Bayesian Analysis (2018) **13**, Number 4, pp. 1261–1313

Bayesian Spatiotemporal Modeling Using Hierarchical Spatial Priors, with Applications to Functional Magnetic Resonance Imaging (with Discussion)

Martin Bezener*, John Hughes†, and Galin Jones‡

Abstract. We propose a spatiotemporal Bayesian variable selection model for detecting activation in functional magnetic resonance imaging (fMRI) settings. Following recent research in this area, we use binary indicator variables for classifying active voxels. We assume that the spatial dependence in the images can be accommodated by applying an areal model to parcels of voxels. The use of parcellation and a spatial hierarchical prior (instead of the popular Ising prior) results in a posterior distribution amenable to exploration with an efficient Markov chain Monte Carlo (MCMC) algorithm. We study the properties of our approach by applying it to simulated data and an fMRI data set.

Keywords: Bayesian variable selection, fMRI, MCMC, spatiotemporal, areal model.

1 Introduction

A typical goal of functional magnetic resonance imaging (fMRI) experiments is to infer the location and magnitude of neuronal activity in response to a stimulus or task. Neuronal activation is not directly observed with fMRI, but the principle of neurovascular coupling allows the use of a proxy. When neurons activate, blood flow increases to active areas of the brain, and an excess of oxygenated hemoglobin is delivered to those regions, causing a change in the magnetic field. This change, which is known as the blood oxygen dependent (BOLD) signal, is captured by the scanner.

Before an fMRI experiment, the image space is partitioned into a rectangular lattice comprising small cubic volume elements, or voxels, of equal size. Partition sizes range from 100,000 to 5,000,000 voxels, with voxel size chosen to balance resolution and time requirements. During the fMRI experiment, a subject lies in the scanner and performs a task (such as finger tapping or talking), or is exposed to an external stimulus (such as watching a movie), while measurements of the BOLD signal at each voxel are captured every two to three seconds at several hundred time points. The data are then preprocessed before statistical analysis to remove artifacts introduced during the data collection process; see, for example, Friston et al. (2007), Huettel et al. (2009), Kaushik et al. (2013), Lazar (2008), Lindquist (2008), Mikl et al. (2008), and Triantafyllou et al.

*Stat-Ease, Inc., martin.bezener@gmail.com
†www.johnhughes.org, jphughesjr@gmail.com
‡School of Statistics, University of Minnesota Twin Cities, galin@umn.edu

 https://doi.org/10.1214/18-BA1108

(2006). We are left with a large amount of data that exhibit spatiotemporal dependence since repeated observations on the same voxel are temporally dependent and neighboring voxels tend to behave similarly.

Standard non-Bayesian methods such as those in Friston et al. (1994), Friston et al. (1995), Friston et al. (2007), Landman et al. (2012), Worsley et al. (1992), Worsley et al. (2002), and Worsley (2003) do not involve explicit spatial modeling. However, these methods are computationally efficient. Bayesian approaches have also received a lot of attention (see, among many others, Bowman et al., 2007; Genovese, 2000; Gössel et al., 2001; Penny et al., 2003, 2005; Quirós et al., 2010; Woolrich et al., 2004; Xia et al., 2009a; Zhang et al., 2014, 2016). Bayesian models can incorporate both spatial and temporal dependence but typically result in complicated high-dimensional posterior distributions and hence can be computationally burdensome to apply. Reviews of Bayesian methods for neuroimaging can be found in Bowman (2014), Friston et al. (2007), Lazar (2008), and Zhang et al. (2015).

One particular Bayesian approach is based on variable selection methods in the linear model. Originally these methods were developed for independent data (George and McCulloch, 1993, 1997), but they have been extended to settings with spatiotemporal dependence in the context of functional neuroimaging (Lee et al., 2014; Smith et al., 2003; Smith and Fahrmeir, 2007). The basic idea is to extend the linear model by incorporating a binary indicator variable for each voxel (to indicate activation), thereby transforming the fMRI activation detection problem into a variable selection problem. To address spatial dependence, Smith and Fahrmeir (2007) employed a binary Ising prior on the field of indicator variables. They do not, however, fully consider the issue of temporal dependence. This is addressed by Lee et al. (2014), who extend the model of Smith and Fahrmeir (2007) to allow prior information about temporal dependence to be incorporated in the variable selection scheme.

The Ising model (Cipra, 1987; Smith and Smith, 2006; Murphy, 2012) is popular in spatial modeling. While attractive from a modeling perspective, its use can create substantial computational difficulties. For example, it often results in a doubly-intractable posterior distribution and hence requires an auxiliary algorithm to estimate intractable normalizing constants (Morris et al., 1996) in Markov Chain Monte Carlo (MCMC) applications. Most algorithms for this are computationally expensive, result in slow mixing, or scale poorly to large problems (Zhou and Schmidler, 2009). Additionally, posterior estimates can be sensitive to the run length of the MCMC sampler (Murphy, 2012). Indeed it has become common to turn to variational inference (see e.g. Blei et al., 2017) in these situations, which does make the computation feasible, but provides inferences based on a difficult-to-quantify approximation to the desired posterior distribution.

We propose a Bayesian spatiotemporal model for detecting activation patterns in fMRI data. 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. This is done by parcellating the image into clusters of voxels and modeling the structure of the spatial dependence using a spatial hierarchical prior (Haran, 2011) for the parcels. The use of parcellation and the spatial hierarchical prior (instead of the Ising prior) results in much more efficient

computation. The modeling advances in conjunction with careful implementation of MCMC methods allows a whole-brain analysis in a fraction of the time required by other methods (cf. Lee et al., 2014) using standard hardware.

The work most closely related to ours was developed by Musgrove et al. (2016). There are many differences between their modeling approach and ours, but there is one significant commonality in that they also use a parcellation scheme. However, they assume that the parcels are independent and hence the posterior factors into independent components. This allows them to use a parallel computing strategy to achieve efficient computation. We do not make any such assumption and do not use parallel computing, yet we achieve similar computational efficiency.

The remainder of this paper is organized as follows. Our Bayesian model is introduced in Section 2. Numerical results are reported in Section 3. We then follow with an analysis of the parcellation in Section 4 and an fMRI data example in Section 5. Details of model fitting, additional simulation results, and an extra data example are given in the supplementary material (Bezener et al., 2018). Final remarks are given in Section 6.

2 The Model

We begin by describing our spatiotemporal model and discussing prior selection. We then describe the posterior distribution and derive the conditional densities used in posterior sampling.

2.1 Model Formulation

Suppose there are $v = 1, \ldots, N$ voxels and a sequence of T_v measurements is taken at voxel v. Also assume that there are p distinct experimental tasks or stimuli of interest. The time series at voxel v is modeled as

$$y_v = Z_v\eta_v + X_v\beta_v + \varepsilon_v, \quad \varepsilon_v \sim N_{T_v}(0, \sigma_v^2\Lambda_v), \tag{1}$$

where X_v is a $T_v \times p$ design matrix of full column rank, β_v is a $p \times 1$ vector of regression coefficients interpreted as activation amplitudes, and ε_v is the error with correlation matrix Λ_v. The matrix Z_v includes covariates that are not of direct interest in the analysis but must be taken into account to facilitate valid inference. These include the baseline signal, long-term drift effects, and other low-frequency noise. The vector η_v contains the coefficients corresponding to Z_v. For the remainder, we assume that the data has been adequately de-trended and preprocessed and drop $Z_v\eta_v$ from (1) to focus on β_v.

We assume an AR(1) error process with the (i, j)th element of Λ_v given by $\rho_v^{|i-j|}$. It is straightforward to incorporate other structures in our modelling framework, however, autoregressive and autoregressive moving average structures are sensible starting points, and are common in neuroimaging applications (see e.g. Lee et al., 2014; Lindquist, 2008; Xia et al., 2009b; Locascio et al., 1997; Monti, 2011; Penny et al., 2003). In particular, Penny et al. (2003) show that low-order autoregressive processes are sufficient as long as drift effects are accounted for in Z_v.

Recall that the BOLD signal is used as a proxy for measuring neuronal activation in fMRI. However, blood does not immediately start flowing to active neurons upon activation nor does it stop as soon as neurons are no longer active. Instead, blood flow is delayed by several seconds after activation and proceeds continuously according to a hemodynamic response function (HRF). A common view is that HRF estimation and activation detection are impossible to disentangle (Makni et al., 2008), but since we focus on the use of a spatial hierarchical prior with our parcellation method, we will follow standard practice and transform the columns of X_v so that the linear model is appropriate. Let $s_{v,j}(t)$ be the stimulus function for task j at voxel v and $h_v(t)$ be the assumed HRF. We follow the standard practice of transforming the columns of X_v by the discretized convolution

$$X_v(t,j) = (s_{v,j} \star h_v)(t) = \sum_{i=0}^{t-d_v} s_{v,j}(t - d_v - i) \cdot h(i)$$

if $t - d_v > 0$ and 0 otherwise, where d_v is a user-specified lag parameter. We will use the canonical HRF, which uses the difference of two gamma densities to model the BOLD response. This issue is discussed in detail in Friston et al. (2007) and Lindquist (2008).

To facilitate variable selection, we extend model (1) by using latent indicator variables $\gamma_v = (\gamma_{v,1}, \gamma_{v,2}, \ldots, \gamma_{v,p})$ to denote which of the p tasks and stimuli results in activation of voxel v. That is, $\beta_{v,j} \neq 0$ if $\gamma_{v,j} = 1$ and $\beta_{v,j} = 0$ if $\gamma_{v,j} = 0$. We rewrite (1) as

$$y_v = X_v(\gamma_v)\beta_v(\gamma_v) + \varepsilon_v, \tag{2}$$

where $\beta_v(\gamma_v)$ is the vector of the nonzero coefficients of β_v and $X_v(\gamma_v)$ is the corresponding design matrix. The activation detection problem in this setting is equivalent to classifying the non-zero $\gamma_{v,j}$ and is therefore a variable selection problem.

2.2 Prior Specifications

Priors for σ_v^2 and ρ_v

Let $\rho = (\rho_1, \rho_2, \ldots, \rho_N)$. We assume the ρ_v are a priori independent so the joint prior on ρ is $\pi(\rho) = \prod_{v=1}^N \pi(\rho_v)$. It seems natural to assume that $\rho_v \sim \text{Uniform}(-1, 1)$, but our experience with fMRI data indicates that priors putting much more mass on the nonnegative part of the interval $(-1, 1)$ are more consistent with a priori scientific expectations. However, we will take an empirical Bayesian approach since being fully Bayesian may result in substantially more computational effort as we explain in the supplementary material. In an effort to provide a balance between computational efficiency and inferential efficacy we assume $\pi(\rho_v) \sim \hat{\rho}_v$ independently. A good candidate for $\hat{\rho}_v$ is its maximum likelihood estimate. This prior is also recommended by Lee et al. (2014) for similar reasons.

Let $\sigma^2 = (\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2)$. We assume the σ_v^2 are a priori independent, so the joint prior on σ^2 is $\pi(\sigma^2) = \prod_{v=1}^N \pi(\sigma_v^2)$. The prior on each σ_v^2 is the standard invariant prior $\pi(\sigma_v^2) \propto \sigma_v^{-2}$.

Prior for $\boldsymbol{\beta_v(\gamma_v)}$

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 γ, σ^2, and ρ, we assume the $\beta_v(\gamma_v)$ are conditionally independent:

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

For each $\beta_v(\gamma_v)$, we use a g prior (Zellner, 1996), which has the form

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

where

$$\begin{aligned} \hat{\beta}_v(\gamma_v) &= [X_v^T(\gamma_v)\Lambda_v^{-1}X_v(\gamma_v)]^{-1}X_v^T(\gamma_v)\Lambda_v^{-1}y_v, \\ \hat{\Sigma}_v &= [X_v^T(\gamma_v)\Lambda_v^{-1}X_v(\gamma_v)]^{-1} \ . \end{aligned} \tag{3}$$

This prior requires the selection of the tuning parameter g, which we set to $g = T_v$, yielding a unit information prior.

Prior for $\boldsymbol{\gamma_v}$

We work directly with the prior probabilities of activation $\pi(\gamma_{v,j} = 1)$. Such an approach has been shown to produce activation maps with better edge-preservation properties and classification accuracies compared to methods that place priors on the activation amplitudes (Smith and Fahrmeir, 2007). Manually specifying a value for each $\pi(\gamma_{v,j} = 1)$ is not feasible unless N is small.

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 assume that the image can be parcellated into G non-overlapping regions. The parcellation should be chosen so that voxels within each region behave similarly due to their location.

We do not require that each region 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 split into many smaller regions. As a practical guideline, we recommend using fewer than $G = 500$ regions for computational reasons. For typical data sets, this means each region will contain from 10 to 400 voxels. Computational issues related to various choices of G are investigated in Section 4.

Let $\gamma_{(j)} = (\gamma_{1,j}, \gamma_{2,j}, \ldots, \gamma_{N,j})$ be the vector all active/inactive voxels under task j. (Although the notation is similar, this is different from γ_v previously defined.) 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

$\mathcal{R}_g$ as the collection of all voxels in region g. We assume that the $\gamma_{v,j}$ are conditionally independent, and so the prior is

$$\pi(\gamma_{(j)} \mid S_{(j)}) = \prod_{g=1}^{G} \prod_{v \in \mathcal{R}_g} \pi(\gamma_{v,j} \mid S_{g,j}) \ .$$

Under this framework, all voxels in region g share the same spatial random effect, and so the prior probability of activation is the same for all voxels in $\mathcal{R}_g$. Note that if $G = N$, each voxel gets its own spatial random effect, which is computationally reasonable when N is small. Given that voxel v lies in $\mathcal{R}_g$, link the prior probabilities to the spatial random effects via the logistic transformation

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

We assume that the $S_{(j)}$ are generated from a Gaussian process so that

$$S_{(j)} \mid \delta_j^2, r_j \overset{\text{ind}}{\sim} N(0, \delta_j^2 \Gamma_j), \tag{5}$$

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

$$\Gamma_j(i,k) = \exp\left(-\frac{\|s_i - s_k\|}{r_j}\right) . \tag{6}$$

Here s_i and s_k denote the centroid coordinates of regions i and k and $||\cdot||$ is the Euclidean distance. In this prior, δ_j^2 is a smoothing parameter which controls the amount of spatial continuity in the $S_{(j)}$ and hence in the $\gamma_{(j)}$. The spatial covariance matrix Γ_j controls the structure and amount of the spatial dependence in the $S_{(j)}$, and therefore, in the $\gamma_{(j)}$. Under this prior, the strength of the spatial dependence between neighboring regions under task j is determined by the range parameter r_j. We require that $r_j > 0$ so that Γ_j is a valid correlation matrix. This prior assumes that regions close to one another will exhibit similar behavior compared to regions further apart. Notice that we allow r_j and δ_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 correlation and smoothness.

In some applications long-range spatial dependence may be of interest. One possibility for addressing this issue is to use an inverse-Wishart prior for Γ_j instead of assuming (6). This is outside the scope of the current paper, and so we do not pursue it further here.

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 $\pi(S \mid \delta^2, r) = \prod_{j=1}^{p} \pi(S_{(j)} \mid \delta_j^2, r_j)$, $\pi(\delta^2) = \prod_{j=1}^{p} \pi(\delta_j^2)$, and $\pi(r) = \prod_{j=1}^{p} \pi(r_j)$.

We consider χ^2 priors for r_j. These 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 correlation parameter.

As r gets larger, images under this prior start to look similar due to the correlation structure in (6). However, we will show through simulation that posterior inferences are robust to the choice of degrees of freedom. We assume the standard invariant prior for δ_j^2, that is, $\pi(\delta_j^2) \propto \delta_j^{-2}$.

Other priors for r_j, such as uniform distributions, are also plausible. In situations where the amount of spatial dependence and smoothness is known a-priori, both r_j and δ_j could fixed to a particular value. This would simplify the model and would substantially speed up model fitting. In our experience, however, this amount of information is usually not available before experimentation, and the uncertainty can be quantified through the use of priors on these hyperparameters.

2.3 Posterior and Conditional Distributions

Combining the results of the previous sections, the posterior distribution is given by

$$\begin{aligned} q(\beta(\gamma), \gamma, S, \delta^2, r, \rho, \sigma^2 \mid y) &\propto p(y \mid \beta(\gamma), \gamma, S, \delta^2, r, \rho, \sigma^2)\pi(\beta(\gamma), \gamma, S, \delta^2, r, \rho, \sigma^2) \\ &\propto p(y \mid \beta(\gamma), \gamma, \rho, \sigma^2)\pi(\beta(\gamma) \mid \gamma, \rho, \sigma^2)\pi(\rho)\pi(\sigma^2) \\ &\quad \times \pi(\gamma \mid S)\pi(S \mid \delta^2, r)\pi(\delta^2)\pi(r) \ . \end{aligned} \tag{7}$$

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 $q(\gamma_{v,j} = 1 \mid y)$ for all v, j, and well as posterior estimates of the spatial correlation parameter $E(r_j \mid y)$ for all j. These quantities cannot be analytically determined from (7) so we use MCMC methods.

