Smooth Scalar-on-Image Regression via Spatial Bayesian Variable Selection

Jeff Goldsmith[1,*], Lei Huang[2], and Ciprian M. Crainiceanu[2]

[1]Department of Biostatistics, Columbia University School of Public Health

[2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

[*]*jeff.goldsmith@columbia.edu*

January 14, 2013

Abstract

We develop scalar-on-image regression models when images are registered multidimensional manifolds. We propose a fast and scalable Bayes inferential procedure to estimate the image coefficient. The central idea is the combination of an Ising prior distribution, which controls a latent binary indicator map, and an intrinsic Gaussian Markov random field, which controls the smoothness of the nonzero coefficients. The model is fit using a single-site Gibbs sampler, which allows fitting within minutes for hundreds of subjects with predictor images containing thousands of locations. The code is simple and is provided in less than one page in the Appendix. We apply this method to a neuroimaging study where cognitive outcomes are regressed on measures of white matter microstructure at every voxel of the corpus callosum for hundreds of subjects.

Keywords: Binary Markov Random Field; Gaussian Markov Random Field; Markov Chain Monte Carlo.

1 Introduction

Sustained exponential growth in computing power and the ability to store massive amounts of information continuously redefines the notions of complexity and scale in reference to modern datasets. The

increased interest in functional data analysis over recent years is largely a response to this trend. In the functional paradigm, trajectories observed over a dense grid are the basic object of investigation and are frequently used as predictors of scalar outcomes. However, studies now routinely collect multidimensional, spatially structured images together with conventional scalar outcomes which require new methods of analysis.

In this paper we consider a regression model relating a scalar response to a two- or three-dimensional predictor image. Our motivation for this work comes from a neuroimaging study relating differences in intracranial white matter microstructure to cognitive disability in multiple sclerosis (MS) patients. In particular, we are interested in the relationship between damage to the corpus callosum, the major white-matter bundle connecting the right and left hemispheres of the brain, and cognitive function in MS patients. From this study we have image predictors that are registered collections fractional anisotropy values measured on three-dimensional images containing $38 \times 72 \times 11 = 30,096$ voxels that include the corpus callosum. At each voxel, we observe a fractional anisotropy value that provides a subject-voxel specific measure of white-matter tissue viability. Our data are illustrated in Figure 1, which shows the three-dimensional image predictor observed for a single subject; similar predictors are observed for the remaining 134 subjects in our dataset. On the left, the corpus callosum is shown in red and the collection of voxels used as an image predictor is shown in a green box. On the right, the image predictor is displayed as a collection of parallel two-dimensional images, or slices, numbered from the most inferior to the most superior with voxels shaded by their fractional anisotropy values (a border shows the outline of the corpus callosum). Accompanying each predictor image is a scalar outcome that measures cognitive function. Our goal in this paper is to investigate the relationship between the scalar outcomes and the images using a single regression model.

Specifically, we pose a scalar-on-image regression model and estimate a smooth coefficient image of the same size as the predictors. From Figure 1, the challenges inherent in posing a scalar-on-image regression model are apparent: first, the number of outcomes (subjects) is dwarfed by the number of predictors (voxels); second, our predictors are observed on a complex, spatially structured multi-dimensional coordinate system; and finally, due to the size of the data sets, methods must be carefully developed with computational efficiency in mind. However, this study and the increasing number of others like it re-

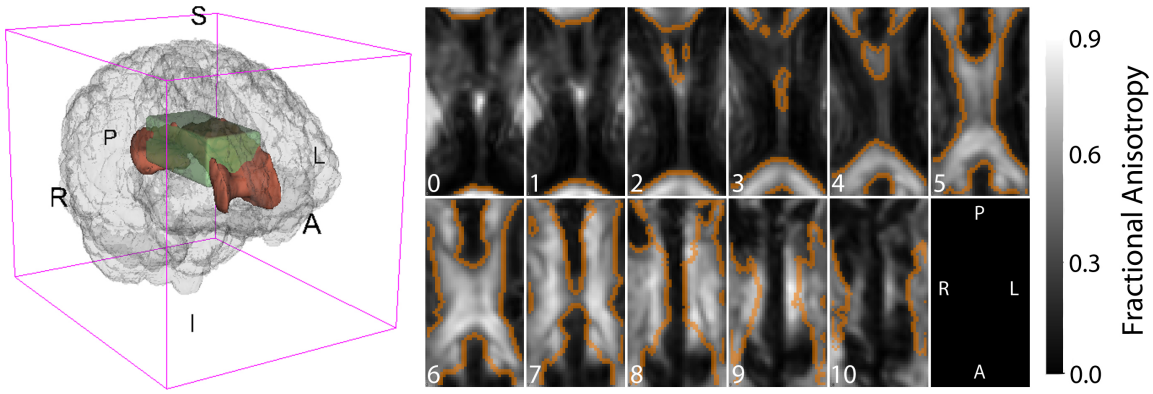

Figure 1: A single predictor image containing 30,096 voxels from the motivating neuroimaging application. On the left, the corpus callosum in shown in red and the collection of voxels used as a predictor is shown in transparent green. A single subject's scan is shown on the right with slices numbered from most inferior to most superior.

quire scalar-on-image regression models. In our proposed method to investigate the relationship between scalar outcomes and image predictors, we postulate that most image locations are not predictive of the outcome and that neighboring locations will have similar effects. Thus we use a latent binary indicator image which dichotomizes image locations as "predictive" or "non-predictive," and promote clustering of these labels. A regression coefficient is estimated at each location; in the resulting coefficient image, non-predictive locations have regression coefficients equal to zero and predictive locations have nonzero coefficients that vary smoothly in space. We implement a fast and simple single-site Gibbs sampler in which the choice between predictive and non-predictive states at a single location also determines the regression coefficient for the current iteration. Code implementing the Gibbs sampler is available in a short appendix.

Our approach to scalar-on-image regression combines prior distributions on the indicator and coefficient images to impose sparsity and smoothness. First, we use an Ising prior distribution to induce sparsity in the large number of image locations used as predictors and promote spatial contiguity in predictive regions by spatially smoothing the probabilities of the latent binary indicator image. Next, an intrinsic Gaussian Markov random field (MRF) prior distribution is used to smooth the nonzero regression coefficients. At each image location the binary choice between "predictive" and "non-predictive" is based on smoothness constraints from the prior distributions and the relative contributions of zero and nonzero regression coefficients to the outcome likelihood. Our single-site Gibbs sampler produces sam-

ples from the joint conditional distribution of the latent indicators and regression coefficients, providing insight into the variability of the estimates. Although our model was designed with the neuroimaging study in mind, the methods apply generally.

Latent binary indicator variables have been used extensively to induce sparsity in regression coefficients in the literature. Mitchell and Beauchamp (1988) used an indicator map to induce a "spike and slab" structure in the regression coefficients, and George and McCulloch (1993) introduced the use of a Gibbs sampler to sweep over the parameter space. Smith and Kohn (1996) proposed marginalizing over the regression coefficients and outcome variance to obtain a posterior density of the indicator variables that depends only on the observed data. Kohn et al. (2001) developed single-site Metropolis–Hastings samplers that decrease the computation burden in the Gibbs sampler. Smith et al. (2003) and Smith and Fahrmeir (2007) used the Ising prior distribution, a binary spatial Markov random field, to induce smoothness in the latent binary indicator variables and identify active regions in fMRI studies; Li and Zhang (2010) explored the impact of the Ising hyperparameters on the number of nonzero regression coefficients, paying particular attention to the transition between small and large models. Regression parameters are typically estimated for locations at which the posterior probability of being predictive survives a threshold or by using Bayesian model averaging. Previous work does not impose smoothness in the regression coefficients, which is implied to be an advantage in some settings (Li and Zhang, 2010). Additionally, these proposed Gibbs samplers grow in computational burden with the square of the number of predictive locations due to a step involving matrix determinants; this makes their application impractical in truly high dimensional cases, such as in our application. The Metropolis–Hastings algorithm used by Smith and Fahrmeir (2007) reduces the frequency with which this step is performed, but does not resolve the quadratic increase in computation time.

Traditionally, the use of latent indicator variables was limited to selecting few influential covariates among many (Ishwaran and Rao, 2005) and to knot selection in semiparametric regression models (Smith and Kohn, 1996; Denison et al., 1998; Kohn et al., 2001). More recently, Smith et al. (2003) and Smith and Fahrmeir (2007) consider a series of many spatially linked linear regressions arising in fMRI data. In this setting, a small linear regression model is proposed at each of many locations on a lattice; the outcomes vary from location to location, but the predictors are the same in each model. The individual

regressions are spatially linked through the inclusion or exclusion of individual predictors. Although it is in three dimensions, the fMRI setting is quite different from our application in that we consider a single regression model with a large, spatially structured three-dimensional covariate rather than a spatially organized collection of simpler linear models. Li and Zhang (2010) employ spatial variable selection in a classic genomics dataset to smooth the binary indicators in a single regression with a one-dimensional covariate, but do not impose smoothness in the nonzero regression coefficients.

The formulation of our problem is linked to the functional regression framework, which is a useful tool for conceptualizing large, spatially structured predictors with associated scalar outcomes. Standard techniques in functional data analysis include principal components decompositions for dimension reduction and penalized spline basis expansions to estimate coefficient functions (Ramsay and Silverman, 2005; Goldsmith et al., 2011a). Both tools are used by Reiss and Ogden (2010) to develop functional principal components regression (FPCR), a state of the art method for estimating the parameters in a scalar-on-image regression model with two-dimensional image predictors which could, in principle, be extended to higher dimensions. We avoid the functional approach for three reasons: i. when the true coefficient image is sparse, penalized spline expansions over-smooth regions of influence and under-smooth elsewhere; ii. in higher dimensions, spline bases become analytically and computationally difficult to use; and iii. data reduction techniques can exclude information that is important in explaining the outcome of interest. Additionally, these methods have not been extended to the three-dimensional setting and their scalability is unclear. A simulation-based comparison of the proposed Bayesian variable selection method and the functional data approach described in Reiss and Ogden (2010) is given in Section 3.

The rest of this paper is organized as follows. Section 2 details the proposed approach to scalar-on-image regression. Simulations motivated by our real-data context are presented in Section 3; the neuroimaging application is presented in Section 4. We conclude with a discussion in Section 5. Theoretical results are given in Appendix A. Code implementing our Gibbs sampler is available in Appendix B; full code for the simulations is part of a web supplement. Additional simulation results are provided in Appendix C.

