Sélection bayésienne de variables pour données longitudinales avec effets différentiels dans le temps : application à l'amélioration génétique

Benjamin Heuclin

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITE DE MONTPELLIER

En Biostatistique

École doctorale I2S – Information, Structures, Systèmes

Unité de recherche UMR 5149 – IMAG – Institut Montpelliérain Alexander Grothendieck

Sélection bayésienne de variables pour données longitudinales avec effets différentiels dans le temps : application à l'amélioration génétique.

Présentée par Benjamin Heuclin

Le 07 juillet 2021

Sous la direction de Catherine Trottier et Frédéric Mortier et co-encadrée par Marie Denis

Devant le jury composé de

Marie Denis	Chargée de recherche	Cirad	Co-encadrante
Frédéric Gosselin	Ingénieur	INRAE	Examinateur
Estelle Kuhn	Directrice de recherche	INRAE	Rapportrice
Benoit Liquet	Professeur	Université de Pau et des Pays de l'Adour	Rapporteur
Jean-Michel Marin	Professeur	Université de Montpellier	Président
Frédéric Mortier	Chargé de recherche	Cirad	Directeur
Leopoldo Sanchez-Rodriguez	Directeur de recherche	INRAE	Examinateur
Catherine Trottier	Maître de Conférences	Université Paul Valéry - Montpellier 3	Directrice

UNIVERSITÉ DE MONTPELLIER

Table des matières

1 Introduction **1**

1.1 Contexte agronomique et génétique 1

1.2 Analyse de données longitudinales 4

1.2.1 Les modèles linéaires mixtes 4

1.2.2 Les modèles à coefficients variants 6

1.3 Sélection bayésienne de variables et régularisation de modèle pour données longitudinales 8

1.3.1 Lois *a priori* pour la sélection de variables et la régularisation de modèles : cas où les paramètres sont non structurés 9

1.3.2 Lois *a priori* pour la sélection de variables et la régularisation de modèles : cas où les paramètres sont structurés 14

1.3.3 Priors de sélection de composantes de la variance et régularisation de modèles linéaires mixtes 18

1.4 Plan de thèse 19

2 Sélection de composantes de la variance dans un modèle linéaire mixte **21**

2.1 Introduction 21

2.2 Article : Continuous shrinkage priors for fixed and random effects selection in linear mixed models: application to genetic mapping 22

3 Sélection de groupe de paramètres ordonnés dans un modèle à coefficients variants **49**

3.1 Introduction 49

3.2 Article : Bayesian varying coefficient model with selection: An application to functional mapping 50

4 Double sélection bayésienne de groupes de variables et de variables en présence d'une forte corrélation entre les variables au sein de chaque groupe 85

4.1 Introduction 85

4.2 Article : Bayesian sparse group selection with indexed regressors within groups: the group fused horseshoe prior 87

4.3 Application à l'identification de positions génétiques influant sur un caractère phénotypique 109

5 Conclusions et perspectives 113

5.1 Contributions par rapport aux objectifs initiaux 114

5.2 Perspectives 116

I

Introduction

Sommaire

1.1 Contexte agronomique et génétique 1

1.2 Analyse de données longitudinales 4

1.2.1 Les modèles linéaires mixtes 4

1.2.2 Les modèles à coefficients variants 6

1.3 Sélection bayésienne de variables et régularisation de modèle pour données longitudinales 8

1.3.1 Lois *a priori* pour la sélection de variables et la régularisation de modèles : cas où les paramètres sont non structurés 9

1.3.2 Lois *a priori* pour la sélection de variables et la régularisation de modèles : cas où les paramètres sont structurés 14

1.3.3 Priors de sélection de composantes de la variance et régularisation de modèles linéaires mixtes 18

1.4 Plan de thèse 19

Ce travail de thèse vise à développer des outils de statistique bayésienne pour l'analyse de données longitudinales dans le domaine de l'agronomie et de l'amélioration génétique. Il sera limité à des données ne contenant pas de données manquantes ainsi qu'à des variables réponses univariées continues.

1.1 Contexte agronomique et génétique

Dans le domaine de l'agronomie, le besoin de produire en grande quantité, tout en garantissant une qualité nutritive et de façon durable face aux changements climatiques, conduit à optimiser les modes de production. Cela passe notamment par le développement de nouvelles variétés plus performantes en cherchant à maximiser à la

fois le rendement, la qualité et la résistance face aux maladies ainsi qu'à minimiser les besoins en eau. Dans ce contexte, les programmes d'amélioration génétique ont pour objectif de sélectionner les meilleurs individus (génotypes) d'une population pour engendrer les générations suivantes. Le succès de ces programmes d'amélioration provient de leur capacité à optimiser un ou plusieurs caractères d'intérêt, en se basant sur des dispositifs expérimentaux et sur l'utilisation de modèles statistiques. Dans le cadre de l'amélioration des espèces pérennes, telles que l'eucalyptus ou le palmier à huile, les schémas de sélection récurrente réciproque sont largement utilisés. Le principe de ces schémas est d'améliorer conjointement deux groupes d'individus, de manière à obtenir des hybrides combinant de façon optimale les caractéristiques des deux groupes parentaux. L'analyse de ces dispositifs a pour objectif principal d'extraire la part génétique de la variation d'un phénotype d'intérêt et de l'isoler de la part des variations environnementales en utilisant les modèles linéaires à effets aléatoires (Linear Mixed Model, LMM) [Lynch et al., 1998].

Dans le cadre de la génétique quantitative avec des plans de croisements, deux modèles sont principalement utilisés selon l'information génétique disponible, le modèle "père-mère" et le modèle "animal" [Sun et al., 2009, Mrode, 2014]. Le modèle "père-mère" permet d'expliquer la variation d'un caractère d'intérêt avec une décomposition de la variance génétique individuelle comme la somme des variances "mère" et "père". Cette décomposition permet d'estimer la valeur génétique individuelle comme la moyenne des variances génétiques héritées de la mère et du père. Le modèle linéaire mixte associé est le suivant :

$$y = X\beta + Z_m M + Z_p P + \varepsilon, \tag{1.1}$$

avec y le vecteur des observations du caractère d'intérêt, X une matrice de variables explicatives, β le vecteur des effets fixes, Z_m et Z_p des matrices de design associées aux effets M et P respectivement. M et P sont considérés comme deux vecteurs à effet aléatoire car ils peuvent être vus comme le résultat d'un échantillonnage aléatoire dans une population parentale plus large. M est l'effet mère supposé de loi normale $N(0, \sigma_m^2 Id)$. P est l'effet père supposé de loi normale $N(0, \sigma_p^2 Id)$. Cette modélisation considère que toutes les mères sont indépendantes, tous les pères sont indépendants et que les mères sont indépendantes des pères. Enfin, ε est le vecteur des résidus supposé distribué selon une loi normale $\mathcal{N}(0, \sigma^2 Id)$.

Dans le contexte de plans d'expériences avec une information pedigree sur plusieurs générations, ce type de modèle n'est pas adapté. Prenons l'exemple de deux individus cousins. Avec le modèle (1.1), la corrélation entre ces deux individus sera nulle car les parents sont supposés tous indépendants entre eux. Or cette corrélation n'est pas nulle. Pour pallier ce problème, les "modèles animaux" ont été développés, initialement pour des études sur les bovins [Lynch et al., 1998, Wilson et al., 2010]. Ils permettent de tenir compte du pedigree en considérant une structure de dépendance génétique entre les individus, aussi appelée structure d'apparentement. Le modèle "animal" est de la forme :

$$y = X\beta + Z_A u_A + \varepsilon, \tag{1.2}$$

avec u_A l'effet aléatoire contenant un niveau pour chaque individu, il suit une loi normale $N(0, \sigma_a^2 A)$ où A est la structure d'apparentement connue. Les structures d'apparentement les plus couramment utilisées sont les structures calculées à partir de la connaissance du pedigree [Mrode, 2014].

Avec l'arrivée des outils de génomique et de génotypage haut débit, les enjeux ont évolué. Il est alors possible d'utiliser l'information moléculaire pour assister la sélection et ainsi accélérer les programmes d'amélioration en identifiant les régions du génome impliquées dans la variation phénotypique du caractère d'intérêt. Toutefois, l'utilisation de données génotypiques soulève des questions méthodologiques nouvelles, en particulier en lien avec la sélection d'effets fixes et d'effets aléatoires.

En effet, depuis quelques années, l'acquisition d'un grand nombre de marqueurs moléculaires a donné accès à une information sur l'ensemble du génome. Ces marqueurs sont intégrés dans les modèles statistiques de génétique quantitative comme des variables à effet fixe. Cependant, la prise en compte des marqueurs sur l'ensemble du génome mène à un sur-ajustement. Effectivement, le nombre de variables devient bien souvent largement supérieur au nombre d'observations. De plus, seules quelques régions du génome apportent de l'information sur le caractère d'intérêt. Une question se pose donc : sélectionner les marqueurs moléculaires qui influencent la variation d'un caractère phénotypique en présence d'information sur le pedigree. Autrement dit, comment faire de la sélection d'effets fixes dans le contexte de modèles à effets aléatoires ?

Des méthodes développées en génétique humaine et plus récemment en génétique des plantes ont permis, à partir de marqueurs moléculaires et/ou de l'information sur le pedigree, d'obtenir une matrice d'apparentement (IBD pour *Identity By Descent*) plus fine soit à l'échelle du génome, soit à chaque position [George et al., 2000, van Eeuwijk et al., 2010, Tisné et al., 2015, Lu et al., 2015, Korontzis et al., 2020]. Ainsi il est possible de décomposer l'effet individuel global comme la somme d'effets individuels en chaque position. Ces structures associées à différentes positions entraînent un grand nombre d'effets aléatoires qui est souvent supérieur au nombre d'observations. Une nouvelle question se pose alors : comment sélectionner les structures qui influencent la variation d'un caractère phénotypique ? Autrement dit, comment sélectionner des effets aléatoires ?

Parallèlement, des technologies de phénotypage haut débit ont fait leur apparition, permettant d'avoir accès à des mesures répétées dans le temps (données longitudinales). Il est alors possible de suivre plusieurs caractères phénotypiques au cours du temps, conjointement à différentes variables environnementales (température, ensoleillement, humidité, pluviométrie, ...) en milieu naturel [Bartholomé et al., 2020]. Mais on peut aussi contrôler complètement l'environnement en serre permettant ainsi de simuler des scénarios environnementaux [Marchadier et al., 2019]. De telles expérimentations ont pour but d'obtenir une compréhension plus fine du comportement d'un ou plusieurs caractères phénotypiques d'intérêt face aux variations environne-

mentales (contrôlées ou non) au travers de modèles statistiques évolués. Il est alors possible d'étudier l'évolution de l'architecture génétique au cours du temps ou encore de mieux comprendre les interactions Génotype $\times$ Environnement (cf chapitres 2 et 3). On peut aussi chercher à étudier comment certains facteurs environnementaux ou écophysiologiques agissent au cours du temps sur certains processus biologiques pour ensuite mieux comprendre comment les cultures s'adaptent face au changement des conditions environnementales. L'analyse statistique de données longitudinales soulève toutefois plusieurs enjeux en lien avec : la prise en compte de la dépendance temporelle entre les observations et les variables, la sélection de variables, et l'estimation d'effets évoluant au cours du temps.

La statistique bayésienne offre un cadre unifié pour répondre à ces différentes problématiques grâce à la flexibilité offerte par la construction de modèles hiérarchiques. Il est alors possible de construire des priors permettant de sélectionner les variables pertinentes tout en tenant compte de diverses structurations des données. C'est l'objectif de ce travail de thèse.

1.2 Analyse de données longitudinales

Comprendre la dynamique de processus d'intérêt constitue un enjeu majeur dans de nombreux domaines. Quels facteurs influencent la croissance des arbres par exemple. Comment l'influencent-ils ? Ont ils un impact restreint à une période de temps, ou au contraire un impact sur l'ensemble de la croissance ? L'acquisition de données mesurées au cours du temps, données longitudinales, permet une compréhension fine de ces processus dynamiques. Toutefois, l'analyse jointe de ces données soulève de nombreuses questions méthodologiques telles que la modélisation de la dépendance entre les observations, ou la modélisation des effets des variables explicatives au cours du temps. Dans cette partie, nous introduisons les modèles les plus couramment utilisés en nous focalisant sur le cas où la variable réponse est continue, avec l'objectif de comprendre l'influence d'un grand nombre de variables explicatives sur la dynamique de la réponse. Nous supposerons de plus que la variable réponse est observée sur différents pas de temps communs à tous les individus. Nous noterons l'indice des individus $i = 1, \ldots, n$, les variables explicatives seront indexées par $j = 1, \ldots, p$ et enfin l'indice des pas de temps sera noté $t = t_1, \ldots, t_T$.

1.2.1 Les modèles linéaires mixtes

Les modèles linéaires mixtes forment une classe de modèles qui permet de structurer les dépendances selon un ou plusieurs facteurs groupant. Dans le contexte de données longitudinales, ces modèles permettent une modélisation de la structure de dépendance entre les observations réalisées sur un même individu. En particulier, les modèles à intercept et pentes aléatoires (RIS pour *random intercept and slopes*), introduits par Laird and Ware [1982], modélisent l'intercept et les pentes en fonction d'un

facteur groupant. Dans la majorité des applications, l'individu est supposé être le facteur groupant. Nous verrons par la suite un autre exemple où le facteur groupant est le temps (cf chapitre 2). Le modèle RIS peut s'exprimer au niveau de l'individu et du temps de la façon suivante :

$$y_{i,t} = \sum_{j=0}^{p} x_{j_{i,t}} \beta_j + \sum_{j=0}^{p} x_{j_{i,t}} \tilde{\beta}_{i_j} + \varepsilon_{i,t}, \tag{1.3}$$

où $y_{i,t}$ est l'observation de l'individu i au temps t, $x_{j_{i,t}}$ est la valeur de la $j^{\text{ème}}$ variable ($j = 0, \ldots, p$) pour l'individu i au temps t avec $x_{0_{i,t}} = 1$ pour l'intercept, β_j le coefficient de régression associé à la $j^{\text{ème}}$ variable avec β_0 l'intercept. $\tilde{\beta}_{i_j}$ est le terme correcteur du coefficient de régression de la $j^{\text{ème}}$ variable pour l'individu i, supposé aléatoire. Ce terme est également nommé pente aléatoire. Le vecteur $\tilde{\beta}_i = (\tilde{\beta}_{i_0}, \ldots, \tilde{\beta}_{i_p})'$ de l'ensemble des intercept et pentes aléatoires de l'individu i est supposé suivre une loi normale d'espérance nulle et de matrice de covariance V inconnue. $\varepsilon_{i,t}$ est le résidu de l'individu i au temps t. Le vecteur $\varepsilon_i = (\varepsilon_{i,1}, \ldots, \varepsilon_{i,t_T})'$ de l'ensemble des résidus de l'individu i est supposé suivre une loi normale d'espérance nulle et de matrice de covariance Σ.