The dimension of the posterior in (7) is $2p(N+1) + 2N + pG$, which in a typical single subject study, can range from tens of thousands to several millions of variables. Clearly, sampling from (7) would be challenging and reducing the dimension of the posterior would be advantageous here. To balance our inferential goals with computation limitations, we instead work with a collapsed Gibbs sampler. This requires that some variables be integrated out from the full posterior. The posterior that allows us to achieve our goals is $q(\gamma, S, r, \rho \mid y)$ which we now derive.

The first step is to integrate out the $\beta(\gamma)$ and σ^2. These parameters are of no interest to the classification problem (below we will discuss how to estimate $\beta(\gamma)$) and integrating them out reduces the dimension of the posterior by $(p+1)N$. In typical settings, this represents a reduction in dimension of approximately 35 to 50%. Note that

$$\begin{aligned} p(y \mid \gamma, \rho) &= \int p(y \mid \beta(\gamma), \gamma, \rho, \sigma^2)\pi(\beta(\gamma) \mid \gamma, \rho, \sigma^2)\pi(\sigma^2) d\beta(\gamma) d\sigma^2 \\ &= \prod_{v=1}^{N} (1+T_v)^{-q_v/2} |\Lambda_v|^{-1/2} K(\rho_v, \gamma_v)^{-T_v/2}, \end{aligned} \tag{8}$$

where $q_v = \sum_{j=1}^{p} \gamma_{v,j}$ denotes the number of non-zero entries in γ_v and

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

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

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

The next step is to integrate out δ^2. It is easy to show that

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

Removing δ^2 avoids expensive matrix operations and is not of direct interest to the classification problem. Then our reduced posterior is

$$q(\gamma, S, r, \rho \mid y) \propto p(y \mid \gamma, \rho)\pi(\rho)\pi(\gamma \mid S)\pi(S \mid r)\pi(r) \ . \tag{9}$$

It would be ideal to integrate out the S to obtain the reduced posterior $q(\gamma, r, \rho \mid y)$. However, the integral

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

is analytically intractable, so we instead sample from (9) and discard the observed S.

Since each conditional posterior density is proportional to the joint posterior in (9), it is straightforward to see that

$$\begin{aligned}
q(\gamma \mid S, r, \rho, y) &\propto \pi(\gamma \mid S) \prod_{v=1}^{N} (1 + T_v)^{-q_v/2} K(\rho_v, \gamma_v)^{-T_v/2}, \\
q(S \mid \gamma, r, \rho, y) &\propto \pi(\gamma \mid S)\pi(S \mid r), \\
q(r \mid S, \gamma, \rho, y) &\propto \pi(S \mid r)\pi(r), \\
q(\rho \mid S, \gamma, r, y) &\propto \pi(\rho) \prod_{v=1}^{N} |\Lambda_v|^{-1/2} K(\rho_v, \gamma_v)^{-T_v/2} \ .
\end{aligned}$$

None of these full conditionals are available in closed form and hence we have to use a component-wise Metropolis-Hastings algorithm. The algorithm is fully specified in the supplementary material as is our method for choosing sensible starting values for the simulation and the method for termination of the simulation.

2.4 Posterior Estimates and Classification of Active Voxels

We now turn our attention to the estimation of the posterior quantities of interest. Recall that one of our goals is to identify active voxels. We do this by computing the posterior probabilities of activation for each voxel $q(\gamma_{v,j} = 1 \mid y)$.

Suppose we draw M samples from the posterior distribution in (9). Let $\{\gamma^{(1)}, \gamma^{(2)}, \ldots, \gamma^{(M)}\}$ denote the obtained MCMC sample of the binary indicator variables. We can estimate the posterior probability that voxel v is activated by task j by

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

It is also straightforward to estimate the amount of spatial dependence in the images under each task. As mentioned previously, the amount of spatial correlation is controlled by r_j. Let $\{r^{(1)}, r^{(2)}, \ldots, r^{(M)}\}$ denote the obtained MCMC sample of the dependence parameters. We use the posterior means as the estimates of the r_j. That is,

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

The magnitude of activation is frequently of interest. Even though we are working with a marginal posterior we can estimate the magnitude through Rao-Blackwellization since

$$E(\beta_v \mid y) = \sum_{\gamma_v} E(\beta_v \mid \gamma_v, y) q(\gamma_v \mid y) \approx \frac{1}{M} \sum_{m=1}^{M} \hat{\beta}(\gamma_v^{(m)}) \ .$$

We choose the number of MCMC samples (that is, the value of M) by assessing the Monte Carlo standard error of the estimates (Flegal et al., 2008; Jones et al., 2006; Vats et al., 2016) with the method of batch means calculated using the `mcmcse` package (Flegal et al., 2017).

In order to classify voxels as active, an activation threshold is necessary. Following Smith and Fahrmeir (2007), Lee et al. (2014), Musgrove et al. (2016), Raftery (1996), and Smith and Smith (2006) we classify voxel v as active under task j if $q(\gamma_{v,j} = 1 \mid y) > 0.8722$. Smith and Fahrmeir (2007) show that this threshold corresponds to a p-value of 0.05. This threshold also yields a posterior error probability of $1 - .8722 = .1278$ which gives an upper bound on the false discovery rate (Käll et al., 2008; Storey, 2003) There are several other decision-theoretic approaches to thresholding; see, for example, Zhang et al. (2016). We find that 0.8722 provides good classification accuracy while keeping the rate of false positives low. This is further investigated in the supplementary material.

3 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 describe the amount of spatial dependence in the images. The second simulation considers the performance of the model when spatial independence is assumed. The final simulation examines the case when there is no activation in the image. The supplementary material contains two additional simulations: a simulation study that assesses the adequacy of the proposed activation threshold of 0.8722, and a classification accuracy study involving a repeated task block experiment.

3.1 Simulation 1

We first consider the model which does not parcelate the image by letting $G = N$. This is a computationally reasonable model when the number of voxels is small or if a region of interest or single slice analysis is desired. We generate fifteen 20×20 active/inactive

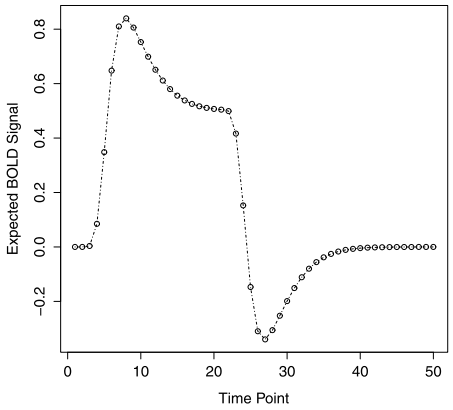


Figure 1: The design matrix used in the simulations.

square images under the proposed spatial hierarchical model when the true value of r is $0.001, 2, 8$, and 20. These values represent images that have spatial correlation structures ranging from near independence to strong spatial correlation. In each case δ^2 was fixed at 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. Denote $\beta_v(\gamma_v) = (\beta_{v,0}, \beta_{v,1})^T$. We assume that $\beta_{v,0} = 300$. When $\gamma_v = 0$, we set $\beta_{v,1} = 0$ and when $\gamma_v = 1$, we consider the cases when $\beta_{v,1} = 3$ and 5. These are typical signal strengths encountered in fMRI data. We select $\sigma_v^2 = 3$ for all v in this simulation. The AR(1) correlation coefficients ρ_v were generated independently from a $U(-1, 1)$ distribution.

We investigate the performance of the model at two prior settings for r, considering both χ_2^2 and χ_8^2 priors.

All tuning parameters were chosen to give acceptance rates between 40 and 55%. Monte Carlo standard errors were computed using the method of batch means (Jones et al., 2006; Vats et al., 2016) using the `mcmcse` package (Flegal et al., 2017) with the batch size set to the root of the number of samples drawn. In each simulation, 150,000 posterior samples were drawn, which resulted in nearly all standard errors being within 2% of the estimated posterior mean.

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 higher signal to noise ratio. We see that the posterior estimates of r are reasonable. When the prior means coincide with the true value of r (at 2 and 8), the posterior means are similar to the true values. We also conclude that the classification accuracies and FPRs are insensitive to the prior degrees of freedom. As the amount of spatial correlation increases, classification accuracies in general tend to increase as

True r	$\hat{r}$	Accuracy (%)	FPR (%)
0.001	1.01 (0.03)	87.95 (0.08)	1.22 (0.03)
2	2.98 (0.06)	89.70 (0.08)	1.20 (0.03)
8	5.10 (0.11)	89.57 (0.08)	1.14 (0.02)
20	5.58 (0.11)	90.31 (0.08)	0.82 (0.02)

(a)

True r	$\hat{r}$	Accuracy (%)	FPR (%)
0.001	3.06 (0.10)	88.18 (0.08)	1.11 (0.03)
2	5.33 (0.10)	89.70 (0.08)	1.21 (0.03)
8	7.46 (0.14)	90.00 (0.08)	1.20 (0.03)
20	7.92 (0.17)	90.13 (0.08)	0.78 (0.02)

(b)

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

True r	$\hat{r}$	Accuracy (%)	FPR (%)
0.001	2.35 (0.06)	97.15 (0.04)	1.77 (0.03)
2	3.89 (0.07)	97.76 (0.04)	1.79 (0.03)
8	6.23 (0.11)	97.91 (0.04)	2.21 (0.03)
20	7.24 (0.12)	98.11 (0.03)	2.21 (0.03)

(a)

True r	$\hat{r}$	Accuracy (%)	FPR (%)
0.001	4.74 (0.13)	97.08 (0.04)	2.07 (0.04)
2	6.37 (0.11)	97.68 (0.04)	1.83 (0.03)
8	8.80 (0.15)	97.88 (0.04)	2.04 (0.03)
20	10.00 (0.17)	98.05 (0.03)	2.46 (0.04)

(b)

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

well due to the fact that generally it is easier to classify images with higher amounts of clustering. We also point out that the differences in classification accuracies as r increases are less pronounced when $\beta_{v,1} = 5$. As the signal increases, most reasonable methods start to perform similarly.

3.2 Simulation 2

It is well-established that ignoring temporal correlation results in poor classification. Monti (2011) provides a detailed explanation and some numerical results are reported by Lee et al. (2014) and Musgrove et al. (2016). However, the performance of models that do not fully consider spatial dependence has not been widely studied. We therefore

True r	Accuracy (%)	FPR (%)
0.001	87.52 (0.08)	1.18 (0.02)
2	87.68 (0.08)	1.07 (0.02)
8	86.35 (0.09)	1.04 (0.02)
20	86.61 (0.09)	0.75 (0.02)

(a)

True r	Accuracy (%)	FPR (%)
0.001	97.01 (0.04)	1.32 (0.03)
2	97.10 (0.04)	1.38 (0.03)
8	96.88 (0.04)	1.44 (0.03)
20	97.03 (0.04)	1.43 (0.03)

(b)

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

investigate what happens when default prior for $\pi(\gamma_v = 1)$ described in Smith and Kohn (1996) is employed. We assume that $\pi(\gamma_v = 1) = 1/2$ for all v and modify our sampler to draw from the posterior distribution

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

We use this modified sampling scheme on the same simulated data sets that were used in Section 3.1.

Table 3 displays the results when the default prior is used for γ. 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 this prior.

As the amount of spatial correlation in the generated images increases, the performance of the default prior gets worse compared to the spatial hierarchical prior. In fact, when $r = 0.001$, the classification accuracies are nearly identical, but are markedly worse when $r = 20$. Another point worth mentioning is that 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 similarly since activation is easier to detect. From this simulation, we conclude that the default prior for γ leads to poor classification accuracy when the spatial correlation in images is high. This prior also does not provide a way of obtaining information about the amount of spatial dependence in the images from the posterior distribution, and therefore has poorer inferential properties compared to the spatial hierarchical prior.

3.3 Simulation 3

We now consider the case when there is no activation in the image. We set all $\gamma_v = 0$ and $\beta_{v,1} = 0$. We examine the situation under both χ^2_2 and χ^2_8 priors.

Prior	Accuracy (%)	FPR (%)
χ^2_2	99.91 (0.01)	0.09 (0.01)
χ^2_8	99.90 (0.01)	0.10 (0.01)

Table 4: Classification accuracies and false positive rates when there is no activation in the images, using both χ^2_2 and χ^2_8 priors.

The results are presented in Table 4. Under both priors we see that the false alarm rate is low, providing evidence that our method will work well even in situations when there is no activation in the images.

4 Parcellation Effects

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 observe how the results vary, primarily focusing on the effects of the parcellation scheme described in Section 2.2.

In this simulation, we use the 20×20 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 3.1, considering both $\beta_{v,1} = 3$ and $\beta_{v,1} = 5$ when $\gamma_v = 1$. We only consider the χ^2_8 prior for r.

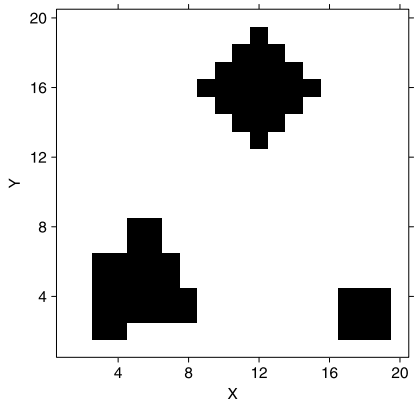
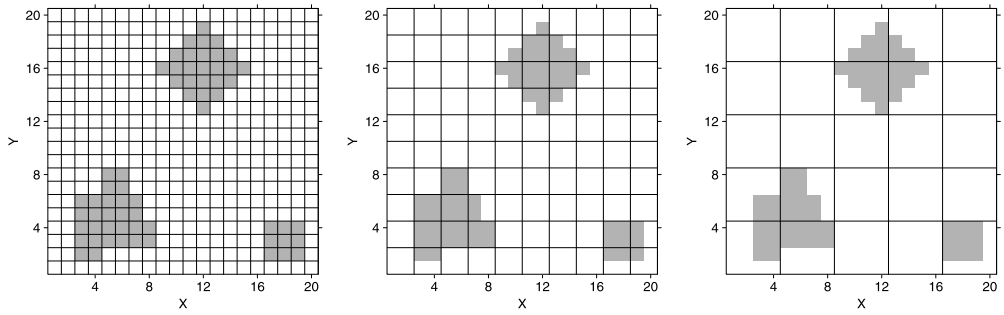


Figure 2: The image used in the parcellation effects simulation.

We use our method when the image is unparcellated ($G = 400$), divided into a grid of 2×2 squares ($G = 100$), and divided into a grid of 4×4 squares ($G = 25$). Figure 3 displays the parcellations considered in this section. We study how the classification accuracy, false positive rate, and false negative rate (FNR) change with the different parcellations.



(a) (b) (c)

Figure 3: The parcellations when (a) $G = 400$, (b) $G = 100$, and (c) $G = 25$.

G	$\hat{r}$	Accuracy (%)	FPR (%)	FNR (%)	Relative Time
400	10.04 (0.14)	97.00 (0.04)	0 (0)	19.35	1.00
100	9.79 (0.07)	97.50 (0.04)	0.30 (0.01)	14.52	0.05
25	9.08 (0.04)	96.50 (0.05)	0.30 (0.01)	20.97	0.01

(a)

G	$\hat{r}$	Accuracy (%)	FPR (%)	FNR (%)	Relative Time
400	9.41 (0.10)	99.50 (0.02)	0 (0)	3.22	1.00
100	9.30 (0.06)	99.25 (0.02)	0.30 (0.01)	3.22	0.05
25	8.70 (0.04)	99.25 (0.02)	0.59 (0.02)	1.61	0.01

(b)

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

Table 5 displays the results of the three parcellation schemes when applied to our image. The classification accuracies are similar under each parcellation scheme, and this is especially true when $\beta_{v,1} = 5$. In addition to the usual accuracy statistics, we also report the computation time relative to the $G = 400$ case and see substantial speed improvements as the number of regions decreases. This is mostly attributed to the reduction in dimension of Γ_j. Figures 4 and 5 display the posterior activation probabilities and classified images at both activation amplitudes under the three parcellation schemes.

One point worth noting is the large decrease in FNR when $G = 100$ in Table 5a. The reason for this decrease can be seen by looking at the parcellation in Figure 3b. When the entire image is subdivided into a grid of 2×2 voxels, the 2×2 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 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 $G = 100$ and $G = 400$. By looking at Figure 3c, we see that some of the 4×4 regions overlap with non-active regions. 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 $\beta_{v,1} = 5$ due to the stronger signal.

