

# Bayesian Comparison of Spatially Regularised General Linear Models

W. Penny<sup>1</sup>, G. Flandin<sup>1</sup>, N. Trujillo-Barreto<sup>2</sup>  
Wellcome Department of Imaging Neuroscience<sup>(1)</sup>,  
University College, London WC1N 3BG.  
Cuban Neuroscience Center, Havana, Cuba<sup>(2)</sup>

July 29, 2005

## Abstract

In previous work [34] we have developed a spatially regularised General Linear Model (GLM) for the analysis of fMRI data which allows for the characterisation of regionally specific effects using Posterior Probability Maps (PPMs). In this paper we show how it also provides an approximation to the model evidence. This is important as it is the basis of Bayesian model comparison and provides a unified framework for Bayesian Analysis of Variance (ANOVA), Cluster of Interest (COI) analyses and the principled selection of signal and noise models. We also provide extensions that implement spatial and anatomical regularisation of noise process parameters.

## 1 Introduction

Functional Magnetic Resonance Imaging (fMRI) using Blood Oxygen Level Dependent (BOLD) contrast is an established method for making inferences about regionally specific activations in the human brain [12]. From measurements of changes in blood oxygenation one uses various statistical models, such as the General Linear Model (GLM) [16], to make inferences about task-specific changes in underlying neuronal activity.

In this paper we propose analysing fMRI using a Bayesian Model Comparison (BMC) framework based on spatially regularised GLMs. Whilst model

comparison can be thought of as a secondary concern, used primarily for fine-tuning, an alternative perspective places it at the heart of the scientific endeavour. This is because in any mature scientific discipline there will be a candidate set of hypotheses. BMC can then be used to update ones beliefs about the competing hypotheses in light of experimental data. A more prosaic example is the analysis of data from factorial experimental designs using Analyses of Variance (ANOVA). This is a mainstay of scientific research [39]. To infer that manipulation of an experimental factor caused a significant effect one compares two models, one with that factor and one without.

In neuroimaging, BMC is used in the analysis of functional integration [33]. This allows inferences to be made about effective connectivity and how that connectivity changes as a function of perceptual or cognitive set. In analyses of functional specialisation BMC has been used to select the optimal order of autoregressive noise models [32].

Model comparison can be implemented using classical or Bayesian inference. In classical inference, however, one is restricted to comparing nested models [18]. Whilst this is sufficient for ANOVA, it is suboptimal in other domains. In this paper we show that a non-nested approach is optimal for the comparison of hemodynamic basis sets.

In previous work we have developed a Bayesian framework that allows inferences to be made about regional activations using Posterior Probability Maps (PPMs) [34]. This has been extended by incorporating a spatial prior embodying our knowledge that evoked responses are spatially contiguous [34]. A key feature of that work is that it provides an approximation to the model evidence.

In this paper we show how the model evidence can be used for model comparison and give details of the necessary computations. This provides a unified framework for Bayesian ANOVAs, Cluster of Interest (COI) analyses and the principled selection of signal and noise models. We also describe extensions to the framework that implement spatial and anatomical regularisation of noise process parameters.

The paper is structured as follows. In section 2.1 we describe Bayesian model comparison and show how the model evidence can be approximated. In section 2.2 we describe our probabilistic model for fMRI and show how approximate inference can proceed. We also describe a variant of the model using ‘tissue-type priors’ that make use of anatomical information. In the results section the method is applied to simulated data to illustrate the properties of COI analysis and to compare nested versus non-nested model com-

parison of hemodynamic basis sets. Results on an event-related fMRI data set illustrate Bayesian selection of signal and noise models and Bayesian ANOVA.

## 2 Theory

### 2.1 Bayesian Model Comparison

#### 2.1.1 Posterior model probabilities

Given a set of probabilistic models indexed by  $m = 1..M$  and a data set  $Y$ , Bayesian Model Comparison (BMC) can be implemented as follows [24]. Firstly, one requires a prior distribution over models,  $p(m)$ . Typically this will be a uniform distribution indicating that no model is favoured a priori. One then needs the evidence for model  $m$ ,  $p(Y|m)$ . The evidence is not straightforward to compute but the next section shows how it can be approximated. From the model prior and model evidence one can then compute the posterior probability of model  $m$  using Bayes rule

$$p(m|Y) = \frac{p(Y|m)p(m)}{\sum_{m'} p(Y|m')p(m')} \quad (1)$$

This posterior distribution can be used in a number of ways. In Bayesian Model Averaging (BMA)  $p(m|Y)$  provides a weighting for combining model predictions. This has been used, for example, to improve EEG source localisation [38].

In this paper we use  $p(m|Y)$  for Bayesian model comparison. In section 4.1.3, for example, we compare the evidence of models with different hemodynamic basis sets. All the examples in this paper assume a uniform prior over models ie.  $p(m) = 1/M$ . The model with the highest posterior probability will therefore also have the highest evidence. This means that the model evidence alone can also be used for model selection.

#### 2.1.2 Approximating the model evidence

Given that model  $m$  has parameters  $\theta$ , the evidence for model  $m$  can be written as

$$p(Y|m) = \frac{p(Y, \theta|m)}{p(\theta|Y, m)} \quad (2)$$

Taking logs, and writing the log-evidence as  $L(m) \equiv \log p(Y|m)$  gives

$$L(m) = \log p(Y, \theta|m) - \log p(\theta|Y, m) \quad (3)$$

If we now take expectations with respect to, what for the moment we will regard as an arbitrary distribution,  $q(\theta|Y, m)$  we get

$$L(m) = \int q(\theta|Y, m) \log p(Y, \theta|m) d\theta - \int q(\theta|Y, m) \log p(\theta|Y, m) d\theta \quad (4)$$

This can be re-arranged as follows

$$L(m) = F(m) + KL[q(\theta|Y, m), p(\theta|Y, m)] \quad (5)$$

where the first term is known as the negative free energy [31]

$$F(m) = \int q(\theta|Y, m) \log \frac{p(Y, \theta|m)}{q(\theta|Y, m)} d\theta \quad (6)$$

and the second term is the Kullback Leibler (KL) divergence [11] which can be written generically for any probability densities  $q(x)$  and  $p(x)$  as

$$KL[q(x), p(x)] = \int q(x) \log \frac{q(x)}{p(x)} dx \quad (7)$$

KL measures the discrepancy between two probability densities. It is equal to zero if the densities are identical and greater than zero otherwise. Because  $KL \geq 0$ , equation 5 tells us that  $L(m) \geq F(m)$ . That is, the log model evidence is bounded below by  $F$ , and the closer  $q(\theta|Y, m)$  is to  $p(\theta|Y, m)$ , the tighter the bound.

Equation 5 describes the fundamental relationship between model evidence, free energy and KL-divergence. This relationship is used in the inference framework known as Variational Bayes (VB) [28]. In VB, the parameters of an approximate posterior density,  $q(\theta|Y, m)$  (see section 2.2.3) are updated to maximise  $F(m)$ . This therefore maximises a lower bound on the model evidence. A number of methods now use this approach in the analysis of neuroimaging data [41, 32, 34, 36, 37].

Model comparison proceeds using  $F(m)$  as a surrogate for the model evidence, under the assumption that the bound in equation 5 is tight. This will be the case if the approximate posteriors are close to the true posteriors.

In our previous work on modelling fMRI time series (see section 5 in [32]) we have used Gibbs sampling to show that this is indeed the case.

It is also possible to approximate the model evidence using sampling methods [18, 5]. In the very general context of probabilistic graphical models, Beal and Ghahramani [5] have shown that the VB approximation of model evidence is considerably more accurate than the Bayesian Information Criterion (BIC) whilst incurring little extra computational cost. Moreover, model selection using VB is of comparable accuracy to a much more computationally demanding method based on Annealed Importance Sampling (AIS) [5].

### 2.1.3 Computing the free energy

In this paper we will use  $F(m)$  as an approximation to the log evidence. This will form the basis of all model comparisons. This quantity can also be expressed in a more convenient form. If we expand the joint density  $p(Y, \theta|m) = p(Y|\theta, m)p(\theta|m)$  and collect terms it can be written

$$F(m) = V(m) - KL[q(\theta|Y, m), p(\theta|m)] \quad (8)$$

where  $V(m)$  is the average log likelihood

$$V(m) = \int q(\theta|Y, m) \log p(Y|\theta, m) d\theta \quad (9)$$

and  $KL[q(\theta|Y, m), p(\theta|m)]$  is the divergence between the approximate posterior and the *prior*. The quantity  $F(m)$  therefore comprises two terms (i) an accuracy term, the average log likelihood, and (ii) a complexity term, the KL divergence. This can be viewed as a complexity term because, as the number of parameters grows (ie. the dimension of  $\theta$ ) so does the KL.

## 2.2 Analysis of fMRI time series

To apply the model comparison framework we need a set of models  $m$  and for each model we must specify a set of parameters,  $\theta$ , a prior distribution of those parameters,  $p(\theta|m)$  and the likelihood  $p(Y|\theta, m)$ . Together, the likelihood and prior define a probabilistic generative model which we describe in the following section.

### 2.2.1 Generative model

We write an fMRI data set consisting of  $T$  time points at  $N$  voxels as the  $T \times N$  matrix  $Y$ . In mass-univariate models [16], these data are explained in terms of a  $T \times K$  design matrix  $X$ , containing the values of  $K$  regressors at  $T$  time points, and a  $K \times N$  matrix of regression coefficients  $W$ , containing  $K$  regression coefficients at each of the  $N$  voxels. The model is written

$$Y = XW + E \quad (10)$$

where  $E$  is a  $T \times N$  error matrix.

It is well known that fMRI data is contaminated with artifacts. These stem primarily from low-frequency drifts due to hardware instabilities, aliased cardiac pulsation and respiratory sources, unmodelled neuronal activity and residual motion artifacts not accounted for by rigid body registration methods [40]. This results in the residuals of an fMRI analysis being temporally autocorrelated.

In previous work we have shown that, after removal of low-frequency drifts using Discrete Cosine Transform (DCT) basis sets, low-order voxel-wise autoregressive (AR) models are sufficient for modelling this autocorrelation [32]. It is important to model these noise processes as parameter estimation becomes less biased [17] and more accurate [32]. Together, DCT and AR modelling can account for long-memory noise processes. Alternative procedures for removing low-frequency drifts include the use of running-line smoothers or polynomial expansions [26].

