The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1217–1244
DOI: 10.1214/16-AOAS907

GENE-PROXIMITY MODELS FOR GENOME-WIDE ASSOCIATION STUDIES[1]

BY IAN JOHNSTON*, TIMOTHY HANCOCK†, HIROSHI MAMITSUKA†,[2] AND LUIS CARVALHO*,[3]

*Boston University** *and Kyoto University*†

Motivated by the important problem of detecting association between genetic markers and binary traits in genome-wide association studies, we present a novel Bayesian model that establishes a hierarchy between markers and genes by defining weights according to gene lengths and distances from genes to markers. The proposed hierarchical model uses these weights to define unique prior probabilities of association for markers based on their proximities to genes that are believed to be relevant to the trait of interest. We use an expectation-maximization algorithm in a filtering step to first reduce the dimensionality of the data and then sample from the posterior distribution of the model parameters to estimate posterior probabilities of association for the markers. We offer practical and meaningful guidelines for the selection of the model tuning parameters and propose a pipeline that exploits a singular value decomposition on the raw data to make our model run efficiently on large data sets. We demonstrate the performance of the model in simulation studies and conclude by discussing the results of a case study using a real-world data set provided by the Wellcome Trust Case Control Consortium.

1. Introduction. A genome-wide association study (GWAS) aims to determine the subset of genetic markers that is most relevant to a particular trait of interest. From a statistical perspective, this task is usually framed as a regression problem where the response variables are measurements of either qualitative traits, for example, a binary value indicating the presence or absence of a disease, or quantitative traits, for example, a person's blood pressure, and the explanatory variables are the number of reference alleles present at each marker, or single nucleotide polymorphism (SNP), as well as other covariates of interest such as age or smoking status. Many linear models [Balding (2006), Stephens and Balding (2009)] have been developed to detect associations between SNPs and traits, but they generally suffer from the "large p, small n" problem where the ratio of the number of predictors, p, to the sample size, n, is on the order of hundreds to thousands [West (2003)]. Moreover, other issues such as collinearity in the covariates due to

Received October 2013; revised September 2015.

[1]Supported in part by the ICR-KU International Short-term Exchange Program for Young Researchers, the JSPS Summer Program 2013 (SP13024), and the NSF EAPSI Program 2013.

[2]Supported in part by JSPS KAKENHI Grant 24300056.

[3]Supported in part by NSF Grant DMS-11-07067.

Key words and phrases. Large p small n, hierarchical Bayes, Pólya–Gamma latent variable.

linkage disequilibrium [LD, Pritchard and Przeworski (2001)], rare variants, and population stratification result in inefficient estimation of model parameters and a loss in statistical power to detect significant associations [Wang et al. (2005)].

A common strategy to overcome the large-p-small-n problem in GWAS is to forgo analyzing the SNPs jointly and to model instead them independently. Although successful GWAS have employed this strategy [Burton et al. (2007)], multiple hypothesis testing leads to an increase in the Type I error, and the necessary correction for this may lead to an overly conservative threshold for statistical significance. Strategies such as grouping SNPs based on proximities to genes [Wu et al. (2010)] or moving windows [Wu et al. (2011)] have been proposed to allow for an increase in power by modeling SNPs jointly, but there is no universal agreement on how to define such windows or groupings. Similarly, other strategies include replacing a group of highly correlated SNPs with only one of its members [Ioannidis, Thomas and Daly (2009)], and removing or collapsing rare variants within a window into a score statistic [Bansal et al. (2010)], but again there is no agreement on how to choose which SNPs to retain or group. Recent approaches aim at gaining more power by pooling information across studies through meta-analysis [Evangelou and Ioannidis (2013)].

Although significant progress has been made on GWAS since 2000, it is still a relevant and challenging problem with goals such as modeling interactions between SNPs, genes, and environment effects that await beyond the obstacles already mentioned [Heard et al. (2010)]. In order to move toward a unifying framework for GWAS that allows for the large-p-small-n problem and the SNP-specific issues to be addressed simultaneously in a principled manner, we propose a novel hierarchical Bayesian model that exploits spatial relationships on the genome to define SNP-specific prior distributions on regression parameters. More specifically, in our proposed setting we model markers *jointly*, but we explore a variable selection approach that uses marker proximity to relevant genomic regions, such as genes, to help identify associated SNPs. Our contributions are the following:

1. We focus on binary traits which are arguably more common to GWAS, for example, case control studies, but more difficult to model due to lack of conjugacy. To circumvent the need for a Metropolis–Hastings step when sampling from the posterior distribution on model parameters, we use a recently proposed data augmentation strategy for logistic regression based on latent Pólya–Gamma random variables [Polson, Scott and Windle (2013)].

2. We perform variable selection by adopting a spike-and-slab prior [George and McCulloch (1993), Ishwaran and Rao (2005)] and propose a principled way to control the separation between the spike and slab components using a Bayesian false discovery rate similar to Whittemore (2007).

3. We use a novel weighting scheme to establish a relationship between SNPs and genomic regions and allow for SNP-specific prior distributions on the model parameters such that the prior probability of association for each SNP is a function

of its location on the chromosome relative to neighboring regions. Moreover, we allow for the "relevance" of a genomic region to contribute to the effect it has on its neighboring SNPs and consider "relevance" values calculated based on previous GWAS results in the literature; for example, see MalaCards (2014).

4. Before sampling from the posterior space using Gibbs sampling, we use an expectation-maximization [EM, Dempster, Laird and Rubin (1977)] algorithm in a filtering step to reduce the number of candidate markers in a manner akin to distilled sensing [Haupt, Castro and Nowak (2011)]. By investigating the update equations for the EM algorithm, we suggest meaningful values to tune the hyperprior parameters of our model and illustrate the induced relationship between SNPs and genomic regions.

5. We derive a more flexible centroid estimator [Carvalho and Lawrence (2008)] for SNP associations that is parameterized by a sensitivity-specificity trade-off. We discuss the relation between this parameter and the prior specification when obtaining estimates of model parameters.

We start by describing previous work and stating our contributions in Section 2. In Section 3 we define our Spatial Boost model and the novel relationship between SNPs and genes. In Section 4 we describe how to fit the model using a combination of a filtering step that exploits an EM filtering step and Gibbs sampling. We provide guidelines for the selection of model tuning parameters in Section 5. We then illustrate the performance of the model on simulated data using real SNPs in Section 6 and apply the model to a real-world GWAS data set provided by the Wellcome Trust Case Control Consortium (WTCCC) in Section 7. Finally, we conclude with a discussion on future extensions to this work in Section 8.

2. Previous and related work. A common solution to large-p-small-n problems is to use penalized regression models such as ridge regression [Hoerl and Kennard (1970)], LASSO [Tibshirani (1996)], or elastic net [Zou and Hastie (2005)]. These solutions can be shown to be equivalent, from a Bayesian perspective, to maximum a posteriori (MAP) estimators under appropriate prior specifications. For instance, for LASSO, the L_1 penalty can be translated into a Laplace prior. However, since LASSO produces biased estimates of the model parameters and tends to select only one parameter in a group of correlated parameters [Zou and Hastie (2005)], it is not suitable for GWAS.

Techniques like group LASSO, fused LASSO [Tibshirani et al. (2005)], or sparse group LASSO [Friedman, Hastie and Tibshirani (2010)] further attempt to account for the structure of genes and markers or linkage disequilibrium by assigning SNPs to groups based on criteria such as gene membership and then placing additional penalties on the L_1 norm of the vector of coefficients for each group, or on the L_1 norm of the difference in coefficients of consecutive SNPs. However, it is difficult to define gene membership universally since genes have varying lengths

and may overlap with each other; moreover, the penalty on the L_1 norm of the difference in consecutive SNPs neglects any information contained in the genomic distance between them.

It may be possible to develop additional, effective penalty terms within models, such as L_1 and L_2, to address the issues present in GWAS data in a penalized regression framework, but because genotypes are more correlated for markers that are close in genomic space due to linkage disequilibrium, the most effective penalties would need to capture the relevance of a particular SNP as a function of its location on the genome. Moreover, since it is typically easier to study the biological function of genes, we are particularly interested in SNPs that lie close to genes [Jorgenson and Witte (2006)]; as a result, the most desirable penalties would likely be SNP-specific. We accomplish this by exploiting biological knowledge about the structure of the genome to set SNP-specific prior distributions on the model parameters in a hierarchical Bayesian model.

Researchers have considered hierarchical Bayesian models for variable selection in GWAS and other large scale problems [e.g. Guan and Stephens (2011), Zhou, Carbonetto and Stephens (2013)]. Some recent models exploit Bayesian methods in particular to allow for data-driven SNP-specific prior distributions [Habier et al. (2011)] which depend on a random variable that describes the proportion of SNPs to be selected. These approaches have adopted a continuous spike-and-slab prior distribution [George and McCulloch (1993), Ishwaran and Rao (2005)] on the model parameters, set an inverse-gamma prior distribution on the variance of the spike component of the prior, and control the difference in the variance of the spike and slab components of the prior using a tuning parameter.

To incorporate external information in a hierarchical Bayesian model, researchers analyzing a different kind of data, gene expression levels, have recently considered relating a linear combination of a set of predictor-level covariates that quantify the relationships between the genes to their prior probabilities of association through a probit link function [Peng et al. (2013)]. This formulation leads to a second-stage probit regression on the probability that any gene is associated with a trait of interest using a set of predictor-level covariates that could be, for instance, indicator variables of molecular pathway membership. In our model, we propose a special case of this formulation tailored for GWAS data where: (i) we use the logit link instead of the probit link, (ii) the predictor-level covariates are spatial weights that quantify a SNP's position on the genome relative to neighboring genes, and (iii) the coefficients of each of the predictor-level covariates are numerical scores that quantify the relevance of a particular gene to the trait of interest.

Fitting a penalized model to a large data set (e.g., $p \geq 100{,}000$) is computationally intense, and thus so is the process of selecting an optimal value for any tuning parameters. Sidestepping this problem, some researchers have had success in applying a suite of penalty terms (e.g., LASSO, Adaptive LASSO, NEG, MCP, LOG) to a *prescreened* subset of markers [Hoffman, Logsdon and Mezey (2013)] and investigating the concordance of significant markers across each of the final

models. Although a prescreening of markers from a marginal regression would ideally retain almost all of the relevant variables, penalized models such as LASSO could likely be improved by using a larger number of SNPs than those which pass an initial screening step (e.g., a genome-wide significance threshold) [Kooperberg, LeBlanc and Obenchain (2010)].

