The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1547–1571
DOI: 10.1214/16-AOAS948

A BAYESIAN PREDICTIVE MODEL FOR IMAGING GENETICS WITH APPLICATION TO SCHIZOPHRENIA[1]

BY THIERRY CHEKOUO[*], FRANCESCO C. STINGO[†,1], MICHELE GUINDANI[‡,1] AND KIM-ANH DO[§,1]

University of Minnesota Duluth[*], University of Florence[†], University of California, Irvine[‡], and University of Texas MD Anderson Cancer Center[§]

Imaging genetics has rapidly emerged as a promising approach for investigating the genetic determinants of brain mechanisms that underlie an individual's behavior or psychiatric condition. In particular, for early detection and targeted treatment of schizophrenia, it is of high clinical relevance to identify genetic variants and imaging-based biomarkers that can be used as diagnostic markers, in addition to commonly used symptom-based assessments. By combining single-nucleotide polymorphism (SNP) arrays and functional magnetic resonance imaging (fMRI), we propose an integrative Bayesian risk prediction model that allows us to discriminate between individuals with schizophrenia and healthy controls, based on a sparse set of discriminatory regions of interest (ROIs) and SNPs. Inference on a regulatory network between SNPs and ROI intensities (ROI–SNP network) is used in a single modeling framework to inform the selection of the discriminatory ROIs and SNPs. We use simulation studies to assess the performance of our method and apply it to data collected from individuals with schizophrenia and healthy controls. We found our approach to outperform competing methods that do not link the ROI–SNP network to the selection of discriminatory markers.

1. Introduction. The advancements in neuroimaging technologies of the last two decades have contributed to an improved understanding of brain function in humans. Functional magnetic resonance imaging (fMRI) techniques in particular have been increasingly used to map neuronal activity because of their relatively low invasiveness, absence of radiation exposure and progressively broad utilization. The statistical analysis of fMRI data has focused primarily on localizing regions of the brain that are activated in response to a task, as well as determining distributed networks associated with different brain functions (brain connectivity) [Bowman (2014), Lindquist (2008)]. In this context, Bayesian approaches have been proposed to incorporate prior knowledge into the analysis, and thus help capture the complex inter-regional spatial correlations typical of fMRI data [see Zhang, Guindani and Vannucci (2015), for a review].

Received March 2015; revised March 2016.

[1]Supported in part by a Cancer Center Support Grant (NCI Grant P30 CA016672).

Key words and phrases. Imaging genetics, fMRI, data integration, Bayesian variable selection, Markov random field, nonlocal prior.

At the same time, recent developments in molecular genetics have lowered the cost of individual genetic profiling, creating the opportunity to collect massive amounts of genetic information and neuroimaging data on the same subjects. As a result, the field of imaging genetics has rapidly emerged as a promising approach for investigating the genetic determinants of the brain mechanisms that underlie an individual's behavior or psychiatric condition [Hariri and Weinberger (2003), Meyer-Lindenberg (2012)]. Ultimately, the objective is to identify specific brain activity characteristics and genetic variants that can be used as biomarkers to assist medical decision-making. However, the complexity, specificity and high-dimensionality of the data present challenges to statistical analysis. On one hand, the large number of variables calls for the use of dimension reduction techniques to identify a sparse set of relevant fMRI features or genetic covariates, leading to a problem of variable selection and multiple decision testing. On the other hand, naive multistep multivariate approaches may lead to results that are difficult to interpret and deprived of direct biological meaning, especially if these approaches cannot incorporate additional biological information at some stage of the analysis [Liu and Calhoun (2014)].

We consider a dataset comprising SNP allele frequencies and fMRI scans on 92 patients diagnosed with schizophrenia and 118 healthy controls from a study conducted by the MIND Clinical Imaging Consortium [MCIC; Chen et al. (2012), Stingo et al. (2013)]. Schizophrenia is often characterized as a disorder of brain connectivity, with symptoms that usually develop slowly over months or years. There are currently no medical tests to diagnose schizophrenia. The diagnosis is typically made based on an interview of the person and family members. The limitations of such symptom-based assessment have long been known in the literature [Weiss (1989); and, more recently, Jacob (2013)]. For early diagnosis and targeted treatment, objective markers to guide clinical practice are highly desirable. In this respect, neuroimaging methods have been widely used to investigate functional brain networks associated with the disease [see, for recent reviews, Calhoun and Hugdahl (2012), Fornito et al. (2012)]. Many studies have also shown that genetic alterations at the mRNA and SNP levels also play important roles in schizophrenia [see, e.g., Chen et al. (2012), Lencz et al. (2007)]. As a matter of fact, Potkin et al. (2015) recently pointed out that focusing on brain imaging data in neuropsychiatry without considering the genetic component may lead to neglecting a huge component of the risk for developing schizophrenia. Therefore, there's a need for integrative models that can identify genetic variants and imaging-based biomarkers associated with the disease [Cao et al. (2013)]. At the same time, for practical clinical relevance and diagnostic purposes, it is essential to develop risk prediction models which can further provide an assessment of an individual probability of being affected by schizophrenia.

Statistical approaches for the analysis of imaging genetics data can be classified as follows. A first group of methods aims to investigate the association between genetic markers and imaging endophenotypes, which are defined as brain imaging

features that highlight the causal links between genes and the phenotypic expression of disorders [Cannon and Keller (2006)]. For this purpose, a variety of techniques, such as group sparse regularization, multifactor dimensionality reduction, principal component analysis, generalized low-rank regression, functional-mixed effects models, independent component analysis (ICA) and clustering methods are commonly employed [Chi et al. (2013), Floch et al. (2012), Hardoon et al. (2009), Lin et al. (2014), Liu et al. (2009), Meda et al. (2012), Vounou, Nichols and Montana (2010), Vounou et al. (2012), Wang et al. (2012b), Zhu et al. (2014)]. A second group of methods employs a two-step approach to detect prognostic markers: first, new genes related to schizophrenia are discovered by studying their association with selected imaging endophenotypes, and then the relevant biomarkers are validated by assessing their association with the disease, often by fitting a simple frequentist logistic model [Chen et al. (2012), Lin, Calhoun and Wang (2014), Potkin et al. (2009)]. A third group of approaches comprises predictive frequentist methods that combine both SNPs and imaging biomarkers as covariates in a single model [Cao et al. (2013), Wang et al. (2012a), Yang et al. (2010)]. In particular, several regularization methods [e.g., L1-Lasso by Friedman, Hastie and Tibshirani (2010), Elastic Net by Zou and Hastie (2005), L1-Support vector machine by Zhang et al. (2006)], as well as boosting algorithms [e.g., LogitBoost, Dettling and Bühlmann (2003)], can be used to identify imaging and/or genetics covariates in the context of binary classification. Other methods, such as the sparse group lasso [Simon et al. (2013)], multikernel and deep network learning methods [Deng and Yu (2013), Sonnenburg et al. (2006)], have been effectively applied in the context of disease prediction [Filipovych, Resnick and Davatzikos (2011), Wang et al. (2012a), Zhang, Huang and Shen (2014)], and belong to this third group of approaches. Finally, a few integrative Bayesian methods have been proposed [Batmanghelich et al. (2013), Stingo et al. (2013)]; these approaches use a hierarchical modeling framework to identify genetic variants associated with disease through the mediation of selected image-based features.

In this manuscript, we propose a more encompassing integrative Bayesian risk prediction model that allows us to discriminate between schizophrenic patients and healthy controls based on a sparse set of discriminatory brain regions of interest (ROIs) and SNPs. More specifically, the model allows for the identification of a regulatory network between SNPs and ROI intensities, thus exploiting the imaging features as an intermediate phenotype. Inference on the ROI–SNP associations is then used, in a single modeling framework, to inform the selection of a set of discriminatory ROIs and SNPs that are mostly associated with an increased probability of schizophrenia through the use of variable selection priors dependent on the inferred regulatory network. For this purpose, we introduce an innovative covariate-dependent Markov random field (MRF) prior to guide the selection of the discriminatory ROIs, whereby the selection probabilities take into account that ROIs highly connected in the ROI–SNP network are likely to be also associated with the clinical outcome, in addition to the spatial dependencies among the ROIs.

We are able to achieve sharper biomarker selection through the specification of nonlocal prior distributions [Johnson and Rossell (2010, 2012)] on the regression coefficients of both the ROI–SNP network and the discriminatory ROIs and SNPs.

With respect to the alternative methods outlined previously, our approach has several advantages. On the one hand, predictive frequentist methods can not easily model the interaction between the two data modalities, so to use the imaging features as intermediate endophenotypes. Often, they also do not account for the spatial dependency among ROIs or imaging features. Furthermore, our novel Bayesian approach is more comprehensive than existing Bayesian methods. In particular, the method by Stingo et al. (2013) is of limited assistance in practical medical decision-making since it fails to directly link genetic and imaging data with the clinical outcome in a convenient risk prediction framework. As a matter of fact, our modeling framework improves on the approaches previously mentioned by assuming that: (i) genetic factors may affect nondiscriminatory brain regions (as endophenotypes); and that (ii) genetic factors may be independently associated with disease status without the mediation of a discriminatory imaging endophenotype. Finally, our method provides a direct assessment of the individual probability of being affected by schizophrenia as a function of the observed fMRI and SNP biomarkers, which can be used to inform targeted therapies. As a matter of fact, by modeling the relationship between a scalar discrete response outcome and a high-dimensional image predictor, our model can also be seen as an extension of recently proposed scalar-on-image regression models [Goldsmith, Huang and Crainiceanu (2014), Goldsmith et al. (2012), Li et al. (2015)] to the more challenging setting of imaging genetics.

In Section 2, we illustrate our proposed modeling approach. In Section 3, we discuss posterior inference and show how our modeling framework can be used to predict the disease status of future subjects, and then aid the diagnosis of schizophrenia. Section 4 presents a few simulation studies in which we found our approach to outperform competing methods that do not link the ROI–SNP network to the selection of discriminatory markers. In Section 5, we discuss the results of our inference on the dataset of schizophrenic patients and healthy controls. Section 6 concludes the paper.

2. Bayesian model specification. Our primary goal is to define a predictive model that accurately predicts the disease status of a subject based on their brain activity (fMRI) and genetic profile (SNPs). In Section 2.1, we introduce a Bayesian linear model that relates the observed ROI intensities and SNP allele frequencies to a binary indicator of the disease status. We then use a direct acyclic graph (DAG) to infer a regulatory network between SNPs and ROIs (Section 2.2) and incorporate this DAG in a novel prior that guides the selection of discriminatory ROIs and SNPs (Section 2.3).

