Fully Bayesian Spatio-temporal Modelling of FMRI Data

FMRIB Technical Report TR03MW2
(A related paper has been accepted for publication in IEEE TMI)

Mark W. Woolrich, Mark Jenkinson, J. Michael Brady 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

We present a fully Bayesian approach to modelling in FMRI, incorporating spatio-temporal noise modelling and haemodynamic response function (HRF) modelling. A fully Bayesian approach allows for the uncertainties in the noise and signal modelling to be incorporated together to provide full posterior distributions of the HRF parameters. The noise modelling is achieved via a non-separable space-time vector autoregressive process. Previous FMRI noise models have either been purely temporal, separable or modelling deterministic trends. The specific form of the noise process is determined using model selection techniques. Notably, this results in the need for a spatially non-stationary and temporally stationary spatial component. Within the same full model, we also investigate the variation of the HRF in different areas of the activation, and for different experimental stimuli. We propose a novel HRF model made up of half-cosines, which allows distinct combinations of parameters to represent characteristics of interest. In addition, to adaptively avoid over-fitting we propose the use of Automatic Relevance Determination priors to force certain parameters in the model to zero with high precision if there is no evidence to support them in the data. We apply the model to three datasets and observe matter-type dependence of the spatial and temporal noise, and a negative correlation between activation height and HRF time to main peak (although we suggest that this apparent correlation may be due to a number of different effects).

1 Introduction

Functional magnetic resonance imaging (FMRI) uses fast MRI techniques to enable studies of dynamic physiological processes at a time scale of seconds. This can be used for spatially localising dynamic processes in the brain, such as neuronal activity. However, to achieve this we need to be able to infer on models of 4-dimensional data. Predominantly, for statistical and computational simplicity, analysis of FMRI data is performed in two-stages. Firstly, the purely temporal nature of the FMRI data is modelled at each voxel independently, before considering spatial modelling on summary statistics from the purely temporal analysis [18, 14]. Clearly, it would be preferable to incorporate the spatial and temporal modelling into one all encompassing model. This would allow for correct propagation of uncertainty between temporal and spatial model parameters. Previous work [1, 8, 26, 29] has taken such a combined spatio-temporal approach. In this work we look to propose an alternative spatio-temporal approach.

There is a wealth of possibility when considering spatio-temporal models for FMRI. A systematic approach is required to determine the most appropriate model. We break down the work into two main areas in this paper. They are spatio-temporal noise modelling and haemodynamic response (HRF) signal modelling.

Previous FMRI spatio-temporal noise models have either been separable [1] or modelling deterministic trends [26]. As we shall see, the assumptions underlying a separable noise model are not well met by FMRI data. In this work, we focus not on longer-term (spatially and temporally) deterministic trends, but on shorter-term correlated noise processes. This results in the proposed use of a novel space-time non-separable autoregressive process to FMRI data.

In the signal modelling we focus on HRF modelling. It is now widely accepted that FMRI modelling requires flexible HRF modelling, with the HRF varying spatially and between subjects. Flexibility in linear modelling has been introduced with the use of basis functions [11, 33]. However, basis functions suffer from a number of limitations. They impose a hard constraint on the allowed HRF shape and often the extent of the constraint is difficult to control and/or interpret. To overcome these problems we use a parameterised HRF approach [21, 25]. Unlike previous work [21, 25] we do not limit ourselves to fixed epoch designs and we propose a novel half-cosine parameterisation to separately represent different shape characteristics of the HRF.

Importantly, and for the first time in FMRI, with such highly parameterised models we introduce the use of Automatic Relevance Determination (ARD) priors, to force parameters with insufficient evidence in the data to support them to be zero with high precision. This adaptively avoids over-fitting at the time of inference on the data.

The use of a Bayesian framework provide us with the ability to probabilistically incorporate prior information (for example, about the expected HRF shape). However, this is not the only reason for adapting the Bayesian approach. To correctly infer on parameters of interest, for example, the HRF parameters or activation height, is not an easy task with the complexity of model proposed here. Indeed, there are no solutions available in the frequentist literature, precluding the use of frequentist null hypothesis testing. However, as in previous work [21, 25, 26, 29] we are able to infer on these complex models by using a fully Bayesian framework. Whilst the fully Bayesian framework is not the only method for dealing with such complex models, it does provide us with a systematic framework for doing full inference by incorporating all of the uncertainty in the model parameters via marginalisation. Hence this approach is of great use, even if non-informative priors were to be assumed for all parameters.

2 Modelling Framework

It is assumed that the observed data, y_{it}, is decomposed into the addition of the response to the stimulus, r_{it}, and some noise process, e_{it}:

$$y_{it} = r_{it} + e_{it} \tag{1}$$

where i indexes the voxel and t indexes the time (or volume). Generally, r_{it} would be a function of the stimulus s_t. Next we need to turn our attention to the modelling of the separate components in equation 1, the response to the stimulus, r_{it}, and the noise process, e_{it}.

3 Spatio-temporal Noise Modelling

There has been extensive work on spatially independent temporal noise modelling. In particular, Friston et al. [12] and Bullmore et al. [3] introduced alternative approaches within a null hypothesis testing general linear model (GLM) framework. This can also be tackled in an *empirical* Bayesian framework to avoid problems with estimating the temporal autocorrelation from the residuals of the model fit [10, 9]. One of the major limitations of the estimation and inference of these techniques, is that the standard equations from the linear modelling statistics literature (e.g. [40, 4]) assume that the parameters associated with the correlation structure of the noise are known exactly. The uncertainty associated with these parameters is not taken into account, resulting in potentially biased statistics.

One solution to this is to consider a fully Bayesian framework. This was the approach taken by Genovese [21], in which he took a full Bayes approach to the modelling of FMRI time series, and inferred using MCMC. The noise is modelled as deterministic drift in the data using cubic splines. Gössl et al. [24] also look to model larger scale drifts over time using a random walk model.

Spatial correlation could arise from several potential sources. Firstly, there is point spread function due to the fact that we are performing finite sampling in k-space. Hence, data from an individual voxel will contain some signal from the tissue around that voxel, an effect compounded by motion. Secondly, there is smoothness introduced by motion correction techniques, in particular due to interpolation. Thirdly, there may be spatially spread physiological effects, which for our purposes would be considered as noise — although these spatial networks of "activity" in supposed rest/null data could be of real interest.

Gössl et al. [26] consider spatial noise in the context of a deterministic spatio-temporal trend model. However, it is yet to be determined whether it is adequate to merely model deterministic drift terms, and then assume that the remaining noise is white.

In this work we will take into account large scale deterministic temporal fluctuations using temporal high-pass filtering as a preprocessing step, and we will model short scale statistical spatial and temporal noise using autoregressive processes. In future work there is the possibility to incorporate large scale deterministic fluctuation modelling into the same model.

3.1 Scale of Variation

Here we are considering both spatial and temporal scale. That is, consideration is given to the scale at which components of variation, or processes, are occurring. We decompose the noise, e_{it}, into:

$$e_{it} = l_{it} + q_{it} \tag{2}$$

where l_{it} is the non-stationary deterministic mean structure or large-scale variation (spatially and temporally), and q_{it} is the stationary short-scale stochastic variation.

3.2 Large Scale Variation

In general, the large-scale variation represents deterministic trends across one or more of the 4 dimensions of the observed data space. Previous work [43, 16] has shown large scale temporal noise processes. We do not incorporate these processes into the model, but instead remove the worst of them by using high-pass filtering as a preprocessing step (i.e. in equation 2 we set $l_{it} = 0$). In this work the high pass filtering cut-off is chosen as the one which removes as much of the temporal low frequencies as possible without degrading the signal.

There is increasing evidence that there is large scale spatial structure in FMRI data possibly attributable to neuronal networks which are not locked to an imposed task [34]. When looking for task related activation these large scale spatial networks are effectively "noise/confounds". Ideally, a spatio-temporal model would attempt to model these confounds out. However, in this work we do not attempt to do so. This is an important area for future work.

3.3 Small Scale Variation

The stationary small-scale variation is modelled as a multivariate Normal noise process:

$$\boldsymbol{q} \sim MVN(0, \boldsymbol{\Sigma}) \tag{3}$$

The covariance matrix $\boldsymbol{\Sigma}$ will be of size $NT \times NT$. This is a large covariance matrix (e.g. $NT = 10000 \times 100 = 10^6$), and hence modelling it will require some simplification to make it computationally feasible.

3.3.1 Separable Models

One approach to simplifying equation 3 is to consider a stationary separable model [28, 1]. This gives:

$$\boldsymbol{q} \sim MVN(0, \boldsymbol{\Gamma} \otimes \boldsymbol{\Lambda}) \tag{4}$$

where $\otimes$ represents the Kronecker product and where $\boldsymbol{\Gamma}$ and $\boldsymbol{\Lambda}$ are the spatial ($N \times N$) and temporal ($T \times T$) covariance matrices respectively. This is where the covariance between two observations is decomposed into the temporal covariance at the lag between the observations multiplied by the spatial covariance at the distance between the observations. Such a decomposition would make computation a lot easier, for example when computing the inverse, which is necessary for all viable inference techniques. The use of equation 4 means that the spatial autocovariance is assumed to be the same at all time points. This seems a reasonable assumption. However, it also requires that the temporal autocovariance is the same at all voxels. Previous studies [43, 13], clearly demonstrated that this was not the case and would be an incorrect assumption to make. Consequently, a separable model is not considered further.

3.3.2 Space-Time Simultaneously specified Auto-Regressive model (STSAR)

The simultaneous auto-regressive approach is analogous to time series auto-regressive modelling [6]:

$$q_u = \sum_{v=1}^{NT} b_{uv} q_v + \epsilon_u \tag{5}$$

where $u \in \{1 \dots NT\}$ indexes a particular voxel and time point, and $\boldsymbol{\epsilon} \sim MVN(0, \boldsymbol{\Lambda})$, giving:

$$\boldsymbol{q} \sim MVN(\boldsymbol{0}, (\boldsymbol{I} - \boldsymbol{B})^{-1} \boldsymbol{\Lambda} (\boldsymbol{I} - \boldsymbol{B}')^{-1}) \tag{6}$$

where $\boldsymbol{B}$ is an $NT \times NT$ matrix whose (u,v)th element is b_{uv} and $\boldsymbol{\Lambda}$ is a diagonal covariance matrix (in this paper the variances along the diagonal of $\boldsymbol{\Lambda}$ vary voxelwise but not through time). Note that for a valid covariance matrix $(\boldsymbol{I} - \boldsymbol{B})$ must be invertible.

In using equation 5, Wikle et al. [42] lagged the spatial correlations by one time point. Then instead of having to consider the entire joint distribution, we can take advantage of conditional independence and factorise as follows:

$$p(\boldsymbol{q}|\boldsymbol{\epsilon}) = \prod_{i}^{N} \prod_{t}^{T} p(q_{it} | \{q_{j\tau}\}, \phi_{\epsilon_i}) \tag{7}$$

where $j \in \{1 \dots N\}, \tau \in \{1 \dots t-1\}$. The reason for lagging the spatial correlations by one time point and creating this conditional independence is to introduce directional acyclic dependency between the q parameters. Avoiding cyclic dependencies will improve the mixing of the MCMC sampling, whilst introducing no loss in the usefulness of the model. This is the approach that we will take in this work.

