Bayesian varying coefficient model with selection: An application to functional mapping

Benjamin Heuclin
IMAG, Univ Montpellier, CNRS, Montpellier, France,
CIRAD, UMR AGAP, F-34398 Montpellier, France.
Frédéric Mortier,
Forêts et Sociétés, Cirad, F-34398 Montpellier, France,
Forêts et Sociétés, Univ Montpellier, Cirad, Montpellier, France.
Catherine Trottier,
Univ Paul-Valéry Montpellier 3, Montpellier, France.
IMAG, Univ Montpellier, CNRS, Montpellier, France.
and Marie Denis,
CIRAD, UMR AGAP, F-34398 Montpellier, France
AGAP, Univ Montpellier, CIRAD, INRAE, Institut Agro, Montpellier, France
E-mail: marie.denis@cirad.fr

Summary. How does the genetic architecture of quantitative traits evolve over time? Answering this question is crucial for many applied fields such as human genetics and plant or animal breeding. In the last decades, high-throughput genome techniques have been used to better understand links between genetic information and quantitative traits. Recently, high-throughput phenotyping methods are also being used to provide huge information at a phenotypic scale. In particular, these methods allow traits to be measured over time, and this, for a large number of individuals. Combining both information might provide evidence on how genetic architecture evolves over time. However, such data raise new statistical challenges related to, among others, high dimensionality, time dependencies, time varying effects. In this work, we propose a Bayesian varying coefficient model allowing, in a single step, the identification of genetic markers involved in the variability of phenotypic traits and the estimation of their dynamic effects. We evaluate the use of spike-and-slab priors for the variable selection with either P-spline interpolation or non-functional techniques to model the dynamic effects. Numerical results are shown on simulations and on a functional mapping study performed on an *Arabidopsis thaliana* (L. Heynh) data which motivated these developments.

Keywords: *Arabidopsis thaliana* (L. Heynh); Functional mapping; Group Spike-and-Slab; P-Splines; Time Varying Parameters; Variable selection; Varying coefficient models.

1. Introduction

Genetic architecture controls part of the variational properties of a phenotype. It has been treated as constant over time while most biological processes of interest are dynamic

by nature (Hansen, 2006). In agronomy, traits such as yield, quality or disease resistance vary over seasons, age of individuals or various environmental conditions. Such variations, so-called phenotypic plasticity, reflect the phenotypic responses of a given genotype to a changing environment and may constitute adaptative processes. Until recently, most analyses of dynamic traits have been based on mapping quantitative trait loci (QTL) at each time point separately. Such analysis does not allow to take into account dependencies between successive measures and can be less powerful to select QTL. It also does not allow the inclusion of external information such as environmental variables in case of identical conditions for all individuals at a given time. To overcome these limitations, new classes of statistical models have been developed to analyze such data. In particular, functional mapping (FM) has been proposed for QTL identification associated with dynamic traits (Ma et al., 2002; Wu et al., 2003; Li and Sillanpää, 2015).

FM is based on simultaneously modeling the dynamic relationship between quantitative traits and genotype information, and the residuals covariance matrix (Li and Wu, 2010). FM relied initially on the assumption that genetic effects are continuous functions (Li and Sillanpää, 2013) and thus appear as a special case of varying coefficient (VC) models (Hastie and Tibshirani, 1993). VC models encompass a broad class of statistical approaches such as generalized additive models (Hastie and Tibshirani, 1986), structured additive regression (STAR) models (Fahrmeir et al., 2004) or time varying parameters (Bitto and Frühwirth-Schnatter, 2019). Parametric methods based on biological knowledge have been initially developed using sigmoid or logistic functions to model the QTL dynamic effects (Ma et al., 2002; Wu et al., 2003). But such assumptions limit the curve flexibility and are restrictive to reflect the underlying processes. To overcome this restriction, non-parametric functional methods have been proposed such as those based on Legendre polynomial (Min et al., 2011; Li et al., 2015), or B-spline (Wang et al., 2008; Gong and Zou, 2012) interpolation techniques. While Legendre polynomial interpolation relies on global function bases that may lead to a decrease of goodness-of-fit when the order of polynomials increases, especially at both ends of the curve, B-splines use local function bases which greatly depend on the number of knots and their positions. Few knots do not provide enough flexibility to capture the variability in the data, while many knots may lead to overfitting. To overcome such limitation, penalization is usually applied to guarantee smoothness of the fitted curves and to limit overfitting (O'Sullivan, 1986, 1988). In particular, P-spline interpolation (Eilers and Marx, 1996) consisting in constraining the coefficients finite differences of adjacent B-splines, has been widely advocated in the FM context (Li and Sillanpää, 2013; Ni et al., 2019). In these previously mentioned approaches, FM was mainly based on the decomposition of a particular functional basis. However, in the VC model context, non-functional methods are an alternative approach consisting in directly modeling the varying coefficients (one parameter per time point without assuming a decomposition in a given functional basis). Such non-functional methods are widely used (Hastie and Tibshirani, 1993; Frühwirth-Schnatter and Wagner, 2010), but an unrestricted estimation does not insure smoothness and leads to overfitting problems (Bitto and Frühwirth-Schnatter, 2019; Franco-Villoria et al., 2019). To overcome these limitations, as mentionned for P-splines, penalization techniques are used. For example, the ℓ_2- or the ℓ_1-norm of the second differences has been proposed to model trends in time series (Kim et al., 2009). From a Bayesian per-

spective, such penalizations are equivalent to defining Gaussian prior distributions (Rue and Held, 2005; Rasmussen and Williams, 2006). For example, the ℓ_2-norm of the first or second differences correspond to first or second order random walk process priors, respectively (Lang and Brezger, 2004). In a genetic context, non-functional methods have been sparsely applied and compared to functional approaches (Li and Sillanpää, 2013; Vanhatalo et al., 2019). In this paper, we propose to evaluate, in a Bayesian framework, the impact of modeling choices focusing either on functional or non-functional approaches, each combined with first or second random walk process priors to model genetic effects over time.

With current technologies, such as high-throughput genotyping, the number of genetic markers may be huge leading to a large set of time varying parameters. To simultaneously analyze all markers and phenotypes observed along time, variable selection methods need to be performed in a FM context. In animal or plant genetics, selection is also crucial to improve breeding programs. Classical variable selection methods focus on a single coefficient. In FM, strategies are slightly different because all the sequences of coefficients associated to a genetic information have to be selected simultaneously. Group variables selection have been developed in such a context. Wang et al. (2008) extended the SCAD penalized approach to grouped longitudinal data and (Li and Sillanpää, 2013; Vanhatalo et al., 2019) adapted stepwise algorithms. In a Bayesian regression model, various variable selection approaches have been proposed. In particular, the Bayesian group LASSO with Legendre interpolation has been investigated by Li et al. (2015). However, in high-dimensional data, this type of approach which shrinks towards zero the effects of irrelevant variables without putting them exactly to zero, leads to biased estimation (Fan and Li, 2001; Kyung et al., 2010) and requires fitting the model in two steps. In time varying parameters, double Gamma prior is advocated (Bitto and Frühwirth-Schnatter, 2019) as proposed by Pérez et al. (2017) in a linear mixed context. In STAR models, Scheipl et al. (2012) proposed the use of a spike-and-slab prior based on mixture of inverse gamma distributions (Ishwaran and Rao, 2005). The spike-and-slab prior is a discrete mixture of two distributions (George and McCulloch, 1993, 1997). The spike distribution is concentrated around zero and models coefficients associated to irrelevant variables while the slab distribution is flat and allows to describe the coefficients of relevant variables (Ishwaran and Rao, 2005; Frühwirth-Schnatter and Wagner, 2010). In this paper, we propose a group spike-and-slab prior with Dirac mass at zero allowing to set to zero non relevant genetic information as proposed in Ghosh and Ghattas (2015); Yang and Narisetty (2020).

To sum up, we propose to use a Bayesian P-spline interpolation or a direct approach with first or second random walk process priors for the functional estimation of genetic and environmental dynamic effects. Both methods are combined with a group spike-and-slab prior for selection of time varying coefficients (functional effects). Our approach allows, in a single step, to estimate complex functions associated to varying coefficients and to select time-varying QTLs associated to phenotypic traits. Section 2 presents the full hierarchical Bayesian models. In section 3, model performances are tested on simulations. Numerical results show that combining penalised functional or non-functional method with a group spike-and-slab prior outperforms existing methods such as B-splines or Legendre interpolation combined with group-LASSO or even with

group spike-and-slab prior. Our approach compared to that of Vanhatalo et al.', also show better performances notably in terms of selection. Finally, section 4 is dedicated to a real case study, investigating the dynamic genetic architecture of shoot growth natural variations for *Arabidopsis thaliana* (L. Heynh) under two water availability conditions.

2. Statistical Models

Let y_{it_k} be the phenotype of individual $i = 1, \ldots, n$ at time t_k $(k = 1, \ldots, T)$. Let $t = (t_1, \ldots, t_T)'$ the time vector and $e^l = \left(e^l_{t_1}, \ldots, e^l_{t_k}, \ldots, e^l_{t_T}\right)'$ be L known environmental variables varying over time but common to all individuals at any given time t_k. Finally let us assume that genotype information, x_{ij}, $j = 1, \ldots, J$, is available for each individual at each of J loci. J is potentially much larger than n. Note that markers are constant over time but vary between individuals. We propose to model the phenotypes according to environmental conditions and genotypes using the following multivariate varying coefficient (VC) model:

$$y_{it_k} = \alpha + \mu(t_k) + \sum_{l=1}^{L} f_l(e^l_{t_k}) + \sum_{j=1}^{J} x_{ij}\beta_j(t_k) + \varepsilon_{it_k}. \tag{1}$$

α is the intercept, μ and f_l are real smooth functions of time and of the l^{th} environmental variable respectively. Note that for the model to be identifiable (Hastie and Tibshirani, 1986), μ and f_l have to be centered. The effect β_j of the j^{th} marker is assumed to be an unknown real smooth function of time. $\varepsilon_i = (\varepsilon_{it_1}, \ldots, \varepsilon_{it_T})'$ is a T-dimensional vector of residuals associated to individual i assumed to follow a multivariate Gaussian distribution, $\mathcal{N}(0, \sigma^2\Gamma)$, with σ^2 the residual variance and Γ the $T \times T$ correlation matrix defined by a first-order autoregressive (AR(1)) structure with unknown parameter ρ (Fahrmeir and Kneib, 2011).

Several functional methods have been proposed to approximate unknown functions (De Boor et al., 1978). Among them, B-spline interpolation is widely used. It consists of writing an unknown function h as a linear combination of B-spline basis functions:

$$h(x) = \sum_{r=1}^{df} B_r(x, \nu)\ c_r$$

where $(B_1(., \nu), \ldots, B_{df}(., \nu))$ is the collection of the ν^{th}-degree B-spline basis functions defined using K knots leading to $(K - 1)$ ordered subintervals on the x-domain and $c = (c_1, \ldots, c_{df})'$ is a vector of unknown B-spline coefficients. *df* is equal to $K + \nu$ and is called the degree of freedom of the B-spline basis. In the following ν and K will be assumed to be equal for all bases. Let us denote B^x the $T \times df$ dimensional matrix where $B^x_{i,r} = B_r(x_i, \nu)$. For $h(.)$ functions to be centered, B^x and c require to be reparametrized (see appendix A.1). In the following, $\widetilde{B}^x$ and $\widetilde{c}$ denote the re-parametrized versions of B^x and c. An accurate use of the B-spline approach strongly depends on the number of knots and the choice of their positions (Eilers and Marx, 1996). A misspecification may lead to over- or under- fits. To overcome these limitations and to introduce smoothness, penalized B-splines (P-splines) have been developed (Eilers and Marx, 1996). The idea

is to penalize the first or second order finite differences in adjacent spline regression coefficients.

Non-functional method presents an alternative to B-spline interpolation. It consists in the discretization of coefficient functions $(\beta_1(t), \ldots, \beta_J(t))$ leading to the estimation of $T \times J$ parameters as in a standard multivariate regression model (Li and Sillanpää, 2013). For smoothness reasons and due to the huge number of parameters, penalized least squares methods have been proposed consisting, as already used in P-spline context, to constrain the first or second differences of successive time regression parameters (Kim et al., 2009; Bruder et al., 2011; Bitto and Frühwirth-Schnatter, 2019; Franco-Villoria et al., 2019).

Finally, using either functional or non-functional methods, equation (1) can be written for individual i over time as

$$y_i = \alpha 1 + \widetilde{B}^t \widetilde{m} + \sum_{l=1}^{L} \widetilde{B}^{e^l} \widetilde{a}_l + \sum_{j=1}^{J} x_{ij} Z b_j + \varepsilon_i, \;\; \varepsilon_i \sim \mathcal{N}(0, \sigma^2 \Gamma) \tag{2}$$

where $y_i = (y_{it_1}, \ldots, y_{it_T})'$ corresponds to the T-dimensional vector of phenotypic values for individual i, $\widetilde{m}$ and $\widetilde{a}_l$ are the $(df - 1)$-dimensional vectors of B-spline coefficients associated to the smooth functions of time and of the l^{th} environmental variable.