2.1. *A predictive model for disease status*. We represent the disease status of n subjects by an $n \times 1$ binary vector $\mathbf{y}$, such that $y_i = 1$ if subject i has been diagnosed with schizophrenia and $y_i = 0$ if subject i is a healthy control. The brain activity is represented by an $n \times G$ matrix $\mathbf{X}$ of ROI-based summaries of BOLD signal intensity as measurements on a set of G ROIs. We denote the set of M available genetic covariates (SNPs) by an $n \times M$ matrix $\mathbf{Z}$. For each SNP, the genotype is coded by the number of minor alleles.

We consider a probit model to relate the ROIs and SNPs to the binary response variable. Alternatively, a logistic model could be employed at the expense of an increased computational cost. In particular, the probit formulation allows to adopt the data augmentation approach of Albert and Chib (1993) by introducing an auxiliary latent variable $\mathbf{y}^\star$, and to express the probit binary regression model on the clinical outcome y_i as a gaussian linear regression model on the auxiliary variables. Hence, our predictive model is defined as

$$\mathbf{y}^\star = \mathbb{1}_n \beta_0 + \mathbf{Z}\boldsymbol{\beta}^{(1)} + \mathbf{X}\boldsymbol{\beta}^{(2)} + \boldsymbol{\nu}, \tag{1}$$

where

$$y_i = \begin{cases} 1, & \text{if } y_i^\star > 0, \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

$\mathbb{1}_n$ is the unit vector of dimension n, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)^T \sim N(0, \mathbf{I}_n)$ is an error term, and $\mathbf{I}_n$ is the identity matrix of size $n \times n$.

Most likely, only a subset of ROIs and SNPs can discriminate between schizophrenia cases and healthy controls among the n subjects. We select discriminatory SNPs and ROIs as biomarkers through the introduction of two binary vectors, $\boldsymbol{\gamma}^{(1)} = (\gamma_1^{(1)}, \ldots, \gamma_M^{(1)})$ and $\boldsymbol{\gamma}^{(2)} = (\gamma_1^{(2)}, \ldots, \gamma_G^{(2)})$, with $\gamma_m^{(1)} = 1$ if SNP m is included in the model and $\gamma_m^{(1)} = 0$ otherwise; similarly, $\gamma_g^{(2)} = 1$ if ROI g is included in the model and $\gamma_g^{(2)} = 0$ otherwise. We use the latent vectors $\boldsymbol{\gamma}^{(1)}$ and $\boldsymbol{\gamma}^{(2)}$ to specify the prior on each regression coefficient in (1) as a scale mixture of a product moment prior (pMOM), described by Johnson and Rossell (2012), and a point mass at zero,

$$p(\beta_m^{(1)} | \boldsymbol{\gamma}^{(1)}, h_1, r) = \gamma_m^{(1)} \mathcal{PM}(\beta_m^{(1)}; r, h_1, 1) + (1 - \gamma_m^{(1)}) \mathcal{I}_0(\beta_m^{(1)}), \tag{3}$$

where $\mathcal{PM}(\beta_m^{(1)}; r, h_1, \sigma^2)$ denotes the pMOM density of parameters r, h_1 and σ^2. The prior on $\boldsymbol{\beta}^{(2)}$, $p(\beta_g^{(2)} | \boldsymbol{\gamma}^{(2)}, h_2, r)$ is defined similarly. A pMOM prior has the following probability density function:

$$\mathcal{PM}(\beta; r, h, \sigma^2) = \frac{1}{(2r-1)!!} (2\pi)^{-0.5} \frac{\beta^{2r}}{(h\sigma^2)^{r+0.5}} \exp\left\{ -\frac{\beta^2}{2h\sigma^2} \right\}, \tag{4}$$

with support on the real line. The parameter r characterizes the order of the distribution and h determines the dispersion around zero. Large values of h correspond

to a prior that is well spread out over the parameter space, and typically encourages the selection of relatively large effects. The pMOM distribution is symmetric at zero and gives a low prior probability to coefficients close to 0, thus eliminating regression models that contain unnecessary explanatory variables, a property which is common to the *nonlocal* prior distributions [Johnson and Rossell (2010, 2012)]. Finally, we assume a pMOM prior on the intercept term β_0 by $p(\beta_0|h_0, r) = \mathcal{PM}(\beta_0; r, h_0, 1)$, with scale parameter h_0.

2.2. *A DAG approach for the ROI–SNP network.* In this section, we aim to identify a regulatory network in which SNPs can affect ROI intensities. This modeling strategy, besides providing interesting insights into the biological mechanisms that characterize schizophrenic patients, yields critical information that will be incorporated into our biomarker selection procedure (see Section 2.3) and will lead to satisfactory prediction performances (see Section 4). We model the ROI–SNP network as a DAG. We assume that each ROI $\mathbf{x}_g$ can be affected only by the SNPs, that is, the arrows can only go from SNPs into ROIs. We assume that the ROIs are independent conditionally upon the SNPs, that is, $\mathbf{x}_g \perp\!\!\!\perp \mathbf{x}_{g'}|\mathbf{Z}$. The likelihood of a DAG can be written as a system of linear equations, where each regression corresponds to an ROI that is potentially affected by all the SNPs:

$$\mathbf{x}_g = \mathbf{Z}\boldsymbol{\beta}_g^{(3)} + \varepsilon_g, \qquad g = 1, \ldots, G, \tag{5}$$

with $\boldsymbol{\varepsilon}_g = (\varepsilon_{1g}, \ldots, \varepsilon_{ng})^T \sim N(0, \sigma_g \mathbf{I}_n)$ indicating the error term. For each ROI, we are interested in identifying a small number of explanatory genetic factors. This goal can be accomplished via a variable selection approach by introducing a binary matrix variable, $\boldsymbol{\Gamma}^{(3)} = (\boldsymbol{\gamma}_1^{(3)}, \ldots, \boldsymbol{\gamma}_G^{(3)})^T = (\gamma_{gm}^{(3)})_{G\times M}$, with $\gamma_{gm}^{(3)} = 1$ if SNP m is related to ROI g and $\gamma_{gm}^{(3)} = 0$, otherwise. Given $\boldsymbol{\Gamma}^{(3)}$, each component of $\boldsymbol{\beta}_g^{(3)}$ follows the mixture distribution (3) with parameters r, h_2 and σ_g^2 on the pMOM density. The matrix $\boldsymbol{\Gamma}^{(3)}$ defines the *ROI–SNP network*. We assume conjugate inverse-gamma priors for the error variances $\sigma_g^2 \sim \text{Inv-Gamma}(\alpha, \psi)$. Further, we capture ROI connectivity via the MRF prior defined in Section 2.3.3.

2.3. *An integrative approach via variable selection priors.* The selection of relevant predictors is a key step in defining a predictive model. In our Bayesian model, three sets of binary indicators define the inclusion of ROI–SNP connections ($\boldsymbol{\Gamma}^{(3)}$), discriminatory SNPs ($\boldsymbol{\gamma}^{(1)}$) and discriminatory ROIs ($\boldsymbol{\gamma}^{(2)}$). We specify innovative prior distributions on these parameters that encourage sparsity, relate the ROI–SNP network to the selection of discriminatory markers and account for the ROI spatial dependencies.

2.3.1. *Selection of the ROI–SNP network.* The binary variable $\gamma_{gm}^{(3)}$ indicates whether there is a relationship between ROI g and SNP m. We assume a Bernoulli

prior on $\gamma_{gm}^{(3)}$, with parameter q_g defined as

$$p(\mathbf{\Gamma}^{(3)}|\mathbf{q}) = \prod_{g=1}^{G}\prod_{m=1}^{M} q_g^{\gamma_{gm}^{(3)}}(1-q_g)^{1-\gamma_{gm}^{(3)}}. \tag{6}$$

The hyperparameters q_g's define the prior probability that is assigned to a connection between ROI g and any given SNP. In our application, we fix these parameters at a small value to encourage the selection of sparse networks. Alternatively, we could place a Beta hyperprior on q_g, yielding an automatic multiplicity penalty since the posterior distribution of q_g will become more concentrated at small values near 0 as the total number of variables increases [Scott and Berger (2010)].

2.3.2. *Selection of discriminatory SNPs.* For the latent SNP selection indicator $\boldsymbol{\gamma}^{(1)}$, we specify a prior distribution that accounts for the ROI–SNP regulatory network as defined by the matrix $\mathbf{\Gamma}^{(3)}$. This prior defines a probabilistic dependency between the ROI–SNP network and the clinical outcome. We model the SNP selection indicators $\boldsymbol{\gamma}^{(1)}$ as a function of $\mathbf{\Gamma}^{(3)}$:

$$P(\boldsymbol{\gamma}^{(1)}|\mathbf{\Gamma}^{(3)}, \nu_1, \tau_1) \propto \exp\bigl(\nu_1 \mathbb{1}_M^T \boldsymbol{\gamma}^{(1)} + \tau_1 \mathbb{1}_G^T \mathbf{\Gamma}^{(3)} \boldsymbol{\gamma}^{(1)}\bigr). \tag{7}$$

The elements of $\boldsymbol{\gamma}^{(1)}$ are stochastically independent given $\mathbf{\Gamma}^{(3)}$, ν_1 and τ_1; for a given SNP m, the inclusion probability is then defined as

$$P(\gamma_m^{(1)} = 1|\mathbf{\Gamma}^{(3)}, \nu_1, \tau_1) = \frac{\exp(\nu_1 + \tau_1 \sum_{g=1}^{G} \gamma_{gm}^{(3)})}{1 + \exp(\nu_1 + \tau_1 \sum_{g=1}^{G} \gamma_{gm}^{(3)})}. \tag{8}$$

This inclusion probability is an increasing function of the number of ROIs connected to each SNP, so as to reflect our hypothesis that SNPs involved in the regulatory network are more likely to be significantly correlated with the clinical outcome. The parameter ν_1 controls the prior inclusion probability and is set to a fixed value that encourages sparsity on the SNP selection. The parameter τ_1 measures the effect of the ROI–SNP network on the SNP selection. When $\tau_1 = 0$, this prior distribution reduces to an independent Bernoulli with probability of success $\exp(\nu_1)/[1 + \exp(\nu_1)]$, the logistic transformation of ν_1. We assume a truncated Normal distribution on τ_1, truncated at zero with mean 0 and variance $\sigma_{\tau_1}^2$. If the data support our hypothesis that SNPs involved in the ROI–SNP network are more likely to be associated with the clinical outcome, we will expect to observe a posterior distribution of τ_1 that gives small probability to values close to zero.

2.3.3. *Selection of discriminatory ROIs.* We define a prior on the ROI binary indicators $\boldsymbol{\gamma}^{(2)}$ that captures two main features of our data: (1) ROIs highly connected in the ROI–SNP network are more likely associated with the clinical outcome; and (2) ROIs located in adjacent areas of the brain show a high level of correlation.