3. Spatial boost model. We perform Bayesian variable selection by analyzing binary traits and using the structure of the genome to dynamically define the prior probabilities of association for the SNPs. Our data are the binary responses $\mathbf{y} \in \{0,1\}^n$ for n individuals and genotypes $\mathbf{x}_i \in \{0,1,2\}^p$ for p markers per individual, where x_{ij} codes the number of *minor* alleles in the ith individual for the jth marker. For the likelihood of the data, we consider the logistic regression:

$$y_i|\mathbf{x}_i, \beta \overset{\text{ind}}{\sim} \mathsf{Bern}\big(\text{logit}^{-1}(\beta_0 + \mathbf{x}_i^\top \beta)\big) \qquad \text{for } i = 1, \ldots, n. \tag{1}$$

We note that GWA studies are usually *retrospective*, that is, cases and controls are selected irrespectively of their history or genotypes; however, as McCullagh and Nelder (1983) point out, coefficient estimates for β are not affected by the sampling design under a logistic regression. Thus, from now on, to alleviate the notation, we extend $\mathbf{x}_i$ to incorporate the intercept, $\mathbf{x}_i = (x_{i0} = 1, x_{i1}, \ldots, x_{ip})$, and also set $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$.

We use latent variables $\theta \in \{0,1\}^p$ and a continuous spike-and-slab prior distribution for the model parameters with the positive constant $\kappa > 1$ denoting the separation between the variance of the spike and the slab components:

$$\beta_j|\theta_j, \sigma^2 \overset{\text{ind}}{\sim} \mathsf{N}\big(0, \sigma^2[\theta_j\kappa + (1-\theta_j)]\big) \qquad \text{for } j = 1, \ldots, p. \tag{2}$$

For the intercept, we set $\beta_0 \sim N(0, \sigma^2\kappa)$ or, equivalently, we define $\theta_0 = 1$ and include $j = 0$ in (2). In the original spike-and-slab prior distribution, the slab component is a normal distribution centered at zero with a large variance or even a diffuse uniform distribution, and the spike component is a degenerate distribution at zero [Mitchell and Beauchamp (1988)]. This setup results in exact variable selection through the use of the θ_j's, since $\theta_j = 0$ would imply that the jth SNP's coefficient is exactly equal to zero. Here we use the continuous version of the spike-and-slab distribution [George and McCulloch (1993), Ishwaran and Rao (2005)] that became more popular by avoiding the spike discontinuities at zero, and thus allowing for a relaxed form of variable selection that lends itself easily to an EM algorithm (see Section 4.1).

For the variance σ^2 of the spike component in (2) we adopt an inverse Gamma prior distribution, $\sigma^2 \sim \mathsf{IG}(\nu, \lambda)$. We expect σ^2 to be reasonably small with high probability in order to enforce the desired regularization that distinguishes selected markers from nonassociated markers. Thus, we recommend choosing ν and λ so that the prior expected value of σ^2 is small.

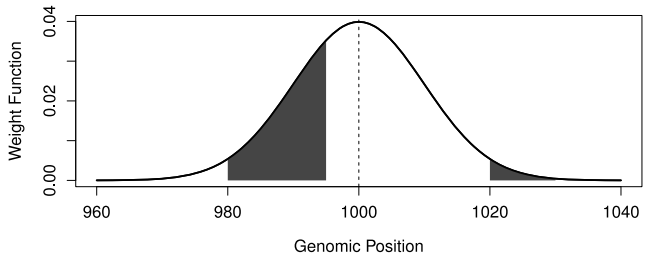

FIG. 1. *Gene weight example*: *for the* j*th SNP at position* $s_j = 1000$ *and two surrounding genes* a *and* b *spanning* $(980, 995)$ *and* $(1020, 1030)$, *we obtain*, *if setting* $\phi = 10$, *weights of* $w_{j,a} = 0.29$ *and* $w_{j,b} = 0.02$, *respectively.*

In the prior distribution for θ_j, we incorporate information from relevant genomic regions. The most common instance of such regions are *genes*, and so we focus on these regions in what follows. Thus, given a list of G genes with gene *relevances* (see Section 3.2 for some choices of definitions), $\mathbf{r} = [r_1, r_2, \ldots, r_G]$, and weights, $\mathbf{w}_j(\phi) = [w_{j,1}, w_{j,2}, \ldots, w_{j,G}]$, the prior on θ_j is

$$\theta_j \overset{\text{ind}}{\sim} \mathsf{Bern}\big(\text{logit}^{-1}\big(\xi_0 + \xi_1 \mathbf{w}_j(\phi)^\top \mathbf{r}\big)\big) \qquad \text{for } j = 1, \ldots, p. \tag{3}$$

The weights $\mathbf{w}_j$ are defined using the structure of the SNPs and genes and aim to account for gene lengths and their proximity to markers as a function of a spatial parameter ϕ, as we see in more detail next.

3.1. *Gene weights*. To control how much a gene can contribute to the prior probability of association for a SNP based on the gene's length and distance to that SNP, we introduce a *range* parameter $\phi > 0$. Consider a gene g that spans genomic positions g_l to g_r, and the jth marker at genomic position s_j; the gene weight $w_{j,g}$ is then

$$w_{j,g} = \int_{g_l}^{g_r} \frac{1}{\sqrt{2\pi\phi^2}} \exp\left\{-\frac{(x - s_j)^2}{2\phi^2}\right\} \mathrm{d}x.$$

Generating gene weights for a particular SNP is equivalent to centering a Gaussian curve at that SNP's position on the genome with standard deviation equal to ϕ and computing the area under that curve between the start and end points of each gene. Figure 1 shows an example. As $\phi \to 0$, the weight that each gene contributes to a particular SNP becomes an indicator function for whether or not it covers that SNP; as $\phi \to \infty$, the weights decay to zero. Intermediate values of ϕ allow then for a variety of weights in $[0, 1]$ that encode *spatial* information about gene lengths and gene proximities to SNPs. In Section 5.1 we discuss a method to select ϕ.

According to (3), it might be possible for multiple, possibly overlapping, genes that are proximal to SNP j to boost θ_j. To avoid this effect, we take two precautions. First, we break genes into nonoverlapping genomic blocks and define the

relevance of a block as the mean gene relevance of all genes that cover the block. Second, we normalize the gene weight contributions to θ_j in (3), $\mathbf{w}_j(\phi)^\top \mathbf{r}$, such that $\max_j \mathbf{w}_j(\phi)^\top \mathbf{r} = 1$. This way, it is possible to compare estimates of ξ_1 across different gene weight and relevance schemes.

3.2. *Gene relevances*. We allow for the further strengthening or diminishing of particular gene weights using gene relevances $\mathbf{r}$. If we set $\mathbf{r} = \mathbf{1}_G$ and allow for all genes to be uniformly relevant, then we have a "noninformative" case. Alternatively, if we have some reason to believe that certain genes are more relevant to a particular trait than others, for instance, on the basis of previous research or prior knowledge from an expert, then we can encode these beliefs through $\mathbf{r}$. In particular, we recommend using either text-mining techniques [e.g., Al-Mubaid and Singh (2010)] to quantify the relevance of a gene to a particular disease based on citation counts in the literature or relevance scores compiled from search hits and citation linking the trait of interest to genes [e.g., MalaCards (2014)].

4. Model fitting and inference. The ultimate goal of our model is to perform inference on the posterior probability of association for SNPs. However, these probabilities are not available in a closed form, and so we must resort to Markov chain Monte Carlo techniques such as Gibbs sampling to draw samples from the posterior distributions of the model parameters and use them to estimate $\mathbb{P}(\theta_j = 1|\mathbf{y})$. Unfortunately, these techniques can be slow to iterate and converge, especially when the number of model parameters is large [Cowles and Carlin (1996)]. Thus, to make our model more computationally feasible, we propose first filtering out markers to reduce the size of the original data set in a strategy similar to distilled sensing [Haupt, Castro and Nowak (2011)], and then applying a Gibbs sampler to only the remaining SNPs.

To this end, we design an EM algorithm based on the hierarchical model above that uses all SNP data simultaneously to quickly find an approximate mode of the posterior distribution on β and σ^2 while regarding θ as missing data. Then, for the filtering step, we iterate between (1) removing a fraction of the markers that have the lowest conditional probabilities of association and (2) refitting using the EM procedure until the predictions of the filtered model degrade. In our analyses we filtered 25% of the markers at each iteration to arrive at estimates β^* and stopped if $\max_i |y_i - \text{logit}^{-1}(\mathbf{x}_i^\top \beta^*)| > 0.5$. Next, we discuss the EM algorithm and the Gibbs sampler, and offer guidelines for selecting the other parameters of the model in Section 5.

4.1. *EM algorithm*. We treat θ as a latent parameter and build an EM algorithm accordingly. If $\ell(\mathbf{y}, \theta, \beta, \sigma^2) = \log \mathbb{P}(\mathbf{y}, \theta, \beta, \sigma^2)$, then for the M-steps on β and σ^2 we maximize the expected log joint $Q(\beta, \sigma^2; \beta^{(t)}, (\sigma^2)^{(t)}) =$

$\mathbb{E}_{\theta|\mathbf{y},X;\beta^{(t)},(\sigma^2)^{(t)}}[\ell(\mathbf{y},\theta,\beta,\sigma^2)]$. The log joint distribution ℓ, up to a normalizing constant, is

$$\begin{aligned}\ell(\mathbf{y},\theta,\beta,\sigma^2) &= \sum_{i=1}^{n} y_i \mathbf{x}_i^\top \beta - \log\bigl(1+\exp\{\mathbf{x}_i^\top\beta\}\bigr) \\ &\quad - \frac{p+1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{j=0}^{p}\beta_j^2\biggl(\frac{\theta_j}{\kappa}+1-\theta_j\biggr) \\ &\quad - (\nu+1)\log\sigma^2 - \frac{\lambda}{\sigma^2},\end{aligned} \tag{4}$$

and so, at the tth iteration of the procedure, for the E-step we just need $\langle\theta_j\rangle^{(t)} \doteq \mathbb{E}_{\theta|\mathbf{y};\beta^{(t)},(\sigma^2)^{(t)}}[\theta_j]$. But since

$$\langle\theta_j\rangle = \mathbb{P}\bigl(\theta_j=1|\mathbf{y},\beta,\sigma^2\bigr) = \frac{\mathbb{P}(\theta_j=1,\beta_j|\sigma^2)}{\mathbb{P}(\theta_j=0,\beta_j|\sigma^2)+\mathbb{P}(\theta_j=1,\beta_j|\sigma^2)},$$

then

$$\operatorname{logit}\langle\theta_j\rangle = \log\frac{\mathbb{P}(\theta_j=1,\beta_j|\sigma^2)}{\mathbb{P}(\theta_j=0,\beta_j|\sigma^2)} = -\frac{1}{2}\log\kappa - \frac{\beta_j^2}{2\sigma^2}\biggl(\frac{1}{\kappa}-1\biggr) + \xi_0 + \xi_1\mathbf{w}_j^\top\mathbf{r} \tag{5}$$

for $j=1,\ldots,p$ and $\langle\theta_0\rangle \doteq 1$.

