Constrained Linear Basis Sets for HRF Modelling Using Variational Bayes

FMRIB Technical Report TR04MW2
(A related paper has been accepted for publication in Neuroimage)

Mark W. Woolrich, Timothy E.J. Behrens and Stephen M. Smith

Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB),
Department of Clinical Neurology, University of Oxford, John Radcliffe Hospital,
Headley Way, Headington, Oxford, UK
Corresponding author is Mark Woolrich: woolrich@fmrib.ox.ac.uk

Abstract

FMRI modelling requires flexible HRF modelling, with the HRF being allowed to vary spatially and between subjects. To achieve this flexibility, voxel-wise parameterised HRFs have been proposed, however inference on such models is very slow. An alternative approach is to use basis functions allowing inference to proceed in the more manageable General Linear Model (GLM) framework. However, a large amount of the subspace spanned by the basis functions produces nonsensical HRF shapes. In this work we propose a technique for choosing a basis set, and then the means to constrain the subspace spanned by the basis set to only include sensible HRF shapes. Penny et al. (2003) showed how Variational Bayes can be used to infer on the GLM for FMRI. Here we extend the work of Penny et al. (2003) to give inference on the GLM with constrained HRF basis functions and with spatial Markov Random Fields on the autoregressive noise parameters. Constraining the subspace spanned by the basis set allows for far superior separation of activating voxels from non-activating voxels in FMRI data. We use spatial mixture modelling to produce final probabilities of activation and demonstrate increased sensitivity on an FMRI dataset.

1 Introduction

An important issue for detecting areas of brain activity in FMRI is the forward modelling of the temporal BOLD response. This forward model predicts what the BOLD response would be if we knew the underlying neural activity.

To allow modelling of BOLD responses to general stimulation types, Friston et al. (1994) introduced the use of convolution models which assume a linear time invariant system. Cohen (1997), Dale and Buckner (1997) and Boynton et al. (1996) provide some evidence that the BOLD response possesses linear characteristics with respect to the stimulation. However, non-linearities are predominant when there are short separations (less than approximately 3 seconds) between stimuli (Friston et al., 1998b). An additional assumption is that the stimulus represents the underlying neural activity. The stimulus (or neural activity) is then convolved with an assumed or modelled impulse response function, known as haemodynamic response function (HRF), to give the assumed BOLD response (Friston et al., 1994).

Genovese (2000) and Gössl et al. (2001) have previously used a Bayesian framework to model the voxel-wise BOLD response to a sustained period of stimulation. The advantage of a Bayesian approach is most obvious in the use of prior experience to justify the prior distributions used for these haemodynamic response parameters. However, the work of Genovese (2000) and Gössl et al. (2001) was restricted to just modelling the response to an epoch of fixed size. Woolrich et al. (2004b) generalised this to general stimulations via parametric modelling of the HRF assuming a linear time invariant system. However, the problem with parametric modelling of the HRF is that the model is difficult to infer upon without slow techniques such as Markov Chain Monte Carlo (MCMC).

Friston et al. (1995) and Josephs et al. (1997) consider voxel-wise linear time invariant system HRF models within the framework of the General Linear Model (GLM). Flexibility to model the HRF is introduced via basis sets. However, a large amount of the subspace spanned by the basis functions produces nonsensical HRF shapes (see figure 4(a)). This is because the conventional GLM will indiscriminately allow all possible linear combinations of the basis set.

In this work we propose a technique for using soft constraints to weight the subspace spanned by the basis set to only include sensible HRF shapes within a linear time-invariant system. The choice of the basis set can be driven by a standard parametric HRF, or a physiologically informed model such as the Balloon Model (Buxton et al., 1998). Using the GLM in a Bayesian framework we can then use priors on the basis function regression parameters to constrain the linear combinations of HRFs to sensible HRF shapes. Penny et al. (2003) showed how Variational Bayes can be used to infer on the GLM for FMRI. Here we extend this work to give inference on the GLM with the constrained HRF basis functions. We also extend the work of Penny et al. (2003) to spatially regularise the autoregressive noise parameters using a Markov Random Field (MRF). We demonstrate that the constrained basis function approach allows for far superior sensitivity, when compared with traditional unconstrained basis function approaches.

1.1 Overview

We start in section 2 by describing how the GLM in a fully Bayesian framework allows us to introduce soft constraints on linear combinations within the GLM. We then discuss how we choose a sensible basis set, and determine the required basis set constraints. In section 3 we describe how we perform approximate Variational Bayesian inference on the model. In section 4 we use the model on artificial null data to demonstrate the effect of the HRF constraints. Then in section 5 we describe how we can use spatial mixture modelling to produce probabilities of activation which takes advantage of the extra sensitivity produced from the HRF constraints. Finally, in section 6 we demonstrate this increased sensitivity on an audio-visual dataset.

2 Model

Here we describe the GLM in the fully Bayesian framework. It will be via the priors on the regression parameters that we propose to constrain the possible linear combinations that are allowed.

Consider that the preprocessed FMRI data at voxel i and at scan t is y_{it} $(i = 1 \dots N,\ t = 1 \dots T)$, the t^{th} $1 \times K$ row of the design matrix, x, is x_t, and β_i is a $K \times 1$ vector of parameter estimates. The preprocessed FMRI data, y, is taken to have been motion corrected and high-pass filtered. The standard general linear model (GLM) is then:

$$y_{it} = x_t \beta_i + \eta_{it} \tag{1}$$

We model the error, η_{it}, as a voxel-wise temporal autoregressive process of order P (AR(P)). We can represent this as:

$$\begin{aligned} \eta_{it} &= \sum_{p=1}^{P} a_{pi} \eta_{i(t-p)} + \epsilon_{it} \\ \epsilon_{it} &\sim N(0, \phi_{\epsilon_i}^{-1}) \end{aligned} \tag{2}$$

where a_{pi} is the p^{th} AR coefficient $(p = 1, \dots, P)$.

2.1 Bayesian framework

In this work we use a Bayesian framework. Equations 1 and 2 form our likelihood. The distribution we are interested in is the full posterior distribution over the model parameters, and depends upon this likelihood and the priors over the unknown parameters in our model via Bayes rule:

$$p(\beta, a, \phi_\epsilon | y) \propto \prod_i p(y_i | \beta_i, a_i, \phi_{\epsilon_i}) p(\beta, a, \phi_\epsilon) \tag{3}$$

We now need to consider the specification of priors over the parameters in our model. A priori we assume independence between the parameters:

$$p(\beta, a, \phi_\epsilon) = p(\beta) p(a) p(\phi_\epsilon) \tag{4}$$

As we shall see, using independent priors allows us to use conjugate priors, which in turns makes the model tractable when using Variational Bayes. However, assuming that we have independence between priors for different parameters does not mean that the parameters will be independent in the posterior. Any dependence between parameters inferred from the data and the likelihood will still be reflected in the joint posterior. For the precision we assume a voxelwise noninformative Gamma prior:

$$p(\phi_\epsilon) = \prod_i p(\phi_{\epsilon_i}) \quad (5)$$

$$\phi_{\epsilon_i} \sim Ga(b_{\epsilon_0}, c_{\epsilon_0}) \quad (6)$$

2.2 Autoregressive Parameters Spatial Prior

In Penny et al. (2003) non-spatial non-informative priors were used on the autoregressive parameters. Previous work has shown that neighbouring voxels have similar temporal autocorrelation (Woolrich et al., 2001; Worsley et al., 2002). Therefore, we want to model the assumption that a priori we expect neighbouring voxels to have similar temporal autocorrelation. To do this we use a Gaussian conditionally specified auto-regressive (CAR) or continuous Markov Random Field (MRF) prior (Cressie, 1993) on each of the P autoregressive parameter maps, i.e. $p(a) = \prod_{p=1}^{P} p(a_p)$ with:

$$p(a_p|\phi_{a_p}) \sim MVN(\mathbf{0}, \phi_{a_p}^{-1} D^{-1}) \quad (7)$$

where MVN denotes a multivariate Normal distribution, D is an $N \times N$ matrix whose (i,j)th element is d_{ij}, and ϕ_{a_p} is the MRF control parameter which controls the amount of spatial regularisation. We set $d_{ii} = 1$, $d_{ij} = -1/N_{ij}$ if i and j are spatial neighbours and $d_{ij} = 0$ otherwise (where N_{ij} is the geometric mean of the number of neighbours for voxels i and j). We also require a hyperprior on ϕ_{a_p}, for which we use a standard noninformative conjugate gamma prior:

$$\phi_{a_p} \sim Ga(b_{a_0}, c_{a_0}) \quad (8)$$

Before we specify the prior for the regression parameters β, we consider how we can use basis functions within this framework.

2.3 Basis Functions

In this work we model the different HRF shapes for different underlying conditions at different voxels using basis functions and assuming a linear time invariant system (Josephs et al., 1997). If we have $e = 1 \ldots N_e$ underlying conditions, and for each condition we have $b = 1 \ldots N_b$ basis functions to model the HRF for that condition for voxel i we can rewrite equation 1 as:

$$y_{it} = \sum_{e=1}^{N_e} \sum_{b=1}^{N_b} \{x_{ebt}\beta_{ieb}\} + \eta_{it} \quad (9)$$

where β_{ieb} is the regression parameter for the b^{th} basis function of the e^{th} underlying condition at voxel i and:

$$x_{eb\tau} = g_{b\tau} \otimes s_{e\tau} \quad (10)$$

where $\otimes$ represents convolution, $g_{b\tau}$ is the b^{th} basis function and $s_{e\tau}$ is the e^{th} stimulus function. Note that τ is an index at higher temporal resolution than t, to capture all of the HRF shape (i.e. for a resolution of $1/\rho$ of a TR, $\tau = 1, \ldots, \rho T$). We obtain x_{ebt} from $x_{eb\tau}$ by sampling every ρ^{th} $x_{eb\tau}$.

2.4 Constraining Basis Function Linear Combinations

Here we are going to describe how we constrain the basis function linear combinations. To do this we need to reparameterise the regression parameters, β_{ieb}, into parameters which describe the shape of the HRF, and parameters which scale these HRF shape parameters, to give the actual fit in the GLM.

Firstly, we specify β_{ie} as the $N_b \times 1$ vector of the regression parameters for the N_b basis functions for the e^{th} underlying condition at voxel i. Then, we reparameterise β_{ie} as being:

$$\beta_{ie} = \bar{D}_{ie} \frac{D_{ie}}{\sqrt{(\sum_b D_{ieb}^2 / N_b)}} \quad (11)$$

where D_{ie} is an $N_b \times 1$ vector of parameters describing the HRF for underlying condition e, and $\bar{D}_{ie}$ is the scalar value representing the scaling of that HRF. We want the scalar $\bar{D}_{ie}$ to contain *all* of our size information. However, left unchecked there is an arbitrary scale factor on vector D_{ie}. We have removed this arbitrary scale factor by normalising the vector using its root mean square. Hence, we now have a normalised vector, $D_{ie}/\sqrt{(}\sum_b D_{ieb}^2/N_b)$, representing the shape of the HRF, and a scalar, $\bar{D}_{ie}$, representing the size of the HRF.