In case of B-spline or P-spline approaches, Z is then equal to B^t and b_j are the df-dimensional vectors of coefficients associated to the j^{th} marker. Otherwise, $Z \equiv Id_T$ where Id_T is the $T \times T$ identity matrix and $b_j = (\beta_{jt_1}, \ldots, \beta_{jt_T})'$.

From a Bayesian perspective, penalties based on the first or second order finite differences on adjacent coefficients correspond to a multivariate first or second order random walk prior (Lang and Brezger, 2004). In the following, prior distribution for $\widetilde{m}$, $\widetilde{a}_l$ or b_j will be assumed to be:

$$\mathcal{N}\left(0, \tau_u (K)^{-1}\right) \tag{3}$$

where τ_u is a variance parameter specific for each group of unknown parameters: τ_m for $\widetilde{m}$, τ_{a_l} for $\widetilde{a}_l$, $l = 1, \ldots, L$, and τ_{b_j} for $b_j, j = 1, \ldots, J$. K is equal to $\widetilde{D}'_m \widetilde{D}_m$, $\widetilde{D}'_{a_l} \widetilde{D}_{a_l}$, $l = 1, \ldots, L$, or $D'D$, where D is the matrix representation of the first and second order finite differentiating operator, $\widetilde{D}_m$ and $\widetilde{D}_{a_l}$ are the associated re-parametrized versions of D (see appendix A.1 for more details).

In order to simultaneously select relevant markers j and estimate their associated effects b_j, group variable selection has to be performed. In a Bayesian regression model, various variable selection approaches have been proposed (O'Hara et al., 2009). In particular, the spike-and-slab prior has been widely and efficiently used (Malsiner-Walli and Wagner, 2011; Ghosh and Ghattas, 2015). The spike-and-slab prior is a discrete mixture of two distributions (George and McCulloch, 1993, 1997). The allocation to both components is controlled by a latent indicator variable γ_j that follows a Bernoulli distribution. Thus, if $\gamma_j = 1$ the coefficient will be assigned to the slab part and the variable will be included in the model. To simultaneously select molecular markers and estimate their effects, we propose to combine the random walk prior (see eq. (3)) of the coefficients with a spike-and-slab prior. In our context, we consider each vector of coefficients as a group and we specify on each vector a multivariate spike-and-slab prior with the random walk prior on the slab component and a Dirac mass at zero (Ghosh

and Ghattas, 2015; Yang and Narisetty, 2020) leading to the following prior:

$$\begin{aligned} b_j|\tau_{b_j},\gamma_j,\sigma^2 &\sim \gamma_j\mathcal{N}(0,\sigma^2(\tau_{b_j}D'D)^{-1})+(1-\gamma_j)\delta(0), \quad j=1,\dots,J \\ \tau_{b_j} &\sim \mathcal{IG}(s,r), \quad \gamma_j\sim\mathcal{B}er(\pi) \quad \text{and} \quad \pi\sim\mathcal{B}eta(1,1) \end{aligned} \tag{4}$$

where $\mathcal{IG}(s,r)$ is the Inverse Gamma distribution with shape and rate respectively equal to s and r. σ^2 is the residual variance, π is the *a priori* inclusion probability and $\mathcal{B}eta(1,1)$ denote the Beta distribution.

Finally, the dynamic QTL mapping model can be expressed as the following Bayesian hierarchical model:

$$\begin{aligned} y_i|\alpha,\widetilde{m},\widetilde{a},b,\rho,\sigma^2 &\sim \mathcal{N}(\alpha+\widetilde{B^t}\widetilde{m}+\sum_{l=1}^{L}\widetilde{B}^{e^l}\widetilde{a}_l+\sum_{j=1}^{J}x_{ij}Zb_j,\sigma^2\Gamma) \\ \alpha &\sim \mathcal{U}_{(-\infty,\infty)} \\ \widetilde{m}|\tau_m &\sim \mathcal{N}(0,(\tau_m\widetilde{D}_m'\widetilde{D}_m)^{-1}) \\ \widetilde{a}_l|\tau_{a_l} &\sim \mathcal{N}(0,(\tau_{a_l}\widetilde{D}_{a_l}'\widetilde{D}_{a_l})^{-1}), \quad l=1,\dots,L \\ b_j|\tau_{b_j},\gamma_j,\sigma^2 &\sim \gamma_j\mathcal{N}(0,\sigma^2(\tau_{b_j}D'D)^{-1})+(1-\gamma_j)\delta(0), \quad j=1,\dots,J \\ \tau_m,\ \tau_{a_l} \text{ and } \tau_{b_j} &\sim \mathcal{IG}(0.1,0.1), \quad l=1,\dots,L \text{ and } j=1,\dots,J \\ \gamma_j &\sim \mathcal{B}er(\pi), \quad j=1,\dots,J \text{ and } \pi\sim\mathcal{B}eta(1,1) \\ \rho &\sim \mathcal{U}_{(-1,1)}, \quad \sigma^2\sim\mathcal{IG}(0.1,0.1) \end{aligned} \tag{5}$$

where $\mathcal{U}_{(-1,1)}$ denotes the uniform distribution on the interval -1 to 1. The use of a Dirac spike may imply reducibility of the Markov chain ($\gamma_j=0$ implies $b_j=0$ and vice versa). To avoid it, it is essential to draw γ from the marginal posterior integrating over the regression coefficients b subject to selection, see Malsiner-Walli and Wagner (2011), Geweke (1996) and Smith et al. (1996). The details of the integration are provided in appendix A.2. This Bayesian hierarchical model (eq. (5)) relies on conditionally conjugate distributions. It allows analytical integration over the regression effects b and thus the development of an efficient Gibbs sampling algorithm (Gilks et al., 1995). The full conditional distributions for the group spike-and-slab prior are given in appendix A.3 and are available on `https://github.com/Heuclin/VCGSS`.

3. Simulations

This section aims to investigate through simulations the performance of the proposed models, by varying different parameters such as the degree of freedom, the residual variance, the number of observations (time steps and individuals), the number of markers, the correlation among them and considering several functional methods (Legendre polynomials (L), B-spline (BS) or P-splines with first or second order difference penalty (PS_1 / PS_2)) and non-functional methods (with first or second order difference penalty (RW_1 / RW_2)) combined with two variable selection priors (group spike-and-slab (GSS) or Bayesian group Lasso (BGL) (Kyung et al., 2010) (see appendix A.3 and A.4 for the full conditional distributions)). We also planned to test the approach proposed by Scheipl

et al. (2012) and implemented in the spikeSlabGAM R-package (Scheipl, 2011). Unfortunately, from computational and modeling perspectives, this was not possible. This method requires indeed data transformation, such as vectorization of matrices and Kronecker products, leading to manipulation of huge matrices, which is particularly the case in the longitudinal context. For example, assuming n = 300 individuals, T = 100 time points, and J = 100 genetic markers, the algorithm crashes on a high performance computer (28 cores, bi processor Intel Xeon E5-2680 v4 2,4 Ghz with 128 Go of RAM). In addition, spikeSlabGAM does not permit to consider residual dependencies within each individual to be structured over time, that may lead to spurious selection (Li and Sillanpää, 2013). In our paper, an AR(1) is used. Assuming independence impacts the variable selection process leading in particular to an increase of false positives. Furthermore, we also compare our different approaches with Vanhatalo et al.'s method that models the functional effects β_j with Gaussian process prior using a Mátern covariance function combined with a stepwise selection approach and taking also into account an AR(1) residual covariance structure. We will refer to this approach as S-GP. Note that in a Bayesian framework, the Legendre interpolation combined with Bayesian group Lasso has been already explored by Li and Sillanpää (2015).

In the following, whatever the number of markers J, only the first four markers are non-zeros and their functional effects are defined as follows:

$$\begin{aligned}
\beta_1(t) &= 4 - 0.08t, \\
\beta_2(t) &= \cos\left(\frac{\pi}{15}(t-25)\right) + \frac{t}{50}, \\
\beta_3(t) &= \frac{60}{25 + (t - \frac{T}{2})^2} \\
\beta_4(t) &= 2 * 1_{t \leq \frac{T}{3}} + 0 * 1_{\frac{2T}{3} < t \leq \frac{2T}{3}} + 1_{t > \frac{2T}{3}}.
\end{aligned} \tag{6}$$

The overall mean function is set to:

$$\mu(t) = 1 + \sin\left(\frac{\pi t}{20}\right). \tag{7}$$

Only one environmental variable is considered:

$$e_t^1 = \cos\left(\frac{\pi}{2}(t-25)\right) + \frac{1}{50}t \tag{8}$$

and its effect on phenotypes is defined for all t as

$$f_1(e_t^1) = 0.5e_t^1 + 0.3(e_t^1)^2. \tag{9}$$

The ratio of false positives (FP) and false negatives (FN) as well as Matthews correlation coefficient (MCC, Matthews (1975)) are recorded to evaluate the selection performances. For the GSS prior, a variable is assumed to be selected if its marginal posterior probability is greater than 0.5. For the BGL prior, a variable is selected if zero does not belong to the credible interval of at least one B-spline or Legendre coefficient. The estimation quality is assessed using the root mean square error (RMSE). For the additive

part $\alpha + \mu(t) + f_1(e_t^1)$, the error is jointly calculated for identifiability reasons. For ease of comparison, RMSEs calculated for each $\beta_j, j = 1, \ldots, 4$, are summed up in a unique value ($RMSE_\beta = \sum_{j=1}^{4} RMSE_{\beta_j}$). All results are based on 100 replications.

Impact of functional and non-functional methods on estimation and prediction performances

Functional methods depend on the degree of freedom (*df*) for the B- and P-spline interpolations and the polynomial degree (d) for the Legendre interpolation. In the following, ν is set to three such that cubic spline basis functions are used. To understand the impact of different methods, we first perform inference with different values of d ranging from 9 to 70, *df* ranging from 9 to 100, and assuming the true model is known (no variable selection, $J = 4$). The sample size n is set to 300, the number of time points T to 100, the residual variance σ^2 to 4 and the residual autocorrelation decay parameter ρ to 0.

Figure 1 presents the RMSEs calculated using the first three smooth effects $\beta_1(t)$, $\beta_2(t)$ and $\beta_3(t)$. It highlights the benefit of coefficient difference penalty. Indeed, among functional methods, the error generated by non penalised methods decreases until 0.118 and then increases. It emphasizes the difficulty to choose the number of polynomial degree / degree of freedom. The P-spline method generates an error that decreases to 0.1 and 0.092 for penalisation of order 1 and 2 respectively, then stabilizes when the degree increases. Thus, it outperforms non penalised methods and avoids overfitting. Finally, penalised non-functional methods perform equally well than non penalised functional methods at optimal degree. Figure 1b presents the RMSE of the piecewise constant effect $\beta_4(t)$. Because of the two jumps, the effect of $\beta_4(t)$ is a complicated task for functional methods, as confirmed here. Indeed the optimal estimations are reached for a degree of freedom equal to the number of time step T and are no better than the estimation generated by non-functional penalised methods. To ensure that the P-spline results showed in Figure 1a are not due to overfitting, a 10-folds cross-validation is performed and predictive RMSEs are given in Figure 1c. This confirms that P-splines are more robust to overfitting.

This simulation has showed that penalised methods outperform non-penalised method and avoid overfitting. Functional penalised methods are suitable for very smooth functions with no function values changing abruptly at any time point. On the contrary, non-functional penalised methods are suitable for more complex functions which can present jumps.

In the following, the *df* for B- or P-splines and d for Legendre interpolation will be fixed at $T/3$.

