Spatial Bayesian Variable Selection Models on Functional Magnetic Resonance Imaging Time-Series Data

Kuo-Jung Lee ♠
Department of Statistics
National Cheng Kung University
kjlee@stat.ncku.edu.tw

Galin L. Jones †
School of Statistics
University of Minnesota
galin@stat.umn.edu

Brian S. Caffo ‡
Department of Biostatistics
Johns Hopkins University
bcaffo@jhsph.edu

Susan Spear Bassett
Department of Psychiatry
Johns Hopkins University
sbassett@jhmi.edu

October 2011; Revised January 2013; October 12, 2013

Abstract

A common objective of fMRI (functional magnetic resonance imaging) studies is to determine subject-specific areas of increased blood oxygenation level dependent (BOLD) signal contrast in response to a stimulus or task, and hence to infer regional neuronal activity. We posit and investigate a Bayesian approach that incorporates spatial and temporal dependence and allows for the task-related change in the BOLD signal to change dynamically over the scanning session. In this way, our model accounts for potential learning effects in addition to other mechanisms of temporal drift in task-related signals. We study the properties of the model through its performance on simulated and real data sets.

♠Research supported by the National Science Foundation
†Research supported by the National Science Foundation and the National Institutes of Health. CRiSM also provided support during a visit to the University of Warwick.
‡Research supported by the the National Institutes of Health.
1 Introduction

Functional neuroimaging experiments typically aim to uncover localized regions in the brain which activate during a task, describe the networks required for a particular brain function or assess physical characteristics that result from disease or trauma. Our focus is on functional magnetic resonance imaging (fMRI) techniques to detect task-related neuronal activation. However, neuronal activation occurs too quickly to be observed directly in fMRI experiments. Despite this, the principle of neurovascular coupling (that is, local neuronal activation is related to changes in cerebral blood flow) allows us to indirectly observe activation via the blood oxygenation level dependent (BOLD) signal contrast.

The typical single-subject fMRI experiment is conceptually straightforward. A subject in an MRI scanner performs a task in response to a stimulus while three-dimensional images of the subject’s brain are captured every 2-3 seconds. However, the data has a complicated structure. Imagine that the image is divided into a regular grid of volume elements, or voxels. The BOLD signal is observed at each voxel at each time point resulting in an enormous amount of data (possibly as many as 40 million observations) exhibiting both spatial and temporal dependence. Moreover, the data tend to be noisy and have a weak signal. Hence accurate and powerful models of single-subject task-related activation would be useful in developing effective imaging biomarkers. Unfortunately there are few, if any, off-the-shelf methods for constructing sensible, computationally feasible statistical models in this situation. Indeed, Lindquist (2008) considered only non-Bayesian approaches and concludes

The size and complexity of the data make it difficult to create a full statistical model for describing its behavior, and a number of shortcuts are required to balance computational feasibility with model efficiency.

In fact, standard non-Bayesian approaches (cf. Chen et al., 2009; Friston et al., 1994; Genovese et al., 2002; Nichols and Holmes, 2002; Worsley, 2003; Worsley et al., 1992) do not produce a full model. However, many of these methods are computationally efficient.

The Bayesian paradigm provides an attractive inferential framework within which to develop models directly incorporating the physical characteristics of the experiment. Hence Bayesian methods in neuroimaging have received a fair amount of recent attention (see e.g Bowman et al., 2008; Caffo et al., 2011; Genovese, 2000; Goldsmith et al., 2012; Smith and Fahrmeir, 2007; Smith et al., 2003; Woolrich et al., 2004; Xia et al., 2009). We use Bayesian variable selection methodology to detect task-specific changes in the BOLD signal. The basic idea is now described while the details are given in Section 2. For voxel \( v = 1, \ldots, N \) let \( y_{v,i} \) be the BOLD image intensity and \( x_{v,i} \) the
transformed stimulus at time $i = 1, \ldots, T_v$. Then we assume a linear model

$$y_{v,i} = z_i^T a_v + x_{v,i} \beta_v + \varepsilon_{v,i}.$$  

The baseline trend $z_i^T a_v$ is a linear combination of basis functions to remove low-frequency stimulus-independent effects. Next, $\beta_v$ is interpreted as the activation amplitude while $x_{v,i}$ is the convolution of a stimulus function $s$ and a parametric hemodynamic response function (HRF) $h$. It is necessary to transform the input functions since for a short-duration stimulus the BOLD response increases above baseline about 2 seconds after the onset of neuronal activity, peaking at about 5-8 seconds and falling below baseline (the undershoot) for about 10 seconds (Aguirre et al., 1998). Typically, $s$ corresponds to a 'boxcar' function indicating the stimulus is active/inactive while the HRF $h$ may correspond to a Poisson or Gamma density (Glover, 1999; Gössl et al., 2001; Smith and Fahrmeir, 2007) in which case we have

$$x_{v,i} = \sum_{k=0}^{i-d_v} h(k, \lambda_v) s_{i-d_v-k}.$$  

The parameters $\lambda_v$ and $d_v$ are usually estimated in a preprocessing step. A more sophisticated approach is to use an HRF consisting of a difference of two Gamma densities, one modeling the undershoot (Friston et al., 1998). Finally, $\varepsilon_{v,i}$ is the measurement error. Modeling the inherent spatial and temporal characteristics may be accomplished by making appropriate distributional assumptions for the $\varepsilon_{v,i}$ and through the choice of prior distributions for the parameters.

Detecting neuronal activation is now equivalent to detecting whether or not the coefficient $\beta_v$ is nonzero, a variable selection problem (George and McCulloch, 1993, 1997). Others have taken similar approaches (Smith and Fahrmeir, 2007; Smith et al., 2003) by modeling spatial association between neighboring voxels, but temporal correlation between the time series for each voxel is often ignored in the interests of computational efficiency. However, in Section 3 we show that this approach leads to lower quality inference. On the other hand, if one proceeds with a naive fully Bayesian approach which incorporates both spatial and temporal dependence it is very easy to encounter posteriors that are computationally intractable even with the most sophisticated modern computational methods (see e.g. Lee, 2010). Thus we are faced with a situation where computational issues drive some of the modeling decisions. In Section 2 we develop our model and pay careful attention to the prior specifications with the dual goals that we accurately reflect the nature of the experiment while mitigating the computational burden.

The posterior based on our prior specifications is typically of an extremely large dimension and is unavailable in closed form. We focus on the use of Markov chain Monte Carlo (MCMC) throughout to perform the required inferential tasks, but MCMC has its challenges. In particular, the high dimension of the posterior means that convergence is essentially impossible to verify using standard diagnostic methods. We derive a general MCMC algorithm which gives asymptotically
correct answers and we investigate the effect of using several different priors on the performance of the MCMC algorithms and the resulting posterior inference.

We develop the Bayesian models and required MCMC algorithms in Section 2 and in Section 3 a simulation study is undertaken to validate the model and estimation procedure. Finally, we present the results of applying our methods to a particular data set in Section 4. Notably, we show how our model can allow for the task-related change in the BOLD signal to change dynamically over the scanning session. In this way, the model accounts for potential learning effects and other mechanisms of temporal drift in task-related signals. In the rest of this section we describe the fMRI experiment and resulting data that motivated our research and which is analyzed in Section 4.

1.1 Experimental fMRI Data

The experimental data are part of a longitudinal study of Alzheimer’s disease (AD) and its correlates. The study has been conducted in two waves of fMRI data collection with a third under way. All subjects provided informed consent, and the study was approved by the Johns Hopkins Medical Institutions Institutional Review Board. We investigate a color/word Stroop paradigm (Stroop, 1935) implemented on a subset of the subjects in the second wave of data collection. To investigate normative values, we only consider right-handed controls in this study, as establishing normative activation is a necessary first step in producing biomarkers. All subjects are older, generally well-educated, and healthy. None of them have any clinically diagnosed neurologic disorders, including AD.

The Stroop task has been used for many years as a test that exploits the conflict between one well-learned or automatic behavior (e.g. reading) and a decision rule that requires this behavior be inhibited. Many previous behavioral studies have established the features of the Stroop task that produce cognitive interference while neuroimaging studies (Bench et al., 1993; Carter et al., 1995; Fisher et al., 1990; Li et al., 2009; Polk et al., 2008; Taylor et al., 1997) have indicated that several brain regions are involved in the performance of the Stroop task, although these imaging studies do not all agree on which brain areas are most centrally involved in resolving Stroop inference.

Each subject in the current investigation performed the following Stroop task: subjects are shown words and subsequently asked to press a button corresponding to the color of the ink when the word is shown. There are three different components of the task:

- **Ink** only; the word is “XXXX” in colored ink; for example, when presented with, XXXX, the subject will be expected to press the button for the color red.