For the scaling parameters we assume a noninformative prior:

$$\begin{aligned} p(\bar{D}) &= \prod_{ie} p(\bar{D}_{ie}) \\ \bar{D}_{ie} &\sim N(0, \phi_{\bar{D}_0}^{-1}) \end{aligned} \tag{12}$$

where the precision, $\phi_{\bar{D}_0}$, is fixed to be very small (1e-6) for all voxels. It is via the prior on D_{ie} that we can constrain the possible linear combinations of basis functions to represent the HRF for an underlying condition. We specify the prior on D_{ie} as:

$$\begin{aligned} p(D) &= \prod_{ie} p(D_{ie}) \\ D_{ie} &\sim MVN(m, C^{-1}) \end{aligned} \tag{13}$$

where m and C will contain the information constraining the possible linear combinations of the basis functions (see section 2.6 for how we set m and C).

2.5 Choosing a Basis Set

Basis sets used previously range from a single canonical HRF plus its temporal derivative to a set of Gamma functions Friston et al. (1998a)). These basis functions are then separately convolved with the known stimulus to give the same number of regressors as there are basis functions for use in the linear model.

Hossein-Zadeh and Ardekani (2002) and Friman et al. (2003) have previously shown how we can generate a basis set using singular value decomposition (SVD). This produces a basis set from samples of the HRF or regressors resulting from a parametric forward model of the haemodynamics. This is the approach we take in this paper.

In this paper we base the basis set on a parameterised model of the HRF. The HRF is parameterised by four half-cosines, requiring six parameters, as illustrated in figure 1.

To complete the HRF parameterised model we need to specify probabilities for the parameter values in the HRF model from which we generate physiologically plausible HRF shapes. With this information we can then draw HRF samples. Figure 2 shows 20 HRF samples drawn from the half-cosine parameterisation using the HRF parameter value probabilities:

$$\begin{aligned} h_1 &\sim \text{Uniform}(0s, 2s) \\ h_2 &\sim \text{Uniform}(2s, 6s) \\ h_3 &\sim \text{Uniform}(2s, 6s) \\ h_4 &\sim \text{Uniform}(2s, 8s) \\ f_1 &= 0 \\ f_2 &\sim \text{Uniform}(0, 0.5) \end{aligned} \tag{14}$$

We strongly emphasise that this is one of many possible choices of obtaining HRF samples. An alternative, attractive approach would be to use a physiologically based model, such as the Balloon model (Friston et al., 2000), with physiologically meaningful parameters which can be given sensible ranges.

Using a probabilistic model of choice, one can obtain a set of samples of the HRF, which represents the space of HRFs expected. The N_H HRF samples, each of length $N_T = \rho T$, can be placed into a $N_T \times N_H$ matrix, W. To obtain a basis set that spans this space of HRFs we perform a Singular Value Decomposition (SVD) on this matrix. This gives us N_H eigenHRFs (of length N_T) and N_H corresponding eigenvalues (describing the power each corresponding eigenHRF explains).

Figure 3 shows the four eigenHRFs (and their corresponding eigenvalues) with the largest eigenvalues obtained from an SVD on $N_H = 1000$ HRF samples of length $N_T = 512$ (resolution of 0.5 seconds) from the half-cosine parameterisation. It is worth noting that the first three eigenHRFs/basis functions look remarkably like the commonly

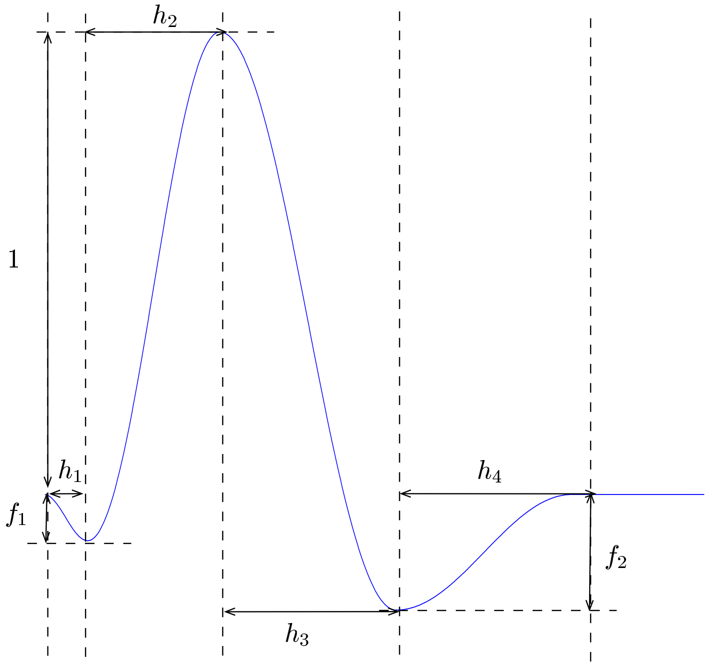


Figure 1: Parameterisation of the HRF into four half-period cosines. There are six parameters.

used canonical, delay (temporal) derivative and width (dispersion) derivative of a Gamma/Gaussian parameterised HRF (in that order).

To form our basis set we need to decide how many of the largest eigenHRFs we want to include. For the rest of this paper we use the $N_b = 3$ largest eigenHRFs as our basis set.

2.6 Determining Basis Set Constraints

In section 2.5 we determined a basis set to use. In this section we describe how we can apply constraints on the linear combinations of the basis set. In previous work Friman et al. (2003) also looked to constrain the possible linear combinations of the basis set, but within the canonical correlation analysis framework. However, they only looked to constrain the linear combination coefficients to be positive. In this work we look to apply a more complete constraint by fitting a multivariate Normal distribution to describe the desired constrained space probabilistically within the GLM framework.

To date, basis sets are used by convolving the constituent basis functions with the known stimulus to give the same number of regressors as there are basis functions stimuli. The resulting regressors are then used in the linear model. This approach corresponds to the model we have described earlier but with $m = 0$ and $C = I$ in equation 13. Figure 4(a) shows example HRF shapes obtained randomly drawn from this basis set when the linear combinations are unconstrained (i.e. with $m = 0$ and $C = I$). If we compare these HRFs with those in figure 2, it is clear that unconstrained linear combinations of our basis set allows for nonsensical HRF shapes.

To provide constraints on the possible linear combinations, we regress the HRF samples we used to obtain our basis set back onto the basis set:

$$W = GR + e \tag{15}$$

where W is the $N_T \times N_H$ matrix containing our N_H HRF samples, G is the $N_T \times N_b$ matrix of our N_b basis functions and $e \sim N(0, \sigma^2 I)$. We can perform standard Ordinary Least Squares (OLS) to obtain an estimate of the $N_b \times N_H$ matrix, R:

$$\hat{R} = (G^T G)^{-1} G^T W \tag{16}$$

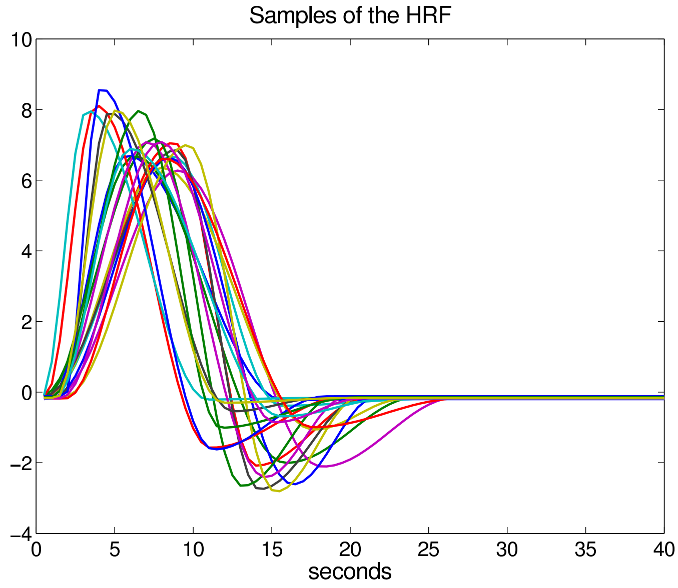


Figure 2: 20 HRF samples drawn from the half-cosine parameterisation

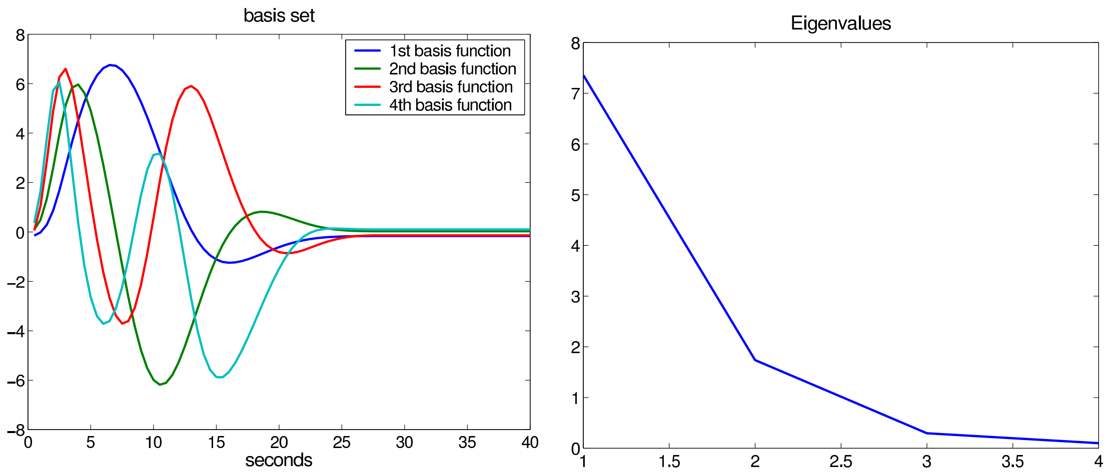
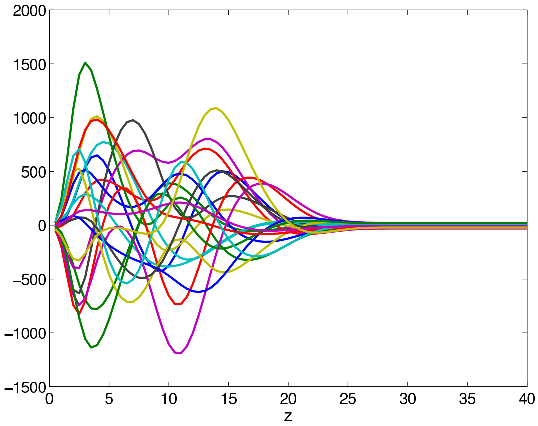
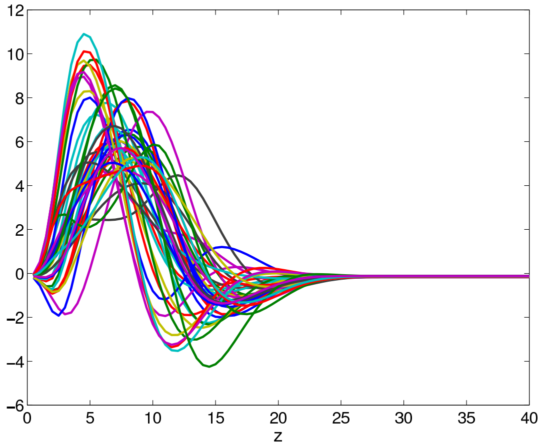


Figure 3: Four eigenHRFs (and their corresponding eigenvalues) with the largest eigenvalues from $N_H = 1000$ HRF samples of length $N_T = 512$ (resolution of 0.5 seconds) from the half-cosine parameterisation.



