HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2017 October 30.

Published in final edited form as:

Stat Med. 2016 October 30; 35(24): 4380–4397. doi:10.1002/sim.6999.

A Bayesian Probit Model with Spatially Varying Coefficients for Brain Decoding using fMRI Data

Fengqing (Zoe) Zhang[a,b,*], **Wenxin Jiang**[a], **Patrick C.M. Wong**[c], and **Ji-Ping Wang**[a,*]

[a]Department of Statistics, Northwestern University, Evanston, IL 60208, USA

[b]Department of Psychology, Drexel University, Philadelphia, PA 19104, USA

[c]Department of Linguistics & Modern Languages, The Chinese University of Hong Kong, Shatin, Hong Kong, PRC

Abstract

Recent advances in human neuroimaging have shown that it is possible to accurately decode how the brain perceives information based only on non-invasive functional magnetic resonance imaging (fMRI) measurements of brain activity. Two commonly used statistical approaches, namely, univariate analysis and multivariate pattern analysis (MVPA) often lead to distinct patterns of selected voxels. One current debate in brain decoding concerns whether the brain's representation of sound categories is localized or distributed. We hypothesize that the distributed pattern of voxels selected by most MVPA models can be an artifact due to the spatial correlation among voxels. Here we propose a Bayesian spatially varying coefficient model, where the spatial correlation is modeled through the variance-covariance matrix of the model coefficients. Combined with a proposed region selection strategy, we demonstrate that our approach is effective in identifying the truly localized patterns of the voxels, while maintaining robustness to discover truly distributed pattern. In addition, we show that localized or clustered patterns can be artificially identified as distributed if without proper usage of the spatial correlation information in fMRI data.

Keywords

Brain decoding; Multivariate pattern analysis; fMRI; Variable selection; Classification

1. Introduction

Recent advances in human neuroimaging have shown that it is possible to accurately decode how the brain perceives information based only on non-invasive functional magnetic resonance imaging (fMRI) measurements of brain activity [1, 2, 3, 4]. Brain decoding is to "read out" what stimuli or mental states are represented by an observed pattern of brain activity. Imagine the scenario that we can tell what someone is currently reading, watching, listening, and thinking based only on the fMRI measurements of brain activity. Such a brain-based decoder would have great usage in unraveling the mechanism of brain in information

*Correspondence to: Fengqing (Zoe) Zhang, Department of Psychology, Drexel University, Philadelphia, PA 19104, USA. fz53@drexel.edu or Ji-Ping Wang, Department of Statistics, 2006 Sheridan Rd, Evanston IL 60208, zwang@northwestern.edu.

perception and learning, as well as in diagnosing of brain deficits, and even building brain-machine interfaces [1, 2]. To perform accurate and efficient brain decoding, more and more attentions have been drawn to statistical analysis of fMRI data [5, 6, 7].

In general, there are two types of statistical methods in human neuroimaging, namely univariate analysis and multivariate pattern analysis (MVPA). A conventional univariate approach typically comprises two steps. In the first step, one uses General Linear Model (GLM) to build regression models to estimate the activation profile for each voxel from the measured fMRI time series. In the second, statistical tests (such as Student's t-test) are applied to compare brain activation status for each of many thousands of voxels independently across different experimental conditions [8, 6]. Brain regions that respond more to one condition than the other are defined by combining all voxels with p-values less than a chosen significance level. As the resolution of fMRI technology improves, the voxel size becomes finer and smaller and hence the difference between experimental conditions at individual voxel level often becomes too small to be detected by conventional univariate approaches [1, 6]. In contrast, decoding-based neuroimaging uses MVPA to combine information from many voxels simultaneously to build machine learning classifiers, which tend to be more powerful in pattern identification for differentiating experimental conditions [1, 5]. For example, if two voxels individually do not carry sufficient information to discriminate different conditions, but jointly they might achieve the power in MVPA. Therefore, the MVPA approach is more sensitive and more informative for distinguishing different mental states than the univariate approach. Decoding-based neuroimaging allows us to determine the brain's presentation of mental contents, rather than overall levels of activation. It differs from the conventional neuroimaging by reversing the classical direction of inference that attempts to explain brain activity from different mental states. Instead of testing whether a single voxel responds more to one condition than another, it asks whether it is possible to predict which of the two or more conditions the subject is in based on the observed pattern of brain activity over a set of voxels. As often only a subset of voxels are functionally essential in differentiating different mental states, voxel selection becomes an important step in MVPA.

The two different types of approaches often lead to distinct patterns of selected voxels, resulting in an ongoing debate between the two different views in cognitive neuroscience [9]. The univariate approach tends to reveal localized brain regions (or clustered voxels in local regions) that are associated with different experimental conditions, whereas the MVPA approach often selects a set of distributed or scattered voxels in the classification models (Figure 1). For example, several groups used the univariate approaches and identified selective regions in the auditory cortex that potentially account for sound categories differentiation [10, 11, 12]. However, Staeren et al. found that sound categories are represented by a set of distributed voxels in the human auditory cortex by using support vector machine with recursive feature elimination (SVM-RFE, one type of MVPA models) [13].

Regardless of which view could be closer to the truth, there is a crucial issue related to the MVPA approach that has not been addressed thoroughly in fMRI literature. A critical step to build an MVPA model concerns selection of a subset of voxels or variables out of thousands

or even more. The fMRI data is spatially correlated in the sense that nearby voxels tend to have similar responses to experiment stimuli. In statistics, multicollinearity typically discourages the correlated variables from being simultaneously selected. For example, regression models are often ill-posed when correlations among variables are high. Zou and Hastie pointed out that LASSO method tends to select one important variable and ignore other variables that are highly correlated with it [14]. Therefore, in the sound category perception example above, even if the true underlying brain pattern is localized, voxels from these important regions may be only partially selected due to their spatial correlation, resulting in distributed patterns. Thus how to effectively account for the spatial correlation becomes critical to recover the true brain response pattern in neuroimaging.

The MVPA approach to decoding-based neuroimaging can be summarized in statistical language as a variable selection problem in classification for high dimensional spatially correlated data. Some existing approaches include SVM-RFE [15, 16], searchlight analysis [17], logistic regression and logistic regression with L_2 regularization [18]. These methods do not explicitly model the spatial correlation in fMRI data. More recently, spatial regularization in multivariate decoding has been implemented, either in a Bayesian [19], or classical framework [20, 21, 22, 23, 24, 25, 26, 27]. The focus of our paper is different from these previous works; we introduce a statistical methodology that is particularly suitable for studying carefully the patterns of selected voxels in fMRI. To account for the spatial structure in the fMRI data, we propose a Bayesian probit model with spatially varying coefficients (BPSV). We develop a region selection strategy and a computational approach through variational Bayesian approximation. In simulation studies and real fMRI data analysis, we shall compare three MVPA models: BPSV which accounts for spatial correlation, and the Elastic Net and SVM-RFE, which do not account for spatial correlation. Our emphasis is *not* on predictive performance (incidentally, all three methods achieve good and comparable classification performance). Rather, we demonstrate that given the same dataset, while all using multivariate approaches, the Elastic Net and SVM-RFE tend to select more distributed voxels (even when the true pattern is localized), while the BPSV approach can correctly pick up the difference, i.e., it identifies more localized regions (and does not make false claims when the true pattern is distributed). This, together with localized findings reported in the univariate studies cited earlier [10, 11, 12], leads us to conjecture that the distributed patterns reported in some MVPA approaches might be an artifact.

2. Statistical methods

2.1. General linear model based univariate approach

In the literature, an additive regression-based model is commonly used for fMRI signal analysis [28]:

$$y_{i,t}=b_{i,t}+f_{i,t}+e_{i,t}, i=1,\ldots,N;t=1,\ldots,T, \quad (1)$$

where $y_{i,t}$ represents the fMRI signal of voxel i at time t, $\{b_{i,t}; t = 1, \ldots, T\}$ denotes the baseline trend, $f_{i,t}$ is the activation profile, and $e_{i,t}$ is the random error. Different models often

vary in the way of specification of these three components and the way in which spatial correlation is taken into account.

In the conventional neuroimaging approach, a General Linear Model (GLM) is often assumed [29]. The first term $b_{i,t}$ in Eq. (1) is modeled as a linear combination of a few simple basis functions. The second term $f_{i,b}$ which is the activation profile, is defined as the product of a scalar activation effect, delta, with a transformed stimulus. The transformed stimulus is a delayed and continuously modified version of the original 0 – 1 stimulus. This transformation is called the hemodynamic response function (HRF), which is typically realized by convoluting the stimulus signal with a Gamma density. Other suggested forms of HRF include discretized Poisson and Gaussian density [30]. The contrasts of these estimated delta values can then be tested using the univariate approach to determine whether a single voxel responds more to one condition than another. This approach seeks to find voxels whose activation time series can be explained by alternations between conditions of interest.

2.2. General framework of MVPA

The MVPA approach inverts the classical direction of inference in the conventional univariate approach. Instead of finding voxels whose activation can be explained by alternations between conditions of interest, it asks whether experimental conditions can be differentiated from one another based on the observed pattern of brain activity.

In general, there are two basic steps for the applications of MVPA in brain decoding, model training and model testing [6]. During the training step, the MVPA model is built by learning a functional relationship between brain response patterns and mental states (experimental conditions) using the training data set. Experimental conditions are usually labeled as discrete values. During the testing step, the newly trained MVPA model is used to classify the experimental conditions of an independent data set (test data) based on the observed brain patterns. Classification accuracy is used for model evaluation.

Consider $\{(y_i, \mathbf{X}_i) : i = 1, \ldots, m\}$, a collection of m independent observations where y_i is a binary response variable representing two experimental conditions, and $\mathbf{X}_i^T = \{x_{i1}, \ldots, x_{in}\}$ is the covariate vector associated with y_i, and $n >> m$. Here, y_i can be zero or one. The observed data $\mathbf{X}_i$ is collected over spatial locations $s_1, \ldots, s_n$. For fMRI data, each $\mathbf{X}_i$ represents the brain activity measured over n voxels in the ith observation. That is the single trial-response estimation (the scalar activation effect) over n voxels by fitting a General Linear Model to each stimulus [15, 31]. The MVPA model first learns a decision function (or a discriminant function), denoted as $f(\mathbf{X}_i)$, which is a scalar function of the input brain response patterns $\mathbf{X}_i$. In the case of linear classification of fMRI responses, $\mathbf{X}_i$ are classified according to the sign of the decision function:

$$f(\mathbf{X}_i) = \mathbf{w}^T \mathbf{X}_i + b,$$

where $\mathbf{w}^T$ is a $1 \times n$ weight vector and b is a bias term or threshold weight. The decision rule for a binary classifier is:

$$y_i=1, \mathbf{X}_i \in \text{class } (+), \text{ if } f(\mathbf{X}_i)>0,$$

$$y_i=0, \mathbf{X}_i \in \text{class } (-), \text{ if } f(\mathbf{X}_i)<0.$$

2.3. Bayesian probit model with spatially varying coefficients

In order to identify the brain's representation of different mental states, voxel selection is an important step in MVPA. Adding more voxels with relevant information for distinguishing different conditions can help to improve classification accuracy, while adding uninformative voxels may simply add noise to the model. In general, there are three main categories of variable selection, wrappers, filters and embedded methods. In the literature, different methods have been proposed to incorporate the spatial information in the GLM based univariate approach. However, careful consideration needs to be given to the voxel selection step of the MVPA approach. Due to the multicollinearity problem, correlated variables are often prevented from simultaneous selection in the classification model. For fMRI data, anatomically the voxels close to each other tend to have similar brain functions, and thus not only the measured activities x_{ij} are correlated (as observed), but also it is believed that they have similar effect on the response variable y_i. Hence, even if the true underlying pattern of brain activity is clustered, voxels from these clustered regions may be only partially selected due to their spatial correlation. This results in the voxels selected by MVPA models tend to be distributed in the brain. In order to accurately identify the brain's representation of different mental states, it is necessary to account for the spatial correlation during the voxel selection step of the MVPA approach. Therefore, we propose to model the data with spatially varying coefficients under a Bayesian framework.

To make the spatial information contained in $\mathbf{X}$ more explicit, we write $\mathbf{X}_i \equiv \mathbf{X}_i(\mathbf{s})$, or $x_{ij} = \mathbf{X}_i(s_j)$. We assume a probit model as follows:

$$y_i|p_i \sim \text{Bernoulli}\,(p_i)\,,$$

$$\Phi^{-1}(p_i)=\mathbf{X}_i(\mathbf{s})^T\boldsymbol{\beta}_1(\mathbf{s})+\beta_0, \quad (2)$$

where Φ^{-1} denotes the inverse of the standard normal cumulative distribution function, β_0 is the intercept, and $\boldsymbol{\beta}_1(\mathbf{s})$ are the spatially varying coefficients. The probit model in Eq. (2) can be alternatively formulated by introducing an auxiliary variable z_i as follows [32]:

$$z_i=\mathbf{X}_i(\mathbf{s})^T\boldsymbol{\beta}_1(\mathbf{s})+\beta_0+\varepsilon_i,$$

where $\boldsymbol{\varepsilon}_i$ are i.i.d. $N(0, 1)$. Let $y_i = 1$ if $z_i > 0$, otherwise $y_i = 0$.

