using PEB methodology. In other words, although the empirical posterior response estimates are good approximations, inferences about them may be slightly capricious. Kass and Steffrey (1989) provide a solution to this problem for conditionally independent hierarchical models, where

$$\begin{aligned}\mathrm{Cov}(\theta|y) &= E\{\mathrm{Cov}\{\theta\}|y, \hat{\lambda}\} + \mathrm{Cov}\{E\{\theta\}|y, \hat{\lambda}\} \\ &\approx C_{\theta|y} + \frac{\partial \eta_{\theta|y}}{\partial \lambda}\,\mathrm{Cov}\{\lambda|y\}\,\frac{\partial \eta_{\theta|y}^{T}}{\partial \lambda},\end{aligned} \tag{20}$$

where the expectation and covariance are with respect to the posterior distribution of λ. This provides for a first order approximation to the second term in the first line of (20), using the delta method (i.e., a first-order Taylor expansion) and the posterior covariance of the hyperparameters (the second line of 20). This is, in principle, feasible but requires the specification of both a likelihood function, of its sufficient statistics, and a prior density function for the hyperparameters. Effectively this means replacing the ML estimate of the hyperparameters with its conditional mean and augmenting the conditional posterior covariance with the second term above. The choice of these likelihood functions and priors can sometimes be quite arbitrary and motivated by mathematical convenience (e.g., conjugate priors). While providing the basis for interesting extensions of PEB for neuroimaging we have chosen to avoid this issue by treating the hyperparameters as fixed effects and eschewing the need to specify a likelihood. Consequently we adhere to an empirical framework, which assumes that the second term in (20) is sufficiently small to be ignored. An interesting possibility is to use a PEB estimator of the hyperparameters themselves. This would lead to a truly recursive algorithm but this is beyond the scope of these papers.

6.2 Priors

A general issue in Bayesian models of spatiotemporal responses is the face validity of the model adopted. For example, in this paper we have assumed that the prior distribution of responses, over voxels, is Gaussian. The ensuing hierarchical model uses the between-voxel variability as a prior variance on the response of any single voxel. Other models could be used, for example mixture models (Everitt and Bullmore, 1999). Mixture models assume that data are generated by a small number of "causes," each parameterized by a different activation level. Inference can be based on the posterior probability that a voxel belongs a particular component of the mixture, where each component is characterized by its estimated activation. In Penny and Friston (2002) we show how mixtures of general linear models can be estimated using EM. Whether a mixture model is more or less appropriate than the hierarchical model presented in this paper is an outstanding question. Indeed knowing which model is best would implicitly resolve some fundamental questions about functional brain architectures. Mixture models can be seen as taking functional segregation to its extreme. Functional segregation posits that a particular cognitive or sensorimotor function is served by functionality that is anatomical segregated in one part of the brain. This implies that the segregated area will respond to experimental challenge and that remaining areas will not. The empirical Bayes model considered above allows for functional specialization assuming that evoked responses are distributed but are expressed much more in specialized areas. This is consistent with perspectives offered by cognitive neuroscience (e.g., parallel distributed processing). Conversely, the hierarchical model represents an extreme of functional integration, in which regional specialization is only quantitative. A simple refinement of the model in this paper could involve super-Gaussian priors that model a small number of highly responsive voxels and a large number of relatively unresponsive voxels. By hyperparameterizing non-Gaussian priors of this sort the de-

gree of segregation or modularity could be ascertained using empirical Bayes through EM.

6.3 Thresholds

A key aspect of Bayesian inference is that classical threshold based on false positive rates or specificity is replaced by a threshold specified in terms of the size of the effect of interest. This could be seen as a problem because in many cases (e.g., T2* data in fMRI) the biophysical or physiological meaning of an activation's size is obscure. However, the converse position is that the whole strength of Bayesian inference stems from the physical meaning, conferred on the inference, through the size of the effect. A Bayesian size threshold enables a much more qualified inference than simply saying the effect was not zero. An important issue here is that, although the absolute size may have no direct relationship to neuronal or synaptic activity, the relative size can. For example, a large (sensory evoked) response in PET would be, typically, 5 ml/dl/min and, for fMRI, about 1% of the signal. A subtle (e.g., cognitive modulation of a sensory evoked response) but substantial effect may correspond to 1 ml/dl/min in PET or 0.2% in fMRI. This "feel" for what constitutes a substantial response is quickly acquired by people doing neuroimaging routinely and yet plays no role in classical inference. Bayesian inference, on the other hand, embeds this useful expertise into the posterior probability.

