Abstract

Functional magnetic resonance imaging (fMRI) has recently become a popular tool for studying human brain activity. Despite fMRI’s widespread use, most existing statistical methods for analyzing fMRI data are problematic. Many methodologies oversimplify the problem for the sake of computational efficiency, often not providing a full statistical model as a result. Other methods are too computationally inefficient to use on large data sets. In this paper, we propose a Bayesian method for analyzing fMRI data that is computationally efficient and provides a full statistical model.

*Research supported by the National Institutes of Health and the National Science Foundation.
1 Introduction

We consider functional magnetic resonance imaging (fMRI) experiments that consist of presenting tasks to a subject in order to detect localized neuronal responses. The goal is to develop imaging biomarkers that describe the nature of the neuronal response. fMRI experiments produce large amounts of noisy correlated data, for which it is difficult to specify computationally tractable statistical models that reflect the nature of the experiment.

Before an fMRI experiment, the image space is partitioned into a rectangular lattice comprising small cuboid volume elements, or voxels, of equal size. Partition sizes can be as large as 500,000 voxels, with the voxel size chosen to balance resolution and scanning time requirements. Neuronal activation occurs quickly and is not observed directly. However, neuronal activation places metabolic demands that result in increased blood flow and an excess of oxygenated hemoglobin. The magnetic properties of oxygen are then exploited, allowing measurement of the so-called blood oxygenation level dependent (BOLD) signal, which is used as a proxy for neuronal activation. It is common to observe BOLD signal intensities for each voxel at several hundred time points.

The relationship between the BOLD signal and the stimulus is complicated. Images are collected every 2–3 seconds, but the BOLD signal increases above baseline two seconds after the start of neuronal activity, peaking after about 5–8 seconds and falling below baseline for roughly ten seconds. The repeated measurements on each voxel in the presence of low-frequency noise suggest within-voxel temporal dependence. In addition, voxels close together behave similarly, and activation tends to come in groups of voxels. Hence the observations are spatially dependent. A further complication is that fMRI data tends to be noisy, and the BOLD signal is usually relatively small (approximately 1% to 5% of baseline). Thus a typical fMRI experiment results in a large amount of noisy data that exhibits complicated spatiotemporal dependence.

Unfortunately, including spatiotemporal dependence in a model typically leads to intractable computation. Hence compromises are made on the modeling side to make the required computing feasible. For example, many standard procedures account for temporal dependence while ignoring spatial dependence, at least in the modeling stage, (Friston et al., 1994; Lindquist, 2008; Worsley et al., 1992, 2002; Worsley, 2003) or account for spatial dependence while ignoring temporal depen-
dence (Genovese, 2000; Smith et al., 2003; Smith and Fahrmeir, 2007). An alternative approach is to specify a full model, but then use a computationally tractable approximation to it. For example, Penny et al. (2003) proposed a Bayesian model that focuses on estimation of the error process of the BOLD signal, and use a variational Bayes approach to facilitate fast posterior sampling.

One of our goals is the development of a Bayesian model that incorporates both spatial and temporal dependence while offering a natural way to directly incorporate the physical characteristics of the experiment and prior anatomical information. However, proceeding naively down the Bayesian path can easily result in posterior distributions that cannot be used even with the most sophisticated Markov chain Monte Carlo (MCMC) algorithms. Thus we are still in a situation where computational considerations must drive some of the modeling decisions.

We propose a Bayesian spatiotemporal model for detecting activation patterns in fMRI data. The method is computationally efficient and the results are easy to interpret.

The remainder of this paper is organized as follows. Our Bayesian model is introduced in Section 2. Following Smith and Fahrmeir (2007) and Lee et al. (2014), we use binary indicator variables for classifying active voxels, but assume that the spatial dependence in the images is governed by an underlying areal model, i.e., a model for spatially aggregated data. This is done by parcellating the image into clusters of voxels and using spatial random effects to capture spatial dependence. A spatial hierarchical prior is used to model the spatial dependence. We show that our proposed methodology allows prior anatomical information to be incorporated in an intuitive manner. In Section 3 we develop an efficient algorithm for fitting the model. In Section 4 we show that our model and algorithm work well on simulated data. In Section 5 we describe a novel benchmark analysis. Two fMRI examples are considered in Section 6, and some final comments are given in Section 7.

2 A Hierarchical Spatiotemporal Model

Let the image intensity at voxel \( v \in \{1, \ldots, N\} \) and time \( t \in \{1, \ldots, T_v\} \) be denoted by \( y_{v,t} \). The physical characteristics of the brain and the BOLD signal suggest that there will be spatial and temporal dependence in the measurements. Spatial dependence occurs because voxels close together
tend to exhibit similar intensities, and activation comes in groups of voxels. We will develop a model that accounts for both sources of dependence under an assumption of space-time separability, i.e., no interaction between space and time.

Set \( y_v = (y_{v,1}, \ldots, y_{v,T_v}) \), and suppose there are \( p \) distinct experimental tasks or stimuli of interest, that \( X_v \) is a \( T_v \times p \) design matrix of full column rank (the structure of \( X_v \) will be discussed below), and \( \beta_v \) is a \( p \times 1 \) vector of regression coefficients. If \( \Lambda_v \) is a \( T_v \times T_v \) matrix having \( (i,j) \)th element \( \rho_v^{i-j} \), and \( \varepsilon_v \sim N_{T_v}(0, \sigma_v^2 \Lambda_v) \), we model the voxel-wise time series as

\[
y_v = X_v \beta_v + \varepsilon_v. \tag{1}
\]

Thus the temporal correlation within each voxel is accounted for by our choice of \( \Lambda_v \). Choosing an appropriate correlation structure is still a source of mild controversy despite having been studied extensively (Monti, 2011). However, autoregressive structures are often effective in fMRI settings (Lee et al., 2014; Lindquist, 2008; Penny et al., 2003). In fact, Lee et al. (2014) study the performance of several correlation structures under model (1), and demonstrate that an AR(1) structure typically provides a good balance between computational efficiency and inferential quality.

The design matrix \( X_v \) is the result of a convolution of the stimulus input function and an assumed hemodynamic response function. Let \( s_{v,j}(t) \) be the stimulus function for task \( j \) at voxel \( v \). Typically, the stimulus function simply indicates whether the stimulus is present or absent at time \( t \), and is represented by a “boxcar” function. The hemodynamic response function is denoted \( h_v(t) \), and models the BOLD response to a stimulus. We will employ the canonical hemodynamic response function, which uses the difference of two gamma densities to model the BOLD response (Friston et al., 1998). Then the elements of \( X_v \) are given by the discretized convolution

\[
X_v(t,j) = (s_{v,j} \ast h_v)(t) = \sum_{i=0}^{t-d_v} s_{v,j}(t-d_v-i) \cdot h_v(i)
\]

if \( t-d_v > 0 \), and zero otherwise, where \( d_v \) is a lag parameter that is estimated in a preprocessing step. See Friston et al. (2007) for much more about the development of \( X_v \).

The regression coefficients are interpreted as activation amplitudes, and detecting neuronal activation is equivalent to detecting the nonzero \( \beta_{v,j} \), which is a variable selection problem. We will address this through a data augmentation scheme (George and McCulloch, 1993). Let \( \gamma_{v,j} \)
be binary random variables such that $\beta_{v,j} \neq 0$ if $\gamma_{v,j} = 1$, and $\beta_{v,j} = 0$ if $\gamma_{v,j} = 0$. Let $\gamma_v = (\gamma_{v,1}, \gamma_{v,2}, \ldots, \gamma_{v,p})$, so that $\beta_v(\gamma_v)$ is the vector of nonzero coefficients from $\beta_v$, and $X_v(\gamma_v)$ is the corresponding design matrix. Then we can express model (1) as

$$y_v = X_v(\gamma_v)\beta_v(\gamma_v) + \varepsilon_v.$$  (2)

Thus far we have addressed the issue of detecting neuronal activation in the presence of temporal correlation while incorporating the nature of the stimulus and the hemodynamic response function, but we have not addressed spatial dependence. This will be handled through the specification of the prior distribution for the $\gamma_v$.