(a) Unconstrained



(b) Constrained

Figure 4: Samples from the basis set: (a) unconstrained with $m = 0$ and $C = I$, (b) constrained with $m = \tilde{m}$ and $C = \tilde{C}$.

We then fit a N_b-dimensional multivariate Normal distribution, $MVN(\tilde{m}, \tilde{C})$, to the matrix $\hat{R}$. The parameters of this multivariate Normal distribution are used to set the parameters in equation 13 (i.e. $m = \tilde{m}$ and $C = \tilde{C}$).

Using the HRF parameter value probabilities in equation 14 for our half-cosine HRF parameterisation, 1000 resulting HRF samples, and the resulting top three eigenHRFs, we obtain the multivariate Normal distribution parameters:

$$\begin{aligned} \tilde{m} &= [0.86, 0.09, 0.01]^T \\ \tilde{C} &= \begin{bmatrix} 0.018 & 0.028 & -0.015 \\ 0.028 & 0.185 & -0.009 \\ -0.015 & -0.009 & 0.030 \end{bmatrix} \end{aligned} \tag{17}$$

Figure 4(b) shows example HRF shapes obtained randomly drawn from this basis set when the linear combinations are constrained with these $m = \tilde{m}$ and $C = \tilde{C}$. We can see that compared with figure 4(a) this is a much more faithful representation of the sensible HRF shapes in figure 2.

It is important to point out that constraining the linear combinations by using a multivariate Normal distribution is an approximation. This is because a multivariate Normal will not capture all of the detail of the distribution in the N_b parameter space. As figure 4(b) shows, whilst we are producing much more sensible shapes than the unconstrained case, there are still some undesirable HRF shapes. The alternative is a fully parametric HRF approach (Woolrich et al., 2004b), which is very slow to infer upon, or a different distribution to a multivariate Normal which can capture the required detail in the N_b parameter space. However, we choose to use a multivariate Normal as it will make inference much more easy to handle within the Variational Bayesian framework, and it does capture the required structure.

3 Inference

The distribution we are interested in inferring upon is the posterior distribution $p(\theta|y)$ (equation 3), where θ is the set of parameters $\{\beta, a, \phi_\epsilon\}$. It is not possible to solve for this distribution analytically. Hence we use the framework introduced to FMRI by Penny et al. (2003) of Variational Bayes. For a general introduction to Variational Methods see Jordan (1999). Using this approach we can approximate a posterior distribution $p(\theta|y)$ with $q(\theta|y)$ by minimising the KL-divergence, or equivalently by maximising the variational free energy, F, between them:

$$F = \int q(\theta|y) \log \frac{p(y, \theta)}{q(\theta|y)} d\theta \tag{18}$$

To maximise this function, we need to ensure that the resulting integrals are tractable. A standard way to help achieve this is to use conjugate priors and to factorise the approximate posterior.

In the modelling section, we parameterised the model in terms of parameters a, ϕ_ϵ, D and $\bar{D}$, and wherever possible specified conjugate priors on them. However, using these parameters and factorising the posterior into

$$q(D, \bar{D}, a, \phi_a, \phi_\epsilon|y) = \prod_p \{q(a_p|y)p(\phi_{a_p}|y)\} \prod_i \{q(D_i|y)q(\bar{D}_i|y)q(\phi_{\epsilon_i}|y)\} \tag{19}$$

is not tractable to Variational Bayes as we can not derive the update equations for $q(D_{ie}|y)$ and $q(\bar{D}_{ie}|y)$. To overcome this problem, instead of using the two parameters D_{ie} and $\bar{D}_{ie}$, we reparameterise to use β_{ie} and $\bar{\beta}_{ie}$ by rewriting equation 11 as:

$$\beta_{ie} = \bar{\beta}_{ie} D_{ie} \tag{20}$$

where:

$$\bar{\beta}_{ie} = \frac{\bar{D}_{ie}}{\sqrt{(\sum_b D_{ieb}^2/N_b)}}. \tag{21}$$

Recall from equation 11 that $\sqrt{(\sum_b D_{ieb}^2/N_b)}$ is the normalisation on the HRF shape vector, D_{ie}, and the scalar, $\bar{D}_{ie}$, represents the size of the HRF. We now have the following prior on β_{ie}:

$$\beta_{ie}|\bar{\beta}_{ie} \sim MVN(\bar{\beta}_{ie} m, \bar{\beta}_{ie}^2 C^{-1}) \tag{22}$$

and a noninformative prior on $\bar{\beta}_{ie}$:

$$\bar{\beta}_{ie} \sim N(0, \phi_{\bar{\beta}_0}^{-1}) \tag{23}$$

where the precision, $\phi_{\bar{\beta}_0}$, is fixed to be very small (1e-6) for all voxels. We now assume the following factorised form for the approximate posterior:

$$q(\beta, \bar{\beta}, a, \phi_\epsilon|y) = \prod_p \{q(a_p|y)p(\phi_{a_p}|y)\} \prod_i \{q(\beta_i, \bar{\beta}_i|y)q(\phi_{\epsilon_i}|y)\} \tag{24}$$

where:

$$\begin{aligned} q(B_i|y) &= MVN(\mu_{B_i}, \Lambda_{B_i}) \\ q(a_p|y) &= N(\mu_{a_p}, \Lambda_{a_p}) \\ q(\phi_{a_p}|y) &= Ga(b_{a_p}, c_{a_p}) \\ q(\phi_{\epsilon_i}|y) &= Ga(b_{\epsilon_i}, c_{\epsilon_i}) \end{aligned} \tag{25}$$

where $B_i = (\beta_i, \bar{\beta}_i)$, where β_i is the $(N_e N_b) \times 1$ vector $[\beta_{i1} \dots \beta_{ie} \dots \beta_{iN_e}]^T$ and $\bar{\beta}_i$ is the $N_e \times 1$ vector $[\bar{\beta}_{i1} \dots \bar{\beta}_{ie} \dots \bar{\beta}_{iN_e}]^T$. Note that we do not fully factorise. We would not expect β_i and $\bar{\beta}_i$ to be independent a posteriori. Therefore, we maintain a combined unfactorised posterior for the two parameters β_i and $\bar{\beta}_i$. However, we have factorised the noise parameter posteriors from the regression parameter posteriors. This assumption helps to make inference tractable using Variational Bayes. Penny et al. (2003) discuss the implications of doing this and show that the error induced by this assumption is negligible for inferring on FMRI data. We also show later (in section 4) that there is negligible error induced when inferring on artificial data.

At this point, the inference is still not fully tractable to Variational Bayes as we can not derive the update equations for $q(B_i|y)$. To overcome this we rewrite the prior in equation 22 as:

$$\beta_{ie} \sim MVN(\bar{\beta}_{ie} m, \phi_{\bar{\beta}_{ie}}^{-1} C^{-1}) \tag{26}$$

where the utility parameter, $\phi_{\bar{\beta}_{ie}}$, is updated as a point estimate equal to one over the current expected value of $\bar{\beta}_{ie}^2$:

$$\begin{aligned} \phi_{\bar{\beta}_{ie}} &= E[\bar{\beta}_{ie}^{-2}]^{-1} \\ &= R_{ie} \mu_{\bar{\beta}_{ie}}^2 + (\Lambda_{B_i}^{-1})_{\bar{\beta}_{ie}} \end{aligned} \tag{27}$$

where R_{ie} is defined by the relationship $\bar{\beta}_{ie} = R_{ie}B_i$, and $(\Lambda_{B_i}^{-1})_{\bar{\beta}_{ie}}$ is the current marginal covariance of $\bar{\beta}_{ie}$. The approximate posterior distributions are now tractable to Variational Bayes. The update rules for the approximate posterior distributions, which iteratively maximises the free energy in equation 18, are given in appendix B.

We can perform standard inference questions on the marginal posterior over β_e, in the same way that we do for the standard use of basis functions in the GLM (i.e. using f-contrasts, see section 3.2). We test the accuracy of the posterior approximations presented in this section using null artificial data in section 4.

The Variational Bayes inference requires approximately 10 iterations and takes approximately 15 minutes (for a whole brain - in-plane resolution 4mm, slice thickness 7mm and 180 volumes) on a 2GHz Intel PC.

3.1 Initialisation

We do not need to initialise the approximate distribution parameters of $q(\beta_i, \bar{\beta}_i|y)$. This is because we can initialise the other approximate distribution parameters and then update $q(\beta_i, \bar{\beta}_i|y)$ first. To allow us to provide sensible initialisation of the other approximate distributions, $q(\phi_{\epsilon_i}|Y)$ and $q(a_p)$, we use the model with the autoregression parameters set to zero ($a_p = 0$) and with the HRF constraints removed ($m = 0$ and $C = I$). This means we can use the standard ordinary least squares (OLS) voxelwise frequentist solution to the GLM to get:

$$\begin{aligned} \hat{\beta}_i &= (x^T x)^{-1} x^T y_i \\ \hat{\eta}_{it} &= y_{it} - x_t \hat{\beta}_i \end{aligned} \tag{28}$$

and

$$S_i = \sum_t (\hat{\eta}_{it})^2 \tag{29}$$

The approximate distribution, $q(\phi_{\epsilon_i}|y)$, is set using frequentist results. The mean is set to $(T-K)/S$ and the variance to $2(T-K)/S^2$. This mean and variance gives the approximate distribution parameters b_{ϵ_i} and c_{ϵ_i} (see appendix A for the conversion).