Impact of priors on variable selection

The second set of simulations aims at comparing BGL and GSS priors under functional and non-functional methods. These different prior combinations are also compared with

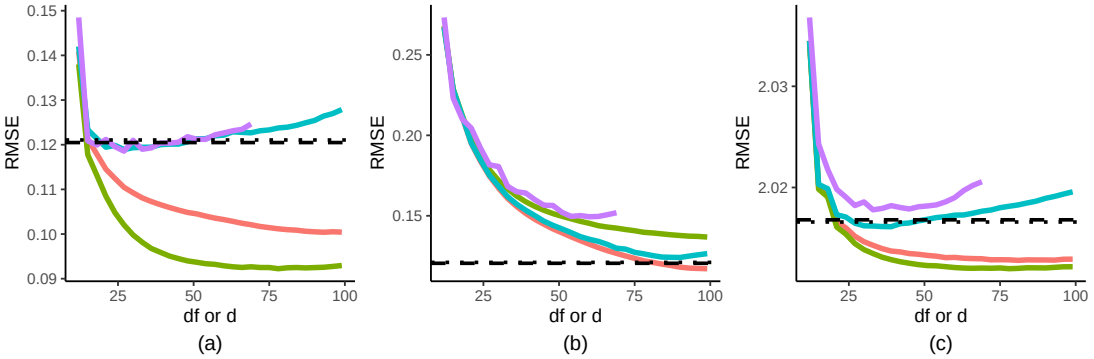


Fig. 1. Panel (a) presents the mean of RMSEs for functional estimation of the smooth effects $\beta_1(t)$, $\beta_2(t)$ and $\beta_3(t)$ for varying number of df and d. Panel (b) presents the RMSE for functional estimation of the piecewise constant effect $\beta_4(t)$ for varying number of df and d. Panel (c) presents the predictive RMSE using 10-folds cross-validation for varying number of df and d. Green, red, blue and purple lines correspond to P-splines 2, P-splines 1, B-splines and Legendre polynomial interpolation respectively. Dashed and dotted black lines correspond to non-functional interpolation with order 1 and 2 respectively.

the stepwise approach of Vanhatalo et al. (2019) combined with Gaussian process using Mátern covariance function to estimate functional effects (S-GP). The number of time points T is set to 100, the number of individuals n is set to 100 or 300 and the number of markers J is set to 3000 or 500 respectively. These scenarios are then coupled with a residual variance σ^2 set to 4 or 16 and a residual autocorrelation decay parameter ρ set to 0.4. When the number of individuals is high and the number of markers is low ($n = 300$ and $J = 500$, columns 1 and 2 in Table 1), BGL and GSS perform equally well regardless of the estimation method used. Both priors allow efficient selection of variables which leads to an MCC close to one. The S-GP approach also performs well with slightly lower MCC when the residual variance increases due to some FN. However, when the sample size is substantially smaller than the number of variables ($n = 100$ and $J = 3000$, columns 3 and 4 in Table 1), BGL and GSS perform differently. BGL fails to select 75% to 100% of the non-zero functions regardless of the estimation method used and leads to a decrease of the MCC down to 0. In order to determine the reasons for this behaviour, we calculated, for BGL combined with P-spline interpolation, the following root mean square errors

(a) between the observations and their predictions

$$RMSE_y = \sqrt{\frac{1}{nT}\sum_{k=1}^{T}\sum_{i=1}^{n}(\hat{y}_{i,t_k} - y_{i,t_k})^2},$$

(b) between the true non-zero functions and their estimations using all markers

$$RMSE_{B^tX} = \sqrt{\frac{1}{nT}\sum_{k=1}^{T}\sum_{i=1}^{n}\sum_{j=1}^{J}(x_{i,j}[B^T\hat{b}_j]_{t_k} - x_{i,j}\beta_j(t_k))^2},$$

Table 1. Matthews correlation coefficient (MCC), False negative (FN) in percentage and RMSE_β obtained using different priors and approaches. Standard deviations are given in brackets.

Criteria	Prior	n=300, J=500, σ^2=4	n=300, J=500, σ^2=16	n=100, J=3000, σ^2=4	n=100, J=3000, σ^2=16
MCC	BGL-PS	0.91 (0.08)	0.9 (0.082)	0.51 (0.041)	0
	BGL-BS	0.99 (0.041)	0.98 (0.046)	0.5 (0)	0
	BGL-L	0.75 (0.099)	0.7 (0.092)	0.5 (0)	0.2 (0.274)
	GSS-L	1 (1)	1 (1)	1 (1)	0.96 (0.962)
	GSS-BS	1 (0)	1 (0)	1 (0)	1 (0.019)
	GSS-PS_1	1 (0)	1 (0)	1 (0)	0.98 (0.044)
	GSS-PS_2	1 (1)	1 (1)	1 (1)	0.94 (0.941)
	GSS-RW_1	1 (0)	0.99 (0.027)	1 (0)	0.87 (0)
	GSS-RW_2	1 (0)	0.99 (0.027)	1 (0)	0.87 (0)
	S-GP	1 (0)	0.89 (0.05)	0.94 (0.063)	0.62 (0.141)
FN	BGL-PS	0	0	73.98 (4.998)	100 (0)
	BGL-BS	0	0	75 (0)	100 (0)
	BGL-L	0	0	75 (0)	90 (13.693)
	GSS-L	0	0	0	7 (7)
	GSS-BS	0	0	0	0.5 (3.536)
	GSS-PS_1	0	0	0	3 (8.207)
	GSS-PS_2	0	0	0	11 (11)
	GSS-RW_1	0	1 (4.949)	0	25 (0)
	GSS-RW_2	0	1 (4.949)	0	25 (0)
	S-GP	0	20.5 (9.702)	7.5 (11.573)	59 (18.736)
RMSE_β	BGL-PS	0.47 (0.083)	0.86 (0.17)	3.48 (0.248)	5.62 (0)
	BGL-BS	0.43 (0.042)	0.69 (0.091)	3.54 (0.065)	5.62 (0)
	BGL-L	0.75 (0.187)	1.53 (0.391)	3.56 (0.108)	4.83 (1.077)
	GSS-L	0.43 (0.429)	0.7 (0.695)	0.63 (0.628)	1.22 (1.224)
	GSS-BS	0.42 (0.022)	0.66 (0.042)	0.6 (0.04)	1.03 (0.1)
	GSS-PS_1	0.38 (0.024)	0.61 (0.041)	0.56 (0.04)	0.96 (0.176)
	GSS-PS_2	0.39 (0.39)	0.66 (0.665)	0.58 (0.578)	1.23 (1.234)
	GSS-RW_1	0.43 (0.024)	0.87 (0.106)	0.74 (0.041)	1.79 (0.054)
	GSS-RW_2	0.42 (0.04)	0.89 (0.131)	0.76(0.043)	1.81 (0.057)
	S-GP	0.44 (0.023)	1.05 (0.204)	0.76 (0.276)	2.87 (0.819)

(c) between the true non-zero functions and their estimations using the markers with true non-zero effects

$$RMSE_{B^tX_1} = \sqrt{\frac{1}{nT}\sum_{k=1}^{T}\sum_{i=1}^{n}\sum_{j=1}^{4}(x_{i,j}[B^T\hat{b}_j]_{t_k} - x_{i,j}\beta_j(t_k))^2},$$

(d) between 0 and the estimation using the markers with true null effects

$$RMSE_{B^tX_0} = \sqrt{\frac{1}{nT}\sum_{k=1}^{T}\sum_{i=1}^{n}\sum_{j=5}^{J}(x_{i,j}[B^T\hat{b}_j]_{t_k})^2}.$$

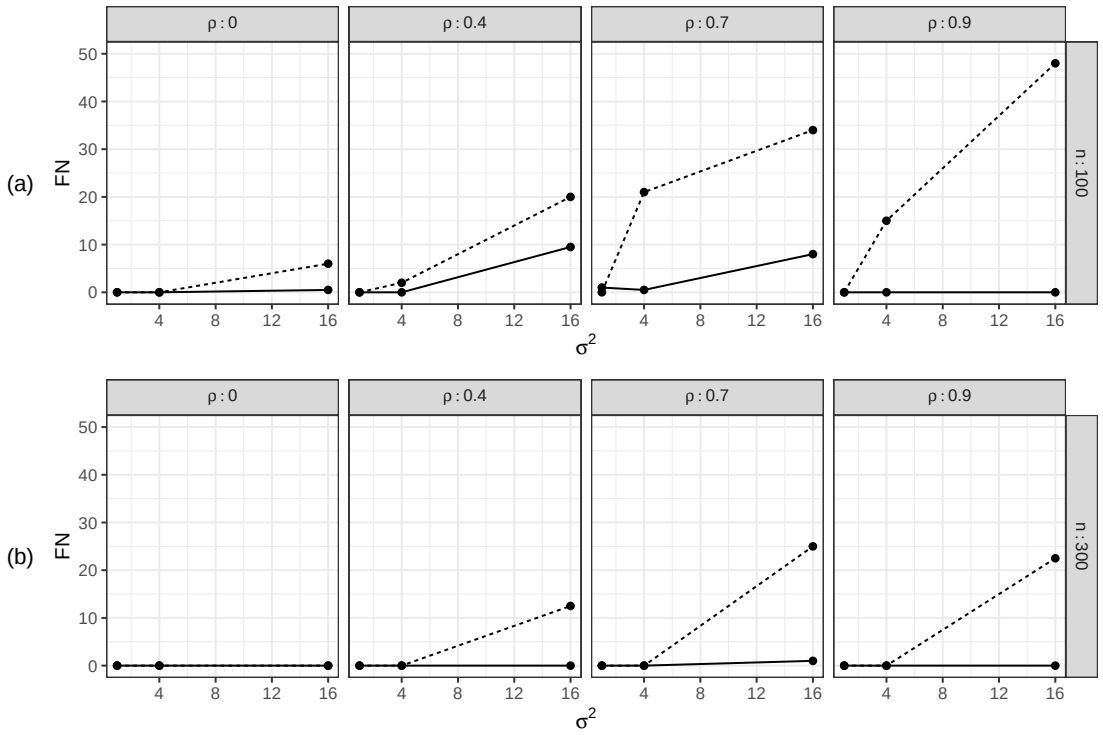
$RMSE_y$ and $RMSE_{B^tX}$ are very similar regardless of the number of individuals and markers (see Table 2). This suggests that even when the model selection fails, the global estimation remains acceptable. However, $RMSE_{B^tX_1}$ and $RMSE_{B^tX_0}$ clearly differ between the two cases ($n = 300,\ J = 500$ vs $n = 100,\ J = 3000$). In the first and more favorable case, both RMSEs are low while for the case where the number of markers is high compared to the number of individuals, the RMSEs increases substantially. In particular, $RMSE_{B^tX_0}$ is high demonstrating a clear over-estimation of the zero components and thus an under-estimation of the true non-zero parts. That is, BGL is not shrinking to zero the 2996 markers with no effect and is estimating them to have low values, while biasing toward zero the estimation of the four markers with true effects.

Table 2. RMSE between the observations and their predictions ($RMSE_y$), between the true non-zero functions and their estimations using all markers ($RMSE_{B^tX}$) or using the markers with true non-zero effects ($RMSE_{B^tX_1}$) and between 0 and the estimation using the markers with true null effects ($RMSE_{B^tX_0}$). All these quantities are obtained using BGL prior combined with P-spline interpolation. X denote the matrix associated to all markers, X_1 the marker matrix associated to the true non-zero effects and X_0 the marker matrix associated to the true zero effects.

n	J	σ^2	$RMSE_y$	$RMSE_{B^tX}$	$RMSE_{B^tX_1}$	$RMSE_{B^tX_0}$
300	500	4	2.64	0.89	0.44	0.93
100	3000	4	2.64	0.97	2.88	2.85

The biased estimations thereby impact the selection. The S-GP approach seems also sensitive to the complexity of the data. Indeed, the S-GP's MCC decreases to 0.62 due to a FN which reaches 59%. It is affected by the ratio of the number of observations to the number of variables and especially by the noise which degrades its selection ability. The selection performance of the GSS prior combined with non-functional methods (GSS-RW_1 / GSS-RW_2) also appears to be slightly affected by the noise when the number of individuals is low. Effectively, these combinations systematically miss variable 3 which is the smallest non-zero effect leading to 25% FN. GSS prior combined with functional method does not present the same comportment despite some false negatives (see Table 1). Li and Sillanpää (2013) showed that the non-functional method performs better when used with a diagonal covariance structure than with AR(1), in the sense that it does not erroneously shrink the effects of any marker toward zero when the number of observations is low and there is high temporal correlation among the residual errors. However, assuming a simple diagonal residual covariance structure tends to significantly underestimate the uncertainty, which may result in including some false positive markers into the variable selection. Therefore, the AR(1) covariance structure might be a more suitable choice. To investigate the limitations of the GSS prior combined with functional and non-functional methods in response to the data complexity, we simulate datasets with 100, 300 or 900 individuals, 20 time points, 500 markers, a residual variance equal to 1, 4 or 16 and a residual autocorrelation decay parameter ρ of 0, 0.4, 0.7 and 0.9. Figure 2 presents the results for GSS prior combined with P-spline interpolation and with non-functional method both with penalty of order 2. The GSS prior combined with non-functional method presents FN which increases with the noise (ρ and σ^2) when the number of observations is low (see Figure 2a) while GSS prior combined with P-spline interpolation does not. This phenomenon is less pronounced when the number of observations increases (see Figure 2b) and disappears totally when the number of individuals is high (n = 900). Thus, non-functional methods assuming AR(1) residual covariance may suffer from lack of statistical power when the data is complex (few observations with high noise) and may have difficulties to identify the correct origin of the observed dependency in this situation. The dimensional reduction caused by functional methods (number of parameters is divided by 3 using P-splines with $df = T/3$) implicitly increases the statistical power. Note that it also reduces the computation time (divided by 10 using $df = T/3$, see Table 4).

Fig. 2. Panel (a) presents the false negative (FN) rate in percentage for $n = 100$. Panel (b) presents the FN rate in percentage for $n = 300$. Black line corresponds to the GSS prior combined with P-spline interpolation and dashed line corresponds to the GSS prior combined with non-functional method both with penalty of order 2.



Finally, the correct selection leads to accurate estimation of parameters (see RMSE_β in Table 1). The RMSE_β in the first scenario where all approaches have a good selection confirms the performance of the different estimation methods. In addition we can see that the Gaussian process method has a comparable performance to the non-functional methods RW_1 and RW_2.

