The Annals of Applied Statistics
2015, Vol. 9, No. 2, 687–713
DOI: 10.1214/15-AOAS818

SPATIAL BAYESIAN VARIABLE SELECTION AND GROUPING FOR HIGH-DIMENSIONAL SCALAR-ON-IMAGE REGRESSION

BY FAN LI[*,1,4], TINGTING ZHANG[†,2,4], QUANLI WANG[*], MARLEN Z. GONZALEZ[†], ERIN L. MARESH[†] AND JAMES A. COAN[3,†]

Duke University and University of Virginia†*

Multi-subject functional magnetic resonance imaging (fMRI) data has been increasingly used to study the population-wide relationship between human brain activity and individual biological or behavioral traits. A common method is to regress the scalar individual response on imaging predictors, known as a scalar-on-image (SI) regression. Analysis and computation of such massive and noisy data with complex spatio-temporal correlation structure is challenging. In this article, motivated by a psychological study on human affective feelings using fMRI, we propose a joint Ising and Dirichlet Process (Ising-DP) prior within the framework of Bayesian stochastic search variable selection for selecting brain voxels in high-dimensional SI regressions. The Ising component of the prior makes use of the spatial information between voxels, and the DP component groups the coefficients of the large number of voxels to a small set of values and thus greatly reduces the posterior computational burden. To address the phase transition phenomenon of the Ising prior, we propose a new analytic approach to derive bounds for the hyperparameters, illustrated on 2- and 3-dimensional lattices. The proposed method is compared with several alternative methods via simulations, and is applied to the fMRI data collected from the KLIFF hand-holding experiment.

1. Introduction. Positive social contact is known to enhance human health and well-being, possibly because it helps to regulate humans' emotional reactivity when facing negative stressors in daily life [Coan, Schaefer and Davidson (2006); Coan, Beckes and Allen (2013); Coan (2010, 2011)]. Conventional studies on social contact primarily focus on its aggregated effect on an entire population. With the common belief that human behavior is controlled by individual mental decisions, which is affected by the immediate environment, it is desirable to investigate emotion regulation activity of the individual brain under different social interaction conditions. Toward this aim, the KLIFF hand-holding psychological experiment [Coan, Schaefer and Davidson (2006)] was conducted. In this experiment,

Received October 2014; revised February 2015.
[1]Supported in part by the U.S. NSF-DMS Grant 1208983.
[2]Supported in part by the U.S. NSF-DMS Grants 1209118 and 1120756.
[3]Supported in part by the National Institute of Mental Health (NIMH) Grant R01MH080725.
[4]Equally contributing authors.
Key words and phrases. Bayesian, Dirichlet Process, fMRI, Ising model, phase transition, scalar-on-image regression, stochastic search, variable selection.

104 pairs—each pair consisting of a male and a female—of mentally and physically healthy young adults in various close relationships including friends and married couples were recruited from a larger representative longitudinal community sample [Allen et al. (2007)]. One participant of each pair was threatened with mild electric shock during a functional magnetic resonance imaging (fMRI) session while either holding a hand of a friend, holding a hand of a stranger or holding no hand at all, in three separate sessions, which represent three different types of social interactions—positive and supportive social interaction with friends, general social interaction with strangers and no social interaction, respectively. At the end of each session, the subjects were asked to rate their feelings of arousal and valence [Russell (1980); Lang et al. (1993)] experienced during the experiment. Arousal and valence are the two dimensions in the framework of emotion fields, representing the extent of excitement and pleasure experienced, respectively [see Bradley and Lang (1994) for more detailed explanation].

To investigate which areas in the brain are predictive of individual's affective feelings in the KLIFF study, we can construct a regression model using subjects' emotion (arousal and valence) measurements as the response, and summaries of the fMRI images in the regions of interests (ROIs) as predictors. This type of regression is often referred to as scalar-on-image (SI) regressions in the literature [Reiss et al. (2011); Huang et al. (2013); Goldsmith, Huang and Crainiceanu (2014)]. SI regressions with predictors from other imaging modalities, such as diffusion tensor imaging (DTI), have also been used in medical and scientific studies [e.g., Reiss et al. (2015)].

The SI regression model in the KLIFF study has several unique characteristics due to the features of fMRI data. First, the sample size is much smaller than the number of predictors, that is, the number of brain voxels (3D cubic volumes in the brain) in the ROIs, which is over 6000 in the KLIFF study. This is known as the "large p, small n" paradigm [West (2003)]. Second, there is rich spatial information between the predictors. Third, neighboring predictors are highly correlated and often have similar but weak effects on the response. Finally, as each voxel accounts for only a tiny area in the brain, it is very likely that the number of significant voxels is much larger than the sample size. The last two characteristics imply that even with all the true voxels being correctly selected, standard regression methods may still not be applicable due to multicolinearity. It is therefore desirable to impose a certain degree of shrinkage or grouping of the regression coefficients so that predictors with similar values can be grouped together, and thus the effective number of selected predictors is smaller than the sample size. Motivated by these considerations, in this article, we propose a Bayesian SI regression model that achieves simultaneous grouping and spatial selection of voxels that are predictive of individual responses. The key to our proposal is to define a joint Ising and Dirichlet Process (Ising-DP) prior for the regression parameters, within the framework of Bayesian stochastic search variable selection [SSVS; George

and McCulloch (1993, 1997)]. The Ising component of the prior utilizes the spatial information between voxels to smooth the selection indicators of neighboring voxels, and the DP component groups the coefficients of voxels with similar effects to improve prediction power and also reduce the posterior computational burden. This method has scientific, statistical and computational advantages over several existing alternative priors.

Bayesian inference has become increasingly popular in fMRI data analysis due to several attractive properties: first, the posterior inference offers direct probabilistic interpretation of the estimates; second, it eschews the multiple-comparison problem faced by classical inference; third, incorporating prior information is straightforward within the Bayesian framework. In particular, Markov Random Fields priors, such as the Ising prior and the Potts prior, have been widely used to account for the spatial information between voxels [e.g., Gössl, Auer and Fahrmeir (2001); Woolrich et al. (2004); Penny, Trujillo-Barreto and Friston (2005); Bowman (2007); Bowman et al. (2008); Derado, Bowman and Kilts (2010); Ge et al. (2014)] and for meta-analysis [e.g., Kang et al. (2011); Yue, Lindquist and Loh (2012)]. Johnson et al. (2013) used a joint Dirichlet Process mixture and Potts prior to achieve simultaneous clustering and selection. Within the SSVS framework, Smith et al. (2003) and Smith and Fahrmeir (2007) used the Ising prior in the context of massive univariate general linear models [GLM, Friston et al. (1995)] for identifying brain regions activated by a stimulus. It is important to stress that the setting in Smith and colleagues is fundamentally different from the SI regression in this paper: the former only involves fMRI time series, without individual scalar outcome, and it deals with selecting and smoothing the coefficients from p one-dimensional regressions (one for each voxel), a setting broadly belonging to multiple testing; whereas our paper deals with variable selection from one p-dimensional regression, a much more challenging task.

Within the SSVS but outside the fMRI literature, there is a stream of recent work on using the Ising prior to incorporate existing structure information between variables under the "large p, small n" paradigm [e.g., Li and Zhang (2010); Stingo et al. (2011); Vannucci and Stingo (2011)]. Moreover, simultaneous selection and clustering in multiple regression was discussed in Tadesse, Sha and Vannucci (2005), Kim, Tadesse and Vannucci (2006) and Dunson, Herring and Engel (2008), but none of those incorporated existing structure between covariates. Another important but under-investigated issue is phase transition in the Ising model [for a review, see Stanley (1987)], which, in the context of variable selection, leads to a drastic change (from nearly none to nearly all) in the number of variables selected given an infinitesimal change in the hyperparameters. And the difficulty and sensitivity in hyperparameter selection increases substantially as the degree of the underlying graph increases. Since the fMRI voxels naturally overlay a 3-dimensional lattice, it is crucial to select hyperparameters that avoid phase transition for valid inference and feasible computation. However, despite being intensively explored in statistical physics, phase transition and the consequent issue

of hyperparameter selection has received relatively little attention in the literature of variable selection. Li and Zhang (2010) derived a ballpark estimate of the phase transition boundary for the Ising prior using mean field theory. But their derivation is solely based on the prior distribution and does not take into account the data or any prior knowledge of the predictors, and thus the resulting range of possible hyperparameters is often very wide. In this article we develop a new analytic approach to derive a tighter boundary of the hyperparameters based on the data and the posterior distribution, and illustrate it on 2- and 3-dimensional lattices.

The rest of the article is organized as follows. Section 2 introduces the new Bayesian model and Section 3 develops an analytic approach to hyperparameter selection. Posterior computation of the model is discussed in Section 4. Section 5 compares the proposed methods with several existing methods through simulations. In Section 6 we apply the proposed method to the KLIFF study to investigate the social regulation of human emotion. Section 7 concludes.

2. The model. We formulate the problem via a standard multiple regression

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\eta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{Y}$ is the $n \times 1$ variable response, for example, the scalar arousal or valence measurement in the KLIFF study; $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ is the $n \times p$ ($p \gg n$) matrix of spatially correlated neuroimaging covariates, for example, the magnitudes of the estimated hemodynamic response function (HRF) of the voxels in the two ROIs in the study; and $\boldsymbol{\varepsilon}$ is the error term with $\boldsymbol{\varepsilon} \sim \mathrm{N}(0, \sigma^2 I_n)$. To focus on the main message, we do not consider design variables, such as age and sex, which can be easily added to the regression.

To select the voxels that are predictive of the response, we adopt the Bayesian SSVS approach that assumes the "spike-and-slab" type of mixture prior for the regression coefficients [Mitchell and Beauchamp (1988); George and McCulloch (1993, 1997); Smith and Kohn (1996)]. Specifically, we define a latent indicator $\gamma_j \in \{0, 1\}$ for each covariate that indicates whether this covariate is included in the model (i.e., whether a voxel is significantly predictive of the response). We let

$$\eta_j = \gamma_j \cdot \beta_j \quad \text{and} \quad \beta_j \sim G,$$

where β_j represents the regression coefficient of predictor j once it is selected, and G is a prespecified probability distribution. Given γ_j and G, η_j are independent following a spike-and-slab prior

$$\eta_j | (\gamma_j, G) \sim (1 - \gamma_j)\delta_0 + \gamma_j G, \tag{2}$$

where δ_0 is a point mass at 0. Our goal is to propose a new joint Ising and DP (Ising-DP) prior, where an Ising prior is imposed on $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$ to incorporate spatial information between voxels, and, in parallel, a Bayesian nonparametric DP prior is imposed on G to achieve grouping of the regression coefficients, as elaborated below.