In this paper, we adopt the approach taken in previous work. For a  $P$ th-order AR model, the likelihood of the data is given by (see equation 10 in [32])

$$p(Y|W, A, \lambda) = \prod_{t=P+1}^T \prod_{n=1}^N \mathbf{N}(y_{tn} - x_t w_n; (d_{tn} - X_t w_n)^T a_n, \lambda_n^{-1}) \quad (11)$$

where  $\mathbf{N}()$  is the Normal density defined in Appendix E1, at the  $n$ th voxel,  $a_n$  is a  $P \times 1$  vector of autoregressive coefficients,  $w_n$  is a  $K \times 1$  vector of regression coefficients and  $\lambda_n$  is the observation noise precision. The vector  $x_t$  is the  $t$ th row of the design matrix and  $X_t$  is a  $P \times K$  matrix containing the previous  $P$  rows of  $X$  prior to time point  $t$ . The scalar  $y_{tn}$  is the fMRI scan at the  $t$ th time point and  $n$ th voxel and  $d_{tn} = [y_{t-1,n}, y_{t-2,n}, \dots, y_{t-P,n}]^T$ . This shows that higher model likelihoods are obtained when the prediction error

$y_{tn} - x_t w_n$  is closer to what is expected from the AR estimate of prediction error. Because  $d_{tn}$  depends on data  $P$  time steps before, the likelihood is evaluated starting at time point  $P + 1$ , thus ignoring the GLM fit at the first  $P$  time points.

The voxel wise parameters  $w_n$  and  $a_n$  are contained in the matrices  $W$  and  $A$  and the voxel-wise precisions  $\lambda_n$  are contained in  $\lambda$ . Appendices A, B and C describe the prior distributions over these parameters. Appendix B, for example, describes a prior over regression coefficients that enforces an automatic spatial regularisation using eg. Low Resolution Tomography (LORETA) or Gaussian Markov Random Field (GMRF) priors. These have been described in detail in previous work [34]. Together, the likelihood and priors define the generative model, which is portrayed graphically in Figure 1. This generative model is identical to that described in our previous work [34], except that we have augmented the model so that the AR coefficients are regularised as described in the next section.

### 2.2.2 AR priors

It is well established that the amount of temporal autocorrelation in fMRI data varies as a function of voxel position. This can be modelled using voxel-wise AR processes [10, 32, 42, 40].

It has also been observed that the autocorrelation varies as a function of tissue type ie. grey matter, white matter or Cerebro-Spinal Fluid (CSF). For example, in AR(1) models, larger coefficients are observed in CSF [32].

It is an open question, however, as to whether tissue type is sufficient to explain the observed spatial variability. In this paper we address this question from a model comparison perspective by comparing two types of model. Each model regularises the estimation of voxel-wise AR coefficients in a different way.

The first type of model uses a ‘tissue-type prior’ which we define as follows. Firstly, each voxel is labelled as belonging to one of  $S$  discrete categories. For example,  $s = \{1, 2, 3\}$  could correspond to the voxel belonging to (1) grey matter, (2) white matter or (3) CSF. This information can be derived from a segmentation of registered structural images [2]. Secondly, archetypal AR coefficient vectors are associated with each category. This is implemented by specifying a Gaussian distribution for each category. Appendix F describes these priors mathematically and shows how the means and precisions of the Gaussians can be estimated from the data.

The second type of model uses a spatial prior that takes into account voxel position. Following Woolrich et al. [41] we use a GMRF prior. This has been shown to improve estimation of AR parameters, especially for the lower order coefficients. This prior is defined mathematically in Appendix C. It is also illustrated in the generative model in Figure 1. Similar spatial regularisation procedures, but based on Gaussian kernels, have been proposed in the context of classical inference [42, 17].

In section 4.1.1 we use simulations to show that our model comparison procedures are capable of detecting the correct type of variation (eg. spatial versus tissue-type). In section 4.2.1 we will see which is a better model for fMRI.

### 2.2.3 Approximate Posteriors

This paper uses the Variational Bayes framework [28] for estimation and inference. This requires the specification of an approximate posterior distribution whose parameters are updated so as to maximise the negative free energy, as described earlier in section 2.1.2.

This paper uses the algorithm described in previous work [34] in which we assume that the approximate posterior factorises over voxels and subsets of parameters. This leads to a set of equations for updating the sufficient statistics of components of the approximate posterior shown in Figure 2. These update equations are also provided in the appendices. These appendices are self-contained except for a number of quantities that are marked out using the ‘tilde’ notation. These are  $\tilde{A}_n, \tilde{b}_n, \tilde{C}_n, \tilde{d}_n$  and  $\tilde{G}_n$  which are all defined in Appendix B of [32].

Derivations of update equations that are new to this paper (see eg. Appendixes C.2, F.2, F.4 and F.5) have been omitted but follow the standard variational approach which is also described in Appendix A of [34].

The central quantity of interest in fMRI analysis is our estimate of effect sizes, embodied in contrasts of regression coefficients. A key update equation in our VB scheme is, therefore, the approximate posterior for the regression coefficients. This is given by equation 25 in the appendix. For the special case of temporally uncorrelated data we have

$$\begin{aligned}\hat{\Sigma}_n &= (\bar{\lambda}_n X^T X + \bar{B}_{nn})^{-1} \\ \hat{w}_n &= \hat{\Sigma}_n (\bar{\lambda}_n X^T y_n + r_n)\end{aligned}\tag{12}$$

where  $\bar{B}$  is a spatial precision matrix and  $r_n$  is the weighted sum of neighboring regression coefficient estimates.

This update therefore indicates that the regression coefficient estimate at a given voxel regresses towards those at nearby voxels. This is the desired effect of the spatial prior and it is preserved despite the factorisation over voxels in the approximate posterior. Equation 12 can be thought of as the combination of a temporal prediction  $X^T y_n$  and a spatial prediction from  $r_n$ . Each prediction is weighted by its relative precision to produce the optimal estimate  $\hat{w}_n$ . In this sense, the VB update rules provide a spatio-temporal deconvolution of fMRI data. Moreover, the parameters controlling the relative precisions,  $\bar{\lambda}_n$  and  $\bar{\alpha}$  are estimated from the data. This means that our effect size estimates derive from an automatically regularised spatio-temporal deconvolution.

#### 2.2.4 Computing the free energy

The negative free energy,  $F(m)$ , will be used to approximate the model evidence and can be computed using the expression in equation 8. It comprises two types of term; the average log likelihood and the KL terms. Appendix D shows how the average log likelihood is computed. For the KL terms we have (dropping the  $m$ 's)

$$\begin{aligned} KL[q(\theta|Y), p(\theta)] &= KL[q(W), p(W)] + KL[q(A), p(A)] \\ &+ KL[q(\lambda), p(\lambda)] + KL[q(\alpha), p(\alpha)] + KL[q(\beta), p(\beta)] \end{aligned} \quad (13)$$

where  $W$  are regression coefficients,  $A$  are AR coefficients,  $\lambda$  are observation noise precisions and  $\alpha$  and  $\beta$  are the spatial regularisation coefficients for  $W$  and  $A$ , respectively (see Figure 1). Appendix E shows in detail how each KL term is computed. It is also possible to re-arrange the computation of  $F(m)$  to make it more efficient, as a number of terms in the average log-likelihood cancel with those in the KL expressions (see eg. [30]). We have not used this re-arrangement, however, as it compromises the readability and extendability of the implementation.

When comparing models with the same type of spatial prior (ie. same  $D_w$  and  $D_a$ ), and the same number of regression coefficients,  $K$ , and the same number of AR coefficients,  $P$ , there is no need to compute terms involving  $\log |D_w|$  or  $\log |D_a|$ . This saves time, especially for slices with large numbers

of voxels. Otherwise, this log-determinant must be computed<sup>1</sup>. This can be implemented by eigendecomposition and then taking the sum of the log of eigenvalues greater than machine precision. This last step is necessary as the matrices are not full rank [34].

Because our approximation to the model evidence depends on the above KL terms, it will also depend on the constants that define the priors at the highest level of the model. These are the priors  $p(\lambda_n), p(\alpha_k)$  and  $p(\beta_p)$  which have associated constants  $q_1, q_2, r_1, r_2, u_1$  and  $u_2$  (see appendix A.1). In previous work, however, we have shown for example that the optimal AR model order is robust to variations in  $\bar{\beta}_p = r_1 r_2$  over several orders of magnitude (see section 5.2 in [32]).

### 2.2.5 Unique contributions

It is possible to decompose the evidence for each model into a sum of unique contributions from each voxel

$$F(m) = \sum_n U_n(m) \quad (14)$$

where

$$\begin{aligned} U_n(m) &= V(n) - KW(n) - KA(n) - KL_{Ga}[q(\lambda_n), p(\lambda_n)] \\ &\quad - \frac{1}{N} (KL[q(\alpha), p(\alpha)] + KL[q(\beta), p(\beta)]) \end{aligned} \quad (15)$$

The computation of these voxel-specific terms is described in detail in appendices D and E.

Breaking the evidence down into contributions from each voxel has two advantages. Firstly, the update equations need only be applied at voxels whose contribution,  $U_n(m)$ , is still increasing. We envisage that this could speed the estimation process, although this has yet to be implemented. Secondly, the differences in voxel-wise contributions between two models can be used to plot Posterior Probability Maps (PPMs). For example, given two models with equal priors the posterior probability of model 2 at voxel  $n$  is given by

$$p(m = 2|Y, n) = \frac{\exp(U_n(2))}{\exp(U_n(1)) + \exp(U_n(2))} \quad (16)$$

---

<sup>1</sup>A promising alternative to GMRF and LORETA priors are the thin plate spline priors defined in [9]. These have the benefit that the determinant has a known algebraic form making computation of the log-determinant much simpler.

Because  $U_n(m)$  is a contribution to the log evidence rather than the log-evidence per se, maps based on  $U_n(m)$  are ‘pseudo’-PPMs rather than PPMs proper. Nevertheless, they should be useful in characterising regionally specific effects. This method is used in the context of Bayesian ANOVAs in section 4.2.2.

The pseudo-PPMs we have defined are conceptually different from PPMs proper. They are not a numerical approximation to proper PPMs. This is because our models are spatially extended and the model evidence is only defined for a slice of spatially extended data. In the absence of spatial correlation pseudo-PPMs will correspond to proper PPMs.