More specifically, we further generalize the prior presented in Section 2.3.2 to include a spatial process on $\boldsymbol{\gamma}^{(2)}$ that takes into account network dependencies within the ROIs so that connected ROIs are more likely to be selected together. We specify the ROI spatial network as follows: two ROIs, g and g', are connected by an edge if the distance between them is among the kth smallest distances from g to other ROIs. We denote by N_g the set of neighbors of ROI g.

Following Besag (1974), we define a symmetric matrix, $\mathbf{B}$, that captures the spatial dependencies between ROIs as follows:

$$b_{gg'} = \exp\left\{-\frac{d(g,g')^2}{2\sigma_r^2}\right\} \qquad \text{if } g' \in N_g \text{ and 0 otherwise,}$$

where $d(g,g')$ is the Euclidean distance between ROIs g and g', computed using the physical location of the ROIs in the brain. In our application, σ_r^2 was chosen to be equal to the average nearest neighbor square distance. We model spatial dependencies via a covariate-dependent MRF prior on the $\gamma_g^{(2)}$'s, defined by

$$\begin{aligned} &P\big(\boldsymbol{\gamma}^{(2)}|\boldsymbol{\Gamma}^{(3)}, \nu_2, \tau_2, \eta_2\big) \\ &\quad\propto \exp\left\{\big(\nu_2 \mathbb{1}_G^T + \tau_2 \mathbb{1}_M^T \boldsymbol{\Gamma}^{(3)T}\big)\boldsymbol{\gamma}^{(2)} + \eta_2 \sum_{g,g'} b_{gg'} \mathcal{I}\big(\gamma_g^{(2)} = \gamma_{g'}^{(2)}\big)\right\}. \end{aligned} \tag{9}$$

Given the binary variables $\gamma_{g'}^{(2)}$, the probability of $\gamma_g^{(2)}$ is an increasing function of the number of SNPs connected to ROI g:

$$\begin{aligned} &P\big(\gamma_g^{(2)}|\boldsymbol{\Gamma}^{(3)}, \big(\gamma_{g'}^{(2)}\big)_{g'\in N_g}\big) \\ &\quad\propto \exp\left(\nu_2\gamma_g^{(2)} + \tau_2 \sum_{m=1}^{M} \gamma_{gm}^{(3)}\gamma_g^{(2)} + 2\eta_2 \sum_{g'\in N_g} b_{gg'} \mathcal{I}\big(\gamma_g^{(2)} = \gamma_{g'}^{(2)}\big)\right). \end{aligned} \tag{10}$$

The parameter ν_2 controls the sparsity of the model, τ_2 measures the effect of the number of SNPs connected to the ROIs, and η_2 controls the strength of the connections in the ROI spatial network. High values of η_2 encourage neighboring features to take on the same $\gamma_g^{(2)}$ value. MRF priors are commonly used to account for spatial dependencies between variables measured in nearby locations, such as neighboring ROIs. The novel MRF prior (9) can be seen as an improved tool, as it accounts for both spatial dependencies and the ROI–SNP network.

We treat ν_2 and η_2 as fixed hyperparameters [Li and Zhang (2010), Stingo, Vannucci and Downey (2012)]. It is known that allowing η_2 to vary can lead to a phase transition problem, that is, the number of $\gamma_g^{(2)} = 1$ undergoes a dramatic change given an infinitesimal change in η_2 [Li and Zhang (2010), Stingo et al. (2011)]. Even if η_2 is fixed in this study, our covariate-dependent MRF includes an additional stochastic parameter τ_2 that links the ROI–SNP network to the biomarker

selection, and a phase transition may occur if the prior on τ_2 is not carefully specified. For that reason, we assume τ_2 to follow a normal distribution with 0 as mean and variance $\sigma^2_{\tau_2}$, truncated on the left by 0 and on the right by r_τ, which represents a problem-dependent phase transition threshold parameter. In Section B of the Supplementary Material, we provide guidelines and an illustrative example to demonstrate how to detect a phase transition value.

3. Posterior inference. For posterior inference, our primary interest is in the estimation of the selection indicators $(\boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)})$, the ROI–SNP network $\boldsymbol{\Gamma}^{(3)}$, and the parameters τ_1 and τ_2, which define the degree of influence of the ROI–SNP network on the selection of the predictive markers. We use a Metropolis–Hastings within Gibbs sampler to sample from the posterior distribution of these parameters. The computational efficiency of our stochastic search algorithm can be improved by integrating the regression coefficients β_0, $\boldsymbol{\beta}^{(1)}$, $\boldsymbol{\beta}^{(2)}$, $\boldsymbol{\beta}^{(3)}$ and the variance parameter $\boldsymbol{\sigma}$ using a standard Laplace approximation [Johnson and Rossell (2012)]. (See Section A in the Supplementary Material for more details.) Therefore, we focus on the marginal posterior distribution of $(\boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)}, \boldsymbol{\Gamma}^{(3)}, \tau_1, \tau_2)$.

The resulting stochastic search Markov chain Monte Carlo (MCMC) algorithm efficiently explores the model space and can quickly find the most probable set of covariates with high posterior probabilities, while spending less time in regions with low posterior probabilities [George and Mcculloch (1997), Stingo et al. (2010)]. Stochastic search variable selection approaches are known to have greater accuracy in binary regression models than standard variable selection methods such as forward, backward or stepwise selection [Swartz, Yu and Shete (2008)].

3.1. *MCMC sampling.* Our algorithm comprises four Metropolis–Hastings (M-H) steps (1)–(4) and a nontrivial step (5) in which samples are drawn from a doubly-intractable distribution, that is, a distribution with an unknown normalizing constant that depends on the sampled parameter.

Here, we succinctly describe our MCMC algorithm:

1. The binary variable selection parameters $\boldsymbol{\gamma}^{(1)}$, $\boldsymbol{\gamma}^{(2)}$ are updated using separate M-H steps.
   A new value for $\boldsymbol{\gamma}^{(1)}$ (or $\boldsymbol{\gamma}^{(2)}$) is proposed as follows: we randomly choose between changing the value of a single component, from 0 to 1 or from 1 to 0, and swapping two components with opposite values. This step explores the model space in order to find relevant SNPs and ROIs, respectively.
2. The parameter $\boldsymbol{\Gamma}^{(3)}$ is updated using M-H steps based on the same type of moves defined for $\boldsymbol{\gamma}^{(1)}$ and $\boldsymbol{\gamma}^{(2)}$. This step explores the model space that defines the ROI–SNP network.
3. An M-H step is used to update the latent variables, $y_n^\star$'s. A $y_n^{\star\text{new}}$ is proposed by an exponential distribution with scale parameter $1/y_n^{\star\text{old}}$, $y_n^{\star\text{new}} \sim \text{Exp}(1/y_n^{\star\text{old}})$ if $y_n = 1$, and $y_n^{\star\text{new}} \sim -\text{Exp}(-1/y_n^{\star\text{old}})$ if $y_n = 0$.

4. An M-H step is used to update τ_1 in (7): a τ_1^{new} is proposed from a truncated normal with mean τ_1^{old} and truncation at 0. The variance of this distribution (before truncation), $h_{\tau_1}^1$, represents a tuning parameter to be set so as to facilitate exploring the parameter space and to induce a good acceptance rate.
5. The interaction parameter τ_2 in (9) is sampled from a doubly-intractable distribution.
   We draw τ_2 from the following density:

$$
\begin{aligned}
p(\tau_2|\boldsymbol{\gamma}^{(2)}, \boldsymbol{\Gamma}^{(3)}) &\propto p(\boldsymbol{\gamma}^{(2)}|\boldsymbol{\Gamma}^{(3)}, \tau_2)p(\tau_2) \\
&\propto \exp\{(\nu_2 \mathbb{1}_G^T + \tau_2 \mathbb{1}_M^T \boldsymbol{\Gamma}^{(3)T})\boldsymbol{\gamma}^{(2)}\}\frac{p(\tau_2)}{Z(\tau_2)},
\end{aligned} \tag{11}
$$

   with $Z(\tau_2)$ being the normalizing constant of the prior distribution of $\boldsymbol{\gamma}^{(2)}$, which is not available analytically. The M-H acceptance ratio depends on two unknown normalizing constants, $Z(\tau_2)$ and $Z(\tau_2^{\text{new}})$. To bypass this issue, we adapt the approach proposed by Atchadé, Lartillot and Robert (2013) to sample from our integrative prior (9). Technical details are presented in Section B of the Supplementary Material.

The MCMC sampler results in lists of the included biomarkers (ROIs, SNPs and ROI–SNP pairs), together with their posterior probabilities. Important biomarkers can be selected by assessing the marginal posterior probabilities $p(\boldsymbol{\gamma}^{(i)}|\mathbf{y}, \mathbf{X}, \mathbf{Z})$, $i = 1, 2$, and $p(\boldsymbol{\Gamma}^{(3)}|\mathbf{y}, \mathbf{X}, \mathbf{Z})$, estimated from the relative frequency of the inclusion of each biomarker in the models visited by the MCMC sampler. Samples from the posterior distribution of τ_1 and τ_2 can be used to infer the effect of the ROI–SNP pairs on the selection of the discriminatory biomarkers.

3.2. *Classification of future cases.* We can use N_{new} further measurements $\mathbf{X}_{\text{new}}$ and $\mathbf{Z}_{\text{new}}$ to predict disease status $\mathbf{y}_{\text{new}}$ for new subjects. We standardize $\mathbf{X}_{\text{new}}$ and $\mathbf{Z}_{\text{new}}$ using the mean and variance from the training data. The latent variables $\mathbf{y}_{\text{new}}^{\star}$ are predicted using a Bayesian model averaging approach [Sha et al. (2004)]:

$$
\hat{\mathbf{y}}_{\text{new}}^{\star} = \sum_{(\boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)})} (\mathbb{1}_n \tilde{\beta}_0 + \mathbf{Z}_{\text{new}} \tilde{\boldsymbol{\beta}}^{(1)} + \mathbf{X}_{\text{new}} \tilde{\boldsymbol{\beta}}^{(2)}) p(\boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)}|\hat{\mathbf{y}}^{\star}, \mathbf{X}, \mathbf{Z}, \hat{\theta}), \tag{12}
$$

where $\hat{\theta} = (\hat{\tau}_1, \hat{\tau}_2, \hat{\boldsymbol{\Gamma}}^{(3)})$ is a posterior estimate of $\theta = (\tau_1, \tau_2, \boldsymbol{\Gamma}^{(3)})$ and $\hat{\Theta} = (\tilde{\beta}_0, \tilde{\boldsymbol{\beta}}^{(1)T}, \tilde{\boldsymbol{\beta}}^{(2)T})^T$ is the posterior mode of $\Theta = (\beta_0, \boldsymbol{\beta}^{(1)T}, \boldsymbol{\beta}^{(2)T})^T$. The latent variable $\mathbf{y}^{\star}$ is set to the mean $\hat{\mathbf{y}}^{\star}$ of the $\mathbf{y}^{\star}$'s, sampled during the MCMC algorithm. The summation in equation (12) is performed over the l models that have the highest posterior probability [Madigan and Raftery (1994)]. Given $\mathbf{y}_{\text{new}}^{\star}$, the corresponding predicted disease binary indicators $\mathbf{y}_{\text{new}}$ for the N_{new} patients can be computed via equation (2) [Sha et al. (2004)]. The predictive probabilities of

disease status can be computed as $\hat{p}(y_i = 1|\mathbf{X}, \mathbf{Z}) \approx \Phi(\hat{y}_i^\star)$, where Φ is the normal cumulative distribution function. An area under the receiver operating characteristics curve (AUC) statistic can be computed using these probabilities on the validation set (denoted as AUCp).