We initialise $q(a_p|y)$ by using the frequentist solution to equation 2, where η_i is set to the residuals, $\hat{\eta}_i$. This gives:

$$\begin{aligned} \mu_{ai} &= (\tilde{\eta}_i \tilde{\eta}_i^T)^{-1} \tilde{\eta}_i \hat{\eta}_0 \\ \Lambda_{ai} &= \sigma_\epsilon^2 (\tilde{\eta}\tilde{\eta}^T)^{-1} \end{aligned} \tag{30}$$

where $\tilde{\eta}_i = [\hat{\eta_{i1}} \ldots \hat{\eta_{iP}}]$, $\hat{\eta}_{ip} = [\hat{\eta}_{i(p+1)} \ldots \hat{\eta}_{iT}]^T$ and where:

$$\sigma_{\epsilon_i}^2 = \sum_t (\eta_{it} - \eta_{i(t-p)}\mu_{ai})^2 \tag{31}$$

We also initialise $b_{a_p} = 1$, $c_{a_p} = 1$ and $\phi_{\bar{\beta}_k} = 1$.

3.2 f-contrasts

Variational Bayes gives us an approximation to the posterior distribution, $p(\beta, a, \phi_\epsilon|y)$. From this we can obtain the approximate marginal posterior distribution, $q(\beta, \bar{\beta}|Y)$, as being a multivariate Normal distribution (equation 25). If we write this as:

$$\begin{bmatrix} \beta_i \\ \bar{\beta}_i \end{bmatrix} \mid y \sim MVN\left(\begin{bmatrix} \mu_{\beta_i} \\ \mu_{\bar{\beta}_i} \end{bmatrix}, \begin{bmatrix} \Lambda_{\beta_i} & \Lambda_{\beta_i\bar{\beta}_i} \\ \Lambda_{\beta_i\bar{\beta}_i} & \Lambda_{\bar{\beta}_i} \end{bmatrix} \right) \tag{32}$$

We can marginalise to get the marginal distribution over the regression parameters, $q(\beta_i|Y)$, as:

$$\beta_i|y \sim MVN(\mu_{\beta_i}, \Lambda_{\beta_i}) \tag{33}$$

We can now use the marginal distribution in equation 33 to perform inference. In this paper we take the approach of using the f-contrast framework traditionally used with basis functions in the frequentist GLM framework (Josephs et al., 1997).

If c is a $(N_e N_b) \times J$ vector representing an f-contrast, we can use the f-contrast framework to compute the normalised power explained by the f-contrast:

$$f = \mu_{\beta_i}^T c (c^T \Lambda_{\beta_i} c)^{-1} c^T \mu_{\beta_i} / J \tag{34}$$

with degrees of freedom J and ∞. As with the use of basis functions in the frequentist GLM framework (Josephs et al., 1997), we lose directionality when doing an f-test. That is, we only investigate the *power* explained by linear combinations of the basis function parameters, regardless of the direction (i.e. whether it is an activation or deactivation). This means that we only look at the positive tail of the f-distribution to find both activations and deactivations.

The alternative to doing a test on β_i would be to ask "What is the probability that $\bar{D}_{ie}$ is greater than zero?". Recall from equation 11 that $\bar{D}_{ie}$, represents the HRF size (activation height). Can we recover the activation height, $\bar{D}_{ie}$, from the parameters we infer upon, β_{ie} and $\bar{\beta}_{ie}$? To do this, we can rewrite equations 20 and 21 to get:

$$\begin{aligned} \bar{D}_{ie} &= \bar{\beta}_{ie} \left(\sum_b \frac{\beta_{ie}^2}{N_b \bar{\beta}_{ie}^2} \right)^{1/2} \\ &= \operatorname{sign}(\bar{\beta}_{ie}) \left(\sum (\beta_{ie}^2 / N_b) \right)^{1/2} \end{aligned} \tag{35}$$

Note that the term $\left(\sum (\beta_{ie}^2 / N_b) \right)^{1/2}$ is the power we are testing when we do the f-test on β_i. Equation 35 tells us that we can use the sign of $\bar{\beta}_{ie}$ to give the sign of $\bar{D}_{ie}$, and therefore the direction of the activation.

We could look to derive the posterior probability of the normalised power explained by the f-contrast in equation 34. Instead, the approach we take in this paper is to convert them to pseudo-z-statistics and then perform spatial mixture modelling on the spatial map of pseudo-z-statistics as described later. The f-to-z transform is carried out by doing an f-to-p-to-z transform (i.e. by ensuring that the probabilities in the tails are equal under the f- and z-distributions for the f- and z-statistics).

We refer to them as *pseudo*-z-statistics as they are not necessarily Normally distributed with zero mean and standard deviation of one under the null hypothesis. This is because they have been obtained by performing Bayesian inference. Whether or not Bayesian inference produces the same null distribution as that in frequentist inference will depend on the form of prior used. As we shall see in section 4 and as we would expect (Penny et al., 2003), if we use noninformative priors we do get approximate equivalence between frequentist and Bayesian inference. However, when we use constrained HRF shape priors we get a different distribution under the null hypothesis. We will see later how we can adjust to this different inference, and take advantage of the extra sensitivity it offers, by using spatial mixture modelling.

4 Artificial null data

Before we consider the use of spatial mixture modelling on the pseudo-z-statistics representing the normalised power explained by an f-contrast, we investigate the empirical null distributions produced from artificial data using the Variational Bayesian inference previously described.

4.1 Methods

We use artificial null data generated from the model, with $\beta = 0$ and $a_p = 0$. The stimulus is taken to be a square wave with period of 60 seconds. We have $T = 200$ scans, a TR of 3 seconds and use an HRF resolution of $\rho = 6$. Using these values we generate time-series for 10000 voxels. We then fit the model with an autoregressive order of $P = 4$ and use a basis set of the top 3 eigenHRFs shown in figure 3. We use two different models, one with no HRF constraints ($m = 0$ and $C = I$) and one with HRF constraints ($m = \tilde{m}$ and $C = \tilde{C}$ as given in equation 17). We use an f-contrast of $[1, 0, 0; 0, 1, 0; 0, 0, 1]$.

4.2 Results

Figure 5(a) shows the histogram of pseudo-z-statistics obtained for the two different models with and without HRF constraints using the Variational Bayesian inference. Figure 5(b) shows the log probability-log probability plots. These show plots of (nominal/theoretical) frequentist FPR against that obtained empirically.

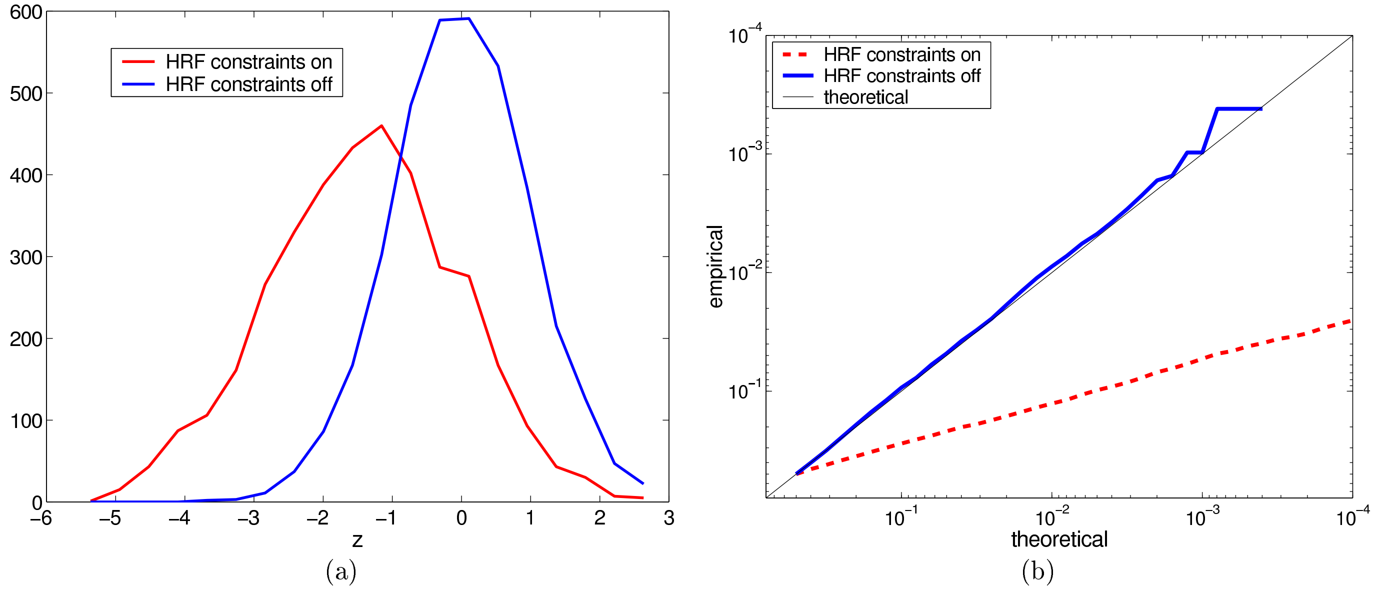


Figure 5: (a) Histogram of pseudo-z-statistics obtained for the two different models with and without HRF constraints. Figure 5(b) Log probability-log probability plots. These show plots of (nominal/theoretical) frequentist FPR against that obtained empirically. The HRF constraints reduce the power in those voxels where the linear combinations of the basis functions do not give sensible HRF shapes; this produces a shift in the null-distribution histogram.

The nominal/theoretical FPR is only applicable to the unconstrained HRF model, as we then have noninformative priors and we would expect the Bayesian inference to be equivalent to frequentist inference. Accordingly, the log probability-log probability plot shows good correspondence between the empirically obtained probabilities under the tail for a given z-statistic, and that which we expect from frequentist theory, for the unconstrained HRF model.

Recall that to make the inference tractable under Variational Bayes, we introduced a utility parameter, $\phi_{\bar{\beta}_{ie}}$, which we update with point estimates an approximation which may have effected the marginal posterior over β_{ie}. The fact that we obtain good correspondence here between our inference and the (for this model, known to be correct) frequentist results provides some validation that the marginal posterior over β_e is *not* significantly affected.

Unlike the unconstrained model, we would not expect the constrained HRF model to give pseudo-z-statistics that conform to frequentist theory. This is because we now have informative HRF shape priors causing Bayesian inference to be different to frequentist inference. The log probability-log probability plot shows that we get probabilities under the tail for a given z-statistic much smaller empirically than if the frequentist GLM solution held true. The histogram in figure 5(a) shows that this is due to a large shift in the histogram to lower pseudo-z-statistics (the mode is at about $z = -1$). The HRF constraints reduce the power in those voxels where the linear combinations of the basis functions do not give sensible HRF shapes; this produces a shift in the null-distribution histogram. When we have activating voxels with HRF shapes which are *not* penalised by the HRF prior constraints then the pseudo-z-statistics will not be reduced. We therefore have extra sensitivity when we use Bayesian inference with informative priors constraining the HRF shape.