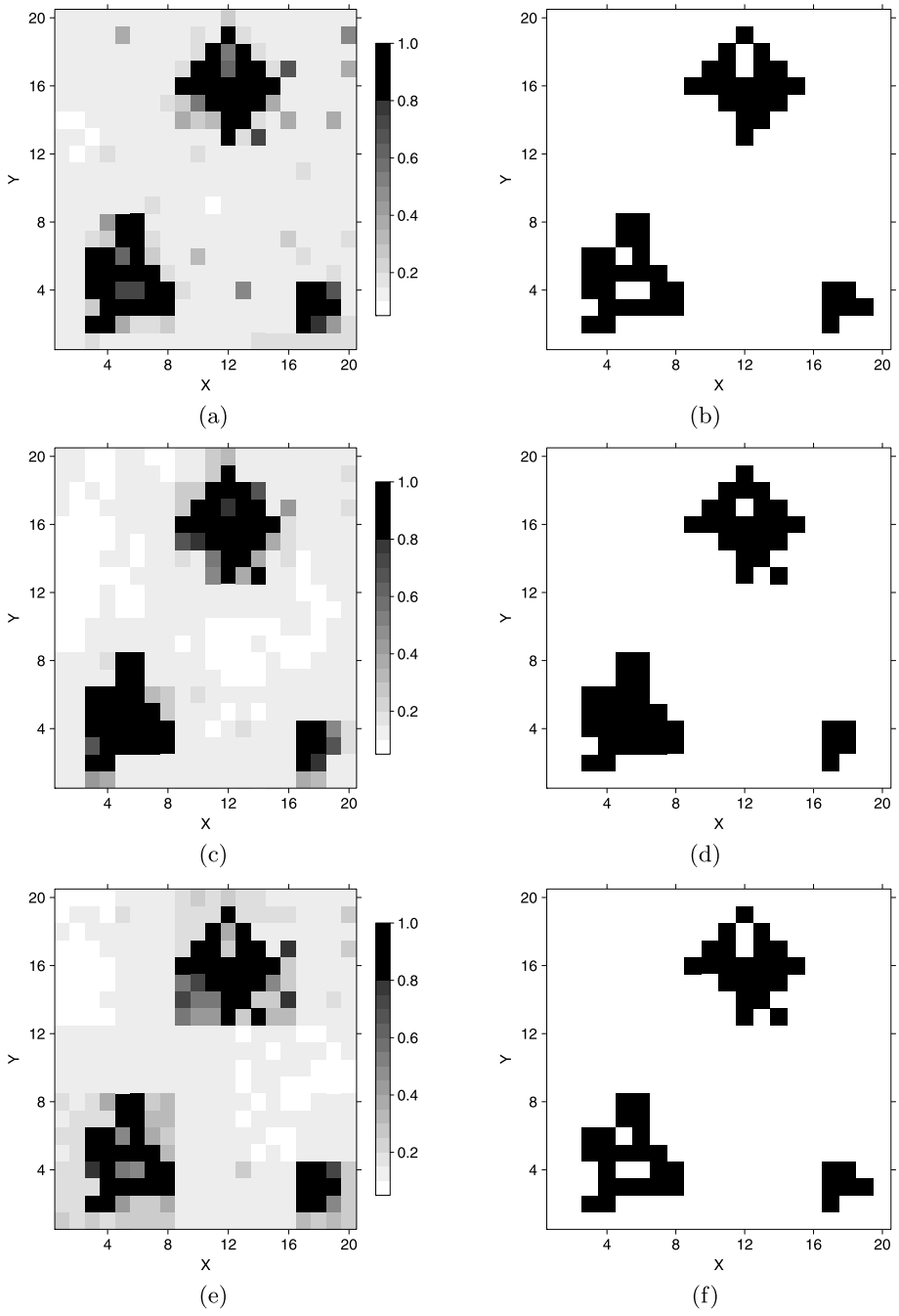


Figure 4: Posterior probabilities of activation and the classified images when (a)(b) $G = 400$, (c)(d) $G = 100$, and (d)(e) $G = 25$ and $\beta_{v,1} = 3$.

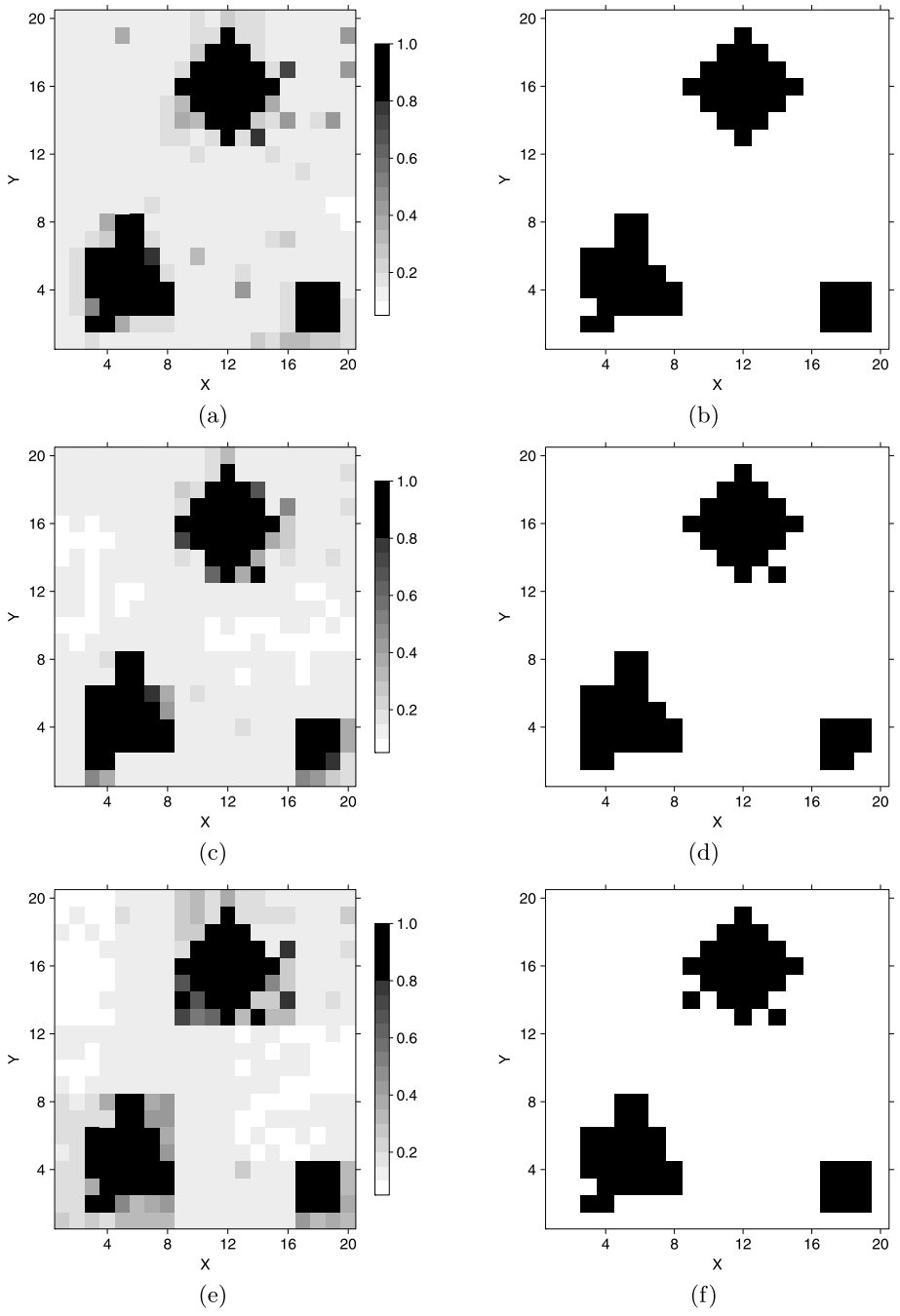


Figure 5: Posterior probabilities of activation and the classified images when (a)(b) $G = 400$, (c)(d) $G = 100$, and (d)(e) $G = 25$ and $\beta_{v,1} = 5$.

We suggest using smaller regions at locations where a priori information suggests activation and reinforces the idea that the regions be chosen according to prior anatomical knowledge. If no prior spatial information is available, we recommend choosing G as large as possible given computational constraints.

5 Data Example

We now demonstrate the proposed methodology on an fMRI data set. A second example, which analyzes the benchmark auditory data from the Wellcome Trust Centre for Neuroimaging, is contained in the supplementary materials.

5.1 Emotion Processing Data

We consider data that 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 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 the regions 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.

The researcher's main goals were to determine (1) which regions of the brain were activated by the two different tasks and (2) did the total amount of brain activation differ under both tasks. This provided them insight into how the brain processes facial expression (emotion) data, compared to how it processed non-emotional (shape) data.

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.

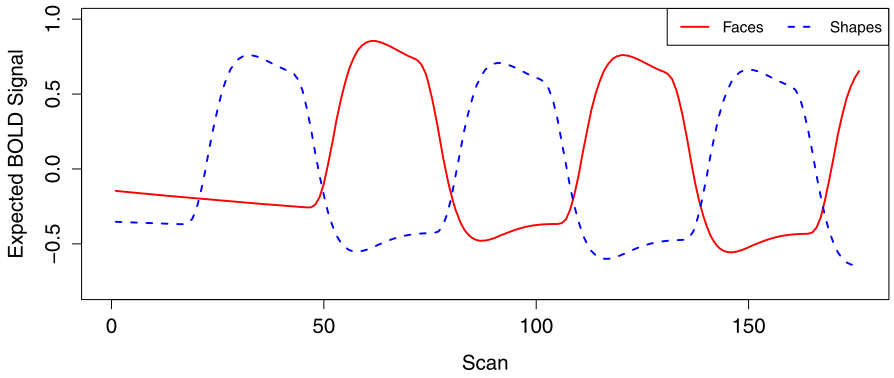


Figure 6: The design matrix used in the HCP analysis.

We use a χ^2_{12} prior for the task-specific spatial dependence parameters r_{face} and r_{shape}. The design matrix used in the statistical analysis is displayed in Figure 6.

Computation

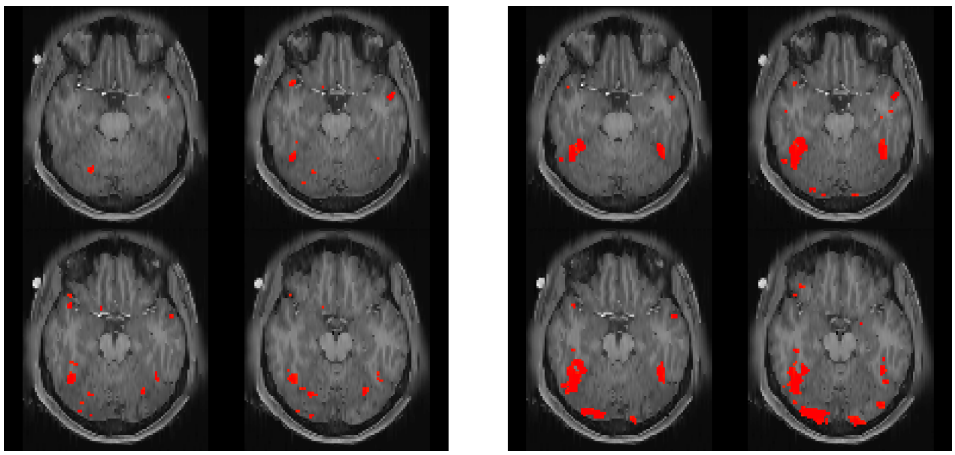
The posterior of interest (9) is intractable, and so we used the MCMC method described in the supplementary material to draw 100,000 MCMC samples. The tuning parameters of the MCMC algorithm were chosen so as to produce acceptance rates close to 50% in the Metropolis-Hastings steps. Standard errors were all less than 2% of their posterior estimates and are therefore omitted in this section. Diagnostic measures used to assess convergence are contained in the supplementary material.

As a final note, this method took slightly more than 3 hours on a single core of an Intel i7-4770 3.5 GHz processor, demonstrating computational feasibility.

Results

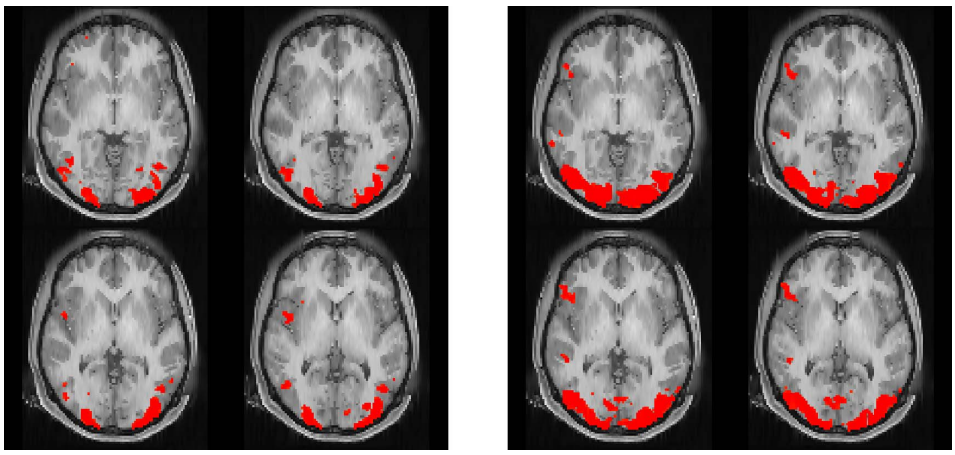
Figures 7–10 display the detected activation in several horizontal slices. Note that each slice is two mm thick and slice 1 contains the topmost region of the brain. Most of the activation occurs in the occipital lobe, which is thought to be responsible for the processing of visual information. Figures 7 and 8 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.



(a) (b)

Figure 7: Neuronal activation in slices 15–18 during the (a) shape blocks and (b) face blocks.

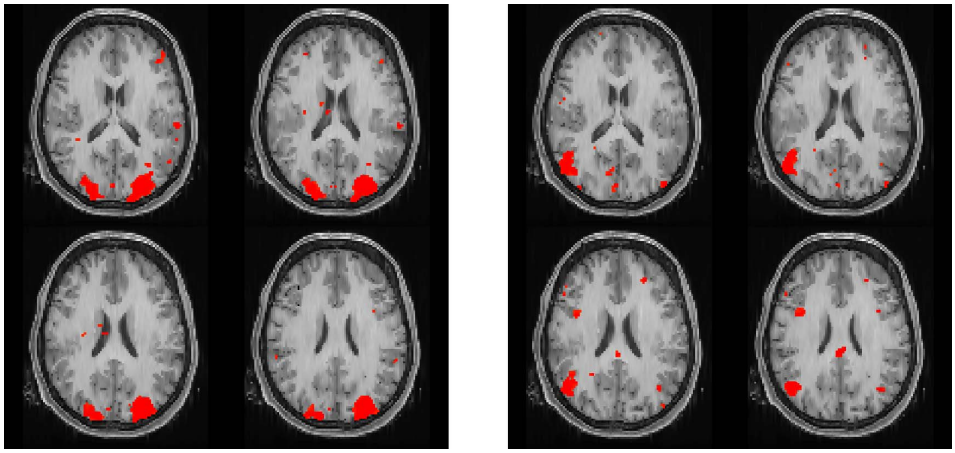


(a) (b)

Figure 8: Neuronal activation in slices 25–28 during the (a) shape blocks and (b) face blocks.

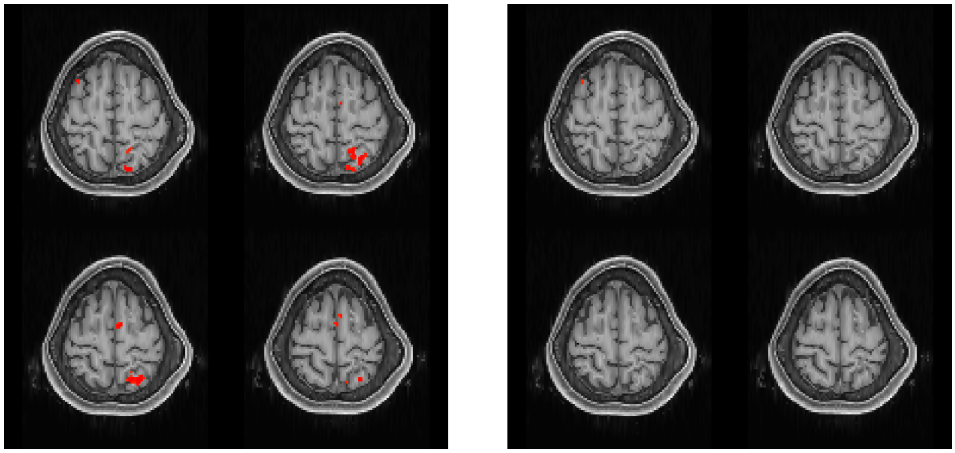
Temporal Correlation

The maximum likelihood estimates of ρ_v in five slices are displayed in Figure 11. We see that temporal correlation tends to be higher in the areas that displayed activation in Figures 7–10, showing that voxel-wise temporal independence is an unreasonable assumption.



(a) (b)

Figure 9: Neuronal activation in slices 36–39 during the (a) shape blocks and (b) face blocks.



(a) (b)

Figure 10: Neuronal activation in slices 57–60 during the (a) shape blocks and (b) face blocks.