3.3.3 STSAR model in FMRI

We assume a temporally stationary but spatially non-stationary temporal component. It seems sensible that the spatial noise structure would be the same through time, hence we use a temporally stationary spatial autoregressive model. Also, the spatial autocorrelations are assumed to be anisotropic, that is we assume that the spatial autocorrelation is different in the three different spatial directions.

Therefore, we use an STSAR which combines a temporally fixed spatial AR of order 1 together with a spatially varying general order temporal AR. It is not clear whether the spatial component should be spatially stationary or non-stationary. Hence, we consider both modelling approaches. In either case we have:

$$q_{it} = \sum_{j \in \mathcal{N}_i} \beta_{ij} q_{j(t-1)} + \sum_{p=1}^{P} \alpha_{pi} q_{i(t-p)} + \epsilon_{it} \tag{8}$$

where β_{ij} is the spatial autocorrelation between voxel i and j at a time lag of one, α_{pi} is the temporal autocorrelation between time point t and $t-p$ at voxel i, and:

$$\epsilon_{it} \sim N(0, \phi_{\epsilon_i}^{-1}) \tag{9}$$

The spatially stationary spatial model sets $\beta_{ij} = \beta_d$ where d is the direction of the link i, j, giving:

$$\begin{aligned} q_{it} &= \sum_{d=1}^{D} \left(\beta_d \sum_{j \in \mathcal{N}_{id}} q_{j(t-1)} \right) \\ &+ \sum_{p=1}^{P} \alpha_{pi} q_{i(t-p)} + \epsilon_{it} \end{aligned} \tag{10}$$

where $j \in \mathcal{N}_{id}$ is the set of neighbouring voxels to voxel i in the d direction, there are $D = 3$ directions. The spatially non-stationary spatial model is the same as equation 8, but with the condition $\beta_{ij} = \beta_{ji}$.

3.3.4 Likelihood

Note that from equations 1 and 2 with $l_{it} = 0$, our data, y_{it}, is related to our spatio-temporal noise process, q_{it}, by:

$$y_{it} = r_{it} + q_{it} \tag{11}$$

where r_{it} is the signal component. This equation, along with equations 7, 8 and 9, describes our likelihood, $p(\boldsymbol{y}|\boldsymbol{r}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\phi_\epsilon})$. We now need to specify the priors on the noise parameters $\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\phi_\epsilon}\}$. For this we assume independence between the priors for these parameters, i.e. $p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\phi_\epsilon}) = p(\boldsymbol{\alpha})p(\boldsymbol{\beta})p(\boldsymbol{\phi_\epsilon})$. It is worth noting that assuming independence in the prior does not impose independence between the same parameters in the posterior distribution.

3.3.5 Autoregressive Parameter Priors

All of the autoregressive parameter priors we describe here appear to allow the autoregressive parameters to range from $-\infty$ to $-\infty$. However, in practice we effectively truncate these priors to only allow values between -1 to 1 by rejecting proposed values outside this range when we perform the MCMC sampling.

If we are using the spatially stationary spatial AR model (equation 10) then we use a noninformative prior of a disperse Gaussian:

$$p(\beta_d|\phi_\beta) \sim N(0, 1/\phi_\beta) \tag{12}$$

with ϕ_β set to a small number (0.001).

For both the non-stationary spatial (β_{ij}) and temporal autoregressive parameters (α_{pi}) we consider three different possible priors. These are now described for α_p; equivalent priors are used for β_{ij}.

Noninformative prior We use a disperse Gaussian prior:

$$p(\alpha_{pi}|\phi_\alpha) \sim N(0, 1/\phi_\alpha) \tag{13}$$

with ϕ_α set to a small number (0.001).

Markov Random Field (MRF) prior We would like to consider a model which spatially regularises the AR parameters. To do this we use a Markov Random Field (MRF) pairwise difference Gaussian prior [37] on the field $\alpha_p \equiv (\alpha_{pi})$:

$$p(\alpha_p|\phi_{\alpha_p}) \quad \propto \quad \phi_{\alpha_p}^{N/2} \exp\left\{ \frac{-\phi_{\alpha_p}}{4} \sum_i \sum_{j \in \mathcal{N}_i} (\alpha_{pi} - \alpha_{pj})^2 \right\} \tag{14}$$

where $j \in \mathcal{N}_i$ is the set of neighbouring voxels to voxel i (for this we use 26-connectivity in 3D). Note that the parameter controlling the amount of spatial regularisation, ϕ_{α_p}, is determined adaptively from the data (see section 3.3.6).

Automatic Relevance Determination (ARD) prior We are proposing to use a general order temporal AR model. The difficulty with this is that different voxels require different orders of temporal AR. The order varies between 0 and 5, but with few voxels with order greater than 2 [43]. Hence, we need some technique to allow the model to automatically adjust to the required AR order at each voxel.

Models with different order ARs have different number of parameters, which is a well known problem for MCMC techniques. One solution is to use reversible jumps [27] or jump diffusion [39] which allow jumps between models of different numbers of parameters. However, we can avoid this added complexity by employing a technique used in Bayesian modelling known as Automatic Relevance Determination (ARD) [36] from the neural network literature. ARD requires the use of a certain type of prior on a parameter whose relevance needs to be determined. The simplest prior to use for this purpose is a Gaussian with zero mean and precision ϕ which is also to be determined or sampled from. If the parameter in question is not required then the precision ϕ will be large, forcing the parameter to be close to zero. The benefit of ARD is that any unnecessary parameters are automatically forced to zero. The disadvantage is that it makes it difficult to incorporate other prior information at the same time as implementing ARD, and hence in this work the use of the ARD model excludes the use of the MRF prior of equation 14. Note that we will also use the ARD prior for automatic relevance determination of the HRF initial dip and post-stimulus undershoot.

The prior for the temporal AR parameters is then:

$$p(\alpha_{pi}|\phi_{\alpha_{pi}}) \sim N(0, 1/\phi_{\alpha_{pi}}) \tag{15}$$

The difference between this prior and the prior in equation 13 is that here the precision, $\phi_{\alpha_{pi}}$, is an unknown hyperparameter and does itself have a hyperprior on it (see next section), whereas in equation 13 the precision, ϕ_α, is fixed at a small value. If there is information in the data to support the existence of the parameter α_{pi}, then $\phi_{\alpha_{pi}} \to 0$, else $\phi_{\alpha_{pi}} \to \infty$.

We also use ARD for the spatial AR parameters β_{ij} in the spatially non-stationary model, to exclude spatial dependency between voxels when it is not relevant.

3.3.6 Precision Parameter Hyperpriors

As we are employing a fully Bayesian approach we do not assume predetermined or known values for precisions in the model. Thus far in the noise modelling, these include the noise precision ϕ_{ϵ_i}, the MRF parameter ϕ_{α_p}, the ARD precisions $\phi_{\beta_{ij}}$ and $\phi_{\alpha_{pi}}$. We use a standard conjugate Gamma hyperprior. For the noise precision we have:

$$\phi_{\epsilon_i}|\tilde{a}_\epsilon, \tilde{b}_\epsilon \sim Ga(\tilde{a}_\epsilon, \tilde{b}_\epsilon) \tag{16}$$

for the MRF precisions:

$$\begin{aligned} \phi_\beta|\tilde{a}_\beta, \tilde{b}_\beta &\sim Ga(\tilde{a}_\beta, \tilde{b}_\beta) \\ \phi_{\alpha_p}|\tilde{a}_\alpha, \tilde{b}_\alpha &\sim Ga(\tilde{a}_\alpha, \tilde{b}_\alpha) \end{aligned} \tag{17}$$

and for the ARD precisions:

$$\begin{aligned} \phi_{\beta_{ij}}|\tilde{a}_\beta, \tilde{b}_\beta &\sim Ga(\tilde{a}_\beta, \tilde{b}_\beta) \\ \phi_{\alpha_{pi}}|\tilde{a}_\alpha, \tilde{b}_\alpha &\sim Ga(\tilde{a}_\alpha, \tilde{b}_\alpha) \end{aligned} \tag{18}$$

where $Ga(a_\phi, b_\phi)$ is the Gamma distribution, and the a_ϕ and b_ϕ are known hyperparameters of the Gamma distribution. Setting $a_\phi = 0$ and $b_\phi = 0$ would be equivalent to a uniform prior $log(\sigma^2 = 1/\phi) \sim U(-\infty, +\infty)$, which would be problematic since it would result in an improper posterior with an infinite spike at $\sigma = 0$. As long as a_ϕ and b_ϕ are set positive then a proper posterior will result.

With no prior information about the variance, a_ϕ and b_ϕ the approach taken is to choose a very disperse prior, i.e. with mean based on an empirical initial estimate and a very large variance so that the choice of mean hardly affects the posterior distribution.

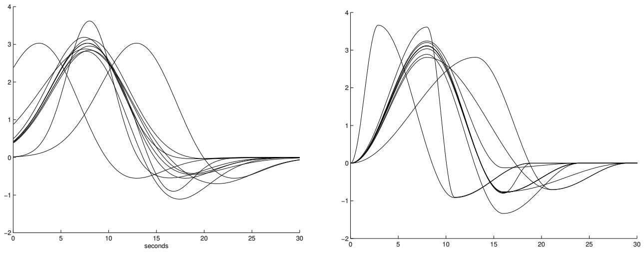
4 Signal Modelling

We now turn our attention to the modelling of r_{it} in equation 1. Typically, the experimenter has knowledge of the timing of stimuli during the experiment and it is this, along with a model of the BOLD response, that can be used in a model-based approach. The approach taken in this work is to separate out the height of the response, a_i, by the normalisation of the assumed response $s_t \otimes h(t; \boldsymbol{\lambda_i})$. We also use a linear, time-invariant model:

$$r_{it} = a_i(s_t \otimes h(t; \boldsymbol{\lambda_i}))/std(s_t \otimes h(t; \boldsymbol{\lambda_i})) \tag{19}$$

where $\otimes$ represents convolution, $h(t; \boldsymbol{\lambda_i})$ is the parameterised impulse response function, or haemodynamic response function (HRF).

The choice of models for the HRF $h(t; \boldsymbol{\lambda_i})$ and for the response size a_i is now addressed.



(a) (b)

Figure 1: 11 evenly spread samples from the prior of the HRF, using (a) two Gaussian model, and (b) the half-cosine HRF model. The prior mean HRF is plotted along with different HRFs each of which have one parameter varying at the $\pm 85^{th}$ percentile of the prior, with the other parameters held at the mean prior values.

4.1 HRF modelling

[21] and [25] have previously used a Bayesian framework to model epochs of the haemodynamic 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.

To allow modelling of BOLD responses to general stimulation types, [18] introduced the use of convolution models assuming a linear time invariant system. [5], [7], [2] and [30] 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 [15]. An additional assumption is that the stimulus represents the underlying neural activity. The stimulus (or neural activity) is then convolved with the assumed or modelled HRF to give the assumed BOLD response.