## 2.1 Prior Specification

Due to the large number of voxels it is easy to select priors that result in computationally intractable posterior distributions. While we will strive to reflect the physical nature of the experiment, we will have to make some choices that balance inferential efficacy with computational feasibility.

### 2.1.1 Priors for $\sigma_v^2$ and $\rho_v$

Let $\rho = (\rho_1, \rho_2, \ldots, \rho_N)$. We assume the $\rho_v$ are a priori independent, and so the joint prior on $\rho$ is $\pi(\rho) = \prod_{v=1}^N \pi(\rho_v)$. Fix $\hat{\rho}_v \in [-1, 1]$. We adopt an empirical Bayes approach, assuming that $\rho_v = \hat{\rho}_v$ with probability 1. We will set $\hat{\rho}_v$ to be the maximum likelihood estimator of $\rho_v$, but other ways of selecting $\hat{\rho}_v$ may be more appropriate in some situations. In what follows, $\hat{\Lambda}_v$ denotes a matrix with $(i,j)$th element $\hat{\rho}_v^{[i-j]}$.

While a fully Bayesian approach is possible, there seems to be little loss in inferential efficacy with the empirical Bayes approach. Indeed, Lee et al. (2014) compared the use of the same empirical Bayes prior to the use of other proper prior distributions on $[-1, 1]$, and found little difference in the quality of inference. In Appendix A we show why fully Bayesian priors may increase the required computing by a non-trivial amount.

Let $\sigma^2 = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2)$. We assume the $\sigma_v^2$ are a priori independent and that each has the
standard invariant prior, so the prior on $\sigma^2$ is

$$\pi(\sigma^2) = \prod_{v=1}^{N} \sigma_{v}^{-2}.$$  

### 2.1.2 Prior for $\beta_v(\gamma_v)$

Let $\beta(\gamma) = (\beta_1(\gamma_1), \beta_2(\gamma_2), \ldots, \beta_N(\gamma_N))$ and $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_N)$. Given $\gamma$, $\sigma^2$, and $\rho$, we assume the $\beta_v(\gamma_v)$ are conditionally independent:

$$\pi(\beta \mid \gamma, \sigma^2, \rho) = \prod_{v=1}^{N} \pi(\beta_v(\gamma_v) \mid \gamma_v, \sigma^2_v, \rho_v).$$

Letting

$$\hat{\beta}_v(\gamma_v) = \left( X^T_v(\gamma_v) \hat{\Lambda}^{-1}_v X_v(\gamma_v) \right)^{-1} X^T_v(\gamma_v) \hat{\Lambda}^{-1}_v y_v$$

$$\hat{\Sigma}_v = \left[ X^T_v(\gamma_v) \hat{\Lambda}^{-1}_v X_v(\gamma_v) \right]^{-1},$$

we use a g-prior (Zellner, 1986) for each $\beta_v(\gamma_v)$, which has the form

$$\beta_v(\gamma_v) \mid \gamma_v, \sigma^2_v, \rho_v \overset{\text{ind}}{\sim} N(\hat{\beta}_v(\gamma_v), T_v \sigma^2_v \hat{\Sigma}_v(\gamma_v)).$$

This is a data-dependent prior since $\hat{\beta}_v(\gamma_v)$ and $\hat{\Sigma}_v$ depend on $y_v$. From a computational standpoint, it is convenient since it allows the $\beta_v(\gamma_v)$ to be easily integrated out when computing marginal posteriors.

### 2.1.3 Prior for $\gamma_v$

We work directly with the prior probabilities of activation $\pi(\gamma_{v,j} = 1)$. This approach has been shown to produce activation maps with better edge-preservation properties and classification accuracies compared to methods that place priors only on the activation amplitudes (Smith and Fahrmeir, 2007).

We begin by assuming that the spatial dependence in the images is governed by an underlying areal model (Cressie, 1993; Haran, 2011; Banerjee et al., 2003), and that the image can be parcellated into $G$ non-overlapping regions. Ideally, the parcellation should be chosen so that the voxels within
a given region behave similarly. A large number of anatomical parcellations known as *atlases* exist, and one of these can be used depending on the goals of the experiment. Additionally, several clustering-based methods are available to choose the parcellation (cf. Ryali et al., 2013). To ensure efficient computation, we recommend using fewer than \( G = 500 \) regions. For typical data sets, this means each region will contain anywhere from 10 to 400 voxels. In Section 5 we investigate computational issues related to various choices of \( G \).

Let \( \gamma(j) = (\gamma_{1,j}, \gamma_{2,j}, \ldots, \gamma_{N,j}) \) be the vector of all active/inactive voxels under task \( j \). The vector \( \gamma(j) \) is often called the *image* (after reshaping) under task \( j \). To model the spatial dependence we introduce spatial random effects \( S(j) = (S_{1,j}, S_{2,j}, \ldots, S_{G,j}) \). Also denote as \( R_g \) the collection of all voxels in region \( g \). We assume that the \( \gamma_{v,j} \) are conditionally independent, and so the prior is

\[
\pi(\gamma(j) | S(j)) = \prod_{g=1}^{G} \prod_{v \in R_g} \pi(\gamma_{v,j} | S_{g,j}).
\]

All voxels in region \( g \) share the same spatial random effect, and hence the prior probability of activation is the same for all voxels in \( R_g \). Given that voxel \( v \) lies in \( R_g \), the next step is to link the prior probabilities to the spatial random effects via the logistic transformation

\[
\gamma_{v,j} | S_{g,j} \overset{\text{ind}}{\sim} \text{Bern} \left( \frac{1}{1 + e^{-(a_{g,j} + S_{g,j})}} \right). \tag{4}
\]

The constants \( a_{g,j} \) are chosen to incorporate prior anatomical information, which is discussed in Section 2.2. We assume that the \( S(j) \) are generated from a Gaussian process so that

\[
S(j) | \delta_j^2, r_j \overset{\text{ind}}{\sim} \mathcal{N}(0, \delta_j^2 \Gamma_j),
\]

where the \((i, k)\)th element of \( \Gamma_j \) is given by

\[
\Gamma_j(i, k) = \exp \left( -\frac{||s_i - s_k||}{r_j} \right). \tag{5}
\]

Here \( s_i \) and \( s_k \) denote the centroid coordinates of regions \( i \) and \( k \), \( || \cdot || \) is the Euclidean distance, and \( \delta_j^2 \) is a smoothing parameter that controls the amount of spatial continuity in the \( S(j) \) and hence in the \( \gamma(j) \). The matrix \( \Gamma_j \) controls the structure and amount of the spatial dependence in the \( S(j) \), and therefore also in the \( \gamma(j) \). The strength of the spatial dependence between neighboring regions under task \( j \) is determined by the range parameter \( r_j \). We assume that \( r_j > 0 \) so that
\( \Gamma_j \) is a valid spatial correlation matrix. This prior assumes that regions close to one another will exhibit similar behavior. Notice that we allow \( r_j \) and \( \delta_j^2 \) to vary across the \( p \) tasks and stimuli in the experiment, since different tasks may result in images with different amounts of spatial dependence and smoothness. This prior structure is discussed extensively by Haran (2011).