To update β and σ^2, we employ conditional maximization steps [Meng and Rubin (1993)], similar to cyclic gradient descent. From (4) we see that the update for σ^2 follows immediately from the mode of an inverse gamma distribution conditional on $\beta^{(t)}$:

$$\bigl(\sigma^2\bigr)^{(t+1)} = \frac{\frac{1}{2}\sum_{j=0}^{p}\bigl(\beta_j^{(t)}\bigr)^2\bigl(\frac{\langle\theta_j\rangle^{(t)}}{\kappa}+1-\langle\theta_j\rangle^{(t)}\bigr)+\lambda}{\frac{p+1}{2}+\nu+1}. \tag{6}$$

The terms in (4) that depend on β come from the log likelihood of $\mathbf{y}$ and from the expected prior on β, $\beta\sim N(0,\Sigma^{(t)})$, where

$$\Sigma^{(t)} = \operatorname{Diag}\biggl(\frac{\sigma^2}{\langle\theta_j\rangle^{(t)}/\kappa+1-\langle\theta_j\rangle^{(t)}}\biggr).$$

Since updating β is equivalent here to fitting a ridge-regularized logistic regression, we exploit the usual iteratively reweighted least squares (IRLS) algorithm [MacCullagh and Nelder (1989)]. Setting $\mu^{(t)}$ as the vector of expected responses with $\mu_i^{(t)} = \operatorname{logit}^{-1}(\mathbf{x}_i^\top\beta^{(t)})$ and $W^{(t)} = \operatorname{Diag}(\mu_i^{(t)}(1-\mu_i^{(t)}))$ as the variance weights, the update for β is then

$$\beta^{(t+1)} = \bigl(X^\top W^{(t)}X + \bigl(\Sigma^{(t)}\bigr)^{-1}\bigr)^{-1}\bigl(X^\top W^{(t)}X\beta^{(t)} + X^\top\bigl(\mathbf{y}-\mu^{(t)}\bigr)\bigr), \tag{7}$$

where we substitute $(\sigma^2)^{(t)}$ for σ^2 in the definition of $\Sigma^{(t)}$.

Rank truncation of design matrix. Computing and storing the inverse of the $(p+1)$-by-$(p+1)$ matrix $X^\top W^{(t)} X + (\Sigma^{(t)})^{-1}$ in (7) is expensive since p is large. To alleviate this problem, we replace X with a rank truncated version based on its singular value decomposition $X = UDV^\top$. More specifically, we take the top l singular values and their respective left and right singular vectors, and so, if $D = \mathrm{Diag}(d_i)$ and $\mathbf{u}_i$ and $\mathbf{v}_i$ are the ith left and right singular vectors, respectively, then

$$X = UDV^\top = \sum_{i=1}^{n} d_i \mathbf{u}_i \mathbf{v}_i^\top \approx \sum_{i=1}^{l} d_i \mathbf{u}_i \mathbf{v}_i^\top = U_{(l)} D_{(l)} V_{(l)}^\top,$$

where $D_{(l)}$ is the lth order diagonal matrix with the top l singular values and $U_{(l)}$ (n-by-l) and $V_{(l)}$ ($(p+1)$-by-l) contain the respective left and right singular vectors. We select l by controlling the mean squared error: l should be large enough such that $\|X - U_{(l)} D_{(l)} V_{(l)}^\top\|_F / (n(p+1)) < 0.01$.

Since $X^\top W^{(t)} X \approx V_{(l)} D_{(l)} U_{(l)}^\top W^{(t)} U_{(l)} D_{(l)} V_{(l)}^\top$, we profit from the rank truncation by defining the (upper) Cholesky factor C_w of $D_{(l)} U_{(l)}^\top W^{(t)} U_{(l)} D_{(l)}$ and $S = C_w V_{(l)}^\top$ so that

$$\begin{aligned}\left(X^\top W^{(t)} X + \left(\Sigma^{(t)}\right)^{-1}\right)^{-1} &\approx \left(S^\top S + \left(\Sigma^{(t)}\right)^{-1}\right)^{-1} \\ &= \Sigma^{(t)} - \Sigma^{(t)} S^\top \left(I_l + S\Sigma^{(t)} S^\top\right)^{-1} S\Sigma^{(t)}\end{aligned} \tag{8}$$

by the Kailath variant of the Woodbury identity [Petersen and Pedersen (2012)]. Now we just need to store and compute the inverse of the lth order square matrix $I_l + S\Sigma^{(t)} S^\top$ to obtain the updated $\beta^{(t+1)}$ in (7).

4.2. *Gibbs sampler.* After obtaining results from the EM filtering procedure, we proceed to analyze the filtered data set by sampling from the joint posterior $\mathbb{P}(\theta, \beta, \sigma^2 | \mathbf{y})$ using Gibbs sampling. We iterate sampling from the conditional distributions

$$[\sigma^2 | \theta, \beta, \mathbf{y}], \qquad [\theta | \beta, \sigma^2, \mathbf{y}] \quad \text{and} \quad [\beta | \theta, \sigma^2, \mathbf{y}]$$

until assessed convergence.

We start by taking advantage of the conjugate prior for σ^2 and draw each new sample from

$$\sigma^2 | \theta, \beta, \mathbf{y} \sim \mathsf{IG}\left(\nu + \frac{p+1}{2}, \lambda + \frac{1}{2} \sum_{j=0}^{p} \beta_j^2 \left(\frac{\theta_j}{\kappa} + 1 - \theta_j\right)\right).$$

Sampling θ is also straightforward: since the θ_j are independent given β_j,

$$\theta_j | \beta, \sigma^2, \mathbf{y} \overset{\text{ind}}{\sim} \mathsf{Bern}(\langle \theta_j \rangle) \qquad \text{for } j = 1, \ldots, p,$$

with $\langle\theta_j\rangle$ as in (5). Sampling β, however, is more challenging since there is no closed-form distribution based on a logistic regression, but we use a data augmentation scheme proposed by Polson, Scott and Windle (2013). This method has been noted to perform well when the model has a complex prior structure and the data have a group structure, and so we believe it is appropriate for the Spatial Boost model.

Thus, to sample β conditional on θ, σ^2, and $\mathbf{y}$, we first sample latent variables ω from a Pólya–Gamma distribution,

$$\omega_i|\beta \sim \mathsf{PG}\big(1, \mathbf{x}_i^\top \beta\big), \qquad i = 1, \ldots, n,$$

and then, setting $\Omega = \operatorname{Diag}(\omega_i)$, $\Sigma = \operatorname{Diag}(\sigma^2(\theta_j\kappa + 1 - \theta_j))$, and $V_\beta = X^\top \Omega X + \Sigma^{-1}$, sample

$$\beta|\omega, \theta, \sigma^2, \mathbf{y} \sim N\big(V_\beta^{-1} X^\top (\mathbf{y} - 0.5 \cdot \mathbf{1}_n), V_\beta^{-1}\big).$$

We note that the same rank truncation used in the EM algorithm from the previous section works here, and we gain more computational efficiency by using an identity similar to (8) when computing and storing V_β^{-1}.

4.3. *Centroid estimation.* To conduct inference on θ, we follow statistical decision theory [Berger (1985)] and define an estimator based on a generalized Hamming loss function $H(\theta, \widetilde{\theta}) = \sum_{j=1}^p h(\theta_j, \widetilde{\theta}_j)$,

$$\begin{aligned} \widehat{\theta}_C &= \operatorname*{arg\,min}_{\widetilde{\theta} \in \{0,1\}^p} \mathbb{E}_{\theta|\mathbf{y}}\big[H(\theta, \widetilde{\theta})\big] \\ &= \operatorname*{arg\,min}_{\widetilde{\theta} \in \{0,1\}^p} \mathbb{E}_{\theta|\mathbf{y}}\left[\sum_{j=1}^p h(\theta_j, \widetilde{\theta}_j)\right]. \end{aligned} \tag{9}$$

We assume that h has symmetric error penalties, $h(0,1) = h(1,0)$, and that $h(1,0) > \max\{h(0,0), h(1,1)\}$, that is, the loss for a false positive or negative is higher than for a true positive and true negative. In this case, we can define a *gain* function g by subtracting each entry in h from $h(1,0)$ and dividing by $h(1,0) - h(0,0)$:

$$g(\theta_j, \widetilde{\theta}_j) = \begin{cases} 1, & \theta_j = \widetilde{\theta}_j = 0, \\ 0, & \theta_j \neq \widetilde{\theta}_j, \\ \gamma \doteq \dfrac{h(1,0) - h(1,1)}{h(1,0) - h(0,0)}, & \theta_j = \widetilde{\theta}_j = 1. \end{cases}$$

Gain $\gamma > 0$ represents a sensitivity-specificity trade-off; if $h(0,0) = h(1,1)$, that is, if true positives and negatives have the same relevance, then $\gamma = 1$.