To account for the spatial effects, we impose a prior distribution for $\boldsymbol{\beta}_1(\mathbf{s})$ as follows:

$$f(\boldsymbol{\beta}_1(\mathbf{s})|\gamma)=N(\mathbf{0},\sigma^2\mathbf{H}(\gamma)), \quad (3)$$

where **H** is a correlation matrix which is allowed to depend on an unknown parameter $\boldsymbol{\gamma}$ and $\boldsymbol{\sigma}^2$ is the scalar variance. This general framework allows different possibilities, and we give an example related to the Matérn correlation function in the Appendix.

2.4. Region selection for the Bayesian probit model

The central task in our problem is to select a small set of voxels that can effectively classify different conditions. Thus we would like to fit a sparse model such that at most locations the regression coefficient $\beta_1(s_j)$ is zero or nearly zero. To this end, we modified the covariance matrix in the prior distribution in Eq. (3) as

$$f(\boldsymbol{\beta}_1(\mathbf{s})|\boldsymbol{\alpha},\gamma)=N(\mathbf{0},\mathbf{D}_{\boldsymbol{\alpha}}\mathbf{H}(\gamma)\mathbf{D}_{\boldsymbol{\alpha}}), \quad (4)$$

where $\mathbf{D}_{\boldsymbol{\alpha}} = diag[\tilde{a}_1\tau_1, \ldots, \tilde{a}_n\tau_n]$, $\tilde{a}_j = 1$ if $\alpha_j = 0$ and $\tilde{a}_j = c_j$ if $\alpha_j = 1$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ is an indicator vector for variable selection. We set τ_j (> 0) small and c_j (always > 1) large to make those $\beta_1(s_j)$ with $\alpha_j = 0$ clustered around 0 with variance τ_j, whereas those $\beta_1(s_j)$ with $\alpha_j = 1$ more dispersed with variance $c_j\tau_j$. Under this setting, the prior of $\beta_1(s_j)$ corresponds to a mixture of two normal distributions, and is more effective to differentiate the two groups of $\beta_1(s_j)$'s [33]. In addition, we assume that $\alpha_1, \ldots, \alpha_n$ are i.i.d. observations from the Bernoulli distribution $P(\alpha_j = 1) = 1 - P(\alpha_j = 0) = p_j = 0.5$. The prior distribution for β_0 is specified as $\pi(\beta_0) = N(0, c_3)$, where $c_3 = 10^5$. The prior of $\boldsymbol{\gamma}$ depends on the model of the correlation function $\mathbf{H}(\boldsymbol{\gamma})$. In the Appendix, we consider an example with a Matérn correlation function with a decay parameter $\boldsymbol{\gamma} > 0$, for which we use $\boldsymbol{\pi}(\boldsymbol{\gamma}) = N(a_2, b_2)$ truncated to be positive. We will discuss the choice of a_2 and b_2 using the empirical semivariogram in the Results Section. All the priors are assumed to be independent. A graphical representation of our BPSV model that depicts the dependency relationship between the variables is presented in Figure 2.

3. Results

In this section we investigate the performance of the proposed BPSV model with both simulated data and real fMRI data. The computation is implemented by a variational Bayesian algorithm (see Appendix). The other two methods to be compared include the elastic net (EN) and support vector machine with recursive feature elimination (SVM-RFE). The EN approach is developed for high dimensional data, which is designed to group important but correlated covariates. We are interested in its performance in the spatially correlated data. The SVM-RFE algorithm has been extensively used in fMRI data analysis [15, 31, 34, 13]. In particular the work by Staeren et al. suggested distributed patterns of sound category representation in the human brain [13]. It is our key interest whether this pattern could be an artifact due to the method itself.

3.1. Simulations

Our simulation studies are designed to evaluate the performance of the proposed BPSV method in identification of regions of voxels even if they are highly correlated in fMRI data. In comparison with other approaches, we are particularly interested in whether localized or clustered patterns can be artificially identified as distributed if without proper usage of the spatial correlation information in fMRI data. Conversely we would like to investigate whether the specified spatial pattern in BPSV could result in false clustered pattern in situations where the true pattern is distributed.

In Simulation I, we considered a 31 × 33 spatial map with $n = 1023$ locations (predictors) in the two dimensional space where the true pattern is clustered at 162 sites with sample size $m = 100$ and 600. The 162 sites with nonzero coefficients are clustered in three rectangular regions. The $\mathbf{X}_i(\mathbf{s})$ were independently simulated from $N(\mathbf{0}, 2\mathbf{H}(\gamma))$. Spatial correlation among these 1023 locations was imposed through the exponential correlation function $\mathbf{H}(\gamma)$ with $\gamma = 2$. The intercept in the true probit model was set as 0.5. The 162 nonzero coefficients $\{\beta_1(s_j)\}$ were simulated from $N_{162}(\mathbf{2}, 1.5\mathbf{H}(\gamma))$. The coefficients for all rest sites were set as zero. The simulated true coefficients $\boldsymbol{\beta}_1(\boldsymbol{s})$ are presented as a heatmap in the top left panel of Figure 3.

With fixed $\boldsymbol{\beta}_1(\mathbf{s})$ and β_0, we simulated nine independent samples (i.e., $(\mathbf{X}, \mathbf{Y})$) at two different sample sizes ($m = 100,600$). The $\boldsymbol{\beta}_1(\mathbf{s})$ estimates are averaged across independent Monte Carlo samples and presented in Figure 3. For SVM-RFE, the presented $\boldsymbol{\beta}_1(\mathbf{s})$ correspond to the weights of each voxel in the linear classifier defined in hyperplanes that best separate the observed data, which are analogous to the regression coefficients (x_i's are all standardized). Nevertheless, the $\boldsymbol{\beta}_1(\mathbf{s})$ values estimated from different methods could be defined in different scales, thus we normalized them by the L_1 norm, namely

$$\frac{1}{n}\sum_{j=1}^{n}|\beta_1(s_j)|.$$

The averages of normalized coefficient estimates from BPSV, EN, and SVM-RFE are presented as heatmaps in Figure 3. At $m = 100$, the BPSV successfully identified the three true clusters, whereas EN and SVM-RFE both identified only two clusters with a more noisy pattern. The increase of sample size results in better definition of the three clusters in all methods, while BPSV clearly provides the best recovery of the three rectangular regions of voxels in the true model.

We further quantified the performance of each method from three different perspectives. First, we calculated the Pearson correlation between the estimated regression coefficients and the true coefficients (In the SVM case, we used the weight parameter instead) [35]. The correlation coefficients averaged over nine replications are summarized in Table 1. Second, as one of the key interests in the fMRI studies is to identify important voxels that are associated with cognitive brain activity, we compared different approaches from the variable selection perspective in the format of the receiver operating characteristics (ROC) curve (Figure 4). Define the true or false positive rate as the fraction of the true positive or false positive features respectively that are selected by the classification model. Instead of a single optimal output, we can plot the true positive rate vs. false positive rate in terms of the

selected features by varying variable selection criteria (tuning parameter values) within each method. For BPSV, each component of $\boldsymbol{\beta}_1(\boldsymbol{s})$ is assumed to arise from a mixture of two normal distributions with different variances indicated by α_j. If $\alpha_j = 0$, then β_j would be probably so small that it would be "safely" estimated by zero. Nevertheless the posterior average rarely gives exact zero. For the variable selection purpose, we need to threshold $\boldsymbol{\beta}_1(\boldsymbol{s})$ to obtain a compact set of predictors in the classification model. Selection of predictors in the classification model can also be achieved by thresholding posterior expectation of $\boldsymbol{\alpha}$ values. We chose the former because thresholding $\boldsymbol{\beta}_1(\boldsymbol{s})$ is more straightforward and consistent with the selection methodologies of the other algorithms. In this case, the sensitivity and specificity are direct functions of the threshold values for $\boldsymbol{\beta}_1(\boldsymbol{s})$. The optimal threshold value of $\boldsymbol{\beta}_1(\boldsymbol{s})$ is selected by cross-validation. The EN model has two tuning parameters λ_1 and λ_2. Zou and Hastie in their paper suggested to pick a (relatively small) grid of values for λ_2 followed by selection of λ_1 using cross-validation [14]. The optimal λ_2 is the one giving the smallest cross-validation error. Hence, in Figure 4, we showed the sensitivities and specificities of the EN model calculated at seven different λ_2 values (0.0001,0.001,0.01,0.1,1,10,100) with each λ_1 chosen by ten-fold cross-validation. The SVM-RFE proceeds with two sequential steps iteratively including classifier construction and recursive feature elimination (RFE). In the classifier construction step, hyperparameters of SVM are selected using a grid search. In the RFE step, all features will be ranked according to their weights computed from SVM classifiers. Sensitivities and specificities of SVM-RFE are thus functions of the number of features included in the classifier from the top list [16]. The optimal number of features included in SVM is selected by cross-validation.

Lastly, we assessed the classification accuracy at the optimal model output. Based on original 18 training samples (half at $m = 100$ and half at $m = 600$) we generated 18 additional test samples from the same model, half at sample size $m = 50$ and half at $m = 300$. Each training sample was paired with a test sample of half sample size. A four-fold cross-validation was carried out to build an optimal model based on the training data, which was subsequently applied to the paired test sample for classification accuracy assessment. For example, for BPSV, the optimal choice of threshold value that achieved best classification accuracy was first obtained by a four-fold cross-validation based on the training sample. Using the optimal threshold value, a new model was built based on the entire training sample. For EN, the optimal model was built based on the optimal choice of (λ_1, λ_2), whereas for SVM-FRFE, the optimal number of features to be selected was determined for building the optimal model. The average classification accuracy over 9 replicated test samples of each sample size is presented in Table 1.

The visual observations from Figure 3 were confirmed by the significantly higher correlation between the true and fitted coefficients (or feature weight in SVM-RFE) and better classification accuracy achieved in BPSV than the other two (Table 1). The ROC curves in Figure 4 illustrate that BPSV can much more effectively identify the true features than SVM-RFE while controlling the false positive rate. We note that the EN approach tends to have strict control of the false positive rate regardless of choice of tuning parameter λ_2. The quantitative measures confirm that the proposed BPSV is more effective than other

approaches in identifying the true clustered features while providing better estimates of regression coefficients and better classification accuracy.

The proposed BPSV model assumes that nearby covariates tend to have similar effects on the response variable. One natural question is, whether violation of this assumption would result in falsely classified clusters when the pattern is truly distributed. Simulation II was intended to investigate the robustness of the proposed approach to model-misspecification. Again we considered a 31 × 33 spatial map with $n = 1023$ locations (predictors) in the two dimensional space where the true pattern involved 50 voxels distributed over the map without spatial correlation (top left panel of Figure 5). Each x_{ij} was independently simulated from $N(0, 1)$. The intercept β_0 was 0.5. The 50 nonzero coefficients $\{\beta_1(s_j)\}$ were independently simulated from $N(0, 2^2)$ whereas for the rest locations, the coefficients $\beta_1(s_j)$ were all set to be zero. As in Simulations I, nine training data sets were simulated at each sample size of $m = 100$ and 600, each of which was paired with a testing sample of half sample size.

At either sample size, the BPSV approach did not generate noticeable artificial "clustered" pattern around true features (Figure 5), suggesting a desirable robustness against the misspecification of spatial correlation in the prior. The correlation and prediction accuracy were comparable among the three approaches (Table 2). Interestingly, the ROC curve suggests that BPSV still has better sensitivity over SVM-RFE (Figure 6).

In Simulation III, we investigated how other spatial correlation structures could affect the performance of the proposed model. Here the spatial correlation was simulated in a discrete fashion while in Simulations I the spatial correlation was simulated to decay continuously over space. The basic setting is similar to Simulation I, where a 31 × 33 layout with 1023 sites is considered, out of which 162 are associated with nonzero coefficients (top left panel of Figure 7). The true coefficients $\boldsymbol{\beta}_1(\boldsymbol{s})$ are identical to Simulations I, while data are generated according to the model with a different error correlation structure [36] as follows:

$$x(s_j) = \mu(s_j) + \varepsilon(s_j), j = 1, \ldots, 1023,$$

where the signals are $\mu(s_j) = 4$ for the 162 sites while for other sites $\mu(s_j) = 0$. The errors $\{\varepsilon(s_j)\}$ have zero-mean, unit-variance and are spatially dependent, by taking $\varepsilon(s_j) = \{\sum_{k \in N} e(s_k)\} / \sqrt{5}$, where N denotes a neighborhood including the center site s_j and four other sites that touch one of the center location's edges. Here $\{e(s_j)\}_{j=1}^{1023}$ are i.i.d. $N(0,1)$. Under this error structure, nearby voxels tend to maintain a positive correlation that decays as a function of distance, but not in an exact exponential form. The training and testing samples were simulated in the same way as previous simulations.

