Graph-structured variable selection with Gaussian Markov random field horseshoe prior

Marie Denis, Mahlet G Tadesse

Graph-structured variable selection with Gaussian Markov random field horseshoe prior

Marie Denis[1] and Mahlet G. Tadesse[2]

[1] *CIRAD, UMR AGAP Institut, Montpellier, France*
[2] *Department of Mathematics and Statistics, Georgetown University, Washington,DC,USA*

Abstract

A graph structure is commonly used to characterize the dependence between variables, which may be induced by time, space, biological networks or other factors. Incorporating this dependence structure into the variable selection procedure can improve the identification of relevant variables, especially those with subtle effects. For example, in genetic and genomic studies, the integration of such information can help identify genomic regions or sets of markers associated with complex traits. The Bayesian approach provides a natural framework to integrate the graph information through the prior distributions. In this work we propose combining two priors that have been well studied separately, the Gaussian Markov random field (GMRF) prior and the horseshoe prior, to perform selection on graph-structured variables. Local shrinkage parameters that capture the dependence between connected covariates are specified for the regression coefficients with the option of incorporating the sign of their empirical correlations. This encourages a similar amount of shrinkage for the regression coefficients while allowing them to have opposite signs. For non-connected variables, a standard horseshoe prior is specified. After evaluating the performance of the method using different simulated scenarios, we analyze the quantitative trait loci mapping study that motivated the proposed method. We also present two other real data applications, one in near-infrared spectroscopy with sequential dependence structure across all wavelengths and the other in gene expression study

with a general dependence structure among transcripts.

Keywords: Gaussian Markov random field, Horseshoe prior, Structured variable selection

1 Introduction

The identification of genetic markers associated with complex traits is an important research problem in genetic and genomic studies. In the *Arabidopsis thaliana* shoot growth study that motivated the proposed method, there is interest in identifying genomic regions associated with particular phenotypes (Marchadier et al., 2019). Univariate analyses that evaluate each marker separately remain the predominant approach, despite the fact that they do not take into account the joint effect of multiple markers and miss markers with small or no marginal effects, thus leading to reduced power and increased false positive detections. Alternative methods to overcome these limitations have been proposed. In particular, penalized regression methods are commonly used to jointly model multiple markers. These methods have further been extended to take into account the dependence between markers. In quantitative trait loci (QTL) mapping and in genome-wide association studies (GWAS), there is strong correlation between contiguous markers, which are generally grouped into haplotype/linkage disequilibrium (LD) blocks. In transcriptomic and other -omic studies, the dependence structure between markers may correspond to experimentally elucidated biological pathways (e.g., metabolic, regulatory or signaling pathways) or may be inferred computationally (e.g., based on co-expression). This has led to penalty functions that encourage the selection of marker-sets, as in group Lasso (Yuan and Lin, 2006). However, the inference of haplotype/LD blocks is subject to uncertainty and it may thus be desirable to account for the dependence without forming groups. One such approach is the elastic net, which combines an L_1 penalty with an L_2 penalty, thereby allowing the selection of sets of highly correlated variables (Zou and Hastie, 2005). For variables with pre-specified ordering, as with adjacent genetic markers, the fused lasso achieves sparsity and local smoothness by penalizing the coefficients and their successive first-order differences with an L_1 penalty (Tibshirani et al., 2005). In situations where markers are grouped into sets, not all markers in a group are relevant and a bi-level selection can be performed (Stingo et al., 2011; Simon et al.,

2013). For network-structured variables, Li and Li (2008, 2010) proposed a graph-constrained estimation method that specifies a penalty defined on the Laplacian matrix of the graph to identify subgroups of connected variables while encouraging smoothness of the regression coefficients over the graph. Pan et al. (2010) extended this approach using a group penalty based on L_γ norm with $\gamma > 1$. In the same spirit, Kim et al. (2013) used a network-based penalty to encourage selection of neighboring variables while allowing different effect sizes between neighbors. Incorporating these dependence structures into statistical models encourages the identification of groups of covariates with subtle individual effects that act jointly on the response variable, thereby increasing the power to detect associations and improving the predictive performance (Li and Li, 2010; Zhou and Zheng, 2013).

In the Bayesian framework, variable selection can be achieved by specifying shrinkage priors on the regression coefficients. These shrinkage priors fall into two broad classes: spike-and-slab priors (George and McCulloch, 1993) and continuous shrinkage priors (Polson and Scott, 2010). Using spike-and-slab priors, the structure information can be incorporated through the priors on the variable selection indicators. Smith and Fahrmeir (2007) used the spatial correlation in brain imaging to specify a binary Markov random field (MRF) or Ising prior for the latent variable selection indicators. In genomic applications, the gene-gene network has been used to specify an Ising prior (Li and Zhang, 2010; Stingo et al., 2011). The dependence structure is viewed as an undirected graph with nodes representing the covariates and edges representing links between the covariates. It should be noted that the Ising prior regulates the smoothness of the binary variable selection indicators over the graph, but not the smoothness of the regression coefficients. A practical challenge in applying the Ising prior is its high sensitivity to the choice of hyperparameters, which results in phase transitions (Stanley, 1987). In variable selection this is characterized by a small change in hyperparameter values leading to a massive increase in the number of selected covariates (Li and Zhang, 2010; Stingo et al., 2011).

With continuous shrinkage priors, Kyung et al. (2010) proposed a Bayesian version of the fused lasso by placing a Laplace prior on the first order differences. Methods that incorporate the covariate structure information at the level of the shrinkage hyperparameters have also been proposed. Rockova and Lesaffre (2014) put forward a Bayesian lasso prior with a gamma hyperprior that incorporates information on pathway membership to encourage simultaneous shrinkage of covariates in the same pathway. Similarly, Chang et al.

(2018) specified a Bayesian Lasso with a log-normal hyperprior to capture the information on pathway membership and used an EM-based approach. The use of a single hyperparameter in the Laplace prior makes it restrictive and this has led to the development of global-local shrinkage priors, which can simultaneously induce shrinkage to zero of small coefficients while leaving large coefficients untouched (Polson and Scott, 2010). Griffin and Brown (2012) achieved this by specifying a multivariate correlated normal-gamma distribution on the regression coefficients to shrink the effects of serially dependent or grouped variables towards each other. Kalli and Griffin (2014) extended this approach for time-varying models by specifying normal-gamma autoregressive process priors for the regression coefficients. The methods above assume that all variables are either sequentially dependent or that variables in a group are fully connected to each other. For general dependence structures, which can be represented by an undirected graph, graph-based priors can be specified to promote graph-structured smoothness among coefficients. Liu et al. (2014) used a graph Laplacian prior that allows both negative and positive partial correlations between pairs of coefficients. Kowal, Matteson and Ruppert (2019) used a global-local shrinkage prior and incorporated the dependence through the local shrinkage hyperparameters.

In this paper, we propose shrinkage priors that incorporate general dependence structures between covariates by combining the efficiency and flexibility of the horseshoe (HS) prior with the appealing connection between GMRFs and undirected graphs. This allows the estimation of smooth coefficient profiles with possible abrupt changes and the selection of sets of dependent markers that may correspond to relevant genomic regions or gene networks. Moreover, in order to allow regression coefficients with different signs between connected variables, the sign of the empirical correlation between the corresponding covariates may be incorporated into the prior. The proposed model extends the work of Faulkner and Minin (2018) for function estimation, to variable selection in the general context of graph-structured covariates. Section 2 presents the Gaussian Markov random field horseshoe prior (HS-GMRF) and the Markov chain Monte Carlo (MCMC) algorithm for posterior sampling. In Section 3, we evaluate the performance of the proposed model using various simulated scenarios and compare the results to the standard HS prior and the spike-and-slab Ising prior. Section 4 presents the results for the shoot growth QTL mapping study in *Arabidopsis thaliana* and two other applications, one in near-infrared spectroscopy with sequential dependence structure across all wavelengths and the other in gene expres-

sion study with a general dependence structure among transcripts. Section 5 concludes the paper with some discussions.

2 Statistical model

2.1 Model formulation

Let $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ be a linear regression model with $\mathbf{y} = (y_1, \ldots, y_n)'$ an $n \times 1$ vector of observations, $\beta = (\beta_1, \ldots, \beta_p)'$ a p-dimensional vector of regression coefficients, $\mathbf{X}$ an $n \times p$ matrix of covariates, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$ an $n \times 1$ vector of residuals assumed to follow a Gaussian distribution $\mathcal{N}_n(0, \sigma^2\mathbf{I}_n)$.

We assume that the structure between the p covariates is encoded by an undirected graph $\mathcal{G} = (V, E)$ where $V = \{1, \ldots, p\}$ is a finite set of vertices or nodes, and E is the set of edges connecting a subset of the $\binom{p}{2}$ vertices. Hereafter, two covariates j and j' are considered neighbors if and only if $(j, j') \in E$. Let $\mathcal{N}(j)$ denote the set of neighbors of j. The graph $\mathcal{G}$ may be rewritten as a disjoint union of I subgraphs such that $\mathcal{G} = \bigcup_{i=1}^{I} \mathcal{G}_i = \bigcup_{i=1}^{I} (V_i, E_i)$ where V_i is a finite set of vertices associated with the subgraph $\mathcal{G}_i$, and E_i is the associated set of edges. Note that it is not necessary to have disjoint subgraphs and I may be equal to 1 (see spectrometric data example in Section 4.2). Let $\mathcal{S}$ denote a set composed of one randomly selected vertex from each disjoint subgraph; the choice of the vertex is arbitrary and does not affect the results (see Supplementary Material Section 5). For covariate j with $\mathcal{N}(j) = \emptyset$, the underlying subgraph called trivial graph reduces to one vertex, which is included in $\mathcal{S}$.