Let us define the marginal posteriors $\pi_j \doteq \mathbb{P}(\theta_j = 1|\mathbf{y})$. The above estimator is then equivalent to

$$\begin{aligned}\widehat{\theta}_C &= \underset{\widetilde{\theta}\in\{0,1\}^p}{\arg\max}\, \mathbb{E}_{\theta|\mathbf{y}}\left[\sum_{j=1}^{p} g(\theta_j, \widetilde{\theta}_j)\right] \\ &= \underset{\widetilde{\theta}\in\{0,1\}^p}{\arg\max} \sum_{j=1}^{p} (1-\widetilde{\theta}_j)(1-\pi_j) + \gamma\widetilde{\theta}_j\theta_j \\ &= \underset{\widetilde{\theta}\in\{0,1\}^p}{\arg\max} \sum_{j=1}^{p} \left(\pi_j - \frac{1}{1+\gamma}\right)\widetilde{\theta}_j,\end{aligned}$$

which can be obtained position-wise,

$$(\widehat{\theta}_C)_j = I\left(\pi_j - \frac{1}{1+\gamma} \geq 0\right). \tag{10}$$

The estimator in (9) is known as the *centroid estimator*; in contrast to maximum a posteriori (MAP) estimators that simply identify the highest peak in a posterior distribution, centroid estimators can be shown to be closer to the mean than to a mode of the posterior space, and so offer a better summary of the posterior distribution [Carvalho and Lawrence (2008)]. Related formulations of centroid estimation for binary spaces in (10) have been proposed in many bioinformatics applications in the context of maximum expected accuracy [Hamada and Asai (2012)]. Moreover, if $\gamma = 1$, then $\widehat{\theta}_C$ is simply a consensus estimator and coincides with the median probability model estimator of Barbieri and Berger (2004).

Finally, we note that the centroid estimator can be readily obtained from MCMC samples $\theta^{(1)}, \ldots, \theta^{(N)}$ since we just need to estimate the marginal posterior probabilities $\widehat{\pi}_j = \sum_{s=1}^{N} \theta_j^{(s)}/N$ and apply them to (10).

5. Guidelines for selecting prior parameters. Since genome-wide association is a large-p-small-n problem, we rely on adequate priors to guide the inference and overcome ill-posedness. In this section we provide guidelines for selecting hyperpriors κ in the slab variance of β, and ϕ, ξ_0, and ξ_1 in the prior for θ.

5.1. *Selecting* ϕ. Biologically, some locations within a chromosome may be more prone to recombination events and consequently to relatively higher linkage disequilibrium. LD can be characterized as correlation in the genotypes, and since we analyze the entire genome, high correlation in markers within a chromosome often results in poor coefficient estimates for the logistic regression model in (1). To account for potentially varying spatial relationships across the genome, we exploit the typical correlation pattern in GWAS data sets to suggest a value for ϕ that properly encodes the spatial relationship between markers and genes in a particular region as a function of genomic distance. To this end, we propose the following procedure to select ϕ:

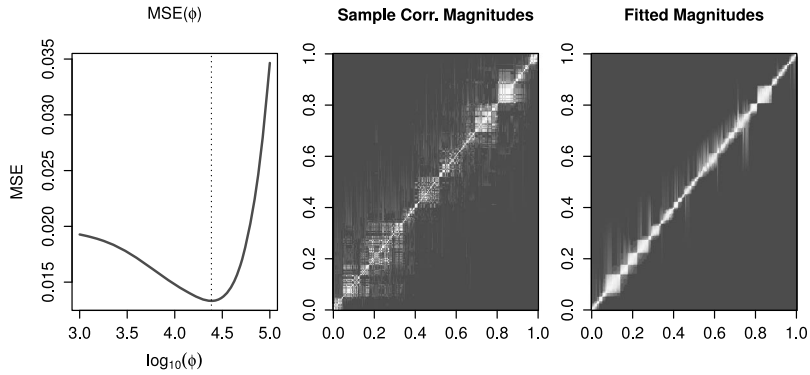

FIG. 2. *Example of selection of ϕ: when using the proposed values of $|\rho_{i,j}|$ to fit the sample correlation magnitudes, we obtain an optimal choice of $\phi = 13{,}530$ for a random window.*

1. Divide each chromosome into regions such that the distance between the SNPs in adjacent regions is at least the average length of a human gene, or 30,000 base pairs [Technology Department Carnegie Library of Pittsburgh (2002)]. The resulting regions will be, on average, at least a gene's distance apart from each other and may possibly exhibit different patterns of correlation.

2. Merge together any adjacent regions that cover the same gene. Although the value of ϕ depends on each region, we want the meaning of the weights assigned from a particular gene to SNPs in the Spatial Boost model to be consistent across regions. As a practical example, by applying the first two steps of the preprocessing procedure on chromosome 1, we obtain 1299 windows of varying sizes ranging from 1 to 300 markers.

3. Iterate over each region and select a value of ϕ that best fits the magnitude of the genotype correlation between any given pair of SNPs as a function of the distance between them. We propose using the normal curve given in the definition of the gene weights to first fit the magnitudes, and then using the mean squared error between the magnitudes in the sample correlation matrix of a region and the magnitudes in the fitted correlation matrix as a metric to decide the optimal value of ϕ. In particular, given two SNPs located at positions s_i and s_j, we relate the magnitude of the correlation between SNPs i and j to the area

$$|\rho_{i,j}|(\phi) = 2\Phi\left(-\frac{|s_i - s_j|}{\phi}\right),$$

where Φ is the standard normal cumulative function.

Figure 2 shows an example of application to chromosome 1 based on data from the case study discussed in Section 6. We note that the mean squared error criterion places more importance on fitting relatively larger magnitudes close to the diagonal of the image matrix, and so there is little harm in choosing a moderate value for ϕ that best fits the magnitudes of dense groups of correlated SNPs in close proximity.

5.2. *Selecting ξ_0 and ξ_1.* According to the centroid estimator in (10), the jth SNP is identified as associated if $\pi_j \geq (1+\gamma)^{-1}$. Following a similar criterion, but with respect to the conditional posteriors, we have $\mathbb{P}(\theta_j = 1|\mathbf{y}, \beta, \sigma^2) = \langle\theta_j\rangle \geq (1+\gamma)^{-1}$, and so, using (5),

$$\operatorname{logit}\langle\theta_j\rangle = -\frac{1}{2}\log\kappa + \xi_0 + \xi_1 \mathbf{w}_j^\top \mathbf{r} + \frac{\beta_j^2}{2\sigma^2}\left(1 - \frac{1}{\kappa}\right) \geq -\log\gamma.$$

After some rearrangements, we see that, in terms of β_j, this criterion is equivalent to $\beta_j^2 \geq \sigma^2 s_j^2$ with

$$s_j^2 \doteq \frac{2\kappa}{\kappa - 1}\left(\frac{1}{2}\log\kappa - \xi_0 - \xi_1 \mathbf{w}_j^\top \mathbf{r} - \log\gamma\right), \tag{11}$$

that is, we select the jth marker if β_j is more than s_j "spike" standard deviations σ away from zero.

This interpretation based on the EM formulation leads to a meaningful criterion for defining ξ_0 and ξ_1: we just require that $\min_{j=1,\ldots,p} s_j^2 \geq s^2$, that is, that the smallest number of standard deviations is at least $s > 0$. Since $\max_{j=1,\ldots,p} \mathbf{w}_j^\top \mathbf{r} = 1$,

$$\min_{j=1,\ldots,p} s_j^2 = \frac{2\kappa}{\kappa - 1}\left(\frac{1}{2}\log\kappa - \xi_0 - \xi_1 - \log\gamma\right) \geq s^2,$$

and so

$$\xi_1 \leq \frac{1}{2}\log\kappa - \xi_0 - \log\gamma - \frac{s^2}{2}\left(1 - \frac{1}{\kappa}\right). \tag{12}$$

For a more stringent criterion, we can take the minimum over κ in the right-hand side of (12) by setting $\kappa = s^2$. When setting ξ_1 it is also important to keep in mind that ξ_1 is the largest allowable gene boost or, better, increase in the log-odds of a marker being associated to the trait.

Since ξ_0 is related to the prior probability of a SNP being associated, we can take ξ_0 to be simply the logit of the fraction of markers that we expect to be associated a priori. However, for consistency, since we want $\xi_1 \geq 0$, we also require that the right-hand side of (12) be non-negative, and so

$$\xi_0 + \log\gamma \leq \frac{1}{2}\log\kappa - \frac{s^2}{2}\left(1 - \frac{1}{\kappa}\right). \tag{13}$$

Equation (13) constrains ξ_0 and γ jointly, but we note that the two parameters have different uses: ξ_0 captures our prior belief on the probability of association and is thus part of the model specification, while γ defines the sensitivity-specificity trade-off that is used to identify associated markers, and is thus related to model inference.

As an example, if $\gamma = 1$ and we set $s = 4$, then the bound in (12) with $\kappa = s^2$ is $\log(s^2)/2 - s^2(1 - 1/s^2)/2 = -6.11$. If we expect 1 in 10,000 markers to be associated, then we have $\xi_0 = \operatorname{logit}(10^{-4}) = -9.21 < -6.11$ and the bound (13) is respected. The upper bound for ξ_1 in (12) is thus 3.10.

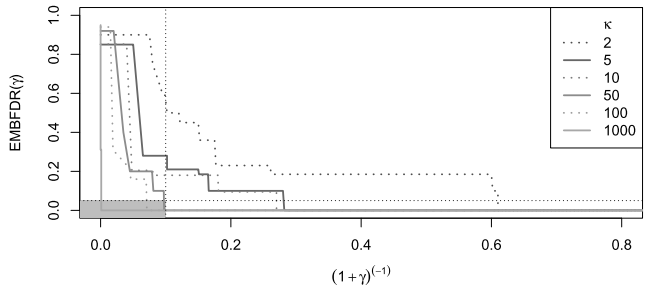

FIG. 3. *When analyzing a data set generated for a simulation study as described in Section 6, we inspect the behavior of the BFDR as a function of γ for various values of κ and see that a choice of $\kappa = 1000$ would be appropriate to achieve a BFDR no greater than 0.05 when using a threshold of $(1+\gamma)^{-1} = 0.1$.*

5.3. *Selecting κ.* We propose using a metric similar to the Bayesian false discovery rate [BFDR, Whittemore (2007)] to select κ. The BFDR of an estimator is computed by taking the expected value of the false discovery proportion under the marginal posterior distribution of θ:

$$\mathrm{BFDR}(\widehat{\theta}) = \mathbb{E}_{\theta|\mathbf{y}}\left[\frac{\sum_{j=1}^{p}\widehat{\theta}_j(1-\theta_j)}{\sum_{j=1}^{p}\widehat{\theta}_j}\right] = \frac{\sum_{j=1}^{p}\widehat{\theta}_j(1-\pi_j)}{\sum_{j=1}^{p}\widehat{\theta}_j}.$$

Since, as in the previous section, we cannot obtain estimates of $\mathbb{P}(\theta_j = 1|\mathbf{y})$ just by running our EM algorithm, we consider instead an alternative metric that uses the conditional posterior probabilities of association given the fitted parameters, $\langle\theta_j\rangle = \mathbb{P}(\theta_j = 1|\mathbf{y}, \widehat{\beta}_{\mathrm{EM}}, \widehat{\sigma}^2_{\mathrm{EM}})$. We call this new metric EMBFDR:

$$\mathrm{EMBFDR}(\widehat{\theta}) = \frac{\sum_{j=1}^{p}\widehat{\theta}_j(1-\langle\theta_j\rangle)}{\sum_{j=1}^{p}\widehat{\theta}_j}.$$

Moreover, by the definition of the centroid estimator in (10), we can parameterize the centroid EMBFDR using γ:

$$\mathrm{EMBFDR}\big(\widehat{\theta}_C(\gamma)\big) = \mathrm{EMBFDR}(\gamma) = \frac{\sum_{j=1}^{p} I[\langle\theta_j\rangle \geq (1+\gamma)^{-1}](1-\langle\theta_j\rangle)}{\sum_{j=1}^{p} I[\langle\theta_j\rangle \geq (1+\gamma)^{-1}]}.$$

We can now analyze a particular data set using a range of values for κ, and subsequently make plots of the EMBFDR metric as a function of the threshold $(1+\gamma)^{-1}$ or as a function of the proportion of SNPs retained after the EM filter step. Thus, by setting an upper bound for a desired value of the EMBFDR, we can investigate these plots and determine an appropriate choice of κ and an appropriate range of values of γ. In Figure 3 we illustrate an application of this criterion. We note that the EMBFDR has broader application to Bayesian variable selection models and can be a useful metric to guide the selection of tuning parameters, in particular, the spike-and-slab variance separation parameter κ.