D'un point de vue matriciel (cf figure 1.1 pour plus de lisibilité), le modèle peut ainsi se réécrire de la façon suivante :

$$y = X\beta + Z\tilde{\beta} + \varepsilon, \tag{1.4}$$

avec y le vecteur de l'ensemble des observations $y = (y_1', y_2', \ldots, y_n')'$, avec $y_i = (y_{i,t_1}, \ldots, y_{i,t_T})'$ le vecteur de longueur t_T des observations pour l'individu i. X la matrice de dimension $(nt_T) \times (p+1)$ contenant les p régresseurs plus le vecteur unitaire : $X = [1, X_1, \ldots, X_p]$ avec $X_j = (X_{j_{1,t_1}}, \ldots, X_{j_{1,t_T}}, \ldots, X_{j_{n,t_1}}, \ldots, X_{j_{n,t_T}})'$ pour $j = 1, \ldots, p$. $\beta = (\beta_0, \ldots, \beta_p)$ est le vecteur de longueur $p+1$ des coefficients de régression. Z est la matrice de design (bloc diagonale) de dimension $(nt_T) \times (n(p+1))$ associée aux pentes et intercept aléatoires et construite à partir de la matrice X. $\tilde{\beta}$ est le vecteur de longueur $n(p+1)$ des pentes et intercept aléatoires. Ce vecteur est la concaténation des vecteurs aléatoires propres à chaque individu $\tilde{\beta} = (\tilde{\beta}_1', \ldots, \tilde{\beta}_n')'$. Ainsi $\tilde{\beta}$ est supposé suivre une loi normale $\mathcal{N}(0, I_n \otimes V)$, où $\otimes$ désigne le produit de Kronecker. Enfin $\varepsilon = (\varepsilon_1', \ldots, \varepsilon_n')'$ est le vecteur de longueur nt_T des résidus supposé suivre une loi normale $\mathcal{N}(0, I_n \otimes \Sigma)$.

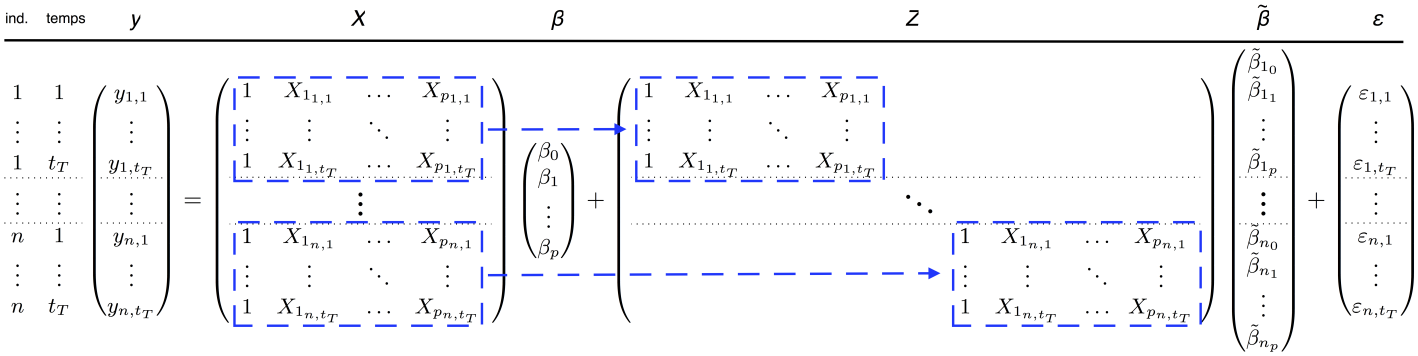


FIGURE 1.1 – *Construction de l'équation (1.4) du modèle à intercept et pentes aléatoires.*

Les structures de dépendance entre les observations sont construites à deux niveaux par les matrices de covariance V et Σ. V permet de structurer les dépendances

entre les effets aléatoires et ainsi de supposer une dépendance entre l'intercept et les pentes. Celle-ci peut être supposée totalement libre ou prendre une forme particulière. Par exemple, dans le contexte de l'analyse des données *arabidospsis thaliana* que nous aborderons dans le chapitre 2, nous supposons que V est structurée en bloc où chaque bloc est associé à un chromosome. Pour ce qui concerne la matrice Σ, elle permet de tenir compte des dépendances résiduelles. Il est commun de supposer les résidus indépendants et cette matrice est alors égale à l'identité. Toutefois, Li and Sillanpää [2013] montrent que celle-ci peut jouer un rôle très important notamment lors de l'étape de sélection des effets fixes. Par la suite, nous considérerons une structure de type autoregressive comme l'ont proposé différents auteurs précédemment [Ma et al., 2002, Fahrmeir et al., 2011, Li and Sillanpää, 2013]. Ce choix sera discuté.

1.2.2 Les modèles à coefficients variants

Les modèles à coefficients variants (VCM pour *Varying Coefficient Model*), introduits par Hastie and Tibshirani [1993], forment une large classe de modèles non-linéaires permettant une modélisation flexible des données. Un VCM consiste à décrire les observations en fonction de deux ensembles de régresseurs $X_1, \dots, X_p$ et $S_1, \dots, S_p$. Il est donné par l'équation générale suivante :

$$y = \mu + X_1.f_1(S_1) + \cdots + X_p.f_p(S_p) + \varepsilon, \tag{1.5}$$

avec $y = (y_{1,t_1}, \dots, y_{1,t_T}, \dots, y_{n,t_1}, \dots, y_{n,t_T})'$, le vecteur de l'ensemble des observations, μ l'intercept, $X_j = (X_{j_{1,t_1}}, \dots, X_{j_{1,t_T}}, \dots, X_{j_{n,t_1}}, \dots, X_{j_{n,t_T}})'$ et
$S_j = (S_{j_{1,t_1}}, \dots, S_{j_{1,t_T}}, \dots, S_{j_{n,t_1}}, \dots, S_{j_{n,t_T}})'$, pour $j = 1, \dots, p$, les vecteurs de longueur nt_T des régresseurs. $f_1, \dots, f_p$ sont des fonctions inconnues supposées continues. L'opérateur "." désigne le produit élément par élément. ε est le vecteur de longueur nt_T des résidus. Il est la concaténation des résidus associés à chaque individu $\varepsilon = (\varepsilon_1', \dots, \varepsilon_n')'$, où $\varepsilon_i = (\varepsilon_{i,t_1}, \dots, \varepsilon_{i,t_T})'$ est un vecteur de longueur t_T supposé suivre une loi normale $\mathcal{N}(0, \Sigma)$, pour $i = 1, \dots, n$. Σ est une matrice de covariance de dimension $t_T \times t_T$ commune à tous les individus. Le vecteur de l'ensemble des résidus est alors supposé suivre une loi normale $\mathcal{N}(0, I_n \otimes \Sigma)$. Un VCM est linéaire par rapport aux régresseurs $X_1, \dots, X_p$ mais leurs effets évoluent de manière non linéaire par rapport aux valeurs des variables $S_1, \dots, S_p$.
La classe des VCMs englobe de nombreux modèles.

- Fixer les variables $S_1, \dots, S_p$ en un unique vecteur correspondant aux indices du temps dans l'équation (1.5) permet de retrouver un **modèle linéaire dynamique** introduit par West et al. [1985]. Ce modèle permet de modéliser les paramètres de régression par des fonctions du temps. Il a été introduit initialement pour l'analyse d'une variable réponse mesurée sur un unique individu à différents pas de temps. Ce modèle a été largement utilisé pour l'analyse de séries temporelles et a motivé différentes extensions, parmi lesquelles on peut citer par exemple les modèles linéaires à tendance (*dynamic linear trend model)* [Hodrick and Prescott, 1997, Kim et al., 2009, Frühwirth-Schnatter and Wagner, 2010, Faulkner and Minin, 2018] largement utilisés pour extraire la tendance d'une série temporelle ou débruiter/lisser un signal.

- Fixer les variables $X_1, \ldots, X_p$ à l'identité dans l'équation (1.5) conduit au **modèle additif** introduit par Hastie and Tibshirani [1990]. Cette classe de modèles permet de considérer une relation non linéaire entre les variables explicatives $S_1, \ldots, S_p$ et la variable réponse y.

- Fixer toutes les fonctions f_j en une unique fonction f et fixer $S_j = j$ permet de retrouver un **modèle de régression sur signaux** dans lequel nous ferions l'hypothèse que les coefficients de régression sont issus de réalisation d'une fonction inconnue plus ou moins lisse [Land and Friedman, 1996, Marx and Eilers, 1999, Tibshirani et al., 2005]. Les modèles de régression sur signaux sont adaptés pour étudier une variable réponse y disposant d'une seule observation par individu que l'on cherche à expliquer en fonction d'un ensemble de régresseurs ordonnés issus de mesures répétées sur t_T pas de temps. Le nombre de variables p correspond alors au nombre de pas de temps t_T. L'ensemble des observations des p/t_T régresseurs d'un même individu forme alors un signal évoluant dans le temps. Ces signaux apparaissent dans de nombreux domaines. En agronomie par exemple, on pourrait être intéressé par mettre en lien un indice de maturité d'un fruit (variable réponse y mesurée une seule fois sur n individus) avec une variable environnementale. Pour chaque individu, on dispose d'une série de p/t_T mesures journalières de la variable environnementale pouvant être vues comme un signal évoluant dans le temps. Cette application sera traitée dans le chapitre 4.

Estimation des effets fonctionnels

L'une des difficultés majeures dans les VCMs est l'estimation des effets fonctionnels. Les fonctions inconnues f_j, $j = 1, \ldots, p$, peuvent être estimées de manière paramétriques ou non-paramétriques [Li and Sillanpää, 2015]. Cette dernière approche permet de ne faire aucune hypothèse sur la forme des courbes. Parmi les approches non-paramétriques, deux sous-classes ont alors émergé dans la littérature, les approches fonctionnelles et non-fonctionnelles.

Les approches fonctionnelles consistent à interpoler une fonction inconnue f_j par une combinaison linéaire de q fonctions de base rangées dans une matrice B_j de dimension $nt_T \times q$ tel que $f_j(S_j) = B_j b_j$ avec b_j le vecteur de poids associés à chacune des fonctions de base. La matrice des fonctions de base B_j est construite à partir de la variable S_j. Chaque ligne l de la matrice B_j est définie par les réalisations des fonctions de base au point S_{j_l} et chaque colonne k est définie par les réalisations de la $k^{\text{ème}}$ fonction de base sur les différentes valeurs du régresseur S_j. Le nombre de fonctions de base étant inférieur au nombre de valeur du régresseur S_j, les approches fonctionnelles permettent ainsi de réduire le nombre de paramètres à estimer. Parmi les fonctions de base les plus couramment rencontrées, on retrouve la base des polynômes de Legendre ou encore les bases splines. Les splines consistent à interpoler des morceaux de polynômes d'ordre 1, 2 ou 3 entre différentes valeurs du régresseur S_j, appelées des noeuds, fixées par l'utilisateur. Les polynômes sont reliés entre eux au niveau des noeuds en imposant une continuité. Cette construction aboutit à un ensemble de m fonctions de

base spline avec m égal au nombre de noeuds plus le degré des polynômes. Différentes fonctions splines existent, on retrouve les fonctions de puissances tronquées, les splines naturelles [Eubank, 1988] ainsi que les B-splines [De Boor et al., 1978, Wahba, 1990, Dierckx, 1995]. Notons que les B-splines sont des fonctions locales permettant ainsi d'éviter les problèmes d'estimation qui peuvent survenir sur les bords du domaine de définition, faisant d'elles les fonctions de base les plus populaires. On pourra se référer au livre de Fitzmaurice et al. [2008], chapitre 11, pour plus de détails sur les différentes bases de splines. Le choix du nombre de noeuds ainsi que leurs positions sont des considérations délicates et ont une forte incidence sur la qualité d'ajustement. Considérer trop peu de noeuds mène à un sous-ajustement, tandis qu'en considérer un trop grand nombre de noeuds mène à un sur-ajustement. Pour palier ce problème, il est alors possible de considérer un nombre relativement large de noeuds équidistants et de contraindre la dérivée (généralement d'ordre deux) de la fonction ainsi ajustée (*smoothing spline* Wahba [1990]), ou encore de contraindre les différences successives des poids b_j (P-spline, Eilers and Marx [1996]). Cette dernière contrainte peut se voir comme une estimation de la dérivée au travers des différences finies et peut être réalisée, d'un point de vue bayésien, à l'aide d'un processus de marche aléatoire placé sur les poids [Lang and Brezger, 2004, Rue and Held, 2005]. Plus de détails sont apportés dans la section 1.3.2.

Les approches non-fonctionnelles permettent d'estimer directement les réalisations de la fonction f_j inconnue supposée continue sur les différentes valeurs du régresseur $f_j(S_j) = b_j$. Notons que cette approche peut être réécrite comme une approche fonctionnelle $f_j(S_j) = B_j b_j$ dans laquelle nous considérons une matrice B_j identité. Le nombre de paramètres à estimer est alors égal à la longueur du régresseur S_j. L'hypothèse de continuité de la fonction f_j implique la présence de lien entre les paramètres successifs contenus dans le vecteur b_j, et doit être prise en compte dans l'estimation de ce dernier. Cette estimation peut se faire de diverses manières, notamment au moyen d'un processus autorégressif ou plus largement à l'aide d'un processus gaussien [Rue and Held, 2005, Rasmussen and Williams, 2006]. Plus de détails sont apportés dans la section 1.3.2.

1.3 Sélection bayésienne de variables et régularisation de modèle pour données longitudinales

L'utilisation des modèles présentés précédemment (section 1.2) pour l'analyse de données longitudinales soulève des questions statistiques notamment en lien avec le sur-ajustement lorsque le nombre de variables explicatives est grand. Par exemple, dans le cadre des VCMs (cf équation (1.5)), un grand nombre de variables conduit à l'estimation d'un grand nombre de fonctions f_j et donc à l'estimation d'un grand nombre de vecteurs de paramètre b_j, $j = 1, \ldots, p$, que cela soit dans le contexte fonctionnel ou non-fonctionnel. Dans le cadre des modèles RIS (équation (1.4)), si le nombre

de covariables p est grand, alors le nombre d'effets aléatoires peut potentiellement être du même ordre de grandeur, voire égal à p. Cela implique que le nombre de composantes de la variance est de l'ordre de p, auquel s'ajoute les covariances entre les effets. Ainsi quelque soit le modèle considéré, le nombre de paramètres peut devenir très important voire même supérieur au nombre d'observations.