At sample size 100, all three true clusters were identified by the BPSV approach, whereas EN and SVM-RFE both identified a noisy distributed pattern (Figure 7). At sample size 600, though all methods provided a better recovery of the true pattern, the BPSV outperformed the other two, evidenced by the higher correlation between $\boldsymbol{\beta}_1(\mathbf{s})$ estimates and their true values, the better classification accuracy (Table 3), and the larger areas under ROC curves

(Figure 8) obtained by BPSV than the other two methods. Compared to the results in Simulation I (Table 1), the change of correlation structure has posed more challenges to all methods. For BPSV, the $\boldsymbol{\beta}_1(\mathbf{s})$ estimates and the classification accuracy were both mildly affected, particularly at the small sample size, resulting in a decline in the correlation between the estimated and true $\boldsymbol{\beta}_1(\mathbf{s})$ (0.64 to 0.51) and the classification accuracy (0.88 to 0.76). A similar effect was also observed for the EN and SVM-RFE approaches. Nevertheless the BPSV approach maintains to be more attractive than EN and SVM-RFE.

3.2. A case study for fMRI data

The development of the BPSV method was motivated by our interest in studying brain activities in language learning and speech disorder settings. To illustrate our method, we applied all three models (EN, SVM-RFE and BPSV) to the real data collected recently in a bilingual fMRI study, where each subject was asked to make semantic congruency judgment during fMRI scanning [37]. There were two experimental conditions, namely, sentence congruency (SC) and sentence violation (SV). Each subject experienced the two conditions for 24 times respectively. The real-time brain activity in the format of a time series was recorded using Siemens Trio 3T MRI scanner. Before applying MVPA models, standard data preprocessing steps were performed using software AFNI [38]. A trial estimate of the brain response at every voxel was obtained by fitting a general linear model for each subject with predictors for two experimental conditions (SC and SV). These predictors were adjusted by convolving with a gamma hemodynamic response function to account for the hemodynamic response delay. The left superior and inferior frontal gyri in the brain are of particular interest in sentence-level semantic integration [37]. Thus our analysis below is based on the region of interest (ROI) with 559 voxels mainly from the left superior and inferior frontal gyri of three subjects.

Before analyzing the real data, we first discuss some diagnostic methods for spatial correlation. Empirical semivariogram plot has been used to diagnose spatial correlation as a function of the spatial distance [39]. Figure 9 presents the semivariance values as a function of distance for the first two simulation settings and the real data. The plot for the real fMRI data resembles that in Simulation I, showing an increasing pattern of semivariance as a function of distance, suggesting a strong spatial correlation. In contrast, a flat semivariance curve for Simulation II suggests spatial independence between locations. Empirical semivariograms can be used to help choose the hyperparameters in the priors [40], particularly here, the choice of the hyperparamters a_2 and b_2 in the prior distribution for the decay parameter γ. Recall that the prior distribution on γ is specified as $N(a_2, b_2)$ truncated to be positive. In Simulation I, the range, that is the distance at which data are no longer autocorrelated, is estimated as about 6 by using standard geostatistical package [41]. For exponential correlation function, the effective range of spatial dependence (the distance at which the correlation drops to 0.05) is determined by $-log(0.05)\gamma$ (that is 3γ) [40]. Therefore, in Simulation I, we set a_2 to be 2 (solving equation $3\gamma = 6$). The variance b_2 should be chosen to give support to possible values of γ and was set to be 0.3^2. In Simulation II, we set a_2 to be 0.4 and b_2 to be 0.3^2 since the empirical semivariograms suggest weak/no spatial correlations.

As in the simulations, we carried out a four-fold cross-validation for each subject. We first divided the data from each subject into four folds. Each of the four folds served as the test sample once and the other three folds served as the training sample. Further, a three-fold cross-validation was carried out to build an optimal model based on the training sample, which was subsequently applied to the corresponding test sample for classification accuracy assessment. Since BPSV was evaluated on the test sample four times in cross-validation for each of the three subjects, a voxel can be selected at most twelve times. Voxels selected by BPSV from the ROI for each subject were combined and mapped back to the three dimensional brain space by using software AFNI. A weight for each voxel was defined as the relative frequency of selection out of twelve. A larger weight is regarded to indicate larger likelihood of active role of the given voxel in differentiating SC and SV conditions. In Figure 10, the selected voxels with weight larger than 0 and less than 0.5 are shown in dark blue while voxels with weight between 0.5 and 1 are plotted in red. The same steps were done for voxels selected by SVM-RFE.

The voxels selected by SVM-RFE and BPSV are shown in the left and right columns in Figure 10 Parts A, B, C and D. These four parts show the Axial and Sagittal images of the brain with selected voxels at the centers of the four largest clusters identified by BPSV respectively. Voxels selected by EN are not represented in the figure because very few voxels (less than 10) were selected each time in the four-fold cross-validation for each subject. Here, a voxel can be considered as a three dimensional cube and a cluster is defined as a set of voxels touching each other's face by AFNI software. The voxels that are selected more often by the model represent a more reliable brain response pattern across repetitions and subjects. Figure 10 suggested that the voxels selected by BPSV tend to be more clustered and be associated with larger weights than voxels selected by SVM-RFE. This demonstrates the advantages of our proposed model on identifying clustered brain regions and consistent brain patterns across different subjects. Classification accuracies for each subject from the four-fold cross-validation are shown in Table 4. Further, we computed the proportions of single voxels out of all selected voxels, and the proportions of voxels in the largest cluster out of all selected voxels as clusteredness indexes to quantify how clustered the selected voxels are (Table 5). Because very few voxels were selected by EN each time in the cross-validation for each subject, clusteredness indexes are not computed for EN. The summary statistics from Tables 4 and 5 again support the observations from Figure 10. Interestingly, despite the substantial difference in the pattern, i.e. distributed vs clustered, the classification accuracy appears to be comparable. One possible explanation is that these three models have different variable selection strategies. Although none of them can achieve perfect classification accuracy, they managed to obtain similar accuracy with different selected variables. This phenomenon is also observed in other studies [42, 43].

4. Conclusion and Discussion

In this paper we have proposed a spatially varying coefficient model under a Bayesian framework with an effective region selection strategy for brain decoding. Compared with other MVPA methods, the major innovation of this approach is to explicitly account for the spatial correlation among voxels through the specification of the prior distribution of regression coefficients for better voxel selection. Our simulation studies demonstrated that

BPSV is effective in selecting clustered true variables even with high correlation while achieving sparsity (Simulation I), with desirable robustness to mis-specification of spatial correlation pattern (e.g., no spatial correlation pattern in Simulation II, and different spatial correlation structure in III). Our simulations also confirmed that the MVPA approaches including EN and SVM-RFE are less capable of identifying truly spatially clustered features in brain-decoding type of problems. In the real data example, the BPSV did suggest a more clustered pattern than the SVM and EN, but without showing significant improvement in terms of the classification accuracy. This could be due to other confounding factors, violations of model assumptions, or could be due to the nature of this problem. In the regression problem, we know that if multiple regressors are highly correlated, dropping some of them will not substantially affect the regression R^2. Likewise in the classification problem addressed in this paper, when multi-colinearity exists, conventional classification models that discourage inclusion of highly correlated variables may still achieve comparable classification accuracy. Other studies in the brain mapping literature also reported that reliability of the spatial patterns identified by the models may change without resulting in change in classification accuracy [24, 25]. Hence the quality of spatial patterns extracted from models cannot be assessed purely by focusing on prediction accuracy The real data example in this paper may present such a case.

In the brain decoding literature, the discrepancy between univariate methods and MVPA methods has been discussed [9]. For example, in the auditory domain, several brain decoding studies investigated the brain's representation of sound categories [11, 44, 13]. Several studies, using univariate analysis, supported a hierarchically organized object-processing pathway along anteroventral auditory cortex [11, 12], while others stressed the importance of distributed representations of auditory objects based on MVPA analysis results [31, 13]. Our proposed BPSV approach, although being an MVPA in nature, provides new clues to support the locally clustered pattern in brain decoding.

We are not the first to notice that MVPA method can also identify localized voxel regions. In an independent work, Gramfort et al. also obtained localized predictive regions (in their Figure 2) based on a total variation penalty [22]. We view such works as mutually augmenting rather than competing with each other. Their work is optimization-flavored, based on setting a penalty on gradients of the regression coefficients. Our work is statistically-flavored, based on modeling correlation of the regression coefficients in the Bayesian framework. Our work is necessary, since otherwise one might think that the localized regions found in [22] were due to an artifact related to their use of the gradient penalty in optimization. This is not the case, since our work, being very different (based on statistically modeling the spatial correlation), can also derive localized regions in an MVPA approach. Also, with only findings on real data (e.g., Figure 2 of [22]), one might suspect that spatial dependence approaches could even force artificially localized regions even when the true pattern is dispersed. We show that this is not the case, by running simulation studies to show that localized regions are detected only when they really exist, i.e., (very importantly,) truly distributed patterns are not falsely claimed to be localized.

Our study of region location is very carefully planned in a Bayesian framework. Our Bayesian framework allows the data to tell objectively how strong or how weak the

correlations should be. In addition, we have utilized a voxel selection mechanism with an independent prior, which allows nearby voxels to be selected independently by data, with no subjective preference in favor of localized or distributed patterns to begin with. This way, we believe that our paper makes a forceful argument that an MVPA model that accounts for spatial dependence is essential to prevent discovery of artificially distributed patterns, which can arise, paradoxically, due to existence of spatial correlations among nearby voxels, which can lead to *exclusion* of nearby voxels due to the multicollinearity phenomenon in regression.

Acknowledgments

This work was supported by the Liu Che Woo Institute of Innovative Medicine at The Chinese University of Hong Kong, the US National Institutes of Health grants R01DC013315, the Research Grants Council of Hong Kong grants 477513 and 14117514, the Health and Medical Research Fund of Hong Kong grant 01120616, and the Global Parent Child Resource Centre Limited. The authors would like to thank Gangyi Feng and Suiping Wang for sharing their data sets. This research was supported in part through the computational resources and staff contributions provided for the Social Sciences Computing cluster (SSCC) at Northwestern University. Recurring funding for the SSCC is provided by Office of the President, Weinberg College of Arts and Sciences, Kellogg School of Management, the School of Professional Studies, and Northwestern University Information Technology. We would also like to thank Jiangtao Gou for useful comments on computing algorithms, and the reviewers for their thorough work.

Appendix: Variational Bayesian

Markov chain monte carlo (MCMC) is an important tool to fit Bayesian models but can require exceedingly long computer time, when applied to large data sets or complex models. Here we consider an appealing approximation, namely Variational Bayesian (VB) for better efficiency. The VB approximation is typically much faster than the full Bayesian inference using MCMC, and in some cases it facilitates the estimation of models that would be otherwise impossible to estimate [45, 46, 47]. VB is an iterative scheme that approximates a full posterior distribution with a factorized set of distributions by maximizing a lower bound on the marginal likelihood [46].

For simplicity, we will rewrite the the concentration matrix as

$$[\mathbf{D}_{\boldsymbol{\alpha}}\mathbf{H}(\gamma)\mathbf{D}_{\boldsymbol{\alpha}}]^{-1}=\boldsymbol{C}_{\boldsymbol{\alpha}}\boldsymbol{Q}\boldsymbol{C}_{\boldsymbol{\alpha}},$$

where $\boldsymbol{Q} = \boldsymbol{H}(\boldsymbol{\gamma})^{-1}$, and $\boldsymbol{C}_{\boldsymbol{\alpha}}=\mathbf{D}_{\boldsymbol{\alpha}}^{-1}=\mathrm{diag}(c_{\alpha_1},\ldots,c_{\alpha_n})$. Let $\mathbf{c}_{\boldsymbol{\alpha}} = (c_{\alpha 1}, \ldots, c_{\alpha n})^T$ denote the vector of the diagonal elements of $\boldsymbol{C}_{\boldsymbol{\alpha}}$. The joint likelihood given the model specified above is:

$$\begin{aligned} f(\boldsymbol{Z},\boldsymbol{\beta}_1(\mathbf{s}),\beta_0,\boldsymbol{\alpha},\gamma|\boldsymbol{Y}) &\propto f(\boldsymbol{Y}|\boldsymbol{Z})f(\boldsymbol{Z}|\boldsymbol{\beta}_1(\mathbf{s}),\beta_0)f(\boldsymbol{\beta}_1(\mathbf{s})|\boldsymbol{\alpha},\gamma)f(\boldsymbol{\alpha})f(\gamma)f(\beta_0) \\ &\propto \prod_{i=1}^{m}[y_i I(z_i\geq 0)+(1-y_i)I(z_i<0)]\frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}[z_i-(\mathbf{X}_i^T\boldsymbol{\beta}_1(\mathbf{s})+\beta_0)]^2\} \\ &\times (2\pi)^{-n/2}|\boldsymbol{C}_{\boldsymbol{\alpha}}\boldsymbol{Q}\boldsymbol{C}_{\boldsymbol{\alpha}}|^{1/2}\exp\{-\frac{1}{2}\boldsymbol{\beta}_1^T(\mathbf{s})[\boldsymbol{C}_{\boldsymbol{\alpha}}\boldsymbol{Q}\boldsymbol{C}_{\boldsymbol{\alpha}}]\boldsymbol{\beta}_1(\mathbf{s})\} \\ &\times f(\gamma)\times\frac{1}{\sqrt{2\pi c_3}}\exp\{-\frac{1}{2c_3}\beta_0^2\}. \end{aligned} \tag{5}$$

Here $f(\boldsymbol{\gamma})$ is the hyperprior for $\boldsymbol{\gamma}$, e.g., $f(\gamma) \propto (\frac{1}{2})^n \frac{1}{\sqrt{2\pi b_2}} \exp\{-\frac{1}{2b_2}(\gamma - a_2)^2\} I(\gamma > 0)$, for a positive decay parameter in the Matérn correlation.