In [11] and [33] HRF models, which are allowed to vary spatially, are considered within the framework of the linear model. Straightforward attempts to allow variation in parameterised forms would be nonlinear, preventing the use of the convenient linear modelling approach. To avoid this problem, variability in the HRF is introduced via basis sets. In [9] an interesting *empirical* Bayes approach is taken to HRF modelling with basis functions, whereby the HRF (and other parameters) within a dataset are probabilistically constrained by datasets from multiple sessions and multiple subjects, by inferring on a hierarchical model which incorporates all of the datasets. Basis sets specify a subspace in which a particular HRF either lies or does not. This represents a hard constraint and often the extent of the constraint is difficult to control and/or interpret. [25] make the point that in this regard using Bayesian prior information is preferable to a basis set approach. Bayesian modelling offers the possibility of soft constraints.

In addition, unlike basis functions, a nonlinear parameterised HRF approach (that the Bayesian framework makes possible), allows interpretation of the parameters in terms of HRF shape characteristics directly. Furthermore, null hypothesis testing in a frequentist framework with basis functions, requires the overall effect for an underlying stimulus of interested to be tested for using f contrasts. These mean that the directionality of the test is lost - something which is very often of interest in FMRI experiments. For these reasons we present a Bayesian approach to linear HRF modelling for *general* stimuli using a novel parameterisation of the HRF with interpretable parameters.

Our proposed form for the HRF is based on observed BOLD responses [23]. This consists of a main response corresponding to an increase in the signal, and a dip in signal before and after the larger increase in signal, possibly reflecting a temporary imbalance between the metabolic activity and blood flow. The dip after the main response is now widely supported, whereas the existence of the early dip as a general phenomenon is still debated.

One possibility would be to use an addition of Gaussians [35]. However, there are a couple of problems evident with a Gaussian HRF model. Firstly, the HRF is not forced to be zero at time $t = 0$. Clearly, this does not reflect what we know physically. This is usually overcome using Gamma functions instead of Gaussians.

The second problem is illustrated by Figure 1(a). This shows an evenly spread 11 samples of the HRF, taken from a sensible 5-dimensional prior probability space. The problem is that there is dependence between some of the HRF characteristics. It is difficult to interpret characteristics when more than one distinct combinations of parameters can affect them. This would also be a problem with the two-parameter Gamma HRF. The clearest example of this problem is the size of the post-stimulus undershoot. It is clear that the post stimulus undershoot size could be affected by a number of different combinations of parameters. Hence, this makes any attempts to investigate undershoot difficult to perform.

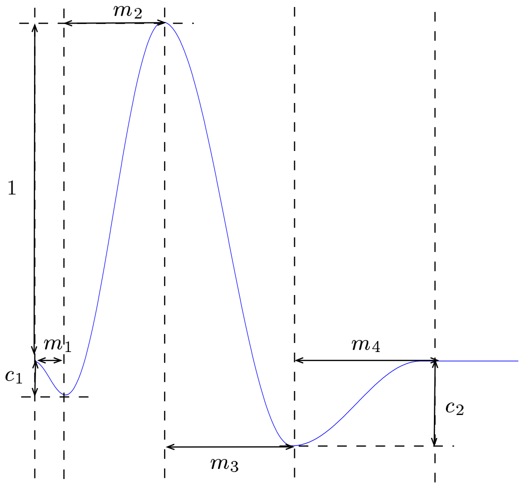


Figure 2: Parameterisation of the HRF into four half-period cosines. There are six parameters.

A solution to both of these problems is to use an alternative parameterisation of the HRF. The one we present here is simply the addition of four half-period cosines. There are six parameters; four are the periods of the four cosines, and the other two are the ratio of the height of the post-stimulus undershoot to the height of the main peak and the ratio of the height of the initial dip to the height of the main peak. Figure 2 shows a schematic of how the HRF is parameterised.

Figure 1(b) shows an evenly spread 11 samples of the HRF without an initial dip ($c_{i1} = 0$ and $m_{i1} = 0$), taken from the resulting 4 dimensional prior probability space using the half-cosine HRF model. A disadvantage with this parameterisation is that its second derivative is discontinuous. However, the range of the HRF parameters are such that sharp transitions in the second derivative are avoided. Hence, sensible looking HRF shapes predominate, as illustrated by figure 1(b).

This parameterisation does clearly impose the constraint that the HRF is zero at $t = 0$. Furthermore, parameters relating to HRF characteristics are independent. As with the Gaussian HRF, another big advantage of the half-cosine HRF model is that it could be parameterised in the frequency domain, hence speeding up the convolution.

When figure 1(b) is compared with figure 1(a), it can be seen how a characteristic of the HRF shape, such as the size of the undershoot, is now controlled by a single parameter.

HRF priors The question remains as to what priors to use on the HRF parameters. There are a number of possibilities. We could use Gaussian priors which probabilistically encode our prior belief about the expected range of shapes of the HRF. We can make the priors as tight or relaxed as we believe. In this work we choose to specify a relaxed range of shapes, using Uniform distributions over a sensible, but otherwise quite wide, range. This restricts us to sensible HRF shapes whilst giving the model full freedom to fit the HRF. This is desirable in our case as we are interested in investigating the HRF characteristics without biasing it with strong priors. The ranges used for the half cosine period parameters are:

$$
\begin{aligned}
m_{i1} &\sim U(0s, 6s) \\
m_{i2} &\sim U(2s, 14s) \\
m_{i3} &\sim U(2s, 14s) \\
m_{i4} &\sim U(2s, 14s)
\end{aligned}
\tag{20}
$$

It is not clear as to whether the data supports the existence of an initial dip or post stimulus undershoot. Hence, we would like to provide a mechanism for allowing the existence of these features to be determined automatically as part of inferring on the model. An approach to this has already been discussed in the context of autoregressive parameters in the noise model. There we use ARD priors, which can adaptively force a parameter to zero if there is no evidence to support it in the data. This is the approach we also take here for the parameters c_1 and c_2:

$$\begin{aligned} \sqrt{c_{i1}} &\sim N(0, 1/\phi_{c_{i2}}) \\ \sqrt{c_{i2}} &\sim N(0, 1/\phi_{c_{i2}}) \end{aligned} \tag{21}$$

4.2 Activation Height Modelling

Clearly, a good deal of consideration could be given to activation height modelling. Typically, within the Bayesian framework people have looked to include spatial models that incorporate the idea of finding activation (or indeed non-activation) in clusters [8, 26], and/or to model the classification of voxels or spatial areas into activation or non-activation [29]. Indeed there are no difficulties in implementing a continuous MRF on the activation height, for example. However, in this work for the purpose of *signal* modelling we choose to treat each of the voxels as independent. This is so that we can assess the information that we can extract, voxel-wise, about the shape of the HRF. We want to be able to assess how much the HRF varies between voxels in the absence of spatial regularisation of the signal.

5 Inference

The overall objective is to obtain the joint posterior distribution of all unobserved parameters in the model, given the observed data. Analytical approaches are not possible with complicated models such as those being considered in this work. Whilst it is possible to perform approximations to the distribution, it is difficult to assess the effect of these approximations on the inference performed. Therefore, the approach taken instead is to use Markov Chain Monte Carlo (MCMC) sampling from the full joint posterior distribution (see [22] and [20] for texts on MCMC).

We are able to use Gibbs sampling for all noise parameters. For the signal parameters (the activation height and HRF parameters) we use single-component Metropolis-Hastings jumps (i.e. we propose separate jumps for each of the parameters in turn). We use Normal proposal distributions with the mean fixed on the current value, and with a scale parameter σ_k for each parameter that is updated every 30 jumps. At the j^{th} update σ_k is updated using:

$$\sigma_k^{j+1} = \sigma_k^j S \frac{(1 + A + R)}{(1 + R)} \tag{22}$$

where A and R are the number of accepted and rejected jumps since the last σ_k update respectively, S is the desired rejection rate, which we fix at 0.5 [22].

When appropriate, parameters are initialised using ordinary least squares. We use a burn-in of 2000 jumps, followed by 2000 further jumps of which every 4th is sampled. Observation of the chains with different (but still sensible) initial conditions confirmed that a burn-in of 2000 samples was sufficient. The HRF parameters are initialised to the middle of their ranges. Details of the sampling used for each of the different parameters are described in the appendix.

6 Noise Model Comparison

Here we are dealing with the adequacy of the noise model in the light of observed data. Determining model adequacy requires assessment of two things: goodness of fit with the data, and penalisation of model complexity.

6.1 Deviance Information Criterion (DIC)

6.1.1 Methods

Ideally, we would use the evidence as a model adequacy measure. The evidence is the probability of getting the data given the model. However, obtaining the evidence is not analytic, and it is not easy to get an accurate estimation of the evidence using MCMC sampling. Instead of using evidence we use the Deviance Information Criterion (DIC), which tackles the issues of goodness of fit and model complexity using an approximate decision-theoretic justification (see [41]). Indeed, the DIC can be shown to be equivalent to the evidence when the deviance is Gaussian. The deviance is defined as the posterior distribution of the log likelihood:

$$D(\theta) = -2logP(y|\theta) + 2logf(y) \tag{23}$$

where $f(y)$ is a standardising term that does not affect model comparison — hence we shall deal with the first term only. The goodness of fit of the model is then summarised by the posterior expectation of the deviance:

$$\bar{D} = E_{\theta|y}[D] \tag{24}$$

and the complexity is given by the expected deviance minus the deviance evaluated at the posterior expectation:

$$\begin{aligned} p_D &= E_{\theta|y}[D] - D(E_{\theta|y}[\theta]) \\ &= \bar{D} - D(\bar{\theta}) \end{aligned} \tag{25}$$

where p_D can be interpreted as the effective number of parameters in the model. These are combined to give the overall DIC:

$$DIC = p_D + \bar{D} \tag{26}$$

where the first term represents the model complexity (the effective number of parameters) and the second term represents the goodness of fit. The attraction of using this measure is that it is trivial to compute when performing MCMC on the model. All that needs to be done is to take samples of the deviance $D(\theta)$ along with samples of θ (which will be done usually anyway) and the terms in equation 26 can be calculated to give the DIC. Note that a good model corresponds to a low DIC. The variations in the models we consider are:

Temporal

- Different temporal AR orders $P = 0, 1, 2, 3$
- Different AR priors:
  - Noninformative (NI)
  - Automatic relevance determination (ARD)
  - Spatial MRF with adaptive smoothness

Spatial

- Spatially stationary (SS) anisotropic spatial AR(1) (i.e. $\beta_{ij} = \beta_d$), compared with a spatial non-stationary (SN) anisotropic spatial AR
- Different spatial AR priors:
  - Noninformative (NI)
  - Automatic relevance determination (ARD)
  - Spatial MRF with adaptive smoothness

The data used is an audio-visual dataset using echo planar images (EPI) 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 discarded to leave $N = 176$ scans and the data was motion corrected using MCFLIRT [31] and high-pass filtered as described in [43]. The stimuli were a reversing checkerboard boxcar stimulus (30 seconds on, 30 seconds off), and an auditory boxcar stimulus (45 seconds on. 45 seconds off). The full signal (HRF) model is used in combination with each of the noise models just described. The different models were fitted to one slice of the data (slice 8 of 20).

6.1.2 Results

The DIC obtained for the different models are shown in table 1.