Dans le contexte bayésien, la loi *a priori*, ou prior, joue un rôle central et permet naturellement une certaine forme de régularisation [Bickel et al., 2006]. Toutefois, le choix de cette loi reste complexe en particulier lorsque le nombre de paramètres (coefficients de régression ou composantes de la variance) est grand, ou que l'on souhaite sélectionner les variables influentes, ou bien encore quand on cherche à forcer les effets vers zéro des variables sans intérêt et cela sans biaiser l'estimation des effets des variables importantes. Dans la suite de cette partie, nous introduisons dans un premier temps les lois *a priori* classiquement usitées pour un ensemble de paramètres non structurés, supposés indépendants entre eux. Dans un deuxième temps, nous présentons les lois *a priori* lorsque les paramètres sont structurés les uns avec les autres, par exemple au travers du temps. Enfin, la troisième et dernière partie montre comment ces lois peuvent être transposées à la question de la sélection des composantes de la variance.

1.3.1 Lois *a priori* pour la sélection de variables et la régularisation de modèles : cas où les paramètres sont non structurés

Dans cette première partie, nous nous positionnons dans le cadre de la régression linéaire :

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 Id), \tag{1.6}$$

où y est le vecteur de longueur n des observations (une seule observation par individu), X une matrice de dimension $n \times p$ contenant la collection des p variables explicatives, β le vecteur de longueur p des coefficients de régression et ε le vecteur de longueur n des résidus supposé gaussien d'espérance nulle et de variance σ^2.

Dans ce contexte, une large littérature existe concernant le choix des lois *a priori* pour sélectionner les variables d'intérêt ou pour réduire à zéro les effets des variables explicatives non pertinentes. La première famille de lois repose sur ce qui est couramment appelé les méthodes *spike-and-slab*. Cette classe de lois *a priori* consiste à supposer que la loi *a priori* des paramètres est un mélange fini de deux distributions, l'une pour la partie *spike*, l'autre pour la partie *slab*. La seconde classe de prior, plus récente, repose sur l'hypothèse que les paramètres suivent une loi continue qui peut se réécrire comme un mélange infini de lois gaussiennes sur la variance ou l'écart-type. Nous désignerons ces lois sous le nom de lois *a priori continues de réduction*.

Priors de type *spike-and-slab*

Les priors de type *spike-and-slab* ont été introduits par Mitchell and Beauchamp [1988]. Ils permettent de classer les paramètres β_j sujets à la sélection en deux catégories : les paramètres non-nuls et nuls. Pour chaque paramètre β_j, une variable latente

binaire γ_j est introduite qui va prendre la valeur 1 si le paramètre est considéré comme non-nul et zéro sinon. En fonction de la valeur prise par γ_j, le paramètre β_j se voit alors assigner une distribution *a priori* diffuse p_{slab} si $\gamma_j = 1$, ou une distribution *a priori* piquée autour de zéro p_{spike} si $\gamma_j = 0$. La loi p_{slab}, centrée en zéro doit avoir une variance τ_1^2 suffisamment large pour permettre aux paramètres β_j de prendre n'importe quelle valeur sur l'axe des réels, tandis que la loi p_{spike}, centrée en zéro est supposée avoir une faible variance τ_0^2 ce qui permet aux β_j de prendre une très faible valeur qui pourra être considérée comme nulle. La variable indicatrice est supposée suivre une distribution de Bernoulli dont le paramètre π peut-être considéré fixe ou aléatoire, le second choix est généralement préféré [Scott and Berger, 2010]. Les variances τ_0^2 et τ_1^2 peuvent également être inférées. Formellement, les lois *a priori* de type *spike-and-slab* peuvent s'écrire de la façon suivante :

$$p(\beta_j|\gamma_j, \tau_1^2, \tau_0^2) = \gamma_j p_{slab}(\beta_j|\tau_1^2) + (1-\gamma_j)p_{spike}(\beta_j|\tau_0^2), \tag{1.7}$$

$$\gamma_j|\pi \sim \mathcal{B}ern(\pi). \tag{1.8}$$

L'utilisation d'un tel prior permet alors d'avoir accès à la probabilité *a posteriori* que le paramètre β_j soit nul ou non et donc à la probabilité d'inclusion de la variable explicative X_j dans le modèle.

Différentes distributions p_{slab} et p_{spike} ont été étudiées dans la littérature. Généralement, la distribution p_{slab} est supposée être une distribution gaussienne. Ce cas particulier est connu sous le non de prior *stochastic search variable selection* [George and McCulloch, 1993, 1997]. La distribution Laplace a également été proposée comme alternative à la distribution gaussienne pour la partie diffuse, p_{slab}, [Ročková and George, 2018]. La distribution p_{spike} peut être identique à la distribution p_{slab} avec une variance τ_0^2 suffisamment petite. Notons qu'une telle distribution p_{spike} ne permet pas de réduire exactement à zéro le paramètre mais offre de très bonnes propriétés algorithmiques [Malsiner-Walli and Wagner, 2018]. L'utilisation d'une masse de Dirac en zéro comme distribution p_{spike} a également été proposée [Smith and Kohn, 1996, Geweke, 1996]. Cette distribution permet de mettre exactement à zéro les paramètres associés aux variables explicatives non pertinentes. Toutefois, cela peut entraîner des problèmes de convergence des algorithmes MCMC. L'utilisation de la vraisemblance intégrée en β_j pour échantillonner γ_j permet d'améliorer la convergence [Malsiner-Walli and Wagner, 2018, van de Schoot et al., 2021].

Bien que l'équation (1.7) reste la forme la plus rencontrée du prior *spike-and-slab*, d'autres paramétrisations ont également été explorées. Kuo and Mallick [1998] proposent par exemple la reparamétrisation $\beta_j = \gamma_j \alpha_j$, où les γ_j sont supposés suivre une loi de Bernoulli de paramètre π et les α_j une loi normale $\mathcal{N}(0, \tau_1^2)$. Cette reparamétrisation implique naturellement une masse de Dirac comme loi p_{spike}. George and McCulloch [1993] reformulent le prior (1.7) comme un mélange d'échelle de loi normale de la forme :

$$\beta_j|\gamma_j, \tau_1^2 \sim \mathcal{N}(0, \gamma_j\tau_1^2), \quad \gamma_j|\pi \sim \pi\delta_1(\gamma_j) + (1-\pi)\delta_{c_0}(\gamma_j). \tag{1.9}$$

Dans le cas où c_0 est fixé à zéro, on retrouve alors une masse de Dirac en zéro pour la loi p_{spike}. Dans le cas où c_0 est faible mais non nul, la loi p_{spike} est alors une loi normale

avec une variance $\tau_0^2 = c_0\tau_1^2$. Ishwaran et al. [2005] proposent une loi inverse-Gamma sur τ_1^2 aboutissant à une variance $v_j^2 = \gamma_j\tau_1^2$ suivant un mélange bimodal continu de loi inverse-Gamma.

Reconnus pour leurs performances, les priors de type *spike-and-slab* ont été largement utilisés avec succès dans différents domaines, notamment en génétique quantitative pour aborder des questions liées à la cartographie de loci de caractères quantitatifs (*QTL mapping*) ou à la génétique d'association pangénomique *GWAS* [Pérez and de Los Campos, 2014, Lu et al., 2015]

Priors *continus de réduction*

Les priors de type *continu de réduction* (CS pour *Continuous Shrinkage priors*) sont une alternative aux priors de type *spike-and-slab* [Park and Casella, 2008, Polson and Scott, 2010]. Ils ont initialement été développés pour obtenir une version bayésienne des méthodes de vraisemblance pénalisée introduites dans le cadre fréquentiste [Kyung et al., 2010]. Ces dernières consistent à minimiser la fonction suivante :

$$\arg\min_\beta l(y,\beta) = \frac{1}{\sigma^2}||y - X\beta||^2 + \nu\sum_{j=1}^{p}\psi(\beta_j^2) \tag{1.10}$$

où ψ est une fonction de pénalisation et ν le paramètre de pénalisation. En effet, l'équation (1.10) peut s'interpréter comme l'opposé du logarithme de la densité *a posteriori* de β dont le prior associé à β_j serait proportionnel à $exp\{-\nu\psi(\beta_j^2)\}$. De façon générale, ces priors reposent sur le formalisme hiérarchique suivant pour $j = 1,\ldots,p$:

$$\beta_j|\theta_j^2 \sim \mathcal{N}(0,\theta_j^2), \tag{1.11}$$
$$\theta_j^2 \sim \mathcal{F}(\theta_j^2; 1/\nu), \tag{1.12}$$

où $\mathcal{F}$ est une certaine loi de distribution de la variance ou de la précision. Par exemple, si tous les θ_j^2 sont identiques et que $\mathcal{F}$ est une loi inverse-Gamma, on retrouve la version bayésienne de la pénalisation ridge [Hoerl and Kennard, 1970]. Si cette loi est égale à une loi exponentielle, on retrouve la version bayésienne de la pénalisation Lasso introduite par Tibshirani [1996].

Remarque. On peut noter d'une part que l'équation (1.11) définit les priors de type *continu de réduction* comme un mélange infini de loi gaussienne, et d'autre part que la loi marginale des paramètres β_j n'est plus gaussienne :

$$\mathcal{L}(\beta_j) = \int \phi\left(0,\theta_j^2\right) g(\theta_j^2)d\theta_j^2$$

où ϕ désigne la fonction de densité de la loi gaussienne, et g la fonction de densité de la loi de θ_j^2.

Pour obtenir de bonnes performances de sélection et de régularisation, un prior de type *continu de réduction* doit satisfaire deux caractéristiques. La première est de permettre d'assigner suffisamment de masse en zéro pour réduire les coefficients de

régression non-pertinents vers zéro. La seconde est d'avoir des queues de distributions suffisamment lourdes pour pouvoir estimer, sans biais, les coefficients non-nuls. Ces deux propriétés ne sont pas respectées, par exemple, dans le cas du Lasso bayésien. Ainsi de nombreux auteurs ont proposé de nouveaux priors tels que le prior *Normal-Exponentiel-Gamma* (NEG) [Griffin and Brown, 2007], le *Normal-Gamma-Double-Pareto* (NGDP) [Armagan et al., 2013], le *Normal-Gamma* (NG) [Griffin et al., 2010], le *Normal-Gamma-Gamma* (NGG) [Griffin et al., 2017], le *Normal-Beta-Prime* (NBP) [Bai and Ghosh, 2019]. La construction hiérarchique de ces priors est détaillée dans le tableau 1.1.

Prior	**Prior sur τ^2**	**Prior sur ω_j^2**	**R package**
NIG	$\tau^2 \sim \mathcal{IG}(s, r)$	$\omega_j^2 = 1$	BGLR [Pérez and de Los Campos, 2014] bayesreg [Makalic and Schmidt, 2016]
Student(d, v)	$\tau^2 = 1$	$\omega_j^2 = \mathcal{IG}(d/2, v\ d/2)$ $v \sim \mathcal{G}(s, r)$	BGLR [Pérez and de Los Campos, 2014]
NE [Park and Casella, 2008]	$\tau^2 \sim \mathcal{IG}(a, b)$	$\omega_j^2 \sim \mathcal{E}xp(1/2)$	BGLR [Pérez and de Los Campos, 2014] bayesreg [Makalic and Schmidt, 2016]
NEG [Griffin and Brown, 2007]	$\tau^2 = 1$	$\omega_j^2 \sim \mathcal{E}xp(z_j/2)$ $z_j \sim \mathcal{G}(a, b)$	
NGDP [Armagan et al., 2013]	$\tau^2 = 1$	$\omega_j^2 \sim \mathcal{E}xp(\lambda_j^2/2)$ $\lambda_j \sim \mathcal{G}(\alpha, \eta)$	
DL [Bhattacharya et al., 2015]	$\tau \sim \mathcal{G}(a\,p, 1/2)$	$\omega_j^2 = \psi_j \phi_j^2$ $\psi_j \sim \mathcal{E}xp(1/2)$ $\phi \sim \mathcal{D}ir(a, \ldots, a)$	dlbayes [Zhang and Li, 2018]
R2D2 [Zhang et al., 2020]	$(\tau^2)^2 \sim \mathcal{G}(b, 1)$	$\omega_j^2 = \psi_j(\sigma^2 \lambda_j/2)^{1/2}$ $\psi_j \sim \mathcal{E}xp(1/2)$ $\lambda_j \vert \xi \sim \mathcal{G}(a, 1)$	
NTPB [Armagan et al., 2011]	$\tau^2 \sim \pi(\tau^2)$	$\omega_j^2 \sim \mathcal{B}'(a, b)$	
NG [Griffin et al., 2010]	$\tau^2 = 1$	$\omega_j^2 \sim \mathcal{G}(\lambda, \gamma)$	
NGG [Griffin et al., 2017]	$\tau^2 = 1$	$\omega_j^2 \sim \mathcal{G}(\lambda, \gamma_j)$ $\gamma_j \sim \mathcal{G}(a, b)$	
HS [Carvalho et al., 2009] [Carvalho et al., 2010]	$\tau \sim \mathcal{C}^+(0, 1)$	$\omega_j \sim \mathcal{C}^+(0, 1)$	bayesreg [Makalic and Schmidt, 2016] horseshoe [van der Pas et al., 2016] fastHorseshoe [Hahn et al., 2016]
HS+ [Bhadra et al., 2017]	$\tau \sim \mathcal{C}^+(0, 1)$	$\omega_j \sim \mathcal{C}^+(0, \eta_i)$ $\eta_i \sim \mathcal{C}^+(0, 1)$	bayesreg [Makalic and Schmidt, 2016]
NBP [Bai and Ghosh, 2019]	$\tau^2 = 1$	$\omega_j^2 \sim \mathcal{B}'(a_p, b)$ $a_p \sim \pi(a_p) 1_{[1/p < a_p < 1]}$	NormalBetaPrime [Bai and Ghosh, 2019]

TABLE 1.1 – *Priors bayésiens de sélection de variables (non-structurées) de type continu de réduction sous la forme de priors globaux-locaux. p est le nombre de variables sujets à la sélection.*

Polson and Scott [2010] proposent une modélisation alternative en supposant que le paramètre θ_j^2 se décompose en un produit de deux termes :

$$\theta_j^2 = \tau^2 \omega_j^2$$

où τ^2 contrôle le niveau de contraction autour de zéro de l'ensemble des coefficients de régression, et ω_j^2 est un paramètre local propre à chaque coefficient qui permet de faire ressortir le coefficient s'il est associé à une variable pertinente. Contrairement à