Impact of the number of individuals and time steps on GSS prior performance

To go a step further and better understand the impact of the number of individuals and time steps on the performance of GSS prior, we consider another set of simulations. In the following, we assume that only three markers have significant and constant effects of $0.1, 0.2$ and 0.3 over time. An additional marker is added with no effects. The number of time points T varies from 1 to 50 and the number of individuals n is set to 100, 300, 500 or 1000. The residual variance σ^2 is fixed to one and the residual autocorrelation decay parameter ρ to 0. We focus on the marginal posterior probabilities of inclusion $(P(\gamma_j = 1|y, X), j = 1, \ldots, 4)$ with all parameters fixed at their true values. Such an approach has already been used by Malsiner-Walli and Wagner (2011) to evaluate the performance of spike-and-slab priors. First, regardless of the number of individuals or time steps, the marker with null effect is never selected (see Figure 3). Next, if we focus on one time step, these simulations confirm that the number of individuals plays a crucial role in variable selection as already mentioned in Malsiner-Walli and Wagner (2011). Increasing the number of individuals leads to a clear improvement of all marginal posterior probabilities. For example, for the strongest effect of 0.3, when the number of individuals goes from 100 to 300 with one time step ($T = 1$), $P(\gamma_3 = 1|y, X)$ increases from 0.44 to 0.92 (see Figures 3a, 3b). For the smallest effect of 0.1, with one time step, $P(\gamma_1 = 1|y, X)$ increases from 0.01 to 0.34 when the number of individuals varies from 100 to 1000 (see Figures 3a, 3d). While increasing the number of individuals improves the posterior probabilities of inclusion, the number of time steps also plays a significant role. Indeed, in the first panel with $n = 100$, the probability of inclusion for the intermediate effect of 0.2 increases from 0.10 for one time step to more than 0.35 using 50 time steps. This phenomenon is more evident when $n = 300$ where $P(\gamma_2 = 1|y, X)$ jumps from 0.52 to 1 when considering around 10 or more time steps, or when $n = 1000$ and $P(\gamma_1 = 1|y, X)$ climbs from 0.01 for one time step to 1 with 20 or more time steps. Thus, combining a high number of individuals with longitudinal data improves the variable selection allowing the detection of small effects while strengthening the confidence in the strongest ones. These results demonstrate the superiority of longitudinal data analyses compared to a separate analysis at each time point.

Impact of correlation between markers

Correlation is a difficult task in practice especially when working with high-throughput genotyping data where the fine discretization of the genome leads to very strong collinearity between markers. So it is important to understand how the GSS prior will perform under this constraint. To study this kind of situation, we consider a new simulated

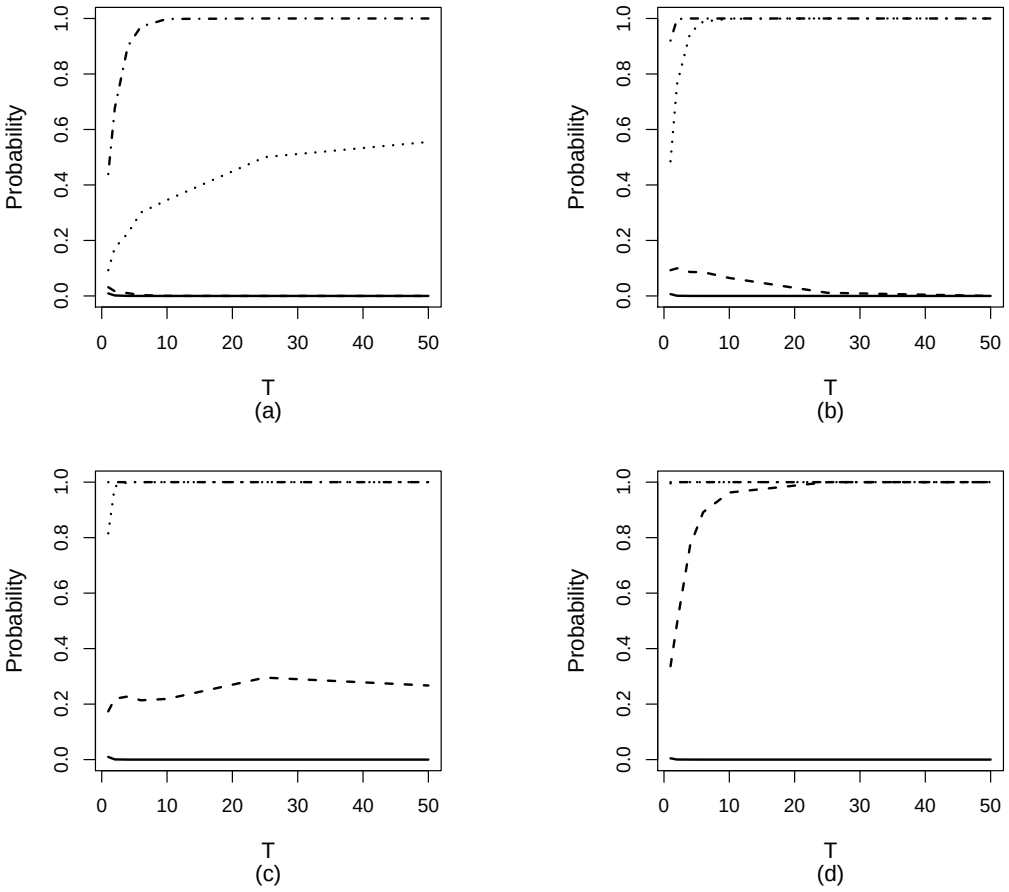


Fig. 3. Marginal probabilities of inclusion for each effect as a function of the number of time points T. Dotted-dashed line, dotted line, dashed line and solid line correspond to effects equal to $0.3, 0.2, 0.1$ and 0 respectively. Figures a, b, c and d are based on 100, 300, 500 and 1000 individuals respectively.

dataset constructed from markers provided from real case study on *Arabidopsis thaliana* (L. Heynh) (Marchadier et al., 2019) presented in section 4. Phenotypic observations y are simulated for 300 individuals over 100 time points from four independent groups of 9 correlated markers. The correlation between adjacent markers within group is set to $0.8, 0.9$ and 0.95 following the data process described in section 4. For the j^{th} group, only the 5^{th} marker has non-zero effect defined by $\beta_j(t)$ in equation (6), $j = 1, 2, 3$ or 4. The residual variance is set to 4 and the residual autocorrelation decay parameter ρ to 0.9.

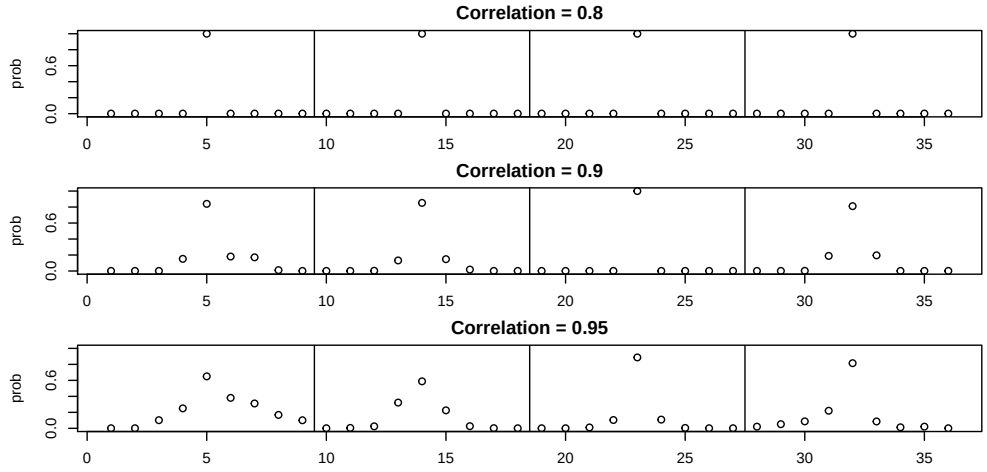


Fig. 4. Marginal probabilities of inclusion for each effect associated to correlated markers within four independent groups.

Figure 4 gives the marginal inclusion probability for each marker under different levels of correlation among them. It shows a clear impact of the correlation among markers on selection. The higher the correlation, the lower the marginal inclusion probabilities of the non-zero markers and the higher the marginal inclusion probabilities of adjacent zero markers. The correlation of 0.95 highlights this fact well. This is due to a switch of selection among markers that are highly correlated (adjacent markers) with the true non-zero markers. This result is in agreement with those of Malsiner-Walli and Wagner (2011) and Ghosh and Ghattas (2015) who have also studied the spike-and-slab prior under collinearity. Thus, when the data present high correlation, approaches using spike-and-slab prior lead to identification of a set of physically related markers defining genomic regions involved for the phenotypic observations. Ghosh and Ghattas (2015) advise against the use of Zellner's g-prior (leading to more false negative) and recommend a routine examination of the correlation matrix and calculation of the joint inclusion probabilities for correlated covariates, in addition to marginal inclusion probabilities, for assessing the importance of covariates.

4. Application

This application aims at disentangling the effects of the complex genetic architecture of shoot growth of *Arabidopsis thaliana* (L. Heynh) (Marchadier et al., 2019) and the impact of soil water conditions (SWC) on its dynamics. The complete phenotypic dataset is freely available at: `https://data.inra.fr/dataset.xhtml?persistentId=doi:10.15454/OCOP9B` (Loudet, 2018). The genotypic dataset is freely available at: `http://publiclines.versailles.inra.fr/page/8`. We focus on the phenotypic trait compactness of a recombinant inbred line (RIL) composed of 358 individuals followed during the vegetative growth from days 8 to 29 after sowing ($T = 21$). Compactness dynamics was observed along time using the high-throughput Phenoscope robot (Tisné et al., 2013). Compactness is the ratio between the projected rosette area and the convex hull area. Two environmental conditions are considered: well-watered (WW) and moderate water deficit (MWD) conditions. WW slowly decreases SWC from 100% on day one to 60% on day five, then maintains that level throughout the experiment. MWD let natural evaporation act until a threshold of 30% humidity is reached (see Figure 5a). The dynamics of compactness according to the two SWC are presented in Figures 5b and 5c. From 113 Single Nucleotide Polymorphisms (SNPs), the parental genotype probabilities were calculated at 538 positions for each individual using the *calc.genoprob* function in R/QTL package (Broman et al., 2003). These probabilites lead to 538 genetic predictors and are referred to "markers" in the following. Markers on different chromosomes are independent (mean correlation between chromosomes lower than 0.05). However, within a chromosome, markers are ordered such that adjacent markers share similar information and are highly correlated. Such dependencies among covariates is known to impact variable selection and parameter estimation as showed on our simulations and by others (Malsiner-Walli and Wagner, 2011; Ghosh and Ghattas, 2015). In order to reduce the collinearity, we process the data as follows: starting from the marker at the first position, we calculate its correlation with the subsequent markers. All markers with correlations greater than 0.95 are discarded and the first marker with a correlation less than 0.95 is retained, defining a new starting point. This procedure is repeated along the genome and results in the selection of 125 markers denoted $X_{0.95}$. Since this correlation threshold is high, we apply the procedure on the subset $X_{0.95}$ using a threshold of 0.7. This results in the selection of 38 markers among the previous 125, which we denote $X_{0.7}$. Selected markers are labelled by their chromosome numbers and their positions separated by an underscore, such that marker 1_1 corresponds to the first position on the first chromosome. Both environmental conditions are initially related to time with a linear decrease over the first few days then become constant for the remainder of the experiment. During the first phase, environmental effects are fully correlated with time. This raises identifiability problems and does not permit to model jointly a time varying intercept and environmental effects. Thus, the environmental factors are not included in the model. In addition, since genotype $\times$ environment interactions are not taken into account, we analyse separately each environmental condition.

In a nutshell, the study data consist of one phenotypic trait (compactness) measured over 21 time points ($T = 21$) on 358 individuals ($n = 358$) under two soil water conditions. We used two sets of covariates $X_{0.70}$ and $X_{0.95}$ containing 38 and 125 markers

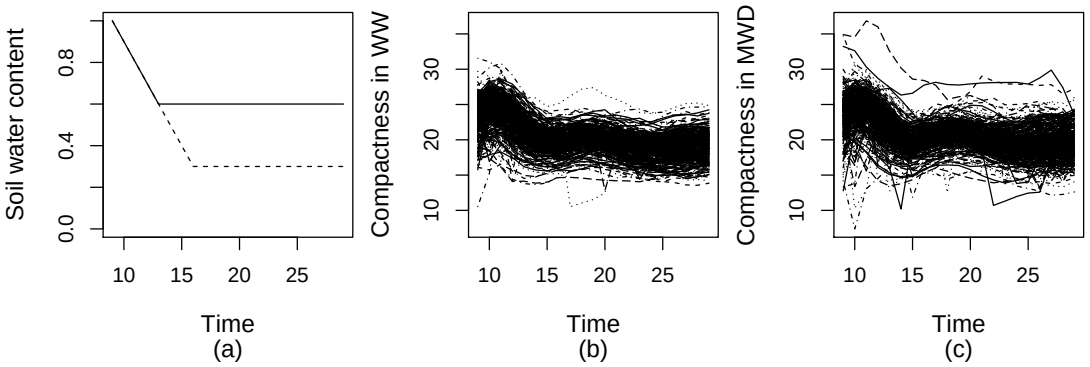


Fig. 5. Panel (a) presents the soil water content under the well-watered (WW) condition in solid line and the moderate water deficit (MWD) conditions in dashed line over time. Panel (b) presents compactness trait observations for the 358 individuals under the WW condition over 21 days. Panel (c) presents compactness trait observations for the 358 individuals under the MWD condition over 21 days.

respectively. The two SWC are analyzed separately to identify differences in the genetic architecture between the conditions. The results are based on 100 MCMC chains initialized at random starting values, each with 1,000,000 iterations, a burn-in of 500,000 and a thinning of ten. Gelman and Rubin's potential scale reduction factors (Gelman et al., 1992) for all continuous parameters and log predictive density (log-likelihood) are close to 1, indicating convergence. More details are presented in the supplementary materials. All output statistics are based on the pooled five million posterior samples.