6 Final Remarks

We proposed a novel spatiotemporal Bayesian variable selection model with a focus on its use in single-subject functional neuroimaging applications. The main advances are the use of a hierarchical spatial prior in conjunction with a parcellation of the image. We demonstrated via simulation that the resulting inferences are insensitive

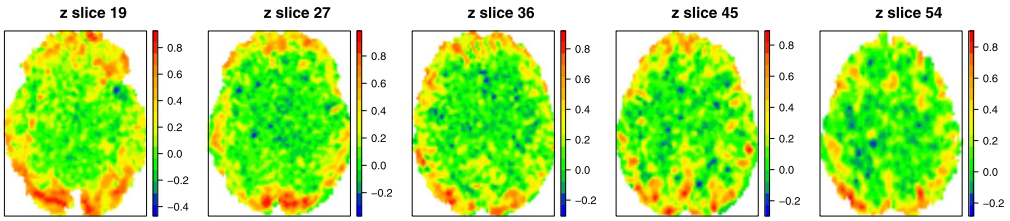


Figure 11: Maximum likelihood estimates of the AR(1) coefficients at five vertical slices in the HCP analysis.

to the parcellation, but we recommend that no more than 500 regions be used for computational reasons. The methodology offers several key advantages over existing procedures. In particular, it adequately models the spatial dependence in fMRI data while avoiding the computational issues encountered under the Ising model. We also paid careful attention to the design and implementation of MCMC methodology, including a method for choosing default starting values and a method for assessing when the MCMC sampling should terminate (described in the supplementary material). The procedure has been shown to be computationally feasible for a whole-brain analysis in that it can be completed in only a few hours on standard hardware.

Despite the fact that our model works well and is computationally reasonable, there are many potential modifications to consider. We have discussed some of these at various points, but we will mention a few more now. For example, the Zellner g-prior for $\beta(\gamma)$ is widely-used, but an obvious alternative is to use spike and slab priors which are common in variable selection problems and if used in conjunction with an inverse-gamma prior on σ_v^2, then it is straightforward to draw from the full conditionals of β and σ^2.

We also assume that the σ_v^2 and ρ_v are independent a priori. This could likely be improved since groups of voxels may have similar values. In Figure 11, there is obvious spatial clustering in the estimates of the ρ_v. Priors with spatial correlation for the σ_v^2 and ρ_v are an appealing possibility for future work.

The issue of functional connectivity has also recently received a great deal of attention. The basic idea behind functional connectivity is to describe long-range dependence in activation patterns throughout the brain across different tasks. While our method assumes distance-based spatial dependence, it can easily be extended to model long-range dependence through the use of an Inverse-Wishart prior for the $S_{(j)}$ instead of the prior described in (5). This method was utilized by Bowman et al. (2007) successfully, albeit in a simpler modeling situation. While intuitively simple, computation and model fitting likely would require substantial effort.

We have limited attention to single-subject experiments. It is natural to want to extend our approach to handle multiple subjects. The biggest obstacle is the increased computational burden that would result. For example, Zhang et al. (2016) recently used a Bayesian variable selection approach for multi-subject data and reported that 1,000 iterations of their MCMC algorithm required seven hours, which makes it infeasible in practical settings. However, they took a modeling approach more closely associated

with the Ising prior approach in Lee et al. (2014) and Smith and Fahrmeir (2007) than with the current paper. Thus we believe that there is room for improvement by using an areal model. Indeed we believe there is a bright future for the development of spatiotemporal Bayesian variable selection areal models for both single-subject and multi-subject neuroimaging applications.

Supplementary Material

Supplemental Material for "Bayesian Spatiotemporal Modeling using Hierarchical Spatial Priors, with Applications to Functional Magnetic Resonance Imaging" (DOI: 10.1214/18-BA1108SUPP; .pdf).

References

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2003). *Hierarchical Modeling and Analysis for Spatial Data*. New York: Chapman and Hall/CRC, 1st edition. MR3362184. 1265

Bezener, M., Hughes, J., and Jones, G. (2018). "Supplemental Material for "Bayesian Spatiotemporal Modeling using Hierarchical Spatial Priors, with Applications to Functional Magnetic Resonance Imaging"." *Bayesian Analysis*. doi: https://doi.org/10.1214/18-BA1108SUPP. 1263

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). "Variational inference: a review for statisticians." *Journal of the American Statistical Association*, 112: 859–877. MR3671776. doi: https://doi.org/10.1080/01621459.2017.1285773. 1262

Bowman, F. D. (2014). "Brain Imaging Analysis." *Annual Review of Statistics and Its Application*, 1: 61–85. 1262

Bowman, F. D., Caffo, B., Bassett, S. S., and Kilts, C. (2007). "A Bayesian Hierarchical Framework for spatial modeling of fMRI data." *NeuroImage*, 39: 146–156. 1262, 1281

Cipra, B. (1987). "An Introduction to the Ising Model." *American Mathematical Monthly*, 94: 937–959. MR0936054. doi: https://doi.org/10.2307/2322600. 1262

Cressie, N. A. (1993). *Statistics for Spatial Data*. New York: Wiley Interscience, Revised edition. MR1239641. doi: https://doi.org/10.1002/9781119115151. 1265

Essen, D. C. V., Smith, S. M., Barch, D. M., Behrens, T. E., Yavoub, E., and Ugurbil, K. (2013). "The WU-Minn Human Connectome Project: An overview." *NeuroImage*, 62–79. 1275

Flegal, J. M., Haran, M., and Jones, G. L. (2008). "Markov Chain Monte Carlo: Can We Trust the Third Significant Figure?" *Statistical Science*, 23: 250–260. MR2516823. doi: https://doi.org/10.1214/08-STS257. 1269

Flegal, J. M., Hughes, J., Vats, D., and Dai, N. (2017). *mcmcse: Monte Carlo Standard Errors for MCMC*. Riverside, CA, Denver, CO, Coventry, UK, and Minneapolis, MN. R package version 1.3–2. 1269, 1270

Friston, K. J., Ashburner, J. T., Kiebel, S. J., Nichols, T. E., and Penny, W. D. (2007). *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. London: Academic Press. 1261, 1262, 1264

Friston, K. J., Holmes, A., Worsley, K. J., Polin, J. B., Frith, C., and Frackowik, R. (1995). "Statistical parametric maps in functional imaging: A general linear approach." *Human Brain Mapping*, 2: 189–210. 1262

Friston, K. J., Worsley, K., Frackowiak, R., Mazziotta, J., and Evans, A. (1994). "Assessing the significance of focal activations using their spatial extent." *Human Brain Mapping*, 1: 210–220. 1262

Genovese, C. R. (2000). "A Bayesian Time-Course Model for Functional Magnetic Resonance Imaging Data." *Journal of the American Statistical Association*, 95: 691–703. 1262

George, E. I. and McCulloch, R. E. (1993). "Variable Selection Via Gibbs Sampling." *Journal of the American Statistical Association*, 88: 881–889. 1262

George, E. I. and McCulloch, R. E. (1997). "Approaches for Bayesian Variable Selection." *Statistica Sinica*, 7: 339–373. 1262

Gössel, C., Auer, D., and Fahrmeir, L. (2001). "Bayesian Spatiotemporal Inference in Functional Magnetic Resonance Imaging." *Biometrics*, 57: 554–562. MR1855691. doi: https://doi.org/10.1111/j.0006-341X.2001.00554.x. 1262

Haran, M. (2011). "Gaussian random field models for spatial data." In Brooks, S. P., Gelman, A. E., Jones, G. L., and Meng, X. L. (eds.), *Handbook of Markov Chain Monte Carlo*, 449–478. London: Chapman and Hall/CRC. MR2858457. 1262, 1265

Hariri, A. R., Mattay, V. S., Tessitore, A., Kolachana, B., Fera, F., Goldman, D., Egan, M. F., and Weinberger, D. R. (2002). "Serotonin Transporter Genetic Variation and the Response of Human Amygdala." *Science*, 297: 400–4003. 1275

Huettel, S. A., Somng, A. W., and McCarthy, G. (2009). *Functional Magnetic Resonance Imaging*. Sunderland, MA: Sinauer Associates. 1261

Jones, G. L., Haran, M., Caffo, B. S., and Neath, R. (2006). "Fixed-width output analysis for Markov chain Monte Carlo." *Journal of the American Statistical Association*, 101: 1537–1547. MR2279478. doi: https://doi.org/10.1198/016214506000000492. 1269, 1270

Käll, L., Storey, J. D., MacCoss, M. J., and Noble, W. S. (2008). "Posterior error probabilities and false discovery rates: two sides of the same coin." *Journal of Proteome Research*, 7: 40–44. 1269

Kaushik, K., Karesh, K., and Suresha, D. (2013). "Segmentation of the white matter from the brain fMRI images." *International Journal of Advanced Research in Computer Engineering and Technology*, 2: 1314–1317. 1261

Landman, B. A., Yang, X., and Kang, H. (2012). "Do we really need robust and altrernative inference methods for brain MRI?" In Yap, P., Liu, T., Shen, D., and Westin, C.

(eds.), *MBIA 2012: Multimodal Brain Image Analysis*, volume 7509 of *Lecture Notes in Computer Science*, 77–93. Berlin: Springer. 1262

Lazar, N. A. (2008). *The Statistical Analysis of fMRI Data*. New York: Springer. MR2597019. 1261, 1262

Lee, K.-J., Jones, G. L., Caffo, B. S., and Bassett, S. S. (2014). "Spatial Bayesian Variable Selection Models on Functional Magnetic Resonance Imaging Time-Series Data." *Bayesian Analysis*, 9: 699–732. MR3256061. doi: https://doi.org/10.1214/14-BA873. 1262, 1263, 1264, 1269, 1271, 1282

Lindquist, M. A. (2008). "The Statistical Analysis of fMRI Data." *Statistical Science*, 23: 439–464. MR2530545. doi: https://doi.org/10.1214/09-STS282. 1261, 1263, 1264

Locascio, J., Jennings, P. J., Moore, C. I., and Corkin, S. (1997). "Time series analysis in the time domain and resampling methods for studies of functional magnetic brain imaging." *Human Brain Mapping*, 168–193. 1263

Makni, S., Idier, J., Vincent, T., Thirion, B., Dehaene-Lambertz, G., and Ciuciu, P. (2008). "A fully Bayesian approach to the parcel-based detection-estimation of brain activity in fMRI." *NeuroImage*, 41: 941–969. 1264

Mikl, M., Mareček, R., Hluštík, P., Pavlicová, M., Drastich, A., Chlebus, P., Brázdil, M., and Krupa, P. (2008). "Effects of spatial smoothing on fMRI group inferences." *Magnetic Resonance Imaging*, 26: 490–503. 1261

Monti, M. M. (2011). "Statistical analysis of fMRI time-series: A critical review of the GLM approach." *Frontiers in Human Neuroscience*, 5. 1263, 1271

Morris, R., Descombes, X., and Zerubia, J. (1996). "The Ising/Potts model is not well suited to segmentation tasks." In *Digital Signal Processing Workshop Proceedings*, 263–265. IEEE. 1262

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge: The MIT Press. 1262

Musgrove, D. R., Hughes, J., and Eberly, L. E. (2016). "Fast, fully Bayesian spatiotemporal inference for fMRI data." *Biostatistics*, 17: 291–303. MR3516001. doi: https://doi.org/10.1093/biostatistics/kxv044. 1263, 1269, 1271

Penny, W., Kiebel, S., and Friston, K. (2003). "Variational Bayesian inference for fMRI time series." *NeuroImage*, 19: 727–741. 1262, 1263

Penny, W. D., Trujillo-Barreto, N. J., and Friston, K. J. (2005). "Bayesian fMRI time series analysis with spatial priors." *NeuroImage*, 24: 350–362. 1262

Quirós, A., Diez, R. M., and Wilson, S. P. (2010). "Bayesian spatiotemporal model of fMRI data using transfer functions." *NeuroImage*, 52: 9995–1004. 1262

Raftery, A. (1996). "Hypothesis Testing and Model Selection." In Gilks, W., Spiegelhalter, D., and Richardson, S. (eds.), *Markov Chain Monte Carlo in Practice*. Lon-

don: Chapman and Hall. MR1397966. doi: https://doi.org/10.1007/978-1-4899-4485-6. 1269

Smith, D. and Smith, M. (2006). "Estimation of Binary Markov Random Fields Using Markov Chain Monte Carlo." *Journal of Computational and Graphical Statistics*, 15: 207–227. MR2252462. doi: https://doi.org/10.1198/106186006X97817. 1262, 1269

Smith, M. and Fahrmeir, L. (2007). "Spatial Bayesian Variable Selection with Application to Functional Magnetic Resonance Imaging." *Journal of the American Statistical Association*, 102: 417–431. MR2370843. doi: https://doi.org/10.1198/016214506000001031. 1262, 1265, 1269, 1282

Smith, M. and Kohn, R. (1996). "Nonparametric regression using Bayesian variable selection." *Econometrics*, 75: 317–343. 1272

Smith, M., Pütz, B., Auer, D., and Fahrmeir, L. (2003). "Assessing brain activity through spatial Bayesian variable selection." *NeuroImage*, 20. 1262

Storey, J. D. (2003). "The positive false discovery rate: a Bayesian interpretation and the q-value." *The Annals of Statistics*, 31: 2013–2035. MR2036398. doi: https://doi.org/10.1214/aos/1074290335. 1269

Triantafyllou, C., Hoge, R., and Wald, L. (2006). "Effect of spatial smoothing on physiological noise in high-resolution fMRI." *NeuroImage*, 32: 551–557. 1261

Vats, D., Flegal, J. M., and Jones, G. L. (2016). "Multivariate output analysis for Markov chain Monte Carlo." *Preprint arXiv:1512.07713*. MR3653667. 1269, 1270

Woolrich, M. W., Jenkinson, M., Brady, J. M., and Smith, S. M. (2004). "Fully Bayesian Spatio-Temporal Modeling of fMRI Data." *IEEE Transactions on Medical Imaging*, 23: 213–231. 1262

Worsley, K. (2003). "Detecting activation in fMRI data." *Statistical Methods in Medical Research*, 12: 401–418. MR2005444. doi: https://doi.org/10.1191/0962280203sm340ra. 1262

Worsley, K., Marrett, S., Neelin, P., and Evans, A. (1992). "A three-dimensional statistical analysis for CBF activation studies in human brain." *Journal of Cerebral Blood Flow and Metabolism*, 12: 900–918. 1262

Worsley, K. J., Liao, C. H., Aston, J., Petre, V., Duncan, G. H., Morales, F., and Evans, A. C. (2002). "A General Statistical Analysis for fMRI Data." *NeuroImage*, 15: 1–15. 1262

Xia, J., Liang, F., and Wang, Y. M. (2009a). "FMRI analysis through Bayesian variable selection with a spatial prior." In *Proceedings of the 6th IEEE International Symposium on Biomedical Imaging*, 714–717. IEEE. 1262

Xia, J., Liang, F., and Wang, Y. M. (2009b). "fMRI analysis through Bayesian variable selection with a spatial prior." *IEEE Int. Symp. on Biomedical Imaging (ISBI)*, 714–717. 1263

Zellner, A. (1996). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions." *In Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti North-Holland/Elsevier*, 233–243. MR0881437. 1265

Zhang, L., Guindani, M., and Vannucci, M. (2015). "Bayesian models for functional magnetic resonance imaging data analysis." *WIREs Computational Statistics*, 7: 21–41. MR3348719. doi: https://doi.org/10.1002/wics.1339. 1262

Zhang, L., Guindani, M., Versace, F., Engelmann, J. M., and Vannucci, M. (2016). "A spatio-temporal nonparametric Bayesian variable selection model of multi-subject fMRI data." *The Annals of Applied Statistics*, 10: 638–666. MR3528355. doi: https://doi.org/10.1214/16-AOAS926. 1262, 1269, 1281

Zhang, L., Guindani, M., Versace, F., and Vannucci, M. (2014). "A spatio-temporal non-parametric Bayesian variable selection model of fMRI data for clustering correlated time courses." *NeuroImage*, 95: 162–175. 1262

Zhou, X. and Schmidler, S. C. (2009). "Bayesian Parameter Estimation in Ising and Potts Models: A Comparative Study with Applications to Protein Modeling." Technical report, Duke University. 1262

Acknowledgments

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.

Invited Discussion

Raquel Prado*

I would like to begin by congratulating the authors for a new and valuable contribution to modeling task-related large-dimensional functional brain imaging data from a Bayesian perspective. My discussion focuses on highlighting some of the modeling and computational aspects of this approach and how it compares to alternative approaches for brain imaging data. I also provide a discussion of possible future extensions.

Modeling features and alternative approaches

The proposed Bayesian approach considers a spatio-temporal model for analyzing functional magnetic resonance images recorded on a subject during a task-related experiment. It uses binary variables for labeling which voxels (volumetric pixels) are active under a particular task. Spatial dependence across voxels is then described using an underlying aereal model on a parcellation of the image. This is one of the main features of the proposed approach, as the parcellation results in substantial dimension reduction that leads to more computationally efficient posterior inference via MCMC. Temporal dependence is induced by considering an autoregressive structure of order one, or AR(1), on the error process, with voxel dependent AR coefficients. Model performance is illustrated in simulation studies. The model is also applied to fMRI data collected as part of the Human Connectome Project (HCP). In particular, fMRI data from a randomly selected subject who completed two tasks arranged in a block designed experiment were analyzed. The main goal of the analysis was to determine which regions of the subject's brain were active under the two different tasks, and if patterns of brain activation were different for the two tasks.