2 Methods

2.1 Scalar-on-Image Regression Model

Assume that for each subject $1 \leq i \leq I$, we observe data of the form $\{y_i, \boldsymbol{w}_i, \boldsymbol{X}_i\}$ where y_i is a scalar outcome, $\boldsymbol{w}_i$ is a vector of scalar covariates and $\boldsymbol{X}_i$ is an image predictor measured over a lattice (a finite, contiguous collection of vertices in a cartesian coordinate system). We formulate the scalar-on-image regression model

$$y_i = \boldsymbol{w}_i^T \boldsymbol{\alpha} + \boldsymbol{X}_i \cdot \boldsymbol{\beta} + \epsilon_i \tag{1}$$

where $\boldsymbol{\alpha}$ is a fixed effects vector, $\boldsymbol{\beta}$ is a collection of regression coefficients defined on the same lattice as the image predictors, $\boldsymbol{X}_i \cdot \boldsymbol{\beta}$ denotes the dot product of $\boldsymbol{X}_i$ and $\boldsymbol{\beta}$, and $\epsilon_i \overset{iid}{\sim} \mathrm{N}\left[0, \sigma_\epsilon^2\right]$. Our goal is to estimate the coefficient image $\boldsymbol{\beta}$ assuming that: $i.$ the signal in $\boldsymbol{\beta}$ is sparse and organized into spatially contiguous regions; and $ii.$ the signal is smooth in non-zero regions. To this end, we also introduce a latent binary indicator image $\boldsymbol{\gamma}$ that designates image locations as either predictive or non-predictive.

Notationally, let β_l and γ_l be the l^{th} image location (pixel or voxel) of the images $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively, and $\boldsymbol{\beta}_{-l}$ and $\boldsymbol{\gamma}_{-l}$ be the images $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ with the l^{th} location removed. Also let δ_l be the neighborhood consisting of all image locations sharing a face (but not a corner) with location l; on a regular lattice in two dimensions, δ_l will contain up to four elements. Let $X_{\cdot l}$ be the length I vector of image values at location l across subjects: $X_{\cdot l}^T = [X_{1,l}, \ldots, X_{I,l}]$. Similarly, let $\boldsymbol{X}_{\cdot(-l)}$ be the collection of images with the l^{th} location removed. We assume that images have been de-meaned, so that the average of each $X_{\cdot l}^T$ is zero. Let $\boldsymbol{X}_{\cdot} \cdot \boldsymbol{\beta}$ be the length I vector consisting of the dot product of each image predictor $\boldsymbol{X}_i$ and with $\boldsymbol{\beta}$: $(\boldsymbol{X}_{\cdot} \cdot \boldsymbol{\beta})^T = [\boldsymbol{X}_1 \cdot \boldsymbol{\beta}, \ldots, \boldsymbol{X}_I \cdot \boldsymbol{\beta}]$. Finally, we define $\boldsymbol{w}$ to be the matrix with rows equal to $\boldsymbol{w}_i^T$

2.2 Parameter Estimation using Single-Site Gibbs Sampler

A combination of Ising and Gaussian MRF priors are used to induce sparsity, spatial clustering and smoothness in the estimate of $\boldsymbol{\beta}$. These priors allow location-specific conditional distributions used in the construction of a single-site Gibbs sampler.

First, we define the latent binary indicator image $\boldsymbol{\gamma}$ so that $\beta_l = 0$ if $\gamma_l = 0$ and $\beta_l \neq 0$ if $\gamma_l = 1$; defined

in this way, $\boldsymbol{\gamma}$ separates the regression coefficients $\boldsymbol{\beta}$ into predictive (nonzero) elements and non-predictive (zero) elements. An Ising prior is used for $\boldsymbol{\gamma}$, so that

$$p(\boldsymbol{\gamma}) = \phi(\boldsymbol{a}, \boldsymbol{b}) \exp\left[\boldsymbol{a} \cdot \boldsymbol{\gamma} + \sum_{l} \left\{ \sum_{l' \in \delta_l} \boldsymbol{b}_l I\left(\gamma_l = \gamma_{l'}\right) \right\}\right] \tag{2}$$

where $\phi(\boldsymbol{a}, \boldsymbol{b})$ is a normalizing constant. The parameters of the Ising distribution $\boldsymbol{a}$ and $\boldsymbol{b}$ control the overall sparsity and interaction between neighboring points, respectively. We fix $\boldsymbol{a}$ and $\boldsymbol{b}$ to be constants (a, b) over all image locations chosen via the cross-validation procedure described in Section 2.5.

Next, we use a Gaussian MRF prior for the predictive regression coefficients. Let

$$\left[\beta_l \mid \gamma_l = 1, \boldsymbol{\beta}_{-l}, \boldsymbol{\gamma}_{-l}\right] \sim \mathrm{N}\left[\overline{\beta}_{\delta_l}, \sigma^2_{\boldsymbol{\beta}}/d_l\right] \tag{3}$$

where $\overline{\beta}_{\delta_l} = \frac{\sum_{l' \in \delta_l} \beta_{l'} \gamma_{l'}}{d_l}$ is the average taken over the neighboring regression coefficients and d_l is the number of elements in δ_l. This specification leads to the posterior conditional distribution

$$\begin{aligned} \left[\beta_l \mid \boldsymbol{y}, \gamma_l = 1, \boldsymbol{\beta}_{-l}, \boldsymbol{\alpha}\right] &\propto \left[\boldsymbol{y} \mid \boldsymbol{\beta}, \gamma_l = 1, \boldsymbol{\alpha}\right] \left[\beta_l \mid \gamma_l = 1, \boldsymbol{\beta}_{-l}\right] \\ &\sim \mathrm{N}\left[\mu_l, \sigma^2_l\right] \end{aligned} \tag{4}$$

where

$$\begin{aligned} \sigma^2_l &= \left(\frac{1}{\sigma^2_\epsilon} X^T_{\cdot l} X_{\cdot l} + \frac{d_l}{\sigma^2_{\boldsymbol{\beta}}}\right)^{-1} \\ \mu_l &= \sigma^2_l \left\{ \frac{1}{\sigma^2_\epsilon} \left(\boldsymbol{y} - \boldsymbol{w}.\boldsymbol{\alpha} - \boldsymbol{X}_{\cdot(-l)} \cdot \boldsymbol{\beta}_{-l}\right)^T X_{\cdot l} + \frac{d_l}{\sigma^2_{\boldsymbol{\beta}}} \overline{\beta}_{\delta_l} \right\} \end{aligned} \tag{5}$$

are the location-specific posterior mean and variance. Selection of the tuning parameters σ^2_ϵ and $\sigma^2_{\boldsymbol{\beta}}$ is considered in Section 2.5. Using the Ising and Gaussian MRF priors, we have the site-specific joint posterior distribution of (γ_l, β_l) given by $p\left\{(\gamma_l = 1, \beta_l = \beta^*) \mid \boldsymbol{y}, \boldsymbol{\beta}_{-l}, \boldsymbol{\gamma}_{-l}\right\} = \frac{1}{1+g_l}$ with

$$\begin{aligned} g_l &= \frac{p\left\{(\gamma_l = 0, \beta_l = 0) \mid \boldsymbol{y}, \boldsymbol{\gamma}_{-l}, \boldsymbol{\beta}_{-l}\right\}}{p\left\{(\gamma_l = 1, \beta_l = \beta^*) \mid \boldsymbol{y}, \boldsymbol{\gamma}_{-l}, \boldsymbol{\beta}_{-l}\right\}} \\ &= \frac{p(\boldsymbol{y} \mid \beta_l = 0, \boldsymbol{\beta}_{-l}) \cdot p(\beta_l = 0 \mid \gamma_l = 0) \cdot p(\gamma_l = 0 \mid \boldsymbol{\gamma}_{-l})}{p(\boldsymbol{y} \mid \beta_l = \beta^*, \boldsymbol{\beta}_{-l}) \cdot p(\beta_l = \beta^* \mid \gamma_l = 1) \cdot p(\gamma_l = 1 \mid \boldsymbol{\gamma}_{-l})}. \end{aligned} \tag{6}$$

Thus, at each image location the joint posterior distribution of the latent binary indicator and regression coefficient is a Bernoulli choice that accounts for prior information through the Ising and MRF distributions as well as the relative impact of a zero and nonzero regression coefficient on the outcome likelihood.

Our full model is given by

$$\begin{aligned} y_i &\sim \mathrm{N}\left[\boldsymbol{w}_i^T\boldsymbol{\alpha} + \boldsymbol{X}_i \cdot \boldsymbol{\beta}, \sigma_\epsilon^2\right] \\ \beta_l &\sim \begin{cases} \delta(0), & \text{if } \gamma_l = 0 \\ \mathrm{N}\left[\overline{\beta}_{\delta_l}, \sigma^2_{\boldsymbol{\beta}}/d_l\right] & \text{if } \gamma_l = 1 \end{cases} \\ \gamma_l &\sim \mathrm{Ising}[a, b] \end{aligned} \tag{7}$$

where $\delta(0)$ is a point-mass at zero. The Ising prior constrains that there are relatively few nonzero regression coefficients and that they are organized into contiguous regions, while the Gaussian MRF prior ensures that coefficients vary smoothly in space. As desired, this combination allows our method to enforce sparsity and smoothness in the coefficient image. Moreover, the Bernoulli choice between zero and nonzero coefficients at each image location depends on the posterior probability (6); the calculation of this probability is straightforward and computationally efficient, as discussed in Section 2.4.

Next, we consider the placement of our method relative to others appearing in the literature. As noted in the Introduction, a functional data approach to the regression of scalars on two-dimensional images has been proposed by Reiss and Ogden (2010), but it is unclear how well this approach will scale to higher dimensions. Our methods are related to the Bayesian variable selection literature through the use of a latent binary indicator image, the Ising prior and a single-site Gibbs sampler (see Smith and Fahrmeir (2007) and Li and Zhang (2010) for recent work). These methods propose an exchangeable prior distribution on nonzero coefficients, but do not impose smoothness in the coefficient image. Further, these methods marginalize over the regression coefficient distribution to obtain a posterior probability $p(\gamma_l = 1 \mid \boldsymbol{y}, \boldsymbol{\gamma}_{-l})$ that does not depend on $\boldsymbol{\beta}$, rather than considering the joint distribution of these parameters. After obtaining marginal estimates of the binary indicator image, estimated regression coefficients are given by averaging $\mathrm{E}[\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{\gamma}]$, the posterior expected value of the coefficient image conditioned on the binary indicator after each iteration of the Gibbs sampler. Because they use only the expected value of the

regression coefficients at each iteration, these methods may ignore some of the variability in the regression coefficients. Marginalization over the regression coefficients also adds complexity to the calculation of the location-specific posterior probability of the binary indicator; see Section 2.4 for a comparison of single-site Gibbs samplers.

2.3 Theoretical Properties

We now consider the theoretical implications of the model specification in (7). First, we note that the location-specific prior distributions $p(\gamma_l \mid \boldsymbol{\gamma}_{-l})$ and $p(\beta_l \mid \gamma_l = 1, \boldsymbol{\beta}_{-l}, \boldsymbol{\gamma}_{-l})$ satisfy the conditions of the Hammersley-Clifford theorem and therefore the existence of joint prior distributions $p(\boldsymbol{\gamma})$ and $p(\boldsymbol{\beta} \mid \boldsymbol{\gamma})$ is guaranteed. Next we consider conditions under which the prior distribution $p(\boldsymbol{\beta} \mid \boldsymbol{\gamma})$ is proper.