We propose incorporating the graph structure into the regression model by specifying a HS prior on the differences of the regression coefficients associated with edges as well as on the regression coefficients associated with vertices in the set $\mathcal{S}$. The proposed hierarchical model, which we refer to as horseshoe Gaussian Markov field (HS-GMRF), is defined as:

$$\begin{aligned}
\mathbf{y}|\beta, \sigma^2 &\sim \mathcal{N}_n(\mathbf{X}\beta, \sigma^2\mathbf{I}_n) \\
\beta_j - s_{jj'}\beta_{j'}|\tau^2_{jj'},\, \lambda^2 &\sim \mathcal{N}(0, \lambda^2\tau^2_{jj'}) \qquad \text{for } (j, j') \in \bigcup_{i=1}^{I} E_i\ , \\
\beta_j|\tau^2_j,\, \lambda^2 &\sim \mathcal{N}(0, \lambda^2\tau^2_j) \qquad \text{for } j \in \mathcal{S} \qquad (1)
\end{aligned}$$

$$\tau_{jj'} \sim \mathcal{C}^+(0,1) \text{ for } (j,j') \in \bigcup_{i=1}^{I} E_i \ , \qquad \tau_j \sim \mathcal{C}^+(0,1) \text{ for } j \in \mathcal{S}$$
$$\lambda|\sigma \sim \mathcal{C}^+(0,\sigma) \ , \qquad \sigma^2 \sim \mathcal{IG}(a_0, b_0)$$

where $\mathcal{C}^+$ denotes the half-Cauchy distribution on the positive reals, $\mathcal{IG}$ is the inverse-gamma density with fixed hyperparameters $a_0, b_0 > 0$. The model induces global smoothness over all regression coefficients via the hyperparameter λ. For connected variables, local adaptivity and smoothness are achieved through the hyperparameters $\tau_{jj':(j,j')\in\bigcup_{i=1}^{I} E_i}$. If it is desired to encourage regression coefficients of negatively correlated variables to take opposite signs, in the same spirit as Monni (2014), the sign of the sample correlation between covariates j and j' may be introduced by setting $s_{jj'} = \text{sign}\{\text{cor}(X_j, X_{j'})\}$; otherwise, $s_{jj'}$ is set to 1. For variables in $\mathcal{S}$, that is, non-connected variables and the randomly selected vertex from each disjoint subgraph, the local adaptivity is controlled by τ_j, as in the standard HS formulation. This ensures the resulting precision matrix is full rank and that the prior is proper (Faulkner and Minin, 2018).

Assuming "independent" increments as in Rue and Held (2005), the joint distribution of β conditionally on λ^2 and $\tau^{\mathbf{2}} = \left(\tau^2_{jj':(j,j')\in\bigcup_{i=1}^{I} E_i}, \tau^2_{j\in\mathcal{S}}\right)'$ is a GMRF distribution:

$$\beta|\tau^{\mathbf{2}}, \lambda^2 \sim \mathcal{N}_p(0, \lambda^2 \mathbf{Q}^{-1}), \tag{2}$$

where $\mathbf{Q}$ is a full rank precision matrix with diagonal elements

$$Q_{jj} = \begin{cases} \frac{1}{\tau_j^2} + \sum_{j'\in\mathcal{N}(j)} s_{jj'} \frac{1}{\tau^2_{jj'}} & \text{if } \ j \in \mathcal{S} \\ \sum_{j'\in\mathcal{N}(j)} s_{jj'} \frac{1}{\tau^2_{jj'}} & \text{otherwise} \end{cases}$$

and off-diagonal elements

$$Q_{jj'} = \begin{cases} -s_{jj'} \frac{1}{\tau^2_{jj'}} & \text{if } \ (j,j') \in \bigcup_{i=1}^{I} E_i \\ 0 & \text{otherwise} \end{cases}.$$

2.2 MCMC implementation

The HS-GMRF model can be fit via Gibbs sampling. The evaluation of the full conditional distributions can be facilitated by relying on reparametrizations of the regression coefficients and of the half-Cauchy distribution:

1. We introduce a q-dimensional vector $\phi = (\phi_1, \ldots, \phi_q)' = \mathbf{C}\beta$ (Martínez-Beneito and Botella-Rocamora, 2019). Here, $q = |E| + |\mathcal{S}|$, with $|E|$ the number of edges and $|\mathcal{S}|$ the number of disjoint subgraphs. If the r-th element of ϕ is $\phi_r = \beta_j - \beta_{j'}$, then the r-th row of the $q \times p$ matrix $\mathbf{C}$ has all 0's except for $c_{rj} = 1$ and $c_{rj'} = -1$. Otherwise, if $\phi_r = \beta_j$ then the j-th row of C has all 0's except for $c_{rj} = 1$. Thus,

$$\phi \sim \mathcal{N}_q(0, \mathbf{\Sigma}_\phi), \qquad \text{with} \quad \mathbf{\Sigma}_\phi = \text{diag}(\lambda^2 \tau^{\mathbf{2}}). \tag{3}$$

2. We use the parametrization of the half-Cauchy as a mixture of inverse-gamma distributions proposed by Makalic and Schmidt (2016):

$$x \sim \mathcal{C}^+(0, A) \qquad \Rightarrow \qquad x^2|a \sim \mathcal{IG}(1/2, 1/a); \;\; a \sim \mathcal{IG}(1/2, 1/A^2).$$

   Thus, the hyperpriors of the global and local shrinkage hyperparameters in (2) can equivalently be written as:

$$\begin{aligned}
\tau_{jj'} \sim \mathcal{C}^+(0,1) \quad \text{for } (j,j') \in \bigcup_{i=1}^{I} E_i \quad &\Rightarrow \quad \left\{ \begin{array}{rcl} \tau^2_{jj'}|\nu_{jj'} & \sim & \mathcal{IG}(1/2, 1/\nu_{jj'}) \\ \nu_{jj'} & \sim & \mathcal{IG}(1/2, 1) \end{array} \right. \\
\tau_j \sim \mathcal{C}^+(0,1) \quad \text{for } j \in \mathcal{S} \quad &\Rightarrow \quad \left\{ \begin{array}{rcl} \tau^2_{j}|\nu_{j} & \sim & \mathcal{IG}(1/2, 1/\nu_{j}) \\ \nu_{j} & \sim & \mathcal{IG}(1/2, 1) \end{array} \right. \\
\lambda|\sigma \sim \mathcal{C}^+(0,\sigma) \quad &\Rightarrow \quad \left\{ \begin{array}{rcl} \lambda^2|\omega & \sim & \mathcal{IG}(1/2, 1/\omega) \\ \omega|\sigma^2 & \sim & \mathcal{IG}(1/2, 1/\sigma^2) \end{array} \right.
\end{aligned} \tag{4}$$

The full conditional distributions are given in the Appendix. Code implementing the HS-GMRF model along with a detailed example are provided in the Supplementary Material.

3 Simulation study

We evaluate the performance of the proposed method using various simulated scenarios. The first simulation mimics the group sequential dependence structure that is typical in genetic analysis with high correlation between adjacent markers along a chromosome and independence between chromosomes. The second set of simulations is designed to be similar to gene expression data, where groups of genes are regulated by one marker that acts as a transcription factor; we consider varying levels of correlation as well as varying effect sizes for the regression coefficients.

3.1 Simulation scenarios

Let $\mathbf{X}$ be the $n \times p$ matrix of predictors divided into G groups

$$\mathbf{X} = [\underbrace{\mathbf{X}_{11}, \ldots, \mathbf{X}_{1k}}_{\mathbf{X_1}}, \underbrace{\mathbf{X}_{21}, \ldots, \mathbf{X}_{2k}}_{\mathbf{X}_2}, \ldots, \underbrace{\mathbf{X}_{G1}, \ldots, \mathbf{X}_{Gk}}_{\mathbf{X}_G}],$$

where $\mathbf{X_g}$ is the $n \times k$ matrix of covariates in group g ($g = 1, \ldots, G$) with row vectors $\mathbf{X}_{i,g} = (X_{i,g1}, \ldots, X_{i,gk})'$ for individual i ($i = 1, \ldots, n$) assumed to follow a multivariate normal distribution $\mathcal{N}_k(0, \mathbf{\Sigma}_g)$. The subgraphs associated with each group g with covariance matrix $\mathbf{\Sigma}_g$ are defined to have $k-1$ edges. We consider a total of $p = 140$ predictors divided into $G = 14$ groups of size $k = 10$ with $n = 100$ samples. A subset of five groups ($g = 1, 3, 5, 8, 10$) are assumed to have non-zero effects. Let $\beta_g = (\beta_{g1}, \ldots, \beta_{gk})'$ for $g = 1, \ldots, G$ be the vector of regression coefficients associated with the g-th group of predictors. The response $\mathbf{y}$ is simulated such that $\mathbf{y} = \sum_{g \in \{1,3,5,8,10\}} \mathbf{X_g} \beta_\mathbf{g} + \varepsilon$ where ε follows a normal distribution with zero mean and variance equal to $\sum_{g \in \{1,3,5,8,10\}} \beta_\mathbf{g}{}' \beta_\mathbf{g}/5$, similarly to Peterson, Stingo and Vannucci (2016). We consider this same scenario with $p = 1000$ covariates divided into $G = 25$ groups of size $k = 40$ and $n = 100$ and present the results in the Supplementary Material (Sections S2 and S3.2).

For the first simulation with group sequential dependence structure, the predictors in each of the G groups are defined such that $X_{ig,j}|X_{ig,j-1} \sim N(\rho X_{ig,j-1}, 1-\rho^2)$, for $j = 2, \ldots, k$, $g \in \{1, \ldots, G\}$ with ρ set to 0.9. We refer to this covariance structure as $\Sigma_{g,\text{seq}}$. The non-zero regression coefficients are simulated from a multivariate normal distribution with covariance $\Sigma_{g,\text{seq}}$.