5 Spatial Mixture Modelling

Using Variational Bayesian inference we have produced a spatial map of pseudo-z-statistics. Since the pseudo-z-statistics for the constrained HRF model give smaller empirical null distribution probabilities than the nominal FPR, we could use the pseudo-z-statistics and infer on them using frequentist probabilities without violating our nominal FPR. However, we would not be taking full advantage of the extra sensitivity of the Bayesian inference when we have informative priors constraining the HRF shape.

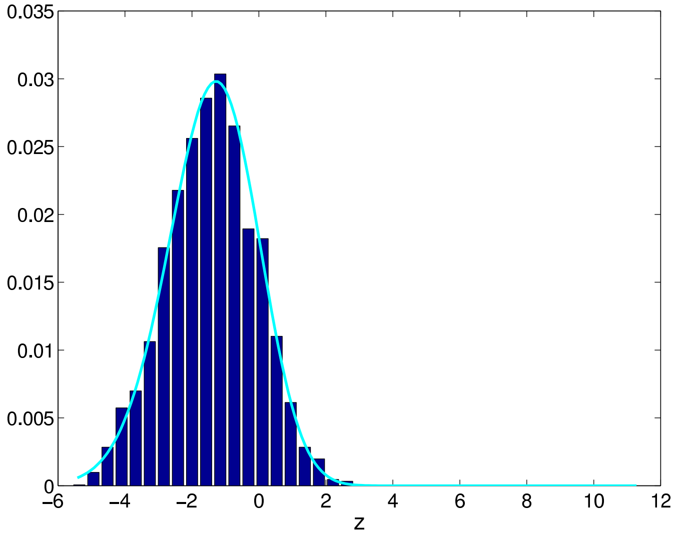


Figure 6: Null distribution of pseudo-z-statistics resulting from the constrained HRF analysis on the artificial null data in section 4. This is the same histogram as that shown in red in figure 5(a). Also shown is the fit to this histogram of a flipped and shifted Gamma distribution.

Therefore, we look to infer on the spatial map of pseudo-z-statistics using the technique of spatial mixture modelling. In mixture modelling we model the spatial map as being made up of voxels which are classified as either activating or non-activating (Everitt and Bullmore, 1999). The activation and deactivation classes have different distributions which describe probabilistically the pseudo-z-statistic values we expect in that class. Since we are estimating the non-activating distribution from the data we can adjust to the shifted non-activation distribution demonstrated in figure 5.

In spatial mixture modelling we augment this histogram information with spatial regularisation of the classification (Hartvig and Jensen, 2000; Woolrich et al., 2004a). Note that this has nothing to do with the way in which we spatially regularise the temporal autocorrelation coefficients in section 2.2. Instead, we a spatially regularising the classification of activating or non-activating voxels. This has been implemented in Woolrich et al. (2004a), within a completely adaptive fully Bayesian framework. Importantly, this approach adaptively determines the amount the classification is spatially regularisation.

Figure 6 shows the null distribution of pseudo-z-statistics resulting from the constrained HRF analysis on the artificial null data in section 4. This distribution is not Normal, in particular it is asymmetric. We therefore model the non-activation distribution in the mixture modelling as a flipped and shifted Gamma distribution. Figure 6 shows the fit of a flipped and shifted Gamma distribution to the null distribution histogram. Recall that with an f contrast we only look at the positive tail of the distribution to find both activations and deactivations. Therefore, we model the activations and deactivations as a single class in the spatial mixture modelling using a straightforward Gamma distribution.

6 FMRI data

6.1 Methods

We use two different datasets. The first is a boxcar visual experiment with a reversing checkerboard boxcar stimulus (30 seconds on, 30 seconds off). The second is a single-event pain experiment, for which the stimulus was a thermal noxious stimuli of 3 seconds duration administered to the dorsum of the volunteer's left hand using an electrical resistor to generate heat with varying ISI (between 30 seconds and 50 seconds).

For both experiments echo planar images (EPI) were acquired using a 3 Tesla system with TR=3 seconds, time to echo (TE) = 30ms, in-plane resolution 4mm and slice thickness 7mm. The first 4 scans were removed and the data was motion corrected using MCFLIRT (Jenkinson et al., 2002) and high-pass filtered as described in Woolrich

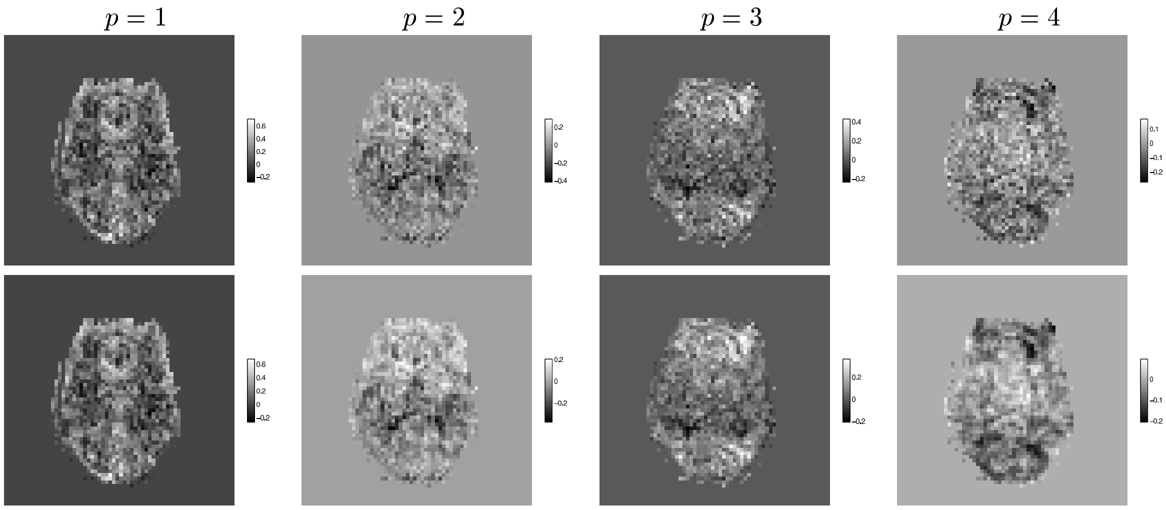


Figure 7: Posterior means of parameters from an autoregressive model of order 4 from the single-event pain dataset. From left to right we have increasing p from 1 to 4. We show this for two different priors on the AR parameters: [top] non-informative non-spatial, and [bottom] spatial MRF.

et al. (2001). The data is *not* spatially smoothed.

We use an HRF resolution of $\rho = 6$ (i.e. 0.5 seconds). We then fit the model with an autoregressive order of $P = 4$ and use a basis set of the top 3 eigenHRFs shown in figure 3. We use two different models, one with no HRF constraints ($m = 0$ and $C = I$) and one with HRF constraints ($m = \tilde{m}$ and $C = \tilde{C}$ as given in equation 17). We use an f-contrast to pick out the linear combinations of the basis functions. The resulting f-statistics are f-to-z converted to produce spatial maps of pseudo-z-statistics. The fully adaptive mixture modelling described in section 5 is then used to provide probabilities of a voxel being activated.

6.2 Results

Figure 7 shows the posterior means of the AR parameters from the single-event pain dataset for two different models: with non-spatial and with spatial MRF AR parameter priors. For the model with spatial MRF priors, the Variational Bayesian inference adaptively determines the amount of spatial regularisation to impose for each of the autoregressive parameter MRFs separately. The spatial maps for $p = 1$ are very similar, but as we increase p the MRF spatial regularisation increases. This shows how the adaptive determination of the amount of spatial regularisation automatically adjusts to avoid overfitting of the high order autoregressive parameters. In terms of inference on the normalised power of the f-contrasts, figure 8(a) suggests that avoiding overfitting and spatially regularising the AR parameters does make a significant difference to the pseudo-z-statistics.

Figure 9 shows the histogram of pseudo-z-statistics obtained for the two different models with and without HRF constraints. For both conditions we can see how the right hand tail, i.e. those voxels which are strongly activating, is relatively unaffected. Whereas, the main body of the histogram, i.e. the background non-activating voxels, is shifted to the left in the same way that it was for the null artificial data.

Figures 10 and 12 show the results of applying the spatial mixture modelling on the unconstrained HRF model. Figures 11 and 13 show the results for the constrained HRF model. Figure 14 shows the difference in voxel classification between the constrained HRF model and the unconstrained HRF model. This difference highlights the increased sensitivity. With the constrained HRF model smaller strength activating voxels have increased probability of being in the activation class. This is because the non-activating class distribution is shifted to lower pseudo-z-statistics when we use the constrained HRF model.

Figures 15 and 16 show samples from the marginal posterior of the HRF at a voxel which is not activating and a voxel which is strongly activating respectively in the pain experiment. In particular, figure 15 highlights the difference between the unconstrained and constrained HRF models for a non-activating voxel. Whereas, in figure 16 the HRF shape conforms to prior expectations, hence there is little different between the unconstrained