Table 1: Table showing the different models considered for model comparison. The columns are: P is the temporal AR order, $p(\alpha_{pi})$ indicates the prior on α_{pi}, β_{ij} indicates the choice of modelling for β_{ij}, $p(\beta_{ij})$ indicates the prior on β_{ij}, DIC is the Deviance Information Criterion (a small value indicates a better model). ARD (automatic relevance determination) or MRF (Markov Random Field) or NI (non-informative) indicates the type of prior used when a relevant parameter is being sampled from. No spatial noise modelling (NS) Spatially stationary (SS) or spatially non-stationary (SN) indicates the choice of modelling for the parameter β_{ij}.

Model	P	$p(\alpha_{pi})$	β_{ij}	$p(\beta_{ij})$	DIC
1	0	NI	NS	NI	9639
2	1	NI	NS	NI	3879
3	3	NI	NS	NI	3975
4	3	ARD	NS	NI	3312
5	3	MRF	NS	NI	2608
6	1	NI	SS	NI	3825
7	1	NI	SN	NI	2452
8	1	NI	SN	ARD	2108
9	3	MRF	SN	ARD	1665
10	3	MRF	SN	MRF	1661

Temporal It can be seen that blindly using a temporal AR(3) is worse than an AR(1) unless an ARD or MRF prior is used. There are clearly benefits in allowing the order to vary *and* in spatially regularising the temporal AR estimates, although, performing these two things in the same model is not straightforward.

Spatial The non-stationary spatial AR model is clearly superior to the stationary AR model. As with the temporal AR model the non-stationary spatial AR model is improved by using an ARD prior, or an MRF prior.

The best spatio-temporal noise model we consider is model 10. This is a temporal AR(3) with MRF prior, and an anisotropic spatially non-stationary spatial AR(1) with MRF prior. This is the noise model we use henceforth in this paper.

7 HRF Inference Evaluation

We use an artificial dataset generated from our signal model to assess the HRF fitting, and in particular to see if the ARD on the ratio parameters (e.g. post-stimulus undershoot size) works as expected.

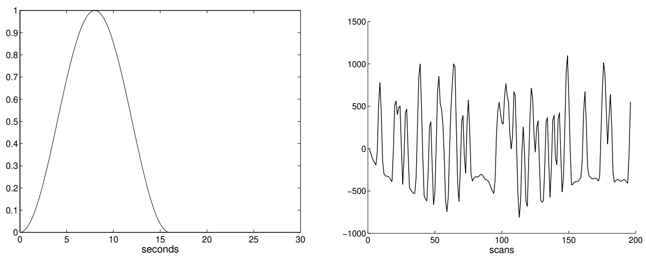
7.1 Methods

We start by generating an artificial stimulus. Inter-stimulus intervals (ISI) between single-events were drawn at random from a Poisson distribution with mean 8 seconds. ISIs were discarded if less than 4 seconds or greater than 16 seconds. Events were then randomly assigned as being stimulus or rest with equal probability.

We then generated two artificial datasets, using two different HRFs. The first HRF is shown in figure 4(a) and has a post stimulus undershoot. The second HRF, shown in figure 3(a), is the same except it has no post stimulus undershoot.

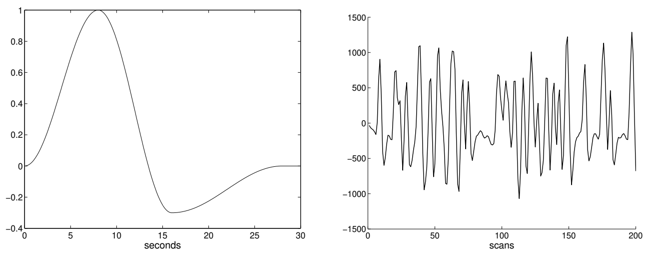
The artificial stimulus was then convolved (at 0.5 second resolution) with the two different HRFs to give the two different artificial signals shown in figures 4(b) and 3(b). The signals were then added to 100 voxels in a 10×10 square section of a slice of a null/rest dataset, to generate two artificial datasets. The null/rest dataset data was obtained with the subject performing no specific task using echo planar images (EPI) 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 discarded to leave $N = 196$ scans and the data was motion corrected using MCFLIRT [31] and high-pass filtered as described in [43].

We can then use the full model (including the best noise model) in two forms on the two datasets. One with the ARD prior on the undershoot size parameter, c_2, and one with a non-informative prior.



(a) (b)

Figure 3: (a) the HRF used to generate the artificial signal without undershoot ($m_1 = 0$, $m_2 = 8$, $m_3 = 8$, $m_4 = 0$, $c_1 = 0$, $c_2 = 0$) (b) artificial signal used in the artificial data generated from convolving the stimulus with the HRF in (a).



(a) (b)

Figure 4: (a) the HRF used to generate the artificial signal with undershoot ($m_1 = 0$, $m_2 = 8$, $m_3 = 8$, $m_4 = 12$, $c_1 = 0$, $c_2 = 0.3$), (b) artificial signal used in the artificial data generated from convolving the stimulus with the HRF in (a).

7.2 Results

Figures 5(a) and 6(a) show the fit at a typical voxel in the dataset generated with and without undershoot respectively. Figures 5(b) and 6(b) show 11 evenly spread samples from the posterior of the HRF for the same voxel.

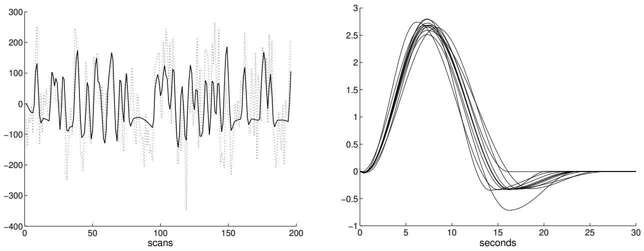
For the four combinations of two datasets (generated with and without undershoot) and different models (with and without ARD prior on the undershoot) we computed the histograms of the mean of the marginal posterior for undershoot size, c_2. We would expect the model fitted without the ARD prior to always fit a post-stimulus undershoot even for the dataset generated without the post-stimulus undershoot. However, the model with the ARD prior should force the undershoot size parameter to close to zero when using the dataset generated without the undershoot, but does fit an undershoot when using the dataset generated with an undershoot. This is exactly what can be seen in figure 7.

8 FMRI data

In this section we use the full spatio-temporal model on three different datasets with distinct stimuli.

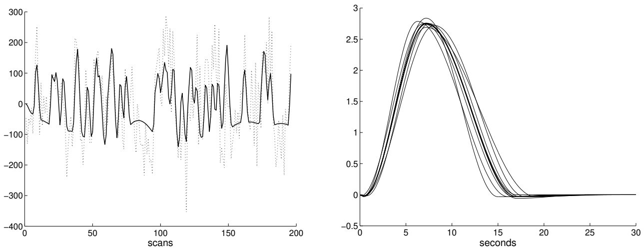
8.1 Methods

We consider three different datasets, all taken using echo planar images (EPI) acquired using a 3 Tesla system with TR=3 seconds, time to echo (TE) = 30ms, in-plane resolution 4mm and slice thickness 7mm. The datasets represent three different types of experimental design: a boxcar design, a well-spaced jittered single-events design, and a randomised single-events design. In all cases, the first 4 scans were removed and the data was motion corrected using MCFLIRT [31] and high-pass filtered as described in [43]. The data is *not* spatially smoothed. The three different datasets are as follows.



(a) (b)

Figure 5: Posterior HRF for artificial activation with undershoot. (a) Mean posterior fit (high-pass filtered data as a broken line, response fit as a solid line). (b) 11 evenly spread samples from the posterior of the HRF. The posterior mean HRF is plotted along with different HRFs each of which have one parameter varying at the $\pm 85^{th}$ percentile of the posterior, with the other parameters held at the mean posterior values.



(a) (b)

Figure 6: Posterior HRF for artificial activation without undershoot. (a) Mean posterior fit (high-pass filtered data as a broken line, response fit as a solid line). (b) 11 evenly spread samples from the posterior of the HRF. The posterior mean HRF is plotted along with different HRFs each of which have one parameter varying at the $\pm 85^{th}$ percentile of the posterior, with the other parameters held at the mean posterior values.

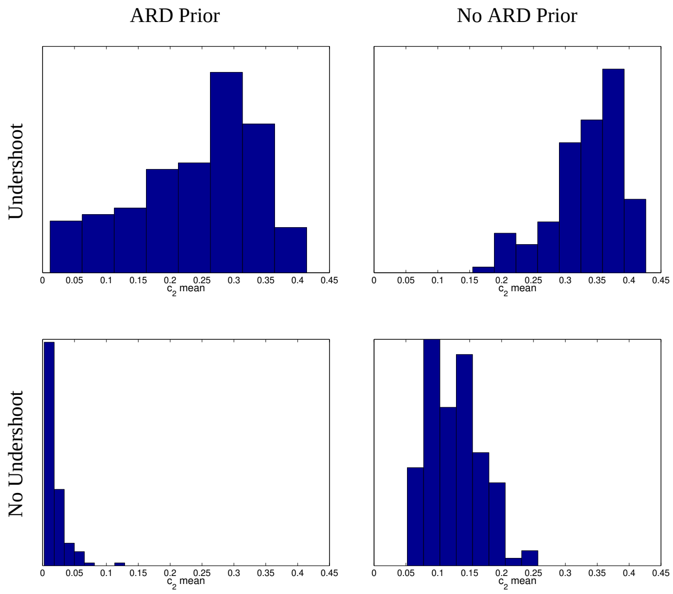


Figure 7: Histograms of the posterior mean of the HRF characteristic, c_2, corresponding to the relative size of the post-stimulus undershoot. [top] Artificial dataset generated with undershoot ($c_{i2} = 0.3$). [bottom] Artificial dataset generated without undershoot. [left] ARD prior. [right] no ARD prior. This illustrates how the ARD prior forces the undershoot to be zero when there is insufficient evidence to support it in the data. Without the ARD prior a non-zero undershoot is inferred when no undershoot actually exists. The ARD prior protects against overfitting.

Boxcar audio-visual experiment The visual stimulus was a reversing checkerboard boxcar stimulus (30 seconds on, 30 seconds off). The auditory stimulus was also a boxcar stimulus (45 seconds on. 45 seconds off).

Jittered single-event pain-audio experiment The pain stimulus was a thermal noxious stimuli of 3 seconds duration, and was administered to the dorsum of the volunteer's left hand using an electrical resistor to generate heat. This device consists of a 1.5 by 2 cm copper sheet, which delivered a very fast ramping heat stimulus (30-60 ^0C in 0.8 seconds) and simultaneously measured the skin temperature under the device. The auditory stimuli lasted for 3 seconds each. Both stimuli had varying ISI (between 30 seconds and 50 seconds).

Randomised single-event visual-visual experiment A flashing checkerboard was presented for 1.25 seconds with ISIs between single-events drawn at random from a Poisson distribution with mean 8 seconds. ISI's were discarded if less than 4 seconds or greater than 16 seconds. Half of the trials were randomly assigned to a black and white checkerboard condition and the others to a blue and yellow checkerboard condition. The black and white checkerboard response was modelled as one regressor and the colour checkerboard response was modelled as another.

8.2 Results - noise

In this section, we examine the characteristics of some of the noise parameters in the model.