For the second set of simulations with non-sequential dependence, we consider two covariance structures for the predictors, which we refer to as $\Sigma_{g,\text{all}}$ and $\Sigma_{g,\text{half}}$, corresponding respectively to a denser and a less dense association. The former, $\Sigma_{g,\text{all}}$, in the same spirit as Li and Li (2008), assumes that the predictors in each of the G groups have the same variance-covariance matrix, $\mathbf{\Sigma}_g$, defined such that $X_{ig,j}|X_{ig,1} \sim N(\rho X_{ig,1}, 1-\rho^2)$, for $j = 2, \ldots, k$, $g \in \{1, \ldots, G\}$. The latter, $\Sigma_{g,\text{half}}$, assumes that only the first half of the groups, $g = 1, \ldots, \lfloor \frac{G}{2} \rfloor$, have this variance-covariance structure, while the predictors in the second half of the groups, $g = \lfloor \frac{G}{2} \rfloor + 1, \ldots, G$, are assumed independent. The correlation ρ is taken to be 0.5 or 0.9. The regression coefficients are set such that the first variable in the group has a large effect size and the remaining variables in the group have small effect sizes. We consider two regression coefficient settings, one with effect sizes of varying

signs in the same group and the other with effects of the same sign in a group, such that $\beta_g = (5, \pm\frac{5}{\sqrt{10}}, \pm\frac{5}{\sqrt{10}}, \underbrace{\frac{5}{\sqrt{10}}, \ldots, \frac{5}{\sqrt{10}}}_{k-3})'$. The predictors associated with regression coefficients with signs that are opposite from that of the first covariate are simulated to be negatively correlated to it, such that $X_{ig,j}|X_{ig,1} \sim \mathcal{N}(-\rho X_{ig,1}, 1-\rho^2)$.

The effects β_g are simulated according to the assumed graph structure of the covariates only in the simulation scenarios with sequential dependence structure. For the non-sequential dependence scenarios, β_g's are fixed to specific values and are not simulated according to the covariance of $\mathbf{X}$. In particular, in the situation where half of the covariates are independent, which we denoted as $\Sigma_{g,\text{half}}$, the β_g's still have the same values as in the scenarios where all covariates within a group are considered dependent, denoted $\Sigma_{g,\text{all}}$.

3.2 Comparison methods

We compare the results of the proposed HS-GMRF prior with several other approaches, including a version that does not incorporate the correlation signs, i.e., setting $s_{jj'} = 1$ (HS-GMRF-nosign), the standard horseshoe prior (HS) that does not account for dependence between variables, and the spike-and-slab Ising prior (SS-Ising) that incorporates the graph information into the latent binary variable selection indicators.

We explored other methods that may be considered competitors to the proposed approach, but we could not use them as they did not accommodate the dependence structures encountered in our simulations or real datasets. For example, the method of Chang et al. (2018) uses the gene membership in pathways and does not take into account the network structure between markers; that is, all genes in a particular pathway are considered to be connected to each other. None of the dependence structures we are considering in this paper can be analyzed with this approach; it cannot handle the sequential dependence with disjoint sets, as in our first simulation scenario and in the *Arabidopsis* QTL mapping study, or the non-sequential dependence with subsets of variables being connected, as in our second simulation scenario and the riboflavin gene expression study, or the fully sequential dependence as in the Tecator data. Another example is the trend filtering method of Wang et al. (2015), which is designed for smoothing parameter estimates over graphs and does not have variable selection as its primary

goal. The implementation in the R package `genlasso` uses a lasso penalty, but the k-fold cross-validation capability for choosing the penalty parameter is only available for sequentially dependent data across all parameters and does not allow covariates to be specified. This approach cannot handle non-sequential dependence or sequential dependence over subsets, but even in the fully sequential setting as in the Tecator data, it is not possible to input a covariate matrix. The only existing method that accommodates all types of dependence structures and performs variable selection, as we propose in this paper, is the method of Peterson et al. (2016), which uses a spike-and-slab Ising prior (SS-Ising).

For all analyses, the columns of $\mathbf{X}$ are scaled to have variance 1. For each method, the MCMC algorithm is run for $6,000$ iterations with the first $1,000$ used as burn-in. The results are averaged over 50 simulated replications. Convergence of the MCMC samplers were assessed using Geweke's convergence diagnostic. Over the 50 simulated replications under each scenario, the Geweke z-scores were in the range $[-2, 2]$ for more than 80% of the replications, indicating that the samples in the first and last part of the chains were drawn from the same stationary distributions. In addition, we evaluated the autocorrelations and the effective sample size estimates for the MCMC chains. All the results indicated that the sampler mixed well, was efficient at exploring the parameter space and there were no indications of a lack of convergence.

For HS, we use the parametrization proposed by Polson and Scott (2010) implemented in the R package `bayesreg` (Makalic and Schmidt, 2016):

$$\begin{aligned} \beta_j|\tau_j^2,\, \lambda,\, \sigma^2 \;&\sim\; \mathcal{N}(0, \tau_j^2\,\lambda\,\sigma^2)\;, & \tau_j \;&\sim\; \mathcal{C}^+(0,1), \qquad j = 1, \ldots, p \\ \lambda \;&\sim\; \mathcal{C}^+(0,1)\;, & \sigma^2 \;&\sim\; \mathcal{IG}(a_0, b_0) \end{aligned} \tag{5}$$

For SS-Ising, we use the implementation of Peterson et al. (2016) with known graph:

$$\begin{aligned} \beta_j|\gamma_j, \sigma^2 \;&\sim\; \gamma_j \cdot \mathcal{N}(0, h_\beta \sigma^2) + (1-\gamma_j)\cdot \delta_0, \qquad j = 1, \ldots, p, \\ p(\gamma|\mathbf{Z}) \;&\propto\; \exp(a\mathbf{1}'\gamma + b\gamma'\mathbf{Z}\gamma) \end{aligned} \tag{6}$$

where $\gamma = (\gamma_1, \ldots, \gamma_p)'$ is a p-dimensional vector of latent binary indicator variables that induce a mixture prior on the β_j's and δ_0 is a point mass distribution at 0. An Ising prior is specified for γ using a matrix representation $\mathbf{Z}$ of the graph with elements 1 for connected variables and 0 elsewhere. The

hyperparameter a controls the sparsity of γ and b controls its smoothness over the graph with larger values of b leading to a phase transition characterized by a substantial increase in the number of selected variables. As recommended in Peterson et al. (2016), we set $a = 0.5$ and $b = -2.75$. Since the columns of $\mathbf{X}$ were scaled to have variance 1, we set $h_\beta = 1$.

3.3 Performance criteria

We assess the various methods in terms of variable selection, estimation accuracy of the regression coefficients, and predictive performance.

In order to determine the selected variables, for the HS-based methods a variable is deemed relevant if the 95% highest posterior density (HPD) interval of its regression coefficient β_j does not contain zero. For SS-Ising, variables with marginal posterior probability of inclusion (PPI) greater than 0.5 are deemed relevant. We then use Matthews correlation coefficient (MCC) (Matthews, 1975), which combines the overall variable selection accuracy in terms of the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN), to evaluate the selection performance. MCC ranges between -1, indicating complete disagreement between the truth and the selection, and $+1$, corresponding to perfect selection, and is defined by:

$$MCC = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}. \tag{7}$$

The mean squared error (MSE) of the regression coefficients is computed to evaluate the overall estimation accuracy:

$$MSE_\beta = \frac{1}{p} \sum_{g=1}^{G} \sum_{j=1}^{k} (\beta_{gj} - \hat{\beta}_{gj})^2, \tag{8}$$

where $\hat{\beta}_{gj}$ is the posterior mean of the regression coefficient for the j-th variable in group g. For the HS-based methods $\hat{\beta}_{gj} = \frac{1}{T} \sum_{t=1}^{T} \hat{\beta}_{gj}^t$ based on the post burn-in MCMC samples. For the SS-Ising implementation of Peterson et al. (2016), the regression coefficients have been integrated out of the model and are estimated post-hoc as:

$$\hat{\beta}_{gj} = \frac{1}{T} \sum_{t=1}^{T} (\mathbf{X}'_{\gamma_t} \mathbf{X}_{\gamma_t} + h_\beta^{-1} \mathbf{I}_{p_{\gamma_t}})^{-1} \mathbf{X}'_{\gamma_t} \mathbf{y} \tag{9}$$

where γ_t is the variable selection indicator vector at iteration t, $\mathbf{X}_{\gamma_t}$ the subset of covariates associated with γ_t, and p_{γ_t} the number of selected covariates at iteration t.

We also examine the coverage probability (CP) and the width of the 95% HPD intervals. For CP, we calculate the proportion of HPD intervals that contain the true regression coefficients, and report the average and standard error across the 50 replications. Similarly, we record the width of the 95% HPD interval for each coefficient, and report the average and standard error across regression coefficients and simulated replications.

The predictive performance is evaluated using the mean squared prediction error (MSPE) on a test set $(\mathbf{y}^{test}, \mathbf{X}^{test})$ of dimension $n \times 1$ and $n \times p$, respectively:

$$MSPE = \frac{1}{n}\sum_{i=1}^{n}(y_i^{test} - \mathbf{X}_i^{test}\hat{\beta})^2, \tag{10}$$

with $\hat{\beta} = (\hat{\beta}_{11}, \ldots, \hat{\beta}_{1k}, \ldots, \hat{\beta}_{G1}, \ldots, \hat{\beta}_{Gk})'$ the p-vector of estimated regression coefficients corresponding to the posterior means for the HS-based methods and defined as in equation (9) for the SS-Ising approach. We set $n = 100$ and consider 50 replications of the test data.