We represent the spatial structure among the fMRI voxels via a graph. Let $i \sim j$ denote that i and j are neighboring voxels. Let $\mathcal{E} = \{(j_1, j_2) : 1 \le j_1 \sim j_2 \le p\}$ be the set of all the neighboring pairs of voxels—the edge set of the underlying graph. Given $\mathcal{E}$, let $\mathbf{a} = (a_1, \ldots, a_p)'$ be a vector and $\mathbf{B} = (b_{j_1,j_2})_{p\times p}$ be a symmetric matrix of real numbers where $b_{j_1,j_2} = 0$ for all $(j_1, j_2) \notin \mathcal{E}$. To incorporate the prior structural information into the model building process, we assume an Ising prior distribution for $\boldsymbol{\gamma}$ [Li and Zhang (2010)] as the first component of the proposed prior:

$$\Pr(\boldsymbol{\gamma}) = \exp\{\mathbf{a}'\boldsymbol{\gamma} + \boldsymbol{\gamma}'\mathbf{B}\boldsymbol{\gamma} - \psi(\mathbf{a}, \mathbf{B})\}, \tag{3}$$

where $\psi(\mathbf{a}, \mathbf{B})$ is the normalizing constant: $\psi(\mathbf{a}, \mathbf{B}) = \log\{\sum_{\boldsymbol{\gamma}\in\{0,1\}^p} \exp(\mathbf{a}'\boldsymbol{\gamma} + \boldsymbol{\gamma}'\mathbf{B}\boldsymbol{\gamma})\}$. If $\mathbf{B} = 0$, then $\psi(\mathbf{a}, \mathbf{B}) = \sum_{j=1}^{p} \log(1 + e^{a_j})$, but in general there is no closed form for ψ. The Ising model is a binary Markov Random Fields model and encourages the formation of clusters of like-valued binary variables.

The hyperparameters $\mathbf{a}$ control the sparsity of $\boldsymbol{\gamma}$. Since we are focused on 2D and 3D lattices, which are regular graphs (i.e., each vertex has the same degree), we do not want to favor a priori the inclusion of any voxel. This is achieved by letting $\mathbf{a} = a\mathbf{1}_p$, where $\mathbf{1}_p = (1, 1, \ldots, 1)' \in \Re^p$. The hyperparameters $\{b_{j_1,j_2}\}$ represent the prior belief on the strength of coupling between the pairs of neighbors (j_1, j_2), and thus control the smoothness of $\boldsymbol{\gamma}$ over $\mathcal{E}$ given $\mathbf{a}$, with larger b_{j_1,j_2} leading to tighter coupling. When $\mathbf{B} = 0$, the prior is the standard i.i.d. Bernoulli for each predictor [George and McCulloch (1993)]. Without specific prior information of the strength of connection between each pair of neighbors, it is natural to assume b_{j_1,j_2}'s to be a constant b. Then $(\mathbf{a}, \mathbf{B})$ reduce to two hyperparameters (a, b), which can be either pre-fixed or assumed to follow some hyperprior distributions.

The Ising prior smoothes the binary selection indicators, but not the regression coefficients. In structured high-dimensional settings like fMRI, neighboring covariates, often highly correlated, tend to have similar effects on the outcome. Intuitively, a certain degree of smoothing or grouping of the coefficients would improve the model fitting, especially when the effects of individual predictors are very weak. We achieve this by imposing a DP prior on G, $G \sim \mathrm{DP}(\alpha, G_0)$, with a precision parameter α and base measure G_0 [Ferguson (1973, 1974); Antoniak (1974)]. Following the sticking-breaking (SB) presentation [Sethuraman (1994)], G can be written as a weighted sum of an infinite number of point masses (atoms):

$$\begin{aligned} G(\cdot) &= \sum_{h=1}^{\infty} w_h \delta_{\theta_h}(\cdot), \qquad \theta_h \overset{\text{i.i.d.}}{\sim} G_0, \\ w_h &= w'_h \prod_{k<h} (1 - w'_k), \qquad w'_h \overset{\text{i.i.d.}}{\sim} \mathrm{Beta}(1, \alpha), \end{aligned} \tag{4}$$

where δ_θ is a point mass at θ. It is clear from (4) that samples from a DP are discrete and the component weights w_h decrease exponentially in expectation. The spike-and-slab prior (2) for each η can then be written as a mixture of an infinite number

of point masses (at 0 and atoms randomly drawn from the base measure G_0):

$$\eta_j|(\gamma_j, \mathbf{w}, \boldsymbol{\theta}) \sim (1-\gamma_j)\delta_0 + \gamma_j \sum_{h=1}^{\infty} w_h \delta_{\theta_h}(\cdot), \tag{5}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_h, \ldots)$ and $\mathbf{w} = (w_1, \ldots, w_h, \ldots)$. The clustering nature of the DP prior can be immediately seen from (5): it classifies the voxels into one cluster of voxels that have no effect on response, and several clusters of the remaining voxels, where the regression coefficients within each cluster are shrunk to be identical. The number of clusters increases automatically as the number of voxels under consideration, p, increases. The precision parameter α governs the number of active components and is assumed to follow a flexible hyper Gamma(1, 1) prior. And we assume the base measure $G_0 = \mathrm{N}(0, v^2)$ with hyperparameter v. In this article, clustering per se is not the primary interest, rather clustering is a means of grouping similar coefficients. There is a clear scientific justification for grouping regression coefficients in this manner, as each predictive brain region usually contains a number of voxels that are of similar (and usually weak) effects on the outcome. Clustering also introduces substantial improvement in posterior computation because instead of sampling the coefficient for each voxel, one only need to sample the common coefficient for each cluster.

Jointly, equations (3), (4) and (5) define the new *Ising-DP spike-and-slab prior*.

3. Selection of hyperparameters. Selection of the hyperparameters a, b in the Ising prior is crucial for both inference and computational feasibility for high-dimensional data. A challenging feature of the Ising prior in the "large p" paradigm is the phase transition behavior in a graph with dimension higher than 1: certain combinations of the hyperparameters a, b lead to the selection of almost all variables and thus induce critical slowdown of the MCMC for posterior computation. This issue cannot be mitigated by simply replacing a and b by a hyperprior, because for a regular graph with even modest degree (say, 3), the range of hyperparameters that do not incur phase transition is narrow. If the domain of the prior is not carefully chosen, it is very likely that little weight is assigned to appropriate hyperparameters, leading to poor posterior results, especially for data with low signal-to-noise ratio (SNR), such as fMRI data. Smith and Fahrmeir (2007) suggested to co-estimate the hyperparameters and the binary indicators in posterior computation. Their method relies on specifying a uniform prior between zero and a prespecified maximum for the smoothing parameter b. However, if the maximum is specified outside the phase transition bounds, the resulting MCMC will still suffer from the critical slowdown. Therefore, finding these phase transition bounds is central to correct specification of hyperparameters for the Ising prior.

Solely based on the prior distribution, Li and Zhang (2010), page 1205, used mean field approximations to derive a ballpark estimate of the phase transition boundary for the Ising prior defined on regular graphs, and illustrated it on a hypertube with degree of 6. However, because this approach does not take into account

the data or any prior knowledge of selection rate, it often results in a very wide range of hyperparameters. The problem becomes even more pronounced when the degree of the graph increases. Below we develop a new method to tighten the bounds on a and b based on the posterior distribution.

The posterior conditional density of $\boldsymbol{\gamma}$ given the rest of parameters is proportional to

$$\mathcal{C}(\boldsymbol{\gamma}) = \exp\left(\mathbf{a}'\boldsymbol{\gamma} + \boldsymbol{\gamma}'\mathbf{B}\boldsymbol{\gamma} - \sum_{i=1}^{n}\big(Y_i - \mathbf{X}_i(\boldsymbol{\beta}\cdot\boldsymbol{\gamma})\big)^2/2\sigma^2\right).$$

In high-dimensional settings, usually it is reasonable to a priori assume sparsity, that is, the proportion of true predictors among the p candidates, π, is much smaller than 1. Intuitively, in order to have only a small proportion of predictors being selected, the mode of $\mathcal{C}(\boldsymbol{\gamma})$ should be larger than $\mathcal{C}(\mathbf{0}_p)$ and attained at a $\hat{\boldsymbol{\gamma}}$ such that the number of nonzero γ's is around $\pi \cdot p$, beyond which $\mathcal{C}(\boldsymbol{\gamma})$ should decrease fast as the number of nonzero γ's increases. Below we form inequalities for a and b based on this intuition.

When all the candidate voxels locate on a lattice, selected voxels give rise to the largest number of neighboring pairs when they form a square in two dimensions or a cube in three dimensions. Therefore, we use squares (2D) or cubes (3D) to approximate the location of the selected $\pi \cdot p$ voxels on the lattice. Let $V = [(\pi \cdot p)^{1/d}]$, where $[c]$ denotes the largest integer no larger than c, and d is the dimension of the lattice, which equals either 2 or 3. For a square containing V^2 voxels, there are $4V^2 - 6V + 2$ neighboring pairs; for a cubic containing V^3 voxels, there are $13V^3 + 28 + 66(V-2) + 51(V-2)^2$ neighboring pairs (derivations are given in Appendix A).

3.1. *Selection on two-dimensional lattice.* We first discuss the two-dimensional lattice. For V^2 selected voxels on a square,

$$\mathbf{a}'\boldsymbol{\gamma} + \boldsymbol{\gamma}'\mathbf{B}\boldsymbol{\gamma} = (a+8b)V^2 - 12bV + 4b.$$

To achieve sparsity, this value needs to decrease fast as V increases, thus we must have $a + 8b < 0$. We also need the conditional density of selecting V^2 voxels to be larger than the null model with zero voxel, that is,

$$\begin{aligned} &-\sum_{i=1}^{n}(Y_i - \bar{Y})^2/2\sigma^2 \\ &\quad\le (a+8b)V^2 - 12bV + 4b - \sum_{i=1}^{n}\big(Y_i - \mathbf{X}_i(\boldsymbol{\beta}\cdot\boldsymbol{\gamma})\big)^2/2\sigma^2. \end{aligned} \tag{6}$$

Since $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ is the total variation of the observed Y, $\sum_{i=1}^{n}(Y_i - \mathbf{X}_i(\boldsymbol{\beta}\cdot\boldsymbol{\gamma}))^2$ is the sum of squared errors, and $\mathbb{E}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i(\boldsymbol{\beta}\cdot\boldsymbol{\gamma}))^2 \approx n\sigma^2$, then $\sum_{i=1}^{n}(Y_i -$

$\bar{Y})^2/2\sigma^2 - \sum_{i=1}^{n}(Y_i - \mathbf{X}_i(\boldsymbol{\beta}\cdot\boldsymbol{\gamma}))^2/2\sigma^2 \approx n \cdot \frac{R^2}{2(1-R^2)}$, where R^2 is the determinant of coefficient in the linear regression of Y versus $\mathbf{X}$. Then inequality (6) is reduced to

$$(a+8b)V^2 - 12bV + 4b > \frac{-n\cdot R^2}{2(1-R^2)}.$$

We now propose two ways to determine R^2 to further tighten the inequality. In the first method, we prespecify the R^2 value that we expect to achieve. Then given V from prior knowledge, obtain bounds on the parameters a and b. For example, if we want at least 50% of variation of Y to be explained by the regression, and at most 5% of 1000 voxels to be selected, we may let $R^2 = 50\%$ and $V = [\sqrt{50}] = 7$, then the inequality becomes $49(a+8b) - 84b + 4b > -n/2$, that is, $312b + 49a > -n/2$. Consequently, the range of a and b is determined by two inequalities: $-8b > a > (-n/2 - 312b)/49$ and $b < n/160$. The second method is to approximate R^2 by a lower bound obtained based on the data: the maximum R^2 among all simple linear regressions of Y versus each single predictor X. We believe such a lower bound is an effective approximation for the problem under study for two reasons. First, by using the DP prior, usually most of the selected voxels should have identical β, effectively converting the multiple regression to a simple linear regression. Second, for fMRI data, spatially close voxels typically have very similar X values, and thus the R^2 value from regressing Y on multiple spatially close predictors is expected to be very similar to that from regressing Y versus a single predictor.