We now place priors on the hyperparameters of the spatial prior. Let \( r = (r_1, r_2, \ldots, r_p) \), \( \delta^2 = (\delta_1^2, \delta_2^2, \ldots, \delta_p^2) \), and \( S = (S(1), S(2), \ldots, S(p)) \). Then we assume that

\[
\pi(S \mid \delta^2, r) = \prod_{j=1}^p \pi(S(j) \mid \delta_j^2, r_j)
\]

\[
\pi(\delta^2) = \prod_{j=1}^p \pi(\delta_j^2)
\]

\[
\pi(r) = \prod_{j=1}^p \pi(r_j).
\]

We assume the standard invariant prior for \( \delta_j^2 \), i.e., \( \pi(\delta_j^2) \propto \delta_j^{-2} \), and \( \chi^2 \) priors for the \( r_j \). The \( \chi^2 \) priors are reasonable priors for two reasons. First, the support of the prior density is the non-negative real line, which coincides with the values of \( r \) we consider. Second, as the prior mean of \( r \) increases, the prior variance also increases, reflecting increasing uncertainty about the spatial dependence parameter.

### 2.2 Incorporating Prior Information

Prior anatomical information can be incorporated in two ways. The first is somewhat indirect, but it is important nonetheless. Note that our proposed methodology does not require that each region of the parcellation contain an equal number of voxels. Therefore, region sizes can be chosen based on prior beliefs regarding activation. If it is known a priori that a large contiguous group of voxels is unlikely to be activated during a particular task, those voxels can be assigned to one large region. On the other hand, if there is an area of uncertainty, the voxels in that area can be broken up into many smaller regions.

The second way to incorporate prior information is through the choice of the constants \( a_{g,j} \) in (4). Suppose that prior anatomical information suggests that the probability of any voxel in region
g being active under task j is \( \pi_{g,j} \). Then we want to select \( a_{g,j} \) so that \( E(\gamma_{v,j} | v \in \mathcal{R}_g) \approx \pi_{g,j} \). Note that
\[
E(\gamma_{v,j} | v \in \mathcal{R}_g) = E \left( \frac{1}{1 + e^{-(a_{g,j} + S_{g,j})}} \right) \text{ set} = \pi_{v,j}. \tag{6}
\]
The expectation in (6) cannot be solved analytically. Although it can be computed using numerical integration or Monte Carlo methods, a quick closed-form solution is often desired. Additionally, the expectation in (6) depends on \( \delta^2_j \). The relationship is complex and can only be explored via simulation. Instead, we use the fact that
\[
\text{median} \left( \frac{1}{1 + e^{-(a_{g,j} + S_{g,j})}} \right) \approx E \left( \frac{1}{1 + e^{-(a_{g,j} + S_{g,j})}} \right) \delta^2_j \tag{7}
\]
and select the \( a_{g,j} \) to satisfy
\[
\text{median} \left( \frac{1}{1 + e^{-(a_{g,j} + S_{g,j})}} \right) = \pi_{g,j}, \tag{8}
\]
which implies \( a_{g,j} = \log(\pi_{g,j}/(1 - \pi_{g,j})) \). A simulation study (not reported here but available as supplementary material) shows that the approximation in (7) is typically satisfactory.

Note that \( E(\gamma_{v,j} | v \in \mathcal{R}_g) = 0.5 \) if \( a_{g,j} = 0 \). In the absence of prior anatomical knowledge, the \( a_{g,j} \) can be set to zero for all \( g \), and \( j \) given a default standard uniform prior.

### 2.3 Posterior Distribution

Combining the results of the previous sections, the full posterior distribution is given by
\[
q(\beta(\gamma), \gamma, S, \delta^2, r, \sigma^2 | y) \propto p(y | \beta(\gamma), \gamma, S, \delta^2, r, \sigma^2) \pi(\beta(\gamma), \gamma, S, \delta^2, r, \sigma^2) \tag{9}
\]
\[
\times p(y | \beta(\gamma), \gamma, \sigma^2) \pi(\beta(\gamma)) \pi(\gamma, \sigma^2) \pi(\sigma^2)
\]
\[
\times \pi(\gamma | S) \pi(S | \delta^2, r) \pi(\delta^2) \pi(r).
\]

Our main goals are to determine which tasks and stimuli result in voxel activation as well as to determine the amount of spatial dependence in the images. Therefore, we need to compute the posterior probabilities of activation \( q(\gamma_{v,j} = 1 | y) \) for all \( v, j \) as well as posterior estimates of the spatial dependence parameter \( E(r_j | y) \) for all \( j \). These quantities cannot be analytically determined from (9), and so we use Markov chain Monte Carlo (MCMC) methods.
The dimension of the posterior in (9) is $2p(N + 1) + N + pG$, which, in a typical single-subject study, can range from tens of thousands to several millions of variables. Note that although we discussed a prior for $\rho$, that prior does not appear in the posterior since it is fixed at $\hat{\rho}_v$ under the empirical Bayes prior. Fortunately, it is sufficient to work with the marginal posterior $q(\gamma, S, r \mid y)$, which we now derive.

The first step is to integrate out $\beta(\gamma)$ and $\sigma^2$, which reduces the dimension by $(p + 1)N$. In typical settings, this represents a reduction in dimension of approximately 35 to 50 percent. Note that

$$p(y \mid \gamma) = \int p(y \mid \beta(\gamma), \gamma, \sigma^2)\pi(\beta(\gamma) \mid \gamma, \sigma^2)d\beta(\gamma)d\sigma^2$$

where $q_v = \sum_{j=1}^p \gamma_{v,j}$ denotes the number of non-zero entries in $\gamma_v$, $\hat{\Lambda}_v$ is the matrix with $(i,j)$th element $\hat{\rho}_{v|i-j}$, and

$$K(\gamma_v) = [y_v - X_v(\gamma_v)\hat{\beta}_v(\gamma_v)]^T\hat{\Lambda}_v^{-1}[y_v - X_v(\gamma_v)\hat{\beta}_v(\gamma_v)],$$

where $\hat{\beta}_v(\gamma_v)$ is defined in (3). Combining the above results gives the marginal posterior:

$$q(\gamma, S, r, \delta^2 \mid y) \propto p(y \mid \gamma)\pi(\gamma \mid S)\pi(S \mid \delta^2, r)\pi(\delta^2)\pi(r).$$

The next step is to integrate out $\delta^2$ since it is not of direct interest to the classification problem. We obtain

$$\pi(S \mid r) = \int \pi(S \mid \delta^2, r)\pi(\delta^2)d\delta^2 = \prod_{j=1}^p |\Gamma_j|^{-1/2} \left[ S^{T}_{(j)} \Gamma_j^{-1} S_{(j)} \right]^{-G/2}.$$

Then the marginal posterior is

$$q(\gamma, S, r \mid y) \propto p(y \mid \gamma)\pi(\gamma \mid S)\pi(S \mid r)\pi(r).$$

It would be ideal to integrate out $S$ to obtain the marginal posterior $q(\gamma, r \mid y)$, but the integral

$$\pi(\gamma \mid r) = \int \pi(\gamma \mid S)\pi(S \mid r)dS$$

is analytically intractable.
3 Markov Chain Monte Carlo

We derive an MCMC algorithm having invariant density \( q(\gamma, S, r \mid y) \). The dimension of \( q(\gamma, S, r \mid y) \) is \( p(N + G + 1) \), which can be extremely high-dimensional. Sampling from such posteriors is challenging, and component-wise sampling schemes (Johnson et al., 2013) are often employed to make the problem manageable. Implementing component-wise MCMC requires the posterior full conditional densities. It is straightforward to show that

\[
q(\gamma \mid S, r, y) \propto \pi(\gamma \mid S) N \prod_{v=1}^{N} (1 + T_v)^{-q_v/2} K(\gamma_v)^{-T_v/2}
\]

\[
q(S \mid \gamma, r, y) \propto \pi(\gamma \mid S) \pi(S \mid r)
\]

\[
q(r \mid S, \gamma, y) \propto \pi(S \mid r) \pi(r).
\]

Now, one complete update of all the parameters looks like

\[
(S, \gamma, r) \rightarrow (S', \gamma, r) \rightarrow (S', \gamma', r) \rightarrow (S', \gamma', r'),
\]

where each update is a Metropolis-Hastings step based on the relevant conditional density. We now turn our attention to describing each of the component-wise updates in detail.