Theorem 1. *If there exists at least one location l for which $\gamma_l = 0$, then $p(\boldsymbol{\beta} \mid \boldsymbol{\gamma})$ is proper.*

A proof of Theorem 1 is given in Appendix A. Note that if the condition of Theorem 1 is not met, then $\gamma_l = 1$ for all locations l. This implies that all image locations are predictive and that all regression coefficients are nonzero. In particular, the prior distribution $p(\boldsymbol{\beta} \mid \boldsymbol{\gamma})$ for the regression coefficients simplifies to the well-known conditional autoregressive prior (Besag, 1974; Gelfand and Vounatsou, 2003). Further, because the binary indicator image $\boldsymbol{\gamma}$ is used to induce sparsity in the regression coefficients, there are expected to be many locations l for which $\gamma_l = 0$ resulting in a proper joint prior distribution.

2.4 Single-Site Gibbs Sampling

We implement a single-site Gibbs sampler to generate iterates from the posterior distribution of $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ using the location-specific posterior probability (6). The computation time needed for each sweep over the image space is linear in the number of locations l and does not depend on the number of nonzero coefficients. Appendix B contains the full `R` implementation of the sampler for a two-dimensional coefficient image.

At each location, we make a Bernoulli choice between the (γ_l, β_l) pairs $(0, 0)$ and $(1, \beta^*)$, where β^* is sampled from the posterior distribution $\left[\beta_l \mid \boldsymbol{y}, \gamma_l = 1, \boldsymbol{\beta}_{-l}, \boldsymbol{\alpha}\right]$ in (4). Let $\boldsymbol{\beta}^0$ be the coefficient image corresponding to the first pair and $\boldsymbol{\beta}^1$ be the coefficient image corresponding to the second pair. Following

(6), the location-specific posterior distribution is $p\left((\gamma_l = 1, \beta_l = \beta^*) \mid \boldsymbol{y}, \boldsymbol{\beta}_{-l}, \boldsymbol{\gamma}_{-l}\right) = \frac{1}{1+g_l}$ where

$$
\begin{aligned}
g_l &= \frac{p(\boldsymbol{y} \mid \beta_l = 0, \boldsymbol{\beta}_{-l}) \cdot p(\beta_l = 0 \mid \gamma_l = 0) \cdot p(\gamma_l = 0 \mid \boldsymbol{\gamma}_{-l})}{p(\boldsymbol{y} \mid \beta_l = \beta^*, \boldsymbol{\beta}_{-l}) \cdot p(\beta_l = \beta^* \mid \gamma_l = 1) \cdot p(\gamma_l = 1 \mid \boldsymbol{\gamma}_{-l})} \\
&= \exp\left[-\frac{1}{2\sigma_\epsilon^2}\left\{(\boldsymbol{y} - \boldsymbol{w}.\alpha - \boldsymbol{X}_{\cdot} \cdot \boldsymbol{\beta}^0)^T(\boldsymbol{y} - \boldsymbol{w}.\alpha - \boldsymbol{X}_{\cdot} \cdot \boldsymbol{\beta}^0) - (\boldsymbol{y} - \boldsymbol{w}.\alpha - \boldsymbol{X}_{\cdot} \cdot \boldsymbol{\beta}^1)^T(\boldsymbol{y} - \boldsymbol{w}.\alpha - \boldsymbol{X}_{\cdot} \cdot \boldsymbol{\beta}^1)\right\}\right. \\
&\quad \left. + \frac{d_l}{2\sigma_{\boldsymbol{\beta}}^2}(\beta^* - \overline{\beta}_{\delta_l})^2 - a + b \sum_{l' \in \delta_l} \{I(\gamma_l = 0) - I(\gamma_l = 1)\}\right] \cdot \sqrt{2\pi \frac{\sigma_{\boldsymbol{\beta}}^2}{d_l}}.
\end{aligned}
\tag{8}
$$

The quantity above illustrates the factors used to distinguish predictive from non-predictive locations: first, the difference in residual sum of squares comparing the nonzero coefficient β^* to a zero coefficient; and second, the conditional probability for the latent binary indicator based on the Ising distribution. The computational cost involved in the calculation of the dot products $\boldsymbol{X}_{\cdot} \cdot \boldsymbol{\beta}$ appearing in (8) can be significantly reduced by noting that most of the operation does not change from location to location: we only need to compute $X_{\cdot l}^T \beta_l^*$ to determine the change in $\boldsymbol{X}_{\cdot} \cdot \boldsymbol{\beta}$ comparing the current location to the previous location. Thus, the Gibbs sampler consists of simple operations which can be quickly executed, and whose computation time does not vary based on the number of nonzero coefficients in $\boldsymbol{\beta}$. The fixed effect vector $\boldsymbol{\alpha}$ is updated after the sweep over all image locations l. A straightforward `R` implementation of the Gibbs sampler to sweep over the latent binary indicator and coefficient images is available in its entirety in a short appendix.

For comparison, Li and Zhang (2010) provides a detailed discussion of the single-site Gibbs sampling scheme used in the marginalizing variable selection methods discussed in Section 1. The most computationally expensive step in these samplers is inverting and calculating the determinant of a $p_i \times p_i$ matrix, where p_i is the number of nonzero coefficients in the model at the ith iteration. The matrix inverse is calculated using a Cholesky or other low-rank update to a matrix available at the previous iteration. In our application, the number of nonzero coefficients was between 5,000 and 8,000 after each sweep of the coefficient image, making this computation impractical. Smith and Fahrmeir (2007) use a Metropolis-Hastings step based on the prior for $\boldsymbol{\gamma}$ to reduce the number of times this step must be carried out. However, two primary issues are that $i.$ the computation time needed for the matrix calculations grows with the square of nonzero coefficients, making large models computationally prohibitive; and $ii.$ to boost computational efficiency this step is implemented in `FORTRAN`, which can hinder the use of these methods on

new datasets.

2.5 Tuning Parameters

In calculation of the location-specific posterior activation probability (8), σ_ϵ^2 determines the impact of the change in the outcome likelihood on the overall activation probability. Similarly, in the posterior distribution of active regression coefficients (4), the parameter $\sigma_{\boldsymbol{\beta}}^2$ is important in determining the posterior mean and variance. Finally, the parameters (a, b) in the Ising prior control the overall sparsity and the degree of smoothing in the activation probabilities. Together, a, b, σ_ϵ^2 and $\sigma_{\boldsymbol{\beta}}^2$ largely influence the shape and sparsity of the estimated coefficient image, and are therefore referred to as tuning parameters.

To select these parameters, we use a five-fold cross validation procedure. That is, we divide our data into five randomly selected groups and choose the collection $(a, b, \sigma_\epsilon^2, \sigma_{\boldsymbol{\beta}}^2)$ that minimizes the quantity

$$\sum_{i=1}^{5} \left\{ \sum_{k \in \text{Group}_i} (y_k - \hat{\alpha} - \boldsymbol{X}_k \cdot \hat{\boldsymbol{\beta}})^2 \right\} \tag{9}$$

where $\hat{\alpha}, \hat{\boldsymbol{\beta}}$ are estimated in each fold without using data from Group_i. This procedure increases the amount of computation time needed for our method, but provides a measure of the predictive power of the resulting coefficient image.

It is important to note that while σ_ϵ^2 is nominally the outcome variance, it acts much more as a smoothing parameter. Because the number of image locations is large, no single location contributes greatly to the dot product $\boldsymbol{X} \cdot \boldsymbol{\beta}$. In the calculation of the posterior activation probability (8), this has the effect that the Ising prior probability overwhelms the contribution of the outcome likelihood unless σ_ϵ^2 is artificially low. The choice of σ_ϵ^2 is important in controlling under- and over-fitting: a choice that is too small exaggerates the effect of each voxel and leads to over-fitting and little sparsity, while a choice that is too large understates the effect of each voxel and leads to uniformly zero regression coefficients.

Several alternative approaches for tuning parameter selection exist. A fully Bayesian approach in Smith and Fahrmeir (2007) imposes hyperprior distributions on the Ising parameters a and b, which are estimated simultaneously in the MCMC simulation. Computation time is increased substantially in this approach due to difficulty in calculating the normalizing constant for the Ising distribution; alternatively,

the parameters can be set subjectively as in Smith et al. (2003). Similarly, one could impose hyperprior distributions on the variances σ_ϵ^2 and $\sigma_{\boldsymbol{y}}^2$ and estimate these parameters via MCMC. Typically one would impose diffuse priors suitable for variance components; however due to the sensitivity of the result on the choice of σ_ϵ^2 noted above, we have found that a restrictive prior for σ_ϵ^2 is needed and that the results are sensitive to the choice of hyperparameters for this prior. Finally, we note that the Ising distribution allows parameters $\boldsymbol{a}$ and $\boldsymbol{b}$ that vary over image locations. Smith and Fahrmeir (2007) therefore construct anatomically informed parameters to induce sparsity more strongly in some regions. Absent scientific justification for an anatomically informed prior distribution, we use constant Ising parameters a and b.

We end this subsection with guidance for finding a suitable range of parameter values for the cross validation procedure. Typically we begin by selecting σ_ϵ^2 due to its importance for over- and under-fitting. The choice for this parameter depends on the scale of predictors and the signal strength, but because unsuitable values lead to extreme behavior in the coefficient image a reasonable value can be found. A useful starting range for the Ising parameter a is $(-4, 0)$, where -4 strongly enforces sparsity and 0 does not. For the Ising parameter b, a starting range is $(0, 2)$ where 2 enforces agreement between neighboring locations and 0 does not. The value for $\sigma_{\boldsymbol{\beta}}^2$ again depends on the scale of the predictors but is often relatively low to induce spatial smoothness in regression coefficients.

3 Simulations

We demonstrate the performance of our method using both two- and three-dimensional predictors using simulations based on our neuroimaging application. In addition to the simulation results presented here, Appendix C contains results for scenarios that demonstrate the effect of registration errors, larger predictive regions, homogeneous regression coefficients, and predictive regions on the boundary of the image.

To generate the predictors in the two-dimensional simulations, we first extract a 50×50 axial slice image $\boldsymbol{X}_i$ from each of the registered patient scans in our dataset. After vectorizing these images, we construct a collection of orthonormal principal components (PCs) $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_{50}\}$ with accompanying eigenvalues $\boldsymbol{\lambda} = \{\lambda_1, ..., \lambda_{50}\}$. For the simulated datasets, subject-specific PC loadings $\boldsymbol{c}_i$ are generated from a Normal distribution: $\boldsymbol{c}_i \sim \mathrm{N}\left[0, \mathrm{diag}(\boldsymbol{\lambda})\right]$. These loadings are used to create simulated 50×50 two-

dimensional predictors using $\boldsymbol{X}_i^S = \sum_{k=1}^{50} c_{ik}\phi_k$, $1 \leq i \leq I$, where $\boldsymbol{X}_i^S$ denotes the i^{th} simulated predictor by transforming the vectors back into 50×50 images. Figure 2 provides an illustration of the method used to construct simulated predictor images. An analogous procedure is used to construct three-dimensional predictors based on $20 \times 20 \times 20$ regions extracted from the frontal cortex. From this procedure, we obtain predictor images with 2500 and 8000 image locations for the two- and three-dimensional simulations, respectively. Thus our simulated datasets retain many of the features of our application, including within-subject spatial correlation and voxel-level variability.