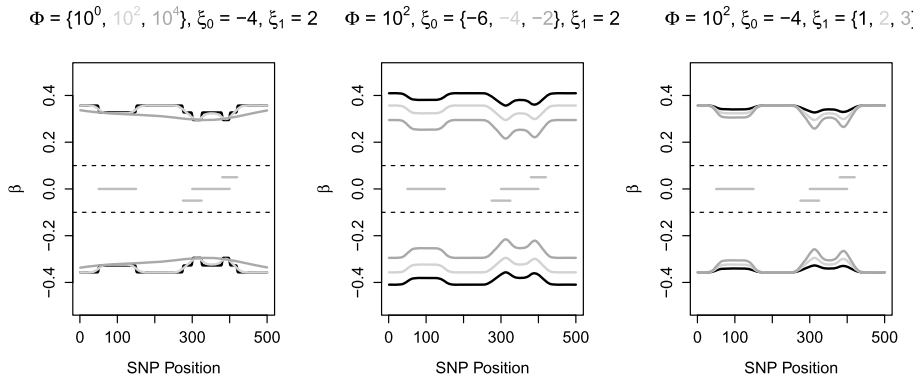

FIG. 4. *We illustrate the effect of varying ϕ, ξ_0, and ξ_1 on the thresholds on the posterior effect sizes, β_j, in a simple window containing a single gene in isolation and a group of three overlapping genes. On the left, we vary ϕ and control the smoothness of the thresholds. In the middle, we vary ξ_0 and control the magnitude of the thresholds or, in other words the number of standard deviations (σ) away from zero at which they are placed. On the right, we vary ξ_1 and control the sharpness of the difference in the thresholds between differently weighted regions of the window. For this illustration, we set $\sigma^2 = 0.01$, $\kappa = 100$, and $\gamma = 1$. We mark the distance σ away from the origin with black dashed lines.*

5.4. *Visualizing the relationship between SNPs and genes.* For a given configuration of κ, γ, and σ^2, we can plot the bounds $\pm\sigma s_j$ on β_j and inspect the effect of parameters ϕ, ξ_0, and ξ_1. SNPs that are close to relevant genes have thresholds that are relatively lower in magnitude; they need a relatively smaller (in magnitude) coefficient to be selected for the final model. With everything else held fixed, as $\phi \to \infty$ the boost received from the relevant genes will decrease to zero and our model will coincide with a basic version of Bayesian variable selection where $\theta_j \overset{\text{i.i.d.}}{\sim} \text{Bern}(\text{logit}^{-1}(\xi_0))$. We demonstrate this visualization on a mock chromosome in Figure 4.

6. Empirical studies. We conduct two simulation studies. First, we compare the performance of our method to other well-known methods including single SNP tests, LASSO, fused LASSO, group LASSO, the penalized unified multiple-locus association (PUMA) suite of Hoffman, Logsdon and Mezey (2013), and the Bayesian sparse linear mixed model (BSLMM) of Zhou, Carbonetto and Stephens (2013). Then we assess the robustness of our method to misspecifications of the range parameter, ϕ, and gene relevances. We describe each study in detail below, but we first explain how the data is simulated in each scenario.

6.1. *Simulation study details.* To provide a fair comparison across methods and to realistically assess the robustness of our method to misspecifications, we designed our simulation study based on real-life genotypical data and current gene and marker annotations. Specifically, to keep a representative LD structure, we

sample whole-chromosome individual genotypes by subsampling individual data provided by the 1000 Genomes Project Consortium et al. (2012); gene weights are computed based on gene lengths and positions in the hg19 reference, while marker positions are taken from actual SNP array designs in the WTCCC studies. We consider two studies: a "noninformative" setup where the gene relevances are uniformly set to one and $\phi = 10^8$ so that marker relevance scores $\mathbf{w}_j(\phi)^\top \mathbf{r}$ are small and close to uniform, and a mild boost effect of $\xi_1 = 1$; and an "informative" study with gene relevances taken as search hit scores from MalaCards (2014) and $\phi = 10^4$, a frequent value when adopting the procedure in Section 5.1, and a stronger boost effect, $\xi_1 = 5$. These two studies are extreme with respect to marker relevance scores—a function of gene relevances and genomic range—and spatial boost effects, and aim at assessing how robust our model and recommended guidelines are. For instance, when fitting all scenarios we take an informative approach by considering the same gene relevances and genomic range from the "informative" scenario, but adopt a conservative approach by setting $\xi_1 = 1$ as in the "noninformative" scenario.

For both studies we simulated two scenarios for the number of markers p: the "small" scenario comprised chromosome 19 with $p = 4199$ markers, and the "large" scenario contained all $p = 28{,}932$ markers in chromosome 2. Chromosomes 19 and 2 are the smallest and largest in terms of number of markers, respectively. We kept the ratio of p to the number of individuals n at 50, representative of real-life studies, so $n = 85$ in the small scenario and $n = 580$ in the large scenario. In all simulations we fix the number of causal markers $m = 10$ and set the baseline log-odds $\xi_0 = \lfloor \text{logit}(m/p) \rfloor$. For each simulation batch, we first sample uniformly at random n individuals from the 1000 Genomes data set, taking their whole chromosome genotypes according to the small or large scenario, and filter out markers with sampled MAF < 0.05, deemed as rare variants, or > 0.50. Next, we sample m causal markers following (3), with marker relevance scores and ξ_1 taken according to a noninformative or informative scenario. Effect sizes β_j are then sampled to reflect the challenging nature of GWAS: $\beta_j|\theta_j \sim \theta_j N(0, 0.25) + (1-\theta_j)N(0, 0.01)$, that is, small effect sizes for causal markers and relatively large coefficients for noisier noncausal effects. Finally, for each replicate within a batch we sample phenotypes according to (1).

In each simulation scenario and data set below we fit the model as follows: we adopt informative gene relevances from MalaCards and $\phi = 10{,}000$, start with conservative values for the baseline log-odds $\xi_0 = \text{logit}(100/p)$ and the gene boost effect $\xi_1 = 1$, and run the EM filtering process until either the predictive performance starts degrading or at most 10 markers remain. We measure predictive performance using a metric similar to posterior predictive loss [PPL; Gelfand and Ghosh (1998)]: if, at the tth EM iteration, $\widehat{y}_i^{(t)} = \mathbb{E}[y_{i,\text{rep}}|\widehat{\beta}_{\text{EM}}^{(t)}, \mathbf{y}]$ is the ith predicted response, the PPL measure under squared error loss is approximated by

$$\text{PPL}(t) = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n} \text{Var}[y_{i,\text{rep}}|\widehat{\beta}_{\text{EM}}^{(t)}, \mathbf{y}] = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 + \widehat{y}_i(1 - \widehat{y}_i).$$

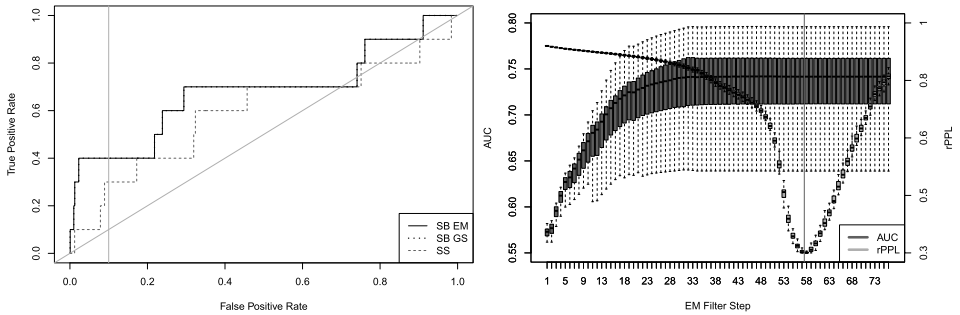

FIG. 5. *Left*: *examples of ROC curves for our method* (*solid line for EM results*, *crosses for Gibbs sampling results*) *and single SNP tests* (*dashed line*); *vertical gray line marks a false positive rate of* 10%. *Right*: *relation between AUC and relative PPL* (*rPPL*) *as the EM filter progresses*, *boxplots over replicates within a simulation batch*; *in this case the EM filter would stop at iteration* 58, *right before rPPL starts increasing*.

The right panel in Figure 5 shows an example of how the relative PPL (rPPL) typically varies as the EM filter advances. We define the rPPL at the tth EM iteration as the ratio between PPL(t) and the PPL of the null model, that is, the model with only the intercept and no marker genotypes as predictors. At the end of the filtering stage we run the Gibbs sampler with $\xi_0 = \text{logit}(m/\widetilde{p})$, where $\widetilde{p}$ is the number of markers retained at the end of the EM filter. Parameter κ is actually elicited at each EM filtering iteration using EMBFDR.