### 2.2.6 Hemodynamic basis sets

It is well known that the shape of the hemodynamic response varies from voxel to voxel and from subject to subject [44]. This is accounted for in the context of GLM analyses by characterising the response using a hemodynamic basis set [12]. For example, Friston et al. [13] proposed the use of a ‘canonical’ basis function composed of a sum of gamma functions. This can be augmented to include two other basis functions, the derivative of the canonical with respect to time and the derivative with respect to dispersion. Together, these basis functions constitute an ‘Informed’ basis set [19]. For event-related designs, other authors have proposed ‘selective averaging’ procedures. These are formally equivalent to the use of a Finite Impulse Response (FIR) basis set [19].

In this paper we consider use of the following hemodynamic basis sets. Unless otherwise specified the basis functions cover a 32s period post-stimulus. We use the seven standard options available in the Statistical Parametric Mapping (SPM) software [1] (i) Inf-1: the canonical response, (ii) Inf-2: the canonical plus temporal derivative, (iii) Inf-3: the canonical plus temporal and dispersion derivatives, (iv) F: a Fourier basis set with 10 sinusoids covering 20s, (v) FH: as (iv) but with Hanning windows, (vi) Gamm3: a set of three Gamma basis functions, (vii) FIR: a finite impulse response with 10 2-second windows.

### 3 Face fMRI data

This paper uses an event-related fMRI data set acquired by Henson et al. [22]. This data and a full description of the experiments and pre-processing are available from <http://www.fil.ion.ucl.ac.uk/spm/data/>. The data were acquired during an experiment concerned with the processing of images of faces [22]. This was an event-related study in which greyscale images of faces were presented for 500ms, replacing a baseline of an oval chequerboard which was present throughout the interstimulus interval (see Figure 3). Some faces were of famous people and were therefore familiar to the subject and others were not. Each face in the database was presented twice. This paradigm is a two-by-two factorial design where the factors are familiarity and repetition. The four experimental conditions are ‘U1’, ‘U2’, ‘F1’ and ‘F2’ which are the first or second (1/2) presentations of images of familiar ‘F’ or unfamiliar ‘U’ faces. The design is shown pictorially in Figure 3.

Images were acquired from a 2T VISION system (Siemens, Erlangen, Germany) which produced T2\*-weighted transverse Echo-Planar Images (EPIs) with BOLD contrast. Whole brain EPIs consisting of 24 transverse slices were acquired every two seconds resulting in a total of T=351 scans. All functional images were realigned to the first functional image using a six-parameter rigid-body transformation. To correct for the fact that different slices were acquired at different times, time series were interpolated to the acquisition time of the reference slice. Images were then spatially normalized to a standard EPI template using a nonlinear warping method [3].

To implement a classical SPM analysis using Random Field Theory one usually spatially smooths the data at this stage [7]. But because our model incorporates a spatial prior that automatically determines the optimal amount of spatial regularization, this smoothing step is unnecessary.

We then computed the global mean value,  $g$ , over all time series, excluding non-brain voxels, and scaled each time series by the factor  $100/g$ . This makes the units of the regression coefficients ‘percentage of global mean value’. Each time series was then high-pass filtered using a set of discrete cosine basis functions with a filter cut-off of 128 seconds.

A structural scan was also acquired. This was normalised to the mean functional image and segmented into grey matter, white matter and CSF using the algorithm described in [2]. Analysis of the functional data was restricted to within-brain voxels, as identified by the structural segmentation.

## 4 Results

### 4.1 Simulated data

#### 4.1.1 Comparing noise models

In this section we compare different assumptions about the prior distribution of AR coefficients using simulated data. The use of spatial GMRF priors assumes that AR coefficients vary smoothly across a slice whereas the use of tissue-type priors assumes that they vary about a small number of typical values. The simulations use first order AR models for simplicity. If the Signal to Noise Ratio (SNR) is sufficiently high, and our approximation,  $F$ , to the model evidence is sufficiently accurate then we should be able to use  $F$  to identify which prior was used to generate the data.

The leftmost column in Figure 4 shows four different profiles of AR(1) coefficients. The first three profiles have one, two and three different typical values and the fourth has values that vary continuously as a function of position.

For each AR(1) profile we generated data as follows. We used a design matrix comprising two regressors, the first being a boxcar with a period of 20 scans and the second a constant. The design matrix,  $X$ , is therefore of dimension  $T \times K$  ( $K = 2$ ) and we chose  $T = 100$  scans. We then used a  $K \times N$  element regression coefficient matrix,  $W$ , whose elements were all set to 0.5. We chose  $N = 64$  so that the coefficients could be reshaped into  $8 \times 8$  images for display purposes. From this, we generated a simulated fMRI signal  $XW$  of dimension  $T \times N$ . We also generated a simulated noise fMRI signal by first generating a Gaussian IID noise sequence with mean zero and precision  $\lambda = 1$ . For each voxel  $n$ , we then introduced a noise correlation as determined by the value of the AR coefficient at that voxel. The simulated data  $Y$  then comprised the simulated fMRI signal plus the simulated fMRI noise. Overall, we generated four data sets having the profile of AR(1) coefficients shown in Figure 4.

We then fitted each data set with four different models. The corresponding estimated profiles are shown in columns two to five in Figure 4. Columns two to four are estimates from tissue-type priors with 1, 2 and 3 different categories, respectively. The fifth column contains estimates from a model with spatial GMRF priors. The final column shows the posterior model probabilities,  $p(m|Y)$ . These posterior probabilities are computed using the constraint

$\sum_i p(m_i|Y) = 1$ , where  $i$  indexes the model and  $p(m_i|Y) \propto p(Y|m_i) \propto F(m_i)$  (under flat model priors  $p(m_i)$ ). These show that the approximation to the model evidence (see section 2.1.2) is sufficiently accurate for the method to correctly detect the type of spatial structure in the data.

The estimate of  $F(m_i)$  entailed designating each voxel to the appropriate ‘tissue-type’ for the first three models. This is a component that is eschewed by the autoregressive model with spatial GMRF priors.

As noted in section 2.2.4, our approximation to the model evidence is dependent on the constants that define the priors at the highest level of the model. We investigated this dependence by repeating the model fitting using different values for the relevant constants. Varying the prior mean precision of AR coefficients  $\bar{\beta} = r_1 r_2$  between 1 and 100 did not have a major effect the choice of optimal model order.

#### 4.1.2 Cluster of interest analysis

If one has a strong prior hypothesis about the potential location of an activation then a Region Of Interest (ROI) analysis can be made. A region comprising a number of voxels is first chosen. This is often identified using localizer contrasts or scans (see eg. [23]). A single time series is then extracted using Principal Component Analysis or Singular Value Decomposition [8], the mean operator or multiplication with a user-specified activation shape [29]. Analysis is then based on this single ‘summary’ time series.

This section describes an alternative approach that we call Cluster of Interest (COI) analysis. Again, a region comprising a number of voxels is first chosen but the analysis is based on all time series in that region. The approach may also be viewed as a Bayesian cluster-level inference as it shares the fundamental property of classical cluster level inference that anatomical specificity is traded off for increased sensitivity [15].

To illustrate the properties of a COI analysis we use simulations based on the experimental design of the face fMRI data in which models with different design matrices are compared. Design matrix 1, a ‘null’ model, comprises a column of 1s to model the mean response at a voxel. Design matrix 2 has a single additional experimental condition which was the presentation of a face regardless of factor or level. This was convolved with the canonical hemodynamic response function. Evidence in favour of a model using design matrix 2 allows one to infer that there was an average response to faces.

Two types of data were generated, type 1 data sets using design matrix

1 and type 2 data sets using design matrix 2. Both types of data were generated for two-dimensional clusters containing  $N = 1, 4, 9, 16$  and 25 voxels. For each size of cluster we generated 500 data sets of each type. Overall,  $2 \times 5 \times 500 = 5,000$  data sets were generated. Both types of data were generated using regression coefficients fixed at unity across the patch. For data type 2 this represents a cluster of voxels that is uniformly active. The observation noise variance was set so that a range of sensitivities would be observed. This was achieved using a voxel-wise Signal to Noise Ratio (SNR) of 0.2 (we define SNR as the ratio of signal to noise standard deviation). This is very small for fMRI. The same observation noise variance was used for both types of data.

For each data set we then fitted four models; design matrices 1 and 2 with spatial GMRF priors and design matrices 1 and 2 with shrinkage priors. These shrinkage priors have been used in previous work [14] and do not make use of spatial information. A cluster was declared to be ‘active’ if the posterior probability of the model using design matrix 2 was greater than 0.999. We then computed the sensitivity and specificity of the inference over the 500 instances for each  $N$ . Overall,  $4 \times 5,000 = 20,000$  models were fitted.

The specificity was found to be 100% for all sizes of cluster and for both types of prior. Figure 5 shows a plot of sensitivity as a function of number of voxels. This indicates that the effect is too weak to be detected at the single voxel level. But the signal is increasingly detectable as the cluster size increases. This shows the power of cluster-level inference. Weak, diffuse signals can be detected at the ‘cluster-level’ that can’t be detected at the ‘voxel-level’. The figure also shows that use of a spatial prior markedly increases this sensitivity. For the cluster containing 9 voxels the sensitivity is increased by over 30%.

In a second set of simulations we repeated the above process but the regression coefficients were set to conform to a non-uniformly activated cluster. A spatial Gaussian shape was chosen. The SNR was again chosen to obtain a range of sensitivities. This was achieved using  $\text{SNR}=0.4$ . We again generated two types of data sets, with 500 instances of each type for each value of  $N$ .

Each data set was then fitted with six models; design matrices 1 and 2 with spatial GMRF priors, design matrices 1 and 2 applied to mean cluster activity, and design matrices 1 and 2 applied to the Principal Component (PC) of cluster activity. The first two models are used to assess the COI approach and the last four to assess two different ROI approaches. Clearly,

the spatial parameters are redundant when modelling univariate summaries of regional responses like the mean or regional eigenvariate.

Clusters were again declared to be ‘active’ if the posterior probability of the model using design matrix 2 was greater than 0.999. Figure 6 shows a plot of sensitivity as a function of number of voxels, for each of the three approaches. Sensitivity reaches a peak for the cluster having 9 voxels. It then falls off due to the Gaussian nature of the spatial activation profile. This is seen most severely for the ROI approaches. Sensitivity is lowest using the mean time series and slightly higher using the PC time series. For the largest cluster, the COI approach is 20 to 30% more sensitive than the ROI approaches.

These properties hold for all types of model comparison, whether it be an inference about a main effect, interaction (see section 4.2.2) or selection of a hemodynamic basis set.