Parcellation and anatomical information

As mentioned above, I think one of the most appealing features of the proposed approach is the parcellation of the image into regions that share some common parameters in order to achieve dimension reduction and more efficient posterior computations. This is particularly important when dealing with whole-brain fMRI analysis. In this sense and as mentioned in the paper, the proposed approach shares similarities with that of Musgrove et al. (2016). The main difference lies in the assumption that the parcels are independent in Musgrove et al. (2016) but dependent here. The main idea here is that, for a given task j and a given parcellation consisting of G regions, all the voxels in the same region $\mathcal{R}_g$, will share the same underlying spatial random effects denoted by $S_{g,j}$. In other words, it is assumed that, for all $v \in \mathcal{R}_g$, the indicators of activation depend

*Department of Statistics, Baskin School of Engineering, Mail-Stop SOE2, University of California Santa Cruz, 1154 High St, Santa Cruz, CA 95060, raquel@soe.ucsc.edu

on the region spatial effect $S_{g,j}$, i.e.,

$$\gamma_{v,j}|S_{g,j} \sim \text{Bern}\left(\frac{1}{1+e^{-S_{g,j}}}\right),$$

and then $S_{(j)} = (S_{1,j}, \ldots, S_{G,j})$ are modeled using a Gaussian process. Section 3 presents some results that illustrate the effect of the number of parcels can have on the activation results via simulation studies, but only equally-sized parcels are considered. No studies are included to show the effect that the shape and dimension of the individual parcels may have in the posterior activation results. Furthermore, it is clear that the parcellation in the aerial spatial model induces edge-type of effects on the posterior probabilities of activations (e.g., see Figure 5 (e)) which may be problematic. Do the authors have any suggestions on how to deal with these effects in practice? It is mentioned that anatomical knowledge could be used to determine the parcellation. This is certainly a very good idea, however, anatomical restrictions will lead to regions that are not equally-sized.

Whole-brain analysis versus analysis of 2D slices

All the simulation studies involve analysis of 2D slices as opposed to full 3D volumes. Also, in the analysis of the single-subject fMRI data, posterior activation results are presented for specific collections of slices, i.e., slices 15–18 in Figure 7, slices 25–28 in Figure 8, slices 36–39 in Figure 9 and slices 57–60 in Figure 10. Why are the results shown for these specific slices? Is there a better way to summarize the results for the full 3D volume? Another interesting question is whether results obtained from separate 2D analyses of the individual slices will be compatible with those obtained from considering a single analysis of the entire 3D image.

HRF choice

The proposed approach assumes a fixed HRF for all voxels and considers the canonical HRF. It has been shown that in some situations the choice of the HRF considerably affects the activation results. In addition, the HRFs may be different for different voxels/regions. There are a number of approaches that consider HRFs estimation and/or joint inference on activation, connectivity and HRFs (e.g., Woolrich et al., 2004; Lindquist et al., 2009; Yu et al., 2016, among others). The focus of this paper is on studying the main features induced by the spatial hierarchical priors and the parcellation in the context of activation detection at the voxel-level. Full posterior HRF inference at this resolution would be too computationally expensive even for single-subject analysis. However, sensitivity studies could be done to assess the effect of the HRF choice in the posterior results. This is the path we have taken in Yu et al. (2018) where we found that for the data sets we considered (complex-valued fMRI as opposed to magnitude only fMRI) the choice of the HRF did not have a huge impact on the activation results. How sensitive are the activation results, especially those obtained for the human fMRI analysis presented in Section 5, to the choice of the HRF function?

Comparison with alternative approaches

No comparisons with alternative approaches are provided in the paper. Have the authors considered formal comparisons with the approach of Musgrove et al. (2016) either in simulation or real settings? Are there substantial differences between the two methods in terms of classification accuracy and false positive rates? I do believe that the incorporation of spatial structure within a Bayesian modeling framework as done in the paper can lead to improved activation results, particularly when this method is compared to some of the methods based on simpler models that are routinely used and implemented in software platforms such as FSL and SPM, however, it is difficult to assess if this is the case given that no comparisons are included.

Multi-subject analysis and additional covariates

The paper briefly discusses future extensions to consider multi-subject models. The proposed modeling approach offers advantages with respect to other approaches currently available for voxel-level fMRI data given the dimension reduction induced by the parcellation. However, before discussing the computational feasibility of the proposed approach in practical settings there are issues related to the actual modeling of the data. For instance, would it make sense to consider the same parcellation across subjects or should parcellations be subject-dependent? Would it be reasonable to have common spatial effects across multiple subjects with additional effects that depend on the individual subjects? Also, in multi-subject analysis there are usually additional covariates for each subject. How do you suggest to incorporate these covariates?

Additional computational and modeling aspects

This paper focuses on obtaining full posterior inference, however, even with the dimension reduction induced by the parcellation full posterior inference is achieved via MCMC and can be very costly in multi-subject studies. Have the authors considered approximations such as those based on variational inference? How about EM-based algorithms for posterior estimation?

Finally, there is growing literature on tensor regression models for analyzing neuroimaging data (see, e.g., Zhou et al., 2013; Li et al., 2018; Guhaniyogi et al., 2017). What are the main advantages/disadvantages and similarities/differences of considering tensor-based models versus the modeling approach considered here?

References

Guhaniyogi, R., Qamar, S., and Dunson, D. (2017). "Bayesian Tensor Regression." *Journal of Machine Learning Research*, 18: 1–31. MR3714242. 1289

Li, X., Xu, D., Zhou, H., and Li, L. (2018). "Tucker tensor regression and neuroimaging analysis." *Statistics in Biosciences*, 1–26. MR3763769. 1289

Lindquist, M., Loh, J., Atlas, L., and Wager, T. (2009). "Modeling the hemodynamic response function in fMRI: Efficiency, bias and mis-modeling." *NeuroImage*, 45: S187–S198. 1288

Musgrove, D., Hughes, J., and Eberly, L. (2016). "Fast, fully Bayesian spatiotemporal inference for fMRI data." *Biostatistics*, 17: 291–303. MR3516001. doi: https://doi.org/10.1093/biostatistics/kxv044. 1287, 1289

Woolrich, M., Behrens, T., and Smith, S. (2004). "Constrained linear basis sets for HRF modeling using variational Bayes." *NeuroImage*, 21: 1748–1761. 1288

Yu, C., Prado, R., Ombao, H., and Rowe, D. (2018). "A Bayesian variable selection approach yields improved detection of brain activation from complex-valued fMRI." *Journal of the American Statistical Association*. 1288

Yu, Z., Prado, R., Burke Quinlan, E., Cramer, S., and Ombao, H. (2016). "Understanding the Impact of Stroke on Barin Motor Function: A hierarchical Bayesian approach." *Journal of the American Statistical Association*, 111: 549–563. MR3538686. doi: https://doi.org/10.1080/01621459.2015.1133425. 1288

Zhou, H., Li, L., and Zhu, H. (2013). "Tensor Regression with Applications in Neuroimaging Data Analysis." *Journal of the American Statistical Association*, 108(502): 540–552. MR3174640. doi: https://doi.org/10.1080/01621459.2013.776499. 1289

Invited Discussion

Per Sidén* and Mattias Villani†,‡

1 Introduction and context

Bezener, Hughes and Jones (BHJ) model brain activity using a Bayesian spatial variable selection approach first proposed in neuroimaging by Smith et al. (2003). Smith et al. (2003) use an Ising prior on the binary activity indicators. BHJ instead follow the approach in Kalus et al. (2014) where a Gaussian Markov Random Field (GMRF) is used to generate the spatially dependent activation indicators, $\Pr(\gamma_v = 1) = \Phi(w_v)$, where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of a standard normal variable, and $w_1, \ldots, w_N$ follow a GMRF; BHJ use the logistic CDF instead of the normal. The novelty in BHJ is the use of a hierarchically structured prior over functionally distinct brain regions, or parcels. Using a Gaussian process (GP) rather than a GMRF, this allows the authors to push the spatial model one level down the hierarchy to obtain spatially correlated random effects on the lower resolution parcel level, with attractive computational properties.

The other main approach for Bayesian spatial modeling in functional magnetic resonance imaging (fMRI) is to directly model the spatial dependence in the activation effects (β) with a GMRF or a GP, so that brain activity varies smoothly over neighboring voxels a priori (Penny et al., 2005).

To place BHJs contribution in context, Table 1 categorizes the various spatial priors proposed for fMRI data along two dimensions: i) the use of brain parcellation, and ii) whether the spatial dependence is directly on the activation effects (β) or on the activity indicators (γ). Note that the categorization treats slice-wise analysis as a special case of parcellation. Table 1 shows that the authors' work fill a previously empty gap in the resulting matrix.

Spatial dependence level	Spatial effects, β	Spatial indicators, γ
Voxel-level independent parcels (slice-wise)	Penny et al. (2005) Harrison et al. (2008) Groves et al. (2009)	Vincent et al. (2010) Lee et al. (2014) Zhang et al. (2014) Musgrove et al. (2016)
Parcel-level constant within parcels	Bowman et al. (2008)	**Bezener et al. (2018)**
Voxel-level whole brain	Harrison and Green (2010) Sidén et al. (2017)	Smith and Fahrmeir (2007)

Table 1: Categorization of spatial priors for fMRI analysis.

*Division of Statistics and Machine Learning, Department of Computer and Information Science, Linköping University, SE-581 83 Linköping, Sweden, per.siden@liu.se

†Division of Statistics and Machine Learning, Department of Computer and Information Science, Linköping University, SE-581 83 Linköping, Sweden, mattias.villani@liu.se

‡Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden

2 Pre-smoothing the data

BHJ follow the rather common practice of spatially smoothing the data as a pre-processing step before the spatial analysis. In classical fMRI analysis, three reasons for pre-smoothing the data are usually given (Lindquist et al., 2008): i) it improves the inter-subject registration for group analyses, ii) it helps satisfy the assumptions of the random field theory often used for multiple comparison correction, and iii) it may increase the signal-to-noise ratio. Points i) and ii) are not applicable to this paper. Regarding the third point, earlier work (for example Penny et al., 2005; Quirós et al., 2010) advocate using the prior to handle the spatial information instead of using pre-smoothing. Moreover, most papers on Bayesian spatial fMRI analysis only model the spatial dependence in the activity and neglect to model the spatial dependence in the noise. With pre-smoothing, this deficiency becomes even more serious as it introduces additional spatial noise.

To illustrate the problem with pre-smoothing, consider the following simple example. We disregard the temporal dimension and consider the model $y_v = \beta_v + \varepsilon_v$, with $\varepsilon_v \sim \mathrm{N}(0, \sigma^2)$ for a given voxel $v \in V$. That is, there is only one observation in each voxel which is a sum of the brain activity signal β_v and iid noise ε_v. We put a Gaussian process (GP) prior on $\{\beta_v\}_{v \in V}$ with a Matern covariance function with $\nu = 1$, and weakly informative priors on the marginal precision τ and the range ρ. The posterior is computed using integrated nested Laplace approximation (INLA) (Rue et al., 2009; Lindgren et al., 2011), using the `r-inla` implementation. The model is estimated on data simulated on a 20×20 grid, for two different true activity patterns; the one from Section 4 in BHJ with $\beta_v = 5$ in active voxels (Figure 1), and an activity pattern that

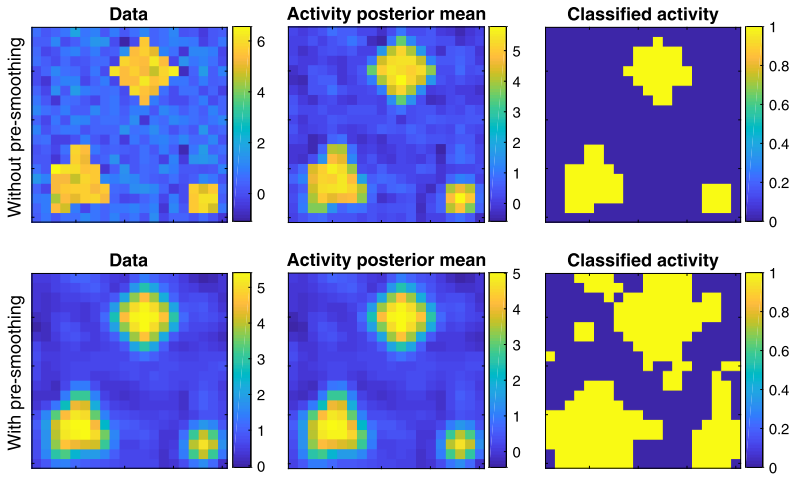


Figure 1: Comparison of results without pre-smoothing (top row) and with pre-smoothing (bottom row) for the activity pattern in Section 4. The figure displays the data (left column), the posterior mean of β_v (middle column), and the posterior classification of active voxels, $I_{\{\Pr(\beta_v > 0 | y) > 0.99\}}$ (right column).

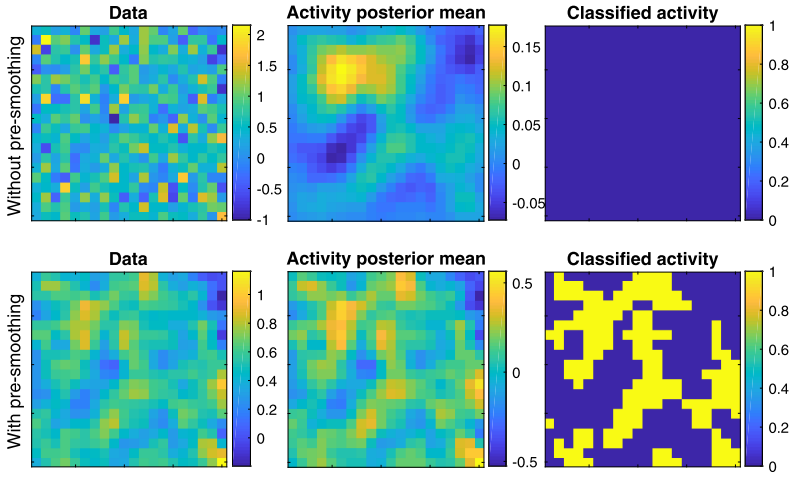


Figure 2: Comparison of results without pre-smoothing (top row) and with pre-smoothing (bottom row) for a zero-everywhere activity pattern. The figure displays the data (left column), the posterior mean of β_v (middle column), and the posterior classification of active voxels, $I_{\{\Pr(\beta_v>0|y)>0.99\}}$ (right column).

is zero everywhere, $\beta_v = 0$ (Figure 2). In both cases, the model is estimated both on the raw data and on the same data after pre-smoothing with a Gaussian kernel with FWHM $= 2$ voxels. The code for this example, and more details on the simulation setups, is available at https://bitbucket.org/psiden/BAdiscussion2018.

Figure 1 shows that the common problem with a GP prior of over-smoothing the edges is substantially worse when data are pre-smoothed, where voxels far away from the truly active areas are classified as active. This is due to the underestimation of the noise variance, since the noise has been pre-smoothed out and is instead interpreted as signal. Figure 2 shows that pre-smoothing the data from a zero-everywhere activity pattern makes the noise being misinterpreted as signal, resulting in many false positives.

Of course, the presented example is much simpler than a real spatio-temporal fMRI data set, and the dynamics of using a variable selection prior is somewhat different than the GP prior used here. Still, we believe that our example clearly illustrates unpleasant side effects from using pre-smoothing when using Bayesian spatial priors. It would be interesting if the authors could comment on the motivation for using pre-smoothing, and on any experiences they might have in applying their proposed model on raw data.

3 Parcellation and edge preservation

The motivation for using a parcellation is usually computational, but it can also have some additional nice side-effects. For example, spatial non-stationarity can be captured with different spatial hyperparameters in different parcels. Also, anatomically different

regions could potentially be separated, if chosen in some clever way. The authors suggest that "the parcellation should be chosen so that voxels within each region behave similarly due to their location", yet they provide no guidance on how to perform the parcellation, nor mention how the parcellation was done for their real data analysis.

BHJ argue that spatial priors for the activity indicators have better edge-preservation properties, compared to priors on the activation effects. There are however some possible objections to this argument. Firstly, some spatial priors for the activity effects are in fact able to preserve edges, for example the geodesic graph Laplacian used in Harrison et al. (2008) and the scale mixture of normals used in Rad et al. (2017). Secondly, the activated areas from some tasks may actually have smooth boundaries, which may be cropped by a binary activation indicator; pre-smoothing the data tends to produce smooth boundaries, as shown in the previous section. Finally, a prior that allows for sharp continuities will be more prone to overfit the noisy fMRI data.

Nevertheless, the edge preserving property of spatial variable selection priors is attractive. However, when the boundaries of the active regions and parcel boundaries do not coincide in the BHJ approach, many voxels will be misclassified, and active and inactive voxels smoothed. That is, edges may not be preserved with the parcel-based BHJ prior, unless we already have knowledge about where the active regions are located, in which case there is no need for an analysis. This issue is demonstrated by BHJ in Section 4, but the highly stylized single slice setup used for illustration does not shed much light on how serious this issue will be in actual whole brain applications. An open problem for parcel-based approaches is how to infer parcels and activity jointly in a computationally tractable way; see e.g. (Chaari et al., 2012) for an attempt.

4 Properties of the hierarchical spatial prior

Interpretation and effect of the spatial hyperparameters

The interpretation of the two key prior hyperparameters, δ and r, could have been more clearly discussed in the article. The standard deviation of the random field, δ, controls the *within parcel* dependence of voxels. This is illustrated by simulating from the BHJ prior in the first column of Figure 3, where the large $\delta = 10$ makes $\Pr(\gamma_v = 1)$ vary wildly from parcel to parcel, giving high within parcel dependence, but low *between parcel* dependence (since $r = 0.05$ is low). Increasing the length scale parameter r increases the between parcel dependence; see the second column of Figure 3. The rightmost column of Figure 3 shows that for small δ, the length scale r becomes ineffective since then $\Pr(\gamma_v = 1) \approx 1/2$ for all v and the γ_v are close to spatially independent.