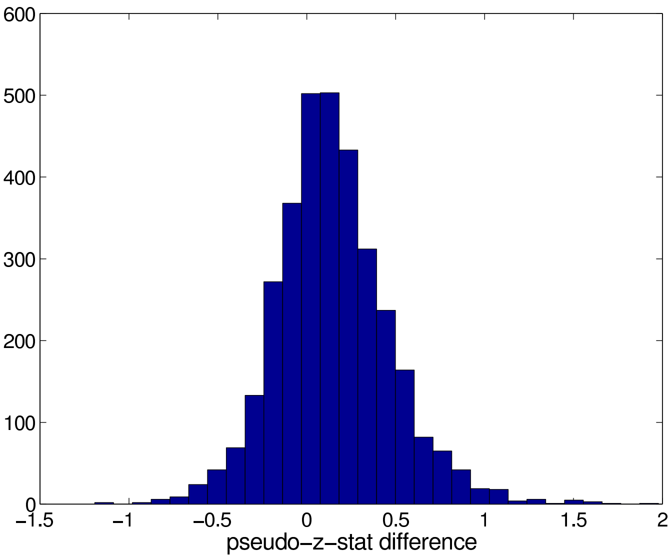
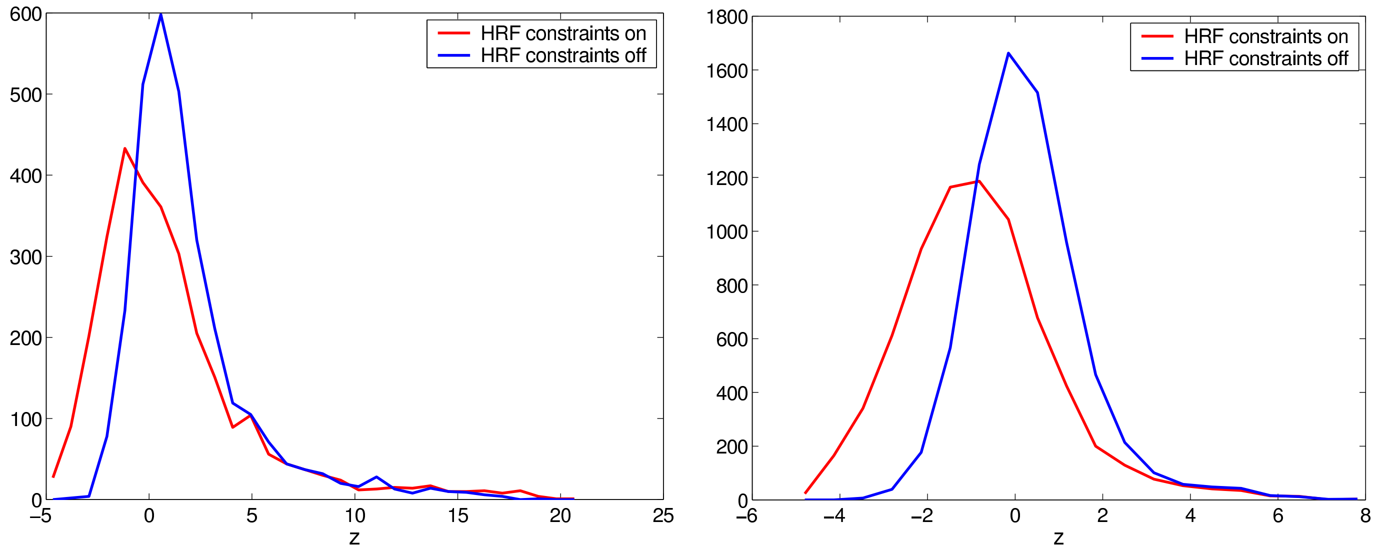


Figure 8: Histogram of the voxelwise difference in pseudo-z-statistics between the model with non-spatial non-informative AR priors and the model with adaptive spatial MRF AR priors for the single-event pain dataset.



(a) Visual boxcar

(b) Pain single-event

Figure 9: Histogram of pseudo-z-statistics obtained for two different models (with and without HRF constraints) on (a) the visual boxcar dataset and (b) the pain single-event dataset.

and constrained HRF models for a strongly activating voxel.

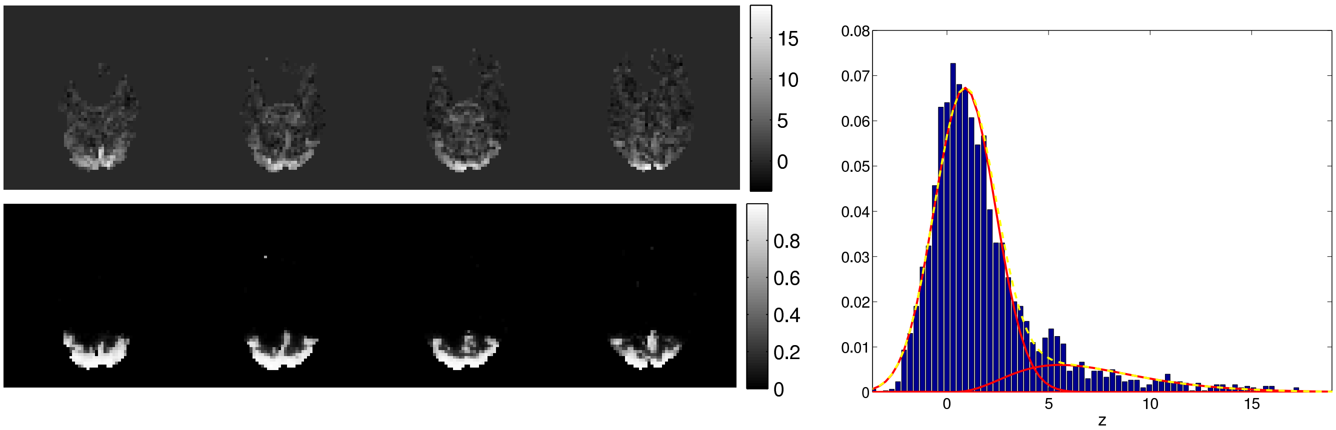


Figure 10: Visual condition. Results of applying the spatial mixture modelling on the unconstrained HRF model. [top] Pseudo-z-statistic spatial maps. [bottom] Probability of being in the activation class. [right] Mixture model fit to histogram of pseudo-z-statistics.

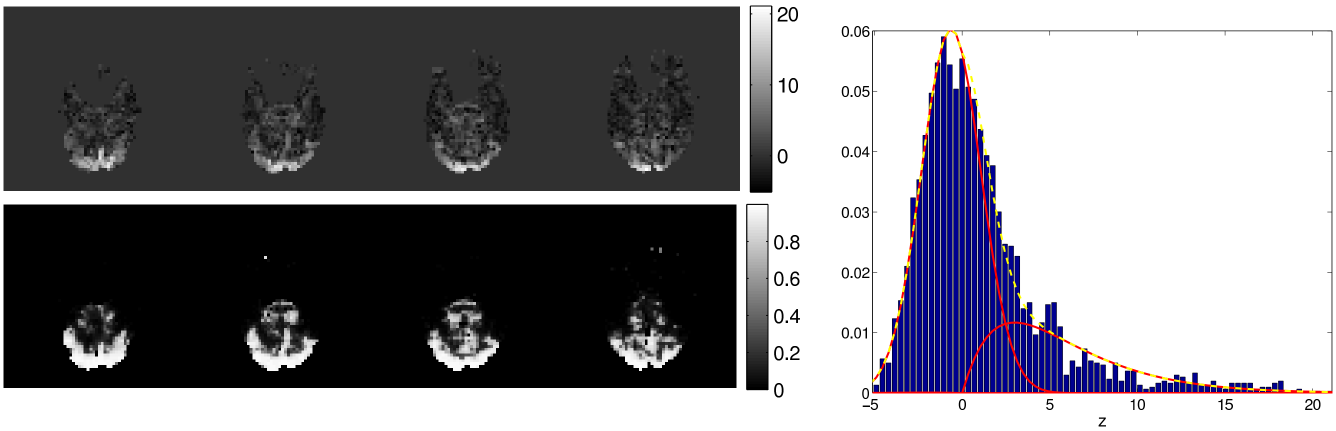


Figure 11: Visual condition. Results of applying the spatial mixture modelling on the constrained HRF model. [top] Pseudo-z-statistic spatial maps. [bottom] Probability of being in the activation class. [right] Mixture model fit to histogram of pseudo-z-statistics ('A' and 'B' mark the pseudo-z-statistics for which voxels with greater pseudo-z-statistics have higher probability of being active than non-active for the constrained HRF model and unconstrained HRF model respectively).

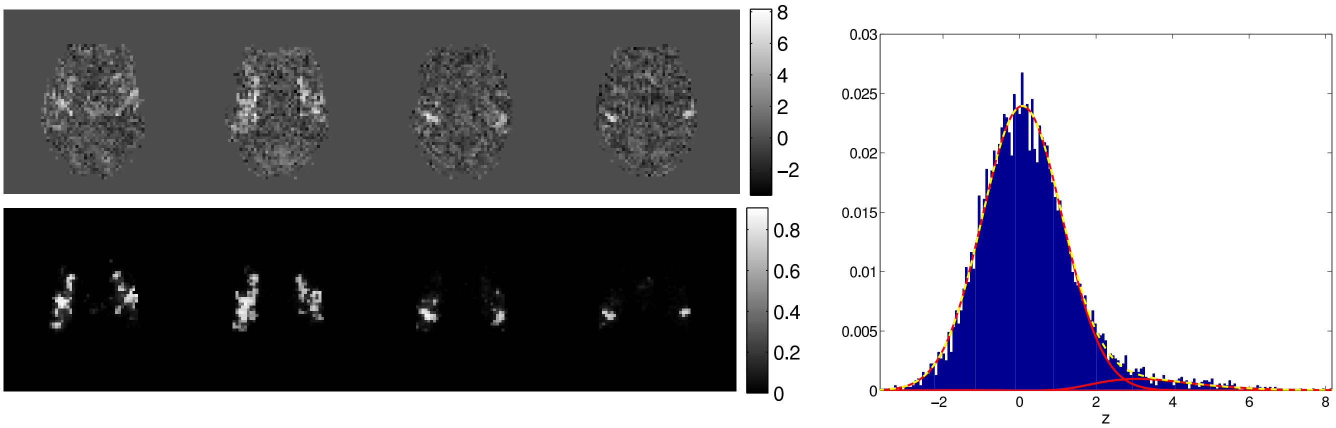


Figure 12: Pain single-event condition. Results of applying the spatial mixture modelling on the unconstrained HRF model. [top] Pseudo-z-statistic spatial maps. [bottom] Probability of being in the activation class. [right] Mixture model fit to histogram of pseudo-z-statistics.

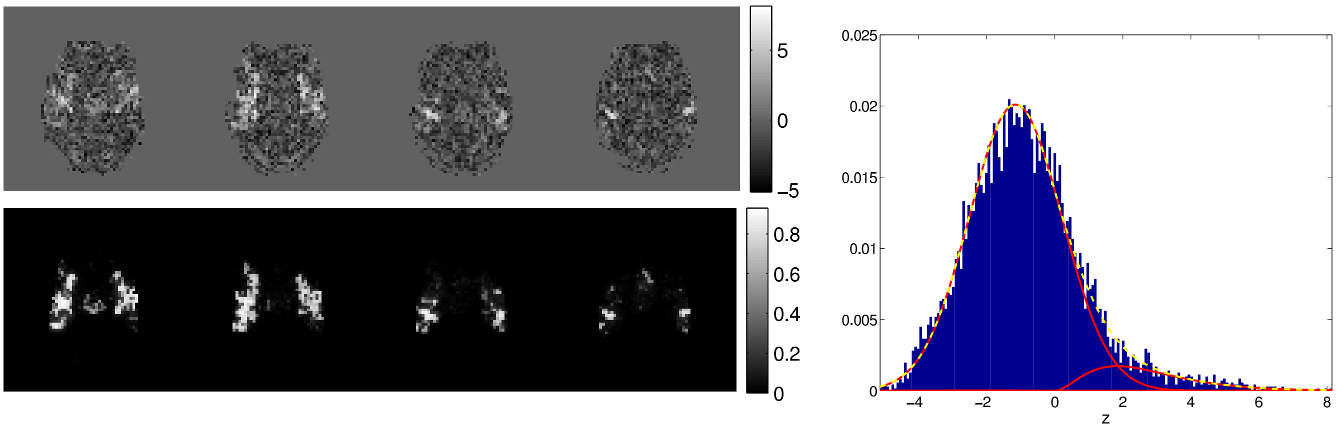


Figure 13: Pain single-event condition. Results of applying the spatial mixture modelling on the constrained HRF model. [top] Pseudo-z-statistic spatial maps. [bottom] Probability of being in the activation class. [right] Mixture model fit to histogram of pseudo-z-statistics ('A' and 'B' mark the pseudo-z-statistics for which voxels with greater pseudo-z-statistics have higher probability of being active than non-active for the constrained HRF model and unconstrained HRF model respectively).