The middle (top to bottom) plots in figure 8 show the first order temporal autoregressive coefficient $\bar{\alpha}_{i1}$ obtained, where $\bar{\alpha}_{i1}$ is the mean of the marginal posterior of $\alpha_i(p=1)$. The temporal autocorrelation can be seen to be spatially non-stationary. EPI slices are shown for comparison and it can be seen that there is increased temporal correlation in grey matter compared with white matter.

The bottom plots in figure 8 show the average spatial autoregressive coefficient $\sum_{j \in \mathcal{N}_i} \bar{\beta}_{ij} / \sum_{j \in \mathcal{N}_i} 1$ obtained, where $\bar{\beta}_{ij}$ is the mean of the marginal posterior of β_{ij}. The spatial correlation can also be seen to be spatially non-stationary, reconfirming the finding of the DIC that a non-stationary spatial noise model is required. It also indicates that, like the temporal autocorrelation, the spatial autocorrelation is spatially correlated with tissue type. There is also clearly increased spatial correlation in grey matter compared with white matter.

8.3 Results - signal

Figure 9 shows the activation maps for the different stimuli from the different datasets. The maps are actually the mean of the marginal posterior distribution of a_i, thresholded to only show those voxels with probability $> 99.9\%$ that $a_i > 0$.

Unsurprisingly, the most efficient stimulus type, the boxcar, produces the strongest activation, and the least efficient stimulus type, the jittered single-event, the weakest activation. Indeed, the audio stimulus of the jittered single-event dataset produces no voxels which pass the threshold used. However, due to the strength of the pain stimulation, there is a good response for that stimulus.

For the boxcar audio-visual stimulus we also performed a standard generalised least squares (GLS) analysis for comparison. The GLS analysis was performed using FSL [19]. The preprocessing was the same as for the Bayesian analysis. FSL [19] performs voxel-wise time-series statistical analysis using local autocorrelation estimation used to prewhiten the data [43]. For each of the stimuli the assumed response was modelled as a fixed Gamma HRF (with mean 6 seconds and standard deviation 3 secs) convolved with the stimulus. A temporal derivative of the assumed response was also included. The resulting z-statistic parametric maps were then thresholded at $p = 0.01$ to compare with the threshold of $p > 0.99$ for the Bayesian analysis.

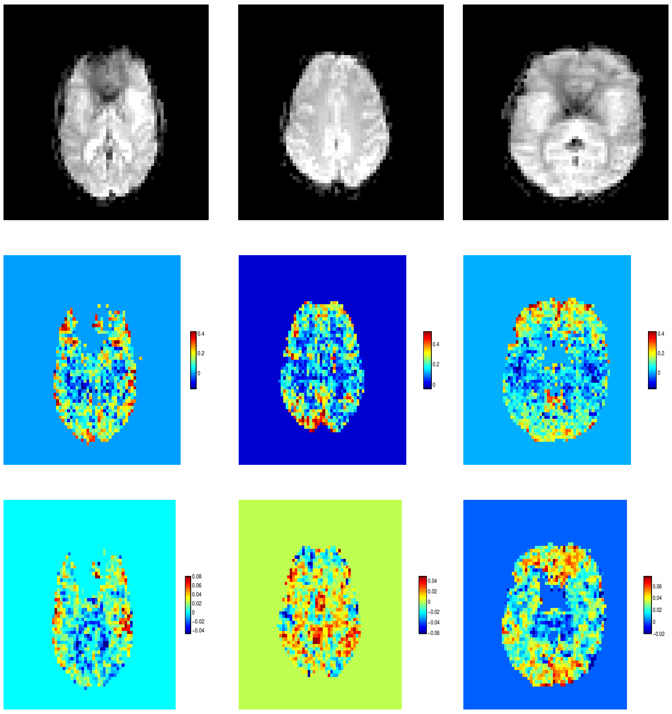


Figure 8: Autoregressive parameters from the [left] boxcar audio-visual dataset. [middle] jittered single-event pain-audio dataset. [right] randomised single-event visual-visual dataset. [top]Four EPI slices.[middle]The first order temporal autoregressive coefficient $\bar{\alpha}_{i1}$ obtained, where $\bar{\alpha}_{i1}$ is the mean of the marginal posterior of α_{i1}.[bottom]The average spatial autoregressive coefficient $\sum_{j \in \mathcal{N}_i} \bar{\beta}_{ij} / \sum_{j \in \mathcal{N}_i} 1$, where $\bar{\beta}_{ij}$ is the mean of the marginal posterior of β_{ij}.

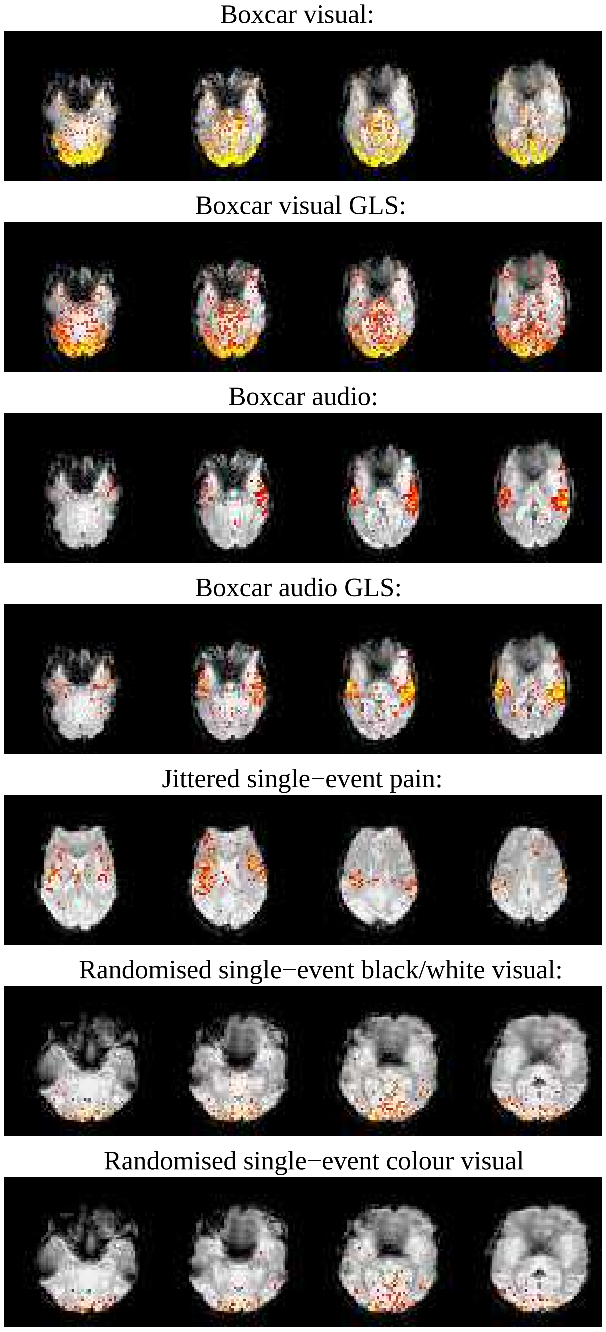


Figure 9: Maps of $\bar{a}_i$ (the mean of the posterior of a_i) for voxels with probability $> 99.9\%$ that $a_i > 0$. The z-statistics resulting from a generalised least squares analysis (thresholded at $p = 0.01$) is shown for comparison for the boxcar dataset.

As with the Bayesian analysis no multiple correction is carried out, although it is worth noting that with the Bayesian approach we do have the *joint* multivariate marginal posterior distribution for the spatial map of α to perform inference over, thereby avoiding multiple comparison issues. However, full inference of activation incorporating these issues, alongside spatial modelling of the activation height, is beyond the scope of this paper and will be addressed in future work.

Figure 10 shows the response fits at strongly activating voxels. The fit corresponds to the marginal posterior mean of activation height and HRF parameters. Figure 10 also shows evenly spread samples from the marginal posterior distribution of the HRF at the same strongly activating voxels. The shapes of the HRFs are similar between conditions of the same design type (e.g. between the visual and audio boxcar), but quite different between the design types (e.g. between boxcar and single-event). The boxcar designs have much quicker and more peaked responses. This is confirmed for all voxels passing the threshold in the histogram of mean posterior time to peak $(m_1 + m_2)$ shown in figure 11.

It is important to appreciate that in the case of the boxcar, the linearity assumption (of convolving the HRF with the stimulus to give the response) is incorrect, there will be nonlinearities present between the underlying neural activity/stimulation and the BOLD response [17]. However, for modelling simplicity it is usual to proceed with that assumption, but it should then be not surprising that the HRF looks considerably different to the single-event HRFs. Furthermore, the boxcar design has far fewer transitions than the single-event designs and hence less chances to estimate rise and fall characteristics of the HRF.

Figure 11 shows histograms of mean posterior time to peak $(m_1 + m_2)$ for all voxels passing the threshold for the different datasets. The mode of the mean posterior time $(m_1 + m_2)$ to peak is about 5 seconds for the boxcar and randomised single-event designs, and about 8 seconds for the jittered single-event design.

Figure 12 shows the scatter plots of the mean of the posterior time to peak $(m_1 + m_2)$ versus the activation height, a_i, for the voxels which are considered as activating. For the boxcar visual or boxcar auditory stimuli, there is an apparent negative correlation between these two parameters, with large activation corresponding to short delays and vice versa. For the voxels which are considered as activating under the single-event stimuli this negative correlation between these two parameters is less clear, although there is still a suggestion of some negative correlation between them particularly for the randomised single-event stimuli. We will discuss this later in the paper.

There is little evidence of a post-undershoot in figure 10, except perhaps for the randomised ISI stimuli, and there is absolutely no evidence of an initial dip for any of the stimuli. This is confirmed for all voxels passing the threshold in the histograms of the initial dip, c_1, and the post-undershoot, c_2, for all datasets shown in figure 13. The ARD prior on the initial dip, and post-stimulus undershoot will force them to zero if there is insufficient evidence for them in the data. It is important to appreciate that this does not necessarily mean that they are not actually present, just that there is insufficient evidence for them in the data when using a voxel-wise *signal* model. Of the three stimulation types, the randomised ISI design gives us the most information to estimate the HRF shape. This is because it provides us with the most transitions between rest and stimulation. Hence, it is perhaps not surprising that there is only evidence for the undershoot in this case. The ability of the randomised ISI to give us better HRF estimation is also illustrated by the tightness of the samples from the posterior HRF.

Any of the samples of the HRF posterior in figure10 can be compared with the samples from the prior HRF in figure 1(b), to show that the introduction of the data decreases the uncertainty in the HRF parameters between the prior and the posterior. This is Bayesian learning.

9 Null Data - Pseudo False Positive Rates

In this section we analyse some null/rest data using the full spatio-temporal model. The intention is to do a test that is similar to the investigation of false positive rates (FPR) in frequentist statistics.