- **Congruence**; the word is the color of the ink; for example, when presented with BLUE; the subject will be expected to press the button for the color blue.
• **Interference**; the word is a different color from the color of the ink; for example, when presented with BLUE, the subject will be expected to press the button for the color red.

This task considers an important cognitive mechanism; specifically, directed attention. Since most people are proficient at reading words, especially so in the highly educated sample under consideration, it takes inhibitory effort to ignore the word and concentrate on the color of the ink. Also, there is a well-documented age effect. That is, younger adults will typically experience a smaller interference effect than older adults. Additionally, interference effects may decrease with practice (Davidson et al., 2003). In this study all subjects are older so we will not model an age effect, but it may be of interest to describe any temporal drift in the response over the scanning session.

The Stroop exam was administered in the scanner in a block design (see e.g. Friston et al., 2007) with a scanning time repetition of two seconds. The battery of Ink, Congruence and Interference were tasks repeated in sequence 3 times with observations taken at a total of 465 time points. We used standard fMRI preprocessing techniques, including slice timing (aligning axial slices via interpolation from the actual acquisition time), coregistration (using affine transformations to spatially normalize subjects images) and spatial smoothing. Data were then registered in standardized space. Template space is $79 \times 95 \times 68$-dimensional with a voxel size of $2\text{mm}^3$.

Imaging data were subsequently masked, vectorized and stacked into a subject-specific matrix of time by voxels that represents our basic analytic structure. The mask was retained for backtransformation and visualization of results.

2 Variable Selection in a Spatio-Temporal Model

It is reasonable to expect that voxels are spatially dependent—voxels close together should behave similarly—and temporally dependent—the nature of the BOLD signal suggests that voxels adjacent in time will have correlated responses and this effect will persist over longer time intervals than the scanning repetition time (Lund et al., 2006; Woolrich et al., 2001; Worsley et al., 2002). The model we develop below incorporates both of these physical characteristics.

Recall that for voxel $v = 1, \ldots, N$ the BOLD image intensity at time $i = 1, \ldots, T_v$ is $y_{v,i}$ and set $y_v = (y_{v,1}, \ldots, y_{v,T_v})^T$. Let $X_v$ be a known $T_v \times p$ matrix of full column rank and $\beta_v = (\beta_{v,1}, \ldots, \beta_{v,p})^T$ be a $p \times 1$ vector of regression coefficients. We assume

$$y_v = X_v \beta_v + \varepsilon_v, \quad \varepsilon_v \sim N_{T_v}(0, \sigma^2_v \Lambda_v).$$

(1)

Consider the structure of the design matrix $X_v$. We can account for long memory trends by introducing factors accounting for hardware related low-frequency drift, residual movement effects,
and aliased physiological noise such as respiration and cardiac pulsation (Friston et al., 2007; Lund et al., 2006). Also, there are columns for each task corresponding to the convolution of the canonical HRF (Friston et al., 2007) with an impulse stimulus function (Lindquist et al., 2009; Rajapakse et al., 1998).

Our goal is detecting neuronal activation in a voxel which corresponds to identifying nonzero $\beta_v$. Let $\gamma_v = (\gamma_{v,1}, \ldots, \gamma_{v,p})^T$ be binary random variables used to indicate whether the voxel is activated by a sequence of input stimuli. That is, the coefficient $\beta_{v,j}$ is equal to zero if $\gamma_{v,j} = 0$ and $\beta_{v,j}$ is nonzero if $\gamma_{v,j} = 1$. The zero of $\gamma_{v,j}$ implies no effect on voxel $v$ is caused by the corresponding experimental task $j$. Therefore, the model in (1) can be written as

$$y_v = X_v(\gamma_v)\beta_v(\gamma_v) + \varepsilon_v, \quad \varepsilon_v \sim N_{T_v}(0, \sigma_v^2 \Lambda_v),$$

where $\beta_v(\gamma_v)$ is the vector of nonzero regression coefficients and $X_v(\gamma_v)$ is the corresponding matrix for a given indicator variable $\gamma_v$. That is, $X_v(\gamma_v)$ includes only the columns of $X_v$ corresponding $\beta_{v,j} \neq 0$, $j = 1, \ldots, p$.

We now turn our attention to specification of prior distributions. First consider the prior on $\beta_v(\gamma_v)$ given $\gamma_v$ for a particular voxel, $v$. We will use Zellner’s $g$-prior (Zellner, 1996). This prior depends on a parameter typically denoted $g$ which effectively controls model selection in that large values often lead to models with only a few large coefficients while small values tend to produce saturated models (Fernández et al., 2000; George, 2000; Liang et al., 2008). We set this parameter equal to $T_v$ yielding the unit information prior which leads to results similar to those if BIC were used. Thus our prior is given by

$$\beta_v(\gamma_v)|y_v, \sigma_v^2, \Lambda_v, \gamma_v \sim N(\hat{\beta}_v(\gamma_v), T_v\sigma_v^2[X_v^T(\gamma_v)\Lambda_v^{-1}X_v(\gamma_v)]^{-1}),$$

where

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

The prior is data-based because the mean depends on $y_v$. This prior often leads to simpler computation than alternative priors. For example, we also investigated the use of multivariate $t$ distributions, but these did not substantively change the results while making the required computation much more challenging (Lee, 2010).

Next we assume the $\sigma_v^2$ are independent and

$$\pi(\sigma_v^2) \propto \frac{1}{\sigma_v^2}.$$  

Note that $\Lambda_v$ allows us to account for the temporal dependence between observations on a given voxel. We will see that the priors for $\Lambda_v$ are critically important to both inferential efficacy and
computational efficiency. Accordingly we will investigate the use of several priors. The first prior assumes no temporal dependence (i.e. $\Lambda_v$ is an identity matrix), which is a common assumption made to achieve computational efficiency (cf. Genovese, 2000; Smith and Fahrmeir, 2007; Smith et al., 2003).

While images are collected every 2 seconds the BOLD signal increases above baseline about 2 seconds after the onset of neuronal activity, peaking at about 5-8 seconds and falling below baseline for about 10 seconds. In addition to the persistence of neuronal activation it has been found that other cyclical neuronal events and artifacts of the measurement process can be responsible for temporal autocorrelation in fMRI settings (see e.g. Locascio et al., 1997). All of this suggests that an autoregressive (AR) or autoregressive moving average (ARMA) dependence might be sensible, but moving average processes, such as MA(1) or MA(2) processes, will be less desirable (see among many others Friston et al., 1995; Lindquist et al., 2009; Xia et al., 2009). In particular, we will focus on using AR(1) dependence so the $(i, j)$ th element of $\Lambda_v$ is $\Lambda_v(i, j) = \rho_v^{i-j}$. We will consider the use of two priors for $\rho$. The first assumes the components of $\rho$ to be independently uniformly distributed between -1 and 1, that is,

$$\pi(\rho) = \prod_{v=1}^N \pi(\rho_v) \propto \prod_{v=1}^N I(-1 \leq \rho_v \leq 1).$$

The second prior we investigate will be based on an empirical Bayes (EB) approach where $\rho$ is estimated with $\hat{\rho}$, the maximum likelihood estimator. This type of EB approach is often referred to as a two-stage solution or prewhitening.

We have investigated the use of other structures such as AR(2), ARMA(1,1) and MA(1) dependence, but using the EB approach with the AR(1) structure seems to be an effective compromise between inferential efficacy and computational efficiency. Indeed the robustness of our method is investigated in Section 3 while we compare the results in our real data application to that of using an empirical Bayes method (all parameters estimated with maximum likelihood) assuming AR(2), ARMA(1,1) and MA(1) dependence in Section 4.

All that remains is to specify a prior for the indicator variables $\gamma_v$. We will use a binary spatial Ising prior that allows us to incorporate anatomical prior information as well as spatial interaction between voxels. Let $\gamma(j) = \{\gamma_{1,j}, \ldots, \gamma_{N,j}\}$ be the vector of indicator variables for regression $j$ over locations $\{1, 2, \ldots, N\}$. We begin by addressing how the spatial interaction is handled. Let $\omega_v,k$ be prespecified constants that allow us to weigh the interaction between neighboring locations on lattices $v$ and $k$ and let $\theta_j$ be the positive parameter to represent the strength of the interaction between any two voxels. If two voxels $v$ and $k$ are neighbors, then we write $v \sim k$ and the spatial interaction is described linearly

$$\theta_j \sum_{v \sim k} \omega_v,k I(\gamma_v,j = \gamma_k,j).$$
The neighborhood structure is defined by the user. Commonly used neighborhood structures are based on the four, eight, or twelve nearest neighbors; see Figure 1. For example, in the simulation study of Section 3 we employ a two-dimensional neighborhood of \( v \) that is defined to contain the directly adjacent vertical and horizontal voxels \( k \). In our main application in Section 4 we employ a three-dimensional neighborhood that also includes the nine voxels immediately above, and nine voxels immediately below, voxel \( v \). However, in any application the user should consider whether it may be more fruitful to consider cliques of higher orders instead of continuing the theme of pairwise interactions only; see Tjelmeland and Besag (1998) where several interesting experiments are conducted. We set the weights to be the reciprocal of the distance between voxel \( v \) and each of the neighboring voxels, but we also assess the robustness of this choice in Section 3.