4. Simulation study. We investigate the performance of our method using simulated data that mimic the characteristics of the MCIC schizophrenia data. The goal of this analysis is threefold: identify the predictive biomarkers, reconstruct the ROI–SNP network, and correctly predict the disease status of the subjects in the validation set.

We generated training and validation sets of 168 and 42 samples, respectively. We retrieved the matrix $\mathbf{Z}$ of the observed SNP data from the MCIC schizophrenia study. This approach ensures that the realistic pattern of correlation across the SNPs is preserved. The ROI–SNP network $\mathbf{\Gamma}^{(3)}$ is generated from an independent Bernoulli distribution, with probability chosen uniformly between 1% and 5%, which on average results in 282 connections (3%) included in the network. The regression coefficients of the connections are set to either 1 or -1; $\beta_{gm}^{(3)} \in \{1, -1\}$ if $\gamma_{gm}^{(3)} = 1$, and 0 otherwise. The matrix $\mathbf{X}$ of $G = 116$ ROIs is generated using independent multivariate distributions as follows:

$$x_{ig} = \sum_{m=1}^{M} z_{im}\beta_{gm}^{(3)} + \varepsilon_{ig}, \tag{13}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_{ig}) \sim \mathcal{N}_{n\times G}(0_{n\times G}, I_n \otimes \Omega^{-1})$ and Ω is the precision matrix and defined as a banded matrix of bandwidth 3, that is, $[\Omega]_{g,g'} = 0$ for $|g - g'| > 3$, $[\Omega]_{g,g'} = \rho$ otherwise for $g \neq g'$, and 1 on the diagonal. This induces a decomposable graph [Zhang, Wiesel and Greco (2013)] with exactly 800 edges, which is equivalent to a precision matrix Ω with only 4% nonzero entries. The parameter ρ defines the degree of partial correlation between the ROIs. When $\rho \neq 0$, the data-generating process of the ROIs differs from our proposed model, in which the ROIs are assumed to be independent in the ROI–SNP network model. We first investigated the performance of our approach on two cases: $\rho \in \{0.1, 0.2\}$. To ensure that our simulation studies encompassed realistic ROI correlation patterns, we also investigated a scenario in which Ω is set as a sparse precision matrix, $\hat{\Omega}_{\mathrm{GL}}$, estimated from the observed ROIs via graphical lasso [Friedman, Hastie and Tibshirani (2008)].

In our simulation studies, we investigated how the prediction and selection performances are affected by the effective size of the association with the predictive biomarkers. Specifically, in *scenario* 1, we set the absolute value of the nonzero coefficients $\beta_m^{(1)}$ and $\beta_g^{(2)}$ to either 0.5, 1 or 1.5. We set the intercept term to $\beta_0 = 1$. We selected 6 ROIs and 5 SNPs to be associated with the clinical outcome. In this scenario, all selected ROIs and SNPs are highly connected in the ROI–SNP network, that is, they all have a large number of connections (6–8), or degrees, in

the ROI–SNP graph. Also, 3 ROIs and 2 SNPs are set to be negatively associated with the clinical endpoint. The analysis of more challenging scenarios, that is, with small effect sizes, is of particular interest in our application context since the SNPs identified from either genome-wide association studies (GWAS) or candidate gene studies have often been shown to explain only a small part of heritability [Shahbaba, Shachaf and Yu (2012)].

In a second simulation study, we considered 3 additional scenarios that allow us to understand how the performances of our models change with respect to the degree of connectivity of the discriminatory markers. In *scenario* 2, we selected 3 ROIs and 3 SNPs that are highly connected (degrees between 6 and 8) in the ROI–SNP network as well as 2 SNPs and 3 ROIs that are less connected in the same network, that is, having at most 1 connection. In *scenario* 3, we selected only 1 ROI and 1 SNP that are highly connected, with degrees 7 and 8 respectively, and 5 ROIs and 4 SNPs that are less connected, with degrees either 0 or 1. In *scenario* 4, we selected 11 markers, all of which had zero or very small degrees (1 or 2) in the ROI–SNP network. The binary outcome and the underlying latent variable are obtained by employing equations (2) and (1). Setting the hyperparameters is discussed in Section C of the Supplementary Material.

4.1. *Results*. To highlight the importance of incorporating the ROI–SNP network in our model, we compare our integrative Bayesian approach for imaging genetics (iBIG) with a Bayesian two-step approach (BTS), defined by setting $\tau_1 = \tau_2 = 0$. BTS fails to link the selection of ROI–SNP pairs to the selection of the discriminatory ROIs and SNPs, and then independently fits the two stages (ROI–SNP network and clinical predictive model) that define our approach. Both methods were run for 80,000 MCMC iterations with a burn-in period of 30,000 iterations to allow our adaptive MCMC algorithm to converge; iBIG and BTS were fitted using the same parameter settings. To assess the convergence of the MCMC algorithm, we ran two MCMC chains for each case with randomly chosen starting points. Details on the MCMC diagnostics are given in Section E of the Supplementary Material.

We further compared our methods with three alternative approaches: the sparse-group lasso (*sGroup-Lasso*) binomial regression of Simon et al. (2013), the L1 support vector machine (*L1-SVM*) of Zhang et al. (2006) and the neural network with model averaging (*avNNnet*) of Ripley (1996). The penalty parameters for *sGroup-Lasso* and L1-SVM were chosen following a 10-fold and 5-fold cross-validation approach as suggested and implemented in the R packages *SGL* and *penalizedSVM*, respectively. The R package *caret* was used to fit the avNNnet method. We also compared our approach with the penalized L1 Lasso [*L1-Lasso*, Friedman, Hastie and Tibshirani (2010)], elastic net [*Net-Lasso*, Zou and Hastie (2005)], multikernel [*MKL*, Sonnenburg et al. (2006)] and deep learning methods [*DNN*, LeCun, Bengio and Hinton (2015)]. The results of these additional comparisons are presented in the Supplementary material Section H. We computed

the F1-measure to assess the performances in terms of variable selection: the F1-measure is defined as $F1 = 2PR/(P + R)$, that is, as the harmonic mean of the precision (P) and the recall, or sensitivity, (R). The F1-measure is commonly used in the presence of a very skewed class imbalance, such as variable selection in high-dimensional sparse data. All three alternative methods yield a list of selected covariates and predict the disease status for the units in the validation set. Concerning our methods, all covariates with marginal posterior probability greater than 0.5 were selected. To evaluate the prediction performance, we reported the AUC and the misclassification error rate (MCE) on the validation set following the strategy defined in Section 3.2.

Both iBIG and BTS perfectly reconstructed the ROI–SNP network, $\mathbf{\Gamma}^{(3)}$, with the AUCs for variable selection close to 1 for all simulation scenarios. Hence, both methods can accurately reconstruct the ROI–SNP network even when partial correlations are of small magnitude. Table 1 shows the variable selection and prediction performances for all competing methods across 9 scenarios: iBIG always performs best, both in terms of variable selection (much higher F1) and prediction (larger AUCp and smaller MCE). As the absolute values of the "effect sizes" increase from 0.5 to 1 or 0.5, the performances of our method improved for all levels of the correlation structure between ROIs (ρ). Results from the second simulation study are summarized in Table 2; for all simulation scenarios, data were generated setting $\rho = 0.1$ and the absolute value of the regression coefficients to 1. The set of discriminatory biomarkers differs between scenarios, and a fair performance comparison can be done only within each scenario. iBIG performs much better than the competing methods when there is a large number of discriminatory biomarkers that are also highly connected in the ROI–SNP network. As expected, we did not observe a significant difference between iBIG and BTS in scenarios 3 and 4, as the number of relevant biomarkers highly connected in the network is only 2 and 0, respectively. Figure 1 shows the density kernel plots of the posterior distribution of the network effect parameters τ_1 and τ_2 for each scenario. The mode of these distributions increased with the number of discriminatory biomarkers involved in the network, which highlights the importance of our prior (9) that relates, through τ_1 and τ_2, the inferred ROI–SNP network to the selection of the discriminatory biomarkers. Note that our approach accounts for the uncertainty on the estimation of the ROI–SNP network, and can flexibly separate the two stages of our approach by assigning values to τ_1 and τ_2 that are close to zero when suggested by the data (such as in scenarios 3 and 4).