3.4 Results

Section 3.4.1 describes the results for the group sequential dependence structure. Section 3.4.2 discusses the results for the non-sequential dependence setting assuming knowledge of the true structure, as may be the case when using established biological pathways. The results for the non-sequential simulated setting where the non-zero regression coefficients in a group have the same sign, as well as the results using an estimated graph when the true dependence structure is unknown are presented in the Supplementary Material (Sections S3 and S4, respectively).

3.4.1 Results for sequential dependence

HS-GMRF yields better selection, better estimation and better prediction. Table 1 reports the mean MCC, MSE and MSPE along with their standard errors over the 50 simulated replications for each of the three methods considered. HS-GMRF has substantially higher MCC values. It also has the lowest

MSEs with lowest variability. As expected, HS, which does not integrate the dependence structure, has the worst performance in terms of MCC, but has lower MSE than SS-Ising despite the improved variable selection performance of the latter. The high MSE values for SS-Ising may be due to the post-hoc estimation of the regression coefficients. In terms of predictive performance, HS-GMRF gives the best result, while the MSPEs of HS and SS-Ising are comparable.

Table 1: Average MCC, MSE and MSPE (with SE) over 50 simulated replications under sequential dependence and non-sequential dependence.

		MCC	MSE	MSPE
		Sequential dependence		
$\Sigma_{g,\text{seq}}$	HS-GMRF	**0.740** (± 0.049)	**0.055** (± 0.010)	**12.763** (± 2.216)
	HS	0.171 (± 0.050)	0.239 (± 0.068)	15.940 (± 3.053)
$\rho = 0.9$	SS-Ising	0.403 (± 0.046)	0.343 (± 0.087)	16.654 (± 3.625)
		Non-sequential dependence		
$\Sigma_{g,\text{half}}$	HS-GMRF	**0.708** (± 0.018)	**0.513** (± 0.067)	**94.871** (± 13.632)
	HS-GMRF-nosign	0.624 (± 0.034)	0.728 (± 0.155)	122.188 (± 21.609)
$\rho = 0.5$	HS	0.240 (± 0.041)	1.009 (± 0.200)	126.252 (± 19.657)
	SS-Ising	0.323 (± 0.054)	1.386 (± 0.204)	149.294 (± 27.384)
$\Sigma_{g,\text{half}}$	HS-GMRF	**0.668** (± 0.046)	**0.541** (± 0.089)	**84.954** (± 14.485)
	HS-GMRF-nosign	0.444 (± 0.117)	1.038 (± 0.259)	99.123 (± 17.694)
$\rho = 0.9$	HS	0.219 (± 0.038)	2.243 (± 0.551)	95.219 (± 19.279)
	SS-Ising	0.312 (± 0.048)	2.359 (± 0.437)	109.387 (± 23.713)
$\Sigma_{g,\text{all}}$	HS-GMRF	**0.989** (± 0.011)	**0.420** (± 0.018)	**66.482** (± 10.084)
	HS-GMRF-nosign	0.866 (± 0.028)	0.697 (± 0.106)	102.078 (± 21.309)
$\rho = 0.5$	HS	0.254 (± 0.028)	1.145 (± 0.276)	107.449 (± 19.642)
	SS-Ising	0.338 (± 0.054)	1.669 (± 0.260)	121.027 (± 18.618)
$\Sigma_{g,\text{all}}$	HS-GMRF	**0.967** (± 0.033)	**0.439** (± 0.062)	**55.755** (± 9.403)
	HS-GMRF-nosign	0.639 (± 0.121)	1.234 (± 0.247)	80.999 (± 16.435)
$\rho = 0.9$	HS	0.211 (± 0.051)	3.310 (± 1.162)	74.425 (± 12.548)
	SS-Ising	0.346 (± 0.051)	3.068 (± 0.65)	71.691 (± 11.253)

HS-GMRF gives better estimation accuracy providing both high coverage probabilities (CP) and narrow HPD intervals. Table 2 presents the CPs and widths of the 95% HPD intervals for the zero and non-zero regression coefficients. For the zero coefficients, the three methods provide similar CPs with all of the 95% HDP intervals containing zero. For the non-zero coefficients, HS yields the highest CP but has very wide intervals. SS-Ising yields narrower intervals than HS, which, in part, can be explained by the fact that the spike-and-slab prior sets exactly to zero the regression coefficients for

non-selected covariates. However, SS-Ising misses the true parameter values as evidenced by its very low CPs.

Table 2: Coverage probability (CP) and width of 95% HPD intervals averaged over the 50 simulated replications under sequential dependence.

	CP of 95% HPD			Width of 95% HPD		
	Overall	β_j's $= 0$	β_j's $\neq 0$	Overall	β_j's $= 0$	β_j's $\neq 0$
HS-GMRF	0.921 (± 0.024)	1.000 (± 0.000)	0.778 (± 0.068)	0.642 (± 0.057)	0.429 (± 0.048)	1.025 (± 0.091)
HS	0.976 (± 0.020)	1.000 (± 0.000)	0.933 (± 0.055)	1.804 (± 0.160)	1.336 (± 0.136)	2.645 (± 0.243)
SS-Ising	0.809 (± 0.034)	1.000 (± 0.000)	0.464 (± 0.094)	0.567 (± 0.115)	0.101 (± 0.064)	1.405 (± 0.281)

In order to gain insights into these results, we examine the posterior densities of three non-zero regression coefficients from one of the groups (Figure 1). For the first panel, HS-GMRF is peaked around the true value with a small bump around zero, while SS-Ising shows a large peak at zero and a small bump at the true value, and HS is peaked around zero with a long tail towards the true value. In the second panel, HS-GMRF is unimodal with the true value in its lower tail, SS-Ising is unimodal around zero and misses the true value, and HS remains concentrated around zero with a long tail towards the true value. In the third panel, HS-GMRF is the only approach that captures the true value with high probability. SS-Ising is concentrated around a non-zero mode that is further away than the true value, and HS exhibits a bimodal posterior density with one mode at zero and the other around the non-zero mode identified by SS-Ising. This is a well-known phenomenon with the horseshoe prior for small to moderate effect sizes (Bhattacharya and Johndrow, 2021). We can clearly see that in the selected cases, HS-GMRF has narrower posterior densities that are away from zero, thereby having narrower HPD intervals and good selection performance. However, in some cases, the true value may fall in the tail area and may not be contained in the 95% HPD interval, explaining its lower CP. On the other hand, the posterior densities with HS span a large range of values that cover the true value as well as zero, leading to high CPs but large HPD interval widths, and hence poor selection and high uncertainty in the estimates.

For further exploration, we examine the estimation of all the regression coefficients in one simulated replication (Figure 2). We note that all the

methods capture the zero coefficients. For the non-zero coefficients, the HS and SS-Ising approaches tend to select a few representatives among a group of sequentially correlated predictors. HS also gives wide HPD intervals that contain both the true value and zero, hence leading to good CPs but wide intervals and poor selection performance. On the other hand, HS-GMRF captures the true coefficient profiles with high accuracy, yielding narrower HPD intervals with good CP. Given the high sequential correlation, it smoothes consecutive coefficients and may sometimes not adjust sufficiently to capture abrupt changes.

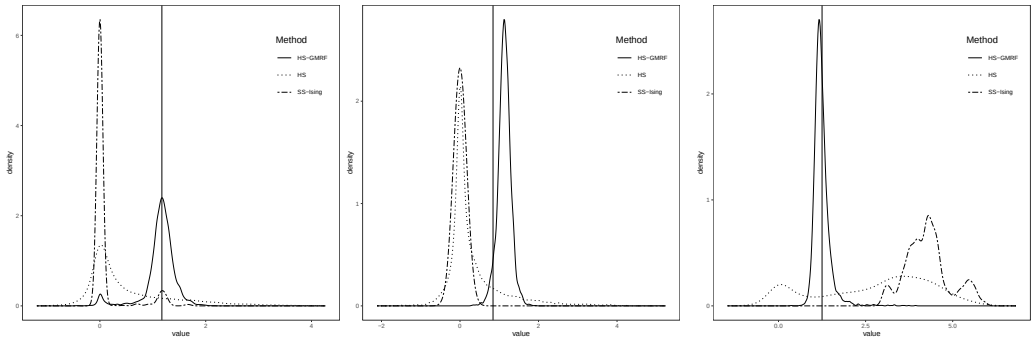


Figure 1: Posterior densities for three non-zero regression coefficients in the sequential dependence simulation. The black vertical bar corresponds to the true β_j value.

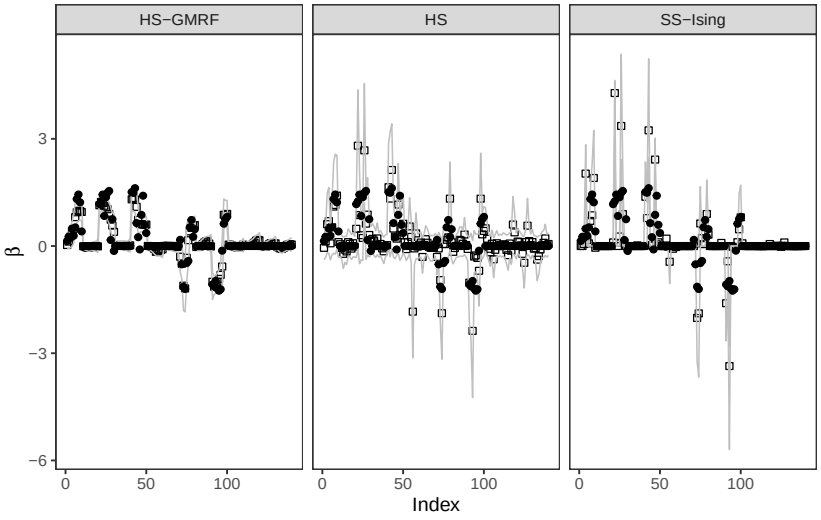


Figure 2: True coefficients (filled black circles) and estimated coefficients (unfilled squares) along with 80% HPD intervals (solid gray lines) for all covariates in one simulated replication under sequential dependence.