Next we specify an “external field” which is meant to incorporate anatomical prior information in a linear combination of parameters

\[
\sum_{v=1}^{N} \alpha_{v,j} \gamma_{v,j}.
\]

The prior on \( \gamma \) is then taken to be

\[
\pi(\gamma|\theta) = \prod_{j=1}^{p} \pi(\gamma(j)|\theta_j),
\]

where

\[
\pi(\gamma(j)|\theta_j) \propto \exp \left\{ \sum_{v=1}^{N} \alpha_{v,j} \gamma_{v,j} + \theta_j \sum_{v \sim k} \omega_{v,k} I(\gamma_{v,j} = \gamma_{k,j}) \right\},
\]

where the \( \alpha_{v,j} \) are chosen to reflect prior knowledge.

Figure 1: Common neighborhood structures in imaging analysis.
Smith and Fahrmeir (2007) consider the problem of incorporating external or anatomical information into the binary spatial Ising prior. We consider their approach in the current setting. To begin we will simplify the presentation by dropping the subscript \( j \) so that we are considering how to choose the parameters \( \alpha_v \) in

\[
\pi(\gamma|\theta) \propto \exp \left\{ \sum_{v=1}^{N} \alpha_v \gamma_v + \theta \sum_{v \sim k} \omega_{v,k} I(\gamma_v = \gamma_k) \right\}.
\]

Now suppose that activation occurs only in some region \( G \) and introduce a indicator \( I_v \) to denote if the voxel \( v \) is in the area \( G \), that is,

\[
I_v = \begin{cases} 
1, & v \in G; \\
0, & \text{otherwise}.
\end{cases}
\]

Since activation only occurs in \( G \) it is reasonable to assume the marginal probability \( p(\gamma_v = 1) = 0 \) where \( v \notin G \). On the other hand, we assume \( p(\gamma_v = 1|I_v) = b_v, v \in G \) where \( b_v \) is specified by the user to reflect prior (including anatomical) knowledge. Further assume that the indicators are a priori independent across voxels. If voxel \( v \in G \), we have

\[
p(\gamma_v = 1) = p(\gamma_v = 1|I_v = 1)p(I_v = 1) = b_v \times p(I_v = 1) := e_v.
\]

We match the marginal prior probabilities \( e_v \) to the external field of the Ising prior when there is no spatial correlation, so that \( \theta = 0 \). In this case the joint density is

\[
\pi(\gamma|\theta = 0) = \exp \left\{ \sum_{v=1}^{N} \alpha_v \gamma_v \right\},
\]

with marginals

\[
\pi(\gamma_v = 1) = \frac{\exp\{\alpha_v\}}{1 + \exp\{\alpha_v\}} = \text{set} e_v, \quad v \in G.
\]

Therefore, an anatomically informed Ising prior is given by setting

\[
\alpha_v = \log \frac{e_v}{1 - e_v}.
\]

However, the drawback of using non-zero external field coefficients in the Ising prior is that \( \theta \) is no longer simply a smoothing parameter. Therefore, we could employ a two-step procedure to address the relationship between \( \theta \) and the marginal probability of activation with non-zero \( \alpha_v \):

**Estimation** Set \( \alpha_1 = \cdots = \alpha_N = 0 \), estimate and obtain \( \hat{\theta} = E(\theta|y) \).

**Refitting** Refit with an anatomically informed Ising prior, that is obtain \( \alpha_1, \ldots, \alpha_N \) as described above, but conditional on a fixed level of smoothing \( \hat{\theta} = E(\theta|y) \).
This leaves only the distribution of the parameter $\theta$ to address. We assume a uniform prior on $\theta = (\theta_1, \ldots, \theta_p)$, i.e., $\pi(\theta) \propto \prod_{j=1}^{p} I(0 < \theta_j < \theta_{\text{max}})$, where $\theta_{\text{max}}$ is a user-specified hyperparameter. In practice, $\theta_{\text{max}} \leq 2$ often suffices; see Møller (2003).

Note that $\theta$, $\rho$, and $\sigma^2$ are \textit{a priori} independent, $\gamma$ conditionally independent, and independence across voxels. Thus the posterior density is characterized by

$$q(\beta(\gamma), \gamma, \rho, \sigma^2|y) \propto p(y|\beta(\gamma), \gamma, \sigma^2, \Lambda_v) \pi(\beta(\gamma)|\sigma^2, \Lambda, \gamma) \pi(\gamma|\theta) \pi(\rho) \pi(\sigma^2) \pi(\theta)$$  \hspace{1cm} (6)

$$\propto \prod_{v=1}^{N} \left[ p(y_v|\beta_v(\gamma_v), \gamma_v, \sigma^2_v, \Lambda_v) \pi(\beta_v(\gamma_v)|y_v, \gamma_v, \sigma^2_v, \Lambda_v) \pi(\sigma^2_v) \right] \pi(\gamma|\theta) \pi(\rho) \pi(\theta) .$$

\subsection{2.1 Posterior Inference}

The posterior quantities of interest are the activation probability and its magnitude, that is, $q(\gamma_{v,j} = 1|y)$ and $E(\beta_v|y)$, respectively. These quantities are analytically intractable and must be approximated with Monte Carlo methods. We will describe a particular MCMC method in the subsequent section. A naive approach would be to construct an MCMC sampler having the full posterior $q(\beta(\gamma), \gamma, \rho, \sigma^2|y)$ as the invariant density. However, this would be computationally prohibitive. An alternative is to use Rao-Blackwellization. Note that

$$E(\beta_v|y) = \sum_{\gamma_v} E(\beta_v|\gamma_v, y) q(\gamma_v|y)$$

and if $\gamma_{-(v,j)}$ denotes the vector of binary regressors excluding $\gamma_{v,j}$, then

$$q(\gamma_{v,j} = 1|y) = \int q(\gamma_{v,j} = 1|\rho_v, \gamma_{-(v,j)}, y) q(\rho_v|y) q(\gamma_{-(v,j)}|y) d\rho_v d\gamma_{-(v,j)} .$$

Thus we only need an MCMC sampler whose invariant distribution is $q(\gamma, \rho|y)$. The MCMC algorithm will be described in the next section. For now, suppose $\{(\gamma_1^{[1]}, \rho_1^{[1]}), (\gamma_2^{[2]}, \rho_2^{[2]}), \ldots, (\gamma_K^{[K]}, \rho_K^{[K]})\}$ is a Monte Carlo sample generated with our MCMC algorithm. Then we can approximate the posterior quantities with

$$E(\beta_v|y) \approx \frac{1}{K} \sum_{k=1}^{K} \hat{\beta}_v(\gamma_v^{[k]})$$  \hspace{1cm} (7)

and

$$q(\gamma_{v,j} = 1|y) \approx \frac{1}{K} \sum_{k=1}^{K} q(\gamma_{v,j} = 1|\gamma_{-(v,j)}^{[k]}, \rho_v^{[k]}, y) := \hat{q}(\gamma_{v,j} = 1|y) .$$  \hspace{1cm} (8)

How large should $K$ be? There are two common approaches to choosing a value of $K$. The first is to fix a value of $K$ before the simulation begins; that is, a fixed-time rule. The disadvantage of such
a rule is that the user then has no control over the Monte Carlo error in the estimates. Another approach is to base the stopping rule on a fixed-width approach which continues the simulation until the Monte Carlo standard error (MCSE) is sufficiently small; see Flegal and Gong (2013), Flegal and Jones (2011), Flegal et al. (2008), and Jones et al. (2006). We will use the method of batch means to calculate asymptotically valid MCSEs.

We need a threshold in order to use \( \hat{q} \) to detect activation. Several authors have suggested using 0.8722, which we will use. That is, an individual voxel is categorized as active if \( \hat{q} > 0.8722 \), otherwise it is considered inactive. Following Raftery (1996), Smith and Fahrmeir (2007) give a clear description of the motivation for this value in the context of a Bayesian spatial model. In Section 3 we compare this threshold value with some other possibilities and show that it is an effective choice.