5. The MCIC schizophrenia dataset. We consider a study on schizophrenia conducted by the MCIC, which is a multi-institutional effort to apply neuroimaging techniques in the study of mental illnesses and brain disorders. The study aimed to identify neural markers for disease onset by using functional imaging combined with clinical characterization and genomic analysis. The data we have

TABLE 1

Simulation results—scenario 1 (30 *replicates*). *Coeff is the effect size in absolute value*, *F*1 *is the F*1*-measure*, *AUCp is the predictive AUC*, *and MCE is the misclassification error. The best values are shown in boldface*

Method	ρ	Coeff	F1	AUCp	MCE
avNNnet	0.1	0.5	0.150 (0.061)	0.759 (0.092)	0.327 (0.089)
sGroup-Lasso	0.1	0.5	0.342 (0.054)	0.814 (0.075)	0.282 (0.070)
L1-SVM	0.1	0.5	0.241 (0.071)	0.812 (0.102)	0.254 (0.074)
iBIG	0.1	0.5	**0.564 (0.120)**	**0.889 (0.051)**	**0.194 (0.047)**
BTS	0.1	0.5	0.340 (0.141)	0.738 (0.122)	0.316 (0.093)
avNNnet	0.1	1	0.145 (0.046)	0.853 (0.049)	0.237 (0.051)
sGroup-Lasso	0.1	1	0.321 (0.050)	0.908 (0.042)	0.204 (0.061)
L1-SVM	0.1	1	0.281 (0.097)	0.920 (0.041)	0.170 (0.052)
iBIG	0.1	1	**0.789 (0.144)**	**0.948 (0.038)**	**0.134 (0.063)**
BTS	0.1	1	0.548 (0.145)	0.894 (0.059)	0.214 (0.080)
avNNnet	0.1	1.5	0.159 (0.04)	0.889 (0.032)	0.201 (0.056)
sGroup-Lasso	0.1	1.5	0.321 (0.044)	0.943 (0.029)	0.168 (0.063)
L1-SVM	0.1	1.5	0.297 (0.086)	0.941 (0.024)	0.154 (0.051)
iBIG	0.1	1.5	**0.821 (0.156)**	**0.974 (0.025)**	**0.099 (0.048)**
BTS	0.1	1.5	0.663 (0.129)	0.922 (0.057)	0.163 (0.075)
avNNnet	0.2	0.5	0.145 (0.069)	0.771 (0.088)	0.318 (0.089)
sGroup-Lasso	0.2	0.5	0.344 (0.056)	0.787 (0.08)	0.310 (0.073)
L1-SVM	0.2	0.5	0.230 (0.051)	0.797 (0.083)	0.279 (0.094)
iBIG	0.2	0.5	**0.566 (0.122)**	**0.873 (0.068)**	**0.221 (0.078)**
BTS	0.2	0.5	0.354 (0.160)	0.736 (0.126)	0.321 (0.114)
avNNnet	0.2	1	0.164 (0.048)	0.844 (0.040)	0.230 (0.048)
sGroup-Lasso	0.2	1	0.336 (0.056)	0.898 (0.042)	0.230 (0.057)
L1-SVM	0.2	1	0.267 (0.077)	0.904 (0.037)	0.204 (0.061)
iBIG	0.2	1	**0.834 (0.127)**	**0.954 (0.039)**	**0.127 (0.059)**
BTS	0.2	1	0.546 (0.132)	0.869 (0.066)	0.234 (0.088)
avNNnet	0.2	1.5	0.155 (0.052)	0.861 (0.043)	0.233 (0.052)
sGroup-Lasso	0.2	1.5	0.329 (0.033)	0.935 (0.036)	0.167 (0.042)
L1-SVM	0.2	1.5	0.304 (0.095)	0.928 (0.034)	0.185 (0.050)
iBIG	0.2	1.5	**0.938 (0.076)**	**0.991 (0.012)**	**0.055 (0.035)**
BTS	0.2	1.5	0.647 (0.140)	0.913 (0.062)	0.161 (0.079)
avNNnet	$\hat{\Omega}_{\mathrm{GL}}$	0.5	0.132 (0.055)	0.803 (0.067)	0.283 (0.092)
sGroup-Lasso	$\hat{\Omega}_{\mathrm{GL}}$	0.5	0.344 (0.048)	0.832 (0.078)	0.254 (0.079)
L1-SVM	$\hat{\Omega}_{\mathrm{GL}}$	0.5	0.223 (0.061)	0.835 (0.122)	0.239 (0.085)
iBIG	$\hat{\Omega}_{\mathrm{GL}}$	0.5	**0.603 (0.146)**	**0.912 (0.070)**	**0.162 (0.068)**
BTS	$\hat{\Omega}_{\mathrm{GL}}$	0.5	0.387 (0.113)	0.821 (0.112)	0.248 (0.102)
avNNnet	$\hat{\Omega}_{\mathrm{GL}}$	1	0.132 (0.055)	0.886 (0.043)	0.217 (0.059)
sGroup-Lasso	$\hat{\Omega}_{\mathrm{GL}}$	1	0.318 (0.052)	0.952 (0.030)	0.137 (0.050)
L1-SVM	$\hat{\Omega}_{\mathrm{GL}}$	1	0.245 (0.069)	0.943 (0.038)	0.144 (0.050)
iBIG	$\hat{\Omega}_{\mathrm{GL}}$	1	**0.715 (0.161)**	**0.971 (0.021)**	**0.090 (0.035)**
BTS	$\hat{\Omega}_{\mathrm{GL}}$	1	0.561 (0.129)	0.939 (0.043)	0.142 (0.057)
avNNnet	$\hat{\Omega}_{\mathrm{GL}}$	1.5	0.105 (0.033)	0.924 (0.041)	0.154 (0.054)
sGroup-Lasso	$\hat{\Omega}_{\mathrm{GL}}$	1.5	0.314 (0.044)	0.972 (0.020)	0.105 (0.047)
L1-SVM	$\hat{\Omega}_{\mathrm{GL}}$	1.5	0.255 (0.083)	0.962 (0.017)	0.111 (0.028)
iBIG	$\hat{\Omega}_{\mathrm{GL}}$	1.5	**0.833 (0.130)**	**0.988 (0.013)**	**0.060 (0.032)**
BTS	$\hat{\Omega}_{\mathrm{GL}}$	1.5	0.654 (0.100)	0.969 (0.039)	0.113 (0.053)

TABLE 2
Simulation results—scenarios 1–4 (30 *replicates*). *F*1 *is the F*1*-measure*, *AUCp is the predictive AUC*, *and MCE is the misclassification error*

Method	Scenario	F1	AUCp	MCE
avNNnet	Scenario 1	0.152 (0.044)	0.841 (0.054)	0.249 (0.059)
sGroup-Lasso	Scenario 1	0.330 (0.050)	0.906 (0.049)	0.208 (0.064)
L1-SVM	Scenario 1	0.294 (0.094)	0.931 (0.026)	0.161 (0.040)
iBIG	Scenario 1	0.789 (0.144)	0.948 (0.038)	0.134 (0.063)
BTS	Scenario 1	0.548 (0.145)	0.875 (0.069)	0.214 (0.080)
avNNnet	Scenario 2	0.236 (0.045)	0.821 (0.061)	0.240 (0.072)
sGroup-Lasso	Scenario 2	0.374 (0.048)	0.896 (0.041)	0.187 (0.051)
L1-SVM	Scenario 2	0.259 (0.070)	0.810 (0.068)	0.251 (0.062)
iBIG	Scenario 2	0.840 (0.157)	0.928 (0.063)	0.138 (0.078)
BTS	Scenario 2	0.724 (0.227)	0.896 (0.094)	0.163 (0.094)
avNNnet	Scenario 3	0.230 (0.062)	0.732 (0.062)	0.330 (0.049)
sGroup-Lasso	Scenario 3	0.387 (0.037)	0.865 (0.036)	0.232 (0.045)
L1-SVM	Scenario 3	0.354 (0.106)	0.826 (0.063)	0.255 (0.069)
iBIG	Scenario 3	0.881 (0.112)	0.902 (0.078)	0.172 (0.082)
BTS	Scenario 3	0.844 (0.115)	0.890 (0.076)	0.183 (0.085)
avNNnet	Scenario 4	0.164 (0.037)	0.767 (0.046)	0.283 (0.053)
sGroup-Lasso	Scenario 4	0.393 (0.049)	0.914 (0.025)	0.159 (0.039)
L1-SVM	Scenario 4	0.364 (0.133)	0.806 (0.043)	0.267 (0.056)
iBIG	Scenario 4	0.891 (0.105)	0.948 (0.052)	0.103 (0.065)
BTS	Scenario 4	0.868 (0.135)	0.946 (0.039)	0.104 (0.042)

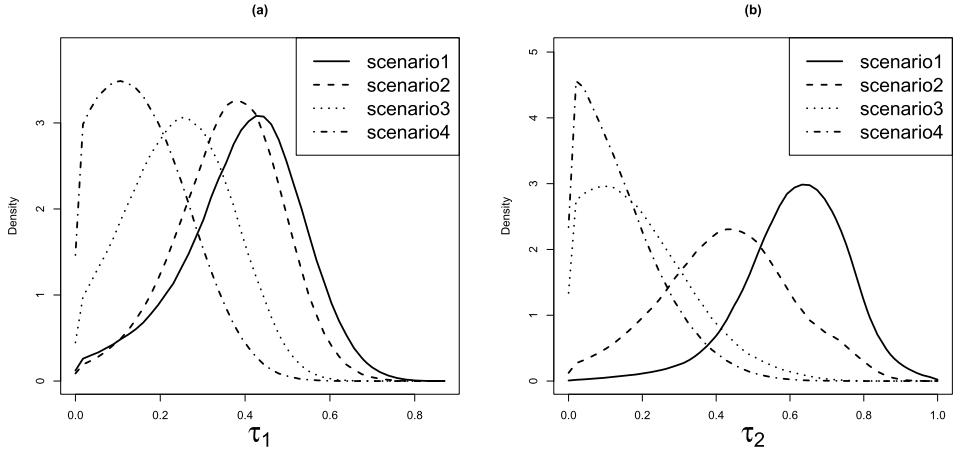

FIG. 1. *Density plots of the effects of the network obtained by combining all the MCMC samples of the* 30 *replicates*. (a) *parameter* τ_1 *and* (b) *parameter* τ_2.

available consist of 92 patients diagnosed with schizophrenia and 118 healthy controls. Prior to inclusion in the study, all healthy participants were free of any medical, neurological or psychiatric illnesses and had no history of substance abuse. The inclusion criteria for patients were based on a diagnosis of schizophrenia, schizophreniform or schizoaffective disorder. Details on data collection and preprocessing, as well as the participants' demographics (sex, age, ethnicity), can be found in Chen et al. (2012) and Gollub et al. (2013), and are also reported in Section I of the Supplementary Materials. In summary, fMRI data were collected from all participants during a sensorimotor task, a block-design motor response to auditory stimulation. In particular, a stimulus-on versus stimulus-off coefficient was obtained for each of the $53 \times 63 \times 46$ voxels that comprised the statistical parametric maps. Each brain image was then segmented into $G = 116$ ROIs according to the Automated Anatomical Labeling (AAL) atlas [Tzourio-Mazoyer et al. (2002)]. ROI-based summaries were obtained by computing the median of the statistical parametric map values of blood oxygenation level-independent (BOLD) measurements from all voxels within each region [Chen et al. (2012)]. In addition to the imaging data, we have measurements available on $M = 81$ genetic covariates (SNPs) for each participant in the study. The SNPs were selected by accessing the online Schizophrenia Research Forum (http://www.schizophreniaforum.org/) and querying for SNPs that had previously been implicated in schizophrenia. We randomly split the MCIC data into a training set and a validation set of 168 (4/5) and 42 (1/5) samples, respectively. To obtain a training set with enough information on both groups, we followed a balanced allocation scheme and randomly selected 94 healthy controls and 74 patients for the training set, and 24 healthy controls and 18 patients for the validation set. We constructed a spatial network among ROIs based on the physical location of the regions in the brain. The three-dimensional spatial coordinates of the centroids of each ROI allowed us to calculate a distance matrix among the ROIs based on the Euclidean distance.