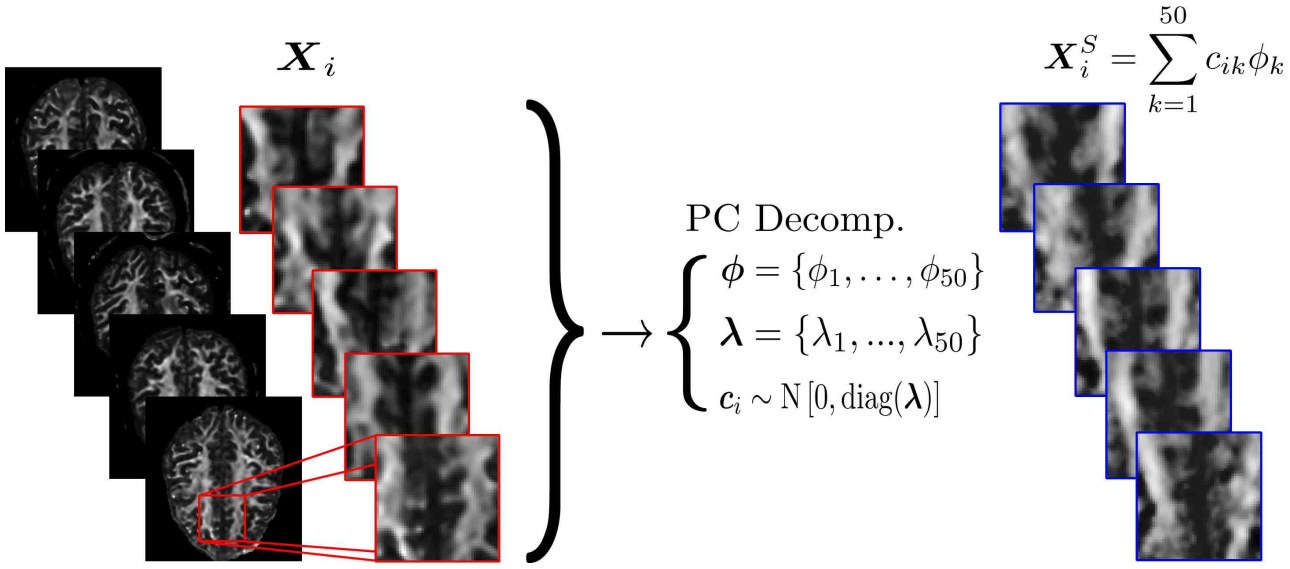

Figure 2: Method used to generated two-dimensional predictors used in the simulation study. First, 50×50 images are extracted from full scans in our neuroimaging application. These images are decomposed into principal components, and simulated predictors are generated by combining parametrically sampling PC scores with the obtained PC images.

For the two-dimensional setting, we construct a coefficient image on the unit square using the densities of bivariate Normal distributions. Let

$$U_1 \sim \mathrm{N}\left[\begin{bmatrix} .2 \\ .3 \end{bmatrix}, \begin{bmatrix} .0025 & 0 \\ 0 & .0015 \end{bmatrix}\right] \text{ and } U_2 \sim \mathrm{N}\left[\begin{bmatrix} .4 \\ .8 \end{bmatrix}, \begin{bmatrix} .002 & -.001 \\ -.001 & .001 \end{bmatrix}\right]$$

with densities f_{U1} and f_{U2} respectively. The coefficient image $\boldsymbol{\beta}$ is given by $\boldsymbol{\beta} = .08 f_{U1} - .05 f_{U2}$ realized on a 50×50 discretized grid to match the dimension of the predictors. In three dimensions, we use the density of a Normal distribution on the unit cube, centered at $(.25, .35, .65)$ and with covariance matrix

diag$(.01, .005, .01)$. Both coefficient images are show in Figure 3. Simulated outcomes y_i^S are given by $y_i^S = \alpha + \boldsymbol{X}_i^S \cdot \boldsymbol{\beta} + \epsilon_i$ where $\alpha = -10$ and $\epsilon_i \sim \text{N}\left[0, \sigma_\epsilon^2\right]$. We consider three levels for the variance σ_ϵ^2: letting $\sigma_{\boldsymbol{y}}^2 = \text{Var}\left(\boldsymbol{X}^S \cdot \boldsymbol{\beta}\right)$ be the sample variance of the simulated outcomes, we choose outcome variances $\sigma_\epsilon^2 = \frac{1}{3}\sigma_{\boldsymbol{y}}^2$, $\sigma_\epsilon^2 = \sigma_{\boldsymbol{y}}^2$, and $\sigma_\epsilon^2 = 3\sigma_{\boldsymbol{y}}^2$, giving three signal-to-noise ratios $\frac{\sigma_{\boldsymbol{y}}^2}{\sigma_\epsilon^2}$.

For each signal-to-noise ratio, we generate 500 datasets in the manner described above for both $I = 100$ and $I = 500$. In the first dataset only we use five-fold cross validation to select the tuning parameters $a, b, \sigma_\epsilon^2, \sigma_{\boldsymbol{\beta}}^2$. Using these values, we fit model (1) on each simulated dataset and obtain estimated coefficient images $\hat{\boldsymbol{\beta}}$ that are the posterior mean of the sampled coefficient images. Because the five-fold cross validation procedure is used only on the first dataset the results may not be fully representative of the proposed method. For each fit, the coefficient image $\boldsymbol{\beta}$ is initialized to zero. We use 250 iterations of the Gibbs sampler and discard the first 100 as burn-in. To provide a comparison for our method, we fit the scalar-on-image model using the functional approach (FPCR) of Reiss and Ogden (2010) discussed in the Introduction and using a Bayesian variable selection approach that imposes an exchangeable prior distribution for the regression coefficients $[\beta_l \mid \gamma_l = 1] \overset{iid}{\sim} \text{N}\left[0, \sigma_{\boldsymbol{\beta}}^2\right]$ (note that the functional approach is currently implemented only for two-dimensional predictor images). As with the proposed approach, all tuning parameters for these methods are estimated using the first dataset only.

To evaluate the estimated coefficient images, we use the mean squared error (MSE) separated by signal in the true coefficient image – regions in which $|\boldsymbol{\beta}| < .05$ are called "nonpredictive", and remaining regions are "predictive" – to provide insight into the method's ability to accurately detect features while inducing sparsity elsewhere. Thus we define $\text{MSE}_1 = \frac{1}{L_1}\sum_{l \in \text{predictive}}(\hat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l)^2$ and $\text{MSE}_0 = \frac{1}{L_0}\sum_{l \in \text{non-predictive}}(\hat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}_l)^2$ with L_1, L_0 as the number of predictive and non-predictive image locations.

Table 1 displays the average MSE taken over all simulated datasets, for each signal-to-noise ratio, sample size, and method (Gaussian MRF prior, exchangeable prior, and FPCR). For perspective on Table 1, in Figure 3 we display estimated coefficient images with typical MSE_1 and MSE_0. Table 1 demonstrates that our method provides good estimates of the coefficient image in predictive and non-predictive regions. Additionally, the use of a Gaussian MRF prior to impose smoothness in regression coefficients provides substantially smaller MSE in predictive regions, typically with the same or smaller MSE in non-predictive

	$I = 100$			$I = 500$		
$\sigma_y^2/\sigma_\epsilon^2 =$	3	1	1/3	3	1	1/3
GMRF – 2D						
MSE_1	0.775	1.263	2.805	0.377	0.605	1.062
MSE_0	0.003	0.027	0.026	0.001	0.005	0.005
Computation	94.8	93.2	96.2	151.0	151.5	156.2
EX. – 2D						
MSE_1	1.461	2.293	3.698	0.910	1.224	1.800
MSE_0	0.015	0.015	0.018	0.002	0.005	0.007
Computation	81.7	78.2	78.6	136.1	136.5	136.7
FPCR – 2D						
MSE_1	1.835	2.482	3.408	1.261	1.475	1.953
MSE_0	0.082	0.079	0.063	0.074	0.079	0.083
Computation	0.5	0.5	0.4	0.8	0.7	0.8
GMRF – 3D						
MSE_1	0.546	0.797	1.151	0.320	0.402	0.599
MSE_0	0.002	0.006	0.015	0.000	0.001	0.005
Computation	328.2	339.4	338.5	550.3	549.6	551.7
EX. – 3D						
MSE_1	0.983	1.218	1.460	0.620	0.809	1.078
MSE_0	0.006	0.010	0.020	0.003	0.005	0.010
Computation	286.4	285.3	285.2	490.9	489.7	494.1

Table 1: Average mean squared error separated by true predictive and non-predictive location, signal-to-noise ratio, sample size, predictor dimension, and estimation technique ("GMRF" labels the Gaussian MRF prior, "EX" labels the exchangeable prior, and "FPCR" the functional approach). Average computation time (in seconds) to fit the model is also shown.

regions. Visual inspection of estimated coefficient images in Figure 3 shows that the goals of sparsity and smoothness are achieved in the estimated coefficients, and confirms that typical estimates accurately recreate the true coefficient image. Results for the FPCR approach are as expected: because smoothness is induced over the full coefficient image using a penalized spline expansion, we observe over-smoothing of the true features and under-smoothing of the non-predictive regions which results in higher MSEs. An interesting consequence of our simulation design is apparent in the estimated three-dimensional coefficient: due to correlation in the simulated predictors resulting from structure in the principal components, the location of the predictive region is somewhat off-center.

To assess the ability of our method to discern between predictive and non-predictive regions, we use the posterior probability of an image location being declared predictive. Specifically, an image location l is defined to be predictive if the posterior mean of $\gamma_l > .05$. Table 2 provides the true positive and true negative rates for each of our simulation designs; from this table, we see that our method accurately

	$I = 100$			$I = 500$		
$\sigma_y^2/\sigma_\epsilon^2 =$	3	1	1/3	3	1	1/3
GMRF - 2D						
True Pos.	0.680	0.627	0.417	0.743	0.689	0.601
True Neg.	0.990	0.969	0.982	0.989	0.982	0.989
EX. - 2D						
True Pos.	0.535	0.428	0.294	0.635	0.561	0.473
True Neg.	0.955	0.968	0.974	0.975	0.970	0.973
GMRF - 3D						
True Pos.	0.589	0.499	0.380	0.672	0.653	0.584
True Neg.	0.992	0.987	0.981	0.996	0.993	0.986
EX. - 3D						
True Pos.	0.558	0.441	0.349	0.705	0.623	0.530
True Neg.	0.956	0.958	0.950	0.959	0.958	0.949

Table 2: True positive and true negative rates for the identifying predictive regions in the estimated coefficient image.

identifies non-predictive regions in the true coefficient image, and typically discerns predictive regions as well. Qualitative inspection of the estimated coefficients shows that many of the false negatives occur at the boundary of predictive regions: at these locations, the true coefficient value is typically small enough that its signal is overwhelmed by the sparsity induced by the Ising prior distribution. In three dimensions, the off-centeredness of the estimated coefficient image noted above also contributes to the relatively low true positive rate. Of course, changing the threshold used to define predictive regions will affect the values in Table 2; a lower threshold will increase the true positive rate and decrease the true negative rate. Because the FPCR approach does not induce sparsity in the coefficient image there is not a suitable definition for true positive and true negative rates, and it is omitted from Table 2. As is expected, the accuracy of our method (both in terms of MSE and in terms of true positive and negative rates) increases as the sample size gets larger and as the signal-to-noise ratio rises, but even for the smaller sample sizes and lower ratios we obtain reasonable estimates.