### 2.2 Bayesian Inference via MCMC Sampling

We need to construct an MCMC sampler which has \( q(\gamma, \rho | y) \) as the invariant density. Unfortunately, this density is not available in closed form since the integral

\[
\int \pi(\gamma | \theta) \pi(\theta) \, d\theta
\]

is analytically intractable. However, it will suffice to create an algorithm having \( q(\gamma, \rho, \theta | y) \) as the invariant density. Let

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

Let \( q_v = \sum_{j=1}^p \gamma_{v,j} \). In Appendix A we show that

\[
q(\gamma, \rho, \theta | y) \propto \pi(\gamma | \theta) \pi(\theta) \prod_{v=1}^N \left(1 + T_v\right)^{-q_v/2} |\Lambda_v|^{-1/2} S(\rho_v, \gamma_v)^{-T_v/2} \cdot
\]

As the target posterior may be truly high-dimensional it is natural to use a component-wise strategy (Johnson et al., 2013). In this case we need the three conditional densities \( q(\gamma | \rho, \theta, y) \), \( q(\rho | \gamma, \theta, y) \) and \( q(\theta | \gamma, \rho, y) \). These are characterized by

\[
q(\gamma | \rho, \theta, y) \propto \pi(\gamma | \theta) \prod_{v=1}^N \left(1 + T_v\right)^{-q_v/2} |\Lambda_v|^{-1/2} S(\rho_v, \gamma_v)^{-T_v/2}
\]

\[
q(\rho | \gamma, \theta, y) = q(\rho | \gamma, y) \propto \pi(\rho) \prod_{v=1}^N |\Lambda_v|^{-1/2} S(\rho_v, \gamma_v)^{-T_v/2}
\]

\[
q(\theta | \gamma, \rho, y) = q(\theta | \gamma, y) \propto \pi(\gamma | \theta) \pi(\theta) \cdot
\]

Note that, given \( \gamma \), \( \theta \) and \( \rho \) are \textit{a posteriori} independent. Let \( \xi = (\theta, \rho) \). Then we can set up a two-variable component-wise sampler that updates \( \gamma \) followed by \( \xi \); that is, if we let \( (\gamma, \xi) \) be the
current state and \((\gamma', \xi')\) be the future state, one step of the MCMC sampler is the composition of two steps and looks like \((\gamma, \xi) \rightarrow (\gamma', \xi) \rightarrow (\gamma', \xi')\).

**Step 1.** Consider updating \(\gamma\) conditional on the values of \(\theta\) and \(\rho\). We will use a component-wise Metropolis-Hastings method. Schematically the transition \(\gamma \rightarrow \gamma'\) consists of \(Np\) steps and will look like

\[
(\gamma_{1,1}, \gamma_{2,1}, \ldots, \gamma_{N,1}, \ldots, \gamma_{1,p}, \gamma_{2,p}, \ldots, \gamma_{N,p}) \rightarrow (\gamma'_{1,1}, \gamma'_{2,1}, \ldots, \gamma'_{N,1}, \ldots, \gamma_{1,p}, \gamma_{2,p}, \ldots, \gamma_{N,p})
\]

\[
\rightarrow (\gamma'_{1,1}, \gamma'_{2,1}, \ldots, \gamma'_{N,1}, \ldots, \gamma'_{1,p}, \gamma'_{2,p}, \ldots, \gamma'_{N,p})
\]

\[
\rightarrow (\gamma'_{1,1}, \gamma'_{2,1}, \ldots, \gamma'_{N,1}, \gamma'_{1,p}, \gamma'_{2,p}, \ldots, \gamma'_{N,p})
\]

Thus we need the density \(q(\gamma_{v,j} | \gamma_{-(v,j)}, \theta, \rho, y)\). Notice that

\[
q(\gamma_{v,j} | \gamma_{-(v,j)}, \theta, \rho, y) \propto (1 + T_v)^{-\gamma_{v,j}/2} S(\rho_v, \gamma_v) T_v^{v,j} / 2 \exp \left\{ \alpha_{v,j} \gamma_{v,j} + \theta_j \sum_{k \in \delta_v} \omega_{v,k} I(\gamma_{k,j} = \gamma_{v,j}) \right\}
\]

Now let \(p_{\gamma_{v,j}}(\gamma_{-(v,j)} | \theta, \rho)\) be a user-specified proposal distribution. Then one of the \(Np\) steps occurs as follows. Let the current state of the Markov chain be

\[
\gamma = (\gamma'_{1,1}, \ldots, \gamma'_{v,j-1}, \gamma_{v,j}, \gamma_{v,j+1}, \ldots, \gamma_{N,p})
\]

Then draw proposal \(\gamma_{v,j}^* \sim p_{\gamma_{v,j}}(\gamma_{-(v,j)} | \theta, \rho)\). Set \(\gamma'_{v,j} = \gamma_{v,j}^*\) with probability the minimum of 1 and the Hastings ratio, otherwise set \(\gamma'_{v,j} = \gamma_{v,j}\). The Hastings ratio is

\[
\frac{(1 + T_v)^{-\gamma_{v,j}/2} S(\rho_v, \gamma_{v,j}) T_v^{v,j} / 2 \exp \left\{ \alpha_{v,j} \gamma_{v,j} + \theta_j \sum_{k \in \delta_v} \omega_{v,k} I(\gamma_{k,j} = \gamma_{v,j}^*) \right\}}{(1 + T_v)^{-\gamma_{v,j}/2} S(\rho_v, \gamma_v) T_v^{v,j} / 2 \exp \left\{ \alpha_{v,j} \gamma_{v,j} + \theta_j \sum_{k \in \delta_v} \omega_{v,k} I(\gamma_{k,j} = \gamma_{v,j}) \right\}} \times \frac{p_{\gamma_{v,j}}(\gamma_{v,j} | \gamma_{-(v,j)}, \theta, \rho)}{p_{\gamma_{v,j}}(\gamma_{v,j}^* | \gamma_{-(v,j)}, \theta, \rho)}.
\]

Notice that no matter which proposal distribution \(p_{\gamma_{v,j}}\) we use, if \(\gamma_{v,j}^* = \gamma_{v,j}\), then the Hastings ratio is 1 and we automatically set \(\gamma'_{v,j} = \gamma_{v,j}^*\).

One possible choice for the proposal distribution is

\[
p_{\gamma_{v,j}}(\gamma_{v,j} | \gamma_{-(v,j)}, \theta, y) = \frac{1}{Z(\theta)} \exp \left\{ \alpha_{v,j} \gamma_{v,j} + \theta_j \sum_{k \in \delta_v} \omega_{v,k} I(\gamma_{k,j} = \gamma_{v,j}) \right\}
\]

from which it can be shown that

\[
p_{\gamma_{v,j}}(\gamma_{v,j} = 1 | \gamma_{-(v,j)}, \theta, y) = \left[ 1 + \exp \left\{ -\alpha_{v,j} + \theta_j \sum_{k \in \delta_v} \omega_{k,v} (1 - 2\gamma_{k,j}) \right\} \right]^{-1}.
\]
This leads to a simplification in the Hastings ratio
\[
\frac{(1 + T_v)^{-\gamma_{v,j}/2}S(\rho_v, \gamma_{-(v,j)}, \gamma^*_{v,j})^{T_v/2}}{(1 + T_v)^{-\gamma_{v,j}/2}S(\rho_v, \gamma^*_{v,j})^{T_v/2}}.
\]

**Step 2.** Consider updating \(\rho\) conditional on \(\gamma\). In this step we again use a component-wise MCMC algorithm. Schematically the transition \(\rho \rightarrow \rho'\) consists of \(N\) steps and will look like
\[
(\rho_1, \rho_2, \ldots, \rho_{N-1}, \rho_N) \rightarrow (\rho'_1, \rho_2, \ldots, \rho_{N-1}, \rho_N)
\]
\[
\quad \rightarrow (\rho'_1, \rho'_2, \ldots, \rho_{N-1}, \rho_N)
\]
\[
\quad \vdots
\]
\[
\quad \rightarrow (\rho'_1, \rho'_2, \ldots, \rho'_{N-1}, \rho'_N)
\]

Thus we need the density \(q(\rho_v|\rho_{-v}, \gamma, y)\). Notice that
\[
q(\rho_v|\rho_{-v}, \gamma, y) = q(\rho_v|\gamma, y) \propto \pi(\rho_v)|\Lambda_v|^{-1/2}S(\rho_v, \gamma_{v})^{T_v/2}.
\]

Let the current state of the Markov chain be \(\rho_v\). Then generate a proposal \(\rho^*_v\) from the density \(p_{\rho_v}(\cdot|\gamma, y)\) and set \(\rho'_v = \rho^*_v\) with probability the minimum of 1 and the Hastings ratio which is given by
\[
\frac{\pi(\rho^*_v)|\Lambda^*_v|^{-1/2}S(\rho^*_v, \gamma_{v})^{T_v/2} p_{\rho_v}(\rho_v|\gamma, y)}{\pi(\rho_v)|\Lambda_v|^{-1/2}S(\rho_v, \gamma_{v})^{T_v/2} p_{\rho_v}(\rho^*_v|\gamma, y)}.
\]

Otherwise, set \(\rho'_v = \rho_v\). One possible choice for the proposal density is a Uniform\((-1, 1)\).