3.2. *Selection on three-dimensional lattice.* Analogously, we can derive the range of a and b for a three-dimensional lattice. For V^3 voxels forming a cubic and $V > 1$,

$$\begin{aligned}\mathbf{a}'\boldsymbol{\gamma} + \boldsymbol{\gamma}'\mathbf{B}\boldsymbol{\gamma} &= (a+26b)(V-2)^3 + 6(a+17b)(V-2)^2 \\ &\quad + 12(a+11b)(V-2) + 8a + 56b.\end{aligned} \tag{7}$$

In order to avoid all predictors being selected, we need $C(\boldsymbol{\gamma}) < 0$ to decrease fast as V increases after certain threshold. For simplicity, we only require $C(\boldsymbol{\gamma})$ to be negative for the maximum possible V, that is, $V = [p^{1/3}]$. For example, in the KLIFF data, p is around 6600 in both ROIs, then $V = 18$ and, consequently, $a < -23b$. In addition, in order to avoid the null model, that is, no voxel being selected, we have

$$\mathbf{a}'\boldsymbol{\gamma} + \boldsymbol{\gamma}'\mathbf{B}\boldsymbol{\gamma} \geq \frac{-n\cdot R^2}{2(1-R^2)}. \tag{8}$$

Given the prespecified R^2 and V, we can obtain the range of a and b satisfying this inequality. Again taking the KLIFF data, for example, $n = 104$, we want at most

1% voxels selected, and the expected R^2 is 0.5. Then $V = [66.7^{1/3}] = 4$, plug this value and $R^2 = 0.5$ into the inequality (8), and we have $a > -14.6b - 0.81$. Combining the previously obtained inequality $a < -23b$, it must be the case that $-23b > -14.6b - 0.81$, so that we have $b < 0.1$. Therefore, for the KLIFF data analysis, we will choose a and b such that $b \leq 0.1$ and $-23b > a > -14.6b - 0.81$.

One potential problem of using (7) to evaluate $\mathbf{a}'\boldsymbol{\gamma} + \boldsymbol{\gamma}'\mathbf{B}\boldsymbol{\gamma}$ in (8) is the overestimation of the number of neighboring pairs of selected voxels, especially when the selected V is larger than 3, which can lead to a very tight range of b and a. We instead propose that as long as there is one predictor whose posterior probability of being selected is larger than that of not selected, the posterior simulation will not be stuck at the null model. Therefore, we can just let $a \geq \frac{-n \cdot R^2}{2(1-R^2)}$, implying $b < \frac{n \cdot R^2}{2 \cdot 23(1-R^2)}$ such that $-23b > \frac{-n \cdot R^2}{2(1-R^2)}$. For one of the real data sets under study, the maximum R^2 across all simple linear regressions is 0.10, then we have $-23b > a > -5.8$ and $b < 0.25$. Given the derived range of hyperparameters, and with the belief that all the true predictors are tightly clustered together, we first choose the largest possible b to induce the most spatial clustering effect; then given the value b, we choose the smallest a within the phase transition boundary to induce sparsity. Such a choice of a also brings computational advantage, because the computational cost of obtaining the regression coefficients decreases with the number of selected predictors in each MCMC iteration. Here, we choose $b = 0.2$ and $a = -4.5$ as the hyperparameters for the Ising prior.

3.3. *Remarks*. The above derivation suggests the following: first, the larger R^2 and the sample size n, and the smaller the degree of the underlying graph (i.e., the average number of neighbors of each candidate predictor), the wider the range of b; and second, the range of a depends on both b and the degree of the graph. Generally, for an Ising model built on a regular graph, given b, a larger degree of the graph leads to smaller a. These are consistent with a general understanding of the effect of prior distributions in Bayesian inference: when R^2 and n are large, indicating a strong SNR and abundant data information, choice of prior is less crucial. On the other hand, if each predictor has many neighbors, then the positive part $\boldsymbol{\gamma}'\mathbf{B}\boldsymbol{\gamma}$ in the prior will give a strong preference to models with many spatially close predictors. Therefore, we need to use a smaller b in order not to impose a strong prior. This also explains, for fixed b, the larger the degree of the graph, the smaller a is required to induce a small prior odds of selecting a large number of predictors.

The degrees of a 2D and 3D lattice are 8 and 26, respectively. Consequently, the range of hyperparameters a and b that avoids phase transition is much tighter in the latter than the former case. Indeed, in the real application, when we assume the Ising prior on a 3D lattice, the results are much more sensitive to the choice of a and b. In general, we find a larger degree of the underlying graph corresponds to substantially more difficult hyperparameter selection and inference, consistent

with the observation made in Li and Zhang (2010). Also, it is crucial to examine $\boldsymbol{\gamma}'\mathbf{B}\boldsymbol{\gamma}$. Nevertheless, when choosing the underlying graph, the concern of the degree of the graph should not outweigh the true physical structure. For example, in fMRI data, we prefer an Ising prior on a 3D lattice than on a 2D lattice, as the latter only accounts for the structure in one slice and ignores the true 3D structure between voxels.

In Bayesian variable selection problems, choice of hyperparameters affects not only posterior selection probabilities, but also computational time, convergence rate and required iteration of MCMC simulations. We found that if very few predictors are selected in each iteration, the DP prior tends to shrink the $\boldsymbol{\beta}$'s of all predictors into one identical value, leading to very sticky MCMC, which offsets the computational advantage per iteration offered by the shrinkage effect of the DP prior. Therefore, besides avoiding the two extreme ends of full selection and zero selection, the trade-off between computation per iteration and convergence rate should be taken into consideration when choosing the hyperparameters.

4. Posterior computation. We use a Gibbs sampler with data augmentation to carry out the posterior inference of the proposed model: $\boldsymbol{\gamma}|-$, $\boldsymbol{\beta}|-$, $\sigma|-$, where "$-$" denotes all the rest of the parameters. Below we describe the outline of the Gibbs sampler but relegate the computational details to Appendix B.

The procedure to update the variance σ, and the indicators $\boldsymbol{\gamma}$, which we update one at a time in a random order in each sweep, is standard. To draw posterior samples of $\boldsymbol{\beta}$, we use an approximate blocked Gibbs sampler based on the truncated stick-breaking process [Ishwaran and Zarepour (2000); Ishwaran and James (2001)]. First choose a conservative upper bound, $H < \infty$ on the number of mixture components potentially occupied by β_j's in the sample. Then introduce latent class indicators for each predictor, $Z_j (\in \{1, \ldots, H\})$ with a multinomial distribution, $Z_j \sim \mathrm{MN}(\mathbf{w})$ where $\mathbf{w} = \{w_1, \ldots, w_H\}$. This associates each predictor in the current iteration with a cluster h in the DP. In the Gibbs sampler, we first augment the cluster membership Z_j and then sample β_j conditional on Z_j.

The main computational gain, especially when p is large, is due to the clustering nature of DP: because all the predictors in one cluster share the same coefficient, we only need to update one β for each cluster within each iteration. It is easy to show the computational order of the posterior computation of one MCMC iteration under the DP prior for β is $O(n \times p \times p_{\mathrm{sel}})$, where p_{sel} is the number of selected predictors (model size) in that iteration. For comparison, we present the corresponding computational order under the standard spike-and-slab prior with Gaussian prior for β, for which there are two general schemes for posterior computation: (i) sample all parameters, β, σ and γ; and (ii) integrate out β and σ under the conjugate setup and only sample γ. In both schemes, the main computational burden is due to the inversion of the covariance matrix, which, even using fast low-rank update algorithms, is of the order $O(n \times p^2)$ and $O(n \times p \times p_{\mathrm{sel}}^2)$,

respectively. When p is very large as in this application, the computational order of the first scheme is prohibitive, and this is the reason that the vast majority of the SSVS literature in high-dimensional settings adopts the second scheme, which, however, does not provide posterior samples of the coefficients β or the variance σ. Moreover, because of the squared term of p_{sel}, even when the average model size is modest (e.g., between 50–100), the second scheme can still incur overwhelming computational cost. In contrast, as shown in the details of the Gibbs sampler in Appendices A and B, the DP prior does not require matrix inversion, yet still provides posterior samples of β's with much lower computational cost.

5. Simulations.

5.1. *Simulation design*. We conduct simulations to examine the performance of the Ising-DP prior and compare with several alternative methods. We simulate data of $n = 104$ subjects (the number of subjects in the real application), each having $p = 1000$ candidate predictors overlaying a $10 \times 10 \times 10$ 3D grid. Each predictor j ($1 \le j \le 1000$) is spatially indexed by $\mathbf{d}_j = (d_j^1, d_j^2, d_j^3)$ for $1 \le d_j^1, d_j^2, d_j^3 \le 10$. To mimic the real data, we let predictors be strongly correlated, and the design matrices of the ith subject $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$ in all the following simulations follow a multivariate normal $\text{MVN}_p(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p) \overset{\text{i.i.d.}}{\sim} \text{Unif}(3, 6)$ and $\Sigma_{j_1 j_2} = 0.8^{|\mathbf{d}_{j_1} - \mathbf{d}_{j_2}|}$, where $|\mathbf{d}_{j_1} - \mathbf{d}_{j_2}| = \sum_{i=1}^3 |d_{j_1}^i - d_{j_2}^i|$. We consider the following four simulation scenarios.

Scenario 1: *One cluster of true predictors, with identical β's*. There is a cluster of $5 \times 5 \times 5$ (125) true predictors ($\gamma_j = 1$) with spatial indices $4 \le d_j^1, d_j^2, d_j^3 \le 8$ located in the center of the 3D cube. The coefficients β of the true predictors are set to 0.6. The response is generated from $Y_i = \sum_j X_{i,j} \beta_j \gamma_j + \varepsilon_i$ with $\varepsilon_i \sim \text{N}(0, 200^2)$ for $i = 1, \ldots, n$, creating a data set with a low SNR 5%—defined as $\mathbb{V}(\mathbf{X}\boldsymbol{\beta})/\mathbb{V}(\boldsymbol{\varepsilon})$. The following scenarios also all have such a low SNR, which is the norm in real fMRI data.

Scenario 2: *One cluster of true predictors, with varying but strongly correlated β's*. We let the coefficients of the true predictors, locating on the same grid as those in scenario 1, vary and follow $\text{MVN}_p(0.6 \times \mathbf{1}_p, \Omega)$, where $\Omega_{j_1 j_2} = 0.1 \times 0.95^{|\mathbf{d}_{j_1} - \mathbf{d}_{j_2}|}$. Therefore, both the observed values and the underlying coefficients of neighboring predictors are strongly correlated.