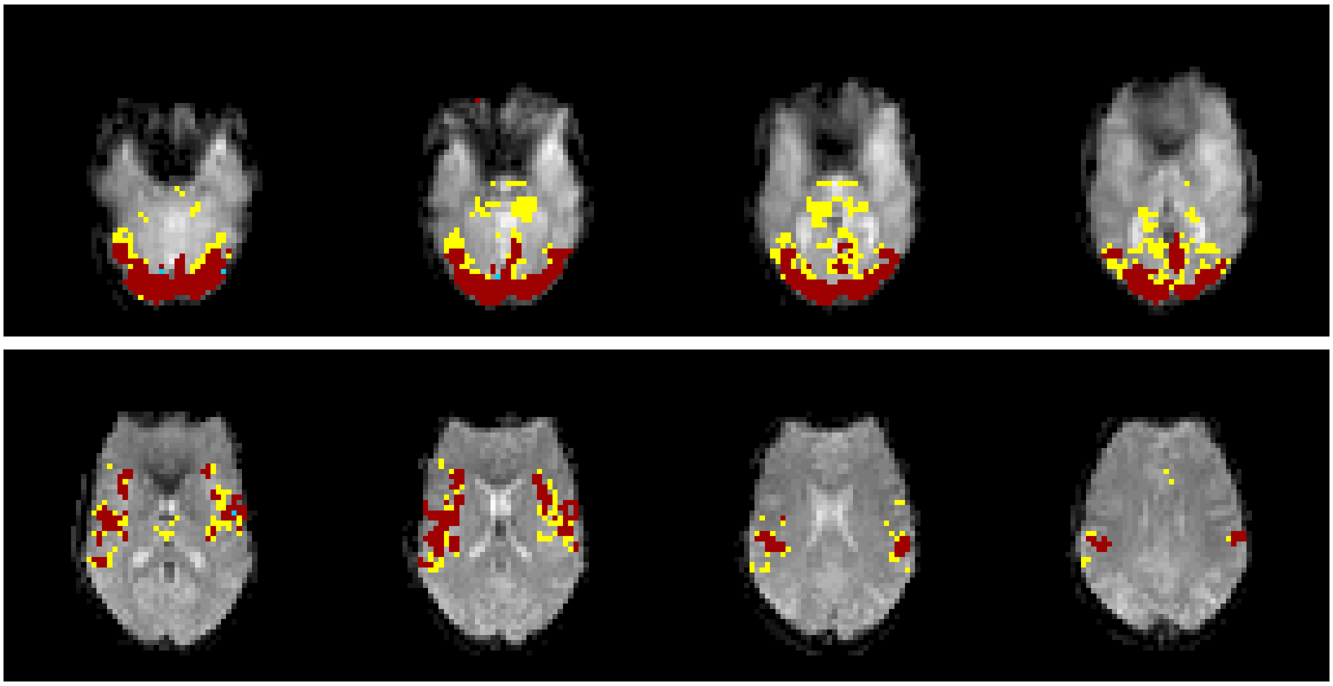
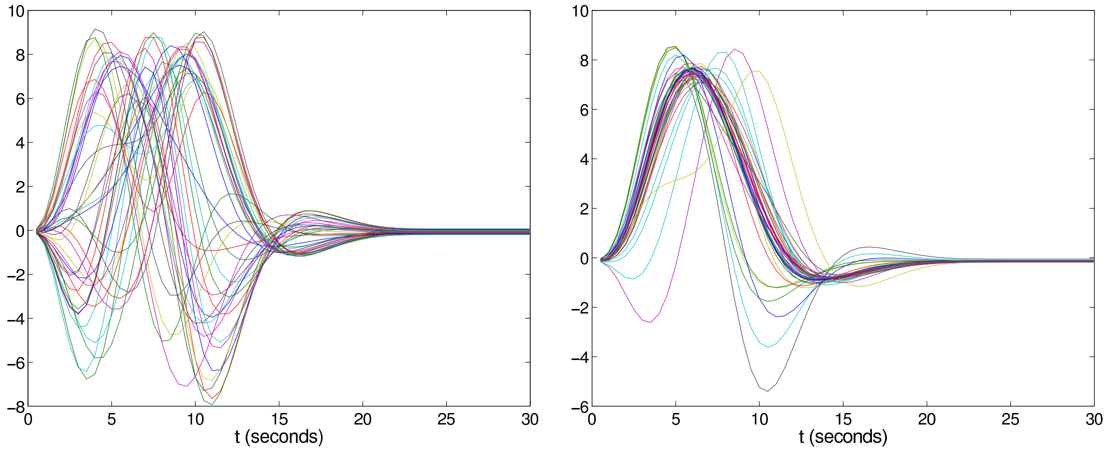
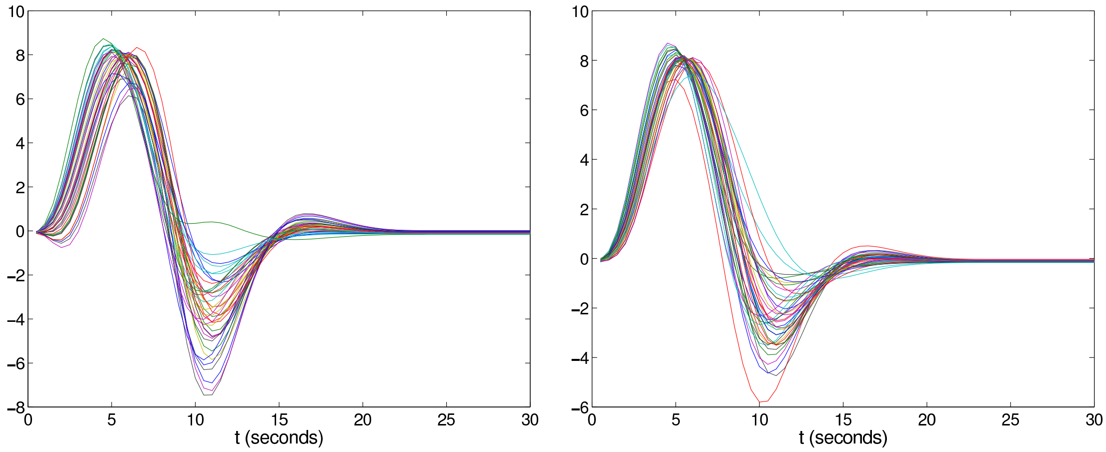


Figure 14: Difference in voxel classification between the constrained HRF model and the unconstrained HRF model. [red] voxels are active for both models, [blue] voxels are active for just the unconstrained HRF model, and [yellow] voxels are active for just the constrained HRF model. Voxels are classified as activating with probability of being in the activation class greater then 0.5. [top] Visual boxcar dataset [Bottom] Pain single-event dataset.



(a) Unconstrained (b) Constrained

Figure 15: Samples from the marginal posterior of the HRF at a single voxel which is not activating in the pain experiment. (a) Unconstrained with $m = 0$ and $C = I$ (b) Constrained with $m = \tilde{m}$ and $C = \tilde{C}$



(a) Unconstrained (b) Constrained

Figure 16: Samples from the marginal posterior of the HRF at a single voxel which is strongly activating in the pain experiment. (a) Unconstrained with $m = 0$ and $C = I$ (b) Constrained with $m = \tilde{m}$ and $C = \tilde{C}$

7 Discussion

In previous work Friman et al. (2003) also looked to constrain the possible linear combinations of the basis set, but within the canonical correlation analysis (CCA) framework. However, they only looked to constrain the linear combination coefficients to be positive. In this work we apply a more complete constraint by fitting a multivariate Normal distribution to describe the desired constrained space probabilistically. A big limitation of the work in Friman et al. (2003) is that they did not address the issue of how to threshold the resulting correlations of the CCA. In contrast, the framework in this paper is the Variational Bayesian GLM framework first used in FMRI by Penny et al. (2003). This framework has the advantage that it takes into account important issues such as temporal autocorrelation in FMRI and at the same time intrinsically produces approximate probability distributions from which inference can take place.

The HRF modelling in this paper all assumes linearity of the HRF. Friston et al. (1998b) produced compelling work addressing the use of basis functions for non-linear HRF modelling using Volterra series. They model the first and second order kernels using Gamma basis functions in a frequentist inferred GLM. In Friston et al. (2000) they derive Volterra kernels from the balloon model (Buxton et al., 1998) and fit them to empirically found Volterra kernels from the frequentist inferred GLM of Friston et al. (1998b). In Friston (2002) they infer on the balloon model parameters directly from the FMRI data in a Bayesian framework. Within the Bayesian framework they can incorporate priors on the balloon model parameters deduced empirically in Friston et al. (2000). This incorporation of the prior information from previous empirical evidence will constrain the balloon model parameters in the same way we constrain the HRF shape basis function parameters in this work. One area of future work is to extend the Variational Bayesian inference in this paper to deal with second order Volterra kernel basis functions and nonlinearities in ways related to the work of Friston et al. (1998b).

8 Conclusions

We have proposed a technique for firstly choosing a basis set, and then the means to constrain the subspace spanned by the basis set to only include sensible HRF shapes within a GLM framework. We can carry out inference using Variational Bayes which also performs adaptive spatial regularisation of temporal autocorrelation. Constraining the subspace spanned by the basis set allows for far superior separation of activating voxels from non-activating voxels in FMRI data. We used spatial mixture modelling to produce final probabilities of activation, and demonstrated on FMRI data the increased sensitivity produced.

Acknowledgements The authors would like to acknowledge support from the UK MRC and EPSRC. Thanks must also go to Heidi Johansen-Berg, Mark Jenkinson and Christian Beckmann for their valuable input. Thanks to Richard Wise, Richard Rogers and Debbie Painter of the Nuffield Dept Anaesthetics, JR Hosp., Oxford Univ for the pain dataset.