### 3.1 Updating \( S \mid \gamma, r, y \)

This sampling step involves \( pG \) updates. Recall that \( S = (S_1, S_2, \ldots, S_p) \), and so we update according to the sequence

\[
(S_1, S_2, \ldots, S_p) \rightarrow (S'_1, S_2, \ldots, S_p) \rightarrow \ldots \rightarrow (S'_1, S'_2, \ldots, S'_p),
\]

and the \( g \) components of each \( S_j \) are updated as

\[
(S_{1,j}, S_{2,j}, \ldots, S_{G,j}) \rightarrow (S'_{1,j}, S_{2,j}, \ldots, S_{G,j}) \rightarrow \ldots \rightarrow (S'_{1,j}, S'_{2,j}, \ldots, S'_{G,j}).
\]

Let \( \gamma_{(g,j)} \) denote the vector of all \( \gamma_{v,j} \) in region \( g \) under task \( j \), and let

\[
\pi(\gamma_{(g,j)} \mid S_{g,j}) = \prod_{v \in \mathcal{R}_g} \pi(\gamma_{v,j} \mid S_{g,j}). \quad (12)
\]
Also let
\[ S_{-g,j} = (S'_{1,j}, \ldots, S'_{g-1,j}, S_{g+1,j}, \ldots, S_{G,j}) \]
and
\[ S_{-g,j} = (S'_{1,j}, \ldots, S'_{j-1,j}, S_{-g,j}, S_{j+1,j}, \ldots, S_{p,j}). \]
Then the conditional density of interest is
\[ q(S_{g,j} \mid S_{-g,j}, \gamma, r, y) = q(S_{g,j} \mid S_{-g,j}, \gamma_{(g,j)}, r_j) \]
\[ \propto \pi(\gamma_{(g,j)} \mid S_{g,j}) \pi(S_{j} \mid r_j). \]

We can use Metropolis-Hastings to sample from \( q(S_{g,j} \mid S_{-g,j}, \gamma, r, y) \). Let \( p_S(S^*_{g,j} \mid S_{g,j}) \) denote the proposal density, and let \( S^*_{g,j} \) denote the proposed value. If
\[ S^*_{(j)} = (S'_{1,j}, \ldots, S'_{g-1,j}, S_{g,j}, S_{g+1,j}, \ldots, S_{G,j}), \]
the Hastings ratio is
\[ \frac{\pi(\gamma_{(g,j)} \mid S^*_{g,j}) \pi(S^*_{j} \mid r_j) p_S(S_{g,j} \mid S^*_{g,j})}{\pi(\gamma_{(g,j)} \mid S_{g,j}) \pi(S_{j} \mid r_j) p_S(S^*_{g,j} \mid S_{g,j})}. \]

We use a random walk proposal \( S^*_{g,j} \sim N(S_{g,j}, \sigma^2_{S_{g,j}}) \), where \( S_{g,j} \) is the current state and \( \sigma^2_{S_{g,j}} \) is a tuning parameter. In this case, the Hastings ratio is
\[ \frac{\pi(\gamma_{(v,j)} \mid S^*_{g,j}) \left[ S^*_{j} \Gamma_j^{-1} S_{j}^* \right]^{-G/2}}{\pi(\gamma_{(v,j)} \mid S_{g,j}) \left[ S_{j}^T \Gamma_j^{-1} S_{j} \right]^{-G/2}}. \]

### 3.2 Updating \( \gamma \mid r, S, y \)

This sampling step consists of \( pN \) updates and is a two-step process. The \( \gamma_{(j)} \) are updated in the following order:

\[ (\gamma(1), \gamma(2), \ldots, \gamma(p)) \rightarrow (\gamma'(1), \gamma(2), \ldots, \gamma(p)) \rightarrow \ldots \rightarrow (\gamma(1), \gamma'(2), \ldots, \gamma(p)). \]

We update the \( N \) elements of each \( \gamma_{(j)} \) one at a time. Therefore, one update of \( \gamma_{(j)} \) looks like

\[ (\gamma_{1,j}, \gamma_{2,j}, \ldots, \gamma_{N,j}) \rightarrow (\gamma'_{1,j}, \gamma_{2,j}, \ldots, \gamma_{N,j}) \rightarrow \ldots \rightarrow (\gamma'_{1,j}, \gamma'_{2,j}, \ldots, \gamma'_{N,j}). \]
If \( \gamma_{-v,j} = (\gamma'_{1,j}, \ldots, \gamma'_{v-1,j}, \gamma_{v+1,j}, \ldots, \gamma_{N,j}) \), the conditional density of interest is
\[
q(\gamma_{v,j} | \gamma_{-v,j}, S_{g,j}, y) \propto \pi(\gamma_{v,j} | S_{g,j})(1 + T_v)^{-\gamma_{v,j}/2}K(\gamma_v)^{-T_v/2}. \tag{14}
\]
It is easy to sample the Bernoulli random variables \( \gamma_{v,j} \) from \( q(\gamma_{v,j} | \gamma_{-v,j}, S_{g,j}, y) \) by simply evaluating (14) at \( \gamma_{v,j} = 0 \) and \( \gamma_{v,j} = 1 \) and normalizing to get the right success probability.

### 3.3 Updating \( r | S, \gamma, y \)

Recall that \( r = (r_1, r_2, \ldots, r_p) \). One update of \( r \) looks like
\[
(r_1, r_2, \ldots, r_p) \rightarrow (r'_1, r_2, \ldots, r_p) \rightarrow \ldots \rightarrow (r'_1, r'_2, \ldots, r'_p).
\]

First, notice that
\[
q(r_j | r_{-j}, S, \gamma, y) = \pi(S_{(j)} | r_j)\pi(r_j). \tag{15}\]

We employ a Metropolis-Hastings step to obtain a draw from \( q(r_j | r_{-j}, S, \gamma, y) \) using a proposal density \( p_r(r_j^* | r_j) \). The Hastings ratio is
\[
\frac{\pi(S_{(j)} | r_j^*)\pi(r_j^*)p_r(r_j^* | r_j)}{\pi(S_{(j)} | r_j)\pi(r_j)p_r(r_j | r_j^*)}.
\]

Once again we use a random walk proposal \( r_j^* \sim N(r_j, \sigma_{r_j}^2) \), where \( r_j \) is current value and \( \sigma_{r_j}^2 \) is a tuning parameter. Let \( \Gamma_j^* \) be the correlation matrix in (5) evaluated at \( r_j^* \). In this case, the Hastings ratio is given by
\[
\frac{|\Gamma_j^*|^{-1/2}[S_{(j)}^{T} \Gamma_j^{*-1} S_{(j)}]^{-G/2} \pi(r_j^*)}{|\Gamma_j|^{-1/2}[S_{(j)}^{T} \Gamma_j^{-1} S_{(j)}]^{-G/2} \pi(r_j)}.
\]
This step is computationally expensive due to the inversion of \( \Gamma_j^* \), especially when \( G \) is large.

### 3.4 Posterior Estimates and Classification of Active Voxels

Recall that we have two inferential goals for each task: (1) identification of active voxels, and (2) estimation of the spatial dependence parameter. Suppose that we draw \( M \) samples \( \{(\gamma^{(m)}, S^{(m)}, r^{(m)})\}_{m=1}^{M} \)
using the MCMC algorithm described above. Notice that we can discard the $S^{(m)}$ since the posterior of interest is $q(\gamma, r \mid y)$.

We identify active voxels by estimating the posterior probability of activation $q(\gamma_{v,j} = 1 \mid y)$ under task $j$ for each voxel with

$$\hat{q}_{v,j} = \frac{1}{M} \sum_{m=1}^{M} \gamma_{v,j}^{(m)}.$$  

We classify voxel $v$ as activated under task $j$ if $\hat{q}_{v,j} > 0.8722$. This activation threshold has been discussed extensively (Smith and Fahrmeir, 2007; Lee et al., 2014; Raftery, 1996; Smith and Smith, 2006), and we further investigate its performance in Section 4.