Selecting relevant markers for WW condition: in the case of low correlations between markers, the selection procedure is highly stable. Figure 6 presents the mean of the marginal posterior inclusion probability for each marker using the PS_2 method across the pooled 10 million posterior samples. Eight markers (1_1, 1_20, 1_110, 2_62, 4_45, 5_33, 5_76 and 5_104) are included in the model with marginal posterior probabilities of one. Seven other markers have a marginal posterior inclusion probabilities lower than one but strictly greater than zero. Among these, for the markers (1_79, 1_97) and (3_14, 3_25) the algorithm tends to switch between the two adjacent markers. Indeed, we first note that the joint inclusion probabilities $\mathbb{P}(\gamma_{1\_79} = 1 \cap \gamma_{1\_97} = 1)$ and $\mathbb{P}(\gamma_{3\_14} = 1 \cap \gamma_{3\_25} = 1)$ are close to zero (lower than 10^{-4}), demonstrating that these two consecutive markers are hardly ever selected simultaneously. Second, the sum of the marginal posterior inclusion probabilities for each pair is equal to one. Thus, the algorithm switches from one marker to another. The three markers 2_47, 3_1 and 3_91 have marginal posterior inclusion probabilities of 0.07, 0.9, 0.97 respectively and have no adjacent markers selected. The switch between included markers can be explained by the pre-selection procedure. Using a threshold of 0.7 and starting from the first position may have led to the removal of other relevant markers or genomic regions, and the retained markers may not actually be relevant but only be close to or encompassing relevant regions. To validate this assumption, GSS-PS_2 is applied to the $X_{0.95}$ dataset.

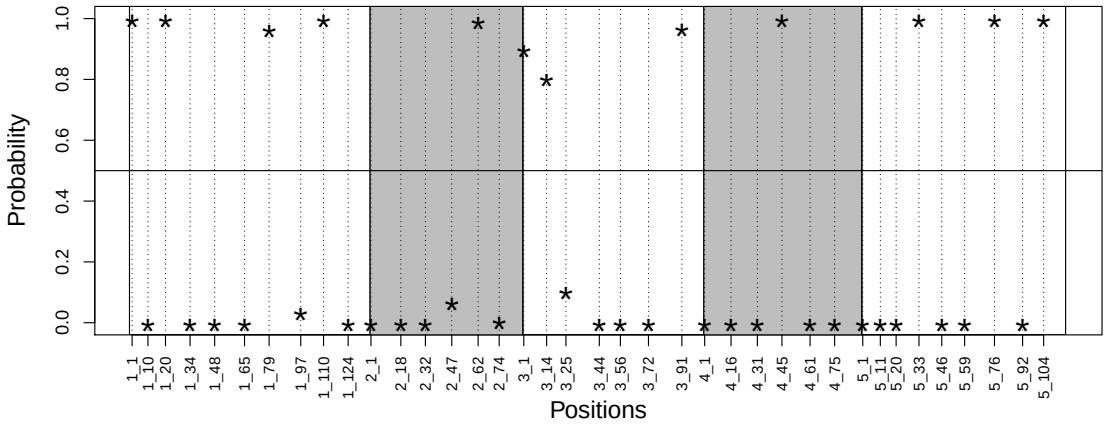


Fig. 6. Marginal posterior inclusion probabilities for the 38 markers in the genetic data $X_{0.7}$ using the PS_2 method. The alternation of white and gray area delimites the 5 chromosomes. A line at 0.5 representing a threshold at 0.5 is plotted.

Revealing genomic regions for WW condition: markers in the $X_{0.95}$ subset are highly correlated but offer a better coverage of the genome. Strong collinearity between covariates can lead to a multimodal posterior distribution and posterior distributions have to be carefully analyzed Ghosh and Ghattas (2015). In particular, it can be troublesome for variable selection where subsets are weakly separable (Rocková and George, 2014). For highly correlated covariates, at a given MCMC iteration, one particular covariate can switch with another as shown on simulations. This phenomenon is classically observed using spike-and-slab priors. However, this drawback can be lifted to identify potential genomic regions involved in phenotypic variations. Applying PS_2 method on the $X_{0.95}$

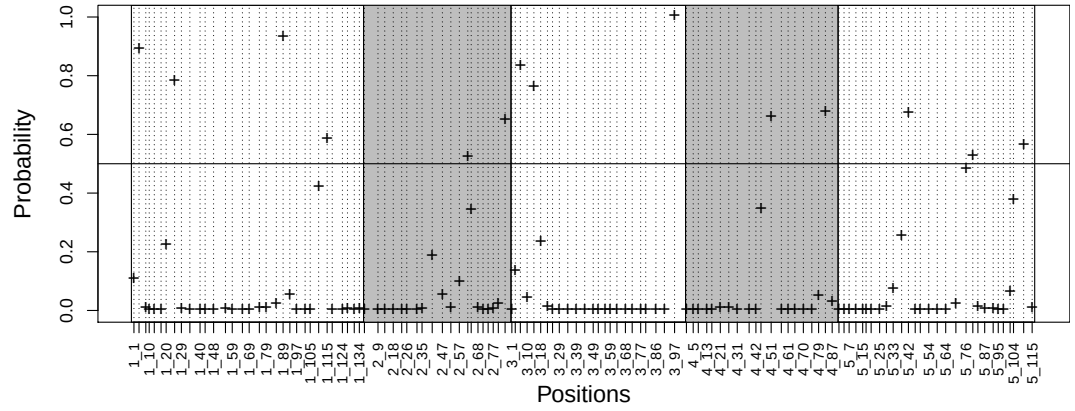


Fig. 7. Marginal posterior inclusion probabilities for the 125 markers of the genetic data $X_{0.95}$ using the PS_2 method. The alternation of white and gray area delimits the five chromosomes. A line at 0.5 representing a threshold at 0.5 is plotted.

subset allows us to check this (see Figure 7). For the $X_{0.70}$ subset, a model which contains

Table 3. Table of the identified relevant regions. Columns 2 and 3 indicate the markers or the range of markers corresponding to regions identified using the PS_2 method on the $X_{0.7}$ and $X_{0.95}$ subsets respectively. Column 4 indicates the markers or the range of markers corresponding to regions identified using the RW_2 method on the $X_{0.95}$ subset. The last column indicates if regions were identified by Marchadier et al. (2019).

Region	$X_{0.70}$ & **PS_2**	$X_{0.95}$ & **PS_2**	$X_{0.95}$ & **RW_2**	**Marchadier et al. (2019)**
1	1_1	1_1 → 1_4	1_4 → 1_8	
2	1_20	1_20 → 1_25	1_20	yes
3	1_79 → 1_97	1_85 →1_93	1_85 →1_89	
4	1_110	1_110 → 1_115		
5	2_62	2_57 → 2_64	2_57 → 2_64	yes
6		2_80 → 2_84		
7	3_1	3_3 → 3_10		yes
8	3_14 → 3_25	3_14 → 3_18		
9	3_97	3_97	3_97	yes
10	4_45	4_45 → 4_51	4_45	yes
11		4_79 → 4_87		
12	5_33	5_33→ 5_42		
13	5_76	5_76 → 5_80	5_64	yes
14	5_104	5_102 → 5_110		yes

12 markers (see Figure 6) is clearly favored with a joint posterior probability of 0.74, while no consensus can be reached based on $X_{0.95}$ as the joint posterior probabilities of the top three models are only $0.027, 0.026$ and 0.022. However and interestingly, the selected positions and models are similar. For example, the first three markers, $1\_1, 1\_2$ and 1_4 are never selected simultaneously ($\mathbb{P}(\gamma_{1\_1} = \ \cap\ \gamma_{1\_2} = 1 \ \cap\ \gamma_{1\_4} = 1) = 0$) but are complementary: $\mathbb{P}(\gamma_{1\_1} = 1) + \mathbb{P}(\gamma_{1\_2} = 1) + \mathbb{P}(\gamma_{1\_4} = 1) = 1$. This phenomenon is observed for most switching positions allowing the delimitation of 14 genetic regions that may be involved in compactness variation (see Table 3). From Table 3 several additional observations can be made. All markers or regions detected using $X_{0.70}$ match those identified with $X_{0.95}$ (see columns 2 and 3 of Table 3). The use of $X_{0.95}$ leads to the selection of two additional regions (regions 6 and 11), and regions 3 and 8 seem narrower with $X_{0.95}$. Thus, a more intensive repartition of markers along the genome, while avoiding extremely high correlations, allows the detection of genetic regions potentially involved in the underlying genetic architecture.

We compare PS_1 and PS_2 methods applied on the subsets $X_{0.70}$ and $X_{0.95}$. The results are identical demonstrating no impact of the order difference penalty (see Figure 8). We also compare the PS_2 and RW_2 methods. The results are different in terms of selection. Indeed, the number of selected markers or regions are lower with RW_2 than PS_2 with for instance 7 regions identified among the 14 of PS_2 using the $X_{0.95}$ subset. The estimation of the residual correlation is roughly equal to 0.9 using all methods. This high correlation seems to influence the selection process when using RW_1 or RW_2 methods, as already observed on simulations.

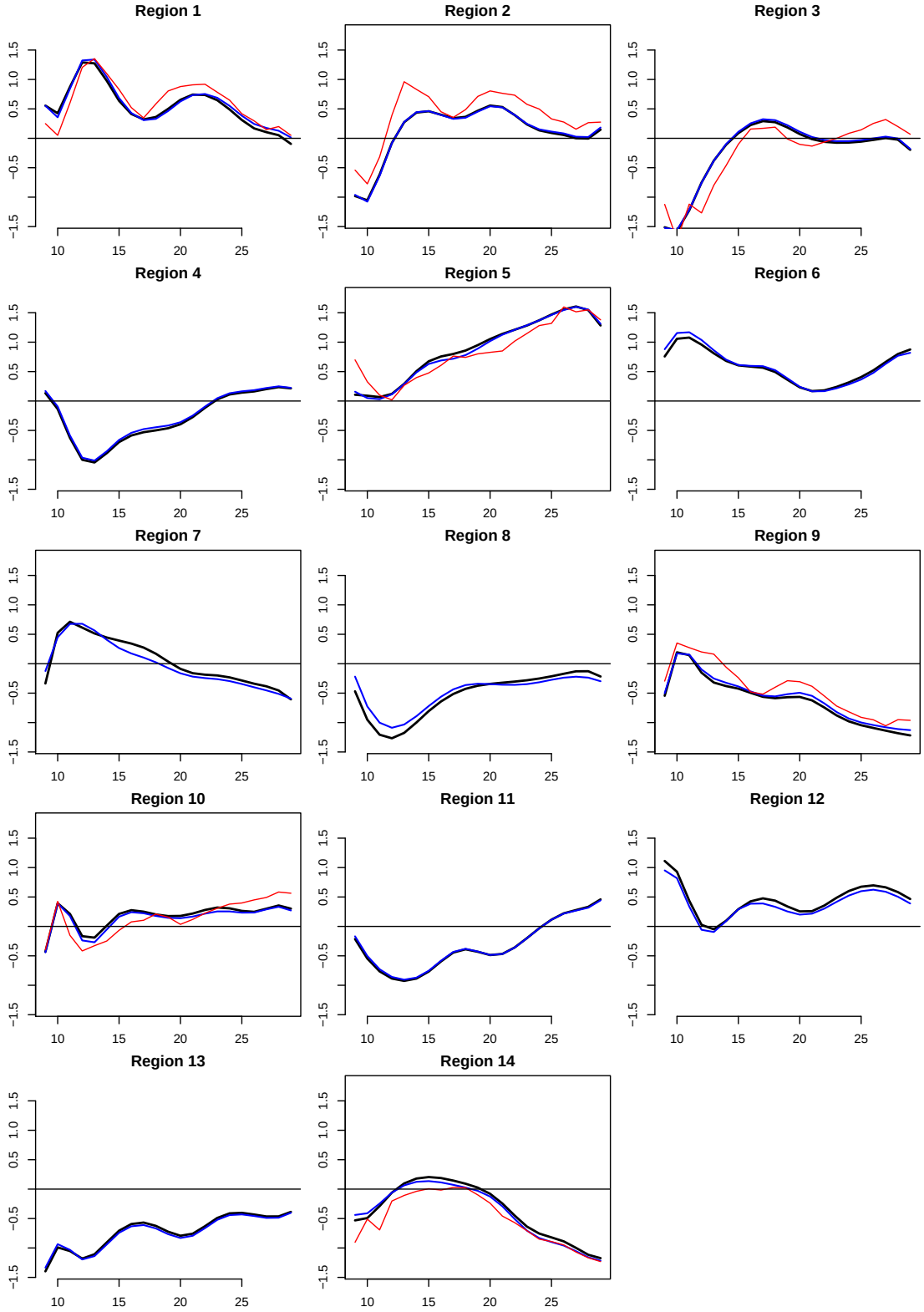


Fig. 8. Estimation of the effect for the marker which has the highest marginal posterior inclusion probability within each region in the $X_{0.95}$ subset. The blue, black, and red lines represent the estimation using the PS_1, PS_2, and RW_2 methods respectively. Plots with box are associated to markers which are identified by Marchadier et al. (2019).