By VB, we would like to approximate the true joint likelihood in Eq. (5) by factored posteriors for different parameters such that

$$p(\boldsymbol{Z}, \boldsymbol{\beta}_1(\mathbf{s}), \beta_0, \boldsymbol{\alpha}, \gamma | \boldsymbol{Y}) \approx q(\boldsymbol{Z}) q(\boldsymbol{\beta}_1(\mathbf{s})) q(\beta_0) q(\boldsymbol{\alpha}) q(\gamma).$$

The approximate posterior for the auxiliary variables $\mathbf{Z}$ follows as

$$q(\boldsymbol{Z}) = \prod_{i=1}^{m} q(z_i) = \prod_{i=1}^{m} N_{z_i}^{y_i}(\mu_i, 1), \quad (6)$$

where $N_{z_i}^{y_i}(\mu_i, 1)$ denotes half normal distribution with $(-\infty, 0]$ or $[0, +\infty)$ truncated for $y_i = 1$ or $y_i = 0$ respectively and $\mu_i = E[\mathbf{X}_i^T \boldsymbol{\beta}_1(\mathbf{s}) + \beta_0] = \mathbf{X}_i^T E[\boldsymbol{\beta}_1(\mathbf{s})] + E[\beta_0]$, evaluated using the current estimates of $\boldsymbol{\beta}_1(\mathbf{s})$ and β_0. In other words, $z_i \sim N_{[0,+\infty)}(\mu_i, 1)$ if $y_i = 1$, otherwise, $z_i \sim N_{(-\infty,0]}(\mu_i, 1)$.

The approximate posterior for $\boldsymbol{\beta}_1(\mathbf{s})$ and β_0 are given by normal distributions as follows:

$$q(\boldsymbol{\beta}_1(\mathbf{s})) = N(\tilde{\boldsymbol{\beta}}_1(\mathbf{s}), \boldsymbol{\Sigma}_{\boldsymbol{\beta}_1(\mathbf{s})}), \quad (7)$$

$$q(\beta_0) = N(\tilde{\beta}_0, \tilde{\sigma}^2_{\beta_0}), \quad (8)$$

where $\tilde{\boldsymbol{\beta}}_1(\mathbf{s}) = [\mathbf{X}\mathbf{X}^T + E(\boldsymbol{C}_{\boldsymbol{\alpha}} \boldsymbol{Q} \boldsymbol{C}_{\boldsymbol{\alpha}})]^{-1} \mathbf{X}(E[\mathbf{Z}] - \mathbf{1}_{m \times 1} E[\beta_0])$, $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_1(s)} = [\mathbf{X}\mathbf{X}^T + E\{\boldsymbol{C}_{\boldsymbol{\alpha}} \boldsymbol{Q} \boldsymbol{C}_{\boldsymbol{\alpha}}\}]^{-1}$. Note that $E\{\boldsymbol{C}_{\boldsymbol{\alpha}} \boldsymbol{Q} \boldsymbol{C}_{\boldsymbol{\alpha}}\} = E_{\boldsymbol{\alpha}}[\mathbf{C}_{\boldsymbol{\alpha}} E_{\boldsymbol{\gamma}}[\mathbf{Q}] \mathbf{C}_{\boldsymbol{\alpha}}] = (E_{\boldsymbol{\gamma}}[\mathbf{Q}] \circ \boldsymbol{\Sigma}_{\mathbf{C}_{\boldsymbol{\alpha}}}) + E[\mathbf{C}_{\boldsymbol{\alpha}}] E_{\boldsymbol{\gamma}}[\mathbf{Q}] E[\mathbf{C}_{\boldsymbol{\alpha}}]$, since $\mathbf{C}_{\boldsymbol{\alpha}} = \mathbf{D}_{\boldsymbol{\alpha}}^{-1}$ is a diagonal matrix, it is easy to find $E[\mathbf{C}_{\boldsymbol{\alpha}}]$ and $\boldsymbol{\Sigma}_{\mathbf{C}_{\boldsymbol{\alpha}}} = var(\mathbf{c}_{\boldsymbol{\alpha}})$ (which is the variance matrix of the vector of the diagonals of $\mathbf{C}_{\boldsymbol{\alpha}}$). The symbol $\circ$ represents element-wise product for two matrices, i.e., $(A \circ B)_{jk} = A_{jk} B_{jk}$ for all j, k.

Derivation of $q(\boldsymbol{\beta}_1(\mathbf{s}))$:

$$\begin{aligned}\ln(q(\boldsymbol{\beta}_1(\mathbf{s})))&=E_{\boldsymbol{Z},\beta_0,\boldsymbol{\alpha},\gamma}[\ln f(\boldsymbol{Y}|\boldsymbol{Z})+\ln f(\boldsymbol{Z}|\boldsymbol{\beta}_1(\mathbf{s}),\beta_0)+\ln f(\boldsymbol{\beta}_1(\mathbf{s})|\boldsymbol{\alpha},\gamma)\\&\quad+\ln f(\boldsymbol{\alpha})+\ln f(\gamma)+\ln f(\beta_0)]+C\\&=E_{\boldsymbol{Z},\beta_0,\boldsymbol{\alpha},\gamma}[\ln f(\boldsymbol{Z}|\boldsymbol{\beta}_1(\mathbf{s}),\beta_0)+\ln f(\boldsymbol{\beta}_1(\mathbf{s})|\boldsymbol{\alpha},\gamma)]+C_1\\&=E_{\boldsymbol{Z},\beta_0,\boldsymbol{\alpha},\gamma}[-\frac{1}{2}(\boldsymbol{Z}-(\mathbf{X}^T\boldsymbol{\beta}_1(\mathbf{s})+1_{m\times1}\beta_0))^T(\boldsymbol{Z}-(\mathbf{X}^T\boldsymbol{\beta}_1(\mathbf{s})+1_{m\times1}\beta_0))\\&\quad-\frac{1}{2}\boldsymbol{\beta}_1^T(\mathbf{s})[\boldsymbol{C}_{\boldsymbol{\alpha}}\boldsymbol{Q}\boldsymbol{C}_{\boldsymbol{\alpha}}]\boldsymbol{\beta}_1(\mathbf{s})]+C_2\\&=-\frac{1}{2}E_{\boldsymbol{Z},\beta_0,\boldsymbol{\alpha},\gamma}[\boldsymbol{\beta}_1^T(\mathbf{s})[\mathbf{X}\mathbf{X}^T+\boldsymbol{C}_{\boldsymbol{\alpha}}\boldsymbol{Q}\boldsymbol{C}_{\boldsymbol{\alpha}}]\boldsymbol{\beta}_1(\mathbf{s})-2\boldsymbol{\beta}_1^T(\mathbf{s})\mathbf{X}(\boldsymbol{Z}-1_{m\times1}\beta_0)]+C_3\\&=-\frac{1}{2}\{\boldsymbol{\beta}_1^T(\mathbf{s})[\mathbf{X}\mathbf{X}^T+E(\boldsymbol{C}_{\boldsymbol{\alpha}}\boldsymbol{Q}\boldsymbol{C}_{\boldsymbol{\alpha}})]\boldsymbol{\beta}_1(\mathbf{s})-2\boldsymbol{\beta}_1^T(\mathbf{s})\mathbf{X}(E[\boldsymbol{Z}]-1_{m\times1}E[\beta_0])\}+C_3\\&=-\frac{1}{2}\{[\boldsymbol{\beta}_1(\mathbf{s})-\tilde{\boldsymbol{\beta}}_1(\mathbf{s})]^T\boldsymbol{\Sigma}_{\boldsymbol{\beta}_1(\mathbf{s})}[\boldsymbol{\beta}_1(\mathbf{s})-\tilde{\boldsymbol{\beta}}_1(\mathbf{s})]\}+C_4\end{aligned}$$

For the approximate posterior $q(\beta_0)$, we have

$$\tilde{\beta}_0=E\{(m+\frac{1}{c_3})^{-1}(\mathbf{1}_{1\times m}\boldsymbol{Z}-\mathbf{1}_{1\times m}\mathbf{X}^T\boldsymbol{\beta}_1(\mathbf{s}))\}=(m+\frac{1}{c_3})^{-1}(\mathbf{1}_{1\times m}E[\boldsymbol{Z}]-\mathbf{1}_{1\times m}\mathbf{X}^TE[\boldsymbol{\beta}_1(\mathbf{s})])$$

and $\tilde{\sigma}^2_{\beta_0}=(m+\frac{1}{c_3})^{-1}$. The expectation is evaluated using the current estimates for the unknown parameters.

For the jth component α_j of the binary vector $\boldsymbol{\alpha}$, we have

$$q(\alpha_j)\propto|c_{\alpha_j}|\exp\{-\frac{1}{2}\mathrm{tr}E(\boldsymbol{\beta}_1(\mathbf{s})\boldsymbol{\beta}_1^T(\mathbf{s}))E_{\boldsymbol{\alpha}_{(j)}}[\boldsymbol{c}_{\boldsymbol{\alpha}}\boldsymbol{c}_{\boldsymbol{\alpha}}^T]\circ E_\gamma\boldsymbol{Q}]\},\quad(9)$$

where $E_{\boldsymbol{\alpha}_{(j)}}$ stands for taking expectation over all the components of $\boldsymbol{\alpha}$ except α_j. Then $q(\alpha_j)$ is the probability function of $Bin(1, q_j)$ where $q_j = q(\alpha_j = 1)/[q(\alpha_j = 1) + q(\alpha_j = 0)]$.

Finally the update the parameter γ in $\boldsymbol{Q} = \boldsymbol{H}(\gamma)^{-1}$ can be obtained by

$$q(\gamma)\propto|\boldsymbol{Q}|^{1/2}\exp\{-\frac{1}{2}\mathrm{tr}E(\boldsymbol{\beta}_1(\mathbf{s})\boldsymbol{\beta}_1^T(\mathbf{s}))E_{\boldsymbol{\alpha}}[\boldsymbol{C}_{\boldsymbol{\alpha}}\boldsymbol{Q}\boldsymbol{C}_{\boldsymbol{\alpha}}]\}f(\gamma).\quad(10)$$

In the above update formulas (i.e. Eq. 6 – 10) only the expectation $E_\gamma\boldsymbol{Q}$ is needed. With a Matérn covariance structure, where the positive decay parameter γ is unknown, and $f(\gamma)$ is the positive truncated $N(a_2, b_2)$, the $E_\gamma\mathbf{Q}$ can be approximated by importance sampling. For example, we can generate γ^k, $k = 1, \ldots, K$ i.i.d. from $N(a_2, b_2)$, then

$$E_{\gamma}\boldsymbol{Q} \approx \frac{\sum_{k=1}^{K}\boldsymbol{H}(\gamma^k)^{-1}|\boldsymbol{H}(\gamma^k)^{-1}|^{1/2}\exp\{-\frac{1}{2}\mathrm{tr}E(\boldsymbol{\beta}_1(\mathbf{s})\boldsymbol{\beta}_1^T(\mathbf{s}))E_{\boldsymbol{\alpha}}[\boldsymbol{C}_{\boldsymbol{\alpha}}\boldsymbol{H}(\gamma^k)^{-1}\boldsymbol{C}_{\boldsymbol{\alpha}}]\}I(\gamma^k>0)}{\sum_{k=1}^{K}|\boldsymbol{H}(\gamma^k)^{-1}|^{1/2}\exp\{-\frac{1}{2}\mathrm{tr}E(\boldsymbol{\beta}_1(\mathbf{s})\boldsymbol{\beta}_1^T(\mathbf{s}))E_{\boldsymbol{\alpha}}[\boldsymbol{C}_{\boldsymbol{\alpha}}\boldsymbol{H}(\gamma^k)^{-1}\boldsymbol{C}_{\boldsymbol{\alpha}}]\}I(\gamma^k>0)}. \quad (11)$$