### 4.1.3 Comparing hemodynamic basis sets

This section describes the properties of nested versus non-nested model comparison in the context of selecting an optimal hemodynamic basis set. As in the previous section, we use simulations based on the experimental design of the face fMRI data. Two types of data set are generated. The design matrix of type 1 data comprises the time series of delta functions indicating the presentation of a face, regardless of factor or level, convolved with Inf-3, the Informed basis set (see section 2.2.6). Type 2 data sets use the same time series of delta functions but convolved with an FIR basis set having 10 time bins. Both types of data used design matrices also containing a constant term.

Type 2 data sets used FIR coefficients that were set to resemble the canonical response but with a pronounced undershoot. These non-canonical undershoots have been observed by the authors in previous work, so make an interesting hypothetical signal. Type 1 data sets were generated by projecting noise-free type 2 data onto the Inf-3 basis set. Type 2 data therefore contains a subtle effect that can be captured by an FIR basis but not by Inf-3. Noise-free versions of each type of data are shown in Figure 7. The difference between these noise-free time series constitutes our signal of interest.

Noise was then added to each type of data set, as in the previous section, such that the noise level would provide a range of sensitivities. This was achieved using SNR=0.6. Again, 500 data sets of each type were generated

for each cluster size. Overall,  $2 \times 5 \times 500 = 5,000$  data sets were generated.

We then fitted three different models to each data set. Model 1 used an Inf3-basis, model 2 an FIR basis and model 3 used an augmented design matrix containing both an Inf3 and FIR basis. Model 1 is therefore ‘nested’ within model 3. Comparing the evidence of model 2 to model 1 constitutes a non-nested model comparison, whereas comparing the evidence of model 3 to model 1 constitutes a nested model comparison. The equivalent nested model comparison in classical inference is a standard approach for comparing basis sets in functional imaging using the ‘extra sum of squares’ principle [21].

Figure 8 plots the sensitivity of nested and non-nested model comparison approaches. Firstly, we note that the subtle undershoot effect cannot be detected at the voxel level. But as the clusters get larger the effect becomes increasingly detectable. Moreover, it is clear from the figure that non-nested model comparison is more sensitive. For the cluster with 9 voxels, for example, the non-nested approach is nearly twice as sensitive.

## 4.2 Face fMRI data

### 4.2.1 Comparing noise models

This section compares spatial versus tissue-type AR priors on the face fMRI data. We used a GLM with a design matrix where each level of each factor is represented separately. Each event type was convolved with the Inf-2 basis set. An additional constant term gives 9 regressors.

We then compared a number of approaches for specifying the AR component of the model. Whilst the model evidence can be used to select the optimal model order, as shown in [32], the focus of this section is on comparing spatial versus tissue-type priors. We therefore used an AR model order of  $P = 1$  for all comparisons.

We compared four different priors. The first three are tissue-type priors which use (i) a single Gaussian for all voxels (ii) three Gaussians, one for voxels in grey matter, one for white matter and one for CSF, (iii) two Gaussians, one for voxels in CSF (as defined using a spatially smoothed and thresholded CSF mask) and the other for the remaining voxels and (iv) a spatial GMRF prior. Prior (iii) was included in an attempt to improve the correspondence between the observed structure in the functional data (AR images) and the segmentation obtained from the structural data. Various spatial smoothing and thresholding operators were fine-tuned so as to obtain the best results

possible.

Figure 9 shows the estimated AR coefficients for priors (iii) and (iv) for selected slices. The corresponding images for priors (i) and (ii) are visually very similar to the prior (iii) images. Estimates using the GMRF prior are smoother, as one would expect. We now turn to a comparison of the model evidence.

Prior (i) had the lowest evidence in 95% slices. This indicates that there is tissue-type structure in the pattern of AR coefficients (as priors (ii) and (iii) had higher evidence). Whilst this is evident from the images themselves and is widely recognised, our framework allows this inference to be made using a statistical test by evaluating the posterior beliefs that correspond to the differences in evidence. This posterior belief was unity for 95% slices.

Prior (ii) had the third highest evidence in 70% slices and prior (iii) had the second highest evidence in 70% of slices. This reflects our extensive efforts to improve the correspondence between the observed structure in the functional data (AR images) and the segmentation obtained from the structural data.

Despite these efforts, however, models with the spatial GMRF prior had the highest evidence in all slices examined. This shows that although tissue-type effects are strong, they are not sufficient to explain the observed spatial variability in temporal autocorrelation.

#### 4.2.2 ANOVA

We now present a Bayesian Analysis of Variance (ANOVA) for the face fMRI data. The presentation of faces conforms to a factorial design with two factors, familiarity and repetition. There are therefore four putative effects of interest (i) the average effect of presenting faces, (ii) the main effect of repetition (iii) the main effect of familiarity and (iv) the interaction between repetition and familiarity.

As described in [39] (see also [20]), ANOVA is fundamentally a model comparison procedure. To test for our putative effects we therefore fitted a number of models to the data. For each model, all experimental conditions were modelled by convolving the appropriate stimulus functions with the Inf-2 hemodynamic basis set. The design matrix of Model 1, a ‘null’ model, comprised a column of 1s to model the mean response at each voxel. Model 2 had a single additional experimental condition, which was the presentation of a face regardless of factor or level. Model 3 had two conditions, first

or second repetition regardless of familiarity. Model 4 had two conditions, familiar or unfamiliar regardless of repetition. Model 5 had a design matrix containing all the conditions from models 3 and 4 (ie. both main effects but no interaction). Model 6 had a ‘full’ design matrix comprising four conditions where each level of each factor is entered separately (ie. all effects).

Voxel-wise contributions to the approximate log-evidence were computed for each model,  $U_n(m)$ . These were then compared to assess putative experimental effects. For (i) the average effect of presenting faces we compared models 1 and 2, for (ii) the main effect of repetition we compared models 2 and 3, for (iii) the main effect of familiarity we compared models 2 and 4 and for (iv) the interaction between repetition and familiarity we compared models 5 and 6.

Figure 10 shows a map of the average effect of presenting faces for selected slices. Figure 11 shows a map of the main effect of repetition. In each figure the top row shows a map of the relevant difference in contribution to the approximate log-evidence,  $U_n(m_1) - U_n(m_2)$ , and the bottom row shows the same but thresholded so that the corresponding posterior probability, computed using equation 16, is greater than 0.999. For this subject there is a large bilateral occipital and fusiform response to the main effect of faces (Figure 10) but no repetition effect (Figure 11). The familiarity and interaction effects were also absent in this subject and so have not been presented.

### 4.2.3 Comparing hemodynamic basis sets

This section compares models with different hemodynamic basis sets. We use the seven options described in section 2.2.6. The basis functions were convolved with stimulus functions corresponding to all four experimental conditions.

The approach is illustrated on COIs from three regions centred on left occipital cortex (LOC)  $x = -45, y = -60, z = -24\text{mm}$ , right occipital cortex (ROC)  $x = 45, y = -66, z = -24\text{mm}$  and sensorimotor cortex  $x = 36, y = -9, z = 66\text{mm}$ . All COIs were 9mm spheres and contained 83, 41 and 33 voxels respectively. We also optimised the Hanning-windowed Fourier and FIR basis sets by selecting the number of time-bins and bin-size that gave the highest model evidence.

Despite this the model evidence favours strongly an Informed basis set for all of the regions, as shown in Figure 12. The set Inf-2 is preferred for LOC and Inf-3 for the others. The differences in log evidence provide posterior

probabilities of unity in favour of the most probable model. We also repeated the analyses with smaller COIs, but the results were much the same, with the informed basis sets always being preferred with high posterior probability ( $\geq 0.95$ ).

## 5 Discussion

We have presented a unified framework for the analysis of fMRI data based on Bayesian comparison of spatially regularised General Linear Models. This allows for Bayesian ANOVAs, Cluster of Interest (COI) analysis and the principled selection of signal and noise models.

COI analysis is similar to an ROI approach but all time series in a region are used rather than a single ‘representative’ time series. Our simulations have shown that, for non-uniformly activated regions the COI approach is substantially more sensitive. The COI approach may also be viewed as a Bayesian cluster-level inference as it shares the fundamental property of classical cluster-level inference that anatomical specificity is traded-off for increase sensitivity. As is the case for classical inference, Bayesian cluster-level inferences are more sensitive to weak, diffuse activations than are voxel-level inferences. For diffusely activated regions this sensitivity increases with number of voxels in the region. Moreover, the use of spatial regularisation increases that sensitivity yet further. Unlike classical cluster-level inferences [15], a primary ‘height’ threshold is not required.

Bayesian ANOVAs can be implemented in two ways (i) using COI analysis or (ii) to produce pseudo Posterior Probability Maps (PPMs) for the whole image volume. The term ‘pseudo’ is used as these probabilities are based on contributions to the model evidence rather than the model evidence per se. These PPMs show voxels expressing overall effects, main effects or interactions.

In previous work we have shown how PPMs can be used to make inferences about regionally specific effects [14]. The approach is based on computing the probability that a contrast of parameter estimates is larger than a user-specified effect size. For example, in primary sensory areas only effect sizes greater than 1% of global brain activity may be deemed biologically relevant. PPMs based on Bayesian ANOVAs, however, do not require the specification of an effect size threshold. This can be viewed as an advantage or a disadvantage depending on your perspective. On the one hand, imagers

are unable to describe the effects that interest them so precisely, but on the other, they have one less parameter to specify.

A disadvantage of the Bayesian ANOVA approach we have described is that a family of models must be fitted which is obviously more computationally demanding than fitting a single model. This is not a problem for a COI analysis where models can be fitted in seconds, but it is potentially a problem for the production of PPMs. Fitting the six models required to produce a full ANOVA took 3 hours of computer time.

An alternative to PPMs here is to make inferences based on F-maps, as is standard in neuroimaging [25]. It is also possible to make ANOVA-like Bayesian inferences about effect sizes based on multivariate Gaussian posteriors. The generic approach is described in [6].

We have also shown how our framework can be used for the principled selection of signal and noise models. We have illustrated its use on a single-subject event-related fMRI study of face processing. The following observations are therefore specific to our analysis of this data. Whether or not they apply generically to fMRI remains to be seen.