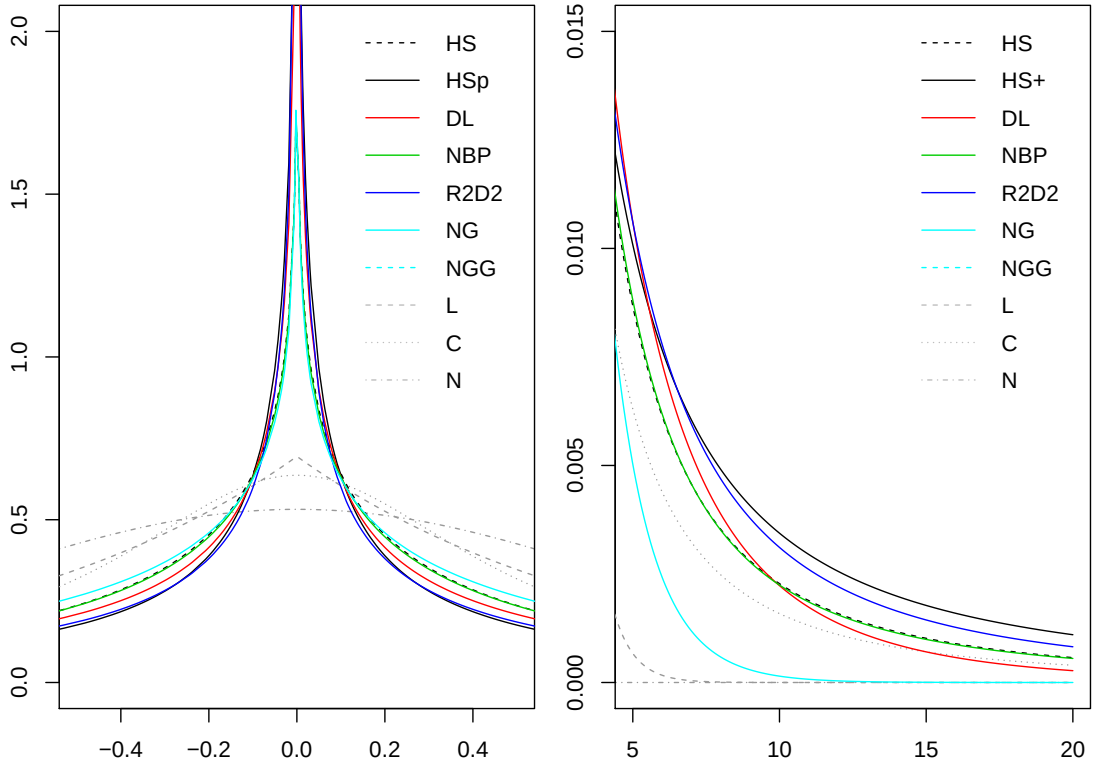


FIGURE 1.2 – *Graphiques des distributions a priori des priors HS, HS+, DL, NBP, R2D2 en comparaison aux priors Normale (N), Cauchy (C) et Laplace (L). Le graphique de gauche permet de visualiser la masse en zéro des priors tandis que le graphique de droite permet de visualiser les queues de distribution des priors.*

l'approche précédente qui fait référence à une approche globale, cette paramétrisation fait référence à une approche locale-globale en référence aux hyper-paramètres locaux ω_j^2 et à l'hyper-paramètre global τ^2. Différentes lois ont été proposées pour modéliser ces paramètres. En particulier, le prior *horseshoe* (HS) [Carvalho et al., 2009, 2010] suppose que les ω_j^2 comme le paramètre τ^2 sont distribués selon une loi demi-Cauchy d'espérance nulle et de variance unitaire. On remarquera que dans le prior HS les paramètres locaux peuvent être mis en relation avec la probabilité d'inclusion obtenue dans les priors de type spike-and-slab. D'autres priors appartenant à cette classe ont également été explorés tels que le prior *Normal-Three-Parameter Beta* (NTPB) [Armagan et al., 2011], le *Dirichlet-Laplace* (DL) [Bhattacharya et al., 2015], le *horseshoe plus* (HS+) [Bhadra et al., 2017] ou plus récemment le R^2*-induced Dirichlet Decomposition* (R2D2) [Zhang et al., 2020]. La construction hiérarchique de ces priors est détaillée dans le tableau 1.1.

Pour conclure cette partie, les priors *continus de réduction* peuvent toujours s'exprimer comme un mélange continu de lois normales :

$$\beta_j|\tau^2, \omega_j^2 \sim \mathcal{N}(0, \tau^2\omega_j^2), \quad j = 1, \ldots, p. \tag{1.13}$$

Selon les choix que l'on fait sur la distribution des paramètres τ^2 et ω_j^2, il est possible de proposer une gamme très importante de lois *a priori* pour les paramètres de régression. Le tableau 1.1 résume de façon synthétique certaines de ces lois selon les distributions assignées aux paramètres τ^2 et ω_j^2. La figure 1.2 permet de visualiser le comportement

autour de zéro ainsi que le poids des queues de distribution des priors HS, HS+, DL, NBP, R2D2.

1.3.2 Lois *a priori* pour la sélection de variables et la régularisation de modèles : cas où les paramètres sont structurés

La notion de données structurées est fréquente en statistique. Cette structuration peut amener à considérer des groupes de paramètres et à vouloir les sélectionner. En génétique, par exemple, les marqueurs moléculaires sont structurés le long du génome. La structuration peut ainsi être au niveau des gènes ou des chromosomes, ou bien encore en fonction des groupes de déséquilibre de liaisons. On est alors en présence d'une structuration de type groupe. Une structuration dans les données peut également impliquer une dépendance entre les paramètres. Par exemple, les paramètres associés à des variables mesurées à des temps successifs sont dépendants. L'utilisation d'un VCM (cf équation (1.5)) fait également apparaître naturellement la notion de structuration. En effet, inclure ou non une variable explicative X_j dans l'équation (1.5) revient à mettre à zéro l'ensemble des paramètres b_j liés à l'approximation dans une approche non-paramétrique de la fonction $f_j(S_j)$. De plus, l'hypothèse d'une fonction f_j continue plus ou moins lisse implique une structuration de dépendance entre les paramètres contenus dans le vecteur b_j.

Différentes approches ont alors été imaginées pour prendre en compte ces structurations. La première forme que nous présentons porte sur une structure en groupe. La seconde concerne une structuration induite par des dépendances (temporelle, spatiale ou selon la distance entre les QTL au sein d'un chromosome). Enfin nous présentons une stratégie pour combiner ces deux formes de structure.

Cas d'une structuration de type groupe : extension des priors *spike-and-slab* et *continus de réduction*

Pour généraliser les priors présentés dans la partie précédente (cf section 1.3.1) à des paramètres associés à des variables groupées, la solution repose sur l'introduction d'un paramètre de variance spécifique au groupe. Celui-ci permet de contrôler l'influence de l'ensemble des variables de chaque groupe séparément. Initialement, ce sont les priors de type *continus de réduction* qui ont été étendus au cas de variables groupées et qui, à nouveau, ont eu pour objectif de mimer les approches développées dans le cas fréquentiste comme le group Lasso [Yuan and Lin, 2006] ou encore sa version adaptative. Posons $\beta_k = (\beta_{k_1}, \ldots, \beta_{k_{p_k}})'$ le vecteur de longueur p_k de paramètres associés au groupe k. τ^2 le paramètre de variance global qui contrôle le niveau global de réduction autour de 0. ξ_k^2 le paramètre de variance spécifique au groupe k qui permet de faire sortir un groupe lorsque celui-ci est pertinent. Un prior de type *groupe continu*

de réduction peut alors s'exprimer de la façon générale suivante :

$$\beta_k|\tau^2,\xi_k^2 \sim \mathcal{N}_{p_k}\left(0,\tau^2\xi_k^2 I_{p_k}\right), \tag{1.14}$$
$$(\tau^2,\xi_k^2) \sim \mathcal{F}\left(\tau^2,\xi_k^2\right), \tag{1.15}$$

où $\mathcal{F}$ est une loi de distribution à spécifier dépendant de paramètres connus ou non. Nous pouvons remarquer que tous les éléments du vecteur β_k ont la même variance et sont indépendants entre eux *a priori*. Des cas particuliers ont été développés. Par exemple, Kyung et al. [2010] proposent le prior *group Normal-Gamma* :

$$\beta_k|\xi_k^2 \sim \mathcal{N}_{p_k}(0,\xi_k^2 I_{p_k}), \quad \xi_k^2|\tau^2 \sim \mathcal{G}\left(\frac{p_k+1}{2},\frac{1}{2\tau^2}\right). \tag{1.16}$$

Ce prior est analogue à la pénalisation *group Lasso* fréquentiste $\lambda\sum_{k=1}^K ||\beta_k||_2$ introduit par Yuan and Lin [2006] avec $\tau^2 = 1/\lambda$. Xu et al. [2015] proposent d'étendre le prior précédent à la double sélection de groupes de variables et de variables en considérant le prior :

$$\beta_k|\xi_k^2,\omega_k^2 \sim \mathcal{N}_{p_k}(0,Q_k^{-1}), \tag{1.17}$$

$$p(\xi_k^2,\omega_{k_1}^2,\ldots,\omega_{k_{p_k}}^2) = c_g(\lambda_1,\lambda_2)\prod_{j=1}^{p_k}\left[(\omega_{k_j}^2)^{-\frac{1}{2}}\left(\frac{1}{\omega_{k_j}^2}+\frac{1}{\omega_{k_j}^2\xi_k^2}\right)^{-\frac{1}{2}}\right](\xi_k^2)^{-\frac{1}{2}} \tag{1.18}$$
$$exp\left\{-\frac{\lambda_1^2}{2}\sum_{j=1}^{p_k}\omega_{k_j}^2+\frac{\lambda_2^2}{2}\xi_k^2\right\},$$

avec $Q_k = diag\left\{\left(\frac{1}{\xi_k^2}+\frac{1}{\omega_{k_j}^2}\right), j=1,\ldots,p_k\right\}$. Nous pouvons remarquer ici que la matrice de précision Q_k est la somme de deux matrices diagonales. La première matrice dépend uniquement d'un paramètre de variance ξ_k^2, propre au groupe k et ayant pour but de réduire à zéro l'ensemble des éléments du groupe s'il n'est pas pertinent. La seconde matrice dépend de paramètres locaux ω_k^2 qui permettent une action plus ciblée en réduisant à zéro certains éléments de β_k. Là encore, ce prior est analogue à la pénalisation *sparse group Lasso* égale à $\lambda_1\sum_{k=1}^K||\beta_k||_2 + \lambda_2\sum_{k=1}^K\sum_{j=1}^{p_k}|\beta_{k_j}|$ introduit par Friedman et al. [2010]. Toutefois, ces deux priors produisent, comme dans le cas non-structuré, des estimations biaisées.

Plus récemment, le prior *spike-and-slab* a également été généralisé au cas de la sélection de groupe de variables. Initialement, l'approche proposée par Scheipl et al. [2012] repose sur la version alternative du *spike-and-slab* introduite par Ishwaran et al. [2005] et présentée dans la section 1.3.1. Scheipl et al. [2012] proposent également d'utiliser la reparamétrisation des paramètres de variance, proposée par Gelman et al. [2006] pour améliorer les performances de mélangeance des chaînes MCMC. La généralisation de l'approche *spike-and-slab* au groupe est abordée par Xu et al. [2015]. Ils proposent de combiner les priors de type *groupe Normal-Gamma* présentés précédemment pour la partie *slab* avec une masse de Dirac pour la partie spike :

$$\beta_k|\xi_k^2 \sim \gamma_k\mathcal{N}_{p_k}(0,\xi_k^2 I_{p_k}) + (1-\gamma_k)\prod_{j=1}^{p_k}\delta(0)$$
$$\xi_k^2|\lambda \sim \mathcal{G}\left(\frac{p_k+1}{2},\frac{\lambda^2}{2}\right) \text{ et } \gamma_k \sim \mathcal{B}ern\left(\pi\right).$$

La loi *a posteriori* de γ_k permet alors d'obtenir la probabilité d'inclusion du $k^{\text{ème}}$ groupe de variables dans le modèle. Liquet et al. [2017] étendent cette approche au cas d'une variable réponse répétée. Xu et al. [2015] proposent également une version sparse du prior *group spike-and-slab* en considérant un prior *spike-and-slab* sur des paramètres de variance locaux avec une loi normale tronquée sur la partie *slab* :

$$\begin{aligned}\beta_{k_j}|\gamma_k^{(1)},\omega_{k_j}^2 &\sim \gamma_k^{(1)}\mathcal{N}(0,\omega_{k_j}^2)+(1-\gamma_k^{(1)})\delta(0), \quad \gamma_k^{(1)}\sim\mathcal{B}ern(\pi_1),\\ \omega_{k_j}^2|\gamma_{k_j}^{(2)} &\sim \gamma_{k_j}^{(2)}\mathcal{N}^{+}(0,\tau^2)+(1-\gamma_{k_j}^{(2)})\delta(0), \quad \gamma_{k_j}^{(2)}\sim\mathcal{B}ern(\pi_2).\end{aligned}$$

Les propriétés statistiques du prior *group spike-and-slab* , notamment sa consistance, ont été démontrées par Yang et al. [2020].

Les priors explicités jusque-là ont été spécifiquement développés pour mettre à zéro ou non des groupes de paramètres. Cependant, ils ne permettent pas de prendre en compte une structure de dépendance entre les paramètres. Pour ce genre de structuration traditionnellement rencontrée dans l'analyse de séries chronologiques ou en geostatistique, une approche repose sur la théorie des processus gaussiens [Rasmussen and Williams, 2006].