6.2. *Comparison simulation study.* In this study, we generated 10 batches of simulated data, each containing 5 replicates, for a total of 50 simulated data sets for each cross-configuration of small and large scenarios by noninformative and informative studies. After simulating the data, we fit our model and compared its performance in terms of area under the receiver operating characteristic (ROC) curve (AUC) to the usual single SNP tests, LASSO, fused LASSO, group LASSO, PUMA, and BSLMM methods. We used the *penalized* package in R to fit the LASSO and fused LASSO models; we used tenfold cross-validation to determine the optimal values for the penalty terms. Similarly, we used the *gglasso* package in R to fit the group LASSO model where we defined the groups such that any two adjacent SNPs belonged to the same group if they were within 10,000 base pairs of each other; we used tenfold cross-validation to determine the optimal value for the penalty term. Finally, we used the authors' respective software packages to fit the PUMA and BSLMM models.

To calculate the AUC for any one of these methods, we took a final ranking of SNPs based on an appropriate criterion (see more about this below), determined the points on the ROC curve using our knowledge of the true positives and the false positives from the simulated data, and then calculated the area under this curve. For our model, we used either the ranking (in descending order) of $\mathbb{E}[\theta_j | \widehat{\beta}_{\text{EM}}, \widehat{\sigma}^2_{\text{EM}}, y]$

for a particular EM filtering step or $\widehat{\mathbb{P}}(\theta_j = 1|y)$ using the samples obtained by the Gibbs sampler; for the single SNP tests, we used the ranking (in ascending order) of the p-values for each marker's test; for LASSO, fused LASSO, and group LASSO, we used the ranking (in descending order) of the magnitude of the effect sizes of the SNPs in the final model; for the other penalized regression models given by the PUMA program, we used the provided software to compute p-values for each SNP's significance in the final model and used the ranking (in ascending order) of these p-values; for BSLMM, we used the ranking (in descending order) of the final estimated posterior probabilities of inclusion for each SNP in the final model.

The left panel in Figure 5 shows examples of ROC curves from a random simulation replicate. Since the interest in GWA studies is focused on low false positive rates, we also evaluate the performance of the methods with respect to *relative* AUC (rAUC) at some false positive rate f, defined simply as the normalized AUC up to FPR f, that is, $\text{rAUC}(f) = \text{AUC}(f)/f$. According to this criterion, the solid ROC curve in Figure 5 has a clearer advantage over the dashed curve at a 10% FPR. The right panel illustrates how the AUC changes with rPPL and as the EM filtering iterations increase.

We summarize the results in Figure 6. With respect to AUC (top four panels), our methods perform better on average than the other methods in the informative scenario, but only comparably in the noninformative scenario. Note that all models have a modest performance due to the challenging nature of the simulation (but our model has improved performance under less stringent scenarios; see supplementary material [Johnston et al. (2016)] for more details). Most of the gain in our methods comes at the beginning of the ROC curves, that is, at low false positive rates, as exemplified in Figure 5. This becomes more evident if we compare relative AUC in the bottom four panels. We note that the gains are present even in the "null" models with $\xi_1 = 0$, so they stem from the joint modeling of markers. Additional gains in rAUC are more apparent in the informative scenario (bottom row).

The seemingly bad performance of our model in the noninformative scenario indicates that the model can be sensitive to poor marker relevance scores, arising either from meaningless gene relevances or a nonrepresentative genomic range ϕ. This is not surprising given that the inference relies on good prior information in the large-p-small-n regimen; however, as the informative scenario suggests, the model is more robust to misspecification of ξ_1, which can be seen as the overall prior strength of relevance scores. The null model, for instance, often offers the best performances in both AUC and relative AUC, which recommends conservative, low values for ξ_1 in initial analyses. Moreover, the EM filtering procedure contributes to further gains in performance even in noninformative scenarios, that is, with misspecified relevance scores. These gains are even more pronounced with larger causal effect sizes relative to noncausal effects, as we show in supplementary material [Johnston et al. (2016)]; that is, we observe that the EM filter becomes

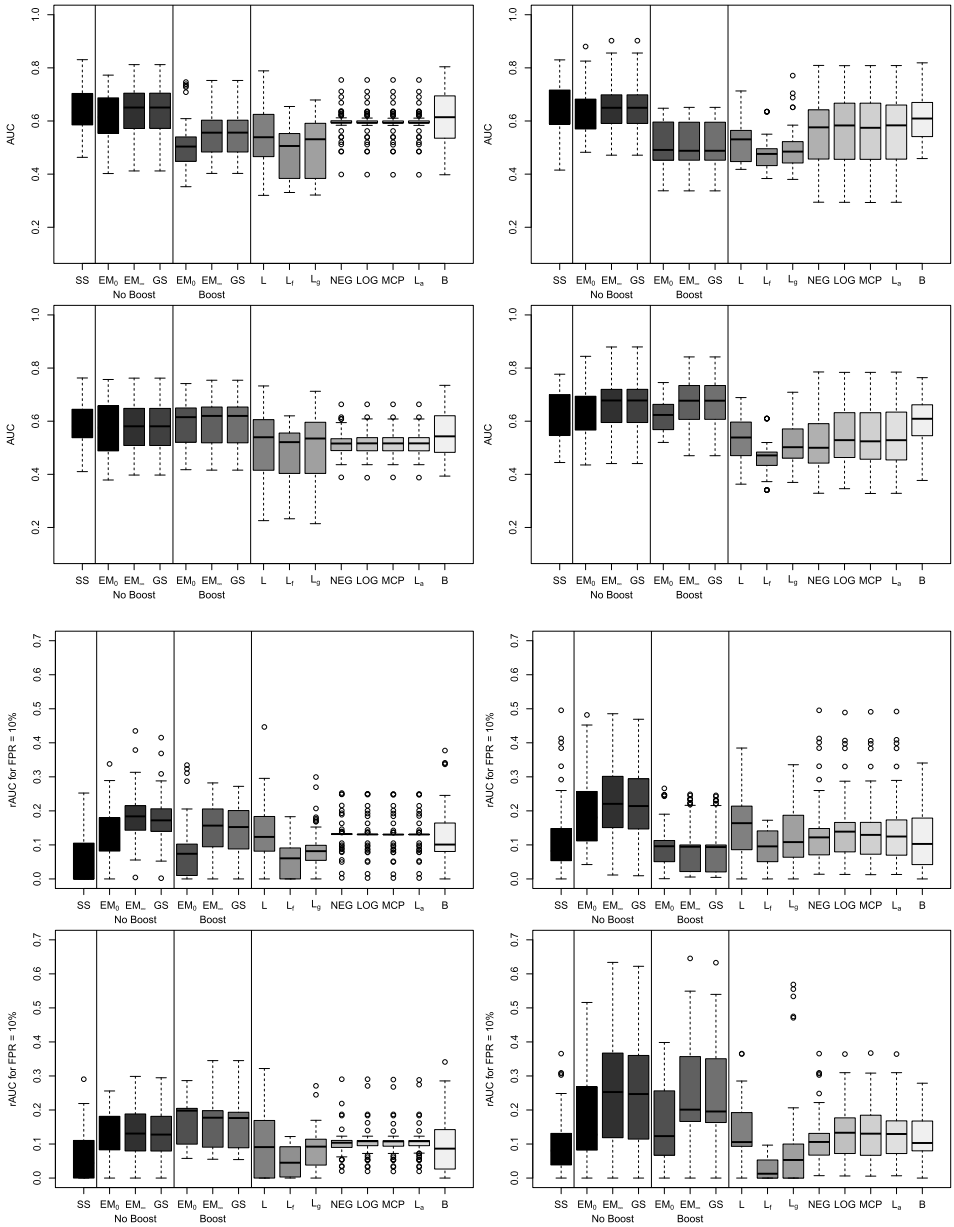

FIG. 6. *Results from the comparison simulation study. Top four panels show AUC, while bottom four panels show relative AUC. Within each block of four panels: left panels show results under noninformative scenario (top) and informative scenario (bottom) for small study, while right panels show respective AUC or rAUC results for the large study. The boxplots in each panel are, left to right, as follows: single SNP (SS) tests; spatial boost "null" (no boost) model at the first (EM_0) and last (EM_∞) iteration of the EM filter and after Gibbs sampling phase; spatial boost "informative" model at first and last EM filtering iterations and after Gibbs sampling phase; LASSO (L); fused LASSO (L_f); grouped LASSO (L_g); PUMA with penalties NEG, LOG, MCP, and adaptive LASSO (L_a); and BSLMM (B).*

more robust to these misspecifications, especially on large scenarios and under low false positive rates, as shown here.

6.3. *Relevance robustness simulation study.* To investigate how robust our model is to misspecifications of gene relevances and genomic range, we focus on their joint effect in defining marker relevance scores and again consider the small (chromosome 19) and large (chromosome 2) scenarios under the informative study, where the gene relevances are more varied. We randomly select one batch from the comparison study to be the reference in each scenario, with its replicates serving as ground truth. We then vary three sampling percentages $\pi \in \{25\%, 50\%, 75\%\}$, and, for each π, we randomly select $\lfloor \pi p \rfloor$ markers uniformly from each reference replicate and sample their relevance scores from the empirical distribution of relevance scores in each scenario. This simulation is replicated 50 times.

Hyperprior parameters ξ_0, ξ_1, and κ were elicited as in Section 6.1. For each simulated replication we then fit our model and assess performance using the AUC, as in the previous study. Figure 7 illustrates the distribution of relevance scores for the markers in chromosomes 19 and 2 and how the performance of the model varies at each sample percentage. The AUC at initial EM iterations degrades with higher percentages for both scenarios, as expected; however, small scenario replicates continue to show a similar pattern at their last EM iterations, as opposed to large scenario replicates that show better and stable results across all percentages. We attribute this discrepancy in robustness to the distribution of relevance scores. As the left panel in Figure 7 shows, the large scenario has a bimodal distribution and low mean score, and so a few markers are relevant; in contrast, the small scenario score distribution has more spread and higher mean, and so many markers can be relevant and influence negatively the fit by calling more false positives. We thus recommend to analyze the resulting distribution of marker relevance scores when eliciting gene relevances and ϕ.

7. Case study. Using data provided by the WTCCC, we analyzed the entire genome (342,502 SNPs total) from a case group of 1999 individuals with rheumatoid arthritis (RA) and a control group of 1504 individuals from the 1958 National Blood Bank data set. For now we addressed the issues of LD, rare variants, and population stratification by analyzing only the SNPs in the Hardy–Weinberg equilibrium [Wigginton, Cutler and Abecasis (2005)] with minor allele frequency greater than 5%. There are 15 SNPs that achieve genome-wide significance when using a Bonferroni multiple testing procedure on the results from a single SNP analysis. Table 1 in supplementary material [Johnston et al. (2016)] provides a summary of these results for comparison to those obtained when using the spatial boost model.