The framework was applied to determine the optimal regularisation method for an AR model of fMRI noise processes. Two types of regularisation were compared (i) a spatial prior which assumes that AR coefficients vary smoothly across the brain and (ii) a ‘tissue-type’ prior which assumes they vary about a small number of tissue-specific values. Bayesian model comparison showed spatial priors to be better. Our results therefore show that tissue type is not sufficient to explain the observed spatial variability in temporal autocorrelation. The largest single source of this variability appears to be the strong autocorrelation observed close to the cerebral arteries, as shown for example in the plots on the left-most side of Figure 9. Unless one has angiographic data, these regions are not easily delineated. The spatial prior approach can, however, automatically accommodate these variations.

The framework was also applied to determine the optimal basis set for describing the hemodynamic response. We have shown that the previously established method of nested-model comparison [21] is sub-optimal. Application of the optimal non-nested framework revealed the ‘informed basis set’ to be the optimal choice in a number of COIs.

We now turn to a discussion of future work. Application of tissue-type-priors to regression coefficients is one simple extension. This would use zero mean Gaussians with prior variances that depend on tissue type. Low prior variances in CSF and white matter could be implemented using equivalents of

the Gamma priors in Appendix F.3. Whether these models would be better than the spatial GMRF priors in this paper (based entirely on functional data) is an issue that can be resolved using the model comparison framework. In a similar vein, we are currently working on spatial-basis set priors, that include eg. wavelets, as an alternative to GMRFs.

## Acknowledgements

W.D. Penny is funded by the Wellcome Trust.

## References

- [1] SPM, Wellcome Department of Imaging Neuroscience. Available from <http://www.fil.ion.ucl.ac.uk/spm/software/>, 2002.
- [2] J. Ashburner and K.J. Friston. Image segmentation. In R.S.J. Frackowiak, K.J. Friston, C. Frith, R. Dolan, K.J. Friston, C.J. Price, S. Zeki, J. Ashburner, and W.D. Penny, editors, *Human Brain Function*. Academic Press, 2nd edition, 2003.
- [3] J. Ashburner and K.J. Friston. Spatial normalization using basis functions. In R.S.J. Frackowiak, K.J. Friston, C. Frith, R. Dolan, K.J. Friston, C.J. Price, S. Zeki, J. Ashburner, and W.D. Penny, editors, *Human Brain Function*. Academic Press, 2nd edition, 2003.
- [4] T. Auranen, A. Nummenmaa, M. Hammalainen, I. Jaaskelainen, J. Lampinen, A. Vehtari, and M. Sams. Bayesian analysis of the neuromagnetic inverse problem with  $l^p$  norm priors. *Neuroimage*, 2005. In Press.
- [5] M. Beal and Z. Ghahramani. The Variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. In J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F. Smith, and M. West, editors, *Bayesian Statistics*, volume 7. Oxford University Press, 2003.
- [6] G.E.P Box and G.C. Tiao. *Bayesian Inference in Statistical Analysis*. John Wiley, 1992.

- [7] M. Brett, W.D. Penny, and S.J. Kiebel. Introduction to random field theory. In R.S.J. Frackowiak, K.J. Friston, C. Frith, R. Dolan, K.J. Friston, C.J. Price, S. Zeki, J. Ashburner, and W.D. Penny, editors, *Human Brain Function*. Academic Press, 2nd edition, 2003.
- [8] C. Buchel and K.J. Friston. Modulation of connectivity in visual pathways by attention: Cortical interactions evaluated with structural equation modelling and fMRI. *Cerebral Cortex*, 7:768–778, 1997.
- [9] M.J. Buckley. Fast computation of a discretized thin-plate smoothing spline for image data. *Biometrika*, 81:247–258, 1994.
- [10] E. Bullmore, M. Brammer, S. Williams, S. Rabe-Hesketh, N. Janot, A. David, J. Mellers, R. Howard, and P. Sham. Statistical methods of estimation and inference for functional MR image analysis. *Magnetic Resonance in Medicine*, 35(2):261–277, 1996.
- [11] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley, 1991.
- [12] R.S.J. Frackowiak, K.J. Friston, C. Frith, R. Dolan, C.J. Price, S. Zeki, J. Ashburner, and W.D. Penny. *Human Brain Function*. Academic Press, 2nd edition, 2003.
- [13] K.J. Friston, P. Fletcher, O. Josephs, A.P. Holmes, M.D. Rugg, and R. Turner. Event-related fMRI: characterizing differential responses. *NeuroImage*, 7:30–40, 1998.
- [14] K.J. Friston, D.E. Glaser, R.N.A. Henson, S.J. Kiebel, C. Phillips, and J. Ashburner. Classical and Bayesian inference in neuroimaging: Applications. *NeuroImage*, 16:484–512, 2002.
- [15] K.J. Friston, A.P. Holmes, J.B. Poline, C.J. Price, and C. Frith. Detecting activations in PET and fMRI: Levels of inference and power. *NeuroImage*, 40:223–235, 1995.
- [16] K.J. Friston, A.P. Holmes, K.J. Worsley, J.B. Poline, C. Frith, and R.S.J. Frackowiak. Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2:189–210, 1995.

- [17] T. Gautama and M.M. Van Hulle. Optimal spatial regularisation of autocorrelation estimates in fMRI analysis. *NeuroImage*, (23):1203–1216, 2004.
- [18] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Chapman and Hall, Boca Raton, 1995.
- [19] R.N.A. Henson. Analysis of fMRI time series. In R.S.J. Frackowiak, K.J. Friston, C. Frith, R. Dolan, K.J. Friston, C.J. Price, S. Zeki, J. Ashburner, and W.D. Penny, editors, *Human Brain Function*. Academic Press, 2nd edition, 2003.
- [20] R.N.A. Henson and W.D. Penny. ANOVAs and SPM. Technical report, Wellcome Department of Imaging Neuroscience, 2003.
- [21] R.N.A. Henson, M.D. Rugg, and K.J. Friston. The choice of basis functions in event-related fMRI. *NeuroImage*, 13(6):127, June 2001. Supplement 1.
- [22] R.N.A. Henson, T. Shallice, M.L. Gorno-Tempini, and R.J. Dolan. Face repetition effects in implicit and explicit memory tests as measured by fMRI. *Cerebral Cortex*, 12:178–186, 2002.
- [23] N. Kanwisher, D. Stanley, and A. Harris. The Fusiform Face Area is Selective for Faces not Animals. *Neuroreport*, 10(1):183–187, 1999.
- [24] R.E. Kass and A.E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.
- [25] S.J. Kiebel. The general linear model. In R.S.J. Frackowiak, K.J. Friston, C. Frith, R. Dolan, K.J. Friston, C.J. Price, S. Zeki, J. Ashburner, and W.D. Penny, editors, *Human Brain Function*. Academic Press, 2nd edition, 2003.
- [26] J. Marchini and B. Ripley. A new statistical approach to detecting significant activation in functional MRI. *Neuroimage*, 12:168–193, 2000.
- [27] R. Pascual Marqui, C. Michel, and D. Lehman. Low resolution electromagnetic tomography: a new method for localizing electrical activity of the brain. *International Journal of Psychophysiology*, pages 49–65, 1994.

- [28] M.Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.
- [29] M.Brett, J.K. Anton, R. Valabregue, and J.B. Poline. Region of interest analysis using an SPM toolbox. In *Eighth International Conference on Functional Mapping of the Human Brain, June 2-6, Sendai, Japan, 2002*. Available on CD-ROM in Neuroimage, Vol 16, No 2, abstract 497.
- [30] J. W. Miskin and D. J. C. MacKay. Ensemble learning for blind source separation. In S.J. Roberts and R. Everson, editors, *ICA: Principles and Practice*. Cambridge University Press., 2000.
- [31] R.M. Neal and G.E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M.I. Jordan, editor, *Learning in Graphical Models*. Kluwer Academic Press, 1998.
- [32] W.D. Penny, S.J. Kiebel, and K.J. Friston. Variational Bayesian Inference for fMRI time series. *NeuroImage*, 19(3):727–741, 2003.
- [33] W.D. Penny, K.E. Stephan, A. Mechelli, and K.J. Friston. Comparing Dynamic Causal Models. *NeuroImage*, 22(3):1157–1172, 2004.
- [34] W.D. Penny, N. Trujillo-Barreto, and K.J. Friston. Bayesian fMRI time series analysis with spatial priors. *NeuroImage*, 24(2):350–362, 2005.
- [35] W. H. Press, S.A. Teukolsky, W.T. Vetterling, and B.V.P. Flannery. *Numerical Recipes in C*. Cambridge, 1992.
- [36] M. Sahani and S. S. Nagarajan. Reconstructing MEG sources with unknown correlations. In L. Saul S. Thrun and B. Schoelkopf, editors, *Advances in Neural Information Processing Systems*, volume 16. MIT, Cambridge, MA, 2004.
- [37] M. Sato, T. Yoshioka, S. Kajihara, K. Toyama, N. Goda, K. Doya, and M. Kawato. Hierarchical Bayesian estimation for MEG inverse problem. *NeuroImage*, 23:806–826, 2004.
- [38] N. Trujillo-Bareto, E. Aubert-Vazquez, and P.A. Valdes-Sosa. Bayesian model averaging in EEG/MEG imaging. *NeuroImage*, 21(4):1300–1319, 2004.

- [39] B.J. Winer, D.R. Brown, and K.M. Michels. *Statistical principles in experimental design*. McGraw-Hill, 1991.
- [40] M. W. Woolrich, B. D. Ripley, M. Brady, and S. M. Smith. Temporal autocorrelation in univariate linear modelling of fMRI data. *NeuroImage*, 14(6):1370–1386, December 2001.
- [41] M.W. Woolrich, T.E. Behrens, and S.M. Smith. Constrained linear basis sets for HRF modelling using Variational Bayes. *NeuroImage*, 21:1748–1761, 2004.
- [42] K. J. Worsley, C. H. Liao, J. Aston, V. Petre, G. H. Duncan, F. Morales, and A. C. Evans. A general statistical analysis for fMRI data. *NeuroImage*, 15(1), January 2002.
- [43] W.Penny and G. Flandin. Bayesian analysis of single-subject fMRI: SPM implementation. Technical report, Wellcome Department of Imaging Neuroscience, 2005.
- [44] E. Zarahn, G.K. Aguirre, and M. D’Esposito. Empirical analysis of BOLD fMRI statistics 1. Spatially unsmoothed data collected under null-hypothesis conditions. *Neuroimage*, 5:179–197, 1997.