Estimating the spatial dependence parameter is done by estimating the posterior mean $E[r_j \mid y]$ for each task $j$ with

$$\hat{r}_j = \frac{1}{M} \sum_{m=1}^{M} r_j^{(m)}.$$  

Assessing the quality of the estimates is best done by calculating a Monte Carlo standard error (Flegal et al., 2008; Flegal and Jones, 2011). In our numerical work we employ the method of batch means (Jones et al., 2006) via the mcmcse package (Flegal and Hughes, 2012) for R (Ihaka and Gentleman, 1996), with the batch size set to the square root of the number of samples drawn.

4 Simulation Study

We now assess the performance of the proposed methodology on simulated data sets. In the first simulation, we focus on the ability of the proposed methodology to classify active voxels, as well as its ability to accurately infer the strength of the spatial dependence in the images. The second simulation considers the performance of the model when spatial independence is assumed. We conclude by checking the performance of our method using various activation thresholds.

4.1 Simulated Data and MCMC Implementation

We generate 15 $20 \times 20$ active/inactive square images under the proposed spatial hierarchical model with $G = N$ and the true value of $r \in \{0.001, 2, 8, 20\}$. These values represent images that have
spatial dependence structures ranging from near independence to strong spatial dependence. In each case we set $\delta^2 = 5$, leading to images with moderate amounts of spatial smoothing.

For each image, a sequence of $T_v = 50$ responses was generated at each voxel from the model in (2). The design matrix $X_v$ used in this simulation is displayed in Figure 1. Let $\beta_v(\gamma_v) = (\beta_v, 0, \beta_v, 1)^T$. We assume that $\beta_{v,0} = 300$. When $\gamma_v = 0$, we set $\beta_{v,1} = 0$, and when $\gamma_v = 1$, we consider $\beta_{v,1} = 3$ and $\beta_{v,1} = 5$, which are typical signal strengths encountered in fMRI data. We select $\sigma_v^2 = 3$ for all $v$ while the AR(1) correlation coefficients $\rho_v$ were generated independently from a Uniform($-1, 1$) distribution.

We will use this simulated data in a few different ways to investigate the performance of our inferential procedure. We employed the component-wise MCMC methodology described in Section 3 to produce 150,000 samples $\{(\gamma^{(m)}, S^{(m)}, r^{(m)})\}$. The tuning parameters for the MCMC algorithm were chosen to produce acceptance rates close to 50%. Other standard diagnostic measures (such as trace plots) were used to ensure that the Markov chain was mixing well. In addition, Monte Carlo standard errors were calculated, and 150,000 samples resulted in all standard errors being within 2% of the estimated posterior mean.
4.2 Simulation 1

We investigate the performance of our model and inferential procedure under two prior settings for $r$. Specifically, we consider both $\chi^2_2$ and $\chi^2_8$ priors. No prior anatomical information is available, and so the constants $a_{g,j}$ in (4) are all set to zero.

<table>
<thead>
<tr>
<th>True $r$</th>
<th>$\hat{r}$</th>
<th>Accuracy (%)</th>
<th>FPR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>1.01 (0.03)</td>
<td>87.95 (0.08)</td>
<td>1.22 (0.03)</td>
</tr>
<tr>
<td>2</td>
<td>2.98 (0.06)</td>
<td>89.70 (0.08)</td>
<td>1.20 (0.03)</td>
</tr>
<tr>
<td>8</td>
<td>5.10 (0.11)</td>
<td>89.57 (0.08)</td>
<td>1.14 (0.02)</td>
</tr>
<tr>
<td>20</td>
<td>5.58 (0.11)</td>
<td>90.31 (0.08)</td>
<td>0.82 (0.02)</td>
</tr>
</tbody>
</table>

(a)

<table>
<thead>
<tr>
<th>True $r$</th>
<th>$\hat{r}$</th>
<th>Accuracy (%)</th>
<th>FPR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>3.06 (0.10)</td>
<td>88.18 (0.08)</td>
<td>1.11 (0.03)</td>
</tr>
<tr>
<td>2</td>
<td>5.33 (0.10)</td>
<td>89.70 (0.08)</td>
<td>1.21 (0.03)</td>
</tr>
<tr>
<td>8</td>
<td>7.46 (0.14)</td>
<td>90.00 (0.08)</td>
<td>1.20 (0.03)</td>
</tr>
<tr>
<td>20</td>
<td>7.92 (0.17)</td>
<td>90.13 (0.08)</td>
<td>0.78 (0.02)</td>
</tr>
</tbody>
</table>

(b)

Table 1: Posterior estimates of $r$, classification accuracies, and false positive rates when $\beta_v,1 = 3$, with (a) $\chi^2_2$ and (b) $\chi^2_8$ priors for $r$.

The posterior estimates of $r$, classification accuracies, and false positive rates (FPR) averaged over the 15 simulated data sets are reported in Tables 1 and 2. We first point out that, on average, classification accuracies are much higher when $\beta_v,1 = 5$, which is expected due to the stronger signal. We see that the posterior estimates of $r$ are reasonable, and that the classification accuracies and FPRs are insensitive to the prior degrees of freedom. As the strength of spatial dependence increases, classification accuracies tend to increase as well due to the fact that it is generally easier to classify images that exhibit greater clustering. Note that the differences in classification accuracies as $r$ increases are less pronounced when $\beta_v,1 = 5$. 

16
Table 2: Posterior estimates of $r$, classification accuracies, and false positive rates when $\beta_{v,1} = 5$, with (a) $\chi^2_2$ and (b) $\chi^2_8$ priors for $r$.

<table>
<thead>
<tr>
<th>True $r$</th>
<th>$\hat{r}$</th>
<th>Accuracy (%)</th>
<th>FPR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>2.35 (0.06)</td>
<td>97.15 (0.04)</td>
<td>1.77 (0.03)</td>
</tr>
<tr>
<td>2</td>
<td>3.89 (0.07)</td>
<td>97.76 (0.04)</td>
<td>1.79 (0.03)</td>
</tr>
<tr>
<td>8</td>
<td>6.23 (0.11)</td>
<td>97.91 (0.04)</td>
<td>2.21 (0.03)</td>
</tr>
<tr>
<td>20</td>
<td>7.24 (0.12)</td>
<td>98.11 (0.03)</td>
<td>2.21 (0.03)</td>
</tr>
</tbody>
</table>

(a)

<table>
<thead>
<tr>
<th>True $r$</th>
<th>$\hat{r}$</th>
<th>Accuracy (%)</th>
<th>FPR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>4.74 (0.13)</td>
<td>97.08 (0.04)</td>
<td>2.07 (0.04)</td>
</tr>
<tr>
<td>2</td>
<td>6.37 (0.11)</td>
<td>97.68 (0.04)</td>
<td>1.83 (0.03)</td>
</tr>
<tr>
<td>8</td>
<td>8.80 (0.15)</td>
<td>97.88 (0.04)</td>
<td>2.04 (0.03)</td>
</tr>
<tr>
<td>20</td>
<td>10.00 (0.17)</td>
<td>98.05 (0.03)</td>
<td>2.46 (0.04)</td>
</tr>
</tbody>
</table>

(b)

4.3 Simulation 2

Models that ignore temporal correlation offer poor classification performance and inferior inference (Monti (2011) provides a detailed discussion, and some numerical results are reported by Lee et al. (2014)). However, the performance of models that do not fully consider spatial dependence has not been widely studied. We therefore investigate what happens when the default prior $\pi(\gamma_v = 1) = 0.5$ for all $v$, described in Smith and Kohn (1996), is employed. This requires us to modify our sampler to draw from the posterior distribution

$$q(\gamma \mid y) \propto p(y \mid \gamma)\pi(\gamma).$$

We use this modified sampling scheme on the same simulated data sets that were obtained in Section 4.1.