3.4.2 Results for known non-sequential dependence

As shown in Table 1, HS-GMRF provides the best results in terms of selection and estimation for all values of ρ and the two covariance structures considered. It enjoys MCC values closer to 1 and lowest MSEs, along with less variability in these quantities. The next best is HS-GMRF-nosign, although its variability in both MCC and MSE increases substantially with larger ρ. A high correlation encourages the regression coefficients associated with correlated covariates to have similar values with the same sign, leading to decreased performance. By incorporating the sign of the correlation among connected covariates into the model, HS-GMRF provides more flexibility by allowing connected covariates to have effects with different signs. As in the previous simulation, HS, which does not integrate the dependence structure, has the worst performance in terms of MCC. The MSEs for HS and SS-Ising are comparable despite the improved variable selection performance of the latter. This, as pointed out above, may be due to the post-hoc estimation of the regression coefficients for SS-Ising. In terms of predictive performance, HS-GMRF gives the best results.

With respect to the covariance structure considered, for the HS-GMRF-

based methods, the scenarios where all the groups have correlated predictors ($\Sigma_{g,\text{all}}$) result in higher MCC compared to situations where only half of the predictors are correlated ($\Sigma_{g,\text{half}}$). Thus, the integration of the structure information between covariates with the HS-GMRF prior is especially helpful as more covariates are connected. Note that although SS-Ising integrates this structure information, the improvement is not observed. As expected, for HS, which does not integrate the dependence structure, similar results are obtained for the two covariance structures. In terms of estimation, the HS-GMRF-based approaches have comparable accuracy across the two covariance structures, while HS and SS-Ising have considerably higher MSEs for the covariance structure $\Sigma_{g,\text{all}}$, especially under high correlation.

The results for the posterior density estimations and their associated CPs and HPD interval widths are, for the most part, similar to those observed in the sequential dependence case; the HS-GMRF-based approaches yield narrower posterior densities with good coverage and fairly accurate estimates of the true regression coefficient values compared to SS-Ising and HS (see Supplementary Table S1). In this setting, where there are regression coefficients with different signs in the same group, we note some differences between HS-GMRF and HS-GMRF-nosign. For the latter, the HPD intervals are relatively wider and less accurate. This indicates that integrating the correlation sign provides a more accurate estimation when the regression coefficients of connected covariates have varying signs.

Table 3: Average MCC and MSE for connected and non-connected covariates over 50 simulated replications under known non-sequential dependence.

	MCC		MSE	
	Connected	Non-connected	Connected	Non-connected
	$\Sigma_{g,\text{half}}$ $\rho = 0.5$			
HS-GMRF	**0.956** (± 0.033)	0.277 (± 0.039)	**0.558** (± 0.061)	**0.469** (± 0.111)
HS-GMRF-nosign	0.810 (± 0.053)	0.264 (± 0.057)	0.913 (± 0.202)	0.542 (± 0.151)
HS	0.237 (± 0.038)	0.244 (± 0.054)	1.464 (± 0.374)	0.553 (± 0.139)
SS-Ising	0.332 (± 0.062)	**0.295** (± 0.096)	2.028 (± 0.372)	0.744 (± 0.208)
	$\Sigma_{g,\text{half}}$ $\rho = 0.9$			
HS-GMRF	**0.883** (± 0.078)	0.278 (± 0.049)	**0.611** (± 0.138)	**0.470** (± 0.091)
HS-GMRF-nosign	0.526 (± 0.177)	0.265 (± 0.053)	1.582 (± 0.465)	0.495 (± 0.112)
HS	0.188 (± 0.043)	0.271 (± 0.046)	3.998(± 1.105)	0.488 (± 0.103)
SS-Ising	0.310 (± 0.047)	**0.304** (± 0.081)	4.055 (± 0.855)	0.662 (± 0.135)

We also examined the variable selection and estimation performance sep-

arately for connected and non-connected predictors (Table 3). For the non-connected variables, all the approaches do a poor job at identifying the relevant variables and have MCC values between 0.25 and 0.30. For the connected variables, the MCC values remain low for HS and SS-Ising. The HS-GMRF approaches, on the other hand, have substantially higher MCC, with HS-GMRF integrating the correlation sign achieving MCC values close to 1. As for MSE, the results for the non-connected covariates are comparable between the different approaches. However, for connected variables, the MSEs are significantly lower for the HS-GMRF based methods, with the approach integrating the correlation signs giving more accurate estimation.

Similar results are observed under the scenario where the non-zero regression coefficients of connected variables have identical signs, except for HS-GMRF-nosign which performs the same as HS-GMRF (see Supplementary Material Section S3.1).

3.4.3 Other remarks

In order to demonstrate that the choice of the representative vertex from each disjoint subgraph can be chosen arbitrarily, we repeated the analysis 20 times, each time picking a different randomly selected vertex from each disjoint subgraph. The results were unchanged as shown in the Supplementary Material (see Section S5).

In terms of computational time, the HS-GMRF approach took about twice as long as the HS to run. For example, for the sequential dependence structure with $p = 1000$ and 6,000 MCMC iterations, HS took 48 minutes while HS-GMRF took 103 minutes. This seems reasonable, given the gains in selection, estimation and prediction obtained with HS-GMRF.

4 Application

In this section we present three applications with different dependence structures. The first one aims to identify genetic regions associated with the shoot growth of *Arabidopsis thaliana* taking into account the dependence between adjacent markers within a chromosome and independence across chromosomes. The second one demonstrates that it is not necessary to have disjoint graphs by examining spectrometric data of food composition with sequential dependence across all wavelengths using the first-order derivatives as covari-

ates. The third one analyzes gene expression data in relation to riboflavin production in *Bacillus subtilis* with genes assumed to have an unknown general graph-structured dependence.

4.1 Shoot growth in Arabidopsis thaliana

Plant growth is a complex trait involving multiple loci. Marchadier et al. (2019) used recombinant inbred lines (RIL) under controlled conditions to study the genetic architecture of shoot growth in *Arabidopsis thaliana*. The phenotype and genotype data are publicly available. We consider the YoxCol RIL set composed of 358 plants at the end of the vegetative growth under well-watered condition. We focus on the rosette compactness phenotype, which is calculated as the ratio of the projected rosette area to the convex hull area. The covariate matrix consists of parental genotype probabilities at 486 loci along the five chromosomes. Marchadier et al. (2019) performed QTL detection using the Multiple QTL Mapping algorithm implemented in the `R/qtl` package, which does not assess the joint effect of multiple markers and does not account for the dependence structure between markers. Here, we apply the proposed HS-GMRF prior to integrate the dependence between adjacent markers on the same chromosome, thus encouraging smoothness across the regression coefficients and selection of contiguous predictors, while different chromosomes are assumed independent. The undirected graph for this dependence structure is represented by a block diagonal matrix with each block corresponding to a chromosome and consisting of a tridiagonal matrix. As consecutive markers are positively correlated, incorporating the sign of the empirical correlation will not be relevant, and the same results are obtained with HS-GMRF and HS-GMRF-nosign. For comparison, we analyzed the datal using HS, SS-Ising and the Bayesian lasso implemented in the R package `BGLR`. We decided against a comparison with fused lasso, as it assumes sequential dependence across all markers and does not allow a separation between chromosomes. One option would be to analyze each chromosome separately, but this would not evaluate the joint effect of markers mapping to different chromosomes.

We assess the predictive performance of the various methods using a five-fold cross-validation (CV) procedure and compute the MSPE. This consists of partitioning the data into five subsamples, using four of the subsamples for training and the left-out set for validation. The models are fit on the training set and the CV-MSPEs are computed based on the prediction for

the validation set :

$$\text{CV-MSEP} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_{test,k}}\sum_{i=1}^{n_{test,k}}\left(y_i^{test,k} - \mathbf{X}_i^{test,k}\hat{\beta}^{(k)}\right)^2,$$

where $(\mathbf{y}^{test,k}, \mathbf{X}^{test,k})$ are the test data associated with the k-th fold, $n_{test,k}$ the corresponding number of samples, and $\hat{\beta}^{(k)}$ are the posterior means of the regression coefficients based on the training set from the k-th fold. For each fold, the MCMC algorithm is run for $6,000$ iterations with the first $1,000$ used as burn-in. The MCMC chains from the $K = 5$ folds are then concatenated to select the relevant predictors. For SS-Ising, the MCMC was run for $10,000$ iterations with $5,000$ burn-in using the default hyperparameter settings.

Table 4 gives the CV-MSPE and the number of selected markers for each approach. HS-GMRF leads to the smallest CV-MSEP of 1.13 followed by the Bayesian lasso with a value of 1.15 then HS with a CV-MSPE of 1.16. SS-Ising provides the largest CV-MSPE of 1.28, which may be due to using the default hyperparameter settings that may not be optimal and estimating the regression coefficients post-hoc. In terms of detection, HS-GMRF selects 42 markers based on 95% HPD intervals and 67 with 90% HPD intervals. HS does not select any marker for any HPD interval considered. For SS-Ising, no marker is selected with a PPI threshold of 0.5, and there are 24 and 121 markers that pass a PPI threshold of 0.2 and 0.1, respectively. Of these, 12 and 39 are contained in the set selected by the HS-GMRF at 90% HPD intervals. No marker was selected with Bayesian lasso with 90% or 95% HPD intervals.