Impact of MWD condition: applying the PS_2 method to compactness measured in MWD condition using the $X_{0.70}$ as well as $X_{0.95}$ subsets reveals no clear impact of the MWD condition on the complex genetic architecture of shoot growth and its dynamics. Among the 12 positions selected in the WW condition using $X_{0.70}$, seven positions are also selected in the MWD condition. Using $X_{0.95}$, 12 genomic regions in the MWD condition overlap with the 14 selected regions in the WW condition. Interestingly, among the 5 positions selected for WW but not MWD using $X_{0.70}$, three positions belong to the 12 shared genomic regions while the two last positions belong to the two unselected regions in MWD. Two hypotheses can explain such differences: (i) a genotype $\times$ environment interaction effect or (ii) an experimental effect. For the PS_2 method, when comparing cumulated effects estimated using the seven shared positions, no difference can be observed between the two conditions (see Figure 9). Moreover, when plotting the effects of the two markers selected in WW condition but not in the MWD condition (see Figure 8, regions 7 and 12), it seems that these two positions impact compactness from the beginning to the end of the experiment. Such results do not support either hypotheses.

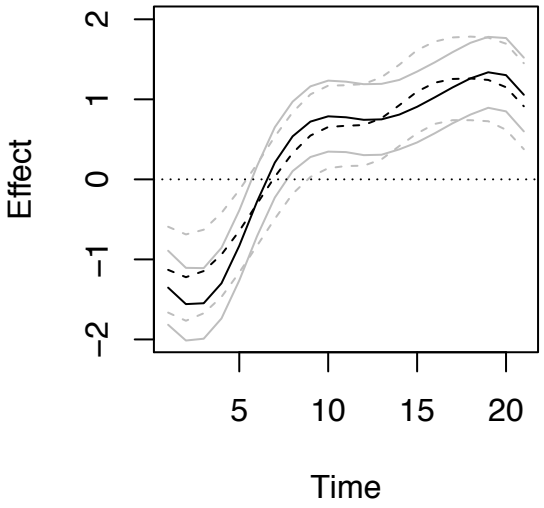


Fig. 9. Cumulative genetic effect of common markers selected in both conditions. The solid line represents the effect for the WW condition and the dashed line represents the effect of MWD conditions. Gray lines represent 95% credible intervals.

Comparative results: in an earlier study, Marchadier et al. (2019) identified in the WW condition eight significant markers involved in compactness variability for the last experimental day ($T = 29$) using a single time analysis. Seven of them match the regions we identified (Table 3, column 6 and Figure 8). Using the PS_2 method, we also identified seven additional regions that were not detected by Marchadier et al. (2019). These additional regions are identified by taking into account the dynamics of the phenotypic trait. Indeed, considering the observations of all individuals over the T times

selects markers which can have an effect only at a few times unlike a single time point analysis as proposed by Marchadier et al. (2019). For example, marker "1_89", which has the highest posterior inclusion probability within the third region (see Figure 8), shows an effect only at the early stage of the vegetative growth process. Thus, it can't be identified using the last day as in Marchadier et al. (2019). Another advantage of considering functional variations of the effects allows a better understanding of the genetic architecture.

Finally using functional methods such as P-spline interpolation compared to non-functional approaches reduces the number of parameters and thus indirectly increases the statistical power.

5. Conclusion

In this article we proposed a Bayesian varying coefficient model with variable selection for studying the dynamic genetic architecture of a complex trait.

The model combines a group spike-and-slab prior for the selection of markers with a P-spline interpolation or direct estimation of time coefficient functions. Both methods use first or second order difference penalty to ensure smoothness of the genetic functional effects. We evaluate the performance of the model through different simulations. We show that our approaches outperform, in terms of estimation as well as prediction, models using B-spline or Legendre interpolation in combination with group spike-and-slab or Bayesian group LASSO priors, as well as the alternative approach of Vanhatalo et al. (2019). P-spline interpolation is more suitable for very smooth genetic effect while direct estimation of time coefficient functions with difference penalty is more suitable for more complex effect with potential jumps. However, simulations demonstrate that direct estimation of time coefficient functions with difference penalty is more sensitive to noise (residual variance and residual time correlation) leading to false negative. P-spline interpolation reduces the number of parameters which indirectly increases the statistical power. Considering a point mass at zero for the spike part of the prior distribution of the regression coefficients improves the selection and thereby the quality of the estimation (George and McCulloch, 1997). Moreover, an investigation of the marginal inclusion probability associated to each covariate reveals the importance of the number of time points in the variable selection performance.

From a practical point of view, we show that a longitudinal approach allows a better detection of relevant markers or genomic regions compared to an approach that analyzes a single time point as proposed in Marchadier et al. (2019). In addition, as classically observed in genetic studies, markers present high correlation, thus requiring pre-selection. In this paper, we considered two correlation thresholds for the pre-selection leading to two subsets of markers considered for the analysis. The first subset with moderate correlation between markers allows a clear identification of positions and the estimation of their associated functional effects. The second, with high correlation among markers and more intensive coverage of the genome, allows the identification of genomic regions but the estimation of their associated effects is unreliable due to identifiability issues. This aspect has been observed on our simulations and was already reported by others (Ghosh and Ghattas, 2015; Malsiner-Walli and Wagner, 2011). Further research is needed for

variable selection in the presence of high collinearity between covariates, for example considering alternative priors such as g-priors (Malsiner-Walli and Wagner, 2011; Ghosh and Ghattas, 2015) or priors defined using the order structure information of markers along the genome.

Finally, more or less complex extensions should be considered. In this work we assumed that time points are common to all individuals. This could be restrictive in some applications. However such assumption could be easily relaxed as done by (Li and Sillanpää, 2015), who defined a B-spline basis for each individual. Moreover, our model considered a time-varying environmental condition and genetic markers to have additive effects. The functional estimation of the genetic effects captures the dynamics associated to each marker. However, the additivity assumption does not permit to determine if these estimated effects are directly related to the physiological processes or to the time-varying environmental condition. Genotype-by-environment (GE) interactions may impact the dynamic genetic architecture of complex traits and the selection procedure. One possible solution for incorporating GE interactions could be the addition of a functional effect depending on the environmental condition for each marker. But such an approach is computationally challenging. Finally, in this paper, only one time-varying environmental condition common to all individuals is considered. Another extension would involve the integration of different environmental conditions for the same genotypes and evaluating GE interactions.

Avaiability of the *Arabidopsis thaliana* (L. Heynh) dataset

The complete phenotypic dataset is freely available on: `https://data.inra.fr/dataset.xhtml?persistentId=doi:10.15454/OCOP9B` (Loudet, 2018). The genotypic dataset is freely available on: `http://publiclines.versailles.inra.fr/page/8`.

Acknowledgement

We thank S. Tisné for the fruitful discussions around *Arabidopsis thaliana* (L. Heynh). We also thank M.G. Tadesse for her helpful comments. M. Denis was partially supported by the European Union's Horizon 2020 research and innovation program under grant agreement No773383. We thank the two reviewers and the associate editor for their numerous valuable comments and suggestions, which substantially improved the paper.

References

Bitto, A. and Frühwirth-Schnatter, S. (2019) Achieving shrinkage in a time-varying parameter model framework. *Journal of Econometrics*, **210**, 75–97.

Broman, K., Wu, H., Sen, and Churchill, G. (2003) R/qtl: Qtl mapping in experimental crosses. *Bioinformatics*, **19**, 889–890.

Bruder, B., Dao, T.-L., Richard, J.-C. and Roncalli, T. (2011) Trend filtering methods for momentum strategies. *Available at SSRN 2289097*.

De Boor, C., De Boor, C., Mathématicien, E.-U., De Boor, C. and De Boor, C. (1978) *A Practical Guide to Splines*, vol. 27. Springer-Verlag New York.

Eilers, P. and Marx, B. (1996) Flexible smoothing with B -splines and penalties. *Statistical Science*, **11**, 89–121.

Fahrmeir, L. and Kneib, T. (2011) *Bayesian Smoothing and Regression for Longitudinal, Spatial and Event History Data*. Oxford University Press.

Fahrmeir, L., Kneib, T. and Lang, S. (2004) Penalized structured additive regression for space-time data: a bayesian perspective. *Statistica Sinica*, **14**, 731–761.

Fan, J. and Li, R. (2001) Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, **96**, 1348–1360.

Franco-Villoria, M., Ventrucci, M. and Rue, H. (2019) A unified view on bayesian varying coefficient models. *Electronic Journal of Statistics*, **13**, 5334–5359.

Frühwirth-Schnatter, S. and Wagner, H. (2010) Stochastic model specification search for gaussian and partial non-gaussian state space models. *Journal of Econometrics*, **154**, 85–100.

Gelman, A., Rubin, D. B. et al. (1992) Inference from iterative simulation using multiple sequences. *Statistical science*, **7**, 457–472.

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

— (1997) Approaches for Bayesian variable selection. *Statistica sinica*, 339–373.

Geweke, J. (1996) Variable selection and model comparison in regression. *In Bayesian Statistics 5*.

Ghosh, J. and Ghattas, A. (2015) Bayesian Variable Selection Under Collinearity. *The American Statistician*, **69**, 165–173.

Gilks, W., Richardson, S. and Spiegelhalter, D. (1995) *Markov Chain Monte Carlo in Practice*. CRC press.

Gong, Y. and Zou, F. (2012) Varying coefficient models for mapping quantitative trait loci using recombinant inbred intercrosses. *Genetics*, **190**, 475–486.

Hansen, T. (2006) The Evolution of Genetic Architecture. *Annual Review of Ecology, Evolution, and Systematics*, **37**, 123–157.

Hastie, T. and Tibshirani, R. (1986) *Generalized Additive Models*, vol. 1. The Institute of Mathematical Statistics.

— (1993) Varying-Coefficient Models. *Journal of the Royal Statistical Society. Series B (Methodological)*, **55**, 757–796.

Ishwaran, H. and Rao, J. S. (2005) Spike and slab variable selection: Frequentist and bayesian strategies. *The Annals of Statistics*, **33**, 730–773.

671 Kim, S.-J., Koh, K., Boyd, S. and Gorinevsky, D. (2009) \ell_1 trend filtering. *SIAM*
672 *review*, **51**, 339–360.

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

675 Lang, S. and Brezger, A. (2004) Bayesian P-Splines. *Journal of Computational and*
676 *Graphical Statistics*, **13**, 183–212.

677 Li, J., Wang, Z., Li, R. and Wu, R. (2015) Bayesian group Lasso for nonparamet-
678 ric varying-coefficient models with application to functional genome-wide association
679 studies. *The Annals of Applied Statistics*, **9**, 640–664.

680 Li, Y. and Wu, R. (2010) Functional mapping of growth and development. *Biological*
681 *Reviews*, **85**, 207–216.

682 Li, Z. and Sillanpää, M. (2013) A Bayesian Nonparametric Approach for Mapping Dy-
683 namic Quantitative Traits. *Genetics*, **194**, 997–1016.

684 — (2015) Dynamic Quantitative Trait Locus Analysis of Plant Phenomic Data. *Trends*
685 *in Plant Science*, **20**, 822–833.

686 Loudet, O. (2018) Raw phenotypic data obtained on the arabidopsis rils with the
687 phenoscope robots (marchadier, hanemian, tisnÃ© et al., 2019). URL: `https:`
688 `//doi.org/10.15454/OCOP9B`.

689 Ma, C.-X., Casella, G. and Wu, R. (2002) Functional Mapping of Quantitative Trait
690 Loci Underlying the Character Process: A Theoretical Framework. *Genetics*, 12.

691 Malsiner-Walli, G. and Wagner, H. (2011) Comparing spike and slab priors for Bayesian
692 variable selection. *Austrian Journal of Statistics*, **40**, 241–264.

693 Marchadier, E., Hanemian, M., Tisne, S., Bach, L., Bazakos, C., Gilbault, E., Haddadi,
694 P., Virlouvet, L. and Loudet, O. (2019) The complex genetic architecture of shoot
695 growth natural variation in Arabidopsis thaliana. *Plos Genetics*, **15**.