The embedded importance sampling in Eq. (11) tends to significantly increase the computing burden in the VB approximation. Thus we consider a further simplified approximation in estimation of $E_{\boldsymbol{\gamma}}\boldsymbol{Q}$. Note that $E_{\boldsymbol{\gamma}}\boldsymbol{Q} = E[\mathbf{H}^{-1}(\boldsymbol{\gamma})] \approx [\mathbf{H}(E[\boldsymbol{\gamma}])]^{-1}$. Then we only need to compute and save $E\boldsymbol{\gamma}$, which is one number, rather than the whole $n \times n$ matrix $E_{\boldsymbol{\gamma}}\boldsymbol{Q}$. The same importance sampling idea as in Eq. (11) can be used to compute $E\boldsymbol{\gamma}$:

$$E(\gamma) \approx \frac{\sum_{k=1}^{K}\gamma^k|\boldsymbol{H}(\gamma^k)^{-1}|^{1/2}\exp\{-\frac{1}{2}\mathrm{tr}E(\boldsymbol{\beta}_1(\mathbf{s})\boldsymbol{\beta}_1^T(\mathbf{s}))E_{\boldsymbol{\alpha}}[\boldsymbol{C}_{\boldsymbol{\alpha}}\boldsymbol{H}(\gamma^k)^{-1}\boldsymbol{C}_{\boldsymbol{\alpha}}]\}I(\gamma^k>0)}{\sum_{k=1}^{K}|\boldsymbol{H}(\gamma^k)^{-1}|^{1/2}\exp\{-\frac{1}{2}\mathrm{tr}E(\boldsymbol{\beta}_1(\mathbf{s})\boldsymbol{\beta}_1^T(\mathbf{s}))E_{\boldsymbol{\alpha}}[\boldsymbol{C}_{\boldsymbol{\alpha}}\boldsymbol{H}(\gamma^k)^{-1}\boldsymbol{C}_{\boldsymbol{\alpha}}]\}I(\gamma^k>0)}. \quad (12)$$

In practice, we found using the simplified approach in Eq. (12) can improve the computing efficiency by a factor of 2–3, while the resulting estimates appear to be very consistent to those from Eq. (11) (results not shown here). For this reason, in our paper, we only present the results obtained based on Eq. (12).

In the numerical studies of this paper, we have used a special class of Matérn correlation function, which coincides with an exponential correlation with decay parameter $\boldsymbol{\gamma}$. Specifically, the ijth element in the correlation matrix, $H_{ij}(\boldsymbol{\gamma})$, represents the correlation between the ith and jth voxels, which is a function of the Euclidean distance d between the ith and jth voxels : $H_{ij}(\boldsymbol{\gamma}) = exp(-d/\boldsymbol{\gamma})$. We choose this structure as an example to illustrate our method, since it is a simplest example that contains an unknown parameter that characterizes the strength of spatial correlations. Allowing more general Matérn correlations would introduce an extra smoothness parameter, which would further increase computational difficulty.

Although the purpose of our paper is on studying the impact of spatial correlations, and not on providing a fast algorithm, there are some possibilities to further speed up the computing. One is to use fixed $\boldsymbol{\gamma}$ in the Matérn correlation function to avoid computing-intensive importance sampling step for $E_{\boldsymbol{\gamma}}(\mathbf{Q})$ in Eq. (9–12). A suitable value for $\boldsymbol{\gamma}$ could be obtained by examining the empirical semivariogram (in this case, the prior for, i.e. $\pi(\boldsymbol{\gamma})$ in the likelihood Eq. (5) should be dropped). Another possibility is to choose some prior for $\boldsymbol{\beta}$ such that the spatial information can be captured while achieving a closed-form posterior when using some suitable hyper-prior. For example, if $\boldsymbol{Q}$ is chosen to be the Gaussian Markov random field prior with an unknown scaling constant $\boldsymbol{\gamma}$, it is possible to use a conjugate

Gamma hyperprior $f(\gamma)$ and derive $q(\gamma)$ analytically. It may also be possible to modify the VB algorithm by updating the regression coefficients for each voxel separately, in order to achieve a linear algorithm for handling more vowels. Such options are under further investigation.

References

1. Haynes JD, Rees G. Decoding mental states from brain activity in humans. Nat. Rev. Neurosci. 2006; 7:523–534. [PubMed: 16791142]
2. Kay KN, Naselaris T, Prenger RJ, Gallant JL. Identifying natural images from human brain activity. Nature. 2008; 452:352–355. [PubMed: 18322462]
3. Mitchell TM, Shinkareva SV, Carlson A, et al. Predicting human brain activity associated with the meanings of nouns. Science. 2008; 320:1191–1195. [PubMed: 18511683]
4. Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, Gallant JL. Reconstructing visual experiences from brain activity evoked by natural movies. Curr. Biol. 2011; 21:1641–1646. [PubMed: 21945275]
5. Tong F, Pratte MS. Decoding patterns of human brain activity. Annu. Rev. Psychol. 2012; 63:483–509. [PubMed: 21943172]
6. Weil RS, Rees G. Decoding the neural correlates of consciousness. Curr. Opin. Neurol. 2010; 23:649–655. [PubMed: 20881487]
7. Degras D, Lindquist MA. A hierarchical model for simultaneous detection and estimation in multi-subject fMRI studies. Neuroimage. 2014; 98:61–72. [PubMed: 24793829]
8. Pereira F, Botvinick M. Information mapping with pattern classifiers: a comparative study. Neuroimage. 2011; 56:476–496. [PubMed: 20488249]
9. Lee YS, Turkeltaub P, Granger R, Raizada RD. Categorical speech processing in Broca's area: an fMRI study using multivariate pattern-based analysis. J. Neurosci. 2012; 32:3942–3948. [PubMed: 22423114]
10. Binder JR, Frost JA, Hammeke TA, et al. Human temporal lobe activation by speech and nonspeech sounds. Cereb. Cortex. 2000; 10:512–528. [PubMed: 10847601]
11. Leaver AM, Rauschecker JP. Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. J. Neurosci. 2010; 30:7604–7612. [PubMed: 20519535]
12. Rauschecker JP, Scott SK. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat. Neurosci. 2009; 12:718–724. [PubMed: 19471271]
13. Staeren N, Renvall H, De Martino F, Goebel R, Formisano E. Sound categories are represented as distributed patterns in the human auditory cortex. Curr. Biol. 2009; 19:498–502. [PubMed: 19268594]
14. Zou H, Hastie T. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B. 2005; 67:301–320.
15. De Martino F, Valente G, Staeren N, Ashburner J, Goebel R, Formisano E. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. Neuroimage. 2008; 43:44–58. [PubMed: 18672070]
16. Guyon I, Weston J, Barnihill S, Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning. Mach. Learn. 2002; 46:389–422.
17. Kriegeskorte N, Goebel R, Bandettini P. Information-based functional brain mapping. Proc. Natl. Acad. Sci. U.S.A. 2006; 103:3863–3868. [PubMed: 16537458]
18. Shinkareva SV, Malave VL, Mason RA, Mitchell TM, Just MA. Commonality of neural representations of words and pictures. Neuroimage. 2011; 54:2418–2425. [PubMed: 20974270]
19. Sabuncu MR, Van Leemput K. Initiative Alzheimer's Disease Neuroimaging. The relevance voxel machine (RVoxM): a self-tuning Bayesian model for informative image-based prediction. IEEE Trans Med Imaging. 2012; 31:2290–2306. [PubMed: 23008245]

20. Michel V, Gramfort A, Varoquaux G, Eger E, Thirion B. Total variation regularization for fMRI-based prediction of behavior. IEEE Trans Med Imaging. 2011; 30:1328–1340. [PubMed: 21317080]
21. Baldassarre L, Mourao-Miranda J, Pontil M. Structured Sparsity Models for Brain Decoding from fMRI Data. Pattern Recognition in NeuroImaging (PRNI), 2012 International Workshop on. 2012:5–8.
22. Gramfort, A.; Thirion, B.; Varoquaux, G. Proceedings of the 2013 International Workshop on Pattern Recognition in Neuroimaging PRNI '13. Washington, DC, USA: IEEE Computer Society; 2013. Identifying Predictive Regions from fMRI with TV-L1 Prior; p. 17-20.
23. Grosenick L, Klingenberg B, Katovich K, Knutson B, Taylor JE. Interpretable whole-brain prediction analysis with GraphNet. Neuroimage. 2013; 72:304–321. [PubMed: 23298747]
24. Rondina J, Hahn T, Oliveira L, et al. Scors: a method based on stability for feature selection and mapping in neuroimaging. IEEE Trans Med Imaging. 2014; 33:85–98. [PubMed: 24043373]
25. Rasmussen PM, Hansen LK, Madsen KH, Churchill NW, Strother SC. Model sparsity and brain pattern interpretation of classification models in neuroimaging. Pattern Recognition. 2012; 45:2085–2100.
26. Ganz M, Greve DN, Fischl B, Konukoglu E. Alzheimer's Disease Neuroimaging Initiative. Relevant feature set estimation with a knock-out strategy and random forests. Neuroimage. 2015; 122:131–148. [PubMed: 26272728]
27. Sotiras A, Resnick SM, Davatzikos C. Finding imaging patterns of structural covariance via Non-Negative Matrix Factorization. Neuroimage. 2015; 108:1–16. [PubMed: 25497684]
28. Smith M, Fahrmeir L. Spatial Bayesian Variable Selection With Application to Functional Magnetic Resonance Imaging. J. Am. Stat. Assoc. 2007; 102:417–431.
29. Friston K, Ashburner J, Frith C, Poline J, Heather J, Frackowiak R. Spatial registration and normalization of images. Human Brain Mapping. 1995; 2:165–189.
30. Bai P, Shen H, Huang JZ, Truong YK. Adaptive statistical parametric mapping for fMRI. Statistics and Its Interface. 2010; 3:33–43.
31. Formisano E, De Martino F, Bonte M, Goebel R. Who Is Saying what Brain-Based Decoding of Human Voice and Speech. Science. 2008; 7:970–973. [PubMed: 18988858]
32. Albert JH, Chib S. Bayesian Analysis of Binary and Polychotomous Response Data. J. Am. Stat. Assoc. 1993; 88:669–679.
33. George EI, McCulloch RE. variable selection visa gibbs sampling. J. Am. Stat. Assoc. 1993; 88:881–889.
34. Formisano E, De Martino F, Valente G. Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning. Magn. Reson. Imaging. 2008; 26:921–934. [PubMed: 18508219]
35. Durrant S, Hardoon DR, Brechmann A, Shawe-Taylor J, Miranda ER, Scheich H. GLM and SVM analyses of neural response to tonal and atonal stimuli: new techniques and a comparison. Conn. Sci. 2009; 21:161–175.
36. Zhang C, Fan J, Yu T. Multiple testing via FDRL for large-scale imaging data. Ann. Stat. 2011; 39:613–642. [PubMed: 21643445]
37. Zhu Z, Feng G, Zhang JX, Li G, Li H, Wang S. The Role of the Left Prefrontal Cortex in Sentence-level Semantic Integration. NeuroImage. 2013; 76:325–331. [PubMed: 23507386]
38. Cox RW. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 1996; 29:162–173. [PubMed: 8812068]
39. Cressie, N. Statistics for spatial data. New York: Wiley; 1993.
40. Finley AO, Banerjee S, Carlin BP. spBayes: an R package for univariate and multivariate hierarchical point-referenced spatial models. J. Stat. Softw. 2007; 19:1–24. [PubMed: 21494410]
41. Pebesma EJ. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 2004; 30:683–691.
42. Celeux G, Martin-Magniette ML, Maugis-Rabusseau C, Raftery AE. Comparing Model Selection and Regularization Approaches to Variable Selection in Model-Based Clustering. ArXiv e-prints. 2013

43. Fan J, Feng Y, Tong X. A road to classification in high dimensional space: the regularized optimal affine discriminant. J. R. Stat. Soc. Ser. B. 2012; 74:745–771.

44. Ley A, Vroomen J, Hausfeld L, Valente G, De Weerd P, Formisano E. Learning of new sound categories shapes neural response patterns in human auditory cortex. J. Neurosci. 2012; 32:13273–13280. [PubMed: 22993443]

45. Brodersen KH, Daunizeau J, Mathys C, Chumbley JR, Buhmann JM, Stephan KE. Variational Bayesian mixed-effects inference for classification studies. Neuroimage. 2013; 76:345–361. [PubMed: 23507390]

46. Girolami M, Rogers S. Variational Bayesian Multinomial Probit Regression with Gaussian Process Priors. Neural Comput. 2006; 18:1790–1817.