Table 4: Results for Arabidopsis thaliana data

Methods	CV-MSPE	Selected genes
HS-GMRF	1.13	42 (95% HPD) 67 (90% HPD)
HS	1.16	0 (95% HPD) 0 (90% HPD)
SS-Ising	1.28	0 (PPI>0.5) 24 (PPI > 0.2) 121 (PPI >0.1)
Bayesian Lasso	1.15	0 (95% HPD) 0 (90% HPD)

Figure 3 plots the estimated regression coefficients along with their 90% HPD intervals for HS-GMRF and HS, as well as the QTLs selected by Marchadier et al. (2019). We observe that HS-GMRF yields smoother coefficient estimates across adjacent loci within the same chromosome relative to HS. We also note that HS has very wide HPD intervals, especially when the estimates are not close to 0. This, as noted in the simulation study, is due to the long-tailed or bimodal posterior densities for small to moderated effects sizes. Furthermore, as observed with the simulations, HS tends to give large coefficient estimates for a few representatives among a group of correlated variables, while HS-GMRF yields smooth estimates with lower magnitudes around these peaks. This can be observed between positions 78 and 100 on chromosome 1, between positions 80 and 95 on chromosome 3, between positions 52 and 62 on chromosome 4, and between positions 4 and 20 on chromosome 5.

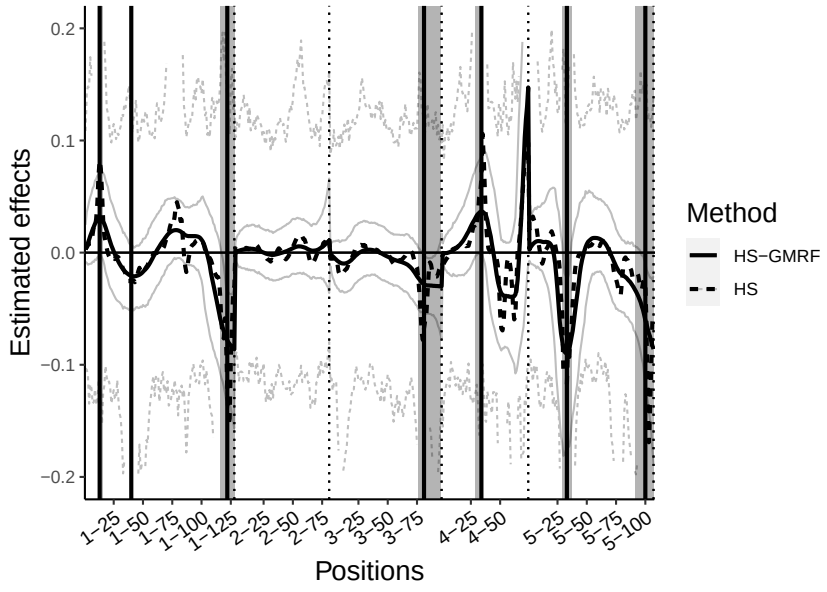


Figure 3: Estimated coefficients for the *Arabidopsis thaliana* data. The vertical dotted lines delimit the chromosomes. Estimates (black solid line) along with 90% HPD intervals (grey solid line), and selected genomic regions (shaded areas) using HS-GMRF. Estimates (black dashed line) along with 90% HPD intervals (grey dashed line) using HS. QTLs selected by Marchadier et al. (2019) (vertical solid lines).

We note that the proposed method picks contiguous markers and leads to the selection of genomic regions, which are represented by grey shaded areas. The solid vertical lines correspond to the QTLs selected by Marchadier et al. (2019), many of which have been experimentally validated. Except for one marker near 1:111650 (position 40 of chromosome 1 in Figure 3), these all fall within the genomic regions selected by HS-GMRF. Several of the identified genomic regions map to genes involved in shoot growth. For example, HS-GMRF selected consecutive markers in genomic region 1:26993011 to 1:29898172 (positions 116 to 128 in Figure 3). This region contains *CML38*, a member of the calmodulin-like proteins that plays important roles in the normal development of *Arabidopsis* and its flowering. Another important gene in this region is *STR1*, a member of sulfurtransferases, involved in *Arabidopsis* embryo and seed development with gene expression that steadily increases as the plant ages. Genomic region 3:19628061 to 3:23411903 (positions 76 to 92 in Figure 3) contains *FGPS3*, a folylpolyglutamate synthetase gene involved in root development that leads to reduction in primary root elongation when disrupted. In genomic region 5:7442381 to 5:8563029 (positions 29 to 37 in Figure 3), we find the *TGH* gene, which is essential for adequate plant development and whose mutation is associated with decreased elongation growth and vascularization, as well as reduced pollen formation.

4.2 Tecator dataset

We analyze the Tecator data, a benchmark for functional data analysis. Briefly, spectra sampled at 100 wavelengths uniformly spaced in the range $850-1050$ nm were recorded for $n = 215$ meat samples on a Tecator Infratec Food and Feed Analyzer (Borggaard and Thodberg, 1992). Recently, Picheny et al. (2019) applied penalized Sliced Inverse Regression (SIR) to relate fat content with the first-order derivatives of the spectra, obtained by finite differences, and identified intervals in the definition domain of the functional predictors that are associated with fat content. Their result highlights that the selection of relevant intervals rather than isolated wavelengths along the spectra improves the interpretability of the estimated coefficients. Here, we formulate the analysis as a regression model relating fat content (response) to the first-order derivatives of the spectra ($p = 99$ predictors) and apply the proposed HS-GMRF prior to integrate the dependence between consecutive predictors, thus encouraging smoothness across the regression coefficients and selection of contiguous predictors. The undirected graph for this dependence

structure would be represented by a tridiagonal matrix and would not be decomposable. As consecutive wavelengths are positively correlated, incorporating the sign of the empirical correlation will not be relevant, and we obtain the same results with HS-GMRF and HS-GMRF-nosign. For comparison, we also fit the model using the standard HS, SS-Ising and the Bayesian Lasso implemented in the R package `BGLR`. In this case, since all variables are sequentially correlated with no separation, it is possible to use the fused Lasso. We therefore applied the Bayesian version proposed by Kyung et al. (2010). For the four continuous shrinkage based approaches, the variable selection relied on the 90% and 80% HPD intervals not containing 0 . For SS-Ising, predictors with a marginal posterior probability of inclusion (PPI) greater than 0.5 or 0.8 were selected.

Table 5 gives the CV-MSPE and the number of selected predictors for each approach. HS-GMRF leads to the smallest CV-MSPE of 6.18 and selects 11 predictors using 90% HPD intervals or 19 with 80% HPD intervals. HS gives the next smallest CV-MSPE of 6.35 and deems only two wavelengths relevant based on 80% HPD intervals, both of which appear among those identified by HS-GMRF. SS-Ising has a relatively larger CV-MSPE of 7.86, which may be due to using the default hyperparameter settings and estimating the regression coefficients post-hoc. It selects four predictors with a 0.5 PPI threshold, all of which are contained in the set selected by the HS-GMRF, and selects none at a PPI threshold of 0.8. The Bayesian Lasso also yields a relatively large CV-MSPE of 7.77 and selects a single predictor, which was also identified as relevant by the other methods. The Bayesian fused Lasso has the largest CV-MSPE at 12.03 and selects five predictors based on 90% HPD intervals, one of which overlaps with HS-GMRF at the same level, or eight predictors based on 80% HPD intervals, four of which overlap with HS-GMRF at the same level.

Table 5: Results for Tecator data

Methods	CV-MSPE	Number of selected predictors
HS-GMRF	6.18	11 (90% HPD) 19 (80% HPD)
HS	6.35	1 (90% HPD) 2 (80% HPD)
SS-Ising	7.86	4 ($PPI > 0.5$) 0 ($PPI > 0.8$)
Bayesian Lasso	7.77	1 (90% HPD) 1 (80% HPD)
Bayesian fused Lasso	12.03	5 (90% HPD) 8 (80% HPD)

Figure 4 displays the estimated regression coefficients along with their 80% HPD intervals for HS-GMRF and HS. We note that HS-GMRF successfully smooths the coefficients over consecutive wavelengths. In contrast, the estimates from HS are jagged. We also note that the 80% HPD interval widths are wider with HS, especially when the estimates are not close to 0. This, as noted in the simulation study, is due to the long-tailed or bimodal posterior densities for small to moderate effect sizes. Furthermore, as observed with the simulations, HS tends to pick out a few representatives among a group of correlated variables. Indeed, HS identified only wavelengths 920 and 940, while HS-GMRF selects consecutive wavelengths around these two. This can be visualized in Figure 5 (a) and (b), which show the first-order derivates at each wavelength with vertical lines indicating those selected using 80% HPD interval with the HS and the HS-GMRF priors, respectively. The two wavelengths identified by HS are those with highest magnitudes located at positions 920 and 940. HS-GMRF, however, picks 19 wavelengths, many of which are adjacent to each other and overlap with the intervals identified by Picheny et al. (2019) (Figure 5(c)).