A more fundamental point is that Bayesian inference does not really depend on any threshold. It is entirely sufficient to report the conditional density of an effect. Under Gaussian assumptions this requires two quantities for each voxel (the conditional mean and covariance). The posterior probability can be viewed as a simple device to reduce the characterization of the posterior density to a single quantity (the posterior probability). This is useful for creating summary maps (PPMs). An alternative would be to present two maps, one of the contrast's conditional expectation and another of its conditional variance. PPMs for any size threshold could be derived from these two moments. As mentioned in Friston *et al.* (2002), the secondary thresholding of the PPM with a further [probability] threshold is useful for comparison with classical inference but is not an intrinsic part of Bayesian inference.

SOFTWARE IMPLEMENTATION

The EM algorithm has been incorporated into the next release of the SPM software (current development version) by providing an assessment of nonsphericity. This allows for analyses of data with inhomogeneity of variance or arbitrary correlations among the errors. At present, the ReML formulation (see Appendix A.2 of Friston *et al.*, 2002) is used to estimate voxel-wise [stationary] nonsphericity, allowing for proper correction to the statistics and degrees of freedom. This formulation is used to estimate serial correlations in fMRI but the approach has also been extended to cover PET and other basic models where nonsphericity in repeated measure designs can be an issue. This is particularly useful in multilevel designs, in which contrasts from one level are taken to a second level, without having to assume sphericity. Because only one EM is required the computational load is small in relation to total analysis time (minutes as opposed to hours).

At the time of writing, computational limitations preclude the routine application of the theory in this paper to all voxels in the search volume. This is due primarily to memory constraints encountered when dealing with the very large matrices at the second level. In principle, the anatomical basis functions, formed by assuming priors of infinite precision, need only be computed once for a given standard anatomical space. This should facilitate routine application without any *ad hoc* preselection of voxels. However, we hope to use the same anatomical basis set in the context of EEG-fMRI integration and this additional constraint means it may be some time before the theory presented in this paper is implemented in SPM software releases.

ACKNOWLEDGMENTS

The Wellcome Trust funded the work reported in this paper.

REFERENCES

Ashburner, J., and Friston, K. J. 1997. Multimodal image co-registration and partitioning—A unified framework. *NeuroImage* **6:** 209–217.

Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc., Ser. B* **57:** 289–300.

Box, G. E. P. 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, E. T., Brammer, M. J., Williams, S. C. R., 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. *Magn. Res. Med.* **35:** 261–277.

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

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

Friston, K. J., Jezzard, P. J., and Turner, R. 1994. Analysis of functional MRI time-series. *Hum. Brain Mapp.* **1:** 153–171.

Friston, K. J., Josephs, O., Rees, G., and Turner, R. 1998. Nonlinear event-related responses in fMRI. *Magn. Reson. Med.* **39:** 41–52.

Friston, K. J., Josephs, O., Zarahn, E., Holmes, A. P., Rouquette, S., and Poline, J.-B. 2000. To smooth or not to smooth? Bias and efficiency in fMRI time-series analysis. *NeuroImage* **12:** 196–208.

Friston, K. J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., and Ashburner, J. 2002. Classical and Bayesian inference in neuroimaging: Theory. *NeuroImage* **16:** 465–483.

Friston, K. J. 2002. Bayesian estimation of dynamical systems: An application to fMRI. *NeuroImage* **16:** 513–530.

Genovese, C. R., Lazar, N. A., and Nichols, T. E. 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage,* submitted.

Geisser, S., and Greenhouse, S. W. 1958. An extension of Box's results on the use of the F distribution in multivariate analysis. *Ann. Math. Stats.* **29:** 885–891.

Hartvig, N. V., and Jensen, J. L. 2000. Spatial mixture modelling of fMRI data. *Hum. Brain Mapp.,* in press.

Henson, R. N. A., Rugg, M. D., Shallice, T., and Dolan, R. J. 2000. Confidence in recognition memory for words: Dissociating right prefrontal roles in episodic retrieval. *J. Cogn. Neurosci.,* in press.

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

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

Kiebel, S., Goebel, R., and Friston, K. J. 2000. Anatomically informed basis functions. *NeuroImage* **11:** 656–667.

Penny, W., and Friston, K. J. 2002. Mixtures of general linear models for neuroimaging. Submitted.

Phillips, C., Rugg, M. D., and Friston, K. J. 1999. Informed spatial basis functions in minimum norm solutions for the electromagnetic source localisation problem. *Biomedizinische technik.* **2:** 87–90.

Purdon, P. L., 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.

Satterthwaite, E. F. 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.

Worsley, K. J., and Friston, K. J. 1995. Analysis of fMRI time-series revisited—Again. *NeuroImage* **2:** 173–181.

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