Notice that when we use the EB approach for \(\rho\) we avoid this step entirely.

**Step 3.** Consider updating \(\theta\) conditional on \(\gamma\). Again we use a component-wise MCMC method. Schematically the transition \(\theta \rightarrow \theta'\) consists of \(p\) steps and looks like
\[
(\theta_1, \theta_2, \ldots, \theta_{p-1}, \theta_p) \rightarrow (\theta'_1, \theta_2, \ldots, \theta_{p-1}, \theta_p)
\]
\[
\quad \rightarrow (\theta'_1, \theta'_2, \ldots, \theta_{p-1}, \theta_p)
\]
\[
\quad \vdots
\]
\[
\quad \rightarrow (\theta'_1, \theta'_2, \ldots, \theta'_{p-1}, \theta'_p)
\]

Thus we need the density \(q(\theta_j|\theta_{-j}, \gamma, y)\). Notice that
\[
q(\theta_j|\theta_{-j}, \gamma, y) = q(\theta_j|\gamma, y) \propto Z_j^{-1}(\theta_j, \alpha_j) \exp \left\{ \theta_j \sum_{v \sim k} \omega_{v,k} I(\gamma_{v,j} = \gamma_{k,j}) \right\} I(0 < \theta_j < \theta_{\text{max}})
\]
where
\[ Z_j(\theta_j, \alpha_j) = \left[ \sum_{\gamma(j)} \exp \left\{ \sum_{v=1}^N \alpha_{v,j} \gamma_{v,j} + \theta_j \sum_{v \sim k} \omega_{v,k} I(\gamma_{v,j} = \gamma_{k,j}) \right\} \right] \]

Let the current state of the Markov chain be \( \theta_j \). Then generate a proposal \( \theta^*_j \) from a proposal density \( p_{\theta_j}(\cdot | \gamma, y) \). Then set \( \theta'_j = \theta^*_j \) with probability the minimum of 1 and the Hastings ratio

\[
\frac{Z_j(\theta_j, \alpha_j) \exp \left\{ \theta^*_j \sum_{v \sim k} \omega_{v,k} I(\gamma_{v,j} = \gamma_{k,j}) \right\} I(0 < \theta^*_j < \theta_{\text{max}}) p_{\theta_j}(\theta_j | \gamma, y)}{Z_j(\theta^*_j, \alpha_j) \exp \left\{ \theta_j \sum_{v \sim k} \omega_{v,k} I(\gamma_{v,j} = \gamma_{k,j}) \right\} I(0 < \theta_j < \theta_{\text{max}}) p_{\theta_j}(\theta_j | \gamma, y)}
\]

The ratio
\[
\frac{Z_j(\theta_j, \alpha_j)}{Z_j(\theta^*_j, \alpha_j)}
\]
is analytically intractable but can be estimated with path sampling (Gelman, 1998) or the Wang-Landau algorithms (Wang and Landau, 2001; Zhang and Ma, 2007), among others.

One choice for a proposal distribution is \( N(\theta_j, \hat{\sigma}^2) \) where \( \theta_j \) is the current state and \( \hat{\sigma} \) will be tuned so that the acceptance rate is roughly 40%.

### 3 Simulation Study

In this section we report the results of a simulation study undertaken to validate the model and estimation procedure described above. In part this is done to investigate what differences arise when different priors are used for \( \rho \). Specifically, we consider the 3 cases where (i) \( \rho_v \sim \text{Uniform}(-1, 1) \), independently; (ii) an empirical Bayes approach is used to estimate \( \rho \) with \( \hat{\rho} \), which is calculated by maximum likelihood; and (iii) taking \( \Lambda_v \) to be the identity matrix for all \( v \). In the second part of this section we continue to use the EB approach and investigate the robustness of our choice for the weights \( \omega_{v,k} \) in our spatial prior and our choice of .8722 for the cutoff for the posterior activation probabilities. The final part of this section is concerned with investigating the robustness of using either (i) an empirical Bayes approach is used to estimate \( \rho \) with \( \hat{\rho} \) or (ii) taking \( \Lambda_v \) to be the identity matrix for all \( v \) when the underlying data is generated with various temporal correlation structures.

We generate 10 data sets from the model based on \( 30 \times 30 \) activated-inactivated images. We do a posteriori inference on these data sets using the full posterior defined at (6). We then compare the various methods based on their ability to estimate the spatial interaction parameter \( \theta \) and their ability to correctly classify voxels as activated or inactivated.
We now describe how we generate the data. Suppose \( \sigma_v = \sigma^2 \) for each \( v = 1, \ldots, N \) and set \((\theta, \sigma^2) = (.7, 3)\). Given \( \theta \), a \( \gamma \) is exactly generated from 
\[
p(\gamma|\theta) = \frac{1}{Z(\theta)} \exp \left\{ \theta \sum_{i\sim j} \omega_{i,j} I(\gamma_i = \gamma_j) \right\}.
\]
Simulating from \( p(\gamma|\theta) \) is not easy but can be done using a perfect sampling technique for Ising models (Propp and Wilson, 1996). This \( \gamma \) is a 30 \times 30 activated-inactivated square image. Given \( \gamma \), we simulate a time-series \( y_v \) in each voxel \( v \) of length 100 from the model 
\[
y_v = X_v(\gamma_v) \beta_v(\gamma_v) + \varepsilon_v, \quad \varepsilon_v \sim N_{100}(0, \sigma_v^2 \Lambda_v).
\]
We assume a 'boxcar' type stimulus function and after convolving with the canonical HRF the design matrix is given in the Figure 2.

![Figure 2](image-url)

**Figure 2:** The design matrix used in the simulation study.

The autoregression coefficient, \( \rho_v \), is generated from Uniform\((-1, 1)\) for each voxel. Let \( \beta_v = (\beta_{v,0}, \beta_{v,1})^T \) where \( \beta_{v,0} \) represents the baseline level and \( \beta_{v,1} \) describes the amplitude of activation in response to a stimulus at each voxel \( v \). We use \( \gamma_{v,j} \) to indicate if \( \beta_{v,j} \) is equal to 0 or not, that is, \( \beta_{v,j} \neq 0 \) if \( \gamma_{v,j} = 1 \); otherwise \( \beta_{v,j} = 0 \). In this simulation, we always assume \( \gamma_{v,0} = 1 \) since \( \beta_{v,0} \) models the baseline level in a human brain. On the other hand, \( \gamma_{v,1} \) can be either 0 or 1. When \( \gamma_{v,1} = 1 \), we set \( \beta_v = (300, 5)^T \); otherwise \( \beta_v = \beta_{v,0} = 300 \). The reason we set \( \beta_{v,1} = 5 \) is that the BOLD contrast is typically fairly small, with activation inducing a signal increase ranging from 1% to 5%.

For each simulated data, the spatial Bayesian variable selection approach is applied to detect the activation and to estimate the spatial coefficient \( \theta \) via its posterior mean. Recall the prior
Table 1: The average Monte Carlo estimates of $\theta$ and corresponding Monte Carlo standard errors (MCSE) based on $1 \times 10^4$ MCMC iterations for each of the 10 simulated data sets.

<table>
<thead>
<tr>
<th>Prior for $\rho$</th>
<th>Estimate $\hat{\theta}$</th>
<th>MCSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform($-1,1$)</td>
<td>0.73</td>
<td>0.0017</td>
</tr>
<tr>
<td>EB</td>
<td>0.74</td>
<td>0.0015</td>
</tr>
<tr>
<td>$\Lambda_v = I_{100}$</td>
<td>0.78</td>
<td>0.0015</td>
</tr>
</tbody>
</table>

for $\beta$ is given in (2). Now consider the binary spatial Ising prior (5). We used a two-dimensional neighborhood which contains the directly adjacent voxel and the horizontal voxels. The weights $\omega_{v,k}$ were taken to be the reciprocal of the distance between voxels $v$ and $k$. Finally, we set $\theta_{\max} = 2$ and since there is no anatomical information in this problem we set $\alpha = 0$. To evaluate the performance of this method, we consider the accuracy of estimating $\theta$ with its posterior mean and the accuracy and false positive rate of identifying activation-inactivation. Note that we classified a voxel as active if $\hat{q} > 0.8722$. Accuracy is defined as the percentage of voxels correctly classified. The false positive rate is the percentage of active voxels falsely identified.