Table 3 displays the results when the default prior is used for $\gamma$. In each case, the classification
accuracies are lower than those when the spatial hierarchical prior is used. False positive rates are similar, which implies that a loss of power is incurred under the default prior.

As the strength of spatial dependence in the generated images increases, the performance of the default prior worsens compared to the spatial hierarchical prior. In fact, when $r = 0.001$, the classification accuracies are nearly identical, but are markedly worse when $r = 20$. Also, the performance decrease under the default prior is less severe when $\beta_{v,1} = 5$. This is expected since as signal strength gets larger, most methods will perform comparably since activation is more obvious and easier to detect. From this simulation, we conclude that the default prior for $\gamma$ leads to poor classification accuracy when the spatial dependence in images is strong. This prior also does not provide a way of obtaining information about the strength of spatial dependence in the images.

<table>
<thead>
<tr>
<th>True $r$</th>
<th>Accuracy (%)</th>
<th>FPR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>87.52 (0.08)</td>
<td>1.18 (0.02)</td>
</tr>
<tr>
<td>2</td>
<td>87.68 (0.08)</td>
<td>1.07 (0.02)</td>
</tr>
<tr>
<td>8</td>
<td>86.35 (0.09)</td>
<td>1.04 (0.02)</td>
</tr>
<tr>
<td>20</td>
<td>86.61 (0.09)</td>
<td>0.75 (0.02)</td>
</tr>
</tbody>
</table>

(a)

<table>
<thead>
<tr>
<th>True $r$</th>
<th>Accuracy (%)</th>
<th>FPR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.001</td>
<td>97.01 (0.04)</td>
<td>1.32 (0.03)</td>
</tr>
<tr>
<td>2</td>
<td>97.10 (0.04)</td>
<td>1.38 (0.03)</td>
</tr>
<tr>
<td>8</td>
<td>96.88 (0.04)</td>
<td>1.44 (0.03)</td>
</tr>
<tr>
<td>20</td>
<td>97.03 (0.04)</td>
<td>1.43 (0.03)</td>
</tr>
</tbody>
</table>

(b)

Table 3: Classification accuracies and false positive rates when the default prior is used for the $\gamma$s and (a) $\beta_{v,1} = 3$ and (b) $\beta_{v,1} = 5$. 
4.4 Simulation 3

Recall that voxels are classified as active if \( \hat{q}_{v,j} > 0.8722 \). We now investigate the validity of this activation threshold. We reclassify activation in the 15 simulated data sets from Section 4.1, which were generated when \( r = 8 \) at both \( \beta_{v,1} = 3 \) and 5. We consider the following three activation thresholds: 0.7946, 0.8722, and 0.9650. The results are reported in Table 4. We conclude that the 0.8722 threshold provides a balance between the overall classification accuracy and the false positive rate.

<table>
<thead>
<tr>
<th>Activation Threshold</th>
<th>Accuracy (%)</th>
<th>FPR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7946</td>
<td>91.47 (0.07)</td>
<td>2.30 (0.04)</td>
</tr>
<tr>
<td>0.8722</td>
<td>90.00 (0.08)</td>
<td>1.20 (0.03)</td>
</tr>
<tr>
<td>0.9650</td>
<td>84.92 (0.09)</td>
<td>0.61 (0.01)</td>
</tr>
</tbody>
</table>

(a)

<table>
<thead>
<tr>
<th>Activation Threshold</th>
<th>Accuracy (%)</th>
<th>FPR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7946</td>
<td>97.64 (0.04)</td>
<td>3.59 (0.03)</td>
</tr>
<tr>
<td>0.8722</td>
<td>97.88 (0.04)</td>
<td>2.04 (0.03)</td>
</tr>
<tr>
<td>0.9650</td>
<td>96.65 (0.05)</td>
<td>1.00 (0.02)</td>
</tr>
</tbody>
</table>

(b)

Table 4: Classification accuracies and false positive rates at various activation thresholds when (a) \( \beta_{v,1} = 3 \) and (b) \( \beta_{v,1} = 5 \) and the \( \chi^2_8 \) prior is used for \( r \).

5 Benchmark Example

In this section we examine the performance of the proposed methodology on a simulated data set representative of fMRI data. This allows us to apply our method under several settings, and visually observe how the results vary, focusing primarily on the effects of the parcellation scheme described in Section 2.1.3.
Consider the $20 \times 20$ image displayed in Figure 2. This image is representative of activation patterns in response to a stimulus (several clusters of active voxels with approximately 10–20% of voxels active overall.) We generate a time series of length $T_v = 50$ from model (2) under the same settings as in Section 4.1, considering both $\beta_{v,1} = 3$ and $\beta_{v,1} = 5$ when $\gamma_v = 1$. We consider only the $\chi^2_8$ prior for $r$.

![Figure 2: The image used in the benchmark simulation. Black indicates activation.](image)

Figure 3 displays the simulated BOLD signal at time point 21 for the design matrix in Figure 1. It is clear that both images are very noisy. While the image corresponding to $\beta_{v,1} = 5$ does vaguely display some areas of activation, it is difficult to see any activation when $\beta_{v,1} = 3$.

We use our method when the image is unparcellated ($G = 400$), divided into a grid of $2 \times 2$ squares ($G = 100$), and divided into a grid of $4 \times 4$ squares ($G = 25$). Figure 4 displays these parcellations. We are interested in studying how the classification accuracy, false positive rate, and false negative rate (FNR) are affected by the choice of parcellation.

Table 5 displays the results for the three parcellation schemes. The classification accuracies are similar for the three schemes, especially when $\beta_{v,1} = 5$. In addition to the usual accuracy statistics, we report the computation time relative to the $G = 400$ case. We see substantial speed improvements as the number of regions decreases. This is mostly attributable to the reduction in dimension of $\Gamma_j$. Figures 5 and 6 display the estimated posterior activation probabilities and classified images, at both activation amplitudes, for the three parcellation schemes.
One point worth noting is the large decrease in FNR when $G = 100$ (Table 5a). The reason for this decrease can be seen by looking at the parcellation in Figure 4b. When the entire image is subdivided into a grid of $2 \times 2$ voxels, the $2 \times 2$ regions fit nearly perfectly within the active cluster in the bottom left-hand corner. This induces a smoothing effect on only the activated voxels, which
<table>
<thead>
<tr>
<th>$G$</th>
<th>$\hat{\tau}$</th>
<th>Accuracy (%)</th>
<th>FPR (%)</th>
<th>FNR (%)</th>
<th>Relative Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>400</td>
<td>10.04 (0.14)</td>
<td>97.00 (0.04)</td>
<td>0 (0)</td>
<td>19.35</td>
<td>1.00</td>
</tr>
<tr>
<td>100</td>
<td>9.79 (0.07)</td>
<td>97.50 (0.04)</td>
<td>0.30 (0.01)</td>
<td>14.52</td>
<td>0.05</td>
</tr>
<tr>
<td>25</td>
<td>9.08 (0.04)</td>
<td>96.50 (0.05)</td>
<td>0.30 (0.01)</td>
<td>20.97</td>
<td>0.01</td>
</tr>
</tbody>
</table>

(a)

<table>
<thead>
<tr>
<th>$G$</th>
<th>$\hat{\tau}$</th>
<th>Accuracy (%)</th>
<th>FPR (%)</th>
<th>FNR (%)</th>
<th>Relative Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>400</td>
<td>9.41 (0.10)</td>
<td>99.50 (0.02)</td>
<td>0 (0)</td>
<td>3.22</td>
<td>1.00</td>
</tr>
<tr>
<td>100</td>
<td>9.30 (0.06)</td>
<td>99.25 (0.02)</td>
<td>0.30 (0.01)</td>
<td>3.22</td>
<td>0.05</td>
</tr>
<tr>
<td>25</td>
<td>8.70 (0.04)</td>
<td>99.25 (0.02)</td>
<td>0.59 (0.02)</td>
<td>1.61</td>
<td>0.01</td>
</tr>
</tbody>
</table>

(b)

Table 5: Results of the benchmark analysis when (a) $\beta_{v,1} = 3$ and (b) $\beta_{v,1} = 5$.

causes nearly the entire area to be classified correctly. When looking at the $G = 25$ case in Table 5a, we see that the FNR is larger than when $G = 100$ or $G = 400$. By looking at Figure 4c, we see that some of the $4 \times 4$ regions include a large number of inactive voxels. This smooths some active and inactive voxels, causing active voxels with weak signal to be classified as inactive. Note that this does not occur when the signal is stronger ($\beta_{v,1} = 5$).