We used the resulting network to define the MRF prior (9). We ran 6 MCMC chains with different starting points, both for iBIG and BTS. We assessed the agreement of the results among the six chains by evaluating the correlation coefficients between the marginal posterior probabilities for biomarker selection. These indicated good concordance between the six chains, with all correlations > 0.9. MCMC diagnostics confirmed that our chains were run for a satisfactory number of iterations (see Section E in the Supplementary Material for details).

By means of 5-fold cross-validation, we compared the prediction performance of both iBIG and BTS to the performance of several competing methods (Table 3). Furthermore, we compared the performance of our integrative model including the ROI–SNP network with simpler models that incorporated only the SNP or ROI covariates. As a general result, for most methods, increased accuracy can be observed when the two data modalities are combined in a single framework, in terms of both MCE and out-of-sample predictions (summarized by AUCp). The

TABLE 3

Comparison of predictive performance on the MCIC data. Empirical standard errors are between parentheses. "NA" is used to indicate that the sparse group lasso is not applicable when using only one type of covariate

Both ROIs and SNPs					
Method	**iBIG**	**BTS**	**L1-Lasso**	**L1-SVM**	**LogitBoost**
AUCp	0.69 (0.03)	0.66 (0.03)	0.62 (0.04)	0.64 (0.05)	0.63 (0.06)
MCE	0.33 (0.02)	0.34 (0.03)	0.36 (0.02)	0.40 (0.02)	0.37 (0.04)
Method	**Net-Lasso**	**sGroup-Lasso**	**MKL**	**avNNnet**	**DNN**
AUCp	0.67 (0.02)	0.67 (0.03)	0.53 (0.01)	0.68 (0.03)	0.55 (0.02)
MCE	0.34 (0.02)	0.36 (0.01)	0.43 (0.00)	0.36 (0.02)	0.46 (0.03)
Only SNPs					
Method	**iBIG***	**BTS***	**L1-Lasso**	**L1-SVM**	**LogitBoost**
AUCp	0.64 (0.04)		0.57 (0.03)	0.62 (0.02)	0.55 (0.02)
MCE	0.45 (0.01)		0.44 (0.01)	0.44 (0.02)	0.49 (0.02)
Method	**Net-Lasso**	**sGroup-Lasso**	**MKL**	**avNNnet**	**DNN**
AUCp	0.61 (0.04)	NA	0.59 (0.01)	0.63 (0.02)	0.53 (0.01)
MCE	0.43 (0.02)	NA	0.44 (0.00)	0.40 (0.01)	0.49 (0.02)
Only ROIs					
Method	**iBIG***	**BTS***	**L1-Lasso**	**L1-SVM**	**LogitBoost**
AUCp	0.66 (0.02)		0.62 (0.02)	0.65 (0.02)	0.62 (0.02)
MCE	0.37 (0.02)		0.40 (0.02)	0.40 (0.02)	0.39 (0.02)
Method	**Net-Lasso**	**sGroup-Lasso**	**MKL**	**avNNnet**	**DNN**
AUCp	0.65 (0.02)	NA	0.53 (0.01)	0.65 (0.02)	0.55 (0.02)
MCE	0.39 (0.02)	NA	0.44 (0.00)	0.37 (0.02)	0.44 (0.00)

*IBIG and BTS are equivalent with only either ROIs or SNPs as covariates.

iBIG method performed generally better than the BTS method (without the ROI–SNP network) and the other competing methods. Small standard errors confirm good stability of the results.

We then investigated the markers selected by our approach. Figures 2 and 3 show the marginal inclusion posterior probabilities for the ROIs, SNPs, and for each connection in the ROI–SNP network. These probabilities can be used to prioritize the relevant ROIs, SNPs and ROI–SNP pairs for further experimental work. Concerning the ROI–SNP network, iBIG identified 22 ROIs connected to 6 SNPs, for a total of 24 ROI–SNP pairs with marginal posterior probability > 0.5 (see

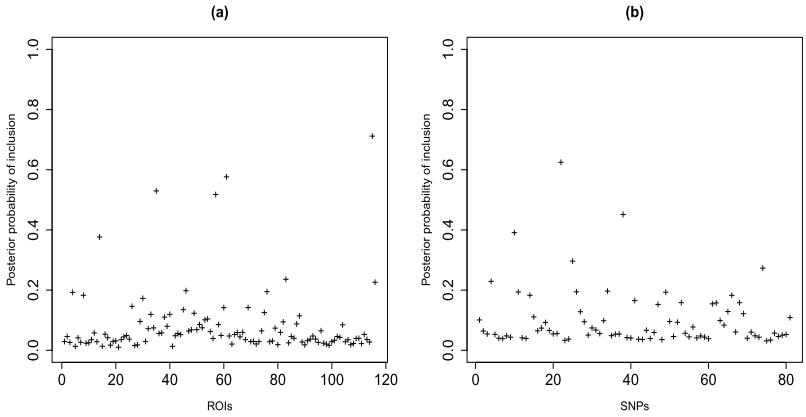

FIG. 2. *Marginal posterior probabilities* (a) *ROI*; (b) *SNP*.

Figure 3). SNPs 4, 11, 18 and 77 are connected to multiple brain regions (15, 2, 2 and 3 regions, respectively). Most of these regions are neighbors in the ROI network. SNP 4 (rs3803300), which is highly connected in the estimated ROI–SNP network, is part of gene AKT1 on chromosome 14, and is known to be associated with schizophrenia [Ikeda et al. (2008), Joo et al. (2009), Xu et al. (2007)]. ROI 35 (left posterior cingulum), which was found to be a discriminatory biomarker, with posterior probability of 0.53, is connected to SNPs 4 and 55 in the estimated ROI–SNP network. Direct relationships between abnormalities of the posterior

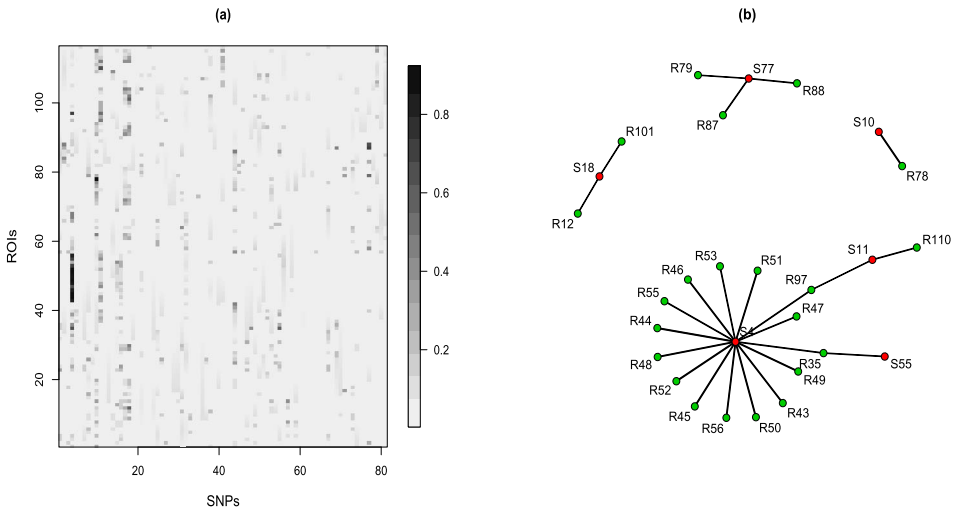

FIG. 3. (a) *Heat map of the marginal posterior probabilities of the ROI–SNP pairs*; (b) *ROI–SNP network using a threshold of* 0.5 *on its marginal posterior probability. Red nodes correspond to SNPs and green nodes correspond to ROIs*.

cingulum and positive symptoms in schizophrenia have been previously reported [Fujiwara et al. (2007)]. ROIs 57, 61 and 115 were also selected as discriminatory, with marginal posterior probabilities of 0.52, 0.57 and 0.71, respectively. ROI 57 (left postcentral gyrus) and ROI 61 (left inferior parietal region) are neighbors in the ROI network and have been associated with schizophrenia [Glahn et al. (2008), Müller et al. (2013), Waltz et al. (2009), Yang et al. (2010)]. To our knowledge, ROI 115, the posterior inferior vermis (lobule IX), has not yet been shown to be related to schizophrenia. However, many studies have shown that an abnormality involving the entire vermis (lobules I–X) may contribute to the pathophysiology of schizophrenia [Levitt et al. (1999), Okugawa, Sedvall and Agartz (2003)]. Three discriminatory SNPs (22, 10 and 38) were also identified by iBIG, with marginal posterior probabilities of 0.62, 0.39 and 0.45, respectively. SNP 22 (rs3737597) is located in gene DISC1 (chromosome 1), a gene which is disrupted in schizophrenia [Kim et al. (2012)]. It was also found to be discriminatory by Stingo et al. (2013) and Yang et al. (2010). This SNP was previously identified as an allele indicating risk of developing the schizophrenic disorder in Finnish families and among the Scandinavian population [Saetre et al. (2008)]. SNP 10 (rs2051632) is in gene ARHGAP18, which has been associated with schizophrenia through both imaging and case-control studies [Potkin et al. (2009)]. SNP 38 (rs194072) in gene GABRB2 has also been associated with schizophrenia susceptibility [Lo et al. (2004), Yu et al. (2006)].

To highlight the importance of our new variable selection prior, we compared the markers selected by iBIG with those obtained from BTS, which does not take into account the ROI–SNP network in the selection of the discriminatory markers (see Section F of the Supplementary Material for details on the BTS results). One main difference pertains to ROI 35, the left posterior cingulate region, which we found to be connected to SNPs 4 and 55: the posterior probability of this region increased from 0.11 (using BTS) to 0.53 (using iBIG).

We observed a similar trend for SNP 10, for which the posterior probability went from 0.24 (using BTS) to 0.39 (using iBIG); SNP 10 is connected to ROI 78 (the right thalamus) in the estimated ROI–SNP network. Both examples highlight the effect of the ROI–SNP network on the selection of discriminatory markers.

6. Conclusion. In this paper, we have proposed a Bayesian predictive model to accurately predict the disease status of schizophrenia for a subject based on a sparse set of imaging and genetic biomarkers. Our Bayesian approach has several innovative characteristics: (1) It is integrative since it combines in a single model both SNP and fMRI data; (2) It employs novel covariate-dependent variable selection priors, which incorporate inference on the ROI–SNP network to select a discriminatory set of biomarkers; and (3) It achieves sharper biomarker selection through the specification of nonlocal prior distributions on the regression coefficients. The performance of the method was evaluated on simulated data and on a dataset collected from individuals diagnosed with schizophrenia that was obtained

from the MIND Clinical Imaging Consortium. Our extensive simulation studies show that including the ROI–SNP network in the selection mechanism enhances both biomarker selection and prediction performances. We investigated the effect of the prior (9) on the posterior inference via simulation studies.