## A Precisions

### A.1 Priors

We use Gamma priors on the precisions  $\alpha$ ,  $\beta$  and  $\lambda$

$$p(\alpha) = \prod_{k=1}^K p(\alpha_k) \tag{17}$$

$$p(\beta) = \prod_{p=1}^P p(\beta_p)$$

$$p(\lambda) = \prod_{n=1}^N p(\lambda_n)$$

$$p(\alpha_k) = \text{Ga}(\alpha_k; q_1, q_2) \tag{18}$$

$$\begin{aligned}
p(\beta_p) &= \text{Ga}(\beta_p; r_1, r_2) \\
p(\lambda_n) &= \text{Ga}(\lambda_n; u_1, u_2)
\end{aligned}$$

where  $\text{Ga}()$  is defined in Appendix E2. Gamma priors were chosen as they are conjugate priors for Gaussian error models. The parameters are set to  $q_1 = r_1 = u_1 = 10$  and  $q_2 = r_2 = u_2 = 0.1$ . These parameters produce Gamma densities with a mean of 1 and a variance of 10. The robustness of model selection to the choice of these parameters is discussed in section 2.2.4 and 4.1.1.

## A.2 Posteriors

The approximate posteriors are also Gamma densities. For the precisions of the regression coefficients we have

$$\begin{aligned}
q(\alpha) &= \prod_{k=1}^K q(\alpha_k) & (19) \\
q(\alpha_k) &= \text{Ga}(\alpha_k; g_k, h_k) \\
\frac{1}{g_k} &= \frac{1}{2} \left( \text{Tr}(\hat{\Sigma}_k D_w) + \hat{w}_k^T D_w \hat{w}_k \right) + \frac{1}{q_1} \\
h_k &= \frac{N}{2} + q_2
\end{aligned}$$

For the precisions of the AR coefficients we have

$$\begin{aligned}
q(\beta) &= \prod_{p=1}^P q(\beta_p) & (20) \\
q(\beta_p) &= \text{Ga}(\beta_p; r_{1p}, r_{2p}) \\
\frac{1}{r_{1p}} &= \frac{1}{2} \left( \text{Tr}(V_p D_a) + m_p^T D_a m_p \right) + \frac{1}{r_1} \\
r_{2p} &= \frac{N}{2} + r_2
\end{aligned}$$

For the precisions on the observation noise we have

$$q(\lambda) = \prod_{n=1}^N q(\lambda_n) \quad (21)$$

$$\begin{aligned}
q(\lambda_n) &= \text{Ga}(\lambda_n; b_n, c_n) \\
\frac{1}{b_n} &= \frac{\tilde{G}_n}{2} + \frac{1}{u_1} \\
c_n &= \frac{T}{2} + u_2
\end{aligned}$$

where  $\tilde{G}_n$  is the expected prediction error defined in Appendix B of [32].

## B Regression coefficients

### B.1 Priors

For the regressions coefficients we have

$$\begin{aligned}
p(W) &= \prod_{k=1}^K p(w_k^T) \\
p(w_k^T) &= N(w_k^T; 0, \alpha_k^{-1} D_w^{-1})
\end{aligned} \tag{22}$$

where  $D_w$  is a spatial precision matrix. This can be set to correspond to eg. a Low Resolution Tomography (LORETA) or Gaussian Markov Random Field (GMRF) prior, as described in earlier work [34]. These priors are specified separately for each slice of data. Specification of 3-dimensional spatial priors (ie. over multiple slices) is desirable from a modelling perspective but is computationally too demanding for current computer technology.

We also write  $w_v = \text{vec}(W)$ ,  $w_r = \text{vec}(W^T)$ ,  $w_v = H_w w_r$  where  $H_w$  is a permutation matrix. This leads to

$$\begin{aligned}
p(W) &= p(w_v) \\
&= N(w_v; 0, B^{-1})
\end{aligned} \tag{23}$$

where  $B$  is an augmented spatial precision matrix given by

$$B = H_w (\text{diag}(\alpha) \otimes D_w) H_w^T \tag{24}$$

This form of the prior will be used in the derivation of KL-divergences in appendix E4.

The above Gaussian priors underly GMRFs and LORETA and have been used previously in fMRI [41] and EEG [27]. They are by no means, however,

the optimal choice for imaging data. In EEG, for example, much interest has focussed on the use of  $L^p$ -norm priors [4] instead of the  $L^2$ -norm implicit in the Gaussian assumption. Additionally, we are currently investigating the use of wavelet priors. This is an active area of research and will be the topic of future publications.

## B.2 Posteriors

We have

$$\begin{aligned}
 q(W) &= \prod_{n=1}^N q(w_n) & (25) \\
 q(w_n) &= \mathbf{N}(w_n; \hat{w}_n, \hat{\Sigma}_n) \\
 \hat{\Sigma}_n &= (\bar{\lambda}_n \tilde{A}_n + \bar{B}_{nn})^{-1} \\
 \hat{w}_n &= \hat{\Sigma}_n (\bar{\lambda}_n \tilde{b}_n^T + r_n) \\
 r_n &= - \sum_{i=1, i \neq n}^N \bar{B}_{ni} \hat{w}_i
 \end{aligned}$$

where  $\bar{B}$  is defined as in equation 24 but uses  $\bar{\alpha}$  instead of  $\alpha$ . The quantities  $\tilde{A}_n$  and  $\tilde{b}_n$  are expectations related to autoregressive processes and are defined in Appendix B of [32]. In the absence of temporal autocorrelation we have  $\tilde{A}_n = X^T X$  and  $\tilde{b}_n^T = X^T y_n$ . The above density can be written as a distribution over  $w_v$

$$\begin{aligned}
 q(W) &= q(w_v) & (26) \\
 &= \mathbf{N}(w_v; \hat{w}_v, \hat{\Sigma}_v)
 \end{aligned}$$

where  $\hat{w}_v^T = [\hat{w}_1^T, \hat{w}_n^T, \dots, \hat{w}_N^T]$  and  $\hat{\Sigma}_v = \text{blkdiag}(\hat{\Sigma}_1, \hat{\Sigma}_n, \dots, \hat{\Sigma}_N)$ . This form of the posterior will be used in the derivation of KL-divergences in appendix E4.

## C AR coefficients

### C.1 Priors

Similarly, for the autoregressive coefficients we have

$$\begin{aligned} p(A) &= \prod_{p=1}^P p(a_p) \\ p(a_p) &= N(a_p; 0, \beta_p^{-1} D_a^{-1}) \end{aligned} \quad (27)$$

Again,  $D_a$  is a user-defined spatial precision matrix,  $a_v = \text{vec}(A)$ ,  $a_r = \text{vec}(A^T)$  and  $a_v = H_a a_r$  where  $H_a$  is a permutation matrix. We can write

$$\begin{aligned} p(A) &= p(a_v) \\ &= N(a_v; 0, J^{-1}) \end{aligned} \quad (28)$$

where  $J$  is an augmented spatial precision matrix

$$J = H_a (\text{diag}(\beta) \otimes D_a) H_a^T \quad (29)$$

This form of the prior will be used in the derivation of KL-divergences in appendix E5.

### C.2 Posteriors

We have

$$\begin{aligned} q(A) &= \prod_{n=1}^N q(a_n) \\ q(a_n) &= \mathbf{N}(a_n; m_n, V_n) \end{aligned} \quad (30)$$

where

$$\begin{aligned} V_n &= (\bar{\lambda}_n \tilde{C}_n + \bar{J}_{nn})^{-1} \\ m_n &= V_n (\bar{\lambda}_n \tilde{d}_n + j_n) \\ j_n &= - \sum_{i=1, i \neq n}^N \bar{J}_{ni} m_i \end{aligned} \quad (31)$$

and  $\bar{J}$  is defined as in equation 29 but  $\bar{\beta}$  is used instead of  $\beta$ . The subscripts in  $\bar{J}_{ni}$  denote that part of  $\bar{J}$  relevant to the  $n$ th and  $i$ th voxels. The quantities

$\tilde{C}_n$  and  $\tilde{d}_n$  are expectations that are defined in equation 50 of [32]. The distribution over  $A$  can be re-written as

$$\begin{aligned} q(A) &= q(a_v) \\ &= \mathbf{N}(a_v; m_v, V_v) \end{aligned} \quad (32)$$

where  $m_v^T = [m_1^T, m_n^T, \dots, m_N^T]$  and  $V_v = \mathbf{blkdiag}(V_1, V_n, \dots, V_N)$ . This form of the posterior will be used in the derivation of KL-divergences in appendix E5.

## D Average likelihood

The average log-likelihood for model  $m$  is given by

$$V(m) = \sum_{n=1}^N V_n(m) \quad (33)$$

where

$$V_n(m) = \frac{T-P}{2} (\psi(c_n) + \log b_n) - \frac{\bar{\lambda}_n}{2} \tilde{G}_n - \frac{T-P}{2} \log 2\pi \quad (34)$$

where  $\psi()$  is the digamma function [35] and the quantity  $\tilde{G}_n$  is defined in equation 77 of [32]. This expression is identical to that given in equation 92 of [32].

## E KL divergences

### E.1 Normal densities

The multivariate Normal density is given by

$$\mathbf{N}(x; \mu, \Sigma) = (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right) \quad (35)$$

The KL divergence for Normal densities  $q(x) = \mathbf{N}(x; \mu_q, \Sigma_q)$  and  $p(x) = \mathbf{N}(x; \mu_p, \Sigma_p)$  is

$$\begin{aligned} KL_N[q(x), p(x)] &= 0.5 \log \frac{|\Sigma_p|}{|\Sigma_q|} + 0.5 Tr(\Sigma_p^{-1} \Sigma_q) \\ &+ 0.5(\mu_q - \mu_p)^T \Sigma_p^{-1} (\mu_q - \mu_p) - \frac{d}{2} \end{aligned} \quad (36)$$

where  $|\Sigma_p|$  denotes the determinant of the matrix  $\Sigma_p$ .

## E.2 Gamma densities

The Gamma density is defined as

$$Ga(x; b, c) = \frac{1}{\Gamma(c)} \frac{x^{c-1}}{b^c} \exp\left(\frac{-x}{b}\right) \quad (37)$$

For Gamma densities  $q(x) = Ga(x; b_q, c_q)$  and  $p(x) = Ga(x; b_p, c_p)$  the KL-divergence is

$$\begin{aligned} KL_{Ga}[q(x), p(x)] &= (c_q - 1)\psi(c_q) - \log b_q - c_q - \log \Gamma(c_q) \\ &+ \log \Gamma(c_p) + c_p \log b_p - (c_p - 1)(\psi(c_q) + \log b_q) + \frac{b_q c_q}{b_p} \end{aligned} \quad (38)$$

where  $\Gamma()$  is the Gamma function [35].