Scenario 3: *Two clusters of true predictors, with identical β's within each cluster*. A more challenging scenario is when there are multiple spatially separated clusters of true predictors. Specifically, we let the true predictors form two clusters: one overlays the grid of $3 \le d_j^1 \le 4, 3 \le d_j^2 \le 4, 3 \le d_j^3 \le 4$, and another overlays the grid of $6 \le d_j^1 \le 9, 6 \le d_j^2 \le 9, 6 \le d_j^3 \le 9$. We set the coefficients β of the predictors in the two clusters to 0.4 and 1, respectively.

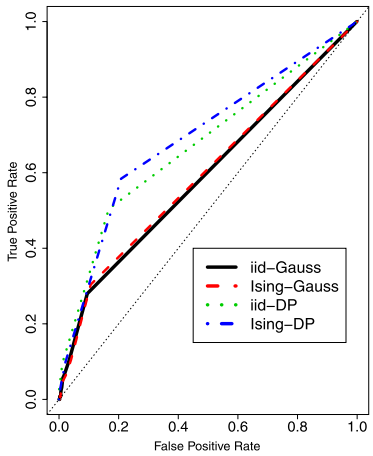
Scenario 4: *Two clusters of true predictors, with varying β's within each cluster*. The true predictors locate on the same grid as those in scenario 3, and one cluster of $\boldsymbol{\beta}$ were generated from $\mathrm{MVN}_p(0.4 \times \mathbf{1}_p, \Omega_1)$ with $\Omega_{1,j_1j_2} = 0.1 \times 0.95^{|\mathbf{d}_{j_1} - \mathbf{d}_{j_2}|}$, and those in the second cluster are from $\mathrm{MVN}_p(1 \times \mathbf{1}_p, \Omega_2)$ with $\Omega_{2,j_1j_2} = 0.1 \times 0.95^{|\mathbf{d}_{j_1} - \mathbf{d}_{j_2}|}$. Variable selection under two-cluster scenarios is challenging: the strong correlation between the predictors outside and inside the clusters renders differentiating nonsignificant predictors, especially those located between the two clusters, from the true ones difficult.

For each of the simulated data set, we fit the regression model (1) with four different priors: (i) i.i.d. Bernoulli prior for γ_j, with a Gaussian prior for the β_j's (this is the standard spike-and-slab prior), referred to as the *i.i.d.-Gaussian* prior; (ii) Ising prior for γ_j, with a Gaussian prior for the β_j's, referred to as the *Ising-Gaussian* prior; (iii) i.i.d. Bernoulli prior for γ_j, with a DP prior for β_j's, referred to as the *i.i.d.-DP* prior; (iv) the *Ising-DP* prior. The hyperparameters (a, b) for the Ising priors are chosen by the proposed approach in Section 3, with $a = -5$ and $b = 0.25$. For the DP priors, we set $H = 20$, $\alpha = 1$ and $v = 10$ such that G_0 is very flat in a wide domain. For each simulated data, we run 10 parallel Gibbs samplers with random start in $\boldsymbol{\gamma}$, each having 20,000 iterations with the first 10,000 ones as burn-in. Posterior computation with the i.i.d.-Gaussian and Ising-Gaussian priors are carried out using the software by Li and Zhang (2010). The main summary statistic, the posterior inclusion probability, is deemed convergent upon inspecting the Gelman–Rubin statistic [Gelman and Rubin (1992)]. In all of our experiments, the 10 simulations lead to highly similar posterior summary statistics.

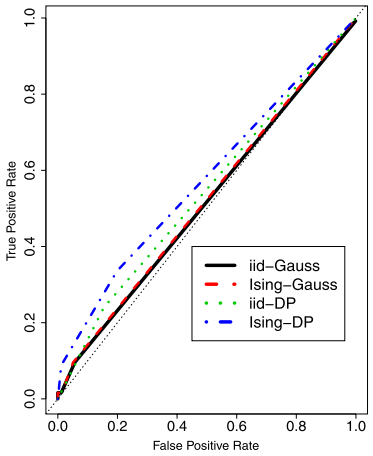
5.2. *Simulation results*. We calculate the posterior inclusion probabilities $\Pr(\gamma_j = 1|\mathbf{Y})$ as the posterior summary statistics, obtained by dividing the number of iterations where $\gamma_j = 1$ over the total number of iterations excluding the burn-in period. To summarize these marginal probabilities, we compute the ROC curve as follows: only those covariates j with $\Pr(\gamma_j = 1|\mathbf{Y})$ greater than a threshold are deemed positives, and those below the threshold are deemed negatives; the ROC curve reflects the pair of true positive rate and false positive rate achieved by varying the calling threshold. The bigger area under the ROC curve (maximum 1), the better the discriminating power of the model.

The ROC curves resulting from the simulations under scenarios 1–2 (one cluster) and 3–4 (two clusters) are presented in the top and bottom panel of Figure 1, respectively. We also calculated the root mean squared error (RMSE) per variable, $(\sum_j (\hat{\beta}_j - \beta_j)^2/p)^{1/2}$, of each prior, summarized in Table 1. In all four simulations, the Ising-DP prior resulted in the best ROC, closely followed by the i.i.d.-DP prior, beating both the i.i.d.-Gaussian and the Ising-Gaussian priors. This pattern is consistent with the RMSEs. Overall, the ROC curves suggest relatively low discriminating power in these simulations, even for the best-performing Ising-DP prior. This is not surprising because variable selection under all four scenarios

One cluster

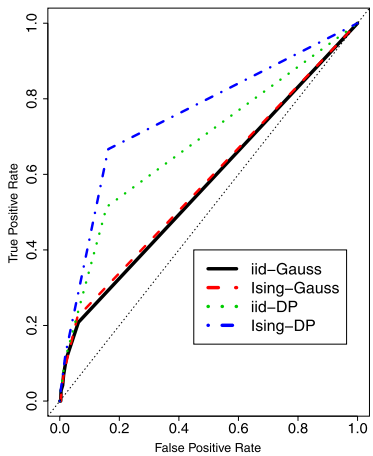


(a) Identical β of true predictors

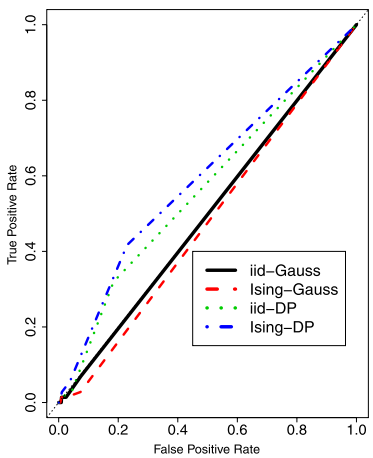


(b) Varying β of true predictors

Two clusters



(c) Identical β of true predictors within a cluster



(d) Varying β of true predictors

FIG. 1. *ROC curves based on the posterior selection probability* $\Pr(\gamma_j = 1|\mathbf{Y})$ *obtained from i.i.d.-Gaussian, Ising-Gaussian, i.i.d.-DP and Ising-DP prior, respectively, under four simulation scenarios.*

is very challenging due to the low SNR, strong correlation between variables and the small-n large-p nature. Indeed, our experience based on more simulations suggests that as the SNR and/or the sample size decreases, performance of all the priors drops, but the Ising-DP prior is the least affected, demonstrating the benefit of introducing additional shrinkage to the coefficients when the signal is weak. In summary, it is evident from these simulations that the Ising-DP prior outperforms

TABLE 1
Root mean squared error (RMSE) per variable, $(\sum_i (\hat{\beta}_i - \beta_i)^2/p)^{1/2}$, *by different priors*

Scenario	I.i.d.-Gaussian	Ising-Gaussian	I.i.d.-DP	Ising-DP
1. One-cluster identical β	0.623	0.599	0.190	0.190
2. One-cluster varying β	0.284	0.283	0.181	0.179
3. Two-cluster identical β	0.311	0.315	0.256	0.250
4. Two-cluster varying β	0.368	0.251	0.235	0.233

the existing alternatives in data with characteristics similar to those of the fMRI data under study.

It is worth noting that in these simulations the DP component appears to impose a stronger clustering effect on performance than the Ising component. One reason is that, as shown in Section 3, when the degree of the graph is large as in the 3D fMRI analysis, the hyperparameter b in the Ising prior used to control the clustering effect has to be set small to avoid phase transition, which consequently limits its clustering effect. Nevertheless, the simulation results suggest that incorporating the spatial information into Bayesian variable selection via the Ising prior still leads to improved selection accuracy than otherwise.

6. Application to the KLIFF study.

6.1. *The data*. We now provide more information on the design of the KLIFF study and the preprocessing procedure. For each of the 104 pairs of participants in a close relationship (referred to as partners hereafter), one of them was randomly selected to be threatened by electric shocks while their brain activities were measured by fMRI in three separate sessions: in one session he/she is holding hands with his/her partner; in the second session, he/she is holding hands with a stranger; in the third session, he/she is alone, holding hands with nobody at all. The three hand-holding conditions mimic three types of social interactions. Each of the three sessions, randomized within each pair of partners, contains 24 trials in random order, half of which are threat cues (a red "X" on a black background) indicating a 20% likelihood of receiving an electric shock to the ankle, and the other half are safety cues (a blue "O" against a black background) indicating no chance of shock. A 3D fMRI scan of the subject's brain was acquired for every 2 seconds in the experiment lasting for 400 s. Overall, fMRI data collected from the KLIFF experiment consist of 104 subjects in 3 sessions at 200 time points for over 100,000 spatially distributed voxels. At the end of each session, the subjects facing the threat were asked to score their arousal and valence feelings experienced during the experiment. Both the arousal and valence measurements range from 1 to 9, encoding feelings from calming/soothing to alert/agitated, and feelings from highly negative/miserable to highly positive/pleased, respectively.