By employing the prior (9), prediction performance is particularly improved when there is a large number of discriminatory biomarkers that are also highly connected in the ROI–SNP network. Our model and simpler approaches, such as BTS, perform equally well if the data do not fully support the existence of a significant ROI–SNP network. When applied to the data from the MIND Clinical Imaging Consortium, the improved performance in variable selection, which results in higher precision and lower false negative and false positive findings, leads to the selection of a set of discriminatory biomarkers that would have been otherwise missed, aiding the interpretation of the results.

Overall, our results have confirmed the complex nature of genetic effects on the functional brain abnormality that is present in schizophrenia. Both biology and improved prediction performances confirm that our modeling assumptions are appropriate.

In our application, the fMRI activation has been summarized by using ROIs defined by the AAL atlas. However, activation or connectivity patterns might vary within predefined ROIs, with resulting loss of signal. To partially obviate this limitation, one might consider finer brain parcellations. In Section J of the Supplementary Materials, we illustrate the results of a simulation where we employ our method on an increased number of brain imaging markers (500 and 1000). Alternatively, one might focus on identifying activity and connectivity patterns localized inside only one or a few regions of interests. Finally, our modeling framework could be applied to other types of dimension reduction techniques. For example, independent component analysis (ICA) provides a grouping of brain activity into regions that share the same response pattern and can be seen as a data-driven approach to brain parcellation [Calhoun, Liu and Adali (2009), Wu et al. (2015)]. The number of components is typically chosen between 10 and 100 [Damaraju et al. (2014), Erhardt et al. (2011)], which is comparable to the number of imaging biomarkers we considered in this manuscript.

Acknowledgments. The authors are grateful to Vince Calhoun of the MIND Clinical Imaging Consortium for making available the data used in this study, to the Editor, two referees and an AE for their comments that greatly improved the quality of the article, and to LeeAnn Chastain for editing assistance.

SUPPLEMENTARY MATERIAL

Supplement to "A Bayesian predictive model for imaging genetics with application to schizophrenia" (DOI: 10.1214/16-AOAS948SUPP; .zip). The supplementary material [Chekouo et al. (2016)] contains details about posterior

computation, hyperparameter settings and sensitivity, data preprocessing, and additional simulation studies and data analyses. The companion MATLAB code is available on The Annals of Applied Statistics website.

REFERENCES

ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. MR1224394

ATCHADÉ, Y. F., LARTILLOT, N. and ROBERT, C. (2013). Bayesian computation for statistical models with intractable normalizing constants. *Braz. J. Probab. Stat.* **27** 416–436. MR3105037

BATMANGHELICH, N., DALCA, A., SABUNCU, M. and GOLLAND, P. (2013). Joint modeling of imaging and genetics. In *Information Processing in Medical Imaging* (J. C. Gee, S. Joshi, K. Pohl, W. M. Wells and L. Zellei, eds.). *Lecture Notes in Computer Science* **7917** 766–777. Springer, Berlin.

BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–236. MR0373208

BOWMAN, F. D. (2014). Brain imaging analysis. *Annu. Rev. Stat. Appl.* **1** 61–85.

CALHOUN, V. D. and HUGDAHL, K. (2012). Cognition and neuroimaging in schizophrenia. *Front. Human Neurosci.* **6** 276.

CALHOUN, V. D., LIU, J. and ADALI, T. (2009). A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *NeuroImage* **45** S163–S172.

CANNON, T. D. and KELLER, M. C. (2006). Endophenotypes in the genetic analyses of mental disorders. *Annu. Rev. Clin. Psychol.* **2** 267–290.

CAO, H., DUAN, J., LIN, D., CALHOUN, V. and WANG, Y.-P. (2013). Integrating fMRI and SNP data for biomarker identification for schizophrenia with a sparse representation based variable selection method. *BMC Medical Genomics* **6 Suppl 3** S2.

CHEKOUO, T., STINGO, F. C., GUINDANI, M. and DO, K. (2016). Supplement to "A Bayesian predictive model for imaging genetics with application to schizophrenia." DOI:10.1214/16-AOAS948SUPP.

CHEN, J., CALHOUN, V. D., PEARLSON, G. D., EHRLICH, S., TURNER, J. A., HO, B.-C., WASSINK, T. H., MICHAEL, A. and LIU, J. (2012). Multifaceted genomic risk for brain function in schizophrenia. *NeuroImage* **61** 866–875.

CHI, E. C., ALLEN, G. I., ZHOU, H., KOHANNIM, O., LANGE, K. and THOMPSON, P. M. (2013). Imaging genetics via sparse canonical correlation analysis. In *Biomedical Imaging* (*ISBI*), 2013 *IEEE* 10*th International Symposium on* 740–743.

DAMARAJU, E., ALLEN, E. A., BELGER, A., FORD, J. M., MCEWEN, S., MATHALON, D. H., CALHOUN, V. D. et al. (2014). Dynamic functional connectivity analysis reveals transient states of dysconnectivity in schizophrenia. *NeuroImage*: *Clinical* **5** 298–308.

DENG, L. and YU, D. (2013). Deep learning: Methods and applications. *Found. Trends Signal Process.* **7** 197–391. MR3295556

DETTLING, M. and BÜHLMANN, P. (2003). Boosting for tumor classification with gene expression data. *Bioinformatics* **19** 1061–1069.

ERHARDT, E. B., RACHAKONDA, S., BEDRICK, E. J., ALLEN, E. A., ADALI, T. and CALHOUN, V. D. (2011). Comparison of multi-subject ICA methods for analysis of fMRI data. *Hum. Brain Mapp.* **32** 2075–2095.

FILIPOVYCH, R., RESNICK, S. M. and DAVATZIKOS, C. (2011). Multi-kernel classification for integration of clinical and imaging data: Application to prediction of cognitive decline in older adults machine learning in medical imaging (K. Suzuki, F. Wang, D. Shen and P. Yan, eds.). *Lecture Notes in Computer Science* **7009** 26–34. Springer, Berlin.

FLOCH, É. L., GUILLEMOT, V., FROUIN, V., PINEL, P., LALANNE, C., TRINCHERA, L., TENENHAUS, A., MORENO, A., ZILBOVICIUS, M., BOURGERON, T., DEHAENE, S., THIRION, B., POLINE, J.-B. and DUCHESNAY, É. (2012). Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares. *NeuroImage* **63** 11–24.

FORNITO, A., ZALESKY, A., PANTELIS, C. and BULLMORE, E. T. (2012). Schizophrenia, neuroimaging and connectomics. *NeuroImage* **62** 2296–2314.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9** 432–441.

FRIEDMAN, J. H., HASTIE, T. and TIBSHIRANI, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33** 1–22.

FUJIWARA, H., NAMIKI, C., HIRAO, K., MIYATA, J., SHIMIZU, M., FUKUYAMA, H., SAWAMOTO, N., HAYASHI, T. and MURAI, T. (2007). Anterior and posterior cingulum abnormalities and their association with psychopathology in schizophrenia: A diffusion tensor imaging study. *Schizophr. Res.* **95** 215–222.

GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–374.

GLAHN, D. C., LAIRD, A. R., ELLISON-WRIGHT, I., THELEN, S. M., ROBINSON, J. L., LANCASTER, J. L., BULLMORE, E. and FOX, P. T. (2008). Meta-analysis of gray matter anomalies in schizophrenia: Application of anatomic likelihood estimation and network analysis. *Biological Psychiatry* **64** 774–781.

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

GOLDSMITH, J., CRAINICEANU, C. M., CAFFO, B. and REICH, D. (2012). Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **61** 453–469. MR2914521

GOLLUB, R. L., SHOEMAKER, J. M., KING, M. D., WHITE, T., EHRLICH, S., SPONHEIM, S. R., CLARK, V. P., TURNER, J. A., MUELLER, B. A., MAGNOTTA, V. et al. (2013). The MCIC collection: A shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia. *Neuroinformatics* **11** 367–388.

HARDOON, D. R., ETTINGER, U., MOURÃO-MIRANDA, J., ANTONOVA, E., COLLIER, D., KUMARI, V., WILLIAMS, S. C. R. and BRAMMER, M. (2009). Correlation-based multivariate analysis of genetic influence on brain volume. *Neurosci. Lett.* **450** 281–286.

HARIRI, A. R. and WEINBERGER, D. R. (2003). Imaging genomics. *Br. Med. Bull.* **65** 259–270.

IKEDA, M., YAMANOUCHI, Y., KINOSHITA, Y., KITAJIMA, T., YOSHIMURA, R., HASHIMOTO, S., O'DONOVAN, M. C., NAKAMURA, J., OZAKI, N. and IWATA, N. (2008). Variants of dopamine and serotonin candidate genes as predictors of response to risperidone treatment in first-episode schizophrenia. *Pharmacogenomics* **9** 1437–1443.

JACOB, A. (2013). Limitations of clinical psychiatric diagnostic measurements. *J. Neurol. Disord.* **2**.

JOHNSON, V. E. and ROSSELL, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 143–170. MR2830762

JOHNSON, V. E. and ROSSELL, D. (2012). Bayesian model selection in high-dimensional settings. *J. Amer. Statist. Assoc.* **107** 649–660. MR2980074

JOO, E. J., LEE, K. Y., JEONG, S. H., ROH, M. S., KIM, S. H., AHN, Y. M. and KIM, Y. S. (2009). AKT1 gene polymorphisms and obstetric complications in the patients with schizophrenia. *Psychiatry Investigation* **6** 102–107.

KIM, J. Y., LIU, C. Y., ZHANG, F., DUAN, X., WEN, Z., SONG, J., FEIGHERY, E., LU, B., RUJESCU, D., CLAIR, D. S., CHRISTIAN, K., CALLICOTT, J. H., WEINBERGER, D. R., SONG, H. and LI MING, G. (2012). Interplay between DISC1 and GABA signaling regulates neurogenesis in mice and risk for schizophrenia. *Cell* **148** 1051–1064.

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. *Nature* **521** 436–444.

Lencz, T., Morgan, T. V., Athanasiou, M., Dain, B., Reed, C. R., Kane, J. M., Kucherlapati, R. and Malhotra, A. K. (2007). Converging evidence for a pseudoautosomal cytokine receptor gene locus in schizophrenia. *Mol. Psychiatry* **12** 572–580.

Levitt, J. J., McCarley, R. W., Nestor, P. G., Petrescu, C., Donnino, R., Hirayasu, Y., Kikinis, R., Jolesz, F. A. and Shenton, M. E. (1999). Quantitative volumetric MRI study of the cerebellum and vermis in schizophrenia: Clinical and cognitive correlates. *Am. J. Psychiatr.* **156** 1105–1107.