47. Grimmer J. An Introduction to Bayesian Inference Via Variational Approximations. Polit. Anal. 2011; 19:32–47.

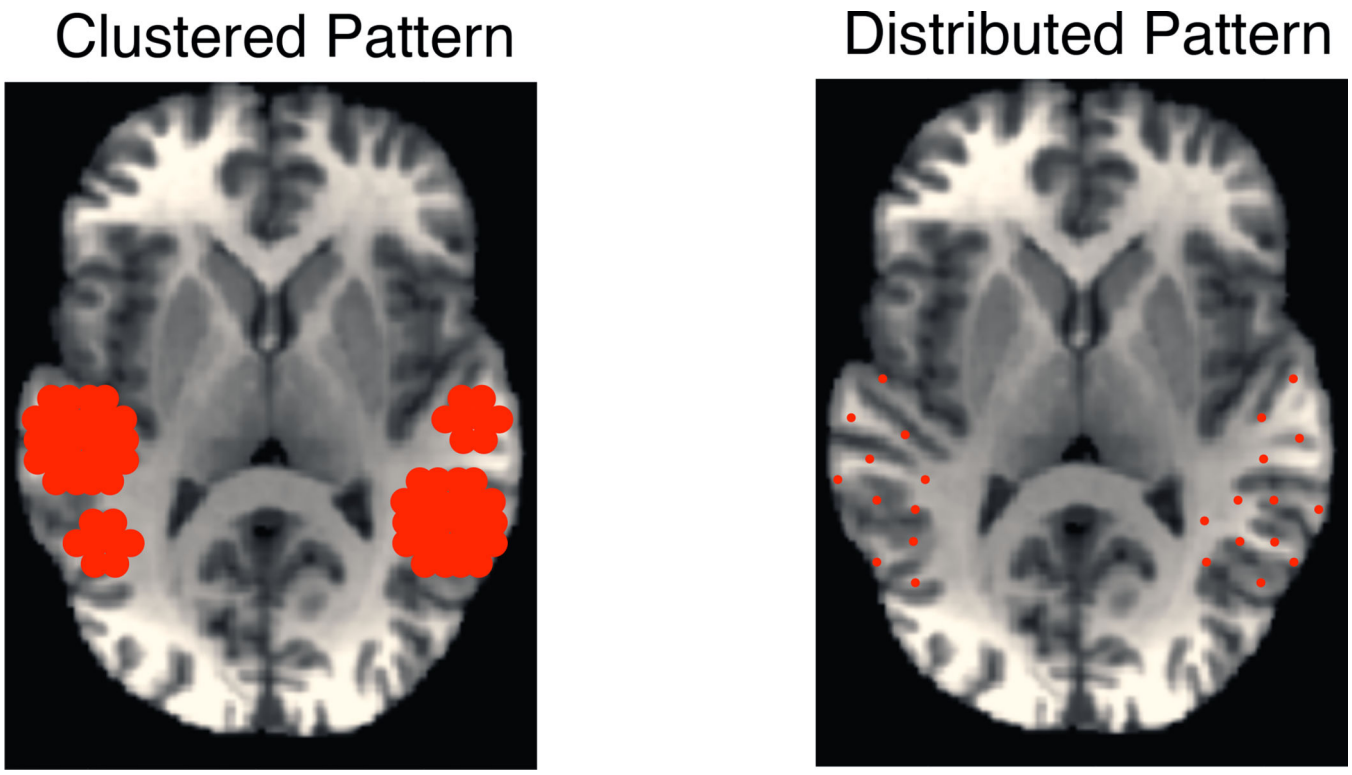

Figure 1.
A simplified diagram illustrating distinct patterns of selected voxels by the univariate approach (left) and the multivariate pattern analysis approach (right).

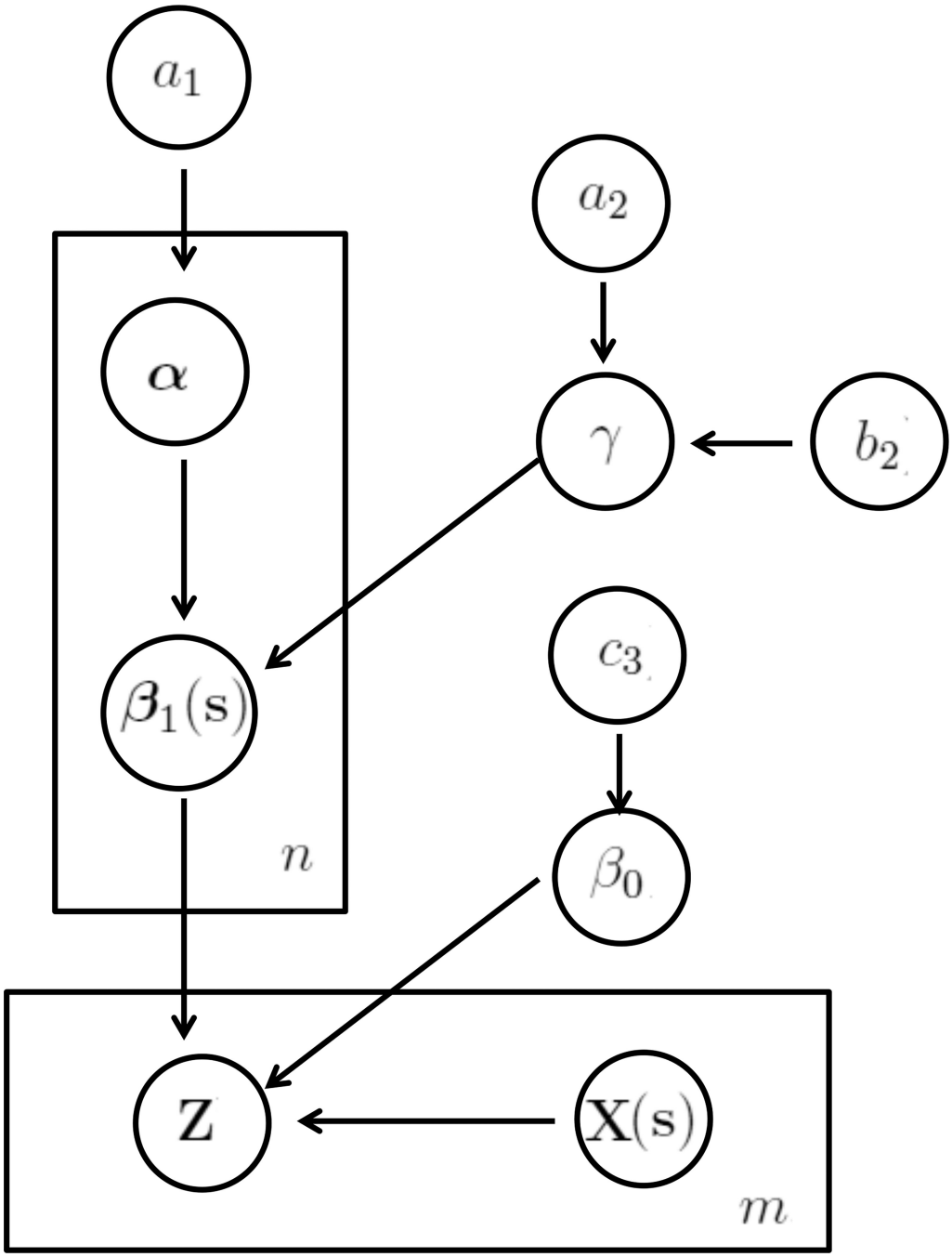

Figure 2.
Graphical representation of our proposed model with m observations and n predictors.

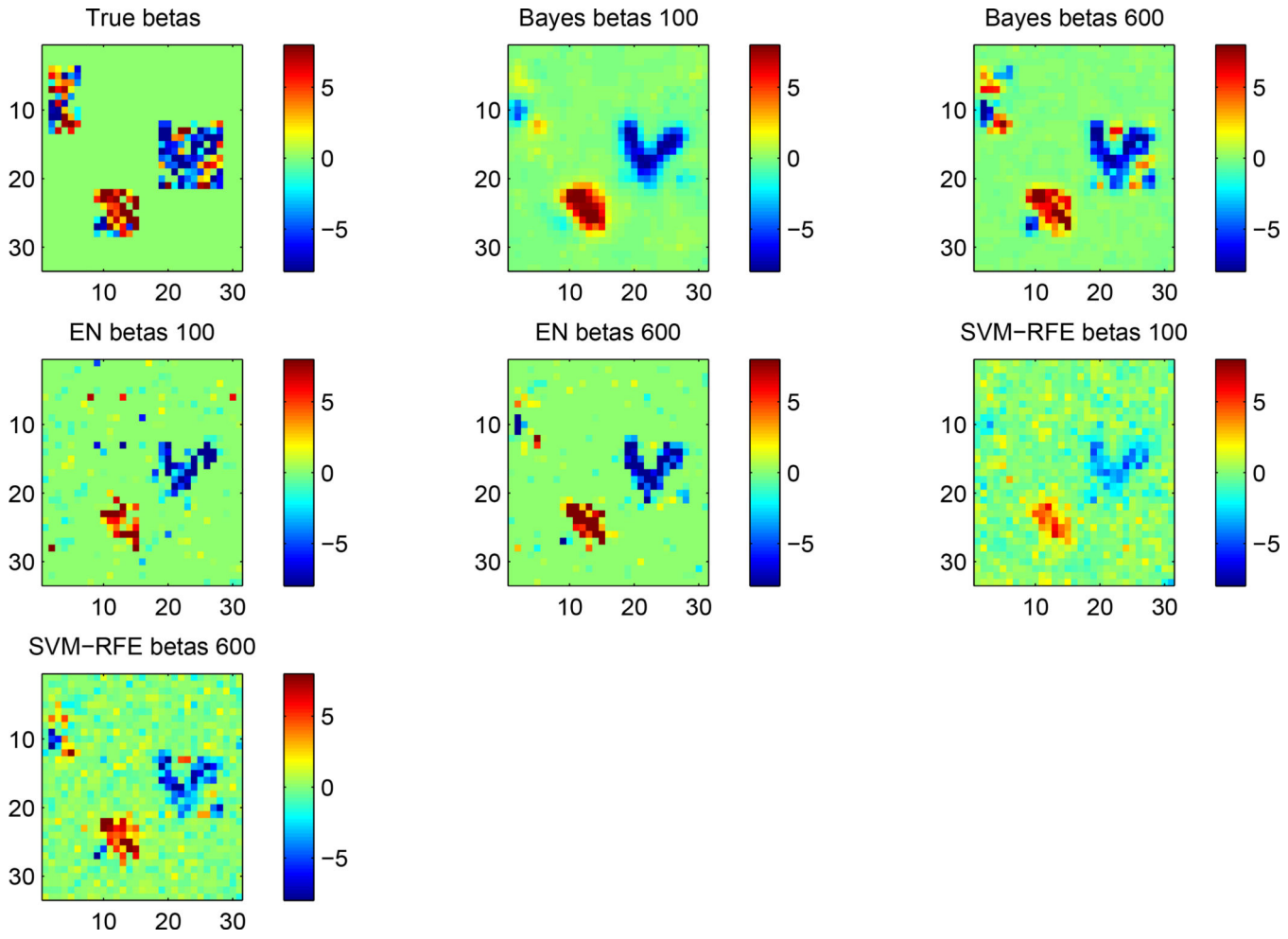

Figure 3.

Comparison of coefficient estimation across models in Simulation I. The averages of normalized coefficients estimated by BPSV, EN, and SVM-RFE with sample sizes 100 and 600 are represented in colors at each location according to the color bar.

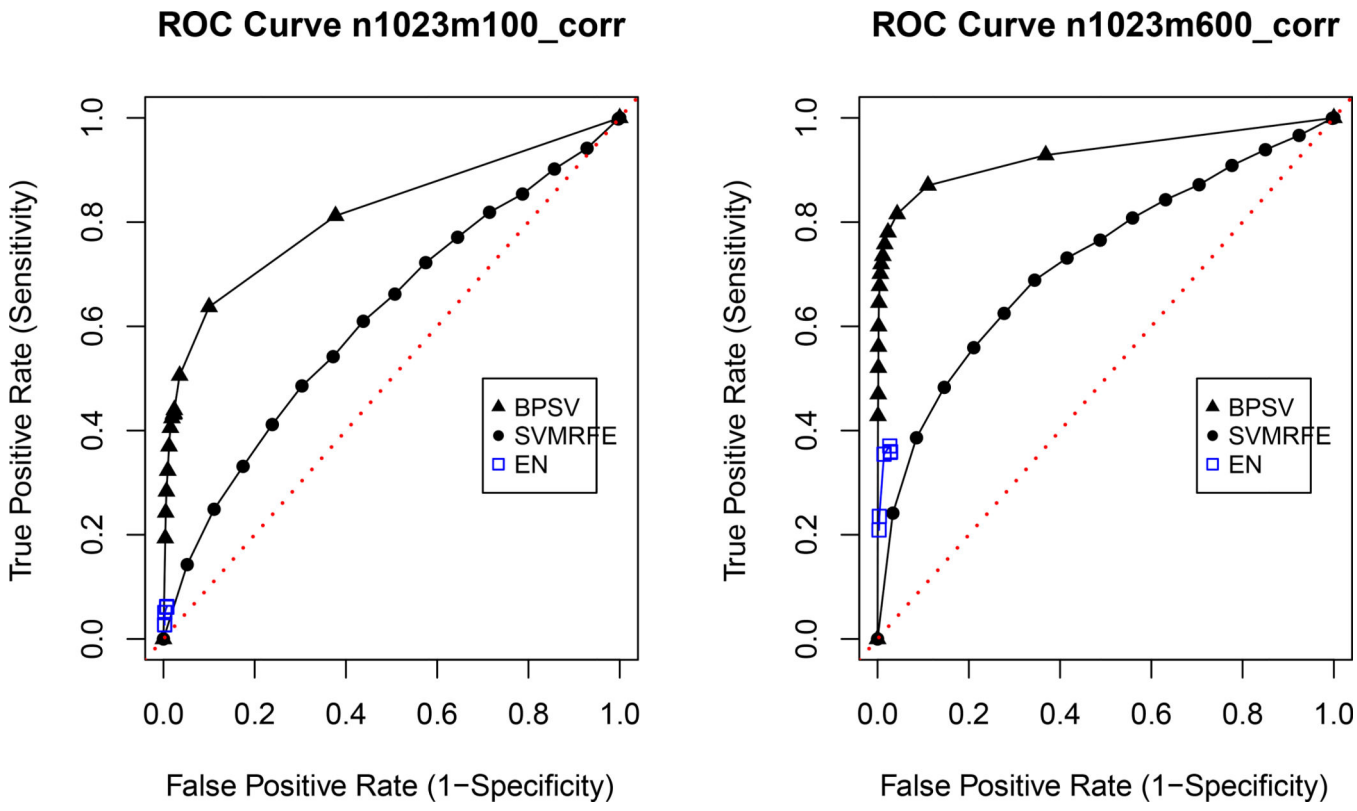

Figure 4.
ROC curves of BPSV, SVM-RFE and EN on voxel selection with different tuning parameter values in Simulation I.