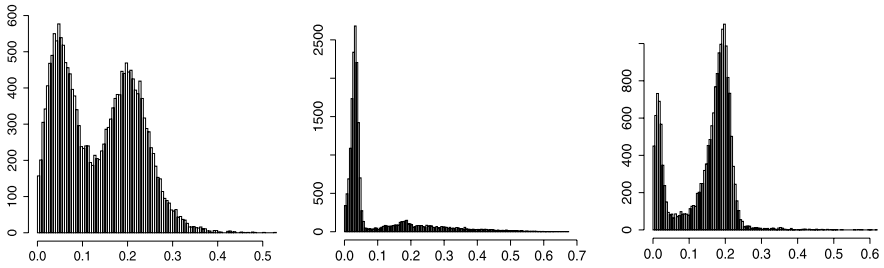
Preprocessing of the fMRI data was carried out via FMRIB's Software Library (FSL) software [Version 5.98; Smith et al. (2004)]. Registration of the images in FLIRT [Jenkinson et al. (2002)] was based on Montreal Neurological Institute (MNI) space. More details of preprocessing can be found in Zhang et al. (2013). ROIs were determined structurally using the Harvard subcortical brain atlas, and were chosen for their likely involvement in affective processing based on previous studies [Maresh, Beckes and Coan (2013)]. In particular, our analysis focuses on two emotion related regions: dorsal anterior cingulate cortex (dACC) and insula, which were commonly implicated in negative affect and threat responding, and whose numbers of voxels are similar, 6666 and 6591, respectively. To obtain the predictors, we conducted massive univariate analysis using the GLM to get scalar summaries of the fMRI time series. Specifically, for every voxel in each ROI, we *used* the semi-parametric GLM approach in Zhang et al. (2013) to estimate the hemodynamic response functions (HRF) corresponding to the threat and safety cues (stimuli), and extracted the height of the HRF estimates, interpreted as the magnitude of brain response to the stimuli of that voxel. We then computed the difference between the estimated magnitudes under the threat cue and the safety cue (baseline) for each voxel as the predictors. In total, for each ROI, we obtained six sets of regression data: two different response variables—valence and arousal scores of the subjects, under each of the three hand-holding conditions, and associated magnitude estimates of each voxel in the ROI collected in the same session as the predictors.

6.2. *Results*. We applied the proposed Bayesian model to the 12 sets of data (6 for each ROI) using the Ising-DP prior on a 3D lattice with hyperparameter $a = -4.5$ and $b = 0.2$ obtained from the method in Section 3. For comparison, we also fit the model with the i.i.d.-Gaussian and the Ising-Gaussian priors. For each regression, 25,000 iterations of MCMC were performed with the first 5000 discarded as burn-in. Convergence of the marginal inclusion probabilities is deemed via the Gelman–Rubin statistics.

Though the number of selected predictors is larger than the sample size in each MCMC iteration, the clustering effect of the DP prior leads to a small number of different β values (less than 10) in most iterations. Among the 12 sets of regressions, we focused on those with (i) reasonably high R-squared values and (ii) top 10% selected voxels having a high proportion of nonzero coefficients with the same sign. The R-squared value for each iteration t is given by

$$R_t^2 = 1 - \mathbb{V}\text{ar}(\mathbf{Y} - \mathbf{X}\boldsymbol{\gamma}_t \cdot \boldsymbol{\beta}_t)/\mathbb{V}\text{ar}(\mathbf{Y}),$$

where $\boldsymbol{\gamma}_t$ and $\boldsymbol{\beta}_t$ are the posterior draws of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, respectively, at the tth iteration. The first criterion requires that a significant proportion of variation of subjects' emotion measurements can be explained by their brain response magnitudes, and the second requires that the majority of the top selected predictors have

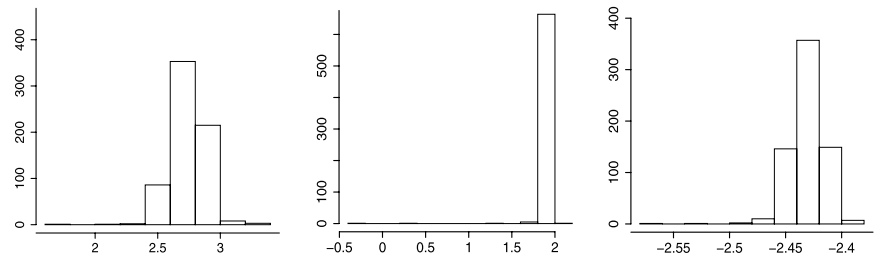


(a) dACC alone arousal (b) Insula alone arousal (c) Insula partner valence

FIG. 2. *R-squared values of the regressions.*

similar and significant effects on the response, matching the substantive knowledge from the existing psychology literature. We found three sets of regressions fit these two criteria: the regression with the arousal measurement under alone condition as the response in dACC and insula, respectively, and the regression with the valence measurement under hand-holding-with-partner condition as the response in insula.

Histograms of the R-squared values and the coefficients of the top 10% selected voxels in these three regressions are displayed in Figures 2 and 3, respectively. We can see that in the regression with arousal under the alone condition as the response in dACC, the R-squared value is larger than 20% in more than 20% of the MCMC draws [Figure 2(a)], and almost all (>99.5%) of the top 10% selected voxels' coefficients are positive in more than 90% of the posterior draws [Figure 3(a)]. The same regression in insula led to similar results [R-squared in Figure 2(b) and coefficients in Figure 3(b)]. The significant positive association between the arousal measurement and brain response magnitudes under the alone condition is consistent with related findings in the literature. First, in a previous



(a) 10% percentile dACC arousal (b) 10% percentile insula arousal (c) 90% percentile insula valence

FIG. 3. *Histograms of* 10% *or* 90% *percentile of the coefficients* (*in scale* 10^{-4}) *of the top* 10% *selected voxels in dACC and insula when regressing subjects' arousal* (*the first two figures*) *or valence* (*the third figure*) *scores versus the magnitude of brain response to threat under the alone or hand-holding-with-partner condition.*

study of the KLIFF data [Zhang et al. (2013)], we found that the brain response to threat stimulus is most active when subjects are alone. This phenomenon can be explained through the social baseline theory [Beckes and Coan (2011); Coan, Beckes and Allen (2013); Coan and Maresh (2014)], which suggests that the human brain assumes proximity to other human beings, and perceives the environment as less threatening during the presence of other people in a close relationship, and thus serving as a default, or baseline, strategy of emotion regulation. This reduces the need to rely on effortful self-regulation in response to threat. On the other hand, when the subjects are alone without any social support, their brains have to use their own energy for emotion regulation, and, consequently, their emotional response is strong, and its association with subjects' emotion measurements is easier to detect in the two emotion-related ROIs. Second, the positive association between brain response and excitement level corresponds with literature showing a role for dACC and insula in both cognitively- and physically-induced arousal [Critchley et al. (2000); Lewis et al. (2007)]. Since the use of electric shock as a threat stimulus causes physical pain and induce subjects' internal awareness of upcoming pain during anticipation of a shock particularly, it is natural that the more active emotion-related ROIs process the stimulus, the more intense and agitated feeling the subjects experience.

We also found significant association between valence and brain response magnitude in insula under hand-holding-with-partner condition [R-squared values shown in Figure 2(c) and coefficients shown in Figure 3(c)]. The negative association has two possible explanations. First, the threat stimulus induces subjects' negative feelings, and the valence and arousal measures are negatively correlated, therefore, the more active the brain responds to the stimulus, the less pleased the subjects' feelings. Second, according to the social baseline theory, humans feel less threatened under the hand-holding-with-partner condition. Thus, subjects' emotion variation is more likely to occur in the valence dimension. We indeed found that the variance of subjects' valence is larger than that of arousal. Moreover, insula is thought to mediate the awareness of internal bodily and emotional states [Craig (2009)] and is related to pain anticipation and intensity [Wiech, Ploner and Tracey (2008)]. Results of the regression under the hand-holding-with-stranger condition are not as stable as the other two regressions, possibly due to the individual differences in cognitive and affective perception of strangers.

In all three regressions, the largest posterior selection probabilities of voxels are around 0.1, and the majority of the probabilities are below 0.05. This is as expected given the very low SNR common in fMRI data. In these situations, arguably, the ranks rather than absolute value of the probabilities are more informative about the selection results. Figures 4, 5 and 6 show the heatmaps of the posterior selection probabilities of the voxels in three slices based on their rank under the Ising-DP (top panel) in these regressions, respectively, in comparison to the corresponding heatmaps under the i.i.d.-Gaussian (middle panel) and the Ising-Gaussian prior (bottom panel). The color scale is arbitrary, with dark red representing the selection

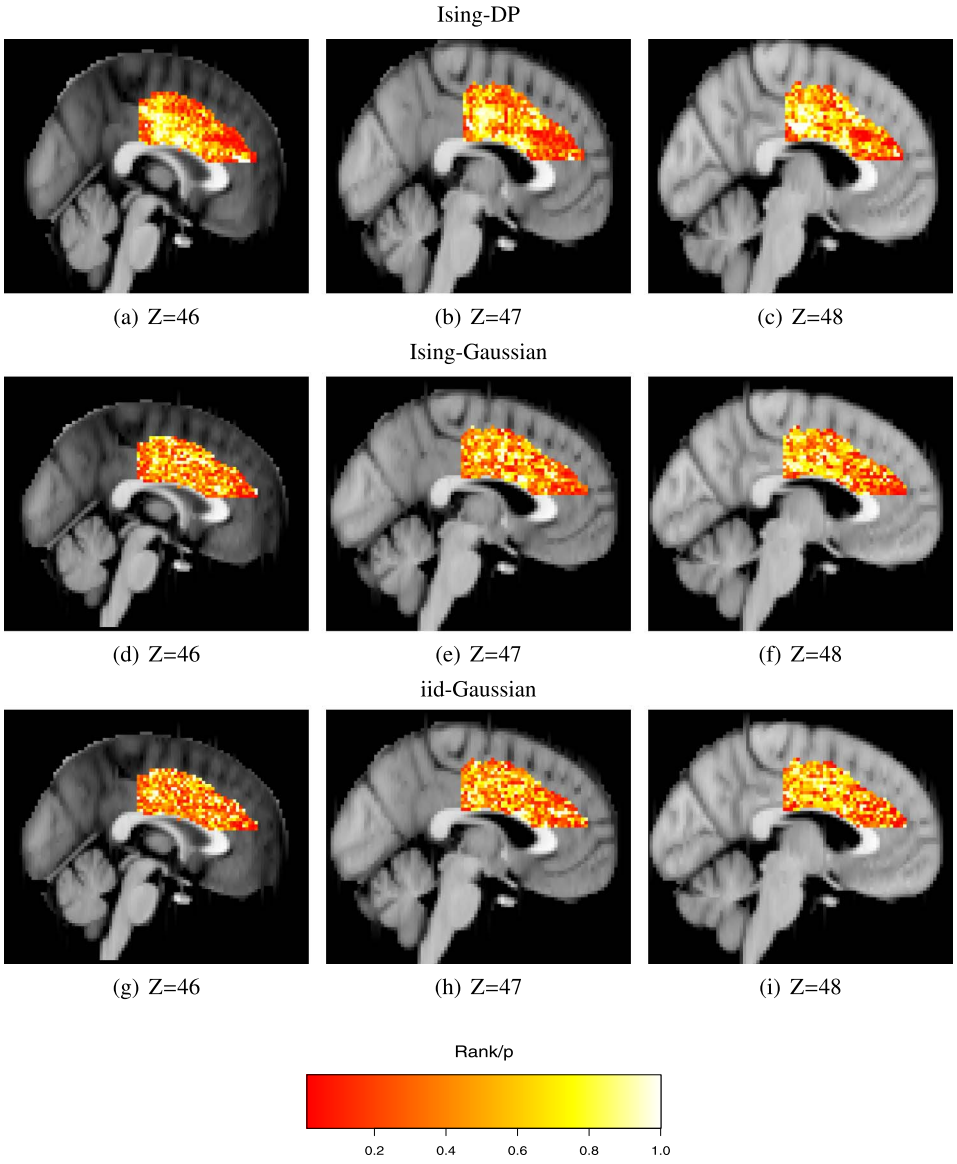


FIG. 4. *Heatmaps of voxels according to the ranks of their posterior inclusion probabilities obtained from Ising-DP, Ising-Gaussian and i.i.d.-Gaussian priors, respectively, in the Bayesian regression of subjects' arousal scores versus the magnitude of brain response to threat of voxels in dACC and insula when subjects are alone.*

probability in the lowest rank and light yellow representing the highest rank. The most striking pattern from these graphs is that the areas with the highest selection probabilities identified by the Ising-DP prior were smoothly located across the