## E.3 Precisions

For  $\alpha$ ,  $\beta$  and  $\lambda$  these divergences are straightforward to calculate as they reduce to a sum of KL-divergences between gamma densities which have a known form and are computationally simple to evaluate.

$$KL[q(\alpha), p(\alpha)] = \sum_{k=1}^K KL_{Ga}[q(\alpha_k), p(\alpha_k)] \quad (39)$$

$$KL[q(\beta), p(\beta)] = \sum_{p=1}^P KL_{Ga}[q(\beta_p), p(\beta_p)] \quad (40)$$

$$KL[q(\lambda), p(\lambda)] = \sum_{n=1}^N KL_{Ga}[q(\lambda_n), p(\lambda_n)] \quad (41)$$

## E.4 Regression coefficients

For the regression coefficients we have

$$\begin{aligned} KL[q(W), p(W)] &= KL_N[q(w_v), p(w_v)] \\ &= -\frac{1}{2} \log |\hat{\Sigma}_v| - \frac{1}{2} \log |\bar{B}| + \frac{1}{2} \text{Tr}(\bar{B} \hat{\Sigma}_v) + \frac{1}{2} \hat{w}_v^T \bar{B} \hat{w}_v - \frac{KN}{2} \end{aligned} \quad (42)$$

The only problematic term here is  $\log |\bar{B}|$ . But we note that (i)  $H_w$  is a permutation matrix so it won't affect determinants and (ii) if  $X$  is an  $m \times m$  matrix and  $Y$  is an  $n \times n$  matrix then  $\det(X \otimes Y) = \det(X)^n \det(Y)^m$ . We can therefore write

$$\log |\bar{B}| = N \sum_{k=1}^K \log \alpha_k + K \log |D_w| \quad (43)$$

We can also write

$$\begin{aligned} \log |\hat{\Sigma}_v| &= \sum_{n=1}^N \log |\hat{\Sigma}_n| \\ \text{Tr}(\bar{B} \hat{\Sigma}_v) &= \sum_{n=1}^N \text{Tr}(\bar{B}_{nn} \hat{\Sigma}_n) \\ \hat{w}_v^T \bar{B} \hat{w}_v &= \sum_{n=1}^N \left( \hat{w}_n^T \bar{B}_{nn} \hat{w}_n + \hat{w}_n^T \sum_{i=1, i \neq n}^N \bar{B}_{ni} \hat{w}_i \right) \end{aligned} \quad (44)$$

We can therefore write the divergence as a sum of ‘unique contributions’ from voxel  $n$

$$\begin{aligned} KL[q(W), p(W)] &= \sum_{n=1}^N KW(n) \\ KW(n) &= -\frac{1}{2} \log |\hat{\Sigma}_n| - \frac{1}{2} \sum_k \log \bar{\alpha}_k - \frac{K}{2N} \log |D_w| + \frac{1}{2} \text{Tr}(\bar{B}_{nn} \hat{\Sigma}_n) \\ &\quad + \frac{1}{2} \left( \hat{w}_n^T \bar{B}_{nn} \hat{w}_n + \hat{w}_n^T \sum_{i=1, i \neq n}^N \bar{B}_{ni} \hat{w}_i \right) - \frac{K}{2} \end{aligned} \quad (45)$$

The subscripts in  $\bar{B}_{ni}$  denote that part of  $\bar{B}$  relevant to voxels  $n$  and  $i$ .

## E.5 Autoregressive coefficients

For the autoregressive coefficients we can use the same approach

$$\begin{aligned} KL[q(A), p(A)] &= KL_N[q(a_v), p(a_v)] \\ &= -\frac{1}{2} \log |V_v| - \frac{1}{2} \log |\bar{J}| + \frac{1}{2} \text{Tr}(\bar{J}V_v) + \frac{1}{2} m_v^T \bar{J} m_v - \frac{PN}{2} \end{aligned} \quad (47)$$

Writing as a sum over ‘unique contributions’ from each voxel gives

$$\begin{aligned} KL[q(A), p(A)] &= \sum_{n=1}^N KA(n) \\ KA(n) &= -\frac{1}{2} \log |V_n| - \frac{1}{2} \sum_p \log \bar{\beta}_p - \frac{P}{2N} \log |D_a| + \frac{1}{2} \text{Tr}(\bar{J}_{nn}V_n) \\ &\quad + \frac{1}{2} \left( m_n^T \bar{J}_{nn} m_n + m_n^T \sum_{i=1, i \neq n}^N \bar{J}_{ni} m_i \right) - \frac{P}{2} \end{aligned} \quad (48)$$

## F AR coefficients with tissue-type priors

### F.1 AR Priors

We introduce the label  $s$ . For example  $s = \{1, 2, 3\}$  could correspond to grey matter, white matter and CSF. We also introduce the indicator function  $\gamma_{ns}$  which is 1 if voxel  $n$  belongs to category  $s$  and zero otherwise.  $N_s = \sum_s \gamma_{ns}$  is the number of voxels in the  $s$ th category.  $S$  is the number of categories. A ‘tissue-type’ prior is then defined as

$$\begin{aligned} p(A) &= \prod_n p(a_n) \\ p(a_n) &= N(a_n; g_n, \beta_n^{-1}) \end{aligned} \quad (49)$$

where

$$\begin{aligned} g_n &= \sum_s \gamma_{ns} a_s \\ \beta_n &= \text{diag} \left( \sum_s \gamma_{ns} \beta_s \right) \end{aligned} \quad (50)$$

and  $a_s$  is the archetypal vector of AR coefficients for voxel type  $s$ , and  $\beta_s$  is the corresponding precision vector. This prior is like a Gaussian mixture model but one where the labelling is known.

In this paper, the parameters  $a_s$  and  $\beta_s$  are estimated from the data on a slice-by-slice basis. We use a Gamma prior on the precisions (see section F.3). For simplicity, there is no prior on  $a_s$ .

## F.2 AR Posteriors

We have

$$\begin{aligned} q(A) &= \prod_n q(a_n) \\ q(a_n) &= \mathbf{N}(a_n; m_n, V_n) \end{aligned} \quad (51)$$

where

$$\begin{aligned} V_n &= (\bar{\lambda}_n \tilde{C}_n + \bar{\beta}_n)^{-1} \\ m_n &= V_n (\bar{\lambda}_n \tilde{d}_n + \bar{\beta}_n \bar{g}_n) \end{aligned} \quad (52)$$

and

$$\begin{aligned} \bar{g}_n &= \sum_s \gamma_{ns} a_s \\ \bar{\beta}_n &= \text{diag} \left( \sum_s \gamma_{ns} \bar{\beta}_s \right) \end{aligned} \quad (53)$$

## F.3 AR precision priors

We define the precision for the  $s$ th structure type and  $p$ th AR coefficient,  $\beta_{sp}$ , as the  $p$ th element of  $\beta_s$ . We then have

$$\begin{aligned} p(\beta) &= \prod_{s=1, p=1}^{S, P} p(\beta_{sp}) \\ p(\beta_{sp}) &= \mathbf{Ga}(\beta_{sp}; r_1, r_2) \end{aligned} \quad (54)$$

## F.4 AR precision posteriors

The posterior is given by

$$\begin{aligned} q(\beta_{sp}) &= \mathbf{Ga}(\beta_{sp}; r_{1sp}, r_{2sp}) \\ \frac{1}{r_{1sp}} &= \frac{1}{2} \sum_n \gamma_{ns} \left( (m_{np} - a_{sp})^2 + V_n(p, p) \right) + \frac{1}{r_1} \end{aligned} \quad (55)$$

$$\begin{aligned}
r_{2sp} &= \frac{N_s}{2} + r_2 \\
\bar{\beta}_{sp} &= r_{1sp} r_{2sp}
\end{aligned}
\tag{56}$$

## F.5 AR means

The archetypal AR coefficient vectors are estimated using

$$a_s = \frac{\sum_{n=1}^N \gamma_{ns} a_n}{N_s}
\tag{57}$$

## F.6 KL Divergences

$$\begin{aligned}
KA(n) &= KL_N[q(a_n), p(a_n)] \\
KL[q(\beta), p(\beta)] &= \sum_{p=1}^P \sum_{s=1}^S KL_{Ga}[q(\beta_{sp}), p(\beta_{sp})]
\end{aligned}
\tag{58}$$

## G Implementation note

The algorithm we have described is implemented in SPM version 5 and can be downloaded from [1]. Computation of a number of quantities (eg.  $\tilde{C}_n$ ,  $\tilde{d}_n$  and  $\tilde{G}_n$  defined in appendices C and D) is now much more efficient than in previous versions [34]. These improvements are described in a separate document [43]. To analyse a single session of data (eg. the face fMRI data) takes about 30 minutes on a typical modern PC.