The average estimates of $\theta$ and corresponding Monte Carlo standard errors (MCSEs) over 10 simulated data are given in Table 1. Recall that the true value for $\theta$ is 0.7. The estimation results are based on $1 \times 10^4$ draws from the posterior using our MCMC algorithm, resulting in all MCSEs being less than 0.005. The results suggest that using a prior which does not include temporal correlation (i.e. $\Lambda_v = I_{100}$) will lead to overestimates of the spatial correlation but there is little difference between either of the other two priors.

Table 2 reports the average (over times and data sets) accuracy and false positive rate for each of the 3 priors. Again we find that ignoring the temporal correlation will lead to inferior inference. Specifically, when $\Lambda_v = I_{100}$ we observed a higher false positive rate. However, the other two priors produce comparable results. It is worth emphasizing that the computation time required for the EB approach is over 300 times less than that required for the Uniform($-1,1$) prior; note that with the EB method Step 2 of the MCMC algorithm is superfluous so inversion of $\Lambda_v$ for each voxel is avoided.

We next use the EB method for estimating $\rho$ and assess the robustness of the procedure to our choice of weights $\omega_{v,k}$ and activation cutoff $\hat{q} > 0.8722$.

Consider our choice for the weights $\omega_{v,k}$, which is reciprocal of Euclidian distance between voxels. We tried half and double these values of $\omega_{v,k}$ in our simulation study. Both show quite similar results to the weights we chose previously. The choice of weights does not appear to significantly impact
Table 2: The accuracy based on $1 \times 10^4$ MCMC iterations for each of the 10 simulated data sets.

<table>
<thead>
<tr>
<th>Prior for $\rho$</th>
<th>$\Lambda_v = I_{100}$</th>
<th>Uniform($-1,1$)</th>
<th>EB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>91.38</td>
<td>97.38</td>
<td>97.16</td>
</tr>
<tr>
<td>False Positive Rate (%)</td>
<td>13.59</td>
<td>0.045</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Table 3: The accuracy and false positive rate on identifying activation corresponding to different weights applied where original means the weights are equal to the reciprocal of the distance between voxels, half for half original weights, double for double original weights.

<table>
<thead>
<tr>
<th>Weight</th>
<th>1/2</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>97.11</td>
<td>97.28</td>
<td>97.25</td>
</tr>
<tr>
<td>False Positive (%)</td>
<td>1.20</td>
<td>1.36</td>
<td>1.34</td>
</tr>
</tbody>
</table>

Table 4: The accuracy of detection of activation using different critical values.

<table>
<thead>
<tr>
<th>Critical value</th>
<th>.7946</th>
<th>.8722</th>
<th>.9650</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>97.44</td>
<td>97.28</td>
<td>95.93</td>
</tr>
<tr>
<td>False Positive (%)</td>
<td>2.30</td>
<td>1.36</td>
<td>0.50</td>
</tr>
</tbody>
</table>

To this point all of our simulated data has followed an autoregressive temporal correlation. We now turn our attention to assessing the robustness of our procedure when this assumption does not hold. Specifically, we simulated 100 images under each of 5 temporal correlation structures—AR(1), AR(2), MA(1), MA(2) and ARMA(1,1). We then fit two models to each data set. One of the models was our EB method assuming an AR(1) temporal correlation and one was a model which
assumed no temporal correlation. The results are presented in Tables 5 and 6. Yet again we see that ignoring the temporal correlation results in inferior inference. It is also clear from these results that the use of an AR(1) structure is quite reasonable in all of these settings.

Table 5: The EB AR(1) model is implemented on 100 images simulated with five different temporal correlation structures. The figures inside of parentheses are the corresponding MCSE.

<table>
<thead>
<tr>
<th>Models</th>
<th>AR(1)</th>
<th>AR(2)</th>
<th>MA(1)</th>
<th>MA(2)</th>
<th>ARMA(1, 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>97.38 (0.97)</td>
<td>96.40 (0.97)</td>
<td>95.98 (0.67)</td>
<td>97.88 (1.75)</td>
<td>96.01 (0.65)</td>
</tr>
<tr>
<td>False Positive Rate (%)</td>
<td>0.68 (0.041)</td>
<td>2.23 (0.077)</td>
<td>0.97 (0.041)</td>
<td>0.98 (0.067)</td>
<td>0.97 (0.043)</td>
</tr>
</tbody>
</table>

Table 6: The model with $\Sigma = I$ is implemented on 100 images simulated with five different temporal correlation structures. The figures inside of parentheses are the corresponding MCSE.

<table>
<thead>
<tr>
<th>Models</th>
<th>AR(1)</th>
<th>AR(2)</th>
<th>MA(1)</th>
<th>MA(2)</th>
<th>ARMA(1, 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>95.11 (1.26)</td>
<td>94.43 (1.47)</td>
<td>98.56 (0.07)</td>
<td>96.88 (0.08)</td>
<td>93.67 (1.65)</td>
</tr>
<tr>
<td>False Positive Rate (%)</td>
<td>8.74 (0.41)</td>
<td>10.22 (0.87)</td>
<td>0.10 (0.004)</td>
<td>6.00 (0.64)</td>
<td>9.79 (1.43)</td>
</tr>
</tbody>
</table>

4 Activation in Experimental fMRI Data

In this section we use the methodology, i.e. the model, estimation procedure and the MCMC sampling method, described earlier to analyze the Stroop data which was described in Section 1.1. Based on the results of Section 3 we will limit attention to using EB methods for $\rho$; recall that we found little difference the quality of inference between the EB approach and the Uniform prior while the computing time for the EB method was over 300 times less. The weights in the binary spatial Ising prior will be given by the reciprocal of the Euclidean distance between neighboring voxels and set $\theta_{\text{max}} = 1$. We used a cutoff of .8772 for detecting activation.

We consider two different models in this section. The first (see Section 4.2) does not allow for changes in the activation patterns over time while the second (see Section 4.3) does. Which model is used should be determined by the inferential goals. If the interest is mainly in which areas are activated by a task, the time-invariant model should be used. But if we want investigate the activation changes over time, the time-varying model is appropriate.
4.1 Design Matrix in the Regression Model

In general, at voxel $v$, the design matrix $X_v$ consists of the effects of no interest, transformed stimuli, and the baseline trend. The transformed stimulus is the convolution of the stimulus function with the assumed HRF, as discussed earlier. In this fMRI experiment, there are 3 types of tasks ("Ink Only", "Congruence", and "Interference") given 3 alternating times. Therefore, this part of the design matrix implemented in this study was obtained by convolving the stimulus function with the canonical HRF. The visualization of the design matrix is given in the Figure 3. The red and black lines in the Figure 3 model the initial effect which is of no interest and the baseline trend during the experiment, respectively. The green, blue, and pink lines model the corresponding effects, "Ink Only", "Congruence”, and ”Interference” on the brain activity during the test.

![Figure 3: The design matrix used for the experimental fMRI data.](image)

4.2 Time-Invarying Activation Patterns

We assume the basic linear model

$$y_v = a_0z_0 + a_1z_1 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \varepsilon_v,$$

where $y_v$ represent a time-series in a particular voxel $v$, each $x_i$ is a transformed input function, $z_i$ is a vector used to remove low-frequency, stimulus independent, effects, the $\beta_i$ are parameters of interest corresponding to different tasks, "Ink Only”, ”Congruence”, and ”Interference”, respectively. The $a_i$ are nonzero nuisance parameters to model the baseline brain signal. The rest of the hierarchical model is specified as described in Section 2 and at the beginning of this section. A sequence of
Table 7: Estimate of the components of $\theta$ based on $1 \times 10^5$ MCMC simulations with MCSE given in the parentheses.

<table>
<thead>
<tr>
<th></th>
<th>$\theta_2$</th>
<th>$\theta_3$</th>
<th>$\theta_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.7476 (0.00006)</td>
<td>0.7520 (0.00006)</td>
<td>0.6181 (0.00007)</td>
</tr>
</tbody>
</table>

binary variables $\gamma_v = \{1, 1, \gamma_3, \gamma_4, \gamma_5, v\}$ is used to indicate if the corresponding parameter is zero or not; recall that we assume both of the $a_i$ are nonzero.