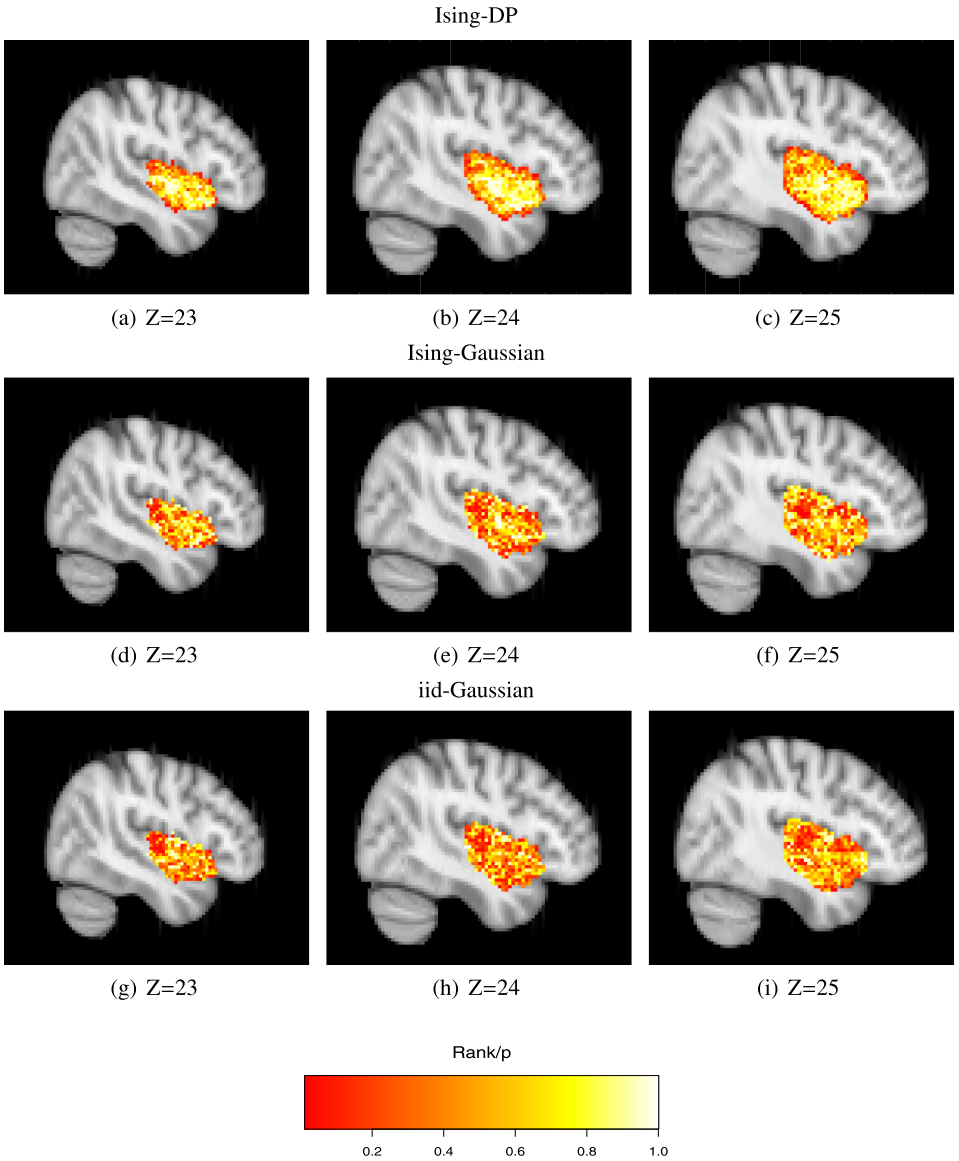


FIG. 5. *Heatmaps of voxels according to the ranks of their posterior inclusion probabilities obtained from Ising-DP, Ising-Gaussian and i.i.d.-Gaussian priors, respectively, in the Bayesian regression of subjects' arousal scores versus the magnitude of brain response to threat of voxels in insula when subjects are alone.*

ROIs, matching the scientific understanding of human brain functions, in contrast to those by the i.i.d.-Gaussian or the Ising-Gaussian prior, which are very diffused and scattered across the entire region.

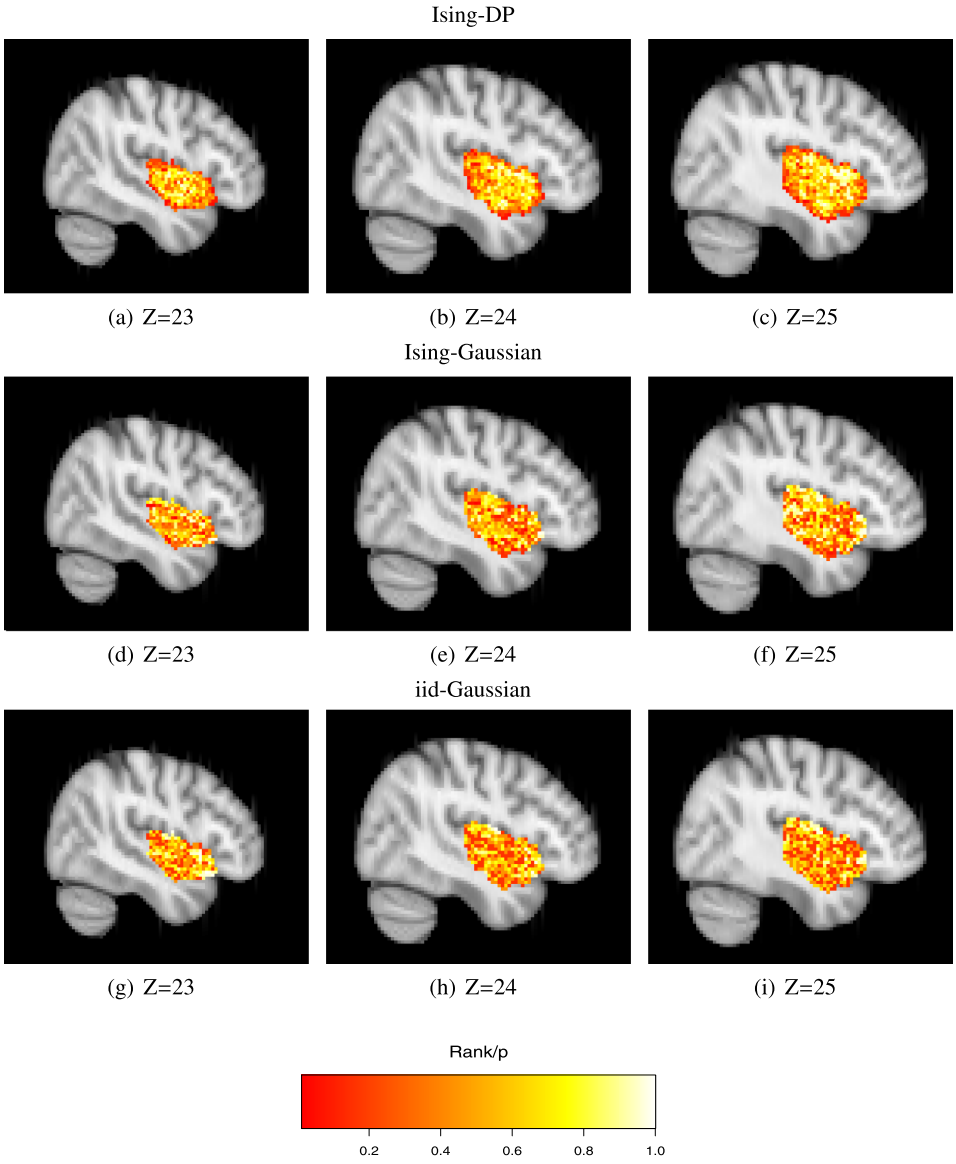
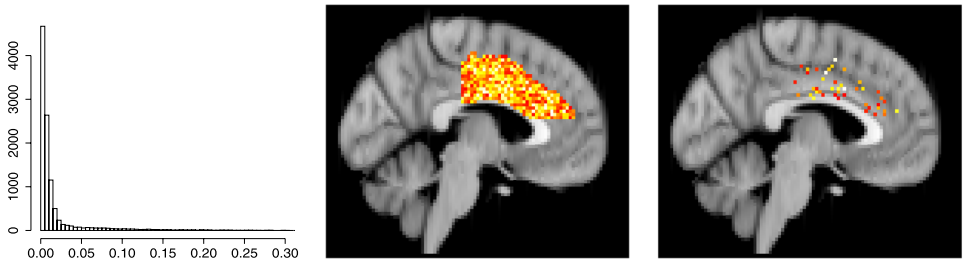


FIG. 6. *Heatmaps of voxels according to the ranks of their posterior inclusion probabilities obtained from Ising-DP, Ising-Gaussian and i.i.d.-Gaussian priors, respectively, in the Bayesian regression of subjects' valence scores versus the magnitude of brain response to threat of voxels in insula when subjects are hand holding with their partners.*

Since the underlying truth is unknown, we use a simulation-based procedure to obtain the sampling distribution of the R-squared values of a null model. Specifically, we simulated, independently of the covariates, a normally distributed re-



(a) Histogram of R-squared

(b) Heatmap of selection probabilities of voxels

(c) Voxels with highest top 10% selection probabilities

FIG. 7. *Regression of simulated response versus brain activity measurements in dACC under alone condition.*

sponse variable with similar variance and range as the observed emotion measurements, and applied the Bayesian model to regress the simulated outcome on the observed covariates in dACC under the alone condition. The histogram of positive R-squared values in the posterior draws of this null model, shown in Figure 7(a), centers around zero, and is distinct from the histograms from the aforementioned three regressions, each of which has a much higher proportion of large R-squared values. In contrast, the histogram of the null model is very similar to those from the remaining nine regressions. As such, we deem there is no statistically significant association between the covariates and the responses in these nine regressions.

7. Discussion. Motivated by the KLIFF hand-holding experiment, in this article we propose a joint Ising-DP prior within the Bayesian SSVS framework to achieve selection and grouping of spatially correlated variables in high-dimensional SI regression models. We developed an analytic approach for deriving the bounds of the hyperparameters to avoid phase transition, a main challenge in methods involving the Ising prior. Though the bounds provided by our method are tighter than the previous mean field bounds, they are still only ballpark estimates and may be wide in graphs with high degrees. A focus of our future research is therefore to improve the method of hyperparameter selection for a more complex graphical structure.