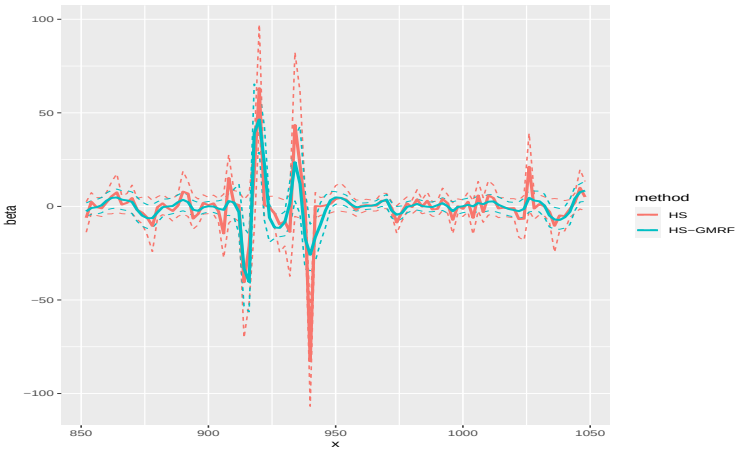
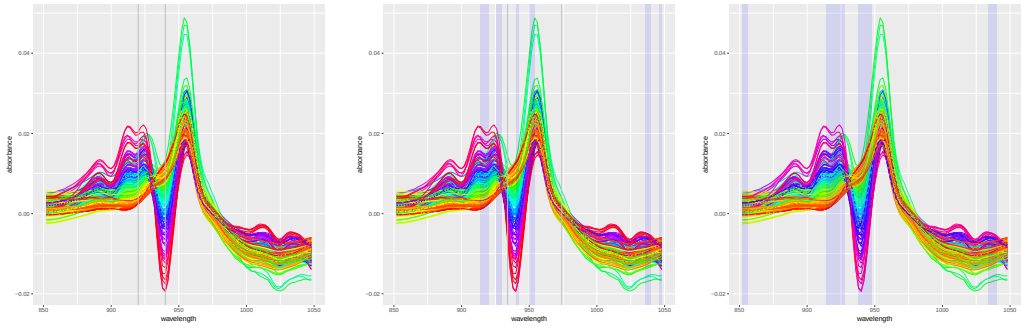


Figure 4: Estimated coefficients with 80% HPD intervals for the HS and HS-GMRF methods in the Tecator data.



(a) (b) (c)

Figure 5: First-order derivatives of raw predictors with vertical bars indicating selected wavelengths in the Tecator data based on (a) 80% HPD intervals for HS, (b) 80% HPD intervals for HS-GMRF, and (c) results from Picheny et al. (2019).

4.3 Riboflavin production in Bacillus subtilis

Riboflavin is an essential micronutrient required for biochemical reactions in all living cells and *Bacillus subtilis* is the most commonly used organism for commercial production of riboflavin. We analyze the riboflavin dataset

available in the R package `hdi`, which contains $p = 4088$ candidate predictors collected on 71 samples. The goal is to identify predictors (gene expressions) associated with riboflavin production in this organism. As there was not sufficient biological knowledge to determine the dependence structure between the markers, we obtained an undirected graph by first clustering the gene expressions using the R package `mclust` then applying within each cluster the graphical lasso approach with the penalty parameter selected by cross-validation using the R package `CVglasso`. This led to 93 subgraphs containing between 2 and 34 markers with 1964 edges, and 3480 subgraphs with singletons. We apply the four methods, HS-GMRF, HS, SS-Ising and Bayesian lasso to the data.

HS-GMRF yields the smallest CV-MSPE, confirming that the integration of the covariate structure along with the correlation sign leads to improved predictive performance (Table 6). The relatively large CV-MSPE for SS-Ising may be due to the default hyperparameter settings not being optimal and the post-hoc approach for estimating the regression coefficients.

Table 6: Results for riboflavin data

Methods	CV-MSPE	Selected genes
HS-GMRF	0.23	6 (70% CI)
HS	0.24	0 (70% CI)
SS-Ising	0.29	2 ($PPI > 0.7$)
Bayesian Lasso	0.27	0 (70% CI)

In terms of variable selection, HS-GMRF identifies six genes based on 70% credible intervals. SS-Ising selects two genes using a PPI cut-off of 0.7. The credible intervals for HS and Bayesian lasso cover 0, leading to no gene being selected. For comparison, in Bühlmann, Kalisch and Meier (2014), several variable selection methods were applied to the riboflavin data: At a 5% family-wise error rate, one approach found no significant gene, while another found one significant gene; based on an approach that controls the false discovery rate, and hence is less stringent, three significant genes were identified. The one common gene, *YXLD*, identified across the methods is also found by HS-GMRF using a 70% credible interval.

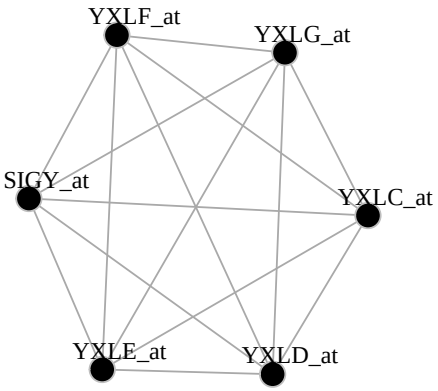


Figure 6: Estimated network for the 6 genes selected by HS-GMRF with 70% credible intervals in the riboflavin data.

Figure 6 displays the estimated network associated with the genes picked by HS-GMRF using 70% credible intervals, which form a clique. The six selected genes, *SIGY, YXLC, YXLD, YXLE, YXLF, YXLG*, comprise the SigY operon, a member of the extracytoplasmic function sigmas that function as regulators of stress. The joint selection of these genes makes biological sense as genes transcribed in a single operon are functionally related and are part of a metabolic pathway.

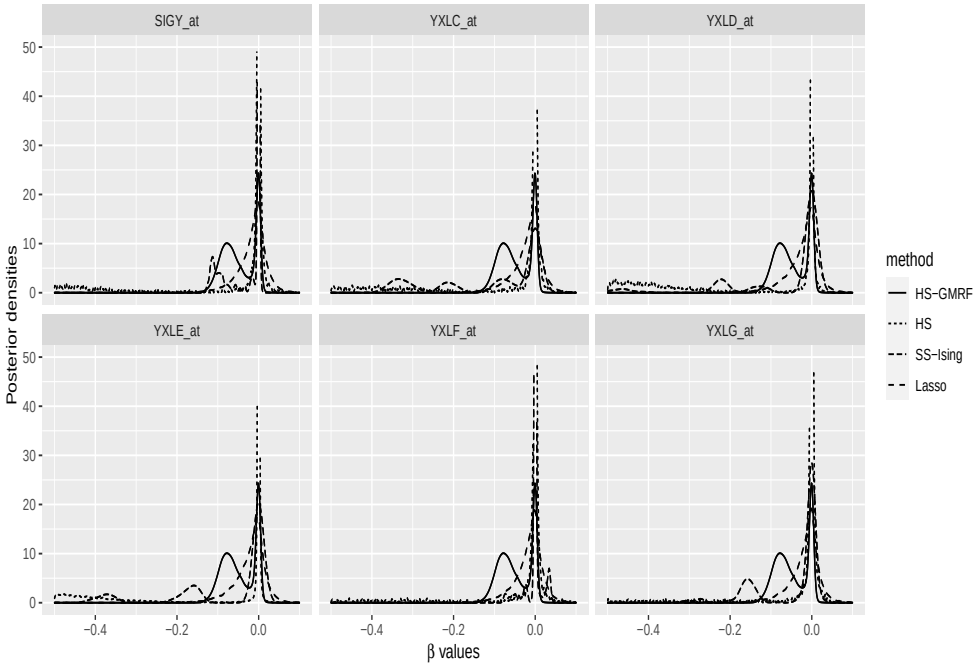


Figure 7: Posterior densities for genes selected with 80% HPD intervals in the Riboflavin data.

Figure 7 shows the posterior density estimates using the HS-based approaches for the regression coefficients associated with the six genes selected with HS-GMRF. In general, the posterior densities estimated with HS-GMRF are shifted away from zero relative to the densities estimated by the other methods. As observed in the simulation study, the posterior densities estimated with HS for moderate non-zero effects are either concentrated around 0 with long tails or exhibit bimodality with the prominent mode around 0. Indeed, some of the genes in Figure 7 exhibit such behaviour, suggesting that they have effect sizes that are not large enough to be detected by HS. By contrast, HS-GMRF leads to unimodal posterior densities centered away from 0 or bimodal posterior densities with much less spike around 0.

5 Conclusion

Incorporating the dependence structure between covariates into variable selection is a research area that has received attention over the past decade.

The contribution of this paper lies in integrating two well studied priors, GMRF and HS, to incorporate various types of dependence structures and achieve improved variable selection. The proposed HS-GMRF prior combines the sparsity inducing property of the horseshoe prior with the smoothing properties of Gaussian Markov random fields, with the option of integrating the correlation sign between covariates to allow regression coefficients with different signs among connected variables. We demonstrate via simulations the improved performance of the HS-GMRF prior in terms of selection, estimation and prediction. In particular, this prior encourages the selection of small to moderate effect sizes that are missed when the dependence structure is ignored, as in a standard HS. In addition, this approach, similarly to other methods that incorporate the dependence or group structure of covariates, encourages the selection of all correlated or connected variables, while standard variable selection methods tend to identify one of the highly correlated variables as they can be considered to be exchangeable. These findings are also supported by the results for the various applications considered, where the HS-GMRF method identifies biologically relevant genomic regions that encompassed experimentally validated markers, and leads to better cross-validated prediction. Although HS-GMRF takes computationally more time than HS, it remains reasonable, especially considering its gains in selection, estimation and prediction.

The HS-GMRF prior provides the flexibility of handling different dependence structures and is widely applicable in settings where the structure can be represented by undirected graphs. In this paper, we considered three applications, one exhibiting subgroups of sequential dependence in QTL mapping, another with sequential dependence across the entire data with near-infrared spectra, and one with a general dependence structure in a transcriptomic study. Other application areas include disease mapping, where variables measured over time at adjacent locations are structured in space and time, and functional MRI (fMRI) studies, characterized by spatial dependence between voxels within anatomical regions of the brain as well as temporal correlation between the serial scans acquired while a subject performs experimental tasks. The method is scalable to large numbers of covariates and further computational gains are obtained in the presence of disjoint subgraphs, as the posterior sampling involves submatrices of smaller dimensions.