When fitting the spatial boost model, we broke each chromosome into blocks and selected an optimal value of ϕ for each block using our proposed method metric, $|\rho_{i,j}|(\phi)$. We used the EMBFDR to select a choice for κ from the set

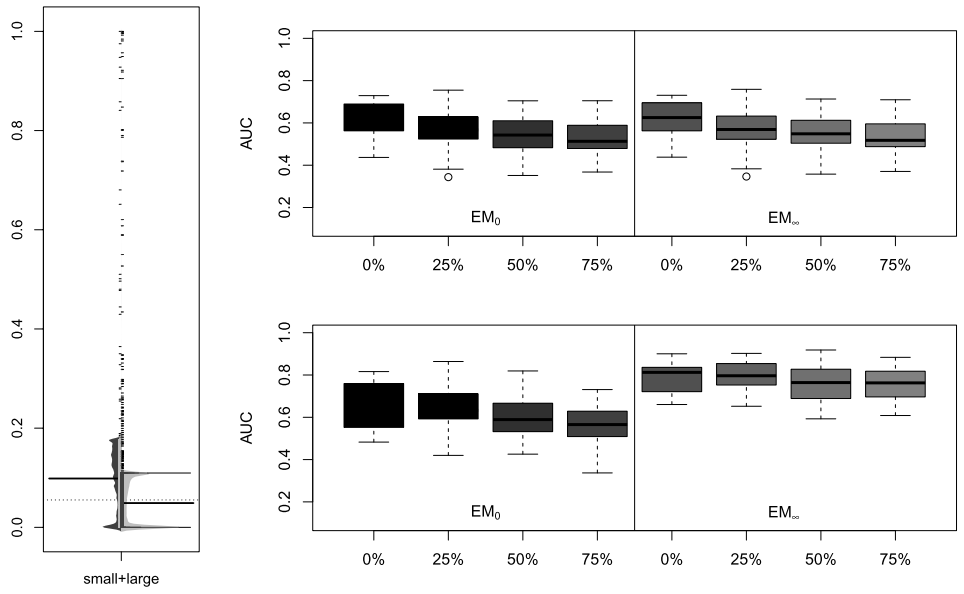

FIG. 7. *Results from the simulation study to assess robustness to marker relevance scores. Left panel shows the distribution of marker relevance score from small and large scenarios, chromosomes* 19 *and* 2*, respectively. Right panels show AUC results for first* (EM_0) *and last* (EM_∞) *filtering iterations across simulation percentages for small* (*top*) *and large* (*bottom*) *scenarios. Null percentages correspond to reference replicates.*

$\{10^2, 10^3, 10^4, 10^5, 10^6\}$ at each step of our model-fitting pipeline so that the BFDR was no greater than 0.05 while retaining no larger than 5% of the total number of SNPs. With a generous minimum standard deviation $s = 1$, we have that trivially $\xi_0 < 0$ from (13), but we set $\xi_0 = -8$ to encode a prior belief that around 100 markers would be associated to the trait on average a priori. The bound on ξ_1 is then $\xi_1 \leq 8$, but we consider log odds-ratio boost effects of $\xi_1 \in \{1, 4, 8\}$. A value of $\xi_1 = 1$ is more representative of low-power GWA studies; however, the larger boost effects offer more weight to our prior information. For comparison, we also fit a model without any gene boost by setting $\xi_1 = 0$ (the "null" model), and also fit two models for each possible value of ξ_1 trying both a noninformative gene relevance vector and a vector based on text-mining scores obtained from MalaCards (2014).

To accelerate the EM algorithm, we rank-truncate X using $l = 3259$ singular vectors; the mean squared error between X and this approximation is less than 1%. We apply the EM filtering 29 times and use PPL to decide when to start the Gibbs sampler. As Figure 8 shows, in all of our fitted models, the PPL decreases slowly and uniformly for the first twenty or so iterations, and then suddenly decreases more sharply for the next five iterations until it reaches a minimum and then begins increasing uniformly until the final iteration, similarly to the behavior depicted in

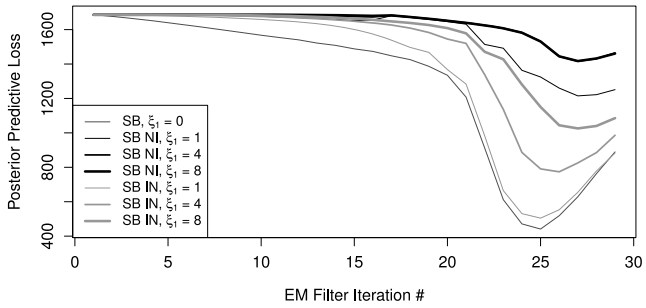

FIG. 8. *Although we run the EM filter until the number of retained markers* <100 (*iteration #*29), *the PPL metric often tells us to keep between* 200 *to* 250 *markers* (*iterations #*25–26).

Figure 5. For comparison to the 15 SNPs that achieve genome-wide significance in the single marker tests, Tables 2–15 in supplementary material [Johnston et al. (2016)] list, for each spatial boost model, the top 15 SNPs at the optimal EM filtering step, that is, the step with the smallest PPL, and the top 15 SNPs based on the posterior samples from our Gibbs sampler when using only the corresponding set of retained markers.

We observe the most overlap with the results of the single SNP tests in our null model where $\xi_1 = 0$ and in our models that use informative priors based on relevance scores from MalaCards. Although there is concordance between these models in terms of the top 15 SNPs, it is noteworthy that we select only a fraction of these markers after running either the EM algorithm or the Gibbs sampler. Based on the results from our simulation study where we observe superior performances for the spatial boost model at low false positive rates, we believe that an advantage of our method is this ability to highlight a smaller set of candidate markers for future investigation.

Indeed, after running our complete analysis, we observe that the usual threshold of 0.5 on $\widehat{\mathbb{P}}(\theta_j = 1|y)$ would result in only the null spatial boost model ($\xi_1 = 0$), the low gene boost noninformative model ($\xi_1 = 1$), and the informative models selecting SNPs for inclusion in their respective final models. The SNPs that occur the most frequently in these final models are the first four top hits from the single SNP tests: rs4718582, rs10262109, rs6679677, and rs664893. The SNP with the highest minor allele frequency in this set is rs6679677; this marker has appeared in several top rankings in the GWAS literature [e.g., Burton et al. (2007)] and is in high LD with another SNP in gene PTPN22 which has been linked to RA [Michou et al. (2007)].

If we consider only the final models obtained after running the EM filter, we see another interesting SNP picked up across the null and informative models: rs1028850. In Figure 9, we show a closer look at the region around this marker and compare the trace of the Manhattan plot with the traces of each spatial boost model's $\mathbb{E}[\theta_j|\widehat{\beta}_{\mathrm{EM}}, \widehat{\sigma}^2_{\mathrm{EM}}, y]$ values at the first iteration of the EM filter. To the best

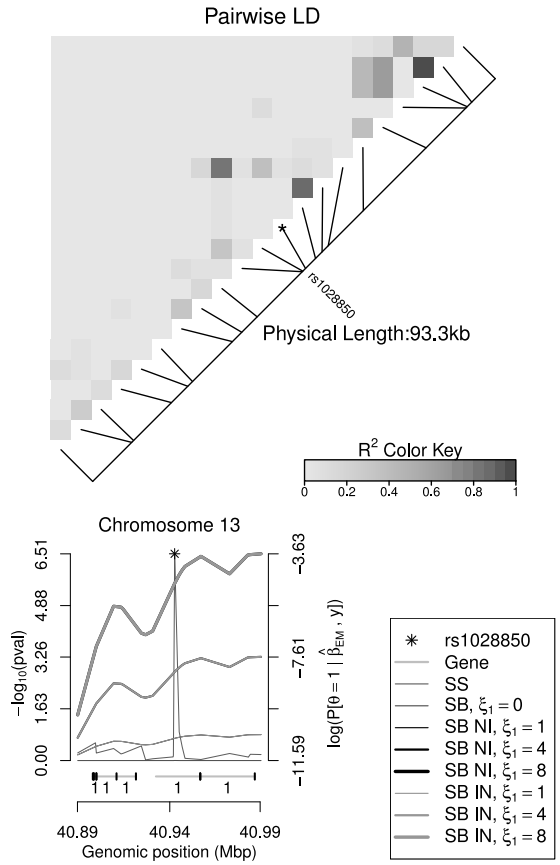

FIG. 9. *Although rs*1028850 *has a relative peak in the Manhattan plot* (*SS*), *it does not achieve genome-wide significance. The spatial boost* (*SB*) *model initially prioritizes markers that are closer to the center of regions rich in genes, but selects rs*1028850 *for inclusion in the final model by the end of the EM filter* (*not shown*) *under several configurations.*

of our knowledge, this marker has not yet been identified as being associated to RA; moreover, it is located inside a nonprotein coding RNA gene, LINC00598, and is close to another gene that has been linked to RA, FOXO1 [Grabiec et al. (2014)].

As we increase the strength of the gene boost term with a noninformative relevance vector, the relatively strong prior likely leads to a mis-prioritization of all SNPs that happen to be located in regions rich in genes. In the supplementary sables (2–15) of supplementary material [Johnston et al. (2016)] we list the lengths of the genes that contain each SNP and we see that indeed the noninformative gene boost models tend to retain SNPs that are near large genes that can offer a generous boost. Perhaps due to prioritizing the SNPs incorrectly in these models, we do not actually select any markers at either the optimal EM filtering step or after running

the Gibbs sampler. However, some of the highest ranking SNPs for these models, rs1982126 and rs6969220, are located in gene PTPRN2, which is interestingly a paralog of PTPN22.

8. Conclusions. We have presented a novel hierarchical Bayesian model for GWAS that exploits the structure of the genome to define SNP-specific prior distributions for the model parameters based on proximities to relevant genes. While it is possible that other "functional" regions are also very relevant—for example, regulatory and highly conserved regions—and that mutations in SNPs influence regions of the genome much farther away—either upstream, downstream, or, through a complex interaction of molecular pathways, even on different chromosomes entirely—we believe that incorporating information about the genes in the immediate surroundings of a SNP is a reasonable place to start.