A major challenge to MCMC-based Bayesian methods in high-dimensional settings is computation. Though the DP prior in our model partially reduces the computational load by clustering the coefficients, computational scalability remains a challenge given the large p. Indeed, currently we are not able to perform a whole brain analysis with $p \approx 100{,}000$. Moreover, the mixing rate of the MCMC of the standard strategy in SSVS of updating one variable at a time may be slow, especially when the DP prior is involved. An attractive direction is to design a block update Gibbs sampling scheme that updates multiple variables at a time, and to parallelize the computation within a block using graphics processing unit (GPU)-based programming [Suchard et al. (2010); Ge et al. (2014)]. The procedure can

be further speeded up by carefully selecting the block so that it matches the underlying block structure.

The Ising prior is a special case of Markov random fields. Kalus, Sämann and Fahrmeir (2014) proposed latent GMRFs via a probit model. The probit-GMRF prior simplifies the calculation of the hyperparameters and does not suffer from the phase transition behavior. However, the main computational hurdle of inversion of a matrix of the size of selected variables remains. Nevertheless, it is possible to combine the DP prior with the probit-GMRF prior to reduce the computation.

Extension to binary and categorical responses is, in principle, straightforward using generalized linear models. Computation is an increased focus, as closed-form posterior conditional distributions are no longer available. The same problem applies with censored survival models. Laplace approximations [Raftery (1996)] are useful, but they usually require gradient methods for iterative computation of posterior modes for each sweep of covariates. A possible improvement can be obtained by exploiting the majorization–minimization/maximization (MM) algorithm [Lange (2008)]—a generalized version of the EM algorithm—for within-model mode computations.

The proposed Ising-DP prior inherently assumes sparsity, that is, only a small portion of the voxels in the ROIs are associated with the individual scalar outcome. This is achieved via a point mass (spike-and-slab) prior for the regression coefficients, resulting in a "hard-thresholding" of the β's. However, in our real application, posterior probabilities of inclusion of nearly all voxels are relatively small, which suggests that an alternative "soft-thresholding" without sparsity—achieved by (spatial adaption of) LASSO-type priors [Park and Casella (2008)]—may be desirable and a worthwhile direction for future investigation.

Though we have focused on fMRI, the proposed model is applicable to other imaging modalities where detailed spatial information between covariates is available, such as DTI or MRI.

Matlab code that implements the method is available at http://faculty.virginia.edu/tingtingzhang/Software.html.

APPENDIX A: CALCULATION OF $\mathbf{a}'\boldsymbol{\gamma} + \boldsymbol{\gamma}'\mathbf{B}\boldsymbol{\gamma}$

1. *Two-dimensional square*. For $V^2 (V > 1)$ voxels on a square, the $(V-2)^2$ voxels in the center all have 8 neighbors, the 4 vertex voxels have 3 neighbors, and the $4 \cdot (V-2)$ voxels on the edge but not vertexes have 5 neighbors. Then, given $\mathbf{a}$ and $\mathbf{B}$ as defined in Section 2, we have

$$\begin{aligned}\mathbf{a}'\boldsymbol{\gamma} + \boldsymbol{\gamma}'\mathbf{B}\boldsymbol{\gamma} &= a \cdot V^2 + b \cdot \big(8 \cdot (V-2)^2 + 4 \cdot 3 + 5 \cdot 4 \cdot (V-2)\big) \\ &= (a+8b)V^2 - 12bV + 4b.\end{aligned}$$

2. *Three-dimensional cube*. For V^3 $(V > 1)$ voxels in a cube, the $(V-2)^3$ voxels in the center all have 26 neighbors, the 8 voxels on the vertex have 7 neighbors,

the $12(V-2)$ voxels on the edge but not vertexes have 11 neighbors, and the $6(V-2)^2$ voxels on the 6 outside faces of the cube but not on the edges have 17 voxels. Then, given $\mathbf{a}$ and $\mathbf{B}$ as defined in Section 2, we have

$$\begin{aligned}
&\mathbf{a}'\boldsymbol{\gamma} + \boldsymbol{\gamma}'\mathbf{B}\boldsymbol{\gamma} \\
&\quad = a \cdot V^3 + b \cdot \big(26(V-2)^3 + 8 \cdot 7 + 12(V-2) \cdot 11 + 6(V-2)^2 \cdot 17\big) \\
&\quad = (a+26b)(V-2)^3 + 6(a+17b)(V-2)^2 + 12(a+11b)(V-2) \\
&\qquad + 8a + 56b.
\end{aligned}$$

APPENDIX B: POSTERIOR DISTRIBUTIONS IN THE GIBBS SAMPLER

1. *Update* $\boldsymbol{\gamma}$. We update the indicator for one voxel γ_j at a time. Let $\boldsymbol{\gamma}_{(-j)} = \{\gamma_l : l \neq j\}$, $I_{(-j)}$ be the set of indices $\{\gamma_l = 1 : l \neq j\}$, $\boldsymbol{\beta}_{(-j)} = \{\beta_l : l \neq j\}$, and $\mathbf{X}_{(-j)}$ be the design matrix corresponding to $\boldsymbol{\beta}_{(-j)}$. The prior probability of $\gamma_j = 1$, $\Pr(\gamma_j = 1|\boldsymbol{\gamma}_{(-j)})$ is $\exp(a + b\sum_{l\in I_{(-j)}}\gamma_l)/(1+\exp(a+b\sum_{l\in I_{(-j)}}\gamma_l))$. By the Bayes rule, the posterior probability of $\gamma_j = 1$ given the data and other parameters is

$$\begin{aligned}
&\Pr(\gamma_j = 1|\boldsymbol{\gamma}_{(-j)}, \boldsymbol{\beta}, \sigma, \mathbf{Y}) \\
&\qquad = \frac{\Pr(\gamma_j = 1|\boldsymbol{\gamma}_{(-j)})}{\Pr(\gamma_j = 1|\boldsymbol{\gamma}_{(-j)}) + F(j|\boldsymbol{\gamma}_{(-j)})^{-1} \cdot \Pr(\gamma_j = 0|\boldsymbol{\gamma}_{(-j)})},
\end{aligned}$$

where $\boldsymbol{\beta} \cdot \boldsymbol{\gamma}$ denotes the dot product between $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and $F(j|\boldsymbol{\gamma}_{(-j)})$ is the Bayes factor,

$$\begin{aligned}
F(j|\boldsymbol{\gamma}_{(-j)}) &= \frac{\Pr(\mathbf{Y}|\gamma_j = 1, \boldsymbol{\gamma}_{(-j)}, \boldsymbol{\beta}, \sigma)}{\Pr(\mathbf{Y}|\gamma_j = 0, \boldsymbol{\gamma}_{(-j)}, \boldsymbol{\beta}, \sigma)} \\
&= \frac{\exp\{-\sum_{i=1}^n (Y_i - \mathbf{X}_i\boldsymbol{\beta} \cdot \boldsymbol{\gamma})^2/2\sigma^2\}}{\exp\{-\sum_{i=1}^n (Y_i - \mathbf{X}_{i,(-j)}\boldsymbol{\beta}_{(-j)} \cdot \boldsymbol{\gamma}_{(-j)})^2/2\sigma^2\}},
\end{aligned}$$

where $\mathbf{X}_{i,(-j)}$ is the ith row of matrix $\mathbf{X}_{(-j)}$.

2. *Update* σ^2. $\sigma^2|- \sim \text{Inv-Gamma}(n/2, \mu_\sigma)$, where $\mu_\sigma = \sum_i (Y_i - \mathbf{X}_i\boldsymbol{\beta} \cdot \boldsymbol{\gamma})^2/2$.

3. *Update* $\boldsymbol{\beta}$. Denote the β_j's in $Z_j = h$ by β^h, and let $\mathbf{X}_i^h = \sum_{j:\gamma_j=1, Z_j=h} X_{ij}$. Note that $\mathbf{X}_i^h = 0$ if $\{j : \gamma_j = 1, Z_j = h\} = \varnothing$. Also, let $\boldsymbol{\beta}^{(-h)} = \{\beta_j : Z_j \neq h\}$, $\boldsymbol{\gamma}^{(-h)} = \{\gamma_j : Z_j \neq h\}$ and $\mathbf{X}^{(-h)} = \{\mathbf{X}_j : Z_j \neq h\}$, respectively, denote the collection of all the β's and the design matrix of the covariates not in cluster h. Then for $h = 1, \ldots, H$,

$$\beta^h|- \sim \mathrm{N}(\mu^h, 1/S^h),$$

with $S^h = \sum_{i=1}^n (\mathbf{X}_i^h)^2/\sigma^2 + 1/v^2$ and $\mu^h = \{\sum_{i=1}^n (Y_i - \mathbf{X}_i^{(-h)}\boldsymbol{\beta}^{(-h)} \cdot \boldsymbol{\gamma}^{(-h)})\mathbf{X}_i^h\}/S^h$. This part can be parallelized (across h).

The posterior cluster membership Z is drawn from a multinomial distribution with

$$\Pr(Z_j = h|\gamma_j = 1, -) = \frac{w_h \exp\{-\sum_{i=1}^n (Y_i - \mathbf{X}_i \boldsymbol{\beta}_{(jh)} \cdot \boldsymbol{\gamma}_{(jh)})^2/2\sigma^2\}}{\sum_{k=1}^H w_k \exp\{-\sum_{i=1}^n (Y_i - X_i \boldsymbol{\beta}_{(jk)} \cdot \boldsymbol{\gamma}_{(jk)})^2/2\sigma^2\}},$$

$$\Pr(Z_j = h|\gamma_j = 0, -) = w_h,$$

where $\boldsymbol{\beta}_{(jh)} = (\beta_1, \ldots, \beta_{j-1}, \beta^h, \beta_{j+1}, \ldots, \beta_p)$ and $\boldsymbol{\gamma}_{(jh)} = (\gamma_1, \ldots, \gamma_{j-1}, 1, \gamma_{j+1}, \ldots, \gamma_p)$ for $h = 1, \ldots, H$ and $j = 1, \ldots, p$. To update the associated weights $\mathbf{w}$, first set $w'_H = 1$ and draw w'_h from $\text{Beta}(1 + \sum_{j: Z_j = h} 1, \alpha + \sum_{j: Z_j > h} 1)$ for each $h \in \{1, \ldots, H-1\}$, then update $w_h = w'_h \prod_{k<h}(1 - w'_k)$.

Acknowledgments. The authors are grateful to the Associate Editor and four reviewers for constructive comments that helped improve the exposition and clarity of the paper, and to Nancy Zhang for insightful discussions. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIMH, the National Institutes of Health or SAMSI.

Part of the project was conducted when Fan Li and Tingting Zhang were research fellows of the Object Data Analysis program of the U.S. Statistical and Applied Mathematical Sciences Institute (SAMSI).

SUPPLEMENTARY MATERIAL

Heatmaps (DOI: 10.1214/15-AOAS818SUPP; .pdf). We provide the heatmaps of the voxels with top 10% highest posterior selection probabilities obtained, resulting from Ising-DP, Ising-Gaussian and i.i.d.-Gaussian priors, respectively, in three regressions [Li et al. (2015)].