Cas d'une structuration de type dépendance : les processus gaussiens

Prendre en compte les dépendances entre les effets d'une variable explicative observée selon une structure donnée, peut s'avérer intéressant pour mieux représenter des phénomènes physiques, biologiques ou environnementaux. Cette problématique a été largement étudiée dans le cadre de l'analyse de données temporelles ou spatiales. Posons $\beta(S)=(\beta_1(S_1),\ldots,\beta_p(S_p))'$ le vecteur de paramètres dépendants, avec S une variable exogène informant de la structuration des paramètres. Par exemple, dans le cas de paramètres évoluant au cours du temps, S serait alors le vecteur des indices du temps. La solution repose sur l'utilisation de lois *a priori* gaussiennes multivariées :

$$\beta(S)\sim\mathcal{N}_p\left(0,\tau^2Q^{-1}(S)\right) \tag{1.19}$$

où $Q(S)$ est une fonction de précision qui permet de refléter les dépendances entre les paramètres $\beta_j(S)$, et τ^2 un paramètre de variance global. Différentes structures de $Q(S)$ ont été explorées. Par exemple, dans le cas où S est un vecteur d'indice de temps, Vanhatalo et al. [2019] proposent l'emploi d'une fonction de covariance de Mátern dans le cadre des VCMs appliqués à la cartographie pangénomique. Dans le cas où les mesures sont réparties régulièrement, une stratégie classiquement utilisée consiste à modéliser la structure de dépendance au travers d'une structure de type autorégressive :

$$Q(S)=\begin{pmatrix} 1 & -\rho & & & \\ -\rho & 1+\rho^2 & -\rho & & \\ & \ddots & \ddots & \ddots & \\ & & -\rho & 1+\rho^2 & -\rho \\ & & & -\rho & 1 \end{pmatrix}. \tag{1.20}$$

où ρ est un paramètre connu ou non. Ce prior conduit alors à considérer la dynamique des paramètres de la façon suivante :

$$\beta_t=\rho\beta_{t-1}+\varepsilon_t, \quad \varepsilon_t\sim\mathcal{N}(0,\tau^2), \quad |\rho|<1, \quad t\geq 2.$$

La structure de type marche aléatoire est également couramment rencontrée dans la littérature. Cette structure peut être vue comme un cas particulier de la structure autorégressive avec $\rho = 1$ [Rue and Held, 2005]. La matrice de précision peut alors être décomposée comme le produit croisé d'une matrice de différence finie $Q(S) = D'D$. Certains auteurs proposent de combiner les priors *continus de réduction*, tels qu'introduits dans la section 1.3.1, aux priors gaussiens multivariés de type marche aléatoire. Des paramètres de variance locaux ω_j^2 propres à chaque différence sont ainsi introduits dans la matrice de précision $Q(\omega^2, S) = D'\Omega^{-1}D$ avec $\Omega = diag\left\{\omega_j^2, j = 1, \ldots, p-1\right\}$. Cette modélisation apporte plus de flexibilité dans l'estimation, permettant à certains paramètres adjacents d'être fusionnés (égaux) ou au contraire éloignés. Ainsi, Rue and Held [2005] considèrent une distribution de Student. Kyung et al. [2010] considèrent une distribution de Laplace. Cette dernière distribution correspond à la pénalisation *fusion Lasso* fréquentiste $\lambda \sum_{t=t_2}^{t_T} |\beta_t - \beta_{t-1}|$ introduite par Land and Friedman [1996]. Faulkner and Minin [2018] considèrent quant à eux un prior *horseshoe*.

Comme indiqué précédemment, cette approche permet de tenir compte d'une structure de dépendance entre les paramètres mais ne permet pas une sélection des paramètres importants. Ročková and George [2018] proposent par exemple un prior *spike-and-slab* sur chaque élément $\beta_j(S)$ combiné à un prior de type autorégressif sur la loi p_{slab} :

$$\beta_j(S)|\gamma_j, \tau_1^2, \tau_0^2 \sim \gamma_j \mathcal{N}(\rho\beta_{j-1}(S), \tau_1^2) + (1-\gamma_j)\mathcal{N}(0, \tau_0^2).$$

Kyung et al. [2010] proposent une version bayésienne de la pénalisation *fused Lasso* fréquentiste $\lambda_1 \sum_{t=t_1}^{t_T} |\beta_t| + \lambda_2 \sum_{t=t_2}^{t_T} |\beta_t - \beta_{t-1}|$ introduite par Tibshirani et al. [2005], en considérant le prior :

$$\beta(S) \sim \mathcal{N}_{t_T}(0, (\Omega_1^{-1} + D'\Omega_2^{-1}D)^{-1}), \tag{1.21}$$
$$\omega_{1_t}^2|\lambda_1 \sim \mathcal{E}xp(\lambda_1/2), \quad t = t_1, \ldots, t_T, \tag{1.22}$$
$$\omega_{2_t}^2|\lambda_2 \sim \mathcal{E}xp(\lambda_2/2), \quad t = t_2, \ldots, t_T, \tag{1.23}$$

avec $\Omega_1 = diag\left\{\omega_{1_t}^2, t = t_1, \ldots, t_T\right\}$ et $\Omega_2 = diag\left\{\omega_{2_t}^2, t = t_2, \ldots, t_T\right\}$. La matrice de précision se décompose comme la somme de deux matrices. La première matrice Ω_1^{-1} permet de réduire à zéro les paramètres β_t non-importants, tandis que la deuxième partie $D'\Omega_2^{-1}D$ permet de prendre en compte la structure de dépendance présente entre les paramètres β_t adjacents.

Cas d'une structuration de type groupe avec dépendance intra-groupe

Certains modèles, tels que les VCM, nous amènent à considérer des groupes de paramètres avec une dépendance intra-groupe. Il est alors nécessaire de tenir compte de cette structuration complexe pour une régularisation optimale des modèles. Il est ainsi possible de combiner les différents priors introduits précédemment pour répondre à ce besoin. Cependant, peu de travaux ont été réalisés dans ce contexte.

Dans le cas d'une estimation P-spline des fonctions $f_j(S_j) = B_j b_j$ de l'équation (1.5) du VCM, b_j est supposé suivre une loi gaussienne multivariée $\mathcal{N}(0, \xi_j^2(D'D)^{-1})$. Scheipl et al. [2012] proposent une reparamétrisation de la forme $B_j b_j = U_j V_j^{\frac{1}{2}} \beta_j$ tel

que $B_j(D'D)^{-1}B_j' = U_jV_jU_j'$, la décomposition SVD avec U_j une matrice orthogonale et V_j une matrice diagonale de valeurs singulières. Le vecteur de paramètres dépendants b_j est alors transformé en un vecteur de paramètres indépendants β_j. Scheipl et al. [2012] décrivent alors un prior *groupe spike-and-slab* sous la forme d'un mélange de loi normale sur la variance tel que discuté précédemment.
Dans le chapitre 3, nous proposons une alternative plus directe, adaptée à l'estimation fonctionnelle (P-spline), ou non-fonctionnelle des fonctions f_j dans le cas où $S_j \equiv S$ est un vecteur d'indice régulier de temps. Nous considérons un prior *group spike-and-slab* avec une distribution gaussienne multivariée ayant une structure de précision de type marche aléatoire sur la partie *slab*, et un produit de masse de Dirac en zéro sur la partie *spike*.

Dans le cas particulier d'une régression sur différents signaux, chaque signal définit un groupe de variables ordonnées. Il est alors parfois nécessaire d'identifier les variables pertinentes, en plus de vouloir identifier des groupes pertinents, tout en prenant en compte l'ordre naturel intra-groupe. Dans cet objectif, Zhang et al. [2015] proposent un prior *group spike-and-slab* combiné à un prior *fused Lasso* utilisé sur la partie *slab* et une masse de Dirac sur la partie *spike*.
Finalement, aucun travail n'a été réalisé jusqu'à présent pour étendre des priors *continus de réduction* au cas de la double sélection de groupes et de variables avec dépendance intra-groupe. Dans le cas de variables ordonnées, un tel prior pourrait ainsi être construit en généralisant la décomposition de la matrice de précision introduite par Kyung et al. [2010] dans le prior *fused Lasso* au cas de groupe de variables. Cette décomposition pourrait ainsi être combinée à des paramètres de variance à trois niveaux permettant de contrôler l'ensemble des coefficients de régression, les groupes de coefficients, ainsi que les coefficients et leurs différences. Dans le chapitre 4, nous proposons un tel prior avec des distributions demi-Cauchy sur les paramètres de variance aboutissant sur un prior *sparse group fused horseshoe*.

1.3.3 Priors de sélection de composantes de la variance et régularisation de modèles linéaires mixtes

La question de la sélection des effets aléatoires est moins développée. Les premières approches pour sélectionner simultanément les effets fixes et les composantes de la variance utilisent des critères de sélection de modèle [Rao and Wu, 1989, Vaida and Blanchard, 2005, Müller et al., 2013, Delattre and Poursat, 2020]. Toutefois, comme dans le contexte des modèles linéaires, les approches de type vraisemblance pénalisée ont également été développées, en particulier lorsque le nombre de prédicteurs augmente. Par exemple, Bondell et al. [2010], Ibrahim et al. [2011] combinent des techniques de vraisemblance pénalisée, utilisant soit une pénalisation Lasso adaptatif ou SCAD en se basant sur la décomposition de Cholesky de la matrice de covariance des effets aléatoires ou sa version modifiée. Dans le contexte de grande dimension, Fan and Li [2012] proposent une approche en deux étapes tandis que Li et al. [2018] proposent une double pénalisation pour estimer et sélectionner les effets fixes et aléatoires dans le contexte longitudinal. La reparamétrisation du LMM par l'utilisation d'une décomposition de Cholesky modifiée de la matrice de covariance des effets aléatoires est initialement proposée dans un cadre bayésien par Chen and Dunson [2003], Kinney and

Dunson [2007]. Une telle technique permet de considérer les paramètres d'écarts types des effets aléatoires comme des paramètres de régression. Dans leur approche, une loi *a priori* de type *spike-and-slab* est utilisée pour les effets fixes et les écarts types, avec une distribution gaussienne tronquée pour la partie diffuse associée aux écarts types. Frühwirth-Schnatter and Tüchler [2008] proposent une approche connexe, modélisant directement les éléments de la décomposition de Cholesky en utilisant un a priori *spike-and-slab* où la distribution de la partie diffuse est une distribution gaussienne sans contrainte. Des approches alternatives, basées sur des priors de type *continus de réduction*, ont également été proposées et comparées pour la sélection des composantes de la variance. On peut citer entre autres les distributions de Student et Cauchy [Gelman et al., 2006, Polson et al., 2012]. Les priors Laplace et Normal-Gamma ont également été longuement discutés dans le cadre d'un modèle à intercept aléatoire [Frühwirth-Schnatter and Wagner, 2011]. Enfin, notons que la question de savoir comment réduire les paramètres de variance vers zéro ne s'est pas posée uniquement dans le contexte des LMMs. Ces objectifs ont été étudiés dans différents contextes statistiques. Dans les VCMs et les modèles d'espace-état, Bitto and Frühwirth-Schnatter [2019], Cadonna et al. [2020] proposent l'utilisation de priors double ou triple Gamma étendant le prior Normal-Gamma [Griffin et al., 2010], ou plus généralement la classe de distribution scaled-Beta [Perez et al., 2017].

Dans le chapitre 2, nous proposons d'étudier le comportement de quatre priors dans le contexte de la cartographie de QTL et fonctionnelle (*functional mapping*).

1.4 Plan de thèse

Après cette introduction, nos contributions sont présentées dans les trois chapitres suivants. Chaque chapitre est constitué d'un article, dont nous faisons une courte introduction visant à en décrire les grandes lignes. Nous terminons enfin par une conclusion et différentes perspectives.

Le chapitre 2 s'intéresse à la sélection de composantes de la variance dans les modèles linéaires mixtes. Dans ce travail, nous montrons l'intérêt de considérer des priors continus à queue lourde, et en particulier nous recommandons l'utilisation de priors de type locaux-globaux (le prior *horseshoe*). Ces résultats sont obtenus en comparant les performances de ce prior à celles d'autres méthodes déjà étudiées dans ce contexte. Cette analyse est réalisée au travers de deux applications génétiques. La première est tournée sur l'identification de QTL à l'aide de matrices d'apparentement (*IBD-QTL mapping*), au travers d'un modèle "animal". La seconde s'intéresse à l'évolution au cours du temps de l'architecture génétique de la compacité des feuilles de l'espèce *arabidopsis thaliana* (L. Heynh), au travers d'un modèle à intercept et pentes aléatoires. Le temps est considéré comme un facteur groupant aléatoire. Ainsi, à chaque pas de temps, les effets des marqueurs sont modélisés par des réalisations d'effets aléatoires supposées structurées à l'aide d'une matrice de corrélation inconnue. Les effets sont supposés indépendants au cours du temps. Un article a été soumis à *Biometrics*.

Le chapitre 3 s'intéresse également à l'évolution au cours du temps de l'architecture génétique d'un caractère d'intérêt, au travers l'utilisation d'un modèle à coefficients va-

riants. Ce modèle permet de décrire l'évolution de l'effet de chaque marqueur comme une fonction du temps, au travers soit d'une interpolation P-splines, soit d'une estimation direct des effets. Pour chaque marqueur, ces deux modélisations font intervenir un groupe de paramètres ordonnées (coefficients P-splines ou les effets au cours du temps directement). Ainsi, l'identification des marqueurs pertinents passe par la sélection de groupe de paramètres ordonnés. Nous proposons alors un prior *group spike-and-slab* combiné à une distribution multivariée gaussienne de type marche aléatoire comme loi diffuse (P_{slab}). Contrairement à la modélisation RIS proposée dans le chapitre précédent, ce prior suppose l'indépendance des effets entre les différents marqueurs à chaque pas de temps, mais suppose une dépendance temporelle entre les effets d'un même marqueur. De plus, il permet d'obtenir un profil d'effets dans le temps homogène pour chaque marqueur. Cette étude a été publiée dans la revue *Journal of the Royal Statistical Society (Series C)*.

Le chapitre 4 est consacré à l'identification de facteurs environnementaux (température, humidité, etc.) agissant sur un processus biologique, ainsi qu'à l'identification des périodes au cours du processus sur lesquels ces facteurs influent. Nous nous sommes notamment intéressés à l'abscission du fruit du *palmier à huile*. Cette problématique peut se conceptualiser comme un problème de double sélection de groupes de paramètres (identification des facteurs) et de paramètres (identifications des période de temps) tout en tenant compte de l'ordre naturel entre eux. Nous proposons alors un prior *group fused horseshoe* pour répondre à cette question. Il généralise ainsi les priors *horseshoe* et *fused Lasso*. Contrairement au prior introduit dans le chapitre précédent, ce prior permet d'estimer des profils d'effets dans le temps pouvant être complexes (changements abrupts, variabilité non homogène) ou encore non-nul uniquement sur certaines périodes. Il permet ainsi de s'adapter à tout type de profil. Nous préparons une soumission de ce travail à la revue *Journal of Agricultural, Biological and Environmental Statistics*.

Dans cette thèse, nous avons adapté, analysé et comparé, dans le contexte de données longitudinales, différentes lois *a priori* classiquement utilisées en régression linéaire. Ces distributions *a priori* permettent de se prémunir du sur-ajustement tout en identifiant les variables pertinentes, qu'elles soient à effet fixe ou aléatoire, et tout en tenant compte de différentes structures de dépendance induites par la nature des données répétées. Nous avons aussi montré pourquoi ces choix offrent une meilleure compréhension des phénomènes biologiques étudiés grâce à une estimation plus efficaces et une modélisation parcimonieuse. L'utilisation de simulations et de nombreux jeux de données a permis de valider l'intérêt et l'efficacité de ces lois *a priori*. Enfin, dans le chapitre 5, nous proposons différentes perspectives concernant l'intérêt de tels priors lorsque la variable réponse n'est plus gaussienne ou que les pas de temps d'observation ne sont pas réguliers. Nous discutons aussi de différents aspects plus méthodologiques notamment en lien avec la prise en compte de la multi-collinéarité entre les variables explicatives, ou encore les interactions génotype$\times$environnement.