This suggests using smaller regions at locations where prior information predicts activation, and reinforces the idea that the regions be chosen according to prior anatomical information when such information is available. If no prior spatial information is available, we recommend choosing $G$ as large as possible, subject to computational constraints.
Figure 5: Estimated posterior probabilities of activation and the classified images when (a, b) $G = 400$, (c, d) $G = 100$, and (d, e) $G = 25$ and $\beta_{v,1} = 3$. 
Figure 6: Estimated posterior probabilities of activation and the classified images when (a, b) $G = 400$, (c, d) $G = 100$, and (d, e) $G = 25$ and $\beta_{v,1} = 5$. 

24
6 Experimental Data Examples

6.1 Auditory Data Set

The first data set we consider is the auditory data set available from the Wellcome Trust Centre for Neuroimaging website. A single subject was studied in this experiment. Before data collection, the image space was partitioned into a $64 \times 64 \times 64$ rectangular lattice with voxel size three mm$^3$. After preprocessing and masking, 70,650 voxels were left to be analyzed. During the experiment, a total of 84 scans were collected on a 2T Siemens scanner, with one scan taken every seven seconds. The experiment was constructed as a block design using 14 blocks of six scans, with each block lasting 42 seconds.

During the odd-numbered blocks, the subject rested. During the even-numbered blocks, the subject listened to bi-syllabic words presented in rapid succession. The scientific goal was to localize the regions of the brain that play a role in processing words with more than one syllable. Before statistical analysis, we preprocessed the data using standard techniques. These included motion correction, normalization, registration, and spatial smoothing (six mm in each direction).

The image space was partitioned into 370 contiguous anatomical regions. Because we expected activation to occur mostly in the parietal and temporal lobes, we used k-means clustering to cluster the voxels in those two areas into 300 regions. The remaining space was clustered into 70 regions.

Since we expected significant spatial dependence in the images, we chose a $\chi^2_{12}$ prior for $r$. The design matrix used in this experiment (displayed in Figure 7) was obtained by convolving the canonical hemodynamic response function with the stimulus function.

6.1.1 Computation

The posterior of interest (11) is intractable, and so we used the MCMC method described in Section 3 to draw 100,000 samples. The tuning parameters of the MCMC algorithm were chosen so as to produce acceptance rates close to 50%. We also used standard diagnostic measures to assess
convergence. For example, we checked the trace plots for the spatial dependence parameter $r$ and a randomly selected subset of 30 of the spatial random effects $S_g$. Figure 8 displays the trace plots for three spatial random effects as well as the trace plot of $r$. There are no obvious signs of convergence problems.

Using the method described in Section 3.4, we then estimated the spatial dependence parameter and posterior probabilities of activation. The batch means method with the batch size set at the square root of the run length was used to calculate Monte Carlo standard errors.
6.1.2 Results

The horizontal slices of the brain showing the bulk of the activation are displayed in Figure 9. We classified voxels as active according to the 0.8722 threshold. Note that each slice is three mm thick and slice 1 is the topmost slice of the brain. The proposed methodology classified 2.6% of voxels as activated during the listening blocks. Most of the activation occurs in the auditory cortex, which is expected since that is the main region responsible for processing auditory information.

We also investigated the amount of activation detected at the two alternative activation thresholds considered in Section 4.4. The results appear in Table 6. The amount of activation detected is similar under the three thresholds, and so our choice of 0.8722 appears to be reasonable.

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Active Voxels (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7946</td>
<td>2.94</td>
</tr>
<tr>
<td>0.8722</td>
<td>2.66</td>
</tr>
<tr>
<td>0.9650</td>
<td>2.14</td>
</tr>
</tbody>
</table>

Table 6: The amount of activation detected at various activation thresholds in the auditory analysis.

The estimate of the spatial dependence parameter is \( \hat{r} = 22.18 (0.05) \), indicating strong spatial dependence in the images. This posterior estimate is consistent with scientifically informed expectations and with the voxel activation patterns displayed in Figure 9.

6.1.3 Model Assessment

The goal here is to consider voxel-level diagnostics to assess the model fit. We checked the assumption of Gaussian errors, as well as the adequacy of the AR(1) error structure, at 30 voxels. Two randomly selected active and inactive voxels are displayed in Figure 10. The first row displays the fitted regression line (in red) as well as the observed BOLD signal. It is clear that the signal is noisy, which is common in fMRI. The second row contains the residual quantile-quantile plots. Finally, the autocorrelation plots of the residuals are in shown in the third row. The plots do not suggest that a higher-order autoregressive process is necessary to model the residual autocorrelation.
We also checked the stability of the results under different parcellations by refitting the model using parcellation sizes of $G = 420$ and $G = 500$ regions. The detected activation patterns were very similar in all cases, although using larger $G$ increased the computational burden substantially.

The empirical Bayes estimates of the $\rho_v$ in five slices of the brain are displayed in Figure 11. Most of the estimated $\rho_v$ are positive, demonstrating that temporal independence is an unreasonable assumption.
Figure 10: Diagnostic plots of four randomly selected voxels in the auditory data analysis.
Figure 11: Empirical Bayes estimates of the AR(1) coefficients at five vertical slices of the brain in the auditory data analysis.
6.2 Human Connectome Project Data

The second data set we analyzed was collected as part of the Human Connectome Project (HCP) (Essen et al., 2013), and aims to evaluate emotional processing. The experiment was a modified version of the design proposed by Hariri et al. (2002), which we now summarize.

The subject laid in a scanner and completed one of two tasks arranged in a block design. In the first task, two faces were displayed in the top half of a screen. One of the faces had a fearful expression, and the other had an angry expression. A third face was displayed in the bottom half of the screen. The third face had either a fearful expression or an angry expression. The subject chose which of the two faces in the top half of the screen matched the expression of the third face in the bottom half of the screen. Each set of faces was displayed for two seconds, after which there was a one-second pause.

The second task was functionally identical to the first task, except that geometric shapes were used instead of faces, and the subject had to choose which of the two shapes in the top half of the screen matched the shape in the bottom of the screen. This task was used as a control. The goal here is to detect which regions of the brain are involved in distinguishing emotional facial expressions, and how the regional activations differ between the two tasks. Both the face and shape blocks were each 18 seconds long, with eight seconds rest between successive task blocks. Each pair of blocks was replicated three times.

A total of 176 scans were collected on a 3T scanner on over 500 subjects. We randomly selected one subject to analyze using our proposed methodology. Before data collection, the image space was partitioned into a 91 \times 109 \times 91 rectangular lattice comprising voxels of size two mm$^3$. After preprocessing and masking, a total of 225,297 voxels remained to be analyzed. We preprocessed the images using standard techniques, and spatially smoothed the images five mm in each direction. Because we expected activation to occur in several regions of the brain, we parcellated the image into 420 regions of approximately equal size.