Marginal prior on activation indicators

Smith and Fahrmeir (2007) emphasize the importance of using a so called external field in their Ising prior for fMRI. The argument is that without this external field, then marginally $\Pr(\gamma_{v,j} = 1) = 1/2$ for all voxels and tasks, which implies a highly implausible large number of active voxels for a typical fMRI study. Moreover, Smith and

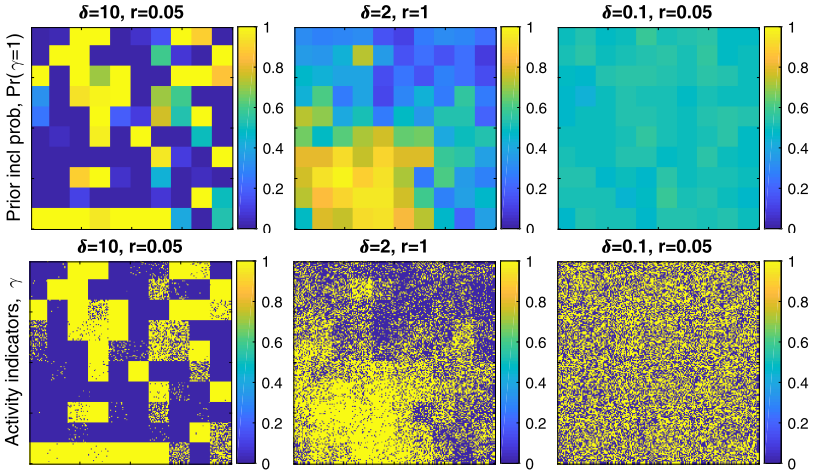


Figure 3: Realizations from the BHJ hierarchical spatial prior for three different sets of hyperparameter values. Both the realization of $\Pr(\gamma_v = 1)$ (first row) and γ_v (second row) are shown. The simulations are over a 200×200 grid of voxels divided into a 10×10 grid of parcels.

Fahrmeir (2007) use anatomical information on the location of gray matter (activations can not occur outside of gray matter) in the external field and argue that the ability to do so is an important advantage of their approach. The prior in (4) and (5) also imply the implausible marginal $\Pr(\gamma_{v,j} = 1) = 1/2$. It seems straightforward however to replace $S_{g,j}$ in the Bernoulli probability in (4) with $\alpha_g I_v + S_{g,j}$, where I_v is a binary indicator for gray matter. The α_g can be specified as in Smith and Fahrmeir (2007) from prior information on the expected proportion of active gray matter voxels in parcel g, or globally if $\alpha_g = \alpha$ for all g. This addition should add only marginally to the computational cost of the single site Metropolis-Hastings updates of the $S_{g,j}$.

Data-based prior and activation smoothing

BHJ follow Smith and Fahrmeir (2007) and use a data-based prior for the effects in active voxels, β_v, centered on the maximum likelihood estimates. The consequences of this violation of Bayes' theorem is kept to a minimum by using a vague unit information prior, and one would think that the mean could equally well have been set to zero. However, the law of total covariance gives

$$\mathrm{Cov}(\beta_u, \beta_v) = \mathrm{E}[\mathrm{Cov}(\beta_u, \beta_v|\gamma_u, \gamma_v)] + \mathrm{Cov}[\mathrm{E}(\beta_u|\gamma_u), \mathrm{E}(\beta_u|\gamma_u)] = \mu_u \mu_v \mathrm{Cov}(\gamma_u, \gamma_v),$$

since $\mathrm{Cov}(\beta_u, \beta_v|\gamma_u, \gamma_v) = 0$ by assumption, and $\mathrm{E}(\beta_u|\gamma_u) = (1 - \gamma_u)0 + \gamma_u \mu_i = \gamma_u \mu_u$, where μ_u is the prior mean of β_u when $\gamma_u = 1$. Hence, only with a non-zero prior mean on β_v would the spatial dependence in the γ_v induce spatial covariance between the β_v in different voxels. This is clearly in stark contrast to a spatial prior on β_v where the spatial covariance is completely separate from the prior mean.

References

Bezener, M., Hughes, J., Jones, G., et al. (2018). "Bayesian Spatiotemporal Modeling Using Hierarchical Spatial Priors, with Applications to Functional Magnetic Resonance Imaging." *Bayesian Analysis*. 1291

Bowman, F. D., Caffo, B., Bassett, S. S., and Kilts, C. (2008). "A Bayesian hierarchical framework for spatial modeling of fMRI data." *NeuroImage*, 39(1): 146–156. 1291

Chaari, L., Forbes, F., Vincent, T., and Ciuciu, P. (2012). "Hemodynamic-informed parcellation of fMRI data in a joint detection estimation framework." In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 180–188. Springer. 1294

Groves, A. R., Chappell, M. A., and Woolrich, M. W. (2009). "Combined spatial and non-spatial prior for inference on MRI time-series." *NeuroImage*, 45(3): 795–809. 1291

Harrison, L. M. and Green, G. G. (2010). "A Bayesian spatiotemporal model for very large data sets." *NeuroImage*, 50(3): 1126–1141. 1291

Harrison, L. M., Penny, W., Daunizeau, J., and Friston, K. J. (2008). "Diffusion-based spatial priors for functional magnetic resonance images." *Neuroimage*, 41(2): 408–423. 1291, 1294

Kalus, S., Sämann, P. G., and Fahrmeir, L. (2014). "Classification of brain activation via spatial Bayesian variable selection in fMRI regression." *Advances in Data Analysis and Classification*, 8(1): 63–83. MR3168680. doi: https://doi.org/10.1007/s11634-013-0142-6. 1291

Lee, K.-J., Jones, G. L., Caffo, B. S., and Bassett, S. S. (2014). "Spatial Bayesian variable selection models on functional magnetic resonance imaging time-series data." *Bayesian Analysis*, 9(3): 699. MR3256061. doi: https://doi.org/10.1214/14-BA873. 1291

Lindgren, F., Rue, H., and Lindström, J. (2011). "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4): 423–498. MR2853727. doi: https://doi.org/10.1111/j.1467-9868.2011.00777.x. 1292

Lindquist, M. A. et al. (2008). "The statistical analysis of fMRI data." *Statistical science*, 23(4): 439–464. MR2530545. doi: https://doi.org/10.1214/09-STS282. 1292

Musgrove, D. R., Hughes, J., and Eberly, L. E. (2016). "Fast, fully Bayesian spatiotemporal inference for fMRI data." *Biostatistics*, 17(2): 291–303. MR3516001. doi: https://doi.org/10.1093/biostatistics/kxv044. 1291

Penny, W. D., Trujillo-Barreto, N. J., and Friston, K. J. (2005). "Bayesian fMRI time series analysis with spatial priors." *NeuroImage*, 24(2): 350–362. 1291, 1292

Quirós, A., Diez, R. M., and Gamerman, D. (2010). "Bayesian spatiotemporal model of fMRI data." *NeuroImage*, 49(1): 442–456. 1292

Rad, K. R., Machado, T. A., Paninski, L., et al. (2017). "Robust and scalable Bayesian analysis of spatial neural tuning function data." *The Annals of Applied Statistics*, 11(2): 598–637. MR3693539. doi: https://doi.org/10.1214/16-AOAS996. 1294

Rue, H., Martino, S., and Chopin, N. (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations." *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2): 319–392. MR2649602. doi: https://doi.org/10.1111/j.1467-9868.2008.00700.x. 1292

Sidén, P., Eklund, A., Bolin, D., and Villani, M. (2017). "Fast Bayesian whole-brain fMRI analysis with spatial 3D priors." *NeuroImage*, 146: 211–225. 1291

Smith, M. and Fahrmeir, L. (2007). "Spatial Bayesian variable selection with application to functional magnetic resonance imaging." *Journal of the American Statistical Association*, 102(478): 417–431. MR2370843. doi: https://doi.org/10.1198/016214506000001031. 1291, 1294, 1295

Smith, M., Pütz, B., Auer, D., and Fahrmeir, L. (2003). "Assessing brain activity through spatial Bayesian variable selection." *NeuroImage*, 20(2): 802–815. 1291

Vincent, T., Risser, L., and Ciuciu, P. (2010). "Spatially adaptive mixture modeling for analysis of fMRI time series." *IEEE transactions on medical imaging*, 29(4): 1059–1074. 1291

Zhang, L., Guindani, M., Versace, F., and Vannucci, M. (2014). "A spatio-temporal nonparametric Bayesian variable selection model of fMRI data for clustering correlated time courses." *NeuroImage*, 95: 162–175. 1291

Acknowledgments

This work was funded by Swedish Research Council (Vetenskapsrådet) grant no. 20135229.

Invited Discussion

Marina Vannucci* and Jeong Hwan Kook†

We would like to congratulate the authors on a very interesting article on the topic of Bayesian models for the analysis of functional magnetic resonance imaging (fMRI) data. Here, in particular, the authors consider a class of spatio-temporal models for the detection of brain regions that activate in response to a stimulus. With respect to existing literature that use binary indicator variables for classifying active voxels, the main advance of the proposed modeling construction is the use of a hierarchical spatial prior in conjunction with a parcellation of the image. This adequately models the spatial dependence in fMRI data. The authors develop a Markov chain Monte Carlo (MCMC) method for posterior inference and show performance of the approach on simulated and real data. In this discussion, we are elaborating on the authors' final remarks on having limited their attention to single-subject data and on the computational burden of the MCMC procedure. In particular, we briefly describe *NPBayes-fMRI* (Kook et al., 2018), a user-friendly `MATLAB` GUI that implements a multi-subject Bayesian spatio-temporal approach for the analysis of task-related brain activity proposed by Zhang et al. (2016). This model formulation specifically accounts for between-subject heterogeneity in neuronal activity via a spatially informed multi-subject nonparametric variable selection prior. Efficient implementation via a Variational Bayes procedure allows to scale the inference to whole-brain analysis. As suggested by Bezener and colleagues, we believe there is room for extending these methods and the software implementation to incorporate areal models.

The `NPBayes-fMRI` software is available for download at: https://github.com/rimehi/NPBayes_fMRI.

1 Bayesian spatio-temporal models for multi-subject fMRI data

For subject $i = 1, \ldots, N$, let $Y_{i\nu} = (Y_{i\nu 1}, \ldots, Y_{i\nu T})^T$ be the vector of the BOLD response data at voxel ν, with $\nu = 1, \ldots, V$. We model the data as

$$Y_{i\nu} = X_{i\nu}\beta_{i\nu} + \varepsilon_{i\nu},\ \varepsilon_{i\nu} \sim N_T(0, \Sigma_{i\nu}), \tag{1}$$

where $X_{i\nu}$ is a known $T \times p$ covariate matrix and $\beta_{i\nu} = (\beta_{i\nu 1}, \ldots, \beta_{i\nu p})^T$ is a $p \times 1$ vector of regression coefficients. Without loss of generality, we center the data and thus do not include the intercept term in the model. Let $X_{i\nu j}$ be the jth column of $X_{i\nu}$. Then $X_{i\nu j}$ is modeled as the convolution of the j-th stimulus pattern with a hemodynamic response function (HRF) (Buxton and Frank, 1997), for example a Poisson HRF or the canonical HRF adopted by Bezener and colleagues.

*Department of Statistics, Rice University, Houston, TX, USA, marina@rice.edu
†Department of Statistics, Rice University, Houston, TX, USA

The error term in (1) is modeled as auto-correlated, specifically long memory and Discrete wavelet transforms (DWT) are employed as a way to decorrelate the data. This is a common approach in the fMRI literature (Fadili and Bullmore, 2002; Meyer, 2003; Sanyal and Ferreira, 2012; Zhang et al., 2014). After applying the DWT to (1) the model in the wavelet domain can be written as

$$Y^*_{i\nu} = \sum_{j=1}^{p} X^*_{i\nu j} \circ \beta_{i\nu j} + \varepsilon^*_{i\nu},\ \varepsilon^*_{i\nu} \sim N_T(0, \Sigma^*_{i\nu}), \tag{2}$$

with $\circ$ the element-by-element (Hadamard) product, and where W is a $T \times T$ matrix corresponding to the wavelet transform, $Y^*_{i\nu} = WY_{i\nu}, X^*_{i\nu} = WX_{i\nu}$, and $\varepsilon^*_{i\nu} = W\varepsilon_{i\nu}$, and with the covariance matrix $\Sigma^*_{i\nu}$ approximately diagonal with elements $\psi_{i\nu}\sigma^2_{imn}$ indicating the variance of the nth wavelet coefficient at the mth scale. We follow the variance progression method of Wornell and Oppenheim (1992) for the wavelet coefficients,

$$\psi_{i\nu}\sigma^2_{imn} = \psi_{i\nu}(2^{\alpha_{i\nu}})^{-m}, \tag{3}$$

with $\psi_{i\nu}$ the innovation variance and $\alpha_{i\nu} \in (0, 1)$ the long memory parameter. This structure encompasses the general fractal process, which includes long memory.

Detecting voxels that activate in response to a stimulus is equivalent to identifying the non-zero regression coefficient $\beta_{i\nu j}$ in model (2). In formulas, let $\gamma_{i\nu j}$ be a binary indicator of whether a given voxel is activated or not, that is, $\gamma_{i\nu j} = 0$ if $\beta_{i\nu j} = 0$ and $\gamma_{i\nu j} = 1$ otherwise. A spiked nonparametric prior is imposed on the coefficients

$$\beta_{i\nu j}|\gamma_{i\nu j}, G_i \quad \sim \quad \gamma_{i\nu j}G_{ij} + (1 - \gamma_{i\nu j})\delta_0, \tag{4}$$

where δ_0 is a point mass at zero and G denotes a known distribution. With multiple subjects, a hierarchical Dirichlet Process (HDP) prior can be specified as the nonparametric slab, inducing clustering among voxels within a subject on one level and between subjects on the second level. This construction enables the model to borrow information from subjects exhibiting similar activation patterns in estimating parameters of interest and also capture spatial correlation among distant voxels. For single-subject analysis, the HDP reduces to a Dirichlet process (DP) prior.

In our construction we capture spatial correlation among neighboring voxels within a subject via a Markov Random Field (MRF) prior imposed on $\gamma_{i\nu j}$,

$$P(\gamma_{i\nu j}|d, e, \gamma_{ikj}) \sim \exp(\gamma_{i\nu j}(d + e \sum_{k \in N_{i\nu}} \gamma_{ikj})),$$

with $N_{i\nu}$ the set of neighboring voxels of voxel ν for subject i, and $p(\gamma_{i\nu}) = \prod_{j=1}^{p} p(\gamma_{i\nu j})$. The sparsity parameter $d \in (-\infty, \infty)$ represents the expected prior number of activated voxels, while the smoothness parameter $e > 0$ controls the probability of identifying a voxel as active based on the activation of the neighboring voxels. As noted by Bezener and colleagues, areal models could be used instead. The prior model is completed by considering a uniform prior distribution on the delay parameter, $\lambda_{i\nu j} \sim \mathrm{U}(u_1, u_2)$, an Inverse Gamma (IG) prior on the innovation variance parameter, $\psi_{i\nu} \sim \mathrm{IG}(a_0, b_0)$, and a Beta distribution on the long memory parameter, $\alpha_{i\nu} \sim \mathrm{Beta}(a_1, b_1)$.

For posterior inference, Zhang et al. (2016) use Variational Bayes (VB) algorithms which, unlike MCMC methods, do not rely on numerical integration. VB methods have been employed successfully in Bayesian models for single-subject fMRI data (Penny et al., 2003; Flandin and Penny, 2007; Harrison and Green, 2010). These methods find an optimal approximation to the posterior that minimizes the Kullback-Leibler (KL) divergence. Typically, VB approaches provide good estimates of means, although they tend to underestimate posterior variances and also to poorly estimate the correlation structure of the data. This can still be an acceptable trade-off for our inferential purposes, as we are only interested in the identification of broad areas of activations. When analytically tractable updates for some of the parameters are not available, the VB algorithm can be combined with importance sampling. See Zhang et al. (2016) for details of the algorithm.

The primary interest of the inference is in the estimation of the selection parameters, γ, and the regression coefficients, β. These can be used to obtain activation maps, by subject and by stimulus. Using the output from the VB algorithm, posterior probabilities of inclusion (PPIs) for stimulus $j, p(\gamma_{i\nu j} = 1)$, for $j = 1, \ldots P$, are approximated as weighted averages of the variational distribution values. Activation maps can then be obtained by thresholding the PPIs using a threshold value to ensure a pre-defined Bayesian false discovery rate (FDR) (Newton et al., 2004). This produces a spatial mapping of the activated brain regions, for each subject. Corresponding posterior β-maps can be calculated by estimating the β coefficients via weighted averages of the variational distribution values, on active voxels. An additional feature of the modeling approach of Zhang et al. (2016) is that the use of the nonparametric HDP prior construction (4) can be exploited to obtain a clustering of the subjects for possible discovery of differential activations. Finally, when analyzing experimental data with multiple stimuli, contrast maps can be produced to compare the effects of different treatments, by subject, by estimating probability maps of the type $p(\beta_j - \beta_{j'} > \kappa)$, with j and j' a pair of stimuli and κ a pre-defined hypothesized value.