By incorporating prior information on relevant genomic regions, we focus on well-annotated parts of the genome and were able to identify, in real data, markers that were previously identified in large studies and highlight at least one novel SNP that has not been found by other models. Clearly, validation via other studies of the importance of this marker for RA should be investigated. In addition, as shown in a simulation study, while logistic regression under the large-p-small-n regimen is challenging, the spatial boost model often outperforms simpler models that either analyze SNPs independently or employ a uniform penalty term on the L_1 norm of their coefficients.

Our main point is that we regard a fully joint analysis of all markers as essential to overcome genotype correlations and rare variants. This approach, however, entails many difficulties. From a statistical point of view, the problem is severely ill-posed, so we rely on informative, meaningful priors to guide the inference. From a computational perspective, we also have the daunting task of fitting a large-scale logistic regression, but we make it feasible by reducing the dimension of both data—intrinsically through rank truncation—and parameters—through EM filtering. Moreover, from a practical point of view, we provide guidelines for selecting hyperpriors, reducing dimensionality, and implement the proposed approach using parallelized routines.

From the simulation studies in Section 6 we can further draw two conclusions. First, as reported by other methods such as PUMA, filtering is important; our EM filtering procedure seems to focus on effectively selecting true positives at low false positive rates. This feature of our method is encouraging, since practitioners are often interested in achieving higher sensitivity by focusing on lower false positive rates. Second, because we depend on good informative priors to guide the selection of associated markers, we rely on a judicious choice of hyperprior parameters, in particular, of the range parameter ϕ and how it boosts markers within neighboring genes that are deemed relevant. It is also important to elicit gene relevances from well-curated databases, for example, MalaCards, and to calibrate prior strength according to how significant these scores are.

We have shown that our model performs better than most variable selection methods, but that it can suffer in the case of severe model misspecification. As a way to flag misspecification, we suggest to check monotonicity in a measure of model fit such as PPL as we filter markers using EM. In addition, refining the EM filtering by using a lower threshold (<0.25) at each iteration can help increase performance, especially at lower false positive rates.

When applying the spatial boost model to a real data set, we were able to confidently isolate at least one marker that has previously been linked to the trait as well as find another novel interesting marker that may be related to the trait. This shows that, although we can better explore associations jointly while accounting for gene effects, the spatial boost model still might lack power to detect associations between diseases and SNPs due to the high correlation induced by linkage disequilibrium. In the future we plan to increase the power even further by extending the model to include a data preprocessing step that attempts to formally correct for the collinearity between SNPs.

Acknowledgements. The authors would like to thank the anonymous reviewers for their valuable suggestions and constructive comments that helped shape a much improved final version of the paper. We would also like to thank the Editors for their generous comments and support during the review process. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from `wtccc.org.uk`.

SUPPLEMENTARY MATERIAL

Extended results tables and figures (DOI: 10.1214/16-AOAS907SUPP; .pdf). We provide figures and tables to summarize the results of additional simulation studies with less stringent effect sizes as well as the findings in our case study.

REFERENCES

1000 GENOMES PROJECT CONSORTIUM et al. (2012). An integrated map of genetic variation from 1092 human genomes. *Nature* **491** 56–65.

AL-MUBAID, H. and SINGH, R. K. (2010). A text-mining technique for extracting gene-disease associations from the biomedical literature. *International Journal of Bioinformatics Research and Applications* **6** 270–286.

BALDING, D. J. (2006). A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.* **7** 781–791.

BANSAL, V., LIBIGER, O., TORKAMANI, A. and SCHORK, N. J. (2010). Statistical analysis strategies for association studies involving rare variants. *Nat. Rev. Genet.* **11** 773–785.

BARBIERI, M. M. and BERGER, J. O. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897. MR2065192

BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, New York. MR0804611

BURTON, P. R., CLAYTON, D. G., CARDON, L. R., CRADDOCK, N., DELOUKAS, P., DUNCANSON, A., KWIATKOWSKI, D. P., MCCARTHY, M. I., OUWEHAND, W. H., SAMANI, N. J. et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. *Nature* **447** 661–678.

CARVALHO, L. and LAWRENCE, C. (2008). Centroid estimation in discrete high-dimensional spaces with applications in biology. *Proc. Natl. Acad. Sci. USA* **105** 3209–3214.

COWLES, M. K. and CARLIN, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *J. Amer. Statist. Assoc.* **91** 883–904. MR1395755

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **39** 1–38. MR0501537

EVANGELOU, E. and IOANNIDIS, J. P. (2013). Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.* **14** 379–389.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2010). A note on the group lasso and a sparse group lasso. Technical report, Stanford Univ., Stanford, CA. Available at arXiv:1001.0736.

GELFAND, A. E. and GHOSH, S. K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika* **85** 1–11. MR1627258

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

GRABIEC, A. M., ANGIOLILLI, C., HARTKAMP, L. M., VAN BAARSEN, L. G., TAK, P. P. and REEDQUIST, K. A. (2014). JNK-dependent downregulation of FoxO1 is required to promote the survival of fibroblast-like synoviocytes in rheumatoid arthritis. *Annals of the Rheumatic Diseases* **74** annrheumdis–2013.

GUAN, Y. and STEPHENS, M. (2011). Bayesian variable selection regression for Genome-wide association studies and other large-scale problems. *Ann. Appl. Stat.* **5** 1780–1815. MR2884922

HABIER, D., FERNANDO, R., KIZILKAYA, K. and GARRIC, D. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics* **12** 186.

HAMADA, M. and ASAI, K. (2012). A classification of bioinformatics algorithms from the viewpoint of maximizing expected accuracy (MEA). *J. Comput. Biol.* **19** 532–549. MR2925546

HAUPT, J., CASTRO, R. M. and NOWAK, R. (2011). Distilled sensing: Adaptive sampling for sparse detection and estimation. *IEEE Trans. Inform. Theory* **57** 6222–6235. MR2857969

HEARD, E., TISHKOFF, S., TODD, J. A., VIDAL, M., WAGNER, G. P., WANG, J., WEIGEL, D. and YOUNG, R. (2010). Ten years of genetics and genomics: What have we achieved and where are we heading? *Nat. Rev. Genet.* **11** 723–733.

HOERL, A. and KENNARD, R. (1970). Ridge regression—Applications to nonorthogonal problems. *Technometrics* **12** 69–82.

HOFFMAN, G. E., LOGSDON, B. A. and MEZEY, J. G. (2013). Puma: A unified framework for penalized multiple regression analysis of gwas data. *PLoS Comput. Biol.* **9** e1003101.

IOANNIDIS, J. P., THOMAS, G. and DALY, M. J. (2009). Validating, augmenting and refining genome-wide association signals. *Nat. Rev. Genet.* **10** 318–329.

ISHWARAN, H. and RAO, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773. MR2163158

JOHNSTON, I., HANCOCK, T., MAMITSUKA, H. and CARVALHO, L. (2016). Supplement to "Gene-proximity models for genome-wide association studies." DOI:10.1214/16-AOAS907SUPP.

JORGENSON, E. and WITTE, J. S. (2006). A gene-centric approach to genome-wide association studies. *Nat. Rev. Genet.* **7** 885–891.

KOOPERBERG, C., LEBLANC, M. and OBENCHAIN, V. (2010). Risk prediction using genome-wide association studies. *Genetic Epidemiology* **34** 643–652.

MacCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models* **37**. Chapman and Hall/CRC press, London.

MalaCards (2014). Genes related to rheumatoid arthritis. Available at http://www.malacards.org/card/rheumatoid_arthritis. [Online. accessed 2014-10-01].

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models*. Chapman & Hall, London. MR0727836

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80** 267–278. MR1243503

Michou, L., Lasbleiz, S., Rat, A.-C., Migliorini, P., Balsa, A., Westhovens, R., Barrera, P., Alves, H., Pierlot, C., Glikmans, E. et al. (2007). Linkage proof for ptpn22, a rheumatoid arthritis susceptibility gene and a human autoimmunity gene. *Proc. Natl. Acad. Sci. USA* **104** 1649–1654.

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

Peng, B., Zhu, D., Ander, B. P., Zhang, X., Xue, F., Sharp, F. R. and Yang, X. (2013). An integrative framework for Bayesian variable selection with informative priors for identifying genes and pathways. *PloS One* **8** e67672.

Petersen, K. B. and Pedersen, M. S. (2012). The matrix cookbook. Technical University of Denmark.

Polson, N. G., Scott, J. G. and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *J. Amer. Statist. Assoc.* **108** 1339–1349. MR3174712

Pritchard, J. and Przeworski, M. (2001). Linkage disequilibrium in humans: Models and data. *American Journal of Human Genetics* **69** 1–14.

Stephens, M. and Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* **10** 681–690.

Technology Department Carnegie Library of Pittsburgh (2002). In The Handy Science Answer Book. Visible Ink Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **58** 267–288. MR1379242

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 91–108. MR2136641

Wang, W. Y., Barratt, B. J., Clayton, D. G. and Todd, J. A. (2005). Genome-wide association studies: Theoretical and practical concerns. *Nat. Rev. Genet.* **6** 109–118.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics*, 7 (*Tenerife*, 2002) 733–742. Oxford Univ. Press, New York. MR2003537

Whittemore, A. S. (2007). A Bayesian false discovery rate for multiple testing. *J. Appl. Stat.* **34** 1–9. MR2345755

Wigginton, J. E., Cutler, D. J. and Abecasis, G. R. (2005). A note on exact tests of Hardy-weinberg equilibrium. *The American Journal of Human Genetics* **76** 887–893.

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J. and Lin, X. (2010). Powerful snp-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics* **86** 929–942.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* **89** 82–93.

Zhou, X., Carbonetto, P. and Stephens, M. (2013). Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genetics* **9** e1003264.

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

I. JOHNSTON
L. CARVALHO
DEPARTMENT OF MATHEMATICS AND STATISTICS
BOSTON UNIVERSITY
111 CUMMINGTON MALL
BOSTON, MASSACHUSETTS 02215
USA
E-MAIL: ianj@math.bu.edu
lecarval@math.bu.edu

T. HANCOCK
BIOSCIENCES RESEARCH DIVISION
DEPARTMENT OF PRIMARY INDUSTRIES
VICTORIA, 5 RING ROAD, BUNDOORA 3086
AUSTRALIA
E-MAIL: timothy.hancock@depi.vic.gov.au

H. MAMITSUKA
INSTITUTE FOR CHEMICAL RESEARCH
KYOTO UNIVERSITY
GOKASHO UJI-CITY, KYOTO 611–0011
JAPAN
E-MAIL: mami@kuicr.kyoto-u.ac.jp