When the true dependence structure between covariates is unknown and needs to be estimated, a two-stage approach that first estimates the under-

lying graph then uses the estimated graph to specify the HS-GMRF prior can be used, as we did in the gene expression study. This works reasonably well, but introduces bias by failing to incorporate the uncertainty associated with the graph estimation. The proposed hierarchical model can be extended by introducing an additional layer that specifies a graphical model for the covariates. This would provide a unified method that estimates the underlying undirected graph while simultaneously performing variable selection and estimation of the regression coefficients as in Peterson et al. (2016). Another possible extension of the proposed method would be to use weighted graphs that capture the strength of the connection between covariates, thereby inducing differential smoothing across regression coefficients. Finally, it may be desirable to encourage the selection of connected covariates without encouraging them to have similar coefficient estimates. This can be achieved by reformulating the proposed method and specifying a MRF prior on the shrinkage hyperparameters, rather than placing a conditional GMRF prior on the regression coefficients.

Acknowledgements

MD was supported by the European Union's Horizon 2020 Research and Innovation programme under the Marie Sklodowska-Curie grant agreement No 840383. This research was completed while MD was hosted by Georgetown University for the grant's outgoing phase.

Appendix

The hierarchical model used for the MCMC implementation is given by:

$$\begin{aligned}
\mathbf{y}|\beta, \sigma^2 &\sim \mathcal{N}_n(\mathbf{X}\beta, \sigma^2\mathbf{I}_n) \\
\beta_j - s_{jj'}\beta_{j'}|\tau^2_{jj'},\, \lambda^2 &\sim \mathcal{N}(0, \lambda^2\tau^2_{jj'}) \text{ for } (j,j') \in \bigcup_{i=1}^{I} E_i \\
\beta_j|\tau^2_j,\, \lambda^2 &\sim \mathcal{N}(0, \lambda^2\tau^2_j) \text{ for } j \in \mathcal{S} \\
\tau^2_{jj'}|\nu_{jj'} &\sim \mathcal{IG}(1/2, 1/\nu_{jj'}) \text{ for } (j,j') \in \bigcup_{i=1}^{I} E_i
\end{aligned}$$

$$
\begin{aligned}
\nu_{jj'} &\sim \mathcal{IG}(1/2, 1) \text{ for } (j, j') \in \bigcup_{i=1}^{I} E_i \\
\tau_j^2|\nu_j &\sim \mathcal{IG}(1/2, 1/\nu_j) \text{ for } j \in \mathcal{S} \\
\nu_j &\sim \mathcal{IG}(1/2, 1) \text{ for } j \in \mathcal{S} \\
\lambda^2|\omega &\sim \mathcal{IG}(1/2, 1/\omega) \\
\omega|\sigma^2 &\sim \mathcal{IG}(1/2, 1/\sigma^2) \\
\sigma^2 &\sim \mathcal{IG}(a_0, b_0)
\end{aligned}
$$

The corresponding full conditional distributions for the model parameters are as follow:

$$
\begin{aligned}
\tau_{jj'}^2|. &\sim \mathcal{IG}(1, \frac{\phi_j^2}{2\lambda^2} + \frac{1}{\nu_{jj'}}) \text{ for } (j, j') \in \bigcup_{i=1}^{I} E_i \\
\nu_{jj'}|. &\sim \mathcal{IG}(1, \frac{1}{\tau_{jj'}^2} + 1) \text{ for } (j, j') \in \bigcup_{i=1}^{I} E_i \\
\beta|. &\sim \mathcal{N}_p(\left(\frac{\mathbf{Q}}{\lambda^2} + \frac{\mathbf{X}'\mathbf{X}}{\sigma^2}\right)^{-1} \frac{\mathbf{X}'\mathbf{y}}{\sigma^2}, \left(\frac{\mathbf{Q}}{\lambda^2} + \frac{\mathbf{X}'\mathbf{X}}{\sigma^2}\right)^{-1}) \\
\tau_j^2|. &\sim \mathcal{IG}(1, \frac{\beta_j^2}{2\lambda^2} + \frac{1}{\nu_j}) \text{ for } j \in \mathcal{S} \\
\nu_j|. &\sim \mathcal{IG}(1, \frac{1}{\tau_j^2} + 1) \text{ for } j \in \mathcal{S} \\
\lambda^2|. &\sim \mathcal{IG}(\frac{p+1}{2}, \frac{\beta'\mathbf{Q}\beta}{2} + \frac{1}{\omega}) \\
\omega|. &\sim \mathcal{IG}(1, \frac{1}{\sigma^2} + \frac{1}{\lambda}) \\
\sigma^2|. &\sim \mathcal{IG}(\frac{n+1}{2} + a_0, \frac{(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)}{2} + \frac{1}{\omega} + b_0)
\end{aligned}
$$

where $\mathbf{Q}$ is the $p \times p$ precision matrix defined in Section 2.1 and $\phi = \mathbf{C}\beta$ is the q-dimensional vector defined in Section 2.2.

References

Bhattacharya, A., Johndrow, J.E., 2021. MCMC for global-local shrinkage priors in high-dimensional settings, in: Tadesse, M.G., Vannucci, M. (Eds.), Handbook of Bayesian variable selection. Chapman and Hall/CRC, pp. 161–178.

Borggaard, C., Thodberg, H.H., 1992. Optimal minimal neural interpretation of spectra. Analytical Chemistry 64, 545–551.

Bühlmann, P., Kalisch, M., Meier, L., 2014. High-dimensional statistics with a view toward applications in biology. Annual Review of Statistics and Its Application 1, 255–278.

Chang, C., Kundu, S., Long, Q., 2018. Scalable bayesian variable selection for structured high-dimensional data. Biometrics 74, 1372–1382.

Faulkner, J.R., Minin, V.N., 2018. Locally adaptive smoothing with Markov random fields and shrinkage priors. Bayesian Analysis 13, 225–252.

George, E.I., McCulloch, R.E., 1993. Variable selection via Gibbs sampling. Journal of the American Statistical Association 88, 881–889.

Griffin, J.E., Brown, P.J., 2012. Structuring shrinkage: some correlated priors for regression. Biometrika 99, 481–487.

Kalli, M., Griffin, J.E., 2014. Time-varying sparsity in dynamic regression models. Journal of Econometrics 178, 779–793.

Kim, S., Pan, W., Shen, X., 2013. Network-based penalized regression with application to genomic data. Biometrics 69, 582–593.

Kowal, D.R., Matteson, D.S., Ruppert, D., 2019. Dynamic shrinkage processes. Journal of the Royal Statistical Society: Series B 81, 781–804.

Kyung, M., Gill, J., Ghosh, M., Casella, G., et al., 2010. Penalized regression, standard errors, and Bayesian lassos. Bayesian Analysis 5, 369–411.

Li, C., Li, H., 2008. Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics 24, 1175–1182.

Li, C., Li, H., 2010. Variable selection and regression analysis for graph-structured covariates with an application to genomics. Annals of Applied Statistics 4, 1498–1516.

Li, F., Zhang, N.R., 2010. Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. Journal of the American Statistical Association 105, 1202–1214.

Liu, F., Chakraborty, S., Li, F., Liu, Y., Lozano, A.C., et al., 2014. Bayesian regularization via graph Laplacian. Bayesian Analysis 9, 449–474.

Makalic, E., Schmidt, D.F., 2016. High-dimensional Bayesian regularised regression with the BayesReg package. arXiv preprint arXiv:1611.06649 .

Marchadier, E., Hanemian, M., Tisne, S., Bach, L., Bazakos, C., Gilbault, E., Haddadi, P., Virlouvet, L., Loudet, O., 2019. The complex genetic architecture of shoot growth natural variation in arabidopsis thaliana. PLoS genetics 15, e1007954.

Martínez-Beneito, M.A., Botella-Rocamora, P., 2019. Disease Mapping: From Foundations to Multidimensional Modeling. CRC Press.

Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure 405, 442–451.

Monni, S., 2014. Bayesian variable selection for correlated covariates via colored cliques. AStA Advances in Statistical Analysis 98, 143–163.

Pan, W., Xie, B., Shen, X., 2010. Incorporating predictor network in penalized regression with application to microarray data. Biometrics 66, 474–484.

Peterson, C.B., Stingo, F.C., Vannucci, M., 2016. Joint Bayesian variable and graph selection for regression models with network-structured predictors. Statistics in Medicine 35, 1017–1031.

Picheny, V., Servien, R., Villa-Vialaneix, N., 2019. Interpretable sparse sir for functional data. Statistics and Computing 29, 255–267.

Polson, N., Scott, J., 2010. Shrink globally, act locally: Sparse Bayesian regularization and prediction, in: Bernardo, J.M. (Ed.), Bayesian Statistics 9. Oxford University Press, pp. 501–538.

Rockova, V., Lesaffre, E., 2014. Incorporating grouping information in Bayesian variable selection with applications in genomics. Bayesian Analysis 9, 221–258.

Rue, H., Held, L., 2005. Gaussian Markov random fields: theory and applications. CRC press.

Simon, N., Friedman, J., Hastie, T., Tibshirani, R., 2013. A sparse-group lasso. Journal of Computational and Graphical Statistics 22, 231–245.

Smith, M., Fahrmeir, L., 2007. Spatial Bayesian variable selection with application to functional magnetic resonance imaging. Journal of the American Statistical Association 102, 417–431.

Stanley, H., 1987. Introduction to Phase Transitions and Critical Phenomena. Oxford University Press, New York.

Stingo, F.C., Chen, Y.A., Tadesse, M.G., Vannucci, M., 2011. Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. Annals of Applied Statistics 5, 1978–2002.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K., 2005. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B 67, 91–108.

Wang, Y.X., Sharpnack, J., Smola, A., Tibshirani, R., 2015. Trend filtering on graphs, in: Artificial Intelligence and Statistics, PMLR. pp. 1042–1050.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B 68, 49–67.

Zhou, H., Zheng, T., 2013. Bayesian hierarchical graph-structured model for pathway analysis using gene expression data. Statistical Applications in Genetics and Molecular Biology 12, 393–412.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: series B 67, 301–320.