The posterior of interest $q(\gamma, \theta | y)$ is intractable so we will use the MCMC algorithm developed in Section 2.2 to obtain $1 \times 10^5$ observed values of the Markov chain. We use the method of batch means (Jones et al., 2006) to calculate MCSEs for the estimated quantities. We also used trace plots and histograms to visually assess convergence of the simulation.

The estimates of the components of the spatial correlation parameter are given in Table 7. Note that the reported MCSEs are small and hence we have precise estimates. Also, the spatial correlations appear quite large.

In Figure 4 we present some results predicting activation. The parietal and occipital lobes are activated during the Stroop task, since the parietal lobe accounts for cognition, information processing, and visual perception and the occipital lobe is the visual processing center which is responsible for control of vision and color recognition. Additionally, the activation regions in the right brain are larger than those in the left brain partly because the left and right hemispheres of the brain process information in different ways and we tend to process information using our dominant side; recall that all subjects were right-handed.

### 4.3 Time-Varying Activation Patterns

In the fMRI experiment described in Section 1, the three different Stroop tasks are repeated three times. The model at (9) fails to serve the goal of studying a possible activation change over time. Here we generalize the model in order to study possible changes to the activation pattern over time. Suppose,

$$ Y_v = a_0 z_0 + a_1 z_1 + \beta_{1,1} x_{1,1} + \beta_{1,2} x_{1,2} + \beta_{1,3} x_{1,3} + \beta_{2,1} x_{2,1} + \beta_{2,2} x_{2,2} + \beta_{2,3} x_{2,3} + \beta_{3,1} x_{3,1} + \beta_{3,2} x_{3,2} + \beta_{3,3} x_{3,3} + \varepsilon_v, $$

(10)

where $x_{i,j}$ is the transformed input function of task $j$ in $i$th trial, $z_i$ is a vector used to remove low-frequency, stimulus independent effects, $\beta_{i,j}$ is the parameter of interest corresponding to the $j$th task in $i$th trial, the $a_i$ are nonzero nuisance parameters to model the baseline of a brain signal.
Figure 4: The top panel is activation maps using an AR(1) temporal dependence. The top row corresponds to performing the "Ink Only" task, the middle one for the "Congruence" task, and bottom one for the "Interference" task.
and $\varepsilon_v$ has a normal distribution, $N(0, \sigma_v^2 \Lambda_v)$. We continue to use the EB method based on the MLE $\hat{\rho}$ for the elements of $\Lambda_v$. A sequence of binary variables is used to indicate if the corresponding parameter is zero or not. In our case, we assume $a_i$ nonzero. Therefore, the variable for voxel $v$ is

$$\gamma_v = \{1, 1, \gamma_3,v, \gamma_4,v, \gamma_5,v, \gamma_6,v, \gamma_7,v, \gamma_8,v, \gamma_9,v, \gamma_{10},v, \gamma_{11},v\}.$$ 

The rest of the model specification is as described in Section 2.

The posterior of interest $q(\gamma, \theta|y)$ is intractable so we will use the MCMC algorithm developed in Section 2.2 to obtain $1 \times 10^5$ observed values of the Markov chain. We use the method of batch means (Jones et al., 2006) to calculate MCSEs for the estimated quantities. We also used trace plots and histograms to visually assess convergence of the simulation.

The activation maps are given in Figures 5, 6, and 7. All these figures show that the partial lobe is activated during these tasks in different phases but the occipital lobe is only activated in the third trial and the frontal lobe is slightly activated in the first trial of congruence and interference tasks and in the second trial of ink only task. The temporal lobe is only activated in the third phase, not in first and second phases. Notice that the size of activation areas when performing the "Interference" task is bigger than that when performing the other two.
Figure 5: Activation maps for the time-varying model in the first trial. The top panel is activation maps when performing "Ink Only" task, the middle one for "Congruence" task, and bottom one for "Interference" task.
Figure 6: Activation maps for the time-varying model in the second trial. The top panel is activation maps when performing "Ink Only" task, the middle one for "Congruence" task, and bottom one for "Interference" task.
Figure 7: Activation maps for the time-varying model in the third trial. The top panel is activation maps when performing "Ink Only" task, the middle one for "Congruence" task, and bottom one for "Interference" task.
4.4 Model Assessment

To keep things manageable we will limit discussion to the model in Section 4.2. Similar results were found for the model in Section 4.3. We performed several voxel-level exploratory analyses to assess model fit. For example, we plotted the raw data versus time with a trend line, residuals as a function of time, quantile-quantile plots to assess our distributional assumption of normality and ACF and partial ACF plots to assess the temporal correlation. We present the results for several randomly selected voxels in the Figures 8 and 9. Consider Figure 8. The first row of each plot is the raw data with a fit line. The second row is the residual multiplied by a inverse squared root of AR(1) covariance and the quantile-quantile plots for the corresponding residuals with a Shapiro–Wilk test given. Even when the Shapiro–Wilk test rejects normality, we find by checking the quantile-quantile plots that most of the points lie around the straight line. Now consider Figure 9 which contains the ACF and PACF plots for the same voxels as in Figure 8. Mostly this shows that an AR(1) structure is reasonable although a couple suggest that a higher order autoregressive structure could be appropriate.

We also performed voxel-wise chi-squared goodness-of-fit tests by comparing the reduced model to the full model at the .1 level, very few (.01%) of the null hypotheses (reduced model) were rejected. In addition, percentages of BIC values over voxels within the brain for reduced model less than that for full model was very small (0.05%). Finally, we examined the nature of the activation patterns with the 3 possible cutoff probabilities (i.e. .9650, .8722 and .7946) introduced in Section 3 and found little practical difference in the patterns. In sum, these analyses suggest that our model fits the data reasonably well.

Recall that in Section 4.2 we used EB methods for all of the \( \rho_v \) and assumed a *priori* that the \( \rho_v \) are independent across voxels. This was done mainly because the simulation study results from Section 3 suggested that the quality of inference was similar when we assumed \( \rho_v \sim \text{Uniform}(−1,1) \), independently. Moreover, when we tried to use the Uniform priors on the experimental fMRI data the computation was prohibitively expensive. The crux of the computational issue can be seen by examining the Hastings ratio for the MCMC step updating \( \rho_v \)

\[
\frac{\pi(\rho_v^*)|\Lambda_v^*|^{-1/2}S(\rho_v^*, \gamma_v)^T + \Lambda_v^{-1}\left((y_v - X_v(\gamma_v)^\beta_v)^T\Lambda_v^{-1}(y_v - X_v(\gamma_v)^\beta_v)\right)}{\pi(\rho_v)|\Lambda_v|^{-1/2}S(\rho_v, \gamma_v)^T + \Lambda_v^{-1}\left((y_v - X_v(\gamma_v)^\beta_v)^T\Lambda_v^{-1}(y_v - X_v(\gamma_v)^\beta_v)\right)}
\]

where \( S(\rho_v, \gamma_v) = (y_v - X_v(\gamma_v)^\beta_v)^T\Lambda_v^{-1}(y_v - X_v(\gamma_v)^\beta_v) \). Although computing the determinant of \( \Lambda_v \) is easy and fast, calculating \( S(\rho_v, \gamma_v) \) is time consuming partly because \( \Lambda_v \) must be inverted. In the current fMRI data example \( \Lambda_v \) is \( 465 \times 465 \) and it must be inverted for each voxel.

However, the assumption of independence of the \( \rho_v \) may not be optimal. Figure 10 shows the MLE for \( \rho_v \) in each voxel. There does seem to be some evidence that nearby voxels tend to have similar values and hence the map may be exhibiting spatial correlation. It would be a potentially
interesting research project to try to account for this through a prior specification on \( \rho \) while maintaining computational feasibility.

Finally, we compare the activation maps produced from our model when different temporal structures, AR(2), ARMA(1,1), and MA(1) are assumed. In each case we used maximum likelihood estimates of the parameters and implemented an empirical Bayes approach. The computational effort for each of these was similar to that of the EB AR(1) model. The results are presented in figures in the Supplemental Material. The activation maps for the AR(1) and AR(2) models are comparable, but the ARMA(1,1) and MA(1) activation maps show larger activated regions. Given the results of the ACF and PACF plots above as well as the results of the simulation study in Section 3 we suspect that the ARMA(1,1) and MA(1) fits are yielding a substantial number of false positives. Overall, the use of an AR(1) temporal dependence seems quite reasonable in this example.