Finally, we include in Table 1 the average computation time needed to fit the scalar-on-image model using each of the three techniques. Both of the Bayesian variable selection methods have computation time that is substantially higher than the functional data approach; the use of the Gaussian MRF prior somewhat increases the computational burden of the sampler compared to the exchangeable prior. This penalty is primarily due to the slightly elevated frequency of nonzero coefficient estimates, which raises

the time spent writing to disk. However, we are able to obtain coefficient estimates within minutes for any simulation scenario considered. Because computation times will differ based on the system used, these values may change in practice and should be used only as a guide.

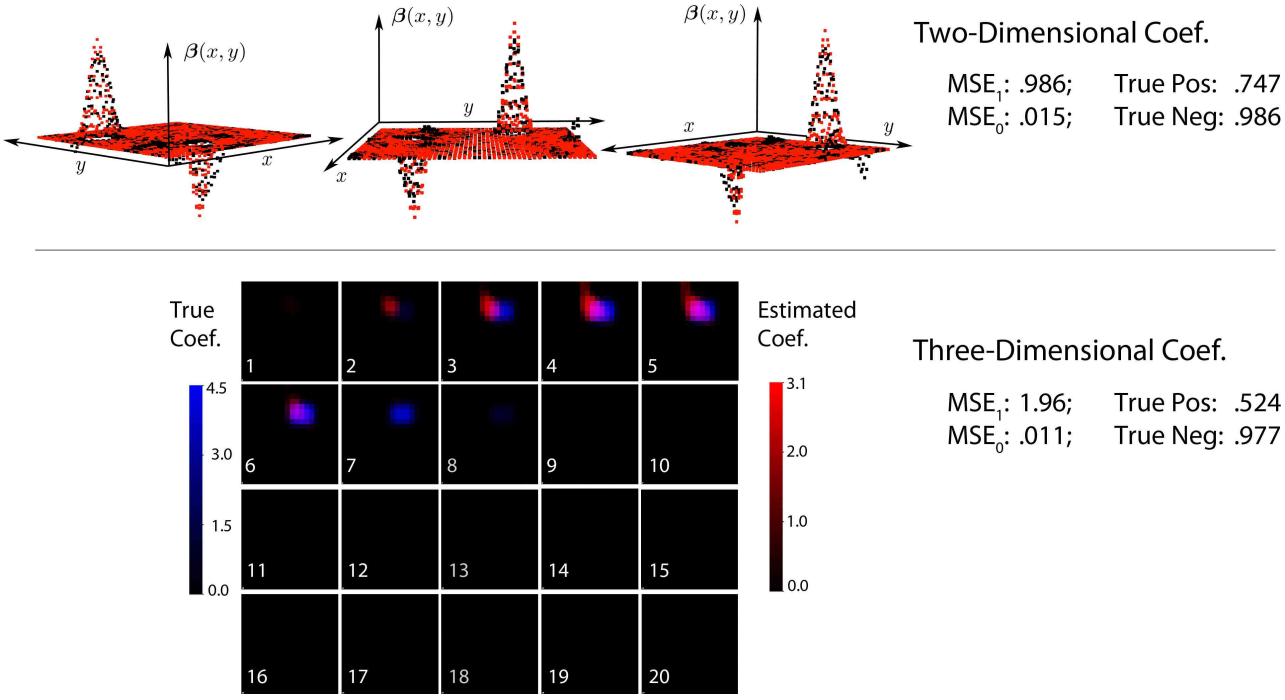

Figure 3: Plot of typical estimated coefficient images from the two- and three-dimensional simulation studies, along with the MSE_1 and MSE_0 values and true positive and negative rates associated with each estimate. The two-dimensional predictor is shown in the form $\boldsymbol{\beta}_l = \beta(x, y)$.

4 Application

Recall our application to a study relating cognitive disability in multiple sclerosis patients to diffusion tensor images. MS is an immune-mediated disease that results in damage to the myelin sheath surrounding white matter axons, which are organized into bundles or tracts. Because myelin is a protective insulation that allows the fast propagation of electrical signals, damage to this sheath disrupts neural activity and can result in severe cognitive and motor disability in affected individuals. To quantify white matter properties, we use diffusion tensor imaging, a magnetic resonance imaging technique that produces detailed images of white matter tissue by tracing the diffusion of water in the brain (Basser et al., 1994, 2000; LeBihan et al., 2001; Mori and Barker, 1999). The demyelination in MS patients occurs primarily

in localized lesions, although some degeneration is distributed throughout white matter tracts. Recent work has demonstrated that accounting for the spatial variation in tract properties improves the prediction of disability compared to using average properties taken over the full tract length (Goldsmith et al., 2011b). However, this work was based on one-dimensional functional summaries of white matter tracts rather than the true three-dimensional anatomical structures. To more completely investigate the relationship between white matter properties and patient outcomes we use a regression model that takes high-dimensional, spatially organized images as predictors.

We focus our attention on the Paced Auditory Serial Addition Test (PASAT), which provides a score between 0 and 60 with higher scores indicating better cognition, as an assessment of cognitive function. We relate this score to the corpus callosum, a major collection of white matter fibers connecting the left and right hemispheres of the brain. Damage to the corpus callosum in MS patients has previously been linked to decreased cognitive performance (Ozturk et al., 2010).

Our dataset consists of PASAT scores, non-image covariates age and sex, and diffusion tensor images of 135 MS patients; the diffusion tensor images are registered across subjects to ensure that image locations are comparable. Registration of images is important in that our regression model assumes that the coefficients are common across subjects; although no registration technique is perfect, major structures (such as the corpus callosum) can be reasonably aligned across subjects. The images provide fractional anisotropy (FA) measures at many voxels – FA is a measure of water diffusion that is indicative of white matter viability. The images consist of $38 \times 72 \times 11$ voxels containing the corpus callosum, which results in roughly 30,000 image locations that potentially predict the scalar PASAT outcome. As shown in Figure 1, we note that these images also include voxels that are not within the corpus callosum. To analyze these data, we first use a standard linear regression model with age and sex as non-image predictors. Next, we implement our proposed scalar-on-image regression using the images as predictors and estimate a three-dimensional coefficient image.

We use a five-fold cross validation procedure to select the tuning parameters a, b, σ^2_ϵ and $\sigma^2_{\boldsymbol{\beta}}$ in the scalar-on-image regressions. For all possible combinations $(a, b, \sigma^2_\epsilon, \sigma^2_{\boldsymbol{\beta}})$ of tuning parameters, each of which is examined on a grid over reasonable values, we fit the scalar-on-image regression model on the training set using chains of length 150 and discarding the first 75 as burn-in. The parameter combination

with the lowest residual sum of squares on the test set averaged over folds is chosen for use on the full dataset. For the full analysis we use chains of length 2500 and discard the first 1000 as burn-in. The latent binary and coefficient images are initialized to be zero at all locations.

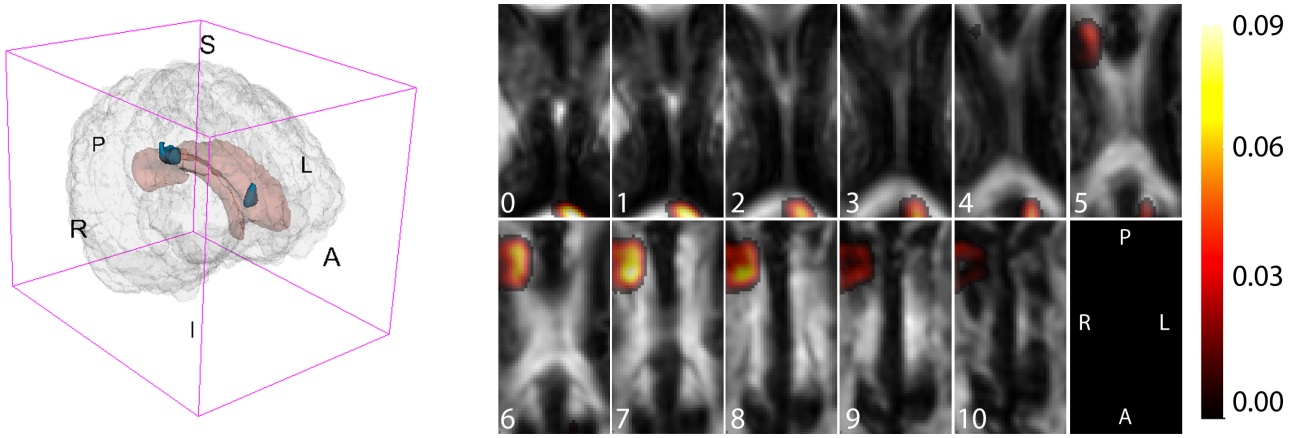

Figure 4: Estimated coefficient image from the scalar-on-image regression analysis. On the left, nonzero coefficients are shown in blue and the corpus callosum is shown in transparent red. On the right, the coefficient image is overlayed on a subject's scan for anatomical reference.

Figure 4 shows the estimated coefficient image from the scalar-on-image regression model. This Figure shows that the coefficient image contains relatively few nonzero regions, that these regions are spatially connected, and that the coefficients vary smoothly in space. From a scientific perspective, we note that the non-zero coefficients are uniformly positive, indicating that subjects with above-average FA values tend to have higher PASAT scores while those with below average FA tend to have lower PASAT scores. Thus, as is expected, degradation of white matter in the corpus callosum is associated with decreased cognitive performance measured by the PASAT score. Moreover, although the image predictors contain voxels both within and without the corpus callosum, the regions of interest in the coefficient image are largely contained within the corpus callosum, which is consistent with scientific expectations. In some areas, the coefficient image appears to extend beyond the white matter. These are possibly driven by registration errors in the images across subjects; other potential explanations are oversmoothing of large effects and the correlation in adjacent image locations.

In Figure 5 are three plots showing the MCMC chains for coefficients at three image locations: one that is predictive, one that is non-predictive, and one on the border between predictive and non-predictive regions. The combination of Ising and Gaussian MRF priors leads to distinctive mixing patterns for the

three chains. For the predictive location, mixing is driven by the Normal distribution of Gaussian MRF prior. For the non-predictive location, the coefficient estimate is zero for most iterations, driven by the sparsity induced by the Ising distribution. The border location displays a mixture of these two behaviors.