REFERENCES

ALLEN, J. P., PORTER, M., MCFARLAND, F. C., MCELHANEY, K. B. and MARSH, P. (2007). The relation of attachment security to adolescents' paternal and peer relationships, depression, and externalizing behavior. *Child Development* **78** 1222–1239.

ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152–1174. MR0365969

BECKES, L. and COAN, J. A. (2011). Social baseline theory: The role of social proximity in emotion and economy of action. *Social and Personality Psychology Compass* **5** 976–988.

BOWMAN, F. D. (2007). Spatiotemporal models for region of interest analyses of functional neuroimaging data. *J. Amer. Statist. Assoc.* **102** 442–453. MR2370845

BOWMAN, F. D., CAFFO, B., BASSETT, S. S. and KILTS, C. (2008). A Bayesian hierarchical framework for spatial modeling of fMRI data. *NeuroImage* **39** 146–156.

BRADLEY, M. M. and LANG, P. J. (1994). Measuring emotion: The self-assessment mankin and the semantic differential. *J. Behav. Ther. Exp. Psychiatry* **25** 49–59.

COAN, J. A. (2010). Adult attachment and the brain. *J. Soc. Pers. Relatsh.* **27** 210–217.

COAN, J. A. (2011). The social regulation of emotion. In *Oxford Handbook of Social Neuroscience* 614–623. Oxford Univ. Press, New York.

COAN, J. A., BECKES, L. and ALLEN, J. P. (2013). Childhood maternal support and social capital moderate the regulatory impact of social relationships in adulthood. *Int. J. Psychophysiol.* **88** 224–231.

COAN, J. A. and MARESH, E. L. (2014). Social baseline theory and the social regulation of emotion. In *The Handbook of Emotion Regulation*, 2nd ed. (J. Gross, ed.) 221–236. The Guilford Press, New York.

COAN, J. A., SCHAEFER, H. S. and DAVIDSON, R. J. (2006). Lending a hand: Social regulation of the neural response to threat. *Psychol. Sci.* **17** 1032–1039.

CRAIG, A. D. (2009). How do you fell now? The anterior insula and human awareness. *Nat. Rev. Neurosci.* **10** 59–70.

CRITCHLEY, H. D., CORFIELD, D. R., CHANDLER, M. P., MATHIAS, C. J. and DOLAN, R. J. (2000). Cerebral correlates of autonomic cardiovascular arousal: A functional neuroimaging investigation in humans. *J. Physiol.* (*Lond.*) **523** 259–270.

DERADO, G., BOWMAN, F. D. and KILTS, C. D. (2010). Modeling the spatial and temporal dependence in fMRI data. *Biometrics* **66** 949–957. MR2758231

DUNSON, D. B., HERRING, A. H. and ENGEL, S. M. (2008). Bayesian selection and clustering of polymorphisms in functionally related genes. *J. Amer. Statist. Assoc.* **103** 534–546. MR2523991

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

FERGUSON, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2** 615–629. MR0438568

FRISTON, K. J., HOLMES, A. P., WORSLEY, K., POLINE, P. J., FRITH, C. and FRACKOWIAK, R. (1995). Statistical parametric maps in functional imaging: A general linear approach. *Hum. Brain Mapp.* **2** 189–210.

GE, T., MÜLLER-LENKE, N., BENDFELDT, K., NICHOLS, T. E. and JOHNSON, T. D. (2014). Analysis of multiple sclerosis lesions via spatially varying coefficients. *Ann. Appl. Stat.* **8** 1095–1118. MR3262547

GELMAN, A. E. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.

GEORGE, E. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

GEORGE, E. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–373.

GOLDSMITH, J., HUANG, L. and CRAINICEANU, C. M. (2014). Smooth scalar-on-image regression via spatial Bayesian variable selection. *J. Comput. Graph. Statist.* **23** 46–64. MR3173760

GÖSSL, C., AUER, D. P. and FAHRMEIR, L. (2001). Bayesian spatiotemporal inference in functional magnetic resonance imaging. *Biometrics* **57** 554–562. MR1855691

HUANG, L., GOLDSMITH, J., REISS, P. T., REICH, D. S. and CRAINICEANU, C. M. (2013). Bayesian scalar-on-image regression with application to association between intracranial DTI and cognitive outcomes. *NeuroImage* **83** 210–223.

ISHWARAN, H. and JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96** 161–173. MR1952729

ISHWARAN, H. and ZAREPOUR, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika* **87** 371–390. MR1782485

JENKINSON, M., BANNISTER, P., BRADY, M. and SMITH, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage* **17** 825–841.

JOHNSON, T. D., LIU, Z., BARTSCH, A. J. and NICHOLS, T. E. (2013). A Bayesian non-parametric Potts model with application to pre-surgical FMRI data. *Stat. Methods Med. Res.* **22** 364–381. MR3190664

KALUS, S., SÄMANN, P. G. and FAHRMEIR, L. (2014). Classification of brain activation via spatial Bayesian variable selection in fMRI regression. *Adv. Data Anal. Classif.* **8** 63–83. MR3168680

KANG, J., JOHNSON, T. D., NICHOLS, T. E. and WAGER, T. D. (2011). Meta analysis of functional neuroimaging data via Bayesian spatial point processes. *J. Amer. Statist. Assoc.* **106** 124–134. MR2816707

KIM, S., TADESSE, M. G. and VANNUCCI, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika* **93** 877–893. MR2285077

LANG, P. J., GREENWALD, M. K., BRADLEY, M. M. and HAMM, A. O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology* **30** 261–273.

LANGE, K. (2008). *Optimization. Springer Texts in Statistics* **95**. Springer, New York.

LEWIS, P. A., CRITCHLEY, H. D., ROTSHTEIN, P. and DOLAN, R. J. (2007). Neural correlates of processing valence and arousal in affective words. *Cereb. Cortex* **17** 742–748.

LI, F. and ZHANG, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *J. Amer. Statist. Assoc.* **105** 1202–1214. MR2752615

LI, F., ZHANG, T., WANG, Q., GONZALEZ, M., MARESH, E. L. and COAN, J. A. (2015). Supplement to "Spatial Bayesian variable selection and grouping for high-dimensional scalar-on-image regression." DOI:10.1214/15-AOAS818SUPP.

MARESH, E. L., BECKES, L. and COAN, J. A. (2013). The social regulation of threat-related attentional disengagement in highly anxious individuals. *Front. Human Neurosci.* **7** 515.

MITCHELL, T. J. and BEAUCHAMP, J. J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83** 1023–1036. MR0997578

PARK, T. and CASELLA, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103** 681–686. MR2524001

PENNY, W. D., TRUJILLO-BARRETO, N. J. and FRISTON, K. J. (2005). Bayesian fMRI time series analysis with spatial priors. *NeuroImage* **24** 350–362.

RAFTERY, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika* **83** 251–266. MR1439782

REISS, P. T., MENNES, M., PETKOVA, E., HUANG, L., HOPTMAN, M. J., BISWAL, B. B., COLCOMBE, S. J., ZUO, X.-N. and MILHAM, M. P. (2011). Extracting information from functional connectivity maps via function-on-scalar regression. *NeuroImage* **56** 140–148.

REISS, P. T., HUO, L., ZHAO, Y., KELLY, C. and OGDEN, R. T. (2015). Wavelet-domain regression and predictive inference in psychiatric neuroimaging. *Ann. Appl. Stat.* **9** 1076–1101.

RUSSELL, J. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol.* **39** 1161–1178.

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. MR1309433

SMITH, M. and FAHRMEIR, L. (2007). Spatial Bayesian variable selection with application to functional magnetic resonance imaging. *J. Amer. Statist. Assoc.* **102** 417–431. MR2370843

SMITH, M. and KOHN, R. (1996). Nonparametric regression using Bayesian variable selection. *J. 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.

SMITH, S. M., JENKINSON, M., WOOLRICH, M. W., BECKMANN, C. F., BEHRENS, T. E. J., JOHANSEN-BERG, H., BANNISTER, P. R., DE LUCA, M., DROBNJAK, I., FLITNEY, D. E., NIAZY, R., SAUNDERS, J., VICKERS, J., ZHANG, Y., DE STEFANO, N., BRADY, J. M. and MATTHEWS, P. M. (2004). In advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage* **23(S1)** 208–219.

STANLEY, H. E. (1987). *Introduction to Phase Transitions and Critical Phenomena*. Oxford Univ. Press, New York.

STINGO, F. C., CHEN, Y. A., TADESSE, M. G. and VANNUCCI, M. (2011). Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. *Ann. Appl. Stat.* **5** 1978–2002. MR2884929

SUCHARD, M. A., WANG, Q., CHAN, C., FRELINGER, J., CRON, A. and WEST, M. (2010). Understanding GPU programming for statistical computation: Studies in massively parallel massive mixtures. *J. Comput. Graph. Statist.* **19** 419–438. MR2758309

TADESSE, M. G., SHA, N. and VANNUCCI, M. (2005). Bayesian variable selection in clustering high-dimensional data. *J. Amer. Statist. Assoc.* **100** 602–617. MR2160563

VANNUCCI, M. and STINGO, F. C. (2011). Bayesian models for variable selection that incorporate biological information. In *Bayesian Statistics* 9 (J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith and M. West, eds.) 659–678. Oxford Univ. Press, Oxford. MR3204022

WEST, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics* 7 (*Tenerife*, 2002) (J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, eds.) 733–742. Oxford Univ. Press, New York. MR2003537

WIECH, K., PLONER, M. and TRACEY, I. (2008). Neurocognitive aspects of pain perception. *Trends Cogn. Sci.* **12** 306–313.

WOOLRICH, M. W., JENKINSON, M., BRADY, J. M. and SMITH, S. M. (2004). Fully Bayesian spatio-temporal modeling of fMRI data. *IEEE Trans. Med. Imag.* **23** 213–231.

YUE, Y. R., LINDQUIST, M. A. and LOH, J. M. (2012). Meta-analysis of functional neuroimaging data using Bayesian nonparametric binary regression. *Ann. Appl. Stat.* **6** 697–718. MR2976488

ZHANG, T., LI, F., BECKES, L. and COAN, J. A. (2013). A semi-parametric model of the hemodynamic response for multi-subject fMRI data. *NeuroImage* **75** 136–145.

F. LI
Q. WANG
DEPARTMENT OF STATISTICAL SCIENCE
DUKE UNIVERSITY
DURHAM, NORTH CAROLINA 27708-0251
USA
E-MAIL: fli@stat.duke.edu
quanli@stat.duke.edu

T. ZHANG
DEPARTMENT OF STATISTICS
UNIVERSITY OF VIRGINIA
CHARLOTTESVILLE, VIRGINIA 22904
USA
E-MAIL: tz3b@virginia.edu

M. Z. GONZALEZ
E. L. MARESH
J. A. COAN
DEPARTMENT OF PSYCHOLOGY
UNIVERSITY OF VIRGINIA
CHARLOTTESVILLE, VIRGINIA 22904
USA
E-MAIL: mzg7uv@virginia.edu
elm2cg@virginia.edu
jcoan@virginia.edu