5 Discussion

We have developed a Bayesian hierarchical model which incorporates both spatial and temporal correlation. In the process we demonstrated how to incorporate a time-varying coefficient that allows for the variety of processes that might impact a BOLD response to a task during scanning. This work has the potential to have a broad and immediate impact on single-subject and -session fMRI task activation studies, which are the common first step when performing an fMRI study. Moreover, recent developments in the clinical usage of fMRI rely on subject- and session-specific investigations to locate and investigate language areas prior to brain surgery (Sabsevitz et al., 2003). In addition, for diseases such as AD, it is hoped that imaging can be used as a diagnostic utility for initiating treatment (Bassett et al., 2006).

For broader goals it would be of interest to extend the model to group studies. One approach would be to mirror existing group studies and adopt a two-stage frequentist/least squares approach (see Friston et al., 2007). That is, use posterior quantities from the subject-specific models in standard second-stage inter-subject regression models. A more satisfying answer would incorporate recent developments in fMRI meta analysis models (Kang et al., 2011), employing a Bayesian second-stage model. Yet another alternative would build a global Bayesian model for groups of whole brain functional images, without relying on two-stage methods. However, such a solution presents numerous modeling and computational challenges. For example, the study of MCMC convergence in such high dimensional settings remains unexplored (present study included). Hence the consequences of having far more parameters than possible MCMC iterations remains unknown. Moreover, any fully Bayesian model for groups must not require loading the full inter-subject data set into memory to be scalable to the ever-increasing scope of fMRI studies. For many of these
Figure 8: Residual check. The 1st, 4th, 7th rows represent the raw data with a fit line, 2nd, 5th, 8th are the corresponding residuals with a spline, and 3rd, 6th, 9th rows are the corresponding QQ plots.
Figure 9: ACF and PACF plots for randomly selected voxels.

Figure 10: Estimates of $\rho_v$ for each voxel.
problems variational Bayesian solutions may be important future directions.

Despite computational and modeling difficulties, we are confident that Bayesian approaches represent an important direction in fMRI, and high-dimensional research in general. Studies such as the present one and others (e.g. Bowman et al., 2008; Caffo et al., 2011; Goldsmith et al., 2012; Smith and Fahrmeir, 2007) demonstrate the practicality, efficacy and feasibility of general Bayesian solutions to these problems.

A Derivation of \( q(\gamma, \rho, \theta|y) \)

\[
q(\gamma, \rho, \theta|y) \propto \int q(\beta(\gamma), \gamma, \rho, \theta, \sigma^2|y) d\beta(\gamma) d\sigma^2
\]

\[
\propto \pi(\gamma|\theta)\pi(\rho)\pi(\theta) \prod_{v=1}^{N} p(y_v|\beta_v(\gamma_v), \gamma_v, \sigma_v^2, \Lambda_v)\pi(\beta_v(\gamma_v)|y_v, \gamma_v, \sigma_v^2, \Lambda_v)\pi(\sigma_v^2) d\beta_v(\gamma_v) d\sigma_v^2
\]

Consider the integrand

\[
p(y_v|\beta_v(\gamma_v), \gamma_v, \sigma_v^2, \Lambda_v)\pi(\beta_v(\gamma_v)|y_v, \gamma_v, \sigma_v^2, \Lambda_v)\pi(\sigma_v^2)
\]

\[
\propto \left[\frac{1}{(1/T_v)X_v^T(\gamma_v)\Lambda_v^{-1}X_v(\gamma_v)}\right]^{1/2} \exp \left\{ -\frac{1}{2\sigma_v^2} \left[ (y_v - X_v(\gamma_v)\beta_v(\gamma_v))^{T}\Lambda_v^{-1}(y_v - X_v(\gamma_v)\beta_v(\gamma_v)) + \frac{1}{T_v} (\beta_v(\gamma_v)\beta_v(\gamma_v))^{T}\Lambda_v^{-1}X_v(\gamma_v)(\beta_v(\gamma_v) - \hat{\beta}_v(\gamma_v)) \right] \right\} (\sigma_v^2)^{-1+(T_v+q_v)/2}(2\pi)^{-q_v/2}
\]

\[
= \left[\frac{1}{|\Lambda_v|}\right]^{1/2} \left(\sigma_v^2\right)^{-1+(T_v+q_v)/2}(2\pi)^{-q_v/2} \times
\]

\[
\times \exp \left\{ -\frac{1}{2\sigma_v^2} \left[ \left(1 + \frac{1}{T_v}\right) \left[ \beta_v(\gamma_v)X_v^T(\gamma_v)\Lambda_v^{-1}X_v(\gamma_v)\beta_v(\gamma_v) - 2y_v^T\Lambda_v^{-1}X_v(\gamma_v)\beta_v(\gamma_v) \right] \right) \right\} \times
\]

\[
\times \exp \left\{ -\frac{1}{2\sigma_v^2} y_v^T \left[ \Lambda_v^{-1} + \frac{1}{T_v}\Lambda_v^{-1}X_v(\gamma_v) (X_v^T(\gamma_v)\Lambda_v^{-1}X_v(\gamma_v))^{-1} X_v^T(\gamma_v)\Lambda_v^{-1} \right] \right\} y_v
\]

Now let \( A_v = X_v^T\Lambda_v^{-1}X_v(\gamma_v) \) and \( C_v = y_v^T\Lambda_v^{-1}X_v(\gamma_v) \). Note that

\[
(\beta_v(\gamma_v) - A_v^{-1}C_v^T)A(\beta_v(\gamma_v) - A_v^{-1}C_v^T) - C_v A_v^{-1}C_v^T = \beta_v^T(\gamma_v)A_v\beta_v(\gamma_v) - 2C_v\beta_v(\gamma_v).
\]
Hence

\[
p(y_v | \beta_v(\gamma_v), \gamma_v, \sigma_v^2, \Lambda_v) \propto (\frac{X_v^T(\gamma_v) \Lambda_v^{-1} X_v(\gamma_v)}{|\Lambda_v|})^{1/2} \left(\sigma_v^2\right)^{-(1+(T_v+q_v)/2)} (2\pi T_v)^{-q_v/2} \times \\
\times \exp \left\{ -\frac{1}{2\sigma_v^2} \left( 1 + \frac{1}{T_v} \right) \left( \beta_v(\gamma_v) - A_v^{-1} C_v^T \right)^T A(\beta_v(\gamma_v) - A_v^{-1} C_v^T) \right\} \times \\
\times \exp \left\{ -\frac{1}{2\sigma_v^2} y_v^T \left[ \Lambda_v^{-1} - A_v^{-1} X_v(\gamma_v) \left( X_v^T(\gamma_v) \Lambda_v^{-1} X_v(\gamma_v) \right)^{-1} X_v^T(\gamma_v) \Lambda_v^{-1} \right] y_v \right\}.
\]

Now

\[
q(\gamma, \theta, \rho | y) \propto \pi(\gamma|\theta) \pi(\rho|\theta) \prod_{v=1}^{N} \left[ (\frac{X_v^T(\gamma_v) \Lambda_v^{-1} X_v(\gamma_v)}{|\Lambda_v|})^{1/2} \right] \left(2\pi T_v\right)^{-q_v/2} \times \\
\int \frac{1}{(\sigma_v^2)^{1+\frac{T_v+q_v}{2}}} \exp \left\{ -\frac{1}{2\sigma_v^2} y_v^T \left[ \Lambda_v^{-1} - A_v^{-1} X_v(\gamma_v) \left( X_v^T(\gamma_v) \Lambda_v^{-1} X_v(\gamma_v) \right)^{-1} X_v^T(\gamma_v) \Lambda_v^{-1} \right] y_v \right\} \times \\
\int \exp \left\{ -\frac{1}{2\sigma_v^2} \left( 1 + \frac{1}{T_v} \right) \left( \beta_v(\gamma_v) - A_v^{-1} C_v^T \right)^T A(\beta_v(\gamma_v) - A_v^{-1} C_v^T) \right\} \frac{1}{(1 + T_v)^{q_v/2} |\Lambda_v|^{1/2}} \times \\
\int \frac{1}{(\sigma_v^2)^{1+\frac{T_v}{2}}} \exp \left\{ -\frac{1}{2\sigma_v^2} y_v^T \left[ \Lambda_v^{-1} - A_v^{-1} X_v(\gamma_v) \left( X_v^T(\gamma_v) \Lambda_v^{-1} X_v(\gamma_v) \right)^{-1} X_v^T(\gamma_v) \Lambda_v^{-1} \right] y_v \right\} \frac{1}{(1 + T_v)^{q_v/2} |\Lambda_v|^{1/2}} \times \\
\propto \pi(\gamma|\theta) \pi(\rho|\theta) \prod_{v=1}^{N} \frac{1}{(1 + T_v)^{q_v/2} |\Lambda_v|^{1/2}} S(p_v, \gamma_v)^{-T_v/2}
\]

References


32