II

Sélection de composantes de la variance dans un modèle linéaire mixte

Sommaire

2.1 Introduction 21

2.2 Article : Continuous shrinkage priors for fixed and random effects selection in linear mixed models: application to genetic mapping . . 22

2.1 Introduction

Les modèles linéaires mixtes sont largement utilisés pour l'analyse de données. Cependant, le choix des facteurs aléatoires ainsi que des interactions possibles entre régresseurs et facteurs aléatoires, peut être délicat et peut mener à un sur-ajustement. Dans ce chapitre, nous nous concentrons sur deux applications en génétique impliquant un grand nombre d'effets aléatoires. La première application prend place dans le contexte de la cartographie IBD-QTL. Ce genre de cartographie consiste à combiner l'information génétique et pedigree, aboutissant sur un ensemble de matrices d'apparentement (IBD) associées à différentes positions sur l'ensemble du génome. L'effet génétique global est alors décomposé en une somme d'effets génétiques locaux. L'identification des QTL reliés à un caractère phénotypique d'intérêt passe alors par une sélection d'effets aléatoires. La seconde application prend place dans le contexte de l'étude de l'évolution de l'architecture génétique d'un caractère au cours du temps. Les technologies actuelles de génotypage haut débit et de phénotypage haut débit nous permettent d'avoir accès à un grand nombre de marqueurs moléculaires ainsi qu'à des mesures répétées d'un ou plusieurs caractères phénotypiques au cours du temps. Il est alors possible de mettre en lien ces deux catégories de données pour étudier cette évolution. L'analyse de telles données peut être réalisée à l'aide d'un modèle à intercept et pentes aléatoires. Dans cette application, nous considérons le facteur temps

comme un effet aléatoire venant corriger l'effet de chaque marqueur à chaque pas de temps. L'identification des QTL ayant un effet sur le caractère phénotypique d'intérêt implique à la fois une sélection des variables à effets fixes et aléatoires associées à chaque marqueur.
Ces deux applications aboutissent à une même question : comment mettre à zéro des effets aléatoires non pertinents ? Une solution consiste à chercher à mettre à zéro les composantes de la variance associées aux effets aléatoires non pertinents. Dans ce chapitre, nous proposons d'étendre le prior *horseshoe* pour la sélection de composantes de la variance en présentant une version pliée (*folded*) de ce prior. Dans le cadre du modèle à intercept et pentes aléatoires, nous utilisons la reparamétrisation de la matrice de covariance des pentes aléatoires proposée par Chen and Dunson [2003] permettant de réécrire cette matrice comme le produit d'une matrice diagonale d'écart-type, de la matrice de corrélation et de la même matrice diagonale d'écart-type. Nous mettons également en œuvre la reparamétrisation angulaire de la matrice de corrélation des pentes aléatoires introduite par Pinheiro and Bates [1996], permettant d'assurer qu'elle soit définie positive. Le prior sinusoïdale introduit par Pourahmadi and Wang [2015] est alors considérer sur les paramètres angulaires.
Nous appliquons les modèles bayésiens hiérarchiques proposés sur les deux jeux de données ayant motivé ce développement. Nous comparons les performances du prior *horseshoe* appliqué aux composantes de la variance, avec les priors *zero-inflated half normal* [Chen and Dunson, 2003, Kinney and Dunson, 2007] et *Cauchy* [Gelman et al., 2006] qui ont également été proposés pour la sélection de composantes de la variance.

2.2 Article : Continuous shrinkage priors for fixed and random effects selection in linear mixed models: application to genetic mapping

L'article a été soumis dans la revue Biometrics.

Biometrics -, 1–11
- 2021

DOI: ...

Continuous shrinkage priors for fixed and random effects selection in linear mixed models: application to genetic mapping

Benjamin Heuclin[1,2,3], Marie Denis[2,3], Catherine Trottier[1,6], Sébastien Tisné[2,3] and Frédéric Mortier[4,5,*]

[1]IMAG, Univ Montpellier, CNRS, Montpellier, France,
[2]CIRAD, UMR AGAP Institut, F-34398 Montpellier, France,
[3]UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, F-34398 Montpellier, France,
[4]Forêts et Sociétés, Cirad, F-34398 Montpellier, France,
[5]Forêts et Sociétés, Univ Montpellier, Cirad, Montpellier, France,
[6]AMIS, Univ Paul-Valéry Montpellier 3, Montpellier, France.
email: frederic.mortier@cirad.fr

Summary: The identification of random factors to include in a linear mixed model is crucial for modeling dependence structures while avoiding over-fitting. Random effects selection can be achieved by shrinking non-relevant variance parameters towards zero. We propose extending the horseshoe prior for variance components selection in a folded version. Motivated by two applications, the folded-horseshoe prior is evaluated either in a genetic breeding or in a functional mapping context. In the latter, we use a polar parametrization of the correlation matrix of random effects, using sinusoidal priors for angular parameters. Finally, we design efficient MCMC algorithms taking advantage of Kronecker product properties. From a statistical point of view, we show that the folded-horseshoe prior outperforms the folded-Cauchy when the number of parameters is close to the sample size. For variance component selection, it performs as well as the folded-spike-and-slab but it is computationally more efficient. We also show the impact of erroneous dependence structures assumptions on the selection and the estimation of variance components. From a genetic point of view, the numerical results highlight the efficiency of the folded-horseshoe prior. In particular, this prior selects molecular markers already identified in these data but also new markers. Finally, we discuss how and why linear mixed models are an interesting alternative to usual functional mapping approaches.

Key words: Angular parametrization; Fixed and random effects selection; Horseshoe prior; Linear mixed model; Quantitative genetics.

1. Introduction

Linear mixed models (LMM) are flexible tools for modeling data from a wide range of data types in various applied fields including ecology and evolution (Bolker et al., 2009; Ives and Helmus, 2011), quantitative genetics (Lynch and Walsh, 1998) or medical research (Brown and Prescott, 2014). An important practical point when using linear mixed models is the choice of random effect components. Choosing which random grouping factors to include is vital to model appropriate dependence structures within data. This issue compounds when the possible number of random effects is large, which also leads to identifiability problems and estimation instability. With new technologies, high-resolution satellite images, high-throughput genotyping/phenotyping techniques, such contexts are now very common. For instance, in quantitative genetics, linear mixed models are commonly used for IBD-QTL mapping (George et al., 2000; van Eeuwijk et al., 2010; Tisné et al., 2015) or gene-set analyses (GSA) (Fridley and Biernacka, 2011), allowing to decompose the global genetic effect as a sum of local/specific effects related to each position/block considered on the genome. In such studies, the genetic effects are associated to random effects that can vary from tens to a few hundreds.

This work has been motivated by two original applications. The first one takes place in the IBD-QTL mapping context and aims at identifying QTLs related to the *oil palm* production (*Elaeis guineensis*, Jacq). In this application, 135 random effects associated to 135 genetic positions are considered. In the functional mapping framework, the second application aims at modeling the effects of 38 genetic markers assumed varying with time, and at selecting those involved in the dynamics of shoot growth of *Arabidopsis thaliana* (L. Heynh) (Marchadier et al., 2019). While providing an alternative modeling approach to the usual functional mapping methods based on non-parametric strategies (Ma et al., 2002), our work also allows to estimate potential dependencies between random effects.

The usual solution for dealing with identifiability and/or inference instability problems is to reduce the number of variables, using model choice procedures based on information criteria (Müller, Scealy, and Welsh, 2013). An alternative strategy relies on regularization approaches (Bickel et al.,

2006). In linear models (LM), regularization procedures have been widely studied and a large set of penalty functions has been proposed (see for example Tibshirani (1996); Desboulets (2018)). Most regularization methods, initially developed in a frequentist context, have been proposed in the Bayesian framework (Kyung et al., 2010). Prior distributions act as penalty terms in the frequentist approach. A set of priors has been extensively developed, among others: spike-and-slab (George and McCulloch, 1993), Bayesian Lasso (Park and Casella, 2008), Elastic-Net (Kyung et al., 2010), normal-gamma (Griffin et al., 2010) or horseshoe (Carvalho et al., 2009) priors.

In the LMM context, literature is less developed. The first approaches to simultaneously select fixed effects and variance components used model choice criteria (Rao and Wu, 1989; Vaida and Blanchard, 2005; Müller et al., 2013; Delattre and Poursat, 2020). As in the LM context, penalized likelihood approaches have also been developed as alternatives, especially when the number of predictors increases. For instance, Bondell, Krishna, and Ghosh (2010) and Ibrahim et al. (2011) combined penalized likelihood techniques, using adaptive lasso or smoothly clipped absolute deviation (SCAD) penalties, with the Cholesky decomposition of the random effects covariance matrix or its modified version. In the high dimensional context, Fan and Li (2012) proposed a two steps approach, while Li et al. (2018) proposed a doubly regularized estimation and selection of fixed and random effects in longitudinal data. The reparametrization of the LMM by the use of a modified Cholesky decomposition of the random effects covariance matrix has initially been put forward in a Bayesian framework by Chen and Dunson (2003). Such a technique allows to consider the standard deviations of random effects as regression parameters. In their approach, a spike-and-slab prior is used for the fixed effects and the standard deviations, with a truncated Gaussian distribution for the slab part associated to the standard deviations. Frühwirth-Schnatter and Tüchler (2008) propose a related approach, modeling directly the Cholesky decomposition elements using a spike-and-slab prior where the slab distribution is an unconstrained Gaussian distribution. As for the fixed effect selection, alternative approaches to spike and-slab priors have also been developed and compared for the selection of variance components. In particular, continuous shrinkage priors, or their mixture versions, such as Student and Cauchy distributions, Bayesian Lasso and normal-gamma priors have been studied in the context of random intercept models (Frühwirth-Schnatter and Wagner, 2011). Finally, we note that the question of shrinking variance parameters towards zero does not raise only in the LMM context. Such objectives have been studied in different statistical contexts. In structured additive regression models (Fahrmeir, Kneib, and Lang, 2004) for instance, groups of fixed effects are selected using a spike-and-slab prior on specific group variance components based on a mixture of inverse gamma distributions (Scheipl, Fahrmeir, and Kneib, 2012). In time varying parameters and state-space models, Bitto and Frühwirth-Schnatter (2019) and Cadonna, Frühwirth-Schnatter, and Knaus (2020) propose the use of double or triple gamma priors extending Normal-Gamma (Griffin et al., 2010) or more generally the scaled-Beta distribution class (Pérez et al., 2017).

In this paper, we propose to combine the horseshoe prior with its folded version to simultaneously select fixed and random effects. We study the performances of the proposed prior through the two applications and discuss results in comparison with the two commonly used alternative priors: the folded Cauchy prior and the folded spike-and-slab prior. In the second application, to model dependency structures between random slopes effects, we apply polar parametrization (Pinheiro and Bates, 1996) using sinusoidal prior on angles (Pourahmadi and Wang, 2015) to ensure symmetry and positive-definiteness of the unknown correlation matrix. The paper is organized as follows. Section 2 presents the general model, priors formulation, and the specific context associated with each of the two applications along with their dedicated models. In section 3, we present the computational aspects of the Bayesian inference in order to optimize the MCMC algorithms. In Section 4, we firstly discuss results obtained by the three priors from a statistical point of view, and we then interpret results from a biological point of view.

2. Model specification and priors formulation

2.1 *General considerations*

LMMs can be expressed in the following general form:

$$y = \boldsymbol{X}\beta + \boldsymbol{Z}\tilde{u} + \varepsilon \tag{1}$$

where y is a n-response vector, $\boldsymbol{X}$ a $n \times (p+1)$-matrix of p covariates with a first unitary column for the intercept, $\boldsymbol{Z}$ a $n \times s$ known sparse random effects design matrix associated to $\tilde{u}$ a s-vector of random effects assumed to be distributed as a multivariate Gaussian distribution with null expectation and covariance matrix denoted by $\boldsymbol{\Omega}$. In the following, $\tilde{u}$ and ε are assumed independent. Such a formulation encompasses a broad set of LMMs. Each model leads to consider different design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$, and variance matrices $\boldsymbol{\Omega}$.

For the variance component model, considering q independent random effects $\tilde{u}_l$, with c_l levels each, following a Gaussian distribution centered on zero with covariance matrix Ω_l ($l = 1, \ldots, q$), then $\boldsymbol{Z} = \bigoplus_{l=1}^{q} Z_l$ with Z_l the l^{th} random effect design matrix, $\tilde{u} = (\tilde{u}_1', \ldots, \tilde{u}_q')'$, $\boldsymbol{\Omega}$ is a block diagonal matrix where each block is the $c_l \times c_l$ matrix Ω_l ($\boldsymbol{\Omega} = bdiag(\Omega_1, \ldots, \Omega_q)$) and $s = \sum_{l=1}^{q} c_l$. Here $\bigoplus$ denotes the column concatenation operator. $\boldsymbol{Z}\tilde{u}$ can be decomposed as a sum of random effects $\boldsymbol{Z}\tilde{u} = Z_1\tilde{u}_1 + \cdots + Z_q\tilde{u}_q$. Two special cases of the variance component model may be considered. The first one is the usual variance component model where $\Omega_l = \lambda_l^2 I_{c_l}$ with I_{c_l} the identity matrix of size c_l. In this case, levels of each random effect are assumed independent. The second one is the animal model where $c_l = n$, $Z_l = I_n$ and $\Omega_l = \lambda_l^2 A_l$, where A_l is a known $n \times n$-IBD matrix.

In the random intercept and slope (RIS) context with one grouping factor with c levels and p covariates, $\boldsymbol{Z} = J \bullet \boldsymbol{X}$ where $\bullet$ is the face-splitting product (row-by-row Kronecker product, see web appendix B for more details), J corresponds to the $n \times c$-0/1-design matrix associated to the random effect and $\boldsymbol{\Omega}$ is a block diagonal matrix such as $\boldsymbol{\Omega} = I_c \otimes \Omega$, where

Ω is a $(p+1) \times (p+1)$ unknown correlation matrix related to dependencies between random intercept and slopes. It is straightforward to extend to q random effects with c_l levels each.

When p and s are large, such models (see equation 1) must be regularized. The already proposed approaches are mainly based on the Cholesky decomposition of $\boldsymbol{\Omega}$ (Chen and Dunson, 2003; Bondell et al., 2010). In this paper, we propose to use this decomposition: $\boldsymbol{\Omega} = \boldsymbol{\Lambda R \Lambda}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix and $\boldsymbol{R}$ the associated correlation matrix. A general LMM model (see equation 1) can then be reformulated as:

$$\begin{aligned} y &= \boldsymbol{X}\beta + \boldsymbol{Z\Lambda}u + \varepsilon \\ &= \left[\boldsymbol{X}, \left(u' \otimes \boldsymbol{Z}\right) P\right] \left[\begin{array}{c} \beta \\ \lambda \end{array}\right] + \varepsilon \end{aligned} \tag{2}$$

where λ is the unique diagonal elements vector of $\boldsymbol{\Lambda}$ and P is the matrix that transforms λ to $vec(\boldsymbol{\Lambda})$ (Ibrahim et al., 2011). Calculation details are presented in web appendix B. Now, u is the vector of Gaussian random effects $\mathcal{N}_s\left(0, \boldsymbol{R}\right)$. Finally, when $\boldsymbol{\Omega}$ is supposed to be unknown (RIS models), we propose to write the correlation matrix $\boldsymbol{R}$ using the polar parametrization (Pinheiro and Bates, 1996; Pourahmadi and Wang, 2015). Such an approach ensures that the correlation matrix R, sampled though the posterior distribution, is a valid symmetric and positive-definite matrix with 1's on the diagonal. More details are given in section 2.3.2 bellow. Finally, ε is a multivariate Gaussian residual vector assumed to be independent of u.