To investigate FPRs in a frequentist framework, a statistic is calculated a number of times from data conforming to the null hypothesis, to build up an experimentally obtained null distribution. The resulting null distribution should correspond to the known theoretical distribution under the null hypothesis. In Bayesian statistics we do not work with theoretical null hypothesis distributions for the statistic of a parameter, but instead we have a probability distribution over a parameter of interest. Nevertheless, we can still use artificial null data to ask: *'If we threshold to assign voxels as "positive" when $p(a_i > 0|y) > 1 - H$, then at what rate do we produce "positives" when we know $a_i = 0$?'* Unlike frequentist statistics we do not require this rate of "false positives" to approximate H, however, it

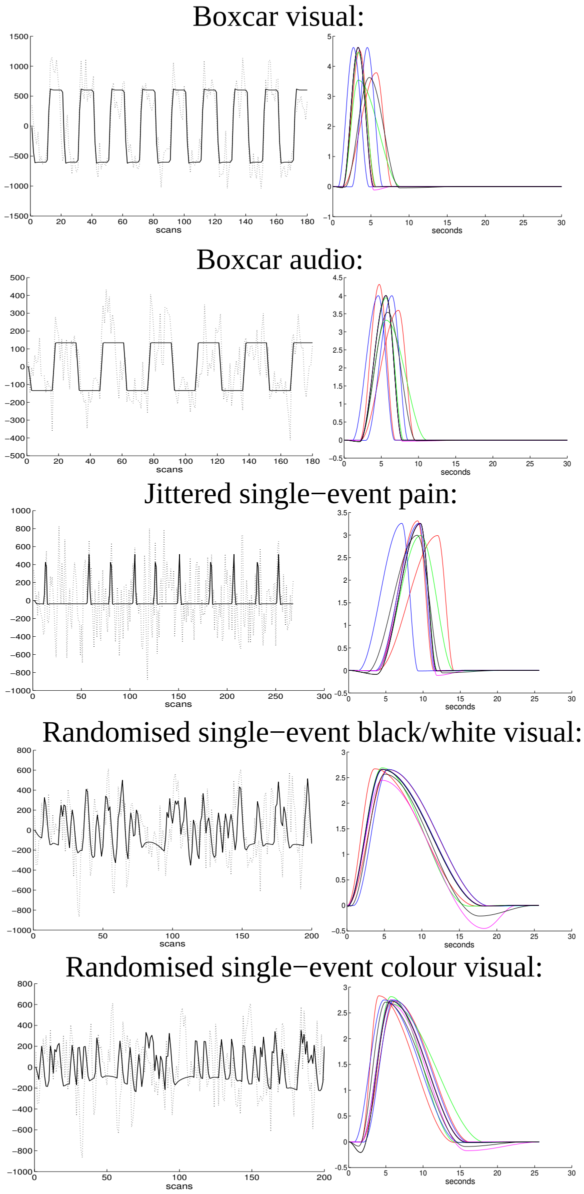


Figure 10: Posterior HRF for a strongly activating voxel in each of the datasets. [left] Mean posterior fit (high-pass filtered data as a broken line, response fit as a solid line). [right] 11 evenly spread samples from the posterior of the HRF. The posterior mean HRF is plotted along with different HRFs each of which have one parameter varying at the $\pm 85^{th}$ percentile of the posterior, with the other parameters held at the mean posterior values.

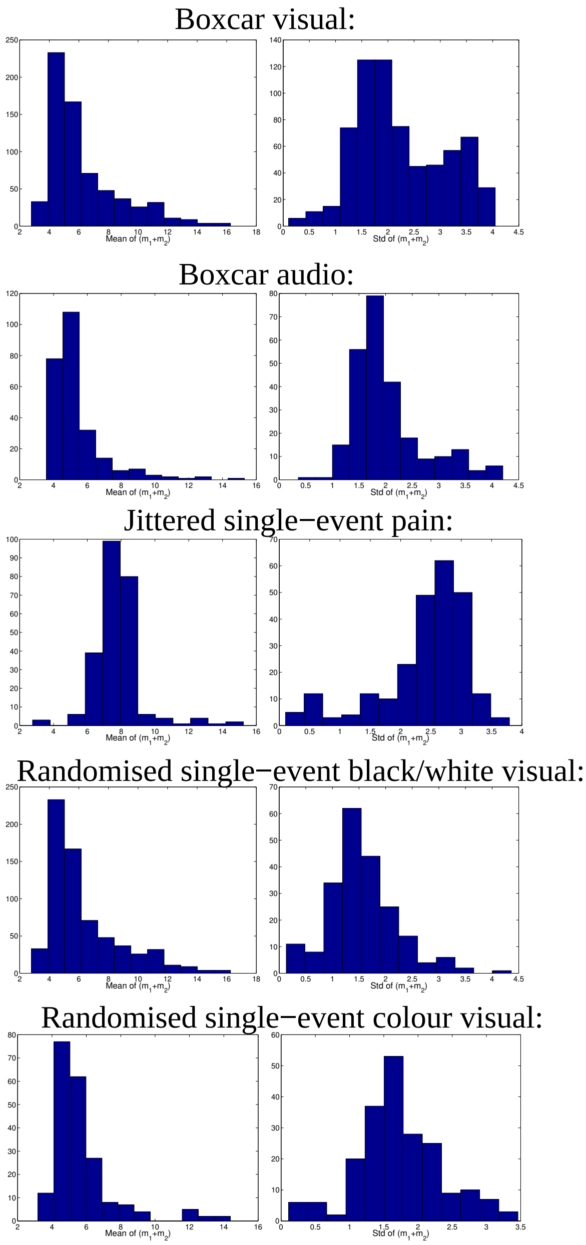


Figure 11: For the voxels which are considered as activating for each of the datasets: [left] Histogram of the posterior mean of the time to peak, $m_1 + m_2$. [right] Histogram of the posterior standard deviation of the time to peak, $m_1 + m_2$.

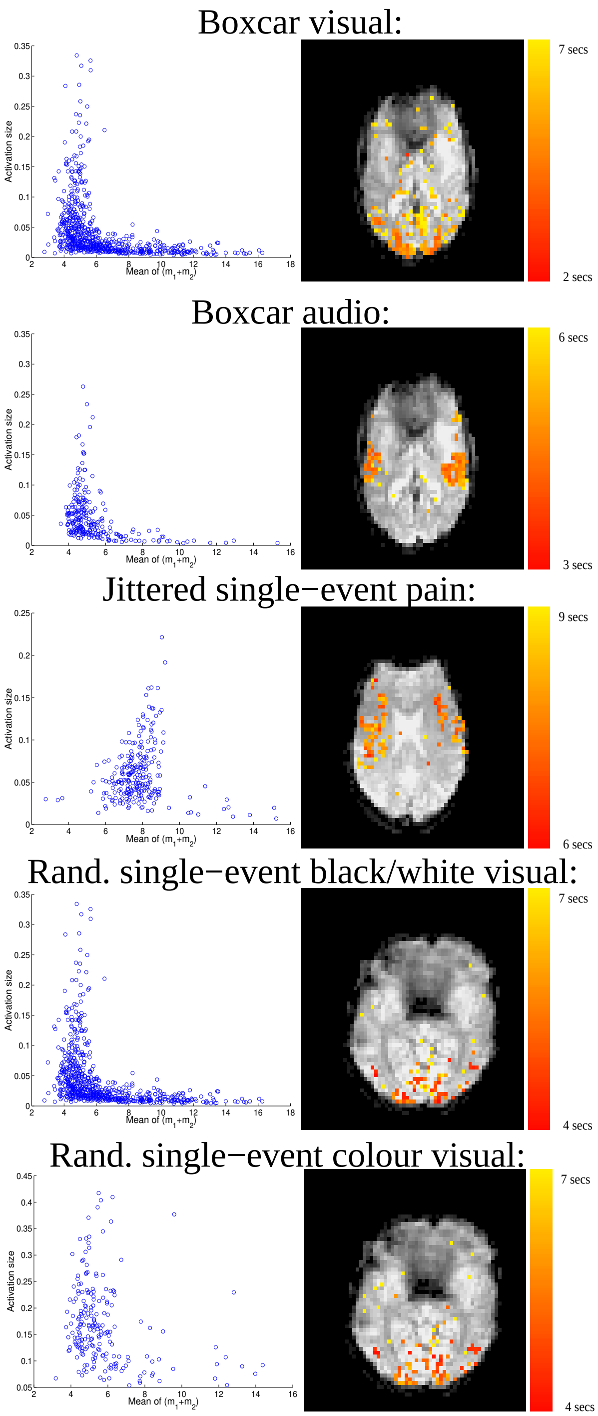
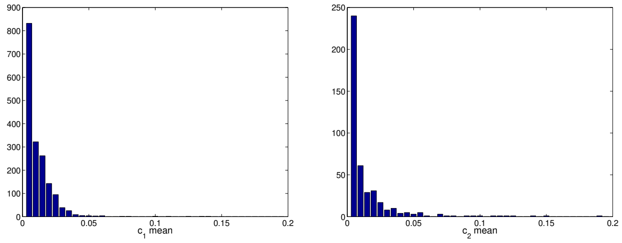
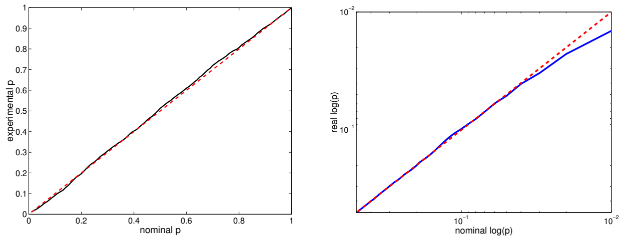


Figure 12: For the voxels which are considered as activating for each of the datasets: [left] Scatter plot of the posterior mean of the time to peak, $m_1 + m_2$, versus the posterior mean activation height, a_i [right] Spatial map of the time to peak, $m_1 + m_2$.



(a) (b)

Figure 13: (a) Histogram of the posterior mean of the HRF characteristic (a) c_1 (initial dip), and (b) c_2 (post-stimulus undershoot) for the voxels which are considered as activating from all datasets.



(a) (b)

Figure 14: (a) Probability-probability (p-p) plot and (b) Log probability-log probability plot, comparing nominal frequentist "pseudo-FPR", H, against the experimentally obtained proportion of voxels with $p(a_i > 0|y) > 1 - H$, obtained from a null/rest dataset for the random ISI single-event design using the full model.

is still interesting to investigate its properties.

9.1 Methods

The dataset used is the null/rest dataset described in section 7, although, here we do not add artificial activation to the data. We also need to decide on an arbitrary assumed response for modelling the signal, for this we use the random ISI design also used in section 7.

We can then use the full model (including the best noise model and full HRF modelling with ARD on the undershoots) on the dataset. For each voxel we obtain the marginal posterior probability over the activation height, $p(a_i|y)$. We can then compute the probability, $p(a_i > 0|y)$, for each voxel. We then compute the number of "positive" voxels N_H with $p(a_i > 0|y) > 1 - H$, and plot N_H/N against H. This is repeated for a range of H.

9.2 Results

The results are shown as a p-p plot in figure 14. There is good correspondence between the nominal frequentist rate and the rate obtained experimentally using the fully Bayesian spatio-temporal inference. This is because we have a non-informative prior on the activation height parameter, a_i.