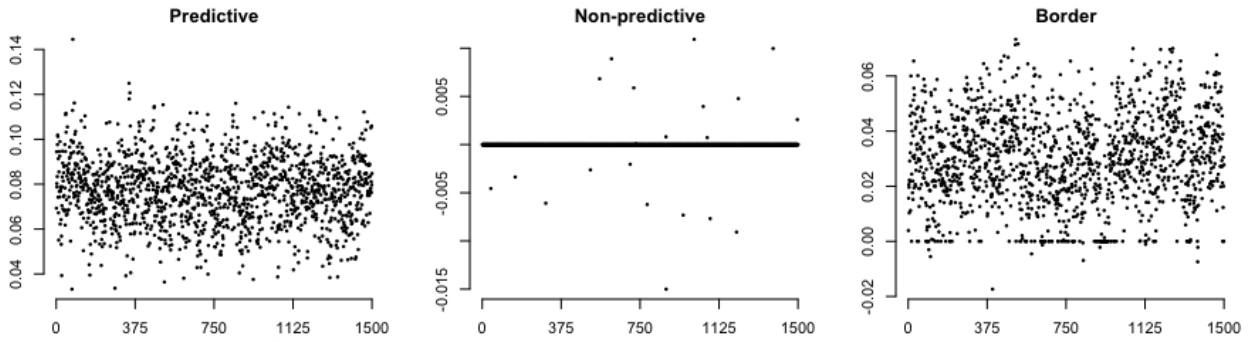

Figure 5: MCMC chains for regression coefficients at three image locations, one predictive, one non-predictive, and one on the border between predictive and non-predictive regions.

In the cross-validation procedure, the average percent of variation in the PASAT outcome explained by our model was 18.4% on the training set and 11.4% on the test set. This indicates some degree of overfitting, which is not surprising given the complexity of the model and the relatively small test set sizes. For the full analysis, the scalar-on-image regression model explains 17.3% of the outcome variance, in line with the results of the cross-validation step. Meanwhile, the standard regression model that uses age and sex as predictors explains only 1.7% of the outcome variance.

It may be surprising that the scalar-on-image regression explains only a small amount of the variability in the outcome. However, there are at least three possible reasons for this weak association. First, the PASAT score is a noisy measure of cognitive function, and is subject to many errors that cannot be captured or controlled during testing. Indeed, multiple sclerosis patients with large differences in disease burden exhibit similar PASAT scores. A more sensitive measure of cognitive function could be more strongly associated with FA values in the corpus callosum. Second, cognitive function is a complex system, and while damage to the corpus callosum certainly contributes to overall patient disability, other regions of the brain also play important roles. Thus, the corpus callosum alone may be insufficient to accurately predict cognitive ability. Finally, potential misregistration of the predictor images could reduce the interpretability of image locations across subjects. Because our model is specified at the voxel level,

such errors would decrease the predictive performance of the regression.

Finally, we implement a standard approach to regression problems in which the predictor is an image. In this voxel-wise regression, we perform a simple linear regression of the PASAT outcome on the FA value at each voxel in the predictor image in turn; the p-values for the slope on FA are shown in Figure 6. A comparison of this p-value image and the coefficient image in Figure 4 shows that the two predictive regions identified by the scalar-on-image approach generally correspond to areas with small p-values in the voxel-wise analysis. However, Figure 6 shows several regions of low p-values that do not appear as predictive in the full regression model; additional analyses find that these regions are highly correlated with those that are predictive, indicating that their low voxel-wise p-values may reflect confounding rather than true association.

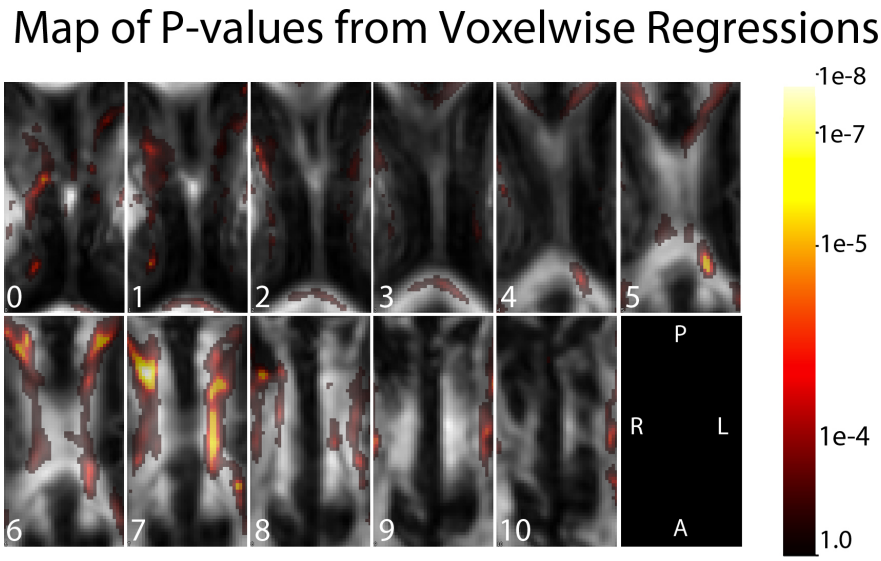

Figure 6: P-values resulting from a voxel-wise analysis of the application data, which poses a simple linear regression of the PASAT on the FA value at each voxel in turn. P-values for the FA slope are overlayed on a subject's scan for anatomical reference.

Our proposed method has several advantages over the voxel-wise approach, including: i. estimating regression coefficients for all locations simultaneously by including the entire image as a predictor; ii. jointly modeling both the predictive status of each location and the regression coefficient through the use of the latent binary and coefficient images; and iii. explicitly inducing spatial smoothness in the posterior probabilities in the binary indicator image and the regression coefficients by posing the Ising and Gaussian MRF prior distributions. Thus, while the voxel-wise regression strategy provides useful

support for the results of our application, it lacks many of the features and insights provided by the more complex approach we propose.

5 Discussion

Incorporating large, spatially structured image predictors in single regression models is a frequently encountered problem in complex data sets. This paper addresses the challenge of scalar-on-image regression in a way that produces smooth coefficient images with spatially contiguous predictive regions. Our method combines Ising and Gaussian MRF prior distributions to achieve sparsity and smoothness in the estimated coefficient. We examine the joint posterior distribution of the regression coefficient and a latent binary indicator at each image location. Importantly, we develop a simple Gibbs sampler that avoids the large matrix inversion that slows down related methods. Simulations indicate that the proposed method performs well, both in terms of classifying image locations as "predictive" and "non-predictive" and in terms of estimating the coefficient image, and demonstrate the usefulness of the Gaussian MRF prior over an exchangeable prior in reducing the mean squared error of the estimated images. Finally, we consider a neuroimaging application in which three-dimensional DTI scans of intracranial white-matter are related to cognitive disability.

The proposed method has several limitations. As noted in our application, it is possible to overfit a model at the expense of prediction on future data. This is not an uncommon issue when the number of parameters vastly exceeds the number of subjects, as is the case in our current setting. The use of multi-fold cross validation to select smoothing parameters may alleviate this problem, but overfitting will need to be evaluated on an application-by-application basis. Our simulation study, which generates predictors based on our neuroimaging application, demonstrates that the correlation in the image predictors that is a hallmark of this data may complicate the identification of regions which influence the outcome. Finally, our development assumes that the coefficient image is sparse, and it is currently unclear what ramifications the sparsity assumption will have when it is inaccurate.

Future work may take several directions. The computational burden of the cross validation may be relieved by the imposition of priors distributions on the tuning parameters. However, for σ_ϵ^2 and σ_β^2, we have found the overall model fit to be sensitive to the choice of hyperparameters, while the appropriate

prior distributions for a and b is unclear; further work is needed to address the estimation of the tuning parameters. Our development has not addressed the possibility of measurement error in the predictor, which can be common in some settings. A joint model of the predictors and outcomes could resolve this issue, but may be computationally expensive. Extending the methods to generalized regression models is clinically important – in our application, the use of brain scans to distinguish multiple sclerosis cases from healthy controls could aid in the earlier detection, diagnosis and treatment of disease. Finally, there is a need for inferential tools to determine the statistical significance of estimated coefficient images.

6 Supplementary Materials

All supplementary materials are contained in a zipped Web Appendix available on the first author's website. Supplements include: A) theoretical results described in Section 2.3; B) `R` code for the single-site Gibbs sampler; C) additional two-dimensional simulations in which the coefficient image contains a single large, homogenous predictive region, and in which the predictors are subject to registration errors; D) a three-dimensional visualization of the simulations results in Section 3; E) a three-dimensional visualization of the simulations results in Supplement B; and F) an additional three-dimensional visualization of the simulations results in Supplement B. Also included are code and generating PC data for the two-dimensional simulations in Section 3.

7 Acknowledgments

The work of Goldsmith and Crainiceanu was supported by Award Number R01NS060910 from the National Institute Of Neurological Disorders And Stroke. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute Of Neurological Disorders and Stroke of the National Institutes of Health. Partial support for Goldsmith's work was provided by Training Grant 2T32ES012871, from the U.S., NIH, National Institute of Environmental Health Sciences.

The authors are grateful to Phil Reiss and Lan Huo for providing software and assistance in implementing the FPCR method in Reiss and Ogden (2010). The authors also thank Daniel Reich and Peter

Calabresi, who were instrumental in collecting the data for this study. Scans were funded by grants from the National Multiple Sclerosis Society and EMD Serono. We are grateful to Vadim Zippunikov and John Muschelli for sharing their expertise in visualization software.

References

BASSER, P., MATTIELLO, J. and LEBIHAN, D. (1994). MR diffusion tensor spectroscopy and imaging. *Biophysical Journal*, **66** 259–267.

BASSER, P., PAJEVIC, S., PIERPAOLI, C. and DUDA, J. (2000). In vivo fiber tractography using DT-MRI data. *Magnetic Resonance in Medicine*, **44** 625–632.

BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, **36** 192–236.

BROOK, D. (1964). On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika*, **51** 481–483.

DENISON, D., MALLICK, B. and SMITH, M. (1998). Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society: Series B*, **60** 333–350.

GELFAND, A. and VOUNATSOU, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, **4** 11–25.

GEORGE, E. and MCCULLOCH, R. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88** 881–889.

GOLDSMITH, J., BOBB, J., CRAINICEANU, C. M., CAFFO, B. and REICH, D. (2011a). Penalized functional regression. *Journal of Computational and Graphical Statistics*, **20** 830–851.

GOLDSMITH, J., CRAINICEANU, C. M., CAFFO, B. and REICH, D. (2011b). Penalized functional regression analysis of white-matter tract profiles in multiple sclerosis. *NeuroImage*, **57** 431–439.

ISHWARAN, H. and RAO, J. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of Statistics*, **33** 730–773.

KOHN, R., SMITH, M. and CHAN, D. (2001). Nonparametric regression using linear combinations of basis functions. *Statistics and Computing*, **11** 313–322.

LEBIHAN, D., MANGIN, J., POUPON, C. and CLARK, C. (2001). Diffusion tensor imaging: Concepts and applications. *Journal of Magnetic Resonance Imaging*, **13** 534–546.

LI, F. and ZHANG, R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association*, **105** 1202–1214.

MITCHELL, T. and BEAUCHAMP, J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, **83** 1023–1032.

MORI, S. and BARKER, P. (1999). Diffusion magnetic resonance imaging: its principle and applications. *The Anatomical Record*, **257** 102–109.

OZTURK, A., SMITH, S., GORDON-LIPKIN, E., HARRISON, D., SHIEE, N., PHAM, D., CAFFO, B., CALABRESI, P. and REICH, D. (2010). MRI of the corpus callosum in multiple sclerosis: association with disability. *Multiple Sclerosis*, **16** 166–177.

RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*. New York: Springer.

REISS, P. and OGDEN, T. (2010). Functional generalized linear models with images as predictors. *Biometrics*, **66** 61–69.

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.

SMITH, M. and KOHN, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, **75** 317–343.

SMITH, M., PÜTZ, B., AUER, D. and FAHRMEIR, L. (2003). Assessing brain activity through spatial Bayesian variable selection. *NeuroImage*, **20** 802–815.

Smooth Scalar-on-Image Regression via Spatial Bayesian Variable Selection: Supplement A

By: Jeff Goldsmith, Lei Huang, Ciprian Crainiceanu

A Theoretical Results

Recall our previous notation that δ_l is the neighborhood consisting of all image locations sharing a face (but not a corner) with location l, and that $d_l = |\delta_l|$ where $|\cdot|$ denotes the cardinality of a set.

Theorem 1. *If there exists at least one location l for which $\gamma_l = 0$, then $p(\boldsymbol{\beta} \mid \boldsymbol{\gamma})$ is proper.*

Proof. Define sets $L = \{l : \boldsymbol{\gamma}_l = 1\}$ and $L^C = \{l : \boldsymbol{\gamma}_l = 0\}$ that partition image locations in predictive and non-predictive locations, and assume that L^C is non-empty. Let $\boldsymbol{\beta}_L$ and $\boldsymbol{\beta}_{L^C}$ be the coefficient image at the predictive and non-predictive locations, respectively; also let $\underset{\sim}{\boldsymbol{\beta}}_L$ be the regression coefficients of predictive locations expressed as a vector.

For $l \in L^C$, $[\beta_l|\gamma_l = 0] \sim \delta(0)$ where $\delta(0)$ is the Dirac delta function. For $l \in L$, $[\beta_l|\boldsymbol{\beta}_{-l}, \boldsymbol{\gamma}_{-l}] \sim \text{N}\left[\frac{\sum_{l' \in \delta_l} \beta_{l'}\gamma_{l'}}{d_l}, \sigma^2_{\boldsymbol{\beta}}/d_l\right]$. Note that $\frac{\sum_{l' \in \delta_l} \beta_{l'}\gamma_{l'}}{d_l} = \frac{\sum_{l' \in (\delta_l \cap L)} \beta_{l'}}{d_l}$ so that the distribution depends only on elements of $\boldsymbol{\beta}_L$. Defining $b_{ll'} = \frac{1}{d_l}$, we have $\frac{\sum_{l' \in (\delta_l \cap L)} \beta_{l'}}{d_l} = \sum_{l' \in (\delta_l \cap L)} b_{ll'}\beta_{l'}$. Following Brook's lemma (Brook, 1964), this prior specification for $[\beta_l|\boldsymbol{\beta}_{-l}, \boldsymbol{\gamma}_{-l}]$ results in the joint density $f(\underset{\sim}{\boldsymbol{\beta}}_L) \propto \exp\left[\frac{-1}{2}\underset{\sim}{\boldsymbol{\beta}}_L^T D^{-1}(I - B)\underset{\sim}{\boldsymbol{\beta}}_L\right]$ where D and B are square matrices of size $|L| \times |L|$ whose rows and columns correspond to the location ordering of $\underset{\sim}{\boldsymbol{\beta}}_L$. D is a diagonal matrix with entries $D_{ll} = \sigma^2_{\boldsymbol{\beta}}/d_l$ and B has elements $B_{ll'} = b_{ll'}$ defined above and is zero elsewhere.

Notice $b_{ll'}d_l = b_{l'l}d_{l'}$, so $D^{-1}(I - B)$ is a symmetric matrix. Next, we show this matrix is positive definite. In the following, the notation $l < l'$ will indicate that the location l precedes location l' in the

ordering of the vector $\underset{\sim}{\boldsymbol{\beta}}_L$. For an arbitrary nonzero vector x of length $|L|$

$$\begin{aligned} x^T D^{-1}(I-B)x &= \sum_{l\in L}\frac{d_l}{\sigma^2_{\boldsymbol{\beta}}}x_l^2 - 2\sum_{l<l'}\frac{d_l}{\sigma^2_{\boldsymbol{\beta}}}b_{ll'}x_l x_{l'} \\ &= \sum_{l\in L}\frac{d_l}{\sigma^2_{\boldsymbol{\beta}}}x_l^2 + \sum_{l<l'}\frac{d_l}{\sigma^2_{\boldsymbol{\beta}}}b_{ll'}(x_l^2 - 2x_l x_{l'} + x_l^2) - \sum_{l<l'}\frac{d_l}{\sigma^2_{\boldsymbol{\beta}}}b_{ll'}(x_l^2 + x_{l'}^2) \\ &= \sum_{l\in L}\frac{d_l}{\sigma^2_{\boldsymbol{\beta}}}x_l^2 + \sum_{l<l'}\frac{d_l}{\sigma^2_{\boldsymbol{\beta}}}b_{ll'}(x_l - x_{l'})^2 - \sum_{l\neq l'}\frac{d_l}{\sigma^2_{\boldsymbol{\beta}}}b_{ll'}x_l^2 \\ &= \sum_{l\in L}\left[\frac{d_l}{\sigma^2_{\boldsymbol{\beta}}}\left(1 - \sum_{\{l':l'\neq l\}} b_{ll'}\right)x_l^2\right] + \sum_{l<l'}\frac{d_l}{\sigma^2_{\boldsymbol{\beta}}}b_{ll'}(x_l - x_l')^2. \end{aligned} \tag{10}$$

Because $b_{ll'}$ is non-negative, we have $\sum_{l<l'}\frac{d_l}{\sigma^2_{\boldsymbol{\beta}}}b_{ll'}(x_l - x_l')^2 \geq 0$. Moreover, for every l the term $\sum_{\{l':l'\neq l\}} b_{ll'} = \frac{1}{d_l}\cdot|\{l' : l' \in \delta_l \cap L\}| \leq 1$. Because we assume L^C is non-empty, there exists at least one location l such that the preceding inequality is strict. Therefore there is at least one location l for which

$$\sum_{l\in L}\left[\frac{d_l}{\sigma^2_{\boldsymbol{\beta}}}\left(1 - \sum_{\{l':l'\neq l\}} b_{ll'}\right)x_l^2\right] > 0$$

and thus $x^T D^{-1}(I-B)x > 0$. So, the matrix $D^{-1}(I-B)$ is symmetric and positive definite, and $[\underset{\sim}{\boldsymbol{\beta}}_L]$ has a proper joint distribution. Finally, we note $f(\boldsymbol{\beta}|\boldsymbol{\gamma}) = f(\underset{\sim}{\boldsymbol{\beta}}_L)\prod_{l\in L^C}\delta(0)$, which is the product of proper densities. □

Smooth Scalar-on-Image Regression via Spatial Bayesian Variable Selection: Supplement B

By: Jeff Goldsmith, Lei Huang, Ciprian Crainiceanu

B R Code for Single-site Sampler

The following code implements the Gibbs sampler used for two-dimensional scalar-on-image regression; straightforward extensions allow for higher dimensions. In this code, the quantities `N1` and `N2` are the dimensions of the image predictors; `xbeta` is the vector of dot products $\boldsymbol{X} \cdot \boldsymbol{\beta}$ and is initialized to 0, as is $\boldsymbol{\beta}$; and `sigeps.inv` and `sigbeta.inv` are $\frac{1}{\sigma^2_\epsilon}$ and $\frac{1}{\sigma^2_{\boldsymbol{\beta}}}$, respectively. This code sweeps over all image locations, at each point making a Bernoulli choice between a zero and nonzero coefficient based on the prior information and the relative impact on the likelihood comparing a zero to a nonzero coefficient. Coefficients `alpha` for fixed effects `FixEf` are updated after the sweep over image locations.

```
for(i in N1:1){
  for(j in 1:N2){
    # define delta neighborhood
    delta = cbind(c(i-1, i+1, i,i), c(j,j, j-1, j+1))
    delta = replace(delta, which(delta<=0), NA)
    delta[,1] = replace(delta[,1], which(delta[,1]>N1), NA)
    delta[,2] = replace(delta[,2], which(delta[,2]>N2), NA)
    dl = sum(complete.cases(delta))
    Xl=X[i,j,]
    xbeta0 = xbeta - betaHat[i,j]*Xl;  txbeta0 = t(xbeta0)

    # generate nonzero coefficient
    betabar = mean(betaHat[delta], na.rm=TRUE)
    Sigl = 1/(sigeps.inv * InProd[i,j]  + dl * sigbeta.inv)
    Mul = Sigl * ( sigeps.inv * (tY - t(FixEf%*%alpha) - txbeta0)%*%Xl
               + dl * sigbeta.inv * betabar   )
    betaStar = rnorm(1, Mul, sqrt(Sigl))

    xbeta1 = xbeta0+Xl*betaStar; txbeta1=t(xbeta1)

    # compute posterior probability g_l
    IND = sum(1-2*z[delta], na.rm=TRUE)
    g = sqrt(2*pi*SigB/dl)*exp((-.5*sigeps.inv)*(txbeta0%*%xbeta0
          + 2*(txbeta1-txbeta0)%*%(Y+FixEf%*%alpha)
          - txbeta1%*%xbeta1) + .5*sigbeta.inv*dl*(betaStar-betabar)^2
          - A[i,j] + b*IND )
    ppos = 1/(1+g)

    # Bernoulli choice based on posterior probability
    if(rbinom(1, 1, ppos)==1) {z[i,j] = 1; betaHat[i,j] = betaStar}
    else {z[i,j] = betaHat[i,j] = 0}
```

```
    # update X * Beta based on current iteration
    xbeta = xbeta0 + betaHat[i,j]*Xl
  }
}
```

Smooth Scalar-on-Image Regression via Spatial Bayesian Variable Selection: Supplement C

By: Jeff Goldsmith, Lei Huang, Ciprian Crainiceanu

C Additional Simulation Results

Here we consider additional scenarios for the two-dimensional simulations described in Section 3 in the main manuscript. We examine two additional cases: first, the coefficient function contains a large, rectangular predictive region on the boundary of the image in which all regression coefficients are equal; second, we examine the impact of registration error on the estimated coefficient image.

C.1 Large, Homogeneous Predictive Region

We construct predictors as in Section 3, using a principal components decomposition of observed 50×50 axial slices to generate simulated images $\boldsymbol{X}_i^S = \sum_{k=1}^{50} c_{ik}\phi_k$. Here $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_{50}\}$ are eigen-images with accompanying eigenvalues $\boldsymbol{\lambda} = \{\lambda_1, ..., \lambda_{50}\}$, and PC loadings $\boldsymbol{c}_i$ are generated $\boldsymbol{c}_i \sim \text{N}\,[0, \text{diag}(\boldsymbol{\lambda})]$. The coefficient function $\boldsymbol{\beta}$ contains a single rectangular (30×15) predictive region in which the regression coefficients β_l are uniformly equal to 1. As in Section 3, we choose three signal-to-noise ratios $\frac{\sigma_{\boldsymbol{y}}^2}{\sigma_\epsilon^2} = 3, 1, 1/3$.