2.2 *Priors formulation*

To achieve the selection of fixed effects (β) and scale parameters (λ), we consider in this work local-global priors (Polson and Scott, 2012; Piironen and Vehtari, 2017). Such priors, initially used for the selection of fixed effects only, consist in a scale mixture of Gaussian distribution on parameters a_j subjects to selection:

$$a_j|\tau^2, \omega_j^2, \sigma^2 \sim \mathcal{N}\left(0, \sigma^2\tau^2\omega_j^2\right), \quad j = 1, \ldots, q, \tag{3}$$

where σ^2 is the residual variance, τ^2 is the global parameter while ω_j^2 are the local ones. τ^2 allows to shrink all coefficients towards zero while local parameters ω_j^2 highlight non-null parameters. Such a prior encompasses a large set of well known priors. Assuming $\omega_j^2 \equiv 1$ leads to global priors such as the well known normal-inverse-gamma (NIG) prior (Gelman et al., 2004), while assuming $\tau^2 \equiv 1$ leads to local priors such as the Laplace prior (Park and Casella, 2008), the student prior (Gelman, 2006) or the normal-gamma prior (Griffin et al., 2010) (see table 1 in web appendix A). Among local-global priors, the horseshoe prior assumes that local, as well as global parameters, are distributed from folded-Cauchy distributions (Carvalho, Polson, and Scott, 2009). The horseshoe prior has demonstrated high performances for the selection of fixed effects, comparable to the spike-and-slab prior (van Erp, Oberski, and Mulder, 2019). In this article, we propose to investigate horseshoe priors to simultaneously select fixed effects and standard deviations. We note that since λ_j is positive, the Gaussian distribution is replaced by a folded-Gaussian distribution $\mathcal{N}^+\left(0, \sigma^2\tau^2\omega_j^2\right)$. The horseshoe prior is systematically applied to fixed effect. For standard deviations, we evaluate the folded horseshoe prior (fHS) addressing two specific questions: (i) what are the performances of such priors to select random effects, and (ii) which impacts on the fixed and random effect estimations. These results are discussed relatively to the use of alternative priors: folded Cauchy (fC) and folded zero-inflated (spike-and-slab, fSS).

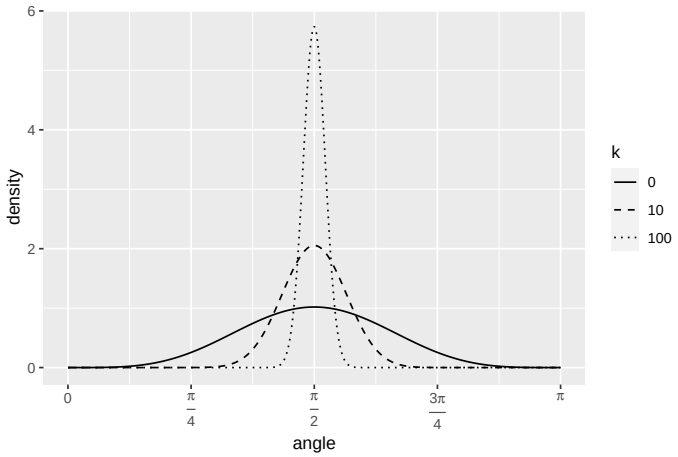


Figure 1: *Prior density distribution of the angle $\theta_{i,j}$ (see equation 5) according to different values of k: 0, 10 of 100.*

In the specific RIS context, random intercept and slopes are commonly assumed to be non-independent through an unknown correlation matrix R such that $\boldsymbol{R} = I_c \otimes R$. Different priors have been proposed (Lewandowski et al., 2009; Pourahmadi and Wang, 2015). Here, we adopt the polar parametrization introduced by Pinheiro and Bates (1996). It consists in the use of a hyperspherical parametrization of the Cholesky factors of the correlation matrix and an appropriate distribution on related angles. Pinheiro and Bates (1996) demonstrate that any correlation matrix R can be factorized as BB', with $B_{1,1} = 1,\ B_{i,1} = cos(\theta_{i,1}),\ i = 2, \ldots, q$ and

$$B_{j,i} = \begin{cases} \prod_{m=1}^{j-1} \sin(\theta_{i,m}) & \text{for } i = j, \\ \cos(\theta_{i,j}) \prod_{l=1}^{j-1} \sin(\theta_{i,l}) & \text{for } 2 \leqslant j \leqslant i-1, \end{cases} \tag{4}$$

Pourahmadi and Wang (2015) proposed the following θ's sinusoidal distribution :

$$\theta_{i,j} \propto sin(\theta)^{2k+(p+1)-j} \mathbb{1}_{0<\theta<\pi},\ i = j+1, \ldots, p, \tag{5}$$

where k is a non-negative constant. This distribution ensures that angles are centered on $\pi/2$ or equivalently that the distribution of R is centered on the identity matrix. Moreover and interestingly, parameter k can be interpreted as a shrinkage parameter (Ghosh, Mallick, and Pourahmadi, 2020). For instance, if $k = 0$ then R is distributed as a uniform distribution on the set of all $(p+1) \times (p+1)$-positive-definite correlation matrices, while if k tends to infinity, the distribution of R tends to a point mass on the unit diagonal $(p+1) \times (p+1)$-matrix (see Figure 1). In this work, k is chosen based on cross-validation procedure.

2.3 *Specific applied contexts*

2.3.1 *The oil palm dataset.* We analyse this data set within the animal model framework and give the results in section 4.1. Indeed, this first application aims at identifying the genomic positions involved in the variability of *oil palm* production traits. A total of 144 palm trees belonging to the breeding program of PalmElit, a Cirad subsidiary and leading *oil palm* breeding company (`www.palmelit.com`), were analyzed. Palm trees were genotyped with 226 molecular markers and 1,007 IBD matrices were estimated on a grid of 3 centimorgan (cM) along the genome (Tisné et al., 2015). Each genetic position l is associated to a random effect u_l with a variance equal to $\lambda_{u_l}^2$ and a correlation matrix A_l equal to the identity-by-descent (IBD) matrix (see equation 6). Then, the identification of the genomic positions is equivalent to the variance components selection. Due to the genetic characteristics of the population, *i.e.* a moderate number of individuals and generations, a subset of 135 genetic positions, spaced 10 cM apart, was considered to avoid a redundant information between consecutive genetic positions. In the next section, we will present the results for the bunch number trait.

As previously explained in subsection 2.1, the animal model can be formulated as follows:

$$y = \mu + u_1\lambda_1 + \cdots + u_q\lambda_q + \varepsilon, \tag{6}$$

where μ is an intercept and u_l is now assumed to follow a Gaussian distribution $\mathcal{N}_n(0, A_l)$, $l = 1, \ldots, q$ and λ_l is the standard deviation associated to u_l. Finally, the matrix version of the animal model is given by:

$$y = \mu + U\lambda + \varepsilon, \tag{7}$$

where U is a $n \times q$-matrix of the concatenation of the random effects $U = \bigoplus_{l=1}^{q} u_l$.
In a fully Bayesian framework, the intercept μ is supposed to be proportional to one and the residual variance σ^2 is supposed to follow an inverse-gamma distribution $\mathcal{IG}(s_{\sigma^2}, r_{\sigma^2})$ (shape and rate parametrization). The Bayesian hierarchical model is presented in web appendix C.

2.3.2 *The arabidopsis thaliana dataset.* We analyse this data set within the RIS model framework and give the results in section 4.2. Indeed, in this second application, we are interested by disentangling the evolution over time of the complex genetic architecture of shoot growth of *Arabidopsis thaliana* (L. Heynh). Data consists of leaf compactness phenotypic trait measured over $T = 21$ time points on $n = 358$ individuals. We use genetic covariates X containing $p = 38$ markers (Marchadier et al., 2019; Heuclin et al., 2020). In the RIS model framework, we consider time as the grouping factor. This model is an alternative approach to the usual non-parametric functional mapping (Ma et al., 2002). It can be expressed as follows:

$$y_{i,t} = x_i\beta + x_i\tilde{u}_t + \alpha_i + \varepsilon_{i,t} \tag{8}$$

where $y_{i,t}$ is the observation of individual i at time t ($i = 1, \ldots, n$ and $t = t_1, \ldots, t_T$). x_i is a $(p+1)$-row vector of p genetic markers (constant over time) associated to the i^{th} individual. The first element is fixed to one and is related to the intercept. β is a $(p+1)$-vector of fixed effects, $\tilde{u}_t$ a $(p+1)$-vector of random intercept and slopes effects assumed to follow a Gaussian distribution $\mathcal{N}_{p+1}(0, \Lambda R \Lambda)$, where Λ is an unknown $(p+1) \times (p+1)$-diagonal matrix of standard deviation and R is an unknown $(p+1) \times (p+1)$-correlation matrix. In this application, R is assumed to be block diagonal where each block is related to one chromosome (Ghosh et al., 2020). The random intercept is also assumed independent from the random slopes. α_i is a Gaussian individual random effect ($\mathcal{N}\left(0, \sigma_\alpha^2\right)$) not subject to selection. $\varepsilon_{i,t}$ corresponds to the residual part such that $\varepsilon_i = (\varepsilon_{i,t_1}, \ldots, \varepsilon_{i,t_T})$ is distributed from a multivariate Gaussian distribution $\mathcal{N}_{t_T}(0, \sigma_e^2\Gamma)$ where Γ is a $t_T \times t_T$-correlation matrix of a first-order autoregressive structure with unknown parameter ρ.

Let $y = (y_{t_1}', \ldots, y_{t_T}')'$ be the concatenation of all measurements over time for all individuals where $y_t = (y_{1,t}, \ldots, y_{n,t})'$. Since the genetic information varies between individuals but is constant over time, $\boldsymbol{X}$ can be simplified such that $\boldsymbol{X} = (\mathbb{1}_{t_T} \otimes X)$ where X is the $n \times (p+1)$-matrix containing the p genetic markers (and the intercept) of all individuals. Matrix J is here equal to $I_{t_T} \otimes \mathbb{1}_n$. The random effects design matrix $\boldsymbol{Z}$ can also be simplified as:

$$\boldsymbol{Z} = I_{t_T} \otimes X. \tag{9}$$

Calculation details are presented in web appendix B. Finally, $\boldsymbol{\Lambda}$ is decomposed as $I_{t_T} \otimes \Lambda$ and $\boldsymbol{R} = I_{t_T} \otimes R$. P is the matrix that transforms λ to $\mathbb{V}ec(\Lambda)$ (or equivalently $\Lambda = diag(\lambda)$ and $\mathbb{V}ec(\Lambda) = P\lambda$). Then, as proposed in section 2.1, this model can be expressed as:

$$y = [\mathbb{1}_{t_T} \otimes X, (U' \otimes X)P] \begin{bmatrix} \beta \\ \lambda \end{bmatrix} + D\alpha + \varepsilon. \tag{10}$$

Calculation details are presented in web appendix B. U is a $(p+1) \times t_T$-matrix of the collection of the t_T reparametrized vectors of random intercept and slopes associated to each time $U = \bigoplus_{k}^{t_T} u_k$ (U follows a matrix Gaussian distribution $\mathcal{MN}_{(p+1)\times t_T}(0, R, I_{t_T})$). $D = \mathbb{1}_{t_T} \otimes I_n$ is the design matrix associated to the individual random effects. $\varepsilon = (\varepsilon_1', \ldots, \varepsilon_{t_T}')'$ is the concatenation of all residuals over time and for all individuals, where $\varepsilon_t = (\varepsilon_{1,t}, \ldots, \varepsilon_{n,t})'$ is a n-vector of residuals associated to all individuals at time t. ε is supposed to follow a Gaussian distribution centered on zero with covariance $\sigma^2\boldsymbol{\Gamma}$ where $\boldsymbol{\Gamma} = \Gamma \otimes I_n$. While the introduction of time random effects allows to capture dependencies between observations within the same time measurement and to model dynamics of genetic effects through the dependence structure. Moreover, introducing a random individual effect combined with a specific residual correlation structure allows to take into account dependencies between measurements over time.
Finally, in a fully bayesian framework, the variance associated to the individual random effect σ_α^2 is supposed to follows an inverse-gamma distribution $\mathcal{IG}(s_\alpha, r_\alpha)$, the residual variance σ^2 is supposed to follows an inverse-gamma distribution $\mathcal{IG}(s_{\sigma^2}, r_{\sigma^2})$ and the autoregressive parameter ρ is supposed to follows a uniform distribution $\mathcal{U}(-1, 1)$. The Bayesian hierarchical model is presented in web appendix C.

3. Computational aspects of the Bayesian inference

Both applications raise computational challenges mainly due to the number of parameters, dependency structures but also the number of latent variables. In the animal model,

the number of parameters is equal to 137 (the intercept, 135 standard deviations and the residual variance) and 19440 unobserved latent variables should be updated (number of elements of U). In the RIS model, the number of parameters is equal to 214 (39 fixed effects, 39 standard deviations, one individual and one residual variances, one autoregressive parameter and 132 angles) and 1176 unobserved latent variables should be updated (number of elements of U). While the animal model looks simpler (with a simple additive form) compared to the RIS model (involving complex unknown dependency structures), both complexities are high and the difference between them are not clear. MCMC algorithms have to be appropriately designed for optimization purposes. These optimizations are achieved by reparametrizing standard deviations and by proposing an efficient sampling scheme to avoid inversion of dense posterior covariance matrices.