2 `NPBayes-fMRI` software design

We now briefly describe the `NPBayes-fMRI` software. Details on parameter setting and input arguments can be found in Kook et al. (2018). `NPBayes-fMRI` comprises of two main interfaces, one for model fitting and one for the visualization of the results, organized as shown in Figure 1.

Model Fitting. For model fitting, the user loads the data and specifies the number of subjects and the type of analysis, that is, whether it should be performed on a single 2D slice or on a 3D whole-brain parcellation. Based on this, the user will be prompted to either define or load additional files. These arguments will be used later for visualization of the results. For both 2D and 3D analyses, the percent signal change normalization and the DWT are applied as part of the model fitting stage. For DWT, Daubechies minimum phase wavelets with 4 vanishing moments are used. The user can choose to run the model with a default parameter setting or to manually set the parameters.

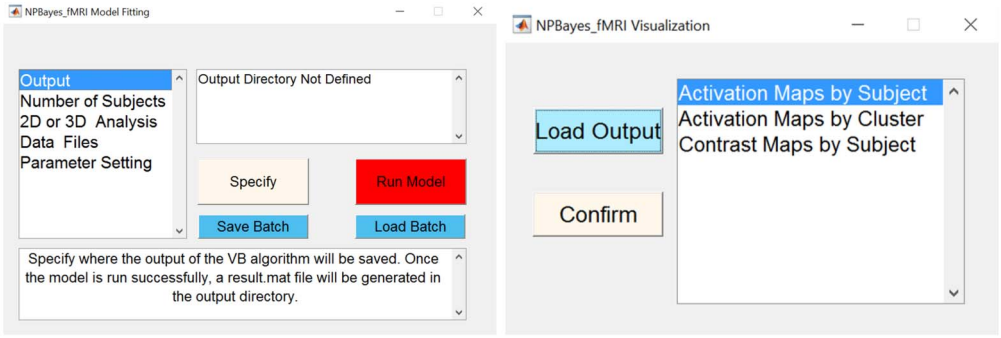


Figure 1: `NPBayes-fMRI`: Main interfaces for Model Fitting and Visualization.

Visualization. This interface is used to visualize the results. It comprises of three components, as briefly described here. For simplicity, we consider 3D data.

- **Activation Maps by Subject**: This function allows the user to view the activation maps, the posterior β-maps and the HRF maps for a single subject. Matlab's built-in colormaps can be selected via a pop-up menu. The PPI threshold and FDR value can be automatically adjusted or set manually by the user. Activation maps are overlayed on top of a reference image. Multiple slices of the brain in one particular orientation for a given stimulus can be viewed as well as single slices for all stimuli at once.

- **Activation Maps by Cluster**: This function is used to view cluster-level activation maps, for a given stimulus and PPI (or FDR) threshold. Clusters are defined based on a dendrogram obtained by applying hierarchical clustering with Ward's linkage method to a dissimilarity matrix defined based on the posterior mean estimates of the non-zero β coefficients.

- **Contrast Maps by Subject**: For multiple stimuli, this function lets the user define a contrast by subject by defining a `Contrast Vector` and `Hypothesis Value`. Once a contrast has been defined, the user can adjust a threshold to view different subjects by entering the subject numbers.

`NPBayes-fMRI` includes data of 30 subjects performing an experiment with three different stimuli. The dataset is part of a pilot study on variability in the cognitive and neural processes involved in reading, conducted at Rice University (Fischer-Baum et al., 2018). Figure 2 shows the posterior β-map for one of the subjects, for stimulus 2, obtained at a PPI threshold of 0.9. Multi-slice sagittal views can also be selected. For stimulus 2 and a PPI threshold of .9, Figure 3 shows the dendrogram, obtained by clustering the posterior β estimates of all 30 subjects, and the cluster-level β-maps when 3 clusters are selected.

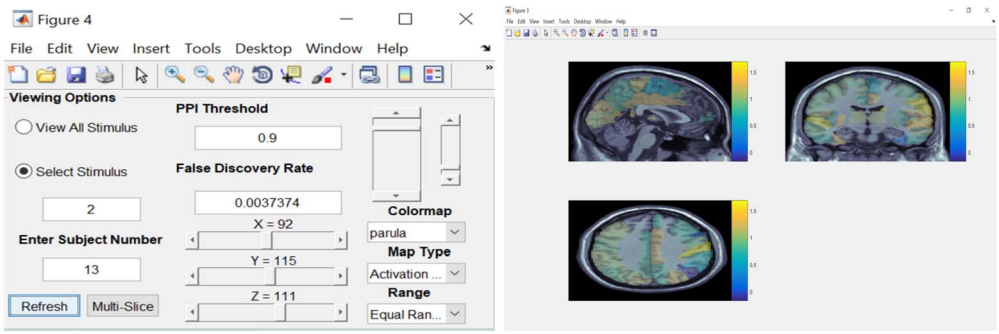


Figure 2: **3D Analysis:** Example of activation β-map, for stimulus 2 and PPI threshold of .9 coordinates $X = 92, Y = 115, Z = 111$. Different locations of the brain can be examined by using the three sliders to control the X, Y, Z coordinates.

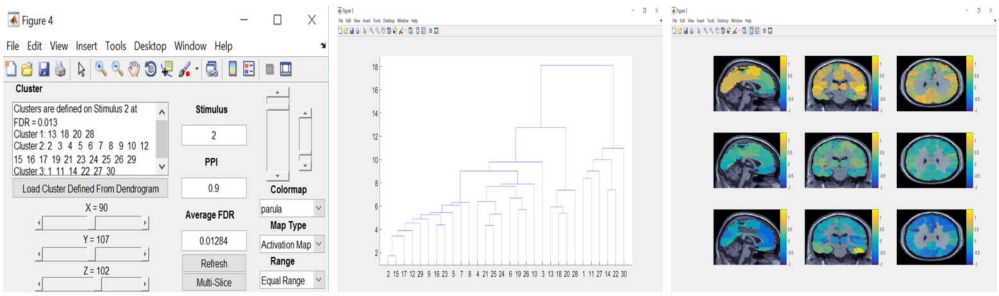


Figure 3: **3D Analysis:** Example of dendrogram (middle), for stimulus 2 and a PPI threshold of .9, and cluster-level β-maps (right), obtained with three clusters. The subject cluster memberships are displayed in the `Cluster` tab of the interface (left).

References

Buxton, R. and Frank, L. (1997). "A model for the coupling between cerebral blood flow and oxygen metabolism during neural stimulation." *Journal of Cerebral Blood Flow & Metabolism*, 17(1): 64–72. 1298

Fadili, M. and Bullmore, E. (2002). "Wavelet-generalised least squares: A new BLU estimator of linear regression models with $1/f$ errors." *NeuroImage*, 15: 217–232. 1299

Fischer-Baum, S., Kook, J., Lee, Y., Ramos-Nunez, A., and Vannucci, M. (2018). "Heterogeneity in the neural and cognitive mechanisms of single word reading." *Frontiers in Human Neuroscience*, 12(271): doi: https://doi.org/10.3389/fnhum.2018.00271. 1301

Flandin, G. and Penny, W. (2007). "Bayesian fMRI data analysis with sparse spatial basis function priors." *NeuroImage*, 34(3): 1108–1125. 1300

Harrison, L. and Green, G. (2010). "A Bayesian spatiotemporal model for very large data sets." *NeuroImage*, 50(3): 1126–1141. 1300

Kook, J., Guindani, M., Zhang, L., and Vannucci, M. (2018). "NPBayes-fMRI: Nonparametric Bayesian general linear models for single- and multi-subject fMRI data." *Statistics in Biosciences*, in press. 1298, 1300

Meyer, F. (2003). "Wavelet-based estimation of a semiparametric generalized linear model of fMRI time-series." *IEEE Transactions on Medical Imaging*, 22(3): 315–322. 1299

Newton, M., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004). "Detecting differential gene expression with a semiparametric hierarchical mixture method." *Biostatistics*, 5(2): 155–176. 1300

Penny, W., Kiebel, S., and Friston, K. (2003). "Variational Bayesian inference for fMRI time series." *NeuroImage*, 19(3): 727–741. 1300

Sanyal, N. and Ferreira, M. (2012). "Bayesian hierarchical multi-subject multiscale analysis of functional MRI data." *NeuroImage*, 63(3): 1519–1531. 1299

Wornell, G. and Oppenheim, A. (1992). "Estimation of fractal signals from noisy measurements using wavelets." *IEEE Transactions on Signal Processing*, 40(3): 611–623. 1299

Zhang, L., Guindani, M., Versace, F., Englemann, J., and Vannucci, M. (2016). "A spatiotemporal nonparametric Bayesian model of multi-subject fMRI data." *Annals of Applied Statistics*, 10(2): 638–666. MR3528355. doi: https://doi.org/10.1214/16-AOAS926. 1298, 1300

Zhang, L., Guindani, M., Versace, F., and Vannucci, M. (2014). "A spatio-temporal nonparametric Bayesian variable selection model of fMRI data for clustering correlated time courses." *NeuroImage*, 95: 162–175. 1299

Contributed Discussion

Max Hinne[*], Ronald J. Janssen[†,§], and Reza Mohammadi[‡]

Bezener, Hughes and Jones (BHJ from here on) propose an elegant Bayesian approach for spatiotemporal modeling of activated voxels in fMRI studies. We would like to discuss two avenues for extending their model that would benefit the functional Magnetic Resonance Imaging (fMRI) research community: automated, data-driven parcellation, rather than using an arbitrary division into a grid, as well as estimation of long-range connectivity, rather than modeling only local spatial correlations.

Data-driven parcellation

BHJ introduce a grid structure with a user-defined resolution to capture the spatial correlation between voxels. For voxels i, j in regions s_i, s_j, the correlation between their activation is a function of the Euclidean distance between the centers of s_i and s_j. This is a convenient approach, as the user can define the resolution of the grid to meet their computational restrictions. However, the cells of the grid have no neurological basis and likely contain many neuronal populations that are responsive to very different tasks.

This issue can be addressed by incorporating data-driven parcellation into the BHJ model (Blumensath et al., 2013). An example of this that would integrate nicely with the BHJ model is CAESAR (Clustered Activation Estimation with Spatial Adjacency Restrictions) (Janssen et al., 2016). CAESAR is a non-parametric model that clusters voxels together based on the similarities of their time courses. Temporal smoothness of the time courses is modeled via a Gaussian process. The clustering itself is inferred using a distance-dependent Chinese restaurant process prior. This distribution enforces spatial contiguity between voxels in a cluster. As a result, only those neuronal populations are grouped together that are nearby on the cortical sheet.

CAESAR was intended for resting-state fMRI data and contains no separate step to learn which voxels are responsive to a particular task; but this is precisely what BHJ offer. Their model can be augmented by replacing the fixed grid parcellation with CAESAR, to learn both the voxel activation and parcellation from the data simultaneously. Because these parcels will now have a functional meaning, the constraints they pose on voxel correlations are now much more interpretable. This also naturally leads to the next extension.

[*]University of Amsterdam, Amsterdam, The Netherlands, m.hinne@uva.nl
[†]Yale University, New Haven, CT, USA, ronald.janssen@yale.edu
[‡]University of Amsterdam, Amsterdam, The Netherlands, a.mohammadi@uva.nl
[§]Olin Neuropsychiatry Research Center, Hartford Hospital, Hartford, CT, USA

Long-range estimation of connectivity

BHJ introduce spatial correlations via random effects distributed according to a Gaussian process, where correlation is a function of spatial proximity. This captures only short-range associations, while ignoring connectivity between spatially segregated regions; a fundamental topic of study for network neuroscience (Behrens and Sporns, 2012; Bassett et al., 2018). BHJ mention that instead of a Gaussian process kernel, connectivity can be modeled using an inverse-Wishart distribution, but do not consider this further. However, given the tremendous amount of interest into precisely this long-range coupling, we think this extension to the BHJ model is very relevant to the community. Here too, existing methodology can be readily integrated with the BHJ model. For instance, in the BaCon (Bayesian Connectomics) framework (Hinne et al., 2014, 2015), hierarchical models were explored to infer long-range connectivity from fMRI data. Here the *G*-Wishart distribution is central to inference (Mohammadi and Wit, 2015), which couples correlated activity with anatomical connections. An additional parameter controls the density of the resulting network, which may be used as regularization or learned using a beta distribution prior (Janssen et al., 2014).

Similar to CAESAR, BaCon was originally designed with applications in resting-state connectivity in mind. For task-based studies, BHJ could use the BaCon approach as prior distributions on the (long- and short-range) spatial correlations. Via this approach, important long-range correlations (consider for example the homotopic inter-hemispheric connections that tend to be quite strong) will further contribute to correctly estimating active voxels.

Integrating existing models for fMRI studies

An important benefit of Bayesian modeling is that newly proposed models can often benefit from preexisting material. In the case of BHJ, the two major limitations of their study have solutions already available in the computational neuroscience community. Obviously, the practical downside of this 'lego-ing' together of Bayesian models is the computational burden this imposes. While this is a fundamental problem that has no end-all solution, we note that smart model and inference choices, such as Gaussian processes with their analytically available posterior (Hyun et al., 2016), automated inference procedures (Kucukelbir et al., 2017) and the spike-and-slab alternative to the *G*-Wishart distribution (Wang, 2015) raise our optimism about fully Bayesian data-driven fMRI analyses.

References

Bassett, D. S., Zurn, P., and Gold, J. I. (2018). "On the nature and use of models in network neuroscience." *Nature Reviews Neuroscience*, 1–13. 1305

Behrens, T. E. J. and Sporns, O. (2012). "Human connectomics." *Current opinion in neurobiology*, 22(1): 144–53. 1305

Blumensath, T., Jbabdi, S., Glasser, M. F., Van Essen, D. C., Uğurbil, K., Behrens, T. E. J., and Smith, S. M. (2013). "Spatially constrained hierarchical parcellation of the brain with resting-state fMRI." *NeuroImage*, 76: 313–324. 1304

Hinne, M., Ambrogioni, L., Janssen, R. J., Heskes, T., and van Gerven, M. A. (2014). "Structurally-informed Bayesian functional connectivity analysis." *NeuroImage*, 86: 294–305. 1305

Hinne, M., Janssen, R. J., Heskes, T., and van Gerven, M. A. (2015). "Bayesian estimation of conditional independence graphs improves functional connectivity estimates." *PLoS Computational Biology*, 11(11): e1004534. 1305

Hyun, J. W., Li, Y., Huang, C., Styner, M., Lin, W., and Zhu, H. (2016). "STGP: Spatio-temporal Gaussian process models for longitudinal neuroimaging data." *NeuroImage*, 134: 550–562. 1305

Janssen, R. J., Hinne, M., Heskes, T., and van Gerven, M. A. J. (2014). "Quantifying uncertainty in brain network measures using Bayesian connectomics." *Frontiers in Computational Neuroscience*, 8(October): 1–10. 1305

Janssen, R. J., Jylänki, P., and van Gerven, M. A. J. (2016). "Let's not waste time: Using temporal information in Clustered Activity Estimation with Spatial Adjacency Restrictions (CAESAR) for parcellating fMRI data." *PLoS ONE*, 11(12): 1–21. 1304

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017). "Automatic differentiation variational inference." *The Journal of Machine Learning Research*, 18(1): 430–474. MR3634881. 1305

Mohammadi, A. and Wit, E. C. (2015). "Bayesian structure learning in sparse Gaussian graphical models." *Bayesian Analysis*, 10(1): 109–138. MR3420899. doi: https://doi.org/10.1214/14-BA889. 1305

Wang, H. (2015). "Scaling it up: stochastic search structure learning in graphical models." *Bayesian Analysis*, 10(2): 351–377. MR3420886. doi: https://doi.org/10.1214/14-BA916. 1305

Contributed Discussion

D. Andrew Brown* and Nicole A. Lazar†

We congratulate the authors on their timely contribution toward fully Bayesian analysis of extremely high-dimensional neuroimaging data with spatiotemporal correlation. Using pre-defined parcellations to reduce the size of the spatial field over which to run Markov chain Monte Carlo (MCMC) is an innovative way to dramatically reduce the computational bottleneck associated with MCMC on this type of data.

The methodology depends crucially on the choice of regions defining the parcellation of the brain. The authors offer some suggestions on how to choose these regions, including anatomically-based regions of interest and k-means clustering. It seems that this issue could be more fully explored. For instance, if a researcher were to use brain anatomy to define the parcels, the regions would have to be segmented separately in each brain that is to be analyzed. Segmentation is a challenging issue in neuroimaging. Broadly speaking, segmentation of a brain is accomplished either through manual delineation, or automatically by, e.g., registering pre-segmented reference images to the target brain image whereby the region labels can be propagated and combined to determine the new segmentation (Iglesias and Sabuncu, 2015). Even if the regions are appropriately defined according to the task being studied in the experiment, delineating these regions in a new brain image is subject to error so that some voxels may be assigned to regions other than what is intended. We wonder how sensitive the results from the proposed model are to misclassified voxels. The simulation study clearly illustrates potentially deleterious effects of grouping together active and inactive voxels. Thus there remains the question of how robust the proposed methodology is to segmentation errors.