	$I = 100$			$I = 500$		
$\sigma_{\boldsymbol{y}}^2/\sigma_\epsilon^2 =$	3	1	1/3	3	1	1/3
GMRF - 2D						
MSE_1	0.172	0.273	0.472	0.079	0.117	0.185
MSE_0	0.006	0.019	0.044	0.005	0.007	0.012
Computation	104.4	101.7	104.8	169.5	167.8	165.6
EX. - 2D						
MSE_1	0.462	0.651	0.778	0.370	0.455	0.550
MSE_0	0.026	0.026	0.030	0.003	0.007	0.020
Computation	86.2	92.8	88.9	141.9	143.9	141.5
FPCR - 2D						
MSE_1	0.175	0.256	0.363	0.103	0.133	0.192
MSE_0	0.034	0.040	0.046	0.026	0.029	0.034
Computation	0.4	0.4	0.4	1.8	1.8	1.8

Table C.1: Average mean squared error separated by true predictive and non-predictive location, signal-to-noise ratio, sample size, predictor dimension, and estimation technique ("GMRF" labels the Gaussian MRF prior, "EX" labels the exchangeable prior, and "FPCR" the functional approach). Average computation time (in seconds) to fit the model is also shown.

For each signal-to-noise ratio we generate 500 datasets for both $I = 100$ and $I = 500$ and apply the proposed method, the related scalar-on-image model with an exchangeable prior on the regression coefficients, and the FPCR approach. Tuning parameters are chosen using the first simulated dataset, and the results below may not be fully representative of the performance of the three methods.

Table C.1 shows the MSE taken over all simulated datasets for each combination of sample size and signal-to-noise ratio. The proposed method provides good estimates of the coefficient image in both predictive and non-predictive regions, and performance improves as sample size increases and as signal-to-noise ratio increases. The proposed method substantially outperforms the related approach that uses an exchangeable prior distribution: given the homogeneity of the predictive region, encouraging spatial smoothness improves estimation. The proposed method often outperforms the FPCR approach as well, with the exception of $I = 100$ and $\frac{\sigma_y^2}{\sigma_\epsilon^2} = 1/3$. While these approaches are much more comparable in terms of MSE here than in the bump-and-slab setting considered in Section 3, there remain significant qualitative differences in estimated coefficient images. As Figure C.1 shows, estimated coefficient images for the proposed method contain many locations that are shrunk entirely to zero due to the sparsity constraint. On the other hand, the FPCR method does not impose sparsity, but induces sparsity both through penalization and a basis representation that avoids voxel-by-voxel estimation.

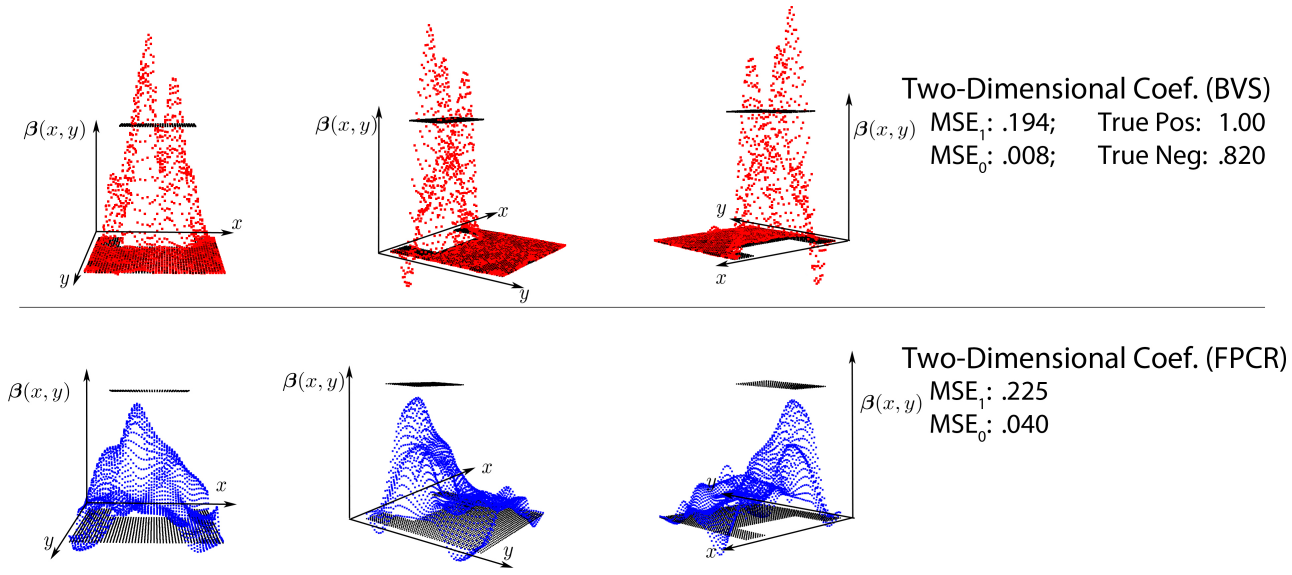

Figure C.1: Plot of typical estimated coefficient images from the simulation with a large, homogeneous predictive region. Estimates from the proposed method (labelled "BVS") and the FPCR method are shown, along with the corresponding MSE_1 and MSE_0. The true coefficient is shown in black.

C.2 Impact of Registration Errors

In this simulation we construct simulated predictors to illustrate the effect of registration error on the proposed procedure. To do so, we modify the previous method for constructing predictors in the following way. From the observed data we extract 56×56 axial slices and perform a principal components decomposition of these images. Simulated images are constructed using $\boldsymbol{X}_i^S = \sum_{k=1}^{50} c_{ik}\phi_k$, where $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_{50}\}$ are eigen-images with accompanying eigenvalues $\boldsymbol{\lambda} = \{\lambda_1, ..., \lambda_{50}\}$, and PC loadings $\boldsymbol{c}_i$ are generated $\boldsymbol{c}_i \sim \mathrm{N}\left[0, \mathrm{diag}(\boldsymbol{\lambda})\right]$. The coefficient function $\boldsymbol{\beta}$ is defined as in Section 3 using bivariate normal density functions and is applied to the middle 50×50 component of each 56×56 simulated predictor to generate outcomes. As in Section 3, we choose three signal-to-noise ratios $\frac{\sigma_{\boldsymbol{y}}^2}{\sigma_\epsilon^2} = 3, 1, 1/3$.

	$I = 100$			$I = 500$		
$\sigma_{\boldsymbol{y}}^2/\sigma_\epsilon^2 =$	3	1	1/3	3	1	1/3
GMRF – 2D						
MSE_1	2.511	3.053	3.781	1.576	1.712	2.214
MSE_0	0.065	0.082	0.218	0.136	0.110	0.143
Computation	112.8	111.7	111.9	163.5	164.9	164.4
EX. – 2D						
MSE_1	2.731	3.126	3.744	1.977	2.182	2.584
MSE_0	0.233	0.194	0.385	0.110	0.120	0.227
Computation	96.8	95.0	95.3	148.9	147.8	147.2
FPCR – 2D						
MSE_1	2.928	3.753	4.635	1.902	2.151	2.784
MSE_0	0.091	0.058	0.027	0.133	0.123	0.099
Computation	0.4	0.4	0.4	2.1	2.1	2.1

Table C.2: Average mean squared error separated by true predictive and non-predictive location, signal-to-noise ratio, sample size, predictor dimension, and estimation technique ("GMRF" labels the Gaussian MRF prior, "EX" labels the exchangeable prior, and "FPCR" the functional approach). Average computation time (in seconds) to fit the model is also shown.

To simulate registration errors, the "observed" predictors in each simulation are 50×50 images that are off-center components of the complete 56×56 generated images; the predictors are randomly shifted horizontally and vertically by $\{-3, -2, -1, 0, 1, 2, 3\}$ voxels with probability $\{.1, .1, .2, .2, .2, .1, .1\}$. Thus, each observed predictor is shifted from the true generating image by a random amount. We also note that due to the size of predictive regions in the coefficient image, these registration errors are substantial.

Table C.2 shows the MSE taken over all simulated datasets for each combination of sample size and signal-to-noise ratio. A comparison with Table 1 comparison, which provides results for the same simula-

tion design without registration error, unsurprisingly indicates that all methods have higher MSEs in the presence of registration error. Moreover, the results indicate that the proposed method outperforms the competing approaches in terms of MSE for predictive regions. The FPCR has comparable and sometimes lower MSE on non-predictive regions; as Figure C.2 shows, this is due to additional regions declared predictive by the proposed method as well as general oversmoothing by the FPCR method. As expected, performance improves for all methods as signal strength rises and as sample size increases.

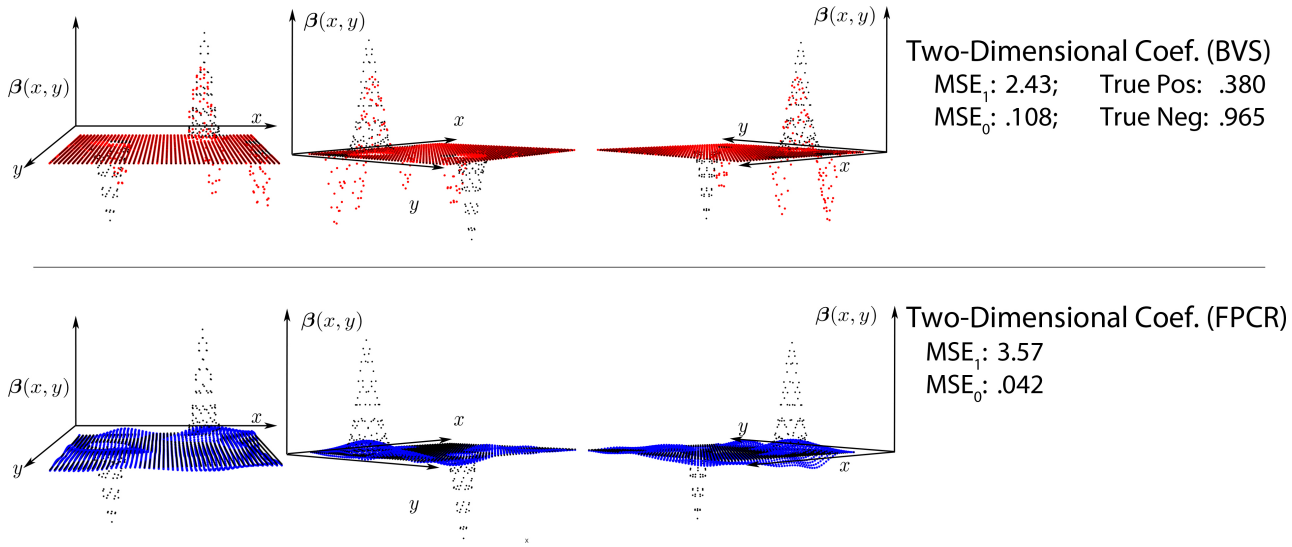

Figure C.2: Plot of typical estimated coefficient images from the simulation with registration errors. Estimates from the proposed method (labelled "BVS") and the FPCR method are shown, along with the corresponding MSE_1 and MSE_0. The true coefficient is shown in black.