## Figure Captions

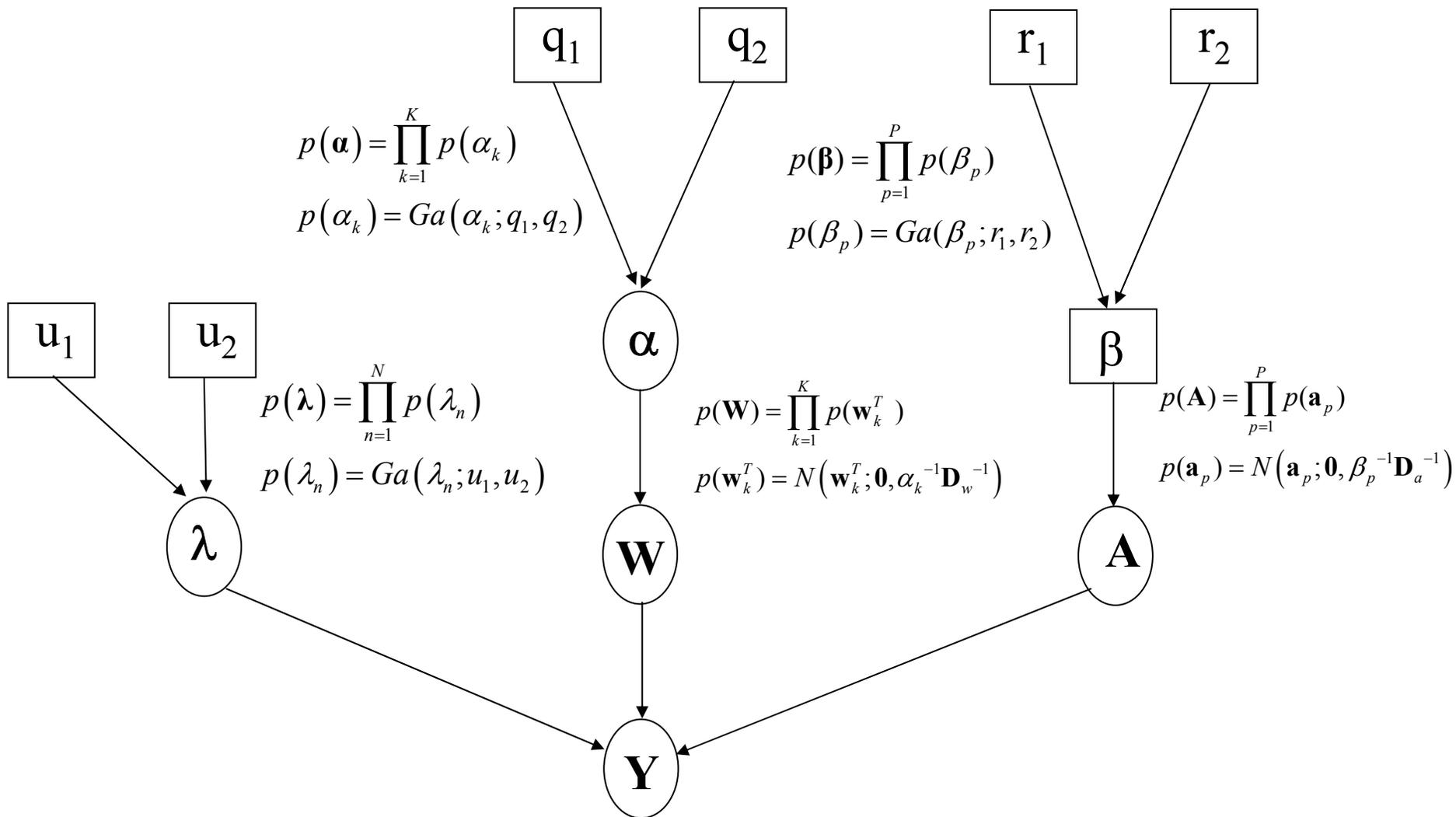
1. **Generative model** The figure shows the probabilistic dependencies underlying our generative model for fMRI data. The quantities in square brackets are constants and those in circles are random variables. The spatial regularisation coefficients  $\alpha$  constrain the regression coefficients  $W$ . The parameters  $\lambda$  and  $A$  define the autoregressive error processes which contribute to the measurements. The spatial regularisation coefficients  $\beta$  constrain the AR coefficients  $A$ . The graph shows that the joint probability of parameters and data can be written

$p(Y, W, A, \lambda, \alpha, \beta) = p(Y|W, A, \lambda)p(W|\alpha)p(A|\beta)p(\lambda|u_1, u_2)p(\alpha|q_1, q_2)p(\beta|r_1, r_2)$ , where the first term is the likelihood and the other terms are the priors. The likelihood is given in equation 11 and the priors are defined in greater detail in Appendices A, B and C.

2. **Approximate posteriors** The full approximate posterior distribution is  $q(W, A, \lambda, \alpha, \beta) = \left[ \prod_{n=1}^N q(w_n)q(a_n)q(\lambda_n) \right] q(\alpha)q(\beta)$ . The boxes in the figure show each component of the approximate posterior along with update equations for their sufficient statistics.
3. **Face paradigm** (a) Experimental stimuli, (b) Time series of stimuli presentation.
4. **AR(1) images for synthetic data.** Each row in this figure corresponds to analysis of a different data set. The leftmost column shows the AR(1) profile used to generate the data. The second, third and 4th columns show AR profiles as estimated by models with tissue-type priors having 1, 2 and 3 (known) discrete levels, respectively. The fifth column shows the estimated profiles from models with spatial GMRF priors. The final column shows bar plots of the posterior model probabilities. The first three bars correspond to models with tissue-type priors having 1, 2 and 3 levels and the fourth bar corresponds to the spatial GMRF model. These results show that our approximation to the model evidence can correctly detect the type of structure in the coefficients.
5. **Cluster of Interest (COI) Analysis for a Uniformly Activated Region** The figure shows of a plot of sensitivity versus number of voxels in the cluster,  $N$ , for models using a spatial prior (circles and solid line) and a shrinkage prior (crosses and dotted line).
6. **Cluster of Interest (COI) Analysis for a Non-Uniformly Activated Region** The figure shows of a plot of sensitivity versus number of voxels in the cluster,  $N$ , for a model using a spatial prior (circles and solid line), ROI analysis using the mean voxel time series (crosses and dotted line) and an ROI analysis using the principal component time series (plusses and dashed line).
7. **Noise-free time series** from type 2 data (thin line), generated from an FIR model, and type 1 data (thick line), generated from a best

fitting Informed basis set model. These data were used to compare the sensitivity of nested versus non-nested model comparison, in the context of selecting an optimal hemodynamic basis set.

8. **Nested versus Non-Nested** The figure shows of a plot of sensitivity versus number of voxels in the cluster,  $N$ , for non-nested model comparison (circles and solid line) versus nested model comparison (crosses and dotted line).
9. **AR(1) images for face data.** The top row shows estimated profiles from a tissue-type prior (smoothed CSF versus other, prior (iii)) and the bottom row shows the estimated profiles from models with spatial GMRF priors. Columns in this figure show results for slices  $z=-27$ ,  $3$ ,  $33$  and  $63\text{mm}$ .
10. **Average effect of faces** The top row shows maps of the difference in contributions to the log evidence,  $U_n(2) - U_n(1)$ , for slices  $z=-24$ ,  $-21$ ,  $-18$  and  $-15\text{mm}$ . The bottom row shows the same map but thresholded so that only effects with a posterior probability greater than  $0.999$  (difference in log evidence =  $4.6$ ) survive.
11. **Main effect of repetition** The top row shows maps of the difference in contributions to the log evidence,  $U_n(3) - U_n(2)$ , for slices  $z=-24$ ,  $-21$ ,  $-18$  and  $-15\text{mm}$ . The bottom row shows the same maps but thresholded so that only effects with a posterior probability greater than  $0.999$  (difference in log evidence =  $4.6$ ) survive.
12. **Comparing hemodynamic basis sets** The bar plots show the log-evidence for each hemodynamic basis set for COIs in (a) Left Occipital Cortex  $x = -45, y = -60, z = -24\text{mm}$ , (b) Right Occipital Cortex  $x = 45, y = -66, z = -24\text{mm}$ , (a) Sensorimotor Cortex  $x = 36, y = -9, z = 66\text{mm}$ . The evidence values have been normalised (by subtraction) so that the minimal log-evidence is zero.



## Regression coefficients, $\mathbf{W}$

$$\begin{aligned}
 q(\mathbf{w}_n) &= N(\mathbf{w}_n; \hat{\mathbf{w}}_n, \hat{\Sigma}_n) \\
 \hat{\mathbf{w}}_n &= \hat{\Sigma}_n (\bar{\lambda}_n \tilde{\mathbf{b}}_n^T + \mathbf{r}_n) \\
 \hat{\Sigma}_n &= (\bar{\lambda}_n \tilde{\mathbf{A}}_n + \bar{\mathbf{B}}_{mm})^{-1} \\
 \bar{\mathbf{B}} &= \mathbf{H}_w (\text{diag}(\bar{\boldsymbol{\alpha}}) \otimes \mathbf{D}_w) \mathbf{H}_w^T \\
 \mathbf{r}_n &= - \sum_{i=1, i \neq n}^N \bar{\mathbf{B}}_{ni} \hat{\mathbf{w}}_i
 \end{aligned}$$

## AR coefficients, $\mathbf{A}$

$$\begin{aligned}
 q(\mathbf{a}_n) &= N(\mathbf{a}_n; \mathbf{m}_n, \mathbf{V}_n) \\
 \mathbf{V}_n &= (\bar{\lambda}_n \tilde{\mathbf{C}}_n + \bar{\mathbf{J}}_{mm})^{-1} \\
 \mathbf{m}_n &= \mathbf{V}_n (\bar{\lambda}_n \tilde{\mathbf{d}}_n + \mathbf{j}_n) \\
 \bar{\mathbf{J}} &= \mathbf{H}_a (\text{diag}(\bar{\boldsymbol{\beta}}) \otimes \mathbf{D}_a) \mathbf{H}_a^T \\
 \mathbf{j}_n &= - \sum_{i=1, i \neq n}^N \bar{\mathbf{J}}_{ni} \mathbf{m}_i
 \end{aligned}$$

## Spatial precisions for $\mathbf{W}$

$$\begin{aligned}
 q(\boldsymbol{\alpha}) &= \prod_{k=1}^K q(\alpha_k) \\
 q(\alpha_k) &= Ga(\alpha_k; \mathbf{g}_k, h_k) \\
 \frac{1}{\mathbf{g}_k} &= \frac{1}{2} \left( \text{Tr}(\hat{\Sigma}_k \mathbf{D}_w) + \hat{\mathbf{w}}_k^T \mathbf{D}_w \hat{\mathbf{w}}_k \right) + \frac{1}{q_1} \\
 h_k &= \frac{N}{2} + q_2 \\
 \bar{\alpha}_k &= \mathbf{g}_k h_k
 \end{aligned}$$

## Spatial precisions for $\mathbf{A}$

$$\begin{aligned}
 q(\boldsymbol{\beta}) &= \prod_{p=1}^P q(\beta_p) \\
 q(\beta_p) &= Ga(\beta_p; r_{1p}, r_{2p}) \\
 \frac{1}{r_{1p}} &= \frac{1}{2} \left( \text{Tr}(\mathbf{V}_p \mathbf{D}_a) + \mathbf{m}_p^T \mathbf{D}_a \mathbf{m}_p \right) + \frac{1}{r_1} \\
 r_{2p} &= \frac{N}{2} + r_2 \\
 \bar{\beta}_p &= r_{1p} r_{2p}
 \end{aligned}$$

## Observation noise

$$\begin{aligned}
 q(\lambda_n) &= Ga(\lambda_n; b_n, c_n) \\
 \frac{1}{b_n} &= \frac{\tilde{G}_n}{2} + \frac{1}{u_1} \\
 c_n &= \frac{T}{2} + u_2
 \end{aligned}$$