10 Conclusions and Discussion

In general, the use of a fully Bayesian approach is a powerful way of considering more reasonable, and often more complex, models whilst guarding against over-fitting and giving correct inference on parameters in the model. We can consider model selection techniques to tune the modelling used and/or we can use techniques such as ARD to adaptively determine the evidence for parameters in the model. ARD is a neat trick to avoid the computational complexities of reversible jumps [27], or similar techniques. This allows us to really explore whether or not there is evidence in the data for the presence of a particular parameter, rather than assuming that there is, and consequently over-fitting and unnecessarily increasing the uncertainty in parameters of interest.

The downside of this approach is that inferring on the models is not analytical and we are required to use techniques such as MCMC. These techniques are time consuming for the large datasets which are encountered in FMRI. On a 2GHz Intel PC the technique takes approximately 6 hours on a single slice of FMRI data. Whilst this is not an obstacle to exploring modelling issues as addressed in this paper, it is realistically an obstacle to using such techniques for "everyday" analysis of FMRI data.

An alternative is to assume approximations to the posterior such as those offered by the framework of Variational Bayes [32]. For example, [38] use Variational Bayes with a multivariate autoregressive temporal model. However, the most common form of Variational Bayes requires conjugate priors and is hence only tractable in the same situations as when Gibbs sampling can be used. For example, in the model used in this paper this would mean that the HRF parameters would be intractable, limiting the choice of HRF modelling to those which would be tractable (e.g. basis functions).

We now discuss some of the issues in noise and signal modelling separately.

10.1 Noise

The noise model consists of a space-time simultaneously specified autoregressive model. We used a model comparison technique in the form of DIC, which balanced model complexity with goodness of fit, to deduce which was the best model out of the ones we considered. This turned out to be a spatially non-stationary, but temporally stationary temporal AR(3) model with a within-matter-type edge-preserving MRF prior on the temporal AR coefficients, combined with a temporally stationary, but spatially non-stationary spatial AR(1) model with a within-matter-type edge-preserving MRF prior on the spatial AR coefficients. We observed matter-type dependence of the spatial and temporal autocorrelation of the noise on three different real FMRI datasets.

It was clear from the DIC model comparisons that use of MRF prior and ARD prior were both beneficial. However, it is currently not clear how to implement both within the same model. Further improvement might include attempting to model the large scale temporal variations l_{it} as in [21, 26]. In this work we remove the worst of these by high-pass filtering as a preprocessing step. If this could be sensibly incorporated into the model then we could be less conservative with the effective high-pass cut-off point and the uncertainty associated with estimating these large scale variations could be taken into account.

10.2 Signal

Previously, parameterised HRFs limited to epochs of boxcar designs have been modelled in a Bayesian framework using MCMC [21, 25]. In this work we introduced a novel half-cosine parameterisation of the HRF, and implemented it in a framework allowing for general stimulation types (boxcar, single-event). We imposed no spatial regularisation of the *signal* to allow an investigation of what can be inferred at each voxel. Whilst the HRF signal model is voxel-wise, it is worth emphasising that the noise model used at the same time is fully spatio-temporal.

The use of the half-cosine parameterised form produces easily interpretable parameters, which is useful for the specification of priors and for interpreting the results. The parameters which represent the size of the initial dip and post undershoot crucially had an ARD prior. An ARD prior will force to zero those parameters that are not supported by the model and the data. This allows us to identify whether or not the data supports the existence of these HRF features on a voxel-wise basis.

One of the results on the HRF characterisation suggested that there is a negative correlation between activation height and HRF time to peak. The idea that activation height is negatively correlated with the HRF time to peak, was also found in [25] for boxcar designs only. However, we need to be careful. The apparent causality between activation height and time to peak is just as likely to be indirect, and merely reflect that for voxels with low activation heights, the uncertainty in the time to peak is larger and hence we get a spread of estimated posterior mean time to peaks around the true value. This is demonstrated in figure 15 with the marginal posterior distribution for a strongly activating voxel having a much tighter distribution than the weakly activating voxel (both from the visual boxcar stimulus). There is another possibility. We may have voxels passing the threshold which are not pure responses to the stimuli. This "confound activation" may be structured noise partially correlated with the assumed response by chance, or response/stimulus related confounds such as motion artefact. These "confound activations" will have apparent HRF shapes spread across a wide range. These could have been avoided by being more restrictive with the HRF shape, however, without knowing the exact shape of the true HRF response a priori, we might then have missed some of the true response to the stimulus. Figure 16 is a schematic suggesting how the

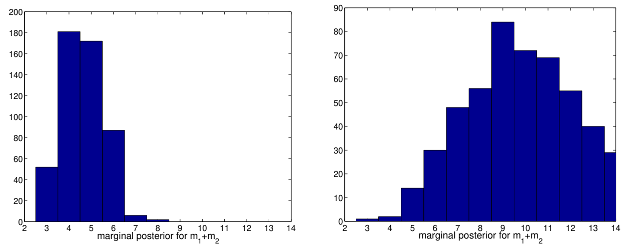


Figure 15: Marginal posterior distribution of the time to peak, $m_1 + m_2$, for the visual boxcar stimulation for (a) a voxel with large activation, (b) a voxel with small activation.

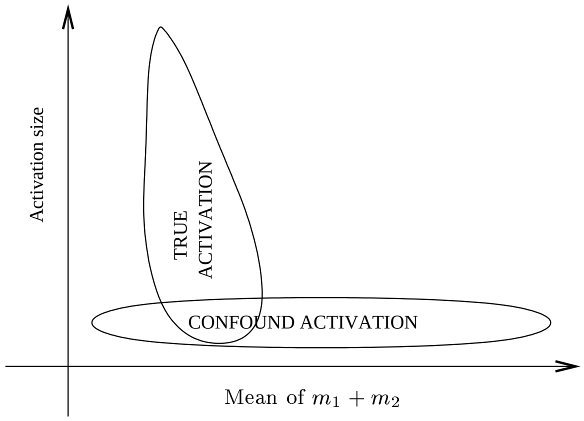


Figure 16: Schematic suggesting how the activation height—time to peak scatter plot maybe made up three different effects (1) The "true activation" is negatively correlated. (2) The uncertainty in the "true activation" increases with smaller activation height. (3) There are voxels with "confound activation" passing the threshold. This may be structured noise randomly correlated with the assumed response, or response/stimulus related confounds such as motion artefact — these "confound activations" will have apparent HRF shapes spread across a wide range. There is no clear way with the current model to distinguish between these effects.

activation height—time to peak $(m_1 + m_2)$ scatter plot maybe made up of all three of these effects. There is no clear way with the current model to distinguish between these effects.

11 Acknowledgements

The authors would like to acknowledge support from the UK MRC and EPSRC. Thanks to Tim Behrens and Christian Beckmann for their input. Thanks to Richard Rogers and Debbie Painter for the pain-audio dataset and to Heidi Johansen-Berg for the randomised single-event dataset.

12 appendix

12.1 Gamma Distribution

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

$$f_{Ga}(x;a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} \tag{27}$$

where $\Gamma(a)$ is the Gamma function. A χ^2 distribution with ν degrees of freedom corresponds to the distribution $Ga(\nu/2, 1/2)$. The b parameter is a scale parameter. The one-parameter gamma distribution corresponds to $Ga(a,1)$. A sample from $Ga(a,b)$ can be obtained by taking a sample from $Ga(a,1)$ and dividing it by b. Note, that a gamma distribution has $mean = a/b$ and $variance = a/b^2$

12.2 Noise Model MCMC Sampling

All parameters in the noise model have full conditionals which can be sampled from. Hence for all noise parameters we employ Gibbs sampling. See [22] or [20] for an introduction to Gibbs sampling.

In deriving some of the full conditional distributions, expressions of the following type are often obtained:

$$p(\theta|A,B) \propto \exp\{-\frac{1}{2}[\theta^2 A - 2\theta B]\} \tag{28}$$

where θ is the parameter whose full conditional we are deriving and A and B are functions of the other parameters in the model. This can be expressed as a Normal distribution by completing the square to give:

$$\theta|A,B \sim N(B/A, 1/A) \tag{29}$$

The full conditions for the noise parameters are as follows.

12.2.1 ϕ_{ϵ_i}

$$\begin{aligned}\phi_{\epsilon_i}|. \quad \sim \quad & Ga\left(\frac{T}{2} + \tilde{a}_\epsilon, \frac{1}{2}\left[\sum_{t=1}^{T}\left(q_{it} - \left(\sum_{j\in\mathcal{N}_i} \beta_{ij} q_{j(t-1)}\right.\right.\right.\right.\\ & \left.\left.\left.\left. + \sum_{p=1}^{P} \alpha_{ip} q_{i(t-p)}\right)\right)^2 + 2\tilde{b}_\epsilon\right]\right)\end{aligned} \tag{30}$$

12.2.2 β_d (Spatially stationary spatial model)

$$\beta_d|. \sim N(B_{\beta_d}/A_{\beta_d}, 1/A_{\beta_d}) \tag{31}$$

where

$$A_{\beta_d} = \sum_{it} \phi_{\epsilon_i}\left(\sum_{j\in\mathcal{N}_{id}} q_{j(t-1)}\right)^2 + \frac{1}{\sigma_\beta^2} \tag{32}$$

$$\begin{aligned}B_{\beta_d} = & \sum_{it} \phi_{\epsilon_i}\left[\left(q_{it} - \sum_{p=1}^{P} \alpha_{ip} q_{i(t-p)}\right.\right.\\ & \left.\left. - \sum_{d'\neq d} \beta_{d'} \sum_{j\in\mathcal{N}_{id'}} q_{j(t-1)}\right) \sum_{j\in\mathcal{N}_{id}} q_{j(t-1)}\right]\end{aligned} \tag{33}$$

12.2.3 β_{ij} (Spatially non-stationary spatial model) with ARD prior

$$\beta_{ij}|. \sim N(B_{\beta_{ij}}/A_{\beta_{ij}}, 1/A_{\beta_{ij}}) \tag{34}$$

where:

$$\begin{aligned}
A_{\beta_{ij}} &= \sum_t (\phi_{\epsilon_i} q^2_{j(t-1)} + \phi_{\epsilon_j} q^2_{i(t-1)}) + \phi_{\beta_{ij}} \\
B_{\beta_{ij}} &= \sum_t \left[\phi_{\epsilon_i} q_{j(t-1)} \left(q_{it} - \sum_{p=1}^{P} \alpha_{ip} q_{i(t-p)} - \right. \right. \\
&\quad \left. \sum_{k \in \mathcal{N}_i : k \neq j} \beta_{ik} q_{k(t-1)} \right) + \phi_{\epsilon_j} q_{i(t-1)} \\
&\quad \left(q_{jt} - \sum_{p=1}^{P} \alpha_{jp} q_{j(t-p)} \right. \\
&\quad \left. \left. - \sum_{k \in \mathcal{N}_j : k \neq i} \beta_{jk} q_{k(t-1)} \right) \right]
\end{aligned} \tag{35}$$

12.2.4 $\phi_{\beta_{ij}}$ for ARD prior

$$\phi_{\beta_{ij}}|. \sim Ga\left(\frac{1}{2} + \tilde{a}_\beta, \frac{\beta_{ij}^2}{2} + \tilde{b}_\beta\right) \tag{36}$$

12.2.5 β_{ij} (Spatially non-stationary spatial model) with MRF prior

$$\beta_{ij}|. \sim N(B_{\beta_{ij}}/A_{\beta_{ij}}, 1/A_{\beta_{ij}}) \tag{37}$$

where:

$$\begin{aligned}
A_{\beta_{ij}} &= \sum_t (\phi_{\epsilon_i} q^2_{j(t-1)} + \phi_{\epsilon_j} q^2_{i(t-1)}) + \phi_\beta \sum_{L_{k\ell} \in \mathcal{N}_{L_{ij}}} 1 \\
B_{\beta_{ij}} &= \sum_t \left[\phi_{\epsilon_i} q_{j(t-1)} \left(q_{it} - \sum_{p=1}^{P} \alpha_{ip} q_{i(t-p)} \right. \right. \\
&\quad \left. - \sum_{k \in \mathcal{N}_i : k \neq j} \beta_{ik} q_{k(t-1)} \right) \\
&\quad + \phi_{\epsilon_j} q_{i(t-1)} \left(q_{jt} - \sum_{p=1}^{P} \alpha_{jp} q_{j(t-p)} \right. \\
&\quad \left. \left. - \sum_{k \in \mathcal{N}_j : k \neq i} \beta_{jk} q_{k(t-1)} \right) \right] \\
&\quad + \phi_\beta \sum_{L_{k\ell} \in \mathcal{N}_{L_{ij}}} \beta_{k\ell}
\end{aligned} \tag{38}$$

where L_{ij} is the link between voxel i and voxel j.