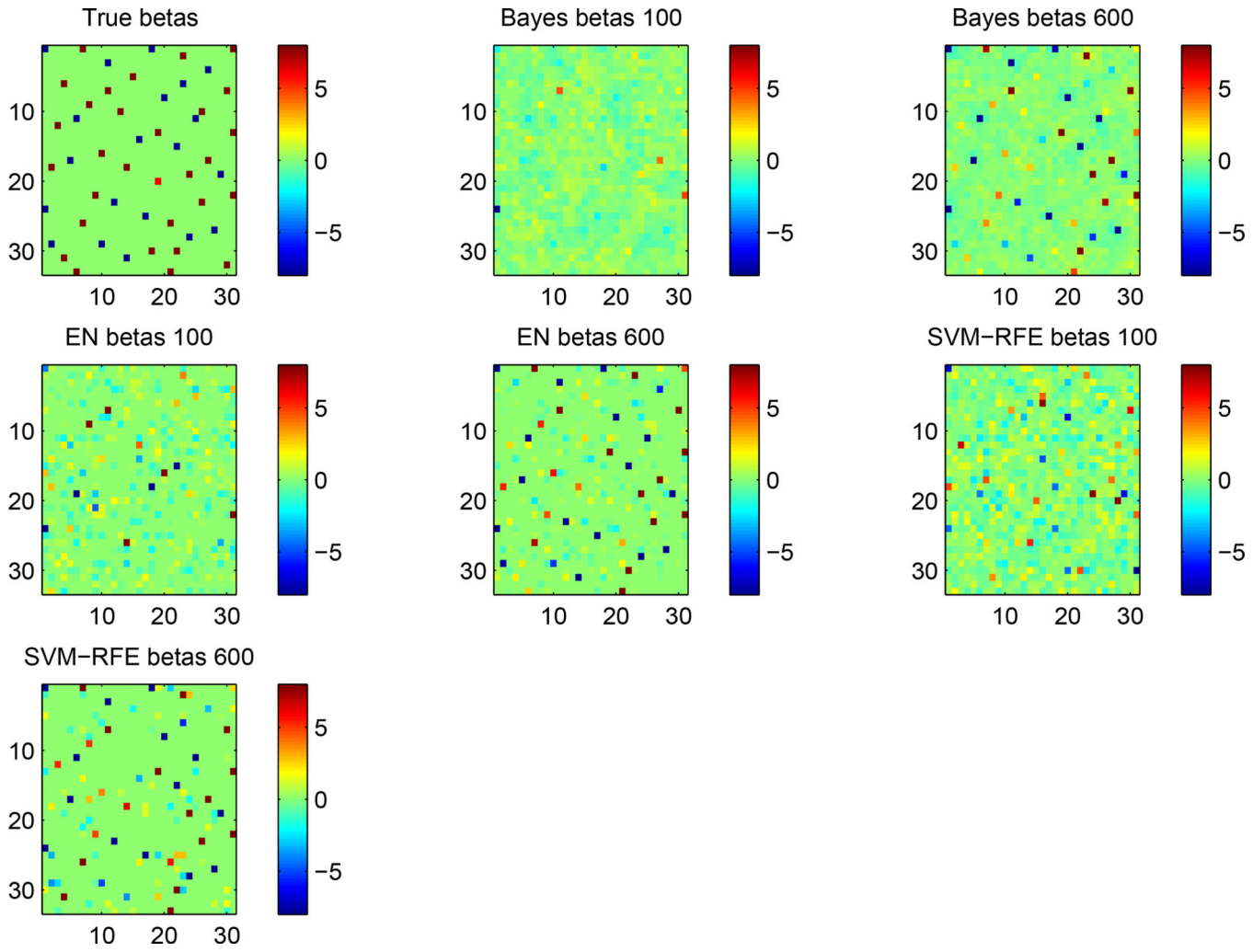

Figure 5.

Comparison of coefficient estimation across models in Simulation II. The averages of normalized coefficients estimated by BPSV, EN, and SVM-RFE with sample sizes 100 and 600 are shown in colors at each location according to the color bar.

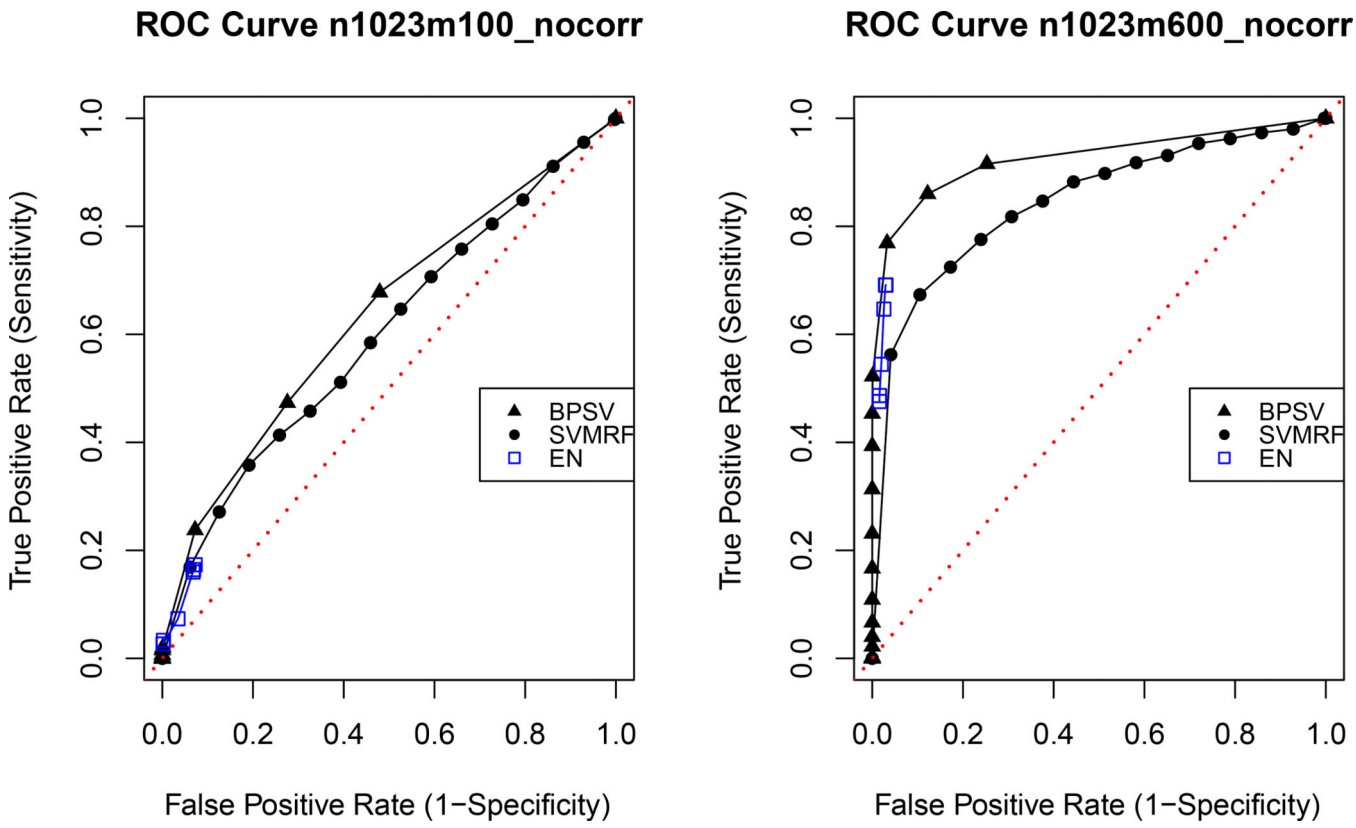

Figure 6.
ROC curves of BPSV, SVM-RFE and EN on voxel selection with different tuning parameter values in Simulation II.

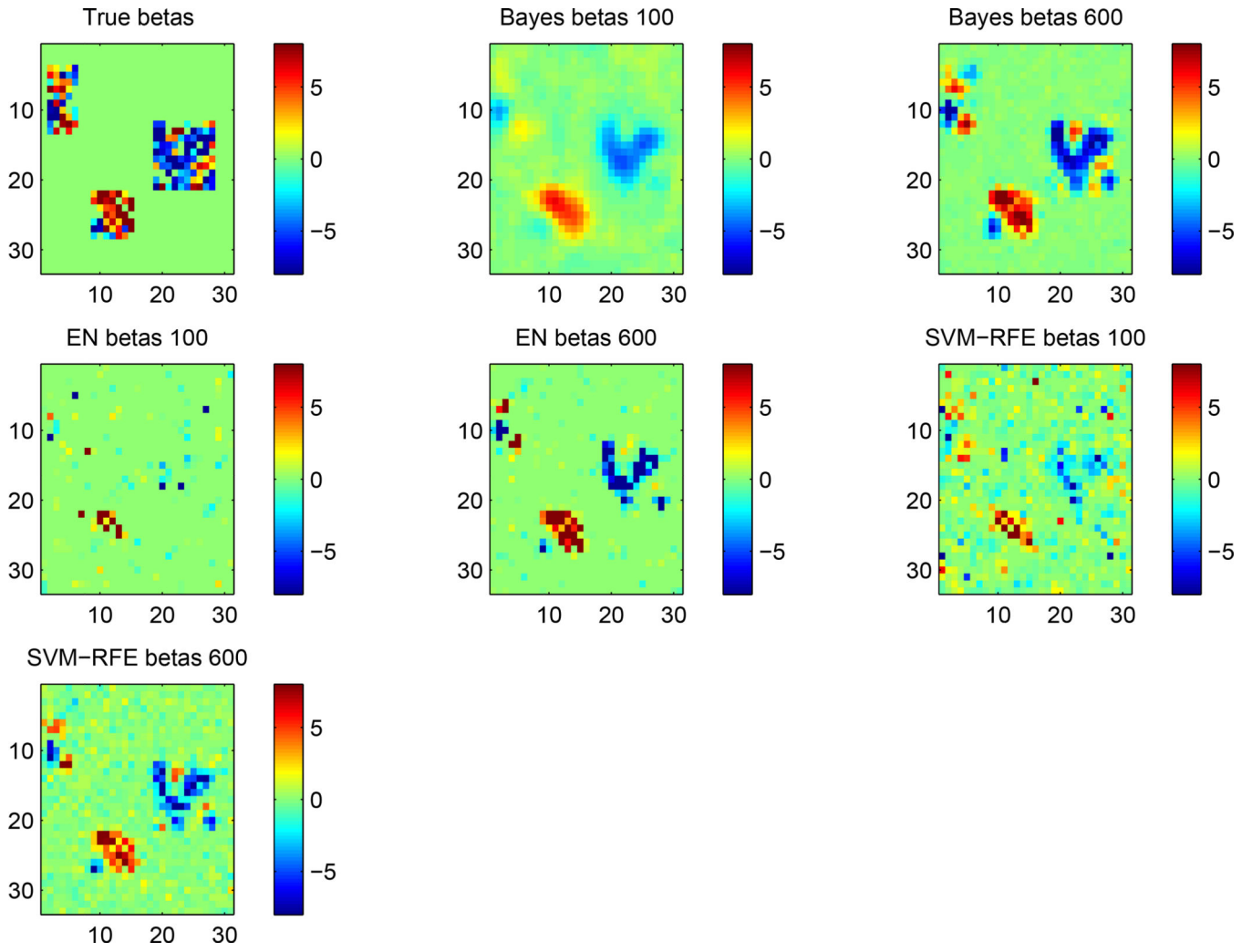

Figure 7.