We use a $\chi^2_{12}$ prior for the task-specific spatial dependence parameters $r_{\text{face}}$ and $r_{\text{shape}}$. The design matrix used in the statistical analysis is displayed in Figure 12.
6.2.1 Computation

The overall procedure is similar to that used for the auditory data, but a description is included here for the sake of completeness. The posterior of interest (11) is intractable, and so we used the MCMC method described in Section 3 to draw 100,000 samples. The tuning parameters of the MCMC algorithm were chosen so as to produce acceptance rates close to 50%. We also used standard diagnostic measures to assess convergence. For example, we checked trace plots of the spatial dependence parameters and a subset of 30 randomly selected spatial random effects under both tasks. Figure 13 displays the plots of three spatial random effects and the spatial dependence parameter \( r \) under both the face and shape blocks. There are no obvious signs of convergence problems.

Using the method described in Section 3.4, we then estimated the spatial dependence parameters and posterior probabilities of activation. The batch means method with the batch size set at the square root of the run length was used to calculate Monte Carlo standard errors.

6.2.2 Results

Figures 14–17 display the detected activation in several horizontal slices of the brain (using an activation threshold of 0.8722, as before). Note that each slice is two mm thick and slice 1 contains

![Figure 12: The design matrix used in the HCP analysis.](image)
the topmost region of the brain. Most of the activation occurs in the occipital lobe, which is thought to be responsible for the processing of visual information. Figures 14 and 15 also show activation in the temporal and frontal lobes during the face blocks.

During the shape blocks, 2.03% of the voxels were declared as active, whereas 3.01% of the voxels were declared active during the face blocks, indicating that more neuronal effort is required to distinguish emotional facial expressions than geometric shapes. The posterior estimate of \( r \) during the shape blocks was \( \hat{r}_{\text{shape}} = 21.56 (0.037) \), and during the face blocks was \( \hat{r}_{\text{face}} = 24.47 (0.041) \). Both tasks had activation patterns with a substantial degree of spatial dependence.
Figure 14: Neuronal activation in slices 15–18 during the (a) shape blocks and (b) face blocks.

Figure 15: Neuronal activation in slices 25–28 during the (a) shape blocks and (b) face blocks.
Figure 16: Neuronal activation in slices 36–39 during the (a) shape blocks and (b) face blocks.

Figure 17: Neuronal activation in slices 57–60 during the (a) shape blocks and (b) face blocks.
6.2.3 Model Assessment

We consider voxel-level diagnostics to assess model fit. We checked the assumption of Gaussian errors, as well as the adequacy of the AR(1) error structure, at 30 randomly selected voxels. Four randomly selected voxels are displayed in Figure 18. The first row displays the fitted regression line (in red) as well as the observed BOLD signal. The leftmost plot is a voxel that was active only under the face task. The middle two plots are voxels that were active under both tasks, and the rightmost plot is a voxel that was not active under either task. The second row shows the quantile-quantile plots, which do not indicate any problems. Finally, the autocorrelation plots of the adjusted residuals are shown in the third row. The plots do not suggest that a higher-order autoregressive process is necessary to model the residual autocorrelation.

The empirical Bayes estimates of $\rho_v$ in five slices of the brain are displayed in Figure 19. We see that temporal correlation tends to be higher in the areas that displayed activation in Figures 14–17, showing that voxel-wise temporal independence is an unreasonable assumption.

7 Final Remarks

We have proposed a Bayesian spatiotemporal model aimed at fMRI applications. The methodology offers several advantages over existing procedures. It allows prior anatomical information to be incorporated directly in the analysis while avoiding the computational issues encountered when the Ising model is used. In particular, the method requires only a few hours to complete on standard hardware.

Despite the fact that our model appears to perform well and is computationally feasible, there are still further improvements that can be considered. For example, we make the standard assumption that the BOLD signal variance is constant at all voxels. This is likely not a reasonable assumption from a scientific standpoint since the BOLD signal at active voxels tends to be noisier than at inactive voxels. Parameterizing $\sigma^2_v$ as a function of time or region may provide further insights regarding the complex underlying spatiotemporal process in fMRI data. Along the same lines, priors taking into account the spatial dependence of both $\sigma^2_v$ and $\rho_v$ are an appealing possibility.
While we have focused on the fMRI applications of our proposed Bayesian spatiotemporal model, the methodology can be used to model other spatiotemporal processes, such as weather or public health data. Although we focus exclusively on single-subject fMRI, extensions to multiple subjects
Figure 19: Empirical Bayes estimates of the AR(1) coefficients at five vertical slices of the brain in the HCP analysis.

are possible (but may be computationally expensive).

Given the recent advances in computing and the availability of large amounts of information about the human brain, Bayesian approaches are a natural way of addressing the fMRI activation detection problem. We remain confident that Bayesian approaches to analyzing fMRI provide valuable insights into the underpinnings of neuronal activation.

8 Acknowledgements

Data were provided (in part) by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

A Updating $\rho \mid S, \gamma, r, y$

Recall that $\rho = (\rho_1, \rho_2, \ldots, \rho_N)$. If we assign a prior $\pi(\rho_v)$ on $[-1, 1]$ independently, it is also necessary to sample from the posterior conditional $q(\rho \mid S, \gamma, r, y)$ in our conditional Metropolis-Hastings algorithm. In this case, the $N$ correlation coefficients are updated using the following scheme:

$$(\rho_1, \rho_2, \ldots, \rho_N) \rightarrow (\rho'_1, \rho_2, \ldots, \rho_N) \rightarrow \ldots \rightarrow (\rho'_1, \rho'_2, \ldots, \rho'_N).$$
Let $\rho_{-v} = (\rho_1, \ldots, \rho_{v-1}, \rho_{v+1}, \ldots, \rho_N)$. The conditional posterior of each $\rho_v$ is

\[
q(\rho_v \mid \rho_{-v}, S, \gamma, r, y) \propto q(\rho_v \mid \gamma_v, y) \\
\propto \pi(\rho_v) |\Lambda_v|^{-1/2} K(\rho_v, \gamma_v)^{-T_v/2},
\]

where

\[
K(\gamma_v, \rho_v) = [y_v - X_v(\gamma_v) \hat{\beta}_v(\gamma_v)]^T \Lambda_v^{-1} [y_v - X_v(\gamma_v) \hat{\beta}_v(\gamma_v)].
\]

We can use a Metropolis-within-Gibbs step here. Suppose $p_\rho(\rho^*_v \mid \rho_v)$ is the proposal density, where $\rho^*$ is the proposed value. Let $\Lambda_v^*(i, k) = \rho_v^*[i-k]$ and $\rho^* = (\rho'_1, \ldots, \rho'_{v-1}, \rho^*_v, \rho_{v+1}, \ldots, \rho_N)$. Then the Hastings ratio is

\[
\frac{\pi(\rho^*_v) |\Lambda_v^*|^{-1/2} K(\rho^*, \gamma_v)^{-T_v/2} p_\rho(\rho_v \mid \rho^*_v)}{\pi(\rho_v) |\Lambda_v|^{-1/2} K(\rho, \gamma_v)^{-T_v/2} p_\rho(\rho^*_v \mid \rho_v)}.
\]

If we use an independence sampler with a Uniform$(-1, 1)$ proposal distribution, the Hastings ratio becomes

\[
\frac{\pi(\rho^*_v) |\Lambda_v^*|^{-1/2} K(\rho^*, \gamma_v)^{-T_v/2}}{\pi(\rho_v) |\Lambda_v|^{-1/2} K(\rho, \gamma_v)^{-T_v/2}}.
\]

Although this step appears computationally intensive since $\Lambda_v$ must be inverted $N$ times, the AR(1) residual error structure significantly reduces the computational burden since a widely-available analytic expression for $\Lambda_v^{-1}$ exists (cf. Christensen et al., 2010). Computing the log-determinant of $\Lambda_v$, as well as evaluating $K(\gamma_v, \rho_v)$, however, are not trivial operations and can contribute significantly to the overall required computation. The empirical Bayes prior in Section 2.1.1 allows this step to be skipped entirely.

**References**