12.2.6 ϕ_β for MRF prior

$$\begin{aligned}\phi_\beta|. \quad \sim \quad & G\left(\frac{N_L}{2} + \tilde{a}_\beta,\right. \\ & \left.\frac{1}{4}\sum_{L_{ij}}\sum_{L_{k\ell}\in\mathcal{N}_{L_{ij}}}\left((\beta_{ij}-\beta_{k\ell})^2\right) + \tilde{b}_\beta\right) \end{aligned} \tag{39}$$

where N_L is the total number of links between voxels.

12.2.7 α_{pi} with ARD prior

$$\alpha_{pi}|. \sim N(B_{\alpha_{pi}}/A_{\alpha_{pi}}, 1/A_{\alpha_{pi}}). \tag{40}$$

where

$$A_{\alpha_{pi}} = \phi_{\epsilon_i}\sum_{t=1}^{T} q_{i(t-p)}^2 + \phi_{\alpha_p} \tag{41}$$

$$\begin{aligned} B_{\alpha_{pi}} = \phi_{\epsilon_i}\sum_{t=1}^{T}&\left[\left(q_{it} - \sum_{d=1}^{D}\left(\sum_{j\in\mathcal{N}_i}\beta_{ij}q_{j(t-1)}\right)\right.\right. \\ & \left.\left. - \sum_{p'\neq p}\alpha_{p'i}q_{i(t-p')}\right) q_{i(t-p)}\right] \end{aligned} \tag{42}$$

12.2.8 $\phi_{\alpha_{pi}}$ for ARD prior

$$\phi_{\alpha_{pi}}|. \quad \sim \quad Ga\left(\frac{1}{2} + \tilde{a}_\alpha, \frac{\alpha_{pi}^2}{2} + \tilde{b}_\alpha\right) \tag{43}$$

12.2.9 α_{pi} with MRF prior

$$\alpha_{pi}|. \sim N(B_{\alpha_{pi}}/A_{\alpha_{pi}}, 1/A_{\alpha_{pi}}) \tag{44}$$

where

$$A_{\alpha_{pi}} = \phi_{\epsilon_i}\sum_{t=1}^{T} q_{i(t-p)}^2 + \phi_{\alpha_p}\sum_{j\in\mathcal{N}_i} 1 \tag{45}$$

$$\begin{aligned} B_{\alpha_{pi}} = \phi_{\epsilon_i}\sum_{t=1}^{T}&\left[\left(q_{it} - \sum_{d=1}^{D}\left(\sum_{j\in\mathcal{N}_i}\beta_{ij}q_{j(t-1)}\right)\right.\right. \\ & \left.\left. - \sum_{p'\neq p}\alpha_{p'i}q_{i(t-p')}\right) q_{i(t-p)}\right] \\ & + \phi_{\alpha_p}\sum_{j\in\mathcal{N}_i}\alpha_{pj} \end{aligned} \tag{46}$$

12.2.10 ϕ_{α_p} for MRF prior

$$\begin{aligned}\phi_\alpha|. \quad \sim \quad & G\left(\frac{N}{2} + \tilde{a}_\alpha, \frac{1}{4}\sum_{i}\sum_{j\in\mathcal{N}_i}\left((\alpha_{pi}-\alpha_{pj})^2\right)\right. \\ & \left. + \tilde{b}_\alpha\right) \end{aligned} \tag{47}$$

12.3 Signal Model MCMC Sampling

12.3.1 λ_{hi}

This is sampled from using Metropolis-Hastings. We propose univariate jumps in the six parameter space of λ_{hi} at each voxel. Metropolis-Hastings requires that the following terms are updated:

$$\prod_t p(q_{it}|\beta, \alpha, q_{i\tilde{t}<t}, \phi_{\epsilon_i}) \tag{48}$$

$$\prod_{t,j\in\mathcal{N}i} p(q_{jt}|\beta, \alpha, q_{i\tilde{t}<t}, \phi_{\epsilon_j}) \tag{49}$$

$$p(\lambda_{hi}) \tag{50}$$

12.3.2 a_i

This is sampled from using Metropolis-Hastings. Although, note that we could have used Gibbs sampling. Metropolis-Hastings requires that the following terms are updated:

$$\prod_t p(q_{it}|\beta, \alpha, q_{i\tilde{t}<t}, \phi_{\epsilon_i}) \tag{51}$$

$$\prod_{t,j\in\mathcal{N}i} p(q_{jt}|\beta, \alpha, q_{i\tilde{t}<t}, \phi_{\epsilon_j}) \tag{52}$$

References

[1] H. Benali, I. Buvat, J.L. Anton, M. Pélégrini, M. Di Paola, J. Bittoun, Y. Burnod, and R. Di Paola. Space-time statistical model for functional MRI image sequences. In J.S. Duncan and G.R. Gindi, editors, *Information Processing in Medical Imaging*, pages 285–298. Kluwer Academic Publishers, 1997.

[2] G.M. Boynton, S.A. Engel, G.H. Glover, and D.J. Heeger. Linear systems analysis of functional magnetic resonance imaging in human V1. *Journal of Neuroscience*, 16:4207–4221, 1996.

[3] E. Bullmore, M. Brammer, S.C.R. 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.

[4] R. Christensen. *Plane Answers to Complex Questions*. Springer, 1996.

[5] M.S. Cohen. Parametric analysis of fMRI data using linear systems methods. *NeuroImage*, 6:93–103, 1997.

[6] N.A.C. Cressie. *Statistics for Spatial Data*. Wiley, New York, 1993.

[7] A.M. Dale and R.L. Buckner. Selective averaging of rapidly presented individual trials using fMRI. *Human Brain Mapping*, 5:329–340, 1997.

[8] X. Descombes, F. Kruggel, and D. Y. von Cramon. Spatio-temporal fMRI analysis using Markov random fields. *IEEE Trans. on Medical Imaging*, 17(6):1028–39, December 1998.

[9] K. J. Friston, D. E. Glaser, R. N. A. Henson, S. Kiebel, C. Phillips, and J. Ashburner. Classical and Bayesian inference in neuroimaging: Applications. *NeuroImage*, 16:484–512, 2002.

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

[11] K.J. Friston, C.D. Frith, R. Turner, and R.S.J. Frackowiak. Characterizing evoked hemodynamics with fMRI. *NeuroImage*, 2:157–165, 1995.

[12] K.J. Friston, A.P. Holmes, J.-B. Poline, P.J. Grasby, S.C.R Williams, R.S.J. Frackowiak, and R. Turner. Analysis of fMRI time series revisited. *NeuroImage*, 2:45–53, 1995.

[13] K.J. Friston, A.P. Holmes, K.J. Worsley, J.-B. Poline, C.D. Frith, and R.S.J. Frackowiak. Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, 2:189–210, 1995.

[14] K.J. Friston, P. Jezzard, and R. Turner. Analysis of Functional MRI time-series. *Human Brain Mapping*, 1:153–171, 1994.

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

[16] K.J. Friston, O. Josephs, E. Zarahn, A.P. Holmes, S. Rouquette, and J-B. Poline. To smooth or not to smooth? *NeuroImage*, 12:196–208, 2000.

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

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

[19] FSL. `http://www.fmrib.ox.ac.uk/fsl`.

[20] D. Gamerman. *Markov Chain Monte Carlo*. Chapman and Hall, London, 1997.

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

[22] W.R. Gilks, S. Richardson, and D.J. Spiegalhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London, 1996.

[23] G.H. Glover. Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage*, 9:416–429, 1999.

[24] C. Gössl, D.P. Auer, and L. Fahrmeir. Dynamic models in fMRI. *Magnetic Resonance in Medicine*, 43:72–81, 2000.

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

[26] C. Gössl, D.P. Auer, and L. Fahrmeir. Bayesian spatio-temporal inference in functional magnetic resonance imaging. *Biometrika*, 2001. Accepted.

[27] P.J. Green. Reversible jump Markov Chain Monte Carlo computation and bayesian model determination. *Biometrika*, 82:711–732, 1995.

[28] N.V. Hartvig. A stochastic geometry model for fMRI data. Technical Report 410, Department of Theoretical Statistics, University of Aarhus, 2000.

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

[30] J. Hykin, R. Bowtell, P. Glover, R. Coxon, L.D. Blumhardt, and P. Mansfield. Investigation of the linearity of functional activation signal changes in the brain using echo planar imaging (EPI) at 3.0 T. In *Proc. of the SMR and ESMRB, Joint Meeting*, page 795, 1995.

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

[32] M.I. Jordan. *Learning in Graphical Models*. MIT Press, 1999.

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

[34] V. Kiviniemi, J.-H. Kantola, J. Jauhiainen, A. Hyvärinen, and O. Tervonen. Independent component analysis of nondetermistic fMRI signal sources. *NeuroImage*, 2003.

[35] F. Kruggel and D.Y. von Cramen. Modeling the hemodynamic response in single-trial functional MRI experiments. *Magnetic Resonance in Medicine*, 42:787–797, 1999.

[36] D.J.C. MacKay. Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6:469–505, 1995.

[37] A. Mollie. *Bayesian mapping of disease*, chapter 20. In [22], 1996.

[38] W.D. Penny and S.J. Roberts. Bayesian multivariate autoregressive models with structured priors. *IEE Proceedings on Vision, Image and Signal Processing.*, 149:33–41, 2002.

[39] D.B. Phillips and A.F.M. Smith. *Bayesian model comparison via jump diffusions*, chapter 13. In [22], 1996.

[40] G.A.F. Seber. *Linear Regression Analysis*. Wiley, 1977.

[41] D.J. Spiegalhalter, N.G. Best, B.P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society*, 64(3):134, 2002.

[42] C.K. Wikle, L.M. Berliner, and N. Cressie. Hierarchical bayesian space-time models. *Environmental and Ecological Statistics*, 5:117–154, 1998.

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