A Gamma Distribution

x has a two-parameter gamma distribution, denoted by $Ga(b, c)$, with parameters b and c, if its density is given by:

$$f_{Ga}(x; b, c) = \frac{1}{\Gamma(c)} \frac{x^{c-1}}{b^c} e^{-x/b} \tag{36}$$

where $\Gamma(c)$ is the Gamma function. Note, that a gamma distribution has $mean = bc$ and $variance = b^2 c$.

B Variational Bayes Updates

We define:

$$B_i = \begin{bmatrix} \beta_i \\ \hat{\beta}_i \end{bmatrix} \tag{37}$$

and:

$$\begin{aligned} \beta_i &= Q B_i \\ \beta_{ie} &= Q_{ie} B_i \\ \bar{\beta}_{ie} &= R_{ie} B_i \end{aligned} \tag{38}$$

$$\tag{39}$$

B.1 Regression coefficient updates

Here we give the update equation for the parameters of the regression coefficient distribution $q(\beta_i, \bar{\beta}_i | y) = MVN(\mu_{B_i}, \Lambda_{B_i})$:

$$\begin{aligned} \mu_{B_i} &= F_{B_i}^{-1} E_{B_i} \\ \Lambda_{B_i} &= F_{B_i}^{-1} \end{aligned} \tag{40}$$

where:

$$\begin{aligned} F_{B_i} &= \gamma_{\epsilon_i} \sum_t \left\{ Q^T x_i^T x_t Q + \sum_p \mu_{a_{pi}} (Q^T x_{t-p}^T x_{t-p} Q - 2 Q^T x_t^T x_{t-p} Q) \right\} \\ &\quad + \sum_e \left\{ \phi_{\bar{\beta}_e}^{(i)} (Q_e^T C Q_e + R_e^T m^T C m R_e - 2 Q_e^T C m R_e) \right\} \\ E_{B_i} &= \gamma_{\epsilon_i} \sum_t \left\{ Q^T x_t^T (y_{it} - \sum_p y_{i(t-p)} \mu_{a_{pi}}) \right. \\ &\quad + \left. \sum_p \mu_{a_{pi}} Q^T x_{t-p}^T (y_{it} - \sum_p y_{i(t-p)} \mu_{a_{pi}}) \right\} \\ \gamma_{\epsilon_i} &= \frac{\Gamma(b_{\epsilon_i}) \Gamma(c_{\epsilon_i} + 1)}{\Gamma(c_{\epsilon_i})} \end{aligned} \tag{41}$$

B.2 Autoregressive parameter updates

Here we give the update equation for the parameters of the autoregressive parameter distribution $q(a_p | y) = N(\mu_{a_p}, \Lambda_{a_p})$:

$$\begin{aligned} \mu_{a_p} &= F_{a_p}^{-1} E_{a_p} \\ \Lambda_{a_p} &= F_{a_p}^{-1} \end{aligned} \tag{42}$$

where:

$$
\begin{aligned}
F_{a_p} &= \sum_t (w_{tp}^T S_\epsilon w_{tp}) + \gamma_{a_p} D \\
E_{a_p} &= \sum_t w_{tp}^T S_\epsilon (y_t - \sum_{q \neq p} w_{tq} \mu_{a_q} - x_t Q \mu_B)
\end{aligned}
\tag{43}
$$

where $S_\epsilon = \mathrm{Diag}(\gamma_{\epsilon_i})$ and $w_{tp} = y_{t-p} - x_{t-p} Q \mu_B$

B.3 MRF precision updates

Here we give the update equation for the MRF precision parameter distribution $q(\phi_{a_p}|y) = Ga(b_{\phi_{a_p}}, c_{\phi_{a_p}})$:

$$
\frac{1}{b_{a_p}} = \frac{1}{2}\left(m_{a_p}^T D m_{a_p} + \mathrm{Trace}(\Lambda_{a_p} D)\right) + \frac{1}{b_{a_0}} \tag{44}
$$

$$
c_{a_p} = N/2 + c_{a_0} \tag{45}
$$

we define γ_{a_p} as:

$$
\gamma_{a_p} = \frac{\Gamma(b_{a_p})\Gamma(c_{a_p}+1)}{\Gamma(c_{a_p})} \tag{46}
$$

Even though the matrix D is very sparse the computation of the $\mathrm{Trace}(\Lambda_{a_p} D)$ term can be very expensive to compute. In particular, this is because $\Lambda_{a_p} = F_{a_p}^{-1}$ and F_{a_p} is a $N \times N$ matrix whose inverse would be very computationally expensive to compute. Therefore, instead of computing this inverse we can compute x in the linear equation:

$$
F_{a_p} x = D \tag{47}
$$

Since D is a positive symmetric definite matrix we can take advantage of the conjugate gradient techniques described in Golub and Van Loan (1996). At each iteration of the conjugate gradient search for x we only need to perform one matrix multiplication of $F_{a_p} x$. The conjugate gradient approach is far quicker than solving for the inverse of F_{a_p} and then multiplying by D.

The conjugate gradient technique takes in an initial guess of x. Hence, as we iterate through the Variational Bayes updates of our approximate posterior distributions, we can store the the value of x from the previous conjugate gradient solution from the previous update of $q(\phi_{a_p}|y)$, and use it as the initialisation of the conjugate gradient search for x at the next update of $q(\phi_{a_p}|y)$. After the first Variational Bayes iteration this makes subsequent conjugate gradient searches very quick to converge.

B.4 Noise precision updates

Here we give the update equation for the parameters of the noise precision distribution $q(\phi_{\epsilon_i}|y) = Ga(b_{\epsilon_i}, c_{\epsilon_i})$:

$$
\begin{aligned}
\frac{1}{b_{\epsilon_i}} &= \frac{1}{2}\Bigg(\sum_t (y_{it} - x_t Q \mu_{B_i} - \sum_p (y_{i(t-p)} - x_{t-p} Q \mu_{B_i}) a_{pi})^2 + \mathrm{Trace}(\Lambda_{\beta_i}^{-1} \sum_t x_t^T x_t) \\
&+ \sum_p \left\{ (\Lambda_{a_p}^{-1})_i \sum_t (y_{i(t-p)} - x_{t-p} Q \mu_{B_i})^2 \right\}\Bigg) + \frac{1}{b_{\epsilon_0}} \\
c_{\epsilon_i} &= (T-1)/2 + c_{\epsilon_0}
\end{aligned}
\tag{48}
$$

we define γ_{ϵ_i} as:

$$
\gamma_{\epsilon_i} = \frac{\Gamma(b_{\epsilon_i})\Gamma(c_{\epsilon_i}+1)}{\Gamma(c_{\epsilon_i})} \tag{49}
$$

References

Boynton, G., Engel, S., Glover, G., and Heeger, D. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. *Journal of Neuroscience*, 16:4207–4221.

Buxton, R., Wong, E., and Frank, L. (1998). Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. *Magnetic Resonance in Medicine*, 39:855–864.

Cohen, M. (1997). Parametric analysis of fMRI data using linear systems methods. *NeuroImage*, 6:93–103.

Cressie, N. (1993). *Statistics for Spatial Data*. Wiley, New York.

Dale, A. and Buckner, R. (1997). Selective averaging of rapidly presented individual trials using fMRI. *Human Brain Mapping*, 5:329–340.

Everitt, B. and Bullmore, E. (1999). Mixture model mapping of brain activation in functional magnetic resonance images. *Human Brain Mapping*, 7:1–14.

Friman, O., Borga, M., Lundberg, P., and Knutsson, H. (2003). Adaptive analysis of fMRI data. *NeuroImage*, 19(3):837–845.

Friston, K., Fletcher, P., Josephs, O., Holmes, A., Rugg, M., and Turner, R. (1998a). Event-related fMRI: Characterizing differential responses. *NeuroImage*, 7:30–40.

Friston, K., Frith, C., Turner, R., and Frackowiak, R. (1995). Characterizing evoked hemodynamics with fMRI. *NeuroImage*, 2:157–165.

Friston, K., Josephs, O., Rees, G., and Turner, R. (1998b). Nonlinear event-related responses in fMRI. *Magnetic Resonance in Medicine*, 39:41–52.

Friston, K., Mechelli, A., Turner, R., and Price, C. (2000). Nonlinear responses in fMRI: the balloon model, Volterra kernels, and other hemodynamics. *NeuroImage*, 12:466–477.

Friston, K., Worsley, K., Frackowiak, R., Mazziotta, J., and Evans, A. (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, 1:214–220.

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

Genovese, C. (2000). A Bayesian time-course model for functional magnetic resonance imaging data (with discussion). *Journal of the American Statistical Association*, 95:691–703.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations*. Johns Hopkins University Press, Baltimore, third edition.

Gössl, C., Auer, D., and Fahrmeir, L. (2001). Bayesian modeling of the haemodynamic response function in BOLD fMRI. *NeuroImage*, 14(1):140–148.

Hartvig, N. and Jensen, J. (2000). Spatial mixture modelling of fMRI data. *Human Brain Mapping*, 11(4):233–248.

Hossein-Zadeh, G. and Ardekani, B. (2002). A signal subspace for modeling the hamodynamic response function in fMRI. In *Eighth Int. Conf. on Functional Mapping of the Human Brain, NeuroImage*.

Jenkinson, M., Bannister, P., Brady, J., and Smith, S. (2002). Improved optimisation for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2):825–841.

Jordan, M. (1999). *Learning in Graphical Models*. MIT Press.

Josephs, O., Turner, R., and Friston, K. (1997). Event-related fMRI. *Human Brain Mapping*, 5:1–7.

Penny, W., Kiebel, S., and Friston, K. (2003). Variational bayesian inference for fMRI time series. *NeuroImage*, 19(3):1477–1491.

Woolrich, M., Behrens, T., Beckmann, C., and Smith, S. (2004a). Mixture models with adaptive spatial regularisation for segmentation with an application to FMRI data. Technical Report TR04MW1, Oxford Centre for Functional Magnetic Resonance Imaging of the Brain, Department of Clinical Neurology, Oxford University, Oxford, UK. Available at http://www.fmrib.ox.ac.uk/analysis/techrep for downloading. Also in submission.

Woolrich, M., Jenkinson, M., Brady, J., and Smith, S. (2004b). Fully Bayesian spatio-temporal modelling of FMRI data. *IEEE Trans. on Medical Imaging*, 22(2).

Woolrich, M., Ripley, B., Brady, J., and Smith, S. (2001). Temporal autocorrelation in univariate linear modelling of FMRI data. *NeuroImage*, 14(6):1370–1386.

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