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. Z., Maresh, E. L. and Coan, J. (2015). Spatial Bayesian variable selection and grouping in high-dimensional scalar-on-image regressions. *Ann. Appl. Stat.* **9** 687–713.

Lin, D., Calhoun, V. D. and Wang, Y.-P. (2014). Correspondence between fMRI and SNP data by group sparse canonical correlation analysis. *Med. Image Anal.* **18** 891–902.

Lin, J.-A., Zhu, H., Mihye, A., Sun, W., Ibrahim, J. G. and for the Alzheimer's Neuroimaging Initiative (2014). Functional-mixed effects models for candidate genetic mapping in imaging genetic studies. *Genet. Epidemiol.* **38** 680–691.

Lindquist, M. A. (2008). The statistical analysis of fMRI data. *Statist. Sci.* **23** 439–464. MR2530545

Liu, J. and Calhoun, V. D. (2014). A review of multivariate analyses in imaging genetics. *Front. Neuroinform.* **8** 29.

Liu, J., Pearlson, G., Windemuth, A., Ruano, G., Perrone-Bizzozero, N. I. and Calhoun, V. (2009). Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA. *Hum. Brain Mapp.* **30** 241–255.

Lo, W.-S., Lau, C.-F., Xuan, Z., Chan, C.-F., Feng, G.-Y., He, L., Cao, Z.-C., Liu, H., Luan, Q.-M. and Xue, H. (2004). Association of SNPs and haplotypes in GABAA receptor beta2 gene with schizophrenia. *Mol. Psychiatry* **9** 603–608.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.* **89** 1535–1546.

Meda, S. A., Narayanan, B., Liu, J., Perrone-Bizzozero, N. I., Stevens, M. C., Calhoun, V. D., Glahn, D. C., Shen, L., Risacher, S. L., Saykin, A. J. and Pearlson, G. D. (2012). A large scale multivariate parallel {ICA} method reveals novel imaging–genetic relationships for Alzheimer's disease in the {ADNI} cohort. *NeuroImage* **60** 1608–1621.

Meyer-Lindenberg, A. (2012). The future of fMRI and genetics research. *NeuroImage* **62** 1286–1292.

Müller, V. I., Cieslik, E. C., Laird, A. R., Fox, P. T. and Eickhoff, S. B. (2013). Dysregulated left inferior parietal activity in schizophrenia and depression: Functional connectivity and characterization. *Front. Human Neurosci.* **7** 268.

Okugawa, G., Sedvall, G. C. and Agartz, I. (2003). Smaller cerebellar vermis but not hemisphere volumes in patients with chronic schizophrenia. *Am. J. Psychiatr.* **160** 1614–1617.

Potkin, S. G., Turner, J. A., Fallon, J. A., Lakatos, A., Keator, D. B., Guffanti, G. and Macciardi, F. (2009). Gene discovery through imaging genetics: Identification of two novel genes associated with schizophrenia. *Mol. Psychiatry* **14** 416–428.

Potkin, S. G., van Erp, T. G. M., Ling, S., Macciardi, F. and Xie, X. (2015). Identifying Unanticipated Genes and Mechanisms in Serious Mental Illness: GWAS Based Imaging Genetics Strategies. 209. Oxford Univ. Press, London.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge Univ. Press, Cambridge. MR1438788

SAETRE, P., AGARTZ, I., FRANCISCIS, A. D., LUNDMARK, P., DJUROVIC, S., KAHLER, A., ANDREASSEN, O. A., JAKOBSEN, K. D., RASMUSSEN, H. B., WERGE, T., HALL, H., TERENIUS, L. and JONSSON, E. G. (2008). Association between a disrupted-in-schizophrenia 1 (DISC1) single nucleotide polymorphism and schizophrenia in a combined Scandinavian case-control sample. *Schizophr. Res.* **106** 237–241.

SCOTT, J. G. and BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38** 2587–2619. MR2722450

SHA, N., VANNUCCI, M., TADESSE, M. G., BROWN, P. J., DRAGONI, I., DAVIES, N., ROBERTS, T. C., CONTESTABILE, A., SALMON, M., BUCKLEY, C. and FALCIANI, F. (2004). Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics* **60** 812–828. MR2089459

SHAHBABA, B., SHACHAF, C. M. and YU, Z. (2012). A pathway analysis method for genome-wide association studies. *Stat. Med.* **31** 988–1000. MR2913874

SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2013). A sparse-group lasso. *J. Comput. Graph. Statist.* **22** 231–245. MR3173712

SONNENBURG, S., RÄTSCH, G., SCHÄFER, C. and SCHÖLKOPF, B. (2006). Large scale multiple kernel learning. *J. Mach. Learn. Res.* **7** 1531–1565. MR2274416

STINGO, F. C., VANNUCCI, M. and DOWNEY, G. (2012). Bayesian wavelet-based curve classification via discriminant analysis with Markov random tree priors. *Statist. Sinica* **22** 465–488. MR2954348

STINGO, F. C., CHEN, Y. A., VANNUCCI, M., BARRIER, M. and MIRKES, P. E. (2010). A Bayesian graphical modeling approach to microRNA regulatory network inference. *Ann. Appl. Stat.* **4** 2024–2048. MR2829945

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

STINGO, F. C., GUINDANI, M., VANNUCCI, M. and CALHOUN, V. D. (2013). An integrative Bayesian modeling approach to imaging genetics. *J. Amer. Statist. Assoc.* **108** 876–891. MR3174670

SWARTZ, M. D., YU, R. K. and SHETE, S. (2008). Finding factors influencing risk: Comparing Bayesian stochastic search and standard variable selection methods applied to logistic regression models of cases and controls. *Stat. Med.* **27** 6158–6174. MR2522315

TZOURIO-MAZOYER, N., LANDEAU, B., PAPATHANASSIOU, D., CRIVELLO, F., ETARD, O., DELCROIX, N., MAZOYER, B. and JOLIOT, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* **15** 273–289.

VOUNOU, M., NICHOLS, T. E. and MONTANA, G. (2010). Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. *NeuroImage* **53** 1147–1159.

VOUNOU, M., JANOUSOVA, E., WOLZ, R., STEIN, J. L., THOMPSON, P. M., RUECKERT, D. and MONTANA, G. (2012). Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. *NeuroImage* **60** 700–716.

WALTZ, J. A., SCHWEITZER, J. B., GOLD, J. M., KURUP, P. K., ROSS, T. J., SALMERON, B. J., ROSE, E. J., MCCLURE, S. M. and STEIN, E. A. (2009). Patients with schizophrenia have a reduced neural response to both unpredictable and predictable primary reinforcers. *Neuropsychopharmacology* **34** 1567–1577.

WANG, H., NIE, F., HUANG, H., RISACHER, S. L., SAYKIN, A. J., SHEN, L. and THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2012a). Identifying disease sensitive and quantitative trait-relevant biomarkers from multidimensional heterogeneous imaging genetics data via sparse multimodal multitask learning. *Bioinformatics* **28** i127–i136.

WANG, H., NIE, F., HUANG, H., KIM, S., NHO, K., RISACHER, S. L., SAYKIN, A. J., SHEN, L. and THE ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (2012b). Identifying quantitative trait loci via group-sparse multitask regression and feature selection: An imaging genetics study of the ADNI cohort. *Bioinformatics* **28** 229–237.

WEISS, K. M. (1989). Advantages of abandoning symptom-based diagnostic systems of research in schizophrenia. *Am. J. Orthopsychiatr.* **59** 324–330.

WU, L., CALHOUN, V. D., JUNG, R. E. and CAPRIHAN, A. (2015). Connectivity-based whole brain dual parcellation by group ICA reveals tract structures and decreased connectivity in schizophrenia. *Hum. Brain Mapp.* **36** 4681–4701.

XU, M.-Q., XING, Q.-H., ZHENG, Y.-L., LI, S., GAO, J.-J., HE, G., GUO, T.-W., FENG, G.-Y., XU, F. and HE, L. (2007). Association of AKT1 gene polymorphisms with risk of schizophrenia and with response to antipsychotics in the Chinese population. *J. Clin. Psychiatry* **68** 1358–1367.

YANG, H., LIU, J., SUI, J., PEARLSON, G. and CALHOUN, V. D. (2010). A hybrid machine learning method for fusing fMRI and genetic data: Combining both improves classification of schizophrenia. *Front. Human Neurosci.* **4** 1–9.

YU, Z., CHEN, J., SHI, H., STOEBER, G., TSANG, S.-Y. and XUE, H. (2006). Analysis of GABRB2 association with schizophrenia in German population with DNA sequencing and one-label extension method for SNP genotyping. *Clin. Biochem.* **39** 210–218.

ZHANG, L., GUINDANI, M. and VANNUCCI, M. (2015). Bayesian models for functional magnetic resonance imaging data analysis. *Wiley Interdiscip. Rev.: Comput. Stat.* **7** 21–41. MR3348719

ZHANG, Z., HUANG, H. and SHEN, D. (2014). Integrative analysis of multi-dimensional imaging genomics data for Alzheimer's disease prediction. *Front. Aging Neurosci.* **6** 1–9.

ZHANG, T., WIESEL, A. and GRECO, M. S. (2013). Multivariate generalized Gaussian distribution: Convexity and graphical models. *IEEE Trans. Signal Process.* **61** 4141–4148. MR3085302

ZHANG, H. H., AHN, J., LIN, X. and PARK, C. (2006). Gene selection using support vector machines with non-convex penalty. *Bioinformatics* **22** 88–95.

ZHU, H., KHONDKER, Z., LU, Z. and IBRAHIM, J. G. (2014). Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. *J. Amer. Statist. Assoc.* **109** 977–990. MR3265670

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 301–320. MR2137327

T. CHEKOUO
DEPARTMENT OF MATHEMATICS
AND STATISTICS
UNIVERSITY OF MINNESOTA DULUTH
1117 UNIVERSITY DRIVE
DULUTH, MINNESOTA 55812
USA
E-MAIL: tchekouo@d.umn.edu

M. GUINDANI
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, IRVINE
BREN HALL 2019
IRVINE, CALIFORNIA 92697-1250
USA
E-MAIL: michele.guindani@UCI.edu

F. C. STINGO
DIPARTIMENTO DI STATISTICA, INFORMATICA,
APPLICAZIONI "G.PARENTI"
UNIVERSITY OF FLORENCE
VIALE MORGAGNI, 59
50134 FLORENCE
ITALY
E-MAIL: f.stingo@disia.unifi.it

K.-A. DO
DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF TEXAS
MD ANDERSON CANCER CENTER
1400 PRESSLER STREET, UNIT 1411
HOUSTON, TEXAS 77030-3722
USA
E-MAIL: kimdo@mdanderson.org