696 Matthews, B. W. (1975) Comparison of the predicted and observed secondary structure
697 of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, **405**,
698 442–451.

699 Min, L., Yang, R., Wang, X. and Wang, B. (2011) Bayesian analysis for genetic archi-
700 tecture of dynamic traits. *Heredity*, **106**, 124–133.

701 Ni, Y., Stingo, F., Ha, M., Akbani, R. and Baladandayuthapani, V. (2019) Bayesian hier-
702 archical varying-sparsity regression models with application to cancer proteogenomics.
703 *Journal of the American Statistical Association*, **114**, 48–60.

704 O'Hara, R. B., Sillanpää, M. J. et al. (2009) A review of Bayesian variable selection
705 methods: What, how and which. *Bayesian analysis*, **4**, 85–117.

706 O'Sullivan, F. (1986) A statistical perspective on ill-posed in- verse problems (with
707 discussion). *Statistical Science*, **1**, 505–527.

— (1988) Fast computation of fully automated log- density and log-hazard estimators. *SIAM Journal on Scientific Computing (SISC)*, **9**, 363–379.

Pérez, M.-E., Pericchi, L. R. and Ramírez, I. C. (2017) The scaled beta2 distribution as a robust prior for scales. *Bayesian Analysis*, **12**, 615–637.

Rasmussen, C. E. and Williams, C. K. (2006) *Gaussian processes for machine learning*, vol. 2. MIT press Cambridge, MA.

Rocková, V. and George, E. (2014) Negotiating multicollinearity with spike-and-slab priors. *Metron*, **72**, 217–229.

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

Scheipl, F. (2011) spikeslabgam: Bayesian variable selection, model choice and regularization for generalized additive mixed models in r. *arXiv preprint arXiv:1105.5253*.

Scheipl, F., Fahrmeir, L. and Kneib, T. (2012) Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association*, **107**, 1518–1532.

Smith, M., Kohn, R. et al. (1996) Nonparametric regression using bayesian variable selection. *Journal of Econometrics*, **75**, 317–344.

Tisné, S., Serrand, Y., Bach, L., Gilbault, E., Ben Ameur, R., Balasse, H., Voisin, R., Bouchez, D., Durand-Tardif, M., Guerche, P., Chareyron, G., Da Rugna, J., Camilleri, C. and Loudet, O. (2013) Phenoscope: an automated large-scale phenotyping platform offering high spatial homogeneity. *The Plant Journal*, **74**, 534–544.

Vanhatalo, J., Li, Z. and Sillanpää, M. (2019) A Gaussian process model and Bayesian variable selection for mapping function-valued quantitative traits with incomplete phenotypic data. *Bioinformatics*.

Wang, L., Li, H. and Huang, J. (2008) Variable Selection in Nonparametric Varying-Coefficient Models for Analysis of Repeated Measurements. *Journal of the American Statistical Association*, **103**, 1556–1569.

Wood, S. (2017) *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.

Wu, R., Ma, C., Zhao, W. and Casella, G. (2003) Functional mapping for quantitative trait loci governing growth rates: A parametric model. *Physiological Genomics*, **14**, 241–249.

Yang, X. and Narisetty, N. N. (2020) Consistent group selection with bayesian high dimensional modeling. *Bayesian Analysis*.

A. Appendix

A.1. Estimation of centered function using interpolation approach

For identifiability reasons in VC models, the h functions to be interpolated for the intercept and the environmental effect have to be centered. This means $\int_{\Re} h(x)dx = 0$ (Hastie and Tibshirani, 1986; Wood, 2017). Let B^x denote the $(T \times df)$-dimensional matrix containing the basis functions calculated at $x = (x_1, .., x_t)'$. Let also denote c a df-dimensional vector of associated coefficients such that

$$h(x) = B^x c. \tag{10}$$

To satisfy the centering constraint on $h(.)$, the sum of the elements of $h(x)$ must be zero $(1'B^x c = 0)$. This can be achieved by a re-parametrisation of B^x and c using a QR decomposition as explained by Wood (2017) in section 1.8.1 and 4.2. Let

$$(1'B^x)' = Q \begin{bmatrix} R \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

the QR decomposition of $(1'B^x)'$ where Q is a $(df \times df)$-dimensional orthogonal matrix and R is a scalar in this case. By taking Z the $df - 1$ last columns of Q we obtain that

$$1'B^x Z = (0 \dots 0).$$

Now, we can rewrite Equation (10) by defining a new $(df - 1)$-dimensional parameters vector $\widetilde{c}$ such that $c = Z\widetilde{c}$ and a new $T \times (df - 1)$ basis functions matrix $\widetilde{B}^x = B^x Z$ leading to $B^x c = \widetilde{B}^x \widetilde{c}$ which satisfies the constraint.
If adjacent coefficients are penalized as in P-spline interpolation, the new parameters $\widetilde{c}$ imply also a re-parametrisation of the matrix of the finite differentiating operator D by $\widetilde{D} = DZ$. Thus $c'D'Dc$ is equal to $\widetilde{c}'\widetilde{D}'\widetilde{D}\widetilde{c}$.

A.2. Detail of the full conditional distribution of γ_k

Let Θ the set of all parameters $\{\alpha, \widetilde{m}, \tau_m, \widetilde{a_1}, \ldots, \widetilde{a_L}, \tau_{a_1}, \ldots, \tau_{a_L}, b_1, \ldots, b_J, \gamma_1, \ldots, \gamma_J, \tau_{b_1}, \ldots, \tau_{b_J}, \pi, \rho, \sigma^2\}$ in the Bayesian hierarchical model (5), Θ_{k_0} and Θ_{k_1} be Θ with $\gamma_k = 0$ and $\gamma_k = 1$ respectively. Let

$$\bar{y_i} = y_i - \alpha 1 - \widetilde{B^t}\widetilde{m} - \sum_{l=1}^{L} \widetilde{B}^{e^l}\widetilde{a}_l - \sum_{j=1}^{J} x_{i,j} Z b_j$$

and

$$\bar{y}_{i_{-k}} = y_i - \alpha 1 - \widetilde{B^t}\widetilde{m} - \sum_{l=1}^{L} \widetilde{B}^{e^l}\widetilde{a}_l - \sum_{j=1;j\neq k}^{J} x_{i,j} Z b_j.$$

$$P(y|\Theta_{k_1} \setminus \{b_k\}) = \int_{\mathbb{R}} P(y|.)P(b_k|\gamma_k = 1)\partial b_k$$

$$= \int_{\mathbb{R}} \frac{1}{(2\pi\sigma^2)^{\frac{nT}{2}}|\Gamma|^{\frac{n}{2}}} exp\Big\{ -\frac{1}{2\sigma^2}\sum_{i=1}^{n} \bar{y}_i'\Gamma^{-1}\bar{y}_i \Big\} \frac{|D'D|^{\frac{1}{2}}}{(2\pi\sigma^2\tau_{b_k}`)^{\frac{df}{2}}} exp\Big\{ -\frac{1}{2\sigma^2\tau_{b_k}} b_k' D'D b_k \Big\}\partial b_k$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{nT}{2}}|\Gamma|^{\frac{n}{2}}} \frac{|D'D|^{\frac{1}{2}}}{(2\pi\sigma^2\tau_{b_j})^{\frac{df}{2}}} exp\Big\{ -\frac{1}{2\sigma^2}\sum_{i=1}^{n} \bar{y}_{i_{-k}}'\Gamma^{-1}\bar{y}_{i_{-k}} \Big\}$$

$$\int_{\mathbb{R}} exp\Big\{ -\frac{1}{2}\Big[b_k' Z' \sum_{i=1}^{n} x_{i,k}\frac{\Gamma^{-1}}{\sigma^2}x_{i,k}Zb_k - b_k'Z'\sum_{i=1}^{n} x_{i,k}\frac{\Gamma^{-1}}{\sigma^2}\bar{y}_{i_{-k}} - \sum_{i=1}^{n}\bar{y}_{i_{-k}}'\frac{\Gamma^{-1}}{\sigma^2}x_{i,k}Zb_k + b_k'\frac{D'D}{\sigma^2\tau_{b_k}}b_k \Big]\Big\}\partial b_k$$

Let $\Sigma_{b_k} = \left(\frac{D'D}{\sigma^2\tau_{b_k}} + \frac{1}{\sigma^2}\sum_{i=1}^{n} x_{i,k}^2 Z'\Gamma^{-1}Z \right)^{-1}$.

$$P(y|\Theta_{k_1} \setminus \{b_k\}) = \frac{1}{(2\pi\sigma^2)^{\frac{nT}{2}}|\Gamma|^{\frac{n}{2}}} \frac{|D'D|^{\frac{1}{2}}}{(2\pi\sigma^2\tau_{b_j})^{\frac{df}{2}}} exp\Big\{ -\frac{1}{2\sigma^2}\sum_{i=1}^{n} \bar{y}_{i_{-k}}'\Gamma^{-1}\bar{y}_{i_{-k}} \Big\}$$
$$exp\Big\{ \frac{1}{2}\sum_{i=1}^{n}(\bar{y}_{i_{-k}}' x_{i,k})\frac{\Gamma^{-1}}{\sigma^2} Z\Sigma_{b_k} Z'\frac{\Gamma^{-1}}{\sigma^2}\sum_{i=1}^{n}(x_{i,k}\bar{y}_{i_{-k}}) \Big\}$$
$$\int_{\mathbb{R}} exp\Big\{ -\frac{1}{2}\Big[\Big(b_k - \Sigma_{b_k}Z'\frac{\Gamma^{-1}}{\sigma^2}\sum_{i=1}^{n}(x_{i,k}\bar{y}_{i_{-k}}) \Big)' \Sigma_{b_k} \Big(b_k - \Sigma_{b_k}Z'\frac{\Gamma^{-1}}{\sigma^2}\sum_{i=1}^{n}(x_{i,k}\bar{y}_{i_{-k}}) \Big) \Big]\Big\}\partial b_k$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{nT}{2}}|\Gamma|^{\frac{n}{2}}} \frac{|D'D|^{\frac{1}{2}}}{(2\pi\sigma^2\tau_{b_j})^{\frac{df}{2}}} exp\Big\{ -\frac{1}{2\sigma^2}\sum_{i=1}^{n} \bar{y}_{i_{-k}}'\Gamma^{-1}\bar{y}_{i_{-k}} \Big\}$$
$$exp\Big\{ \frac{1}{2}\sum_{i=1}^{n}(\bar{y}_{i_{-k}}' x_{i,k})\frac{\Gamma^{-1}}{\sigma^2} Z\Sigma_{b_k} Z'\frac{\Gamma^{-1}}{\sigma^2}\sum_{i=1}^{n}(x_{i,k}\bar{y}_{i_{-k}}) \Big\} (2\pi)^{\frac{df}{2}}|\Sigma_{b_k}|^{\frac{1}{2}}$$

$$P(\gamma_k = 1|\Theta \setminus \{b_k, \gamma_k\}) = \frac{P(y|\Theta_{k_1} \setminus \{b_k\})P(\gamma_k = 1)}{P(y|\Theta_{k_1} \setminus \{b_k\})P(\gamma_k = 1) + P(y|\Theta_{k_0} \setminus \{b_k\})P(\gamma_k = 0)}$$
$$= \frac{R}{1+R}$$

with

$$R = \frac{P(y|\Theta_{k_1} \setminus \{b_k\})P(\gamma_k = 1)}{P(y|\Theta_{k_0} \setminus \{b_k\})P(\gamma_k = 0)}$$

$$= \frac{\pi \frac{|D'D|^{\frac{1}{2}}(2\pi)^{\frac{df}{2}}|\Sigma_{b_k}|^{\frac{1}{2}}}{(2\pi\sigma^2)^{\frac{nT}{2}}|\Gamma|^{\frac{n}{2}}(2\pi\sigma^2\tau_{b_j})^{\frac{df}{2}}} exp\left\{ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\bar{y}'_{i_{-k}}\Gamma^{-1}\bar{y}_{i_{-k}} \right\} exp\left\{ \frac{1}{2}\sum_{i=1}^{n}(\bar{y}'_{i_{-k}}x_{i,k})\frac{\Gamma^{-1}}{\sigma^2}Z\Sigma_{b_k}Z'\frac{\Gamma^{-1}}{\sigma^2}\sum_{i=1}^{n}(x_{i,k}\bar{y}_{i_{-k}}) \right\}}{(1-\pi)\frac{1}{(2\pi\sigma^2)^{\frac{nT}{2}}|\Gamma|^{\frac{n}{2}}} exp\left\{ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}\bar{y}'_{i_{-k}}\Gamma^{-1}\bar{y}_{i_{-k}} \right\}}$$

$$= \frac{\pi}{1-\pi}|D'D|^{\frac{1}{2}}|\Sigma_{b_k}|^{\frac{1}{2}}\frac{1}{(\sigma^2\tau_{b_k})^{\frac{df}{2}}} exp\left\{ \frac{1}{2}\sum_{i=1}^{n}(\bar{y}'_{i_{-k}}x_{i,k})\frac{\Gamma^{-1}}{\sigma^2}Z\Sigma_{b_k}Z'\frac{\Gamma^{-1}}{\sigma^2}\sum_{i=1}^{n}(x_{i,k}\bar{y}_{i_{-k}}) \right\}$$