Comparison of coefficient estimation across models in Simulation III. The averages of normalized coefficients estimated by BPSV, EN, and SVM-RFE with sample sizes 100 and 600 are shown in colors at each location according to the color bar.

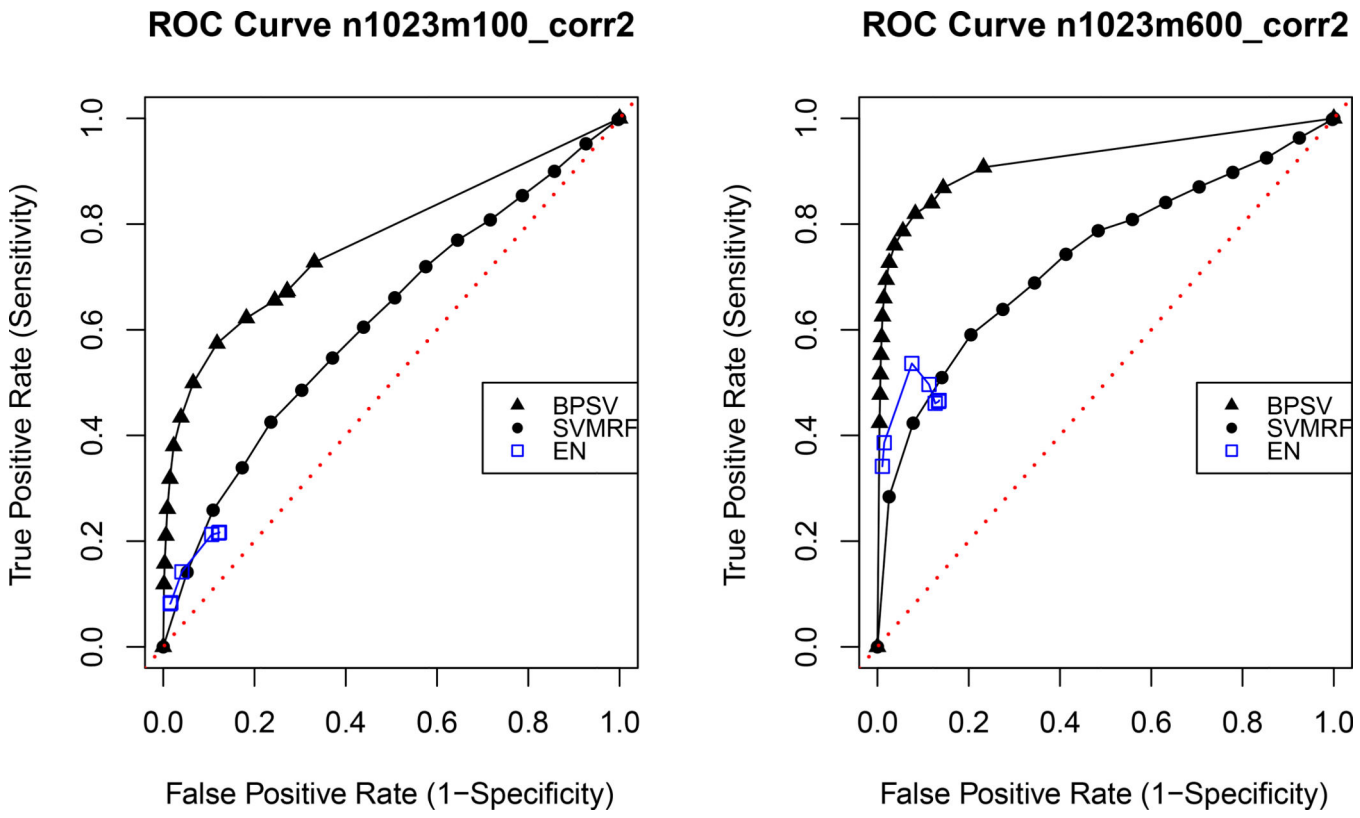

Figure 8.
ROC curves of BPSV, SVM-RFE and EN on voxel selection with different tuning parameter values in Simulation III.

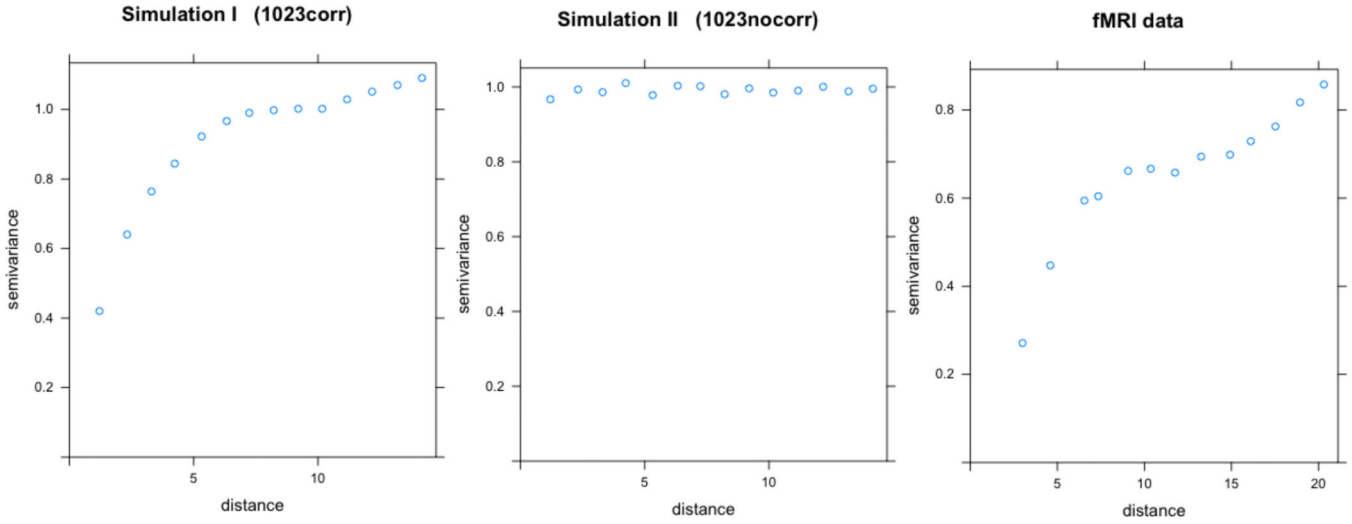

Figure 9.
Empirical semivariograms for synthetic data from the first two simulations and real fMRI data.

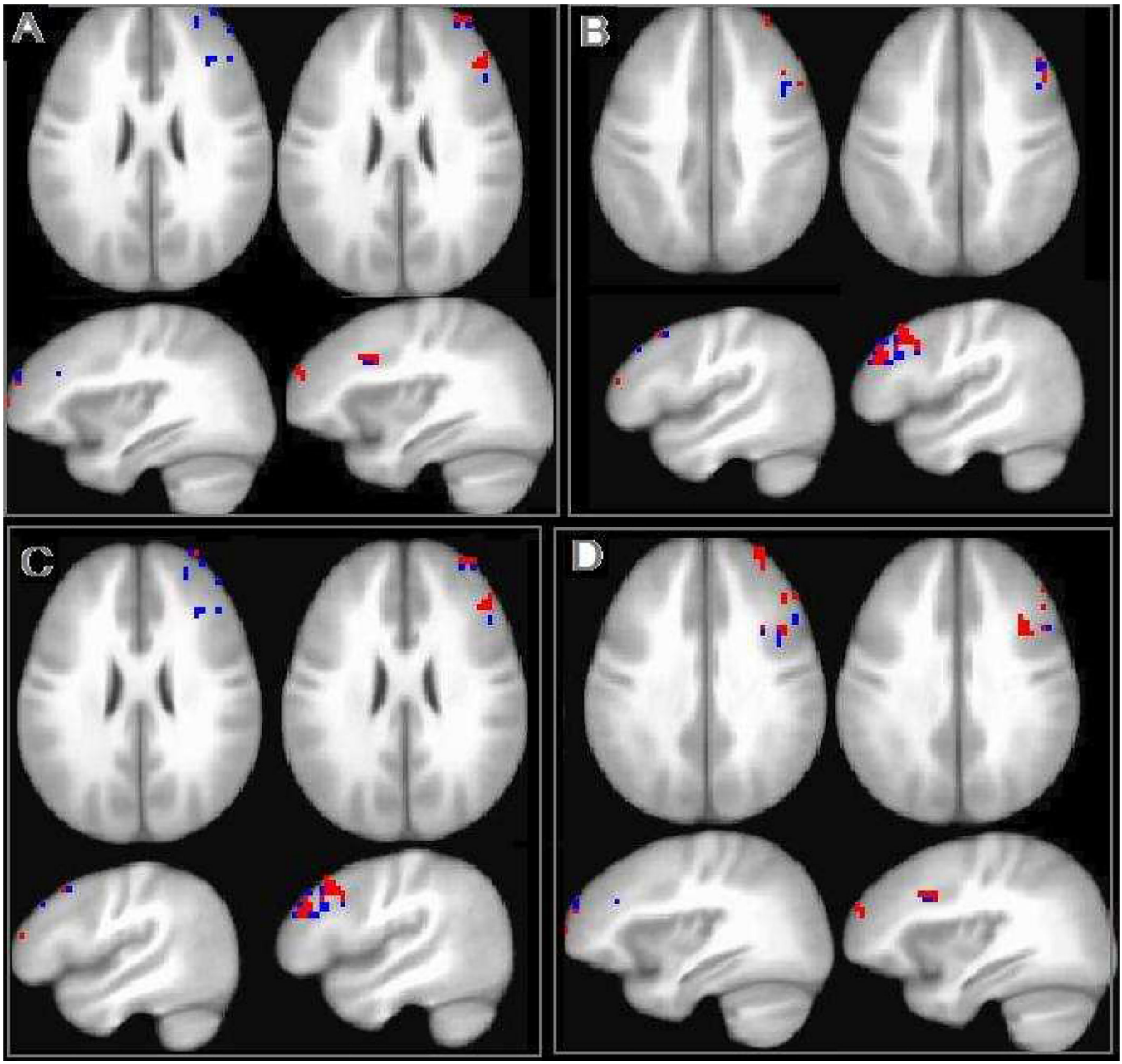

Figure 10.

The voxels selected by SVM-RFE (left column) and BPSV (right column) from all three subjects. Parts A, B, C and D show the Axial and Sagittal images of the brain with selected voxels at the centers of the four largest clusters identified by BPSV respectively. Voxels with weights larger than 0 and less than 0.5 are shown in dark blue while voxels with weights between 0.5 and 1 are plotted in red.

Table 1

Model comparison in Simulation I.

	correlation coefficient (classification accuracy) m=100	correlation coefficient (classification accuracy) m= 600
EN	0.19 (0.76)	0.60 (0.83)
SVM-RFE	0.31 (0.78)	0.57 (0.79)
BPSV	0.64 (0.88)	0.90 (0.93)

NOTE: The presented is the Pearson correlation between true regression coefficients and the fitted regression coefficients (EN, BPSV) or weights (SVM-REF) averaged over 9 replications. The number inside the parentheses is the averaged mean classification accuracy obtained through cross-validation.

Table 2

Model comparison in Simulation II.

	correlation coefficient (classification accuracy) m=100	correlation coefficient (classification accuracy) m= 600
EN	0.20 (0.50)	0.84 (0.82)
SVM-RFE	0.21 (0.54)	0.81 (0.80)
BPSV	0.28 (0.56)	0.86 (0.83)

NOTE: Same as in Table 1.

Table 3

Model comparison in Simulation III.

	correlation coefficient (classification accuracy) m=100	correlation coefficient (classification accuracy) m= 600
EN	0.16 (0.60)	0.63 (0.82)
SVM-RFE	0.26 (0.66)	0.59 (0.81)
BPSV	0.51 (0.76)	0.83 (0.92)

NOTE: same as in Table 1.

Table 4

Classification accuracy comparison in fMRI data analysis.

	BPSV	SVM-RFE	EN
Subject 1	0.65 (0.04)	0.63 (0.21)	0.69 (0.14)
Subject 2	0.71 (0.05)	0.69 (0.17)	0.69 (0.13)
Subject 3	0.63 (0.14)	0.63 (0.16)	0.60 (0.13)

NOTE: Numbers outside and inside the parentheses are means and standard deviations of classification accuracies from the four-fold cross-validation.

Table 5

Clusteredness indexes of the selected voxels in fMRI data analysis.

	Proportion of single voxels	Proportion of voxels in the largest cluster
Subject 1	0.06 (0.15)	0.37 (0.19)
Subject 2	0.06 (0.29)	0.27 (0.11)
Subject 3	0.05 (0.23)	0.29 (0.24)

NOTE: The presented are the averaged results from the four-fold cross-validation. Numbers outside and inside the parentheses are results from BPSV and SVM-RFE respectively.