Related to the previous point is the need to balance "correct" partitioning of the brain with computational effort of the MCMC algorithm. Is there some way to determine an optimal partitioning subject to an upper limit on the number of regions in the parcellation? The paper illustrates the benefit of grouping together as many active voxels as possible to boost their signal for the selection procedure, but this signal is attenuated if too many inactive cases are included in the parcel. In practice, how can one balance the desired signal boosting against the risk of misaligned regions?

Other questions we have concern the prior specification in the Gaussian processes, $S_j \sim N(0, \delta_j^2 \Gamma_j)$, where $\Gamma_j \equiv \Gamma(r_j)$ depends on a correlation length parameter r_j. The authors suggest using a conventional (improper) Jeffreys prior for the variance, $\pi(\delta_j^2) = \delta_j^{-2}$. However, Gelman (2006) discusses potential risks of using this prior, or a $\text{Ga}(\epsilon, \epsilon)$ approximation, at this level in the hierarchy. There it is shown that even supposedly vague priors can have a disproportionate effect on the posterior distribution. While the model considered by Gelman (2006) is quite different than what is considered

*School of Mathematical and Statistical Sciences, Clemson University, 220 Parkway Drive, Clemson, SC 29634, USA, ab7@clemson.edu

†Department of Statistics, University of Georgia, 310 Herty Drive, Athens, GA 30602, USA, nlazar@stat.uga.edu

in this work, it is not unreasonable to suspect that the same problem could be present. It seems that prior specification is important here, as the authors' simulation results (e.g., Tables 1 and 2) suggest that the data contain only limited information about the parameters governing the Gaussian processes. Perhaps a subjective prior on the correlation length r_j could be elicited by reparameterizing with $\rho_j = \exp(-1/r_j)$ and taking $\rho_j \sim \text{Beta}(\alpha_1, \alpha_2)$. This parameterization, in our opinion, is more interpretable and thus easier to model. Of course, the correlation length issue could be circumvented altogether if a Gaussian Markov random field (GMRF; Rue and Held, 2005) were used in place of the Gaussian process. Indeed, Gaussian processes are appealing for modeling spatial correlation due to their use of explicit covariance functions and the formal prediction rules that follow (e.g., kriging), and there is generally no formal covariance function associated with a GMRF. Nevertheless, such a model might still be sufficient for the functional magnetic resonance imaging application in which there is no need for spatial prediction at unobserved locations. A GMRF could potentially ease the computational burden, as well, due to its typically sparse precision matrix. The ability to assess a posteriori correlation between regions would not be lost, since one could simply study the distribution of Moran's I statistic in place of the posterior of r_j.

Lastly, we remark that the authors' methodology is similar to the multi-resolution Bayesian variable selection approach proposed by Zhao et al. (2018). Their approach initially searches over a coarse partition of the brain image, and then searches through partitions of successively refined resolutions only within those larger regions that are identified as potentially interesting. Some of the techniques in that work could potentially be applicable for drawing inference in the model proposed here.

References

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis*, 1(3): 515–533. MR2221284. doi: https://doi.org/10.1214/06-BA117A. 1307

Iglesias, J. E. and Sabuncu, M. R. (2015). "Multi-atlas segmentation of biomedical images: A survey." *Medical Image Analysis*, 24(1): 205–219. 1307

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*. Boca Raton: Chapman & Hall/CRC. MR2130347. doi: https://doi.org/10.1201/9780203492024. 1308

Zhao, Y., Kang, J., and Long, Q. (2018). "Bayesian multiresolution variable selection for ultra-high dimensional neuroimaging data." *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(2): 537–550. 1308

Acknowledgments

DAB is partially supported by National Science Foundation (NSF) grants CMMI-1563435, EEC-1744476, and OIA-1826715. NAL is partially supported by NSF grant IIS-1607919.

Contributed Discussion

Francesca Gasperoni* and Alessandra Luati†

The paper introduces a Bayesian spatiotemporal model for identifying and assessing the strength of neural activations in task based functional magnetic resonance imaging (fMRI) data. A key feature of the model is the inclusion of explanatory indicator variables in a regression framework at voxel level, so that a detection problem, which is a core challenge in fMRI analysis, may be treated as a variable selection problem. In this context, the authors recognise the relevance of taking into account both spatial and temporal correlations and, by extending previous works of Smith and Fahrmeir (2007) and Lee et al. (2007), assume temporally correlated errors and consider an areal model to account for spatial dependence, instead of the Ising model, thus avoiding Bayesian computation inefficiency.

We enjoyed reading the paper and agree with the authors that it is reasonable to assume temporal dependence among repeated observations in the same voxels. In this discussion, we would like to comment on three aspects of the paper, related with the time series analysis of fMRI data: preprocessing, to remove artifacts; estimation of a drift term, accounting for low frequency movements; specification of the dynamics in the error term, i.e. the temporal correlation.

Preprocessing is an essential component in the analysis of fMRI data, to remove noise arising from many different sources such as motion artifacts, instrumental noise or scanner instabilities, see Tovar et al. (2015). If preprocessing accounts for artifacts, drift components may account for low frequency fluctuations of neuronal origin, usually addressed in resting state fMRI analysis. Detrending is typically obtained by low-order polynomial models, high-pass filters, spline functions, or wavelets, and it can be part of the preprocessing stage or not. We prefer to separately consider preprocessing and baseline drift estimation, as in Zhang and Yu (2008), where voxelwise semiparametric inference for fMRI is discussed; the drift term is estimated non parametrically and interpreted, as in the present paper, as a nuisance parameter. As far as temporal correlation is concerned, most of the models used for the time series analysis of fMRI assume a stationary Gaussian distribution for the noise term. However, there is still a considerable debate on the dynamic properties of fMRI and AR(p) or ARMA(p, q) errors have been considered, as Lee et al. (2007) widely discuss, as well as change point methods, as an alternative to stationarity, see Aston and Kirsch (2012). Lund (2006) concluded that no commonly accepted model for noise in fMRI exists and that regressors may whiten the noise as well as high pass filters.

This is actually the main point we raise here: specification of the time dynamics heavily depends on preprocessing and detrending, as low frequency components, whatever their origin, if not adequately accounted for in the preprocessing stage or in the drift term, will enter in the model in the noise term.

*MOX, Department of Mathematics, Politecnico di Milano, francesca.gasperoni@polimi.it
†Department of Statistics, University of Bologna, alessandra.luati@unibo.it

The authors acknowledge this point, as their motivation for a low order, namely first order, autoregressive model, AR(1), for the noise term, lies in the fact that low frequency or cyclical features are accounted for in the term $Z_v\eta_v$ of (1). If the component $Z_v\eta_v$ were absent, then higher order autoregressive processes would be necessary to capture low frequency fluctuations, such as AR(2), ARMA(2,1) or even ARMA(1,1). However, the authors preprocess the data by spatial smoothing (section 5.1) and assume that the drift component is preprocessed as well (section 2.1). This assumption is not fully clear to us, although we understand that the term $Z_v\eta_v$ is then considered as a negligible nuisance component and model (2) is applied in the paper, both in the simulation study and in the real data analysis.

We conclude our comment by addressing the reproducibility of the model in the case when resting state fMRI data are of interest. Spontaneous activations, as opposed to task-based activations, are getting more and more attention in the neurological literature (Cole et al., 2010) and it would be promising to investigate whether and how the proposed model is applicable to this framework.

References

Aston, J. and Kirsch, C. (2012). "Evaluating stationarity via change-point alternatives with applications to fMRI data." *Annals of Applied Statistics*, 6(4): 1906–1948. MR3058688. doi: https://doi.org/10.1214/12-AOAS565. 1309

Cole, D. M., Smith, S. M., and Beckmann, C. F. (2010). "Advances and pitfalls in the analysis and interpretation of resting-state FMRI data." *Frontiers in Systems Neuroscience*, 4: 8. 1310

Lee, K., Jones, G., Caffo, B., and Bassett, S. S. (2007). "Spatial Bayesian Variable Selection Models on Functional Magnetic Resonance Imaging Time-Series Data." *Bayesian Analysis*, 9(3): 699–732. MR3256061. doi: https://doi.org/10.1214/14-BA873. 1309

Lund, T. E. e. (2006). "Non-white noise in fMRI: Does modelling have an impact?" *Neuroimage*, 29(4): 1639–1651. 1309

Smith, M. and Fahrmeir, L. (2007). "Spatial Bayesian Variable Selection with Application to Functional Magnetic Resonance Imaging." *Journal of the American Statistical Association*, 102(507): 417–431. MR2370843. doi: https://doi.org/10.1198/016214506000001031. 1309

Tovar, D., Wang, Z., and Rajan, S. (2015). "A Rotational Cylindrical fMRI Phantom for Image Quality Control." *Plos One*, 10(12): 1–15. 1309

Zhang, C. and Yu, T. (2008). "Semiparametric Detection of Significant Activation for Brain fMRI." *Annals of Statistics*, 38(4): 1693–1725. MR2435453. doi: https://doi.org/10.1214/07-AOS519. 1309

Acknowledgments

We would like to thank Lucia Paci for kindly bringing the paper to our attention.

Rejoinder

Martin Bezener[*], John Hughes[†], and Galin Jones[‡]

We extend our sincere appreciation to the discussants for their efforts at a thought-provoking discussion. The discussants will be referred to by initials: Brown and Lazar (BL), Gasperoni and Luati (GL), Hinne, Janssen, and Mohammadi (HJP), Prado (P), Sidén and Villani (SV), and Vanucci and Kook (VK).

Our goal was developing a computationally efficient, yet scientifically reasonable Bayesian approach to task-based fMRI. This was motivated in part by our prior work in Lee et al. (2014) which was a good start, but still too computationally intensive for large fMRI problems. We think that we largely achieved our goal, but, as noted by the discussants, there is still work to be done with many possible refinements of the proposed model. Moreover, we are clearly some ways from attaining any sort of consensus Bayesian approach to task-based fMRI analysis.

The main topics raised by the discussants include the following.

Parcellation. We introduced parcellations in order to facilitate efficient computation by reducing the dimension of the spatial component of the model. Absent computational constraints it is optimal to have a parcel for each voxel. However, we saw massive computational gains from a coarser partition with little loss in inferential efficacy.

HJM consider using data-driven methods for parcellation and SV asks about how the parcellation was done for the HCP data. We endorse the idea behind the suggestion of HJM, having used clustering methods in our HCP data examples. We note that while the parcels are two-dimensional in the simulation study, they are three-dimensional in our data examples.

Throughout we used equal-sized parcels. It is unclear to us if having different sized parcels will impact the computational effort, but it does seem plausible that they could impact the posterior inference. For example, if there is some a priori information about activation and it were incorporated through a finer partition in the relevant region, then it wouldn't be surprising to get better classification.

We leave it as an open question as to how to best construct the parcels. Should it be that they are all equally-sized, or chosen by anatomical considerations, or chosen in a data-driven fashion? Should we use a mulitresolution approach as suggested by BL? As noted by P and BL this question doesn't get easier to answer if the models are extended to incorporate multiple subjects.

Temporal Correlation. We agree with GL who note that there is "no commonly accepted model for noise in fMRI". We believe the AR(p) and ARMA (p, q) structures are reasonable starting points; see also Monti (2011) for further discussion on the nature

[*]Stat-Ease, Inc., martin.bezener@gmail.com
[†]www.johnhughes.org, jphughesjr@gmail.com
[‡]School of Statistics, University of Minnesota Twin Cities, galin@umn.edu

temporal correlation in fMRI. There is nothing in our setup that prevents the use of alternative models for the temporal correlation; it simply wasn't our main focus.

Comparison to other methods. P asks about comparisons with other approaches. The work most closely associated with ours is that of Musgrove et al. (2016). The two methods were directly compared in Bezener et al. (2018) where we found that the method in Musgrove et al. (2016) can achieve slightly better computational efficiency whereas our work results in more accurate classification of voxels.

Priors. HJM, BL, and SV all comment on the choice of priors. This is another area where there are open questions that merit further investigation. In our development of the current work we tried several variations on the priors eventually chosen. Our choice of priors was largely guided by our overall goal of balancing inference and computation.

SV ask about replacing $S_{g,j}$ with $\alpha_g I_v + S_{g,j}$ in the Bernoulli in (4). This was investigated in an earlier version of our paper, but it was not impactful and space constraints forced us to remove it; but see Bezener (2015) for more. As we mention in the paper, we agree with HJM in that the use of an inverse Wishart to capture information about connectivity would be an interesting extension to our methods. This is mentioned in the conclusion of our paper as a potential direction for future research.

We do not generally share BL's pessimism about Jeffrey's priors, but their suggestion to consider a subjective prior on the correlation length is one that deserves further investigation. We'll go one step further and say that while we use a mixture of data-driven priors, improper priors, and subjective priors, we believe that the use of fully Bayesian (subjective) models have been underdeveloped in the fMRI literature.

Another possible modification of our method would be to replace the prior for ρ_v in Section 2.2 with a proper prior that incorporates anatomical information. We chose a point-mass prior to avoid expensive computations in the inversion of Λ_v. However, we later realized that Λ_v^{-1} has a closed form equation if an AR(1) process is assumed and does not require numerical inversion at each MCMC sampling step. This modification potentially will add information for little additional computation.

Edge Effects. Both P and SV bring up the issue of edge effects. SV observes that (1) there exist priors on activation amplitudes that are able to preserve edges and (2) edges themselves may be smooth in some settings. We agree with the first point. We were using the argument made in numerous other papers (cf. Smith et al., 2003; Smith and Fahrmeir, 2007) as a starting point. In a non-detection problem, the priors mentioned by SV may certainly be more appropriate. With regards to the second point, while there are settings in which activation tapers off gradually, in many applications a binary decision must be made, in which case that decision would still need to be made using posterior activation amplitudes. Regardless, this is an interesting point and we agree that the effects of the parcellation is an open issue.

P and SV consider what happens if the edges of the parcellation don't line up with the edge of an active area. We agree that this can influence posterior inference. Note that this is the case in our simulation in an attempt to display the worst-case scenario, but yet the edges are still well preserved and also any "holes" in the bottom left active area of Figure 5(e) are "filled in" that would otherwise be misclassified as inactive.

Spatial Smoothing. GL and SV bring up the issue of spatial smoothing. In particular we thank SV for the detailed demonstration included in their discussion. Spatial smoothing, while common, is somewhat controversial. However, our method does not require the use of smoothing, as the spatial random effects essentially act to smooth the field of binary indicator variables. Therefore we do not use smoothing in any of our simulations. For example, our results in Section 3.3 of our paper agree with the simulations in bottom row of Figure 2 of SV's discussion. Furthermore, none of the false positive rates in our simulations appear to be inflated, at least to the degree demonstrated in SV's simulations.

In the data example of Section 5, we used a small amount of spatial smoothing. We first tried our method without spatial smoothing and the posterior activation maps were unusually disjoint and choppy according to a neuroscientist we consulted. We offer two points to address this. First, despite trying our best to recreate fMRI noise in a simulated environment, real fMRI data is more complicated, noisy, full of artifacts, missing values etc. Second, we chose a distance based parcellation–with a more anatomically informed parcellation, the spatial random effects may have done a better job of smoothing out the posterior without having to use pre-smoothing.

References

Bezener, M. (2015). "Bayesian Spatiotemporal Modeling Using Hierarchical Spatial Priors with Applications to Functional Magnetic Resonance Imaging." Ph.D. thesis, University of Minnesota. MR3347037. 1312

Bezener, M., Eberly, L. E., Hughes, J., Jones, G. L., and Musgrove, D. R. (2018). "Bayesian spatiotemporal modeling for detecting activation via functional magnetic resonance imaging." In Härdle, W. K., Horng-Shing, H., and Shen, X. (eds.), *Handbook of Big Data Analytics*, 485–501. Springer. 1312

Lee, K.-J., Jones, G. L., Caffo, B. S., and Bassett, S. S. (2014). "Spatial Bayesian Variable Selection Models on Functional Magnetic Resonance Imaging Time-Series Data." *Bayesian Analysis*, 9: 699–732. MR3256061. doi: https://doi.org/10.1214/14-BA873. 1311

Monti, M. M. (2011). "Statistical analysis of fMRI time-series: A critical review of the GLM approach." *Frontiers in Human Neuroscience*, 5. 1311

Musgrove, D. R., Hughes, J., and Eberly, L. E. (2016). "Fast, fully Bayesian spatiotemporal inference for fMRI data." *Biostatistics*, 17: 291–303. MR3516001. doi: https://doi.org/10.1093/biostatistics/kxv044. 1312

Smith, M. and Fahrmeir, L. (2007). "Spatial Bayesian Variable Selection with Application to Functional Magnetic Resonance Imaging." *Journal of the American Statistical Association*, 102: 417–431. MR2370843. doi: https://doi.org/10.1198/016214506000001031. 1312

Smith, M., Pütz, B., Auer, D., and Fahrmeir, L. (2003). "Assessing brain activity through spatial Bayesian variable selection." *NeuroImage*, 20. 1312