A.3. Full conditional distributions for group spike-and-slab prior

Let Θ the set of all parameters $\{\alpha, \widetilde{m}, \tau_m, \widetilde{a_1}, \ldots, \widetilde{a_L}, \tau_{a_1}, \ldots, \tau_{a_L}, b_1, \ldots, b_J, \gamma_1, \ldots, \gamma_J, \tau_{b_1}, \ldots, \tau_{b_J}, \pi, \rho, \sigma^2\}$ in the Bayesian hierarchical model (5), $\bar{y_i} = y_i - \alpha 1 - \widetilde{B^t}\widetilde{m} - \sum_{l=1}^{L} \widetilde{B}^{e^l}\widetilde{a}_l - \sum_{j=1}^{J} x_{i,j} Z b_j$ and $\bar{y}_{i_{-k}} = y_i - \alpha 1 - \widetilde{B^t}\widetilde{m} - \sum_{l=1}^{L} \widetilde{B}^{e^l}\widetilde{a}_l - \sum_{j=1;j\neq k}^{J} x_{i,j} Z b_j$.

$$\alpha|. \sim N_1\left(\Sigma_\alpha 1' \frac{\Gamma^{-1}}{\sigma^2} \sum_{i=1}^{n} (\bar{y_i} + \alpha 1), \Sigma_\alpha\right) \quad \text{with } \Sigma_\alpha = \left(n 1' \frac{\Gamma^{-1}}{\sigma^2} 1\right)^{-1}$$

$$\widetilde{m}|. \sim \mathcal{N}\left(\Sigma_{\widetilde{m}} \sum_{i=1}^{n} \widetilde{B^t}' \frac{\Gamma^{-1}}{\sigma^2} (\bar{y_i} + \widetilde{B^t}\widetilde{m}), \Sigma_{\widetilde{m}}\right) \quad \text{with}$$

$$\Sigma_{\widetilde{m}} = \left(\frac{\widetilde{D}'_m \widetilde{D}_m}{\tau_m} + \frac{n}{\sigma^2} \widetilde{B^t}' \Gamma^{-1} \widetilde{B^T}\right)^{-1}$$

$$\tau_m|. \sim \mathcal{IG}\left(\frac{df}{2} + 0.001, \frac{1}{2}\widetilde{m}' \widetilde{D}'_m \widetilde{D}_m \widetilde{m} + 0.001\right)$$

$$\widetilde{a_k}|. \sim \mathcal{N}\left(\Sigma_{\widetilde{a_k}} \sum_{i=1}^{n} \widetilde{B}^{e^{k'}} \frac{\Gamma^{-1}}{\sigma^2} (\bar{y_i} + \widetilde{B}^{e^k}\widetilde{a_k}), \Sigma_{\widetilde{a_k}}\right) \text{ with}$$

$$\Sigma_{\widetilde{a_k}} = \left(\frac{\widetilde{D}'_{a_k} \widetilde{D}_{a_k}}{\tau_{a_k}} + \frac{n}{\sigma^2} \widetilde{B}^{e^{k'}} \Gamma^{-1} \widetilde{B}^{e^k}\right)^{-1}, k = 1, \ldots, L$$

$$\tau_{a_k}|. \sim \mathcal{IG}\left(\frac{df}{2} + 0.001, \frac{1}{2}\widetilde{a_k}' \widetilde{D}'_{a_k} \widetilde{D}_{a_k} \widetilde{a_k} + 0.001\right), \quad k = 1, \ldots, L$$

$$b_k|. \sim \gamma_k \mathcal{N}\left(\Sigma_{b_k} \sum_{i=1}^{n} x_{i,j} B^{t'} \frac{\Gamma^{-1}}{\sigma^2} (\bar{y_i} + x_{i,k} Z b_k), \Sigma_{b_k}\right) + (1 - \gamma_k)\delta \text{ with}$$

$$\Sigma_{b_k} = \left(\frac{D'D}{\sigma^2 \tau_{b_k}} + \frac{1}{\sigma^2} \sum_{i=1}^{n} x_{i,k}^2 Z' \Gamma^{-1} Z\right)^{-1}, k = 1, \ldots, J$$

$$P(\gamma_k = 1 | \Theta \setminus \{b_k, \gamma_k\}) \sim \frac{R}{1+R} \qquad \text{with}$$

$$R = \frac{\pi}{1-\pi} |D'D|^{\frac{1}{2}} |\Sigma_{b_k}|^{\frac{1}{2}} \frac{1}{(\sigma^2 \tau_{b_k})^{\frac{df}{2}}} exp\left\{\frac{1}{2} \sum_{i=1}^{n} (\bar{y}'_{i_{-k}} x_{i,k}) \frac{\Gamma^{-1}}{\sigma^2} Z \Sigma_{b_k} Z' \frac{\Gamma^{-1}}{\sigma^2} \sum_{i=1}^{n} (x_{i,k} \bar{y}_{i_{-k}})\right\}$$

$$\tau_{b_k}|. \sim \mathcal{IG}\left(\frac{df}{2} + 0.001, \frac{1}{2\sigma^2} b'_k D' D b_k + 0.001\right), \quad k = 1, \ldots, J$$

$$\pi|. \sim \mathcal{B}eta(1 + |\gamma|, 1 + J - |\gamma|)$$

$$\rho|. \sim |\Gamma|^{-\frac{n}{2}} exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \bar{y_i}' \Gamma^{-1} \bar{y_i}\right\} \mathbb{1}_{(-1<\rho<1)}$$

$$\sigma^2|. \sim \mathcal{IG}\left(0.001 + \frac{1}{2} nT + \frac{1}{2} df \sum_{j=1}^{J} \gamma_j, 0.001 + \frac{1}{2} \sum_{j=1}^{J} b_j' D' D b_j \eta_j + \frac{1}{2} \sum_{i=1}^{n} \bar{y_i}' \Gamma^{-1} \bar{y_i}\right)$$

A.4. Bayesian group Lasso

A.4.1. Hierarchical model

$$
\begin{aligned}
y_i|\alpha,\widetilde{m},\widetilde{a},b,\rho,\sigma^2 &\sim \mathcal{N}(\alpha+\widetilde{B^t}\widetilde{m}+\sum_{l=1}^{L}\widetilde{B}^{e^l}\widetilde{a}_l+\sum_{j=1}^{J}x_{i,j}Zb_j,\sigma^2\Gamma)\\
\alpha &\sim \mathcal{U}_{(-\infty,\infty)}\\
\widetilde{m}|\tau_m &\sim \mathcal{N}(0,(\tau_m\widetilde{D}'_m\widetilde{D}_m)^{-1})\\
\widetilde{a}_l|\tau_{a_l} &\sim \mathcal{N}(0,(\tau_{a_l}\widetilde{D}'_{a_l}\widetilde{D}_{a_k})^{-1}),\quad l=1,\ldots,L\\
b_j|\eta_j,\sigma^2 &\sim \mathcal{N}(0,\sigma^2\tau_j^2(D'D)^{-1}),\quad j=1,\ldots,J\\
\tau_j^2|\lambda^2 &\sim \mathcal{G}\left(\frac{df+1}{2},\frac{\lambda^2}{2}\right),j=1,\ldots,J\\
\tau_m,\quad \tau_{a_l}\ \text{ and }\ \lambda^2 &\sim \mathcal{G}(0.001,0.001)\ \text{ and }\ l=1,\ldots,L\\
\rho &\sim \mathcal{U}_{(-1,1)}\ \text{ and }\ \sigma^2\sim\mathcal{IG}(0.001,0.001)
\end{aligned}
\tag{11}
$$

A.4.2. Full conditional distributions

Let Θ the set of all parameters $\{\alpha,\widetilde{m},\tau_m,\widetilde{a_1},\ldots,\widetilde{a_L},\tau_{a_1},\ldots,\tau_{a_L},b_1,\ldots,b_J,\tau_1^2,\ldots,\tau_J^2,\lambda,\rho,\sigma^2\}$ in the Bayesian hierarchical model (11) and $\bar{y_i}=y_i-\alpha 1-\widetilde{B^t}\widetilde{m}-\sum_{l=1}^{L}\widetilde{B}^{e^l}\widetilde{a}_l-\sum_{j=1}^{J}x_{i,j}Zb_j$

$$
\begin{aligned}
\alpha|. &\sim N_1\left(\Sigma_\alpha 1'\frac{\Gamma^{-1}}{\sigma^2}\sum_{i=1}^{n}(\bar{y_i}+\alpha 1),\Sigma_\alpha\right)\ \text{ with } \Sigma_\alpha=\left(n1'\frac{\Gamma^{-1}}{\sigma^2}1\right)^{-1}\\
\widetilde{m}|. &\sim \mathcal{N}\left(\Sigma_{\widetilde{m}}\sum_{i=1}^{n}\widetilde{B^t}'\frac{\Gamma^{-1}}{\sigma^2}(\bar{y_i}+\widetilde{B^t}\widetilde{m}),\Sigma_{\widetilde{m}}\right)\quad \text{with}\\
&\Sigma_{\widetilde{m}}=\left(\tau_m\widetilde{D}'_m\widetilde{D}_m+\frac{n}{\sigma^2}\widetilde{B^t}'\Gamma^{-1}\widetilde{B^T}\right)^{-1}\\
\tau_m|. &\sim \mathcal{G}\left(\frac{df}{2}+0.001,\frac{1}{2}\widetilde{m}'\widetilde{D}'_m\widetilde{D}_m\widetilde{m}+0.001\right)\\
\widetilde{a_k}|. &\sim \mathcal{N}\left(\Sigma_{\widetilde{a_k}}\sum_{i=1}^{n}\widetilde{B}^{e^{k'}}\frac{\Gamma^{-1}}{\sigma^2}(\bar{y_i}+\widetilde{B}^{e^k}\widetilde{a_k}),\Sigma_{\widetilde{a_k}}\right)\ \text{with}\\
&\Sigma_{\widetilde{a_k}}=\left(\tau_{a_k}\widetilde{D}'_{a_k}\widetilde{D}_{a_k}+\frac{n}{\sigma^2}\widetilde{B}^{e^{k'}}\Gamma^{-1}\widetilde{B}^{e^k}\right)^{-1},\quad k=1,\ldots,L\\
\tau_{a_k}|. &\sim \mathcal{G}\left(\frac{df}{2}+0.001,\frac{1}{2}\widetilde{a_k}'\widetilde{D}'_{a_k}\widetilde{D}_{a_k}\widetilde{a_k}+0.001\right),\quad k=1,\ldots,L
\end{aligned}
$$

779

$$
\begin{aligned}
b_k|. &\sim \mathcal{N}\left(\Sigma_{b_k}\sum_{i=1}^{n} x_{i,j}B^{t\prime}\frac{\Gamma^{-1}}{\sigma^2}(\bar{y_i}+x_{i,k}Zb_k),\Sigma_{b_k}\right) \text{ with}\\
&\quad \Sigma_{b_k}=\left(\frac{D'D}{\tau_k^2\sigma^2}+\frac{1}{\sigma^2}\sum_{i=1}^{n}x_{i,k}Z'\Gamma^{-1}Z\right)^{-1}, \quad k=1,\ldots,J\\
\frac{1}{\tau_k^2}|. &\sim \mathcal{I}-\mathcal{G}\text{aussian}\left(\sqrt{\frac{\sigma^2\lambda^2}{{b_k}'D'Db_k}},\lambda^2\right), \quad k=1,\ldots,J\\
\lambda^2|. &\sim \mathcal{G}\left(\frac{Jdf+J}{2}+0.001,\sum_{j=1}^{J}\frac{\tau_j^2}{2}+0.001\right)\\
\rho|. &\sim |\Gamma|^{-\frac{n}{2}}exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\bar{y_i}'\Gamma^{-1}\bar{y_i}\right\}\mathbb{1}_{(-1<\rho<1)}\\
\sigma^2|. &\sim \mathcal{IG}\left(0.001+\frac{1}{2}nT+\frac{1}{2}df\sum_{j=1}^{J}\gamma_j,0.001+\frac{1}{2}\sum_{j=1}^{J}{b_j}'D'Db_j\eta_j+\frac{1}{2}\sum_{i=1}^{n}\bar{y_i}'\Gamma^{-1}\bar{y_i}\right)
\end{aligned}
$$

780

Table 4. Computational time (in minutes) obtained using different priors.

Prior	n=300, J=500, σ^2=4	n=300, J=500, σ^2=16	n=100, J=3000, σ^2=4	n=100, J=3000, σ^2=16
BGL-PS BGL-BS BGL-L	8 (0.5)	8 (0.5)	67 (1)	66 (2)
GSS-L GSS-BS	8 (1)	8 (1)	60 (5)	60 (5)
GSS-PS_1 GSS-PS_2	16 (5)	16 (5)	120 (10)	120 (10)
GSS-RW_1 GSS-RW_2	282 (9)	281 (10)	1500 (150)	1500 (150)
S-GP	68 (13)	61 (9)	26 (6)	11 (4)