The first difficulty relies on sampling the standard deviations λ_j according to their full conditional distributions. These distributions are proportional to a non-central multivariate folded-Gaussian distribution. Such a distribution does not have a closed form and cannot easily be sampled. To overcome this challenge, following Gelman's work, we propose to reparametrize λ_l as $sign(\xi_l)\xi_l$ where ξ_l is a parameter which can be positive or negative. It follows that ξ_l is distributed from a Gaussian distribution (and not from a folded-Gaussian): $\xi_l \sim \mathcal{N}(0, \sigma^2\tau^2\omega_l^2)$. Thus, to sample a standard deviation λ_l from its full conditional distribution $p(\lambda_l|.)$, we can more simply (i) sample ξ_l from its full conditional distribution $p(\xi_l|.)$ which is a Gaussian distribution and then (ii) compute $\lambda_l = sign(\xi_l)\xi_l$. Demonstrations are provided for both models in web appendix C.
High dimensionality causes a second issue to arise. Indeed, at each iteration of the MCMC algorithms, the random effects sampling step involves either the inversion of q $n \times n$-dense covariance matrices for the animal model (complexity $O(qn^3)$) or one $t_T(p+1) \times t_T(p+1)$-matrix for the RIS model (complexity in $O((t_T(p+1))^3)$). However, these covariance matrices have the form $\Sigma_u = (aA + bI)^{-1}$ (after a reparametrization under the RIS model), which is the inverse of the addition of a dense matrix A and a unit identity matrix (a and b are scalars, a, b and A depend on the specific context). This form is very convenient because SVD of the dense matrix A can be used to compute the Cholesky decomposition of Σ_u efficiently. Thus, to sample a random effect u from its full conditional distribution of the form $p(u|.) \sim \mathcal{N}(\Sigma_u h, \Sigma_u = (aA + bI)^{-1})$ (where h is a vector), we can (i) compute $A = WDW$, the SVD of A, where W is an orthogonal matrix of singular vectors and D is a diagonal matrix of singular values, (ii) compute $L = W(aD+bI)^{-1/2}$, the Cholesky decomposition of Σ_u, (iii) sample z from a standard Gaussian distribution and then (iv) compute $u = L(z + L'h)$. For the animal model context, dense matrices A are known IBD matrices and SVD can be computed only once at the beginning of the algorithm. Thus, the complexity of the sample scheme is in $O(qn^2)$. For the RIS model context, matrix A is unknown. However, using specific reformulations of matrices $\boldsymbol{B}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Omega}$ as Kronecker products, matrix A can be reformulated as a Kronecker product of two matrices and SVD of A can be computed using SVDs of both matrices. Thus, the complexity of the sample scheme is $O((p+1)^3)$ if $p+1 > t_T$, $O(t_T^3)$ otherwise. Such algebraic simplifications considerably accelerate MCMC algorithms.
The third challenge, specific to the RIS context, is related to the sampling of fixed effects, β, and of random individual effects, from their full conditional Gaussian distributions. Again, algebraic simplifications based on reformulations of $\boldsymbol{X}$, D and $\boldsymbol{\Omega}$ matrices as Kronecker products allow the simplification of posterior covariance matrices and highly increase the speed of MCMC algorithms.

All these manipulations allow to deal with full conditional posterior distributions and to propose an efficient Gibbs sampler algorithm for the animal model (see web appendix C) or a faster Metropolis-within Gibbs algorithm in the RIS context. A Metropolis-Hasting step is proposed to update angle parameters associated to the correlation matrix between random intercept and random slopes (see web appendix C). All results presented in the next section, are based on 3 MCMC chains initialized at random starting values, each with 50,000 iterations, a burn-in of 10,000 iterations and a thinning of ten. All output statistics are based on the pooled 120,000 posterior samples. The Gelman and Rubin's Potential Scale Reduction Factors (PSRF) statistics (Gelman et al., 1992) is used to evaluate chains convergence. For standard deviation parameters, estimation is based on the posterior median.

4. Results

In the next subsections, we show that the fHS prior distribution is efficient to infer and select fixed effects and variance component parameters. As expected, when the number of parameters is large compared to the number of observations (first application), the fC prior does not shrink enough parameters towards zero, leading to clear over-fitting. We highlight that fHS and fSS priors perform similarly to select variance components as it has been shown in the multivariate linear context. In the second application, where the number of parameters is low compared to the number of observations, we show that the three priors perform well and no criteria, based on cross-validation procedure, allows to favour one more than the other.

4.1 *The oil palm dataset (animal model)*

Statistical results. Considering the algorithm does not converge, we adopt a fC prior for standard derivations as an alternative to the commonly used inverse-Gamma prior for variance parameters. The fC prior is not dedicated to selection but should allow for better model regularization than the inverse-Gamma. However, results show that even this prior does not shrink enough towards zero leading to a systematic bias in the estimations (see figure 2), with posterior medians varying around 0.17. The fC prior leads to over-fitting, which is particularly noticeable when analyzing the residual variance: it is estimated around zero (see figure 3) and it has a notable impact on the converge of the Gibbs

sampling algorithm by leading to a PSRF greater than 2 for a few continuous parameters. Comparatively, the fHS prior exhibits a very different behavior. It shrinks towards zero most standard deviations and let some of them be far from zero. Thus, it enables the selection of random effects and improves the MCMC convergence (PSRFs are always close to one for all continuous parameters). In this application, we propose the selection of variance components representing at least 0.05 percent of the total phenotypic variance (0.0023 or equivalently a threshold of 0.048 on standard deviations). This threshold leads to select 10 random effects (see figure 2 and table 1). The use of the fSS prior, with marginal inclusion posterior probability threshold equal to 0.1, leads to the selection of 7 standard deviations (see figure 2 and table 1). Six markers are commonly selected by fHS or fSS priors. The selection of variance components is comparable. Such similarities have already been observed in the selection of the fixed effects. Interestingly, the four markers selected using the fHS prior that are not selected using the fSS prior, have also been reported to impact phenotypic variability in different studies.

Thus, the fHS prior seems to efficiently shrink towards zero the non-relevant random effects while properly estimating relevant parameters. Moreover, it presents better computational performances than the fSS prior. Indeed, computational time for the fHS prior is twice faster than the fSS prior (40 and 80 minutes respectively for 50,000 iterations). Then, the fHS prior should clearly be promoted in a high dimensional quantitative genetic context.

Biological interpretation. We turn to biological interpretations by focusing on the results obtained by the fHS prior. Comparing with the Tisné et al. (2015) study that analyzed the same data using maximum likelihood ratio tests combined with a forward approach, all but one position identified in the former study were found. Surprisingly, the common positions were all identified at the 0.1% threshold selection, but none for the 0.05% selection. This could be due to the genetic design of the population studied derived from a breeding pedigree with unequal contributions of contrasted genetic groups: among the 144 palm trees, 73% were from La Mé (LM) genetic background, 15% from Yangambi (YBI) and 3% from their combination. Several other studies analyzed both genetic backgrounds with different genetic designs and common genetic markers. Billotte et al. (2010), with 25% LM and 25% YBI, found four common positions including two at the 0.05% threshold, Ukoskit et al. (2014), with 50% YBI, four common positions including two at the 0.05% threshold and Seng et al. (2016), three common positions including two at the 0.05% threshold. The ability of selecting positions corresponding to YBI QTL that were segregating in a minor fraction of the population indicates that the method evaluated in this study performs well even with unbalanced genetic designs and rare allele segregations. This result highlights that a multivariate approach increases the power of detection of subtle effects.

4.2 *The arabidopsis thaliana dataset (RIS model)*

In this second application, markers are labelled by their chromosome numbers and their positions (within the whole dataset of 538 markers) separated by a dash, such that marker 1-2 corresponds to the second position on the first chromosome. This notation was used by Heuclin et al. (2020) and will be used for comparison purposes. We compare our results with those of Heuclin et al. (2020), which used a non parametric functional mapping method, but also with the approach of Marchadier et al. (2019), which is based on a stepwise strategy. For the three approaches, PSRF statistics of all continuous parameters are lower than 1.1 indicating chains' convergence.

Selection of fixed effects.
Fixed effects are considered selected if zero does not belong to their credible intervals, leading to three selected markers (2-32, 2-62 and 5-104). Whatever the prior for variance components, the selection of fixed effects using HS priors performs well and provides the same results (see table 2).

Selection of variance components and impact of the correlation matrix between random effects.
Using fC or fHS priors, a threshold representing 0.1 percent of the total phenotypic variance (0.0068, or equivalently a threshold of 0.083 on standard deviations) is used. For the fSS prior, a threshold of 0.5 is considered. While the time random intercept is systematically included in the model, the number of selected random slopes varies according to priors but also to the correlation matrix prior we use between random effects. In particular, selection appears sensitive to the assumption we make on the angle shrinkage global parameter k which plays an important role on the correlation prior distribution (see equation 5). For example, when k tends to infinity (identity case) the number of selected random slopes is equal to 13, 10 and 10 for fC, fHS and fSS priors respectively. When k is fixed to one, numbers increase to 24, 18 and 11 respectively. To choose the most appropriate k value for each prior, a 10 cross-validation scheme is performed. The log pointwise predictive density (*lppd*, Gelman, Hwang, and Vehtari (2014)), related to $k = 1, 3, 5, 7, 10$ and for independence assumption, are reported on table 3. Small differences can be observed. For parsimony reasons, random effects are assumed independent. In this example, the fC prior leads to select only few more markers (13) than the fHS or fSS (10) (see figure 4 and table 2). Differences between priors are less pronounced than in the animal context where the use of the fC prior leads to an estimation of the residual variance close to zero and then to over-fitting problems. Here, residual variance is slightly lower using the fC prior (2), than using fHS or fSS priors (2.5). These differences cannot be used to evidence one prior rather than another. To decide if a prior can be promoted, we compare *lppd* between models (see table 3). But results are very close and no clear conclusion can be drawn from these results.

Selection of variance components and impact of the residual correlation.
The selection of fixed effects is not impacted by the residual correlation matrix, on the contrary this dependency structure impacts the selection of variance components. Such conclusions have already been observed in the functional mapping context (Ma et al., 2002; Li and Wu, 2010; Heuclin et al., 2020). When we compare selection of random effects taking into account an AR(1) residual correlation structure or assuming independence between residuals, the number of selected markers differs. It considerably increases with the

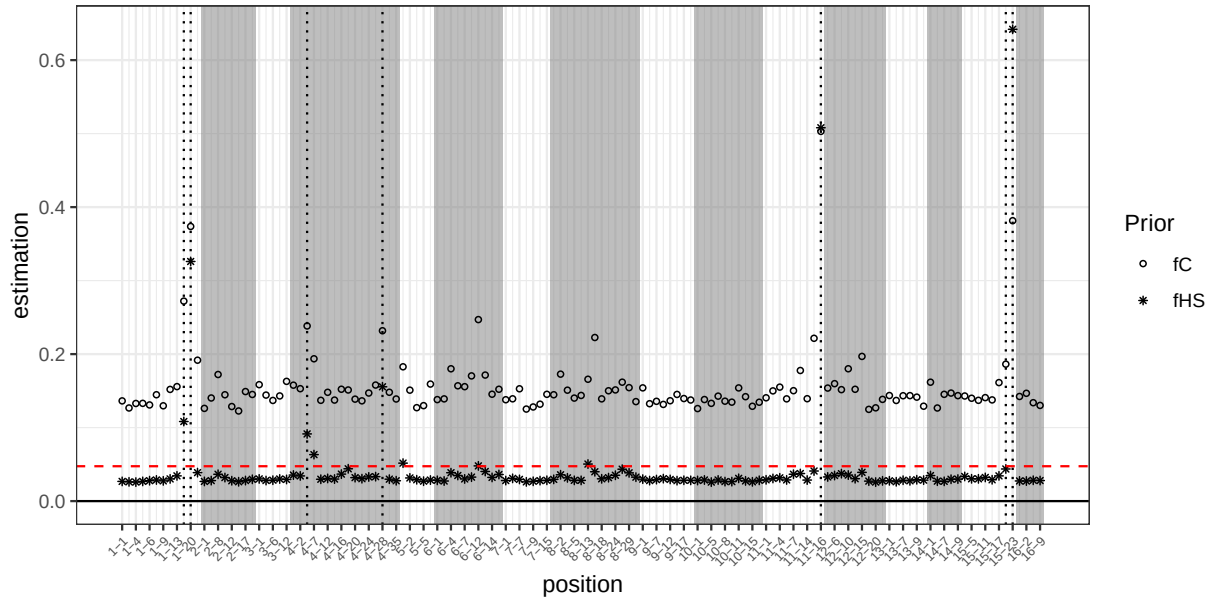


Figure 2: *Posterior median of standard deviation parameters λ_l for folded horseshoe (fHS) and folded Cauchy (fC) priors on the oil palm trees dataset. Vertical dotted lines correspond to the selected positions using the fSS prior with posterior marginal probability of inclusion upper than a threshold of 0.1. The horizontal red dashed line corresponds to a threshold of 0.048 which is the root of 0.05% of the response variance. The alternated white and grey areas delimit the 16 chromosomes.*

Table 1: *Selected standard deviation parameters λ_l for the oil palm trees dataset using folded horseshoe (fHS) and folded spike-and-slab (fSS) priors.*

Chromosome	**1**	**4**	**5**	**6**	**8**	**11**	**15**	**nb**
HS	1-17, 1-20	4-5, 4-7, 4-28	5-1	6-12	8-13	11-16	15-23	10
SS	1-17, 1-20	4-5, 4-28				11-16	15-21, 15-23	7

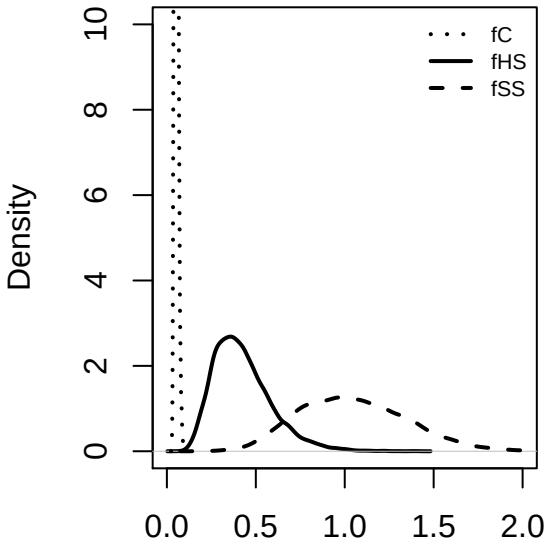


Figure 3: *Posterior density of the residual variance parameter for folded Cauchy (fC), folded horseshoe (fHS) and folded spike-and-slab (fSS) priors on the oil palm trees dataset.*

independence assumption, leading to potential over-fitting problems. Indeed, the 10 cross-validation *lppd*, considering an AR(1) residual correlation structure and the fC prior, is equal to -924, while considering an independent residual structure, it is equal to -1148. Here, results are clear and the residual correlation structure has to be included in the model.

Comparison with previous studies.
The initial study identified eight markers using the last time measurement and a forward likelihood ratio test approach (Marchadier et al., 2019). In a recent work, Heuclin et al. (2020) reanalyses this data proposing a non-functional mapping technique combined with group spike-and-slab and taking into account the full phenotypic profile over time. They identified the same eight markers but also highlighted five more effects. In our current analysis, all positions already identified by the previous approaches are selected except two on chromosome three (3-1 and 3-25), compared to Heuclin et al.'s approach. And we select two extra positions (2-47 and 5-46). Moreover, in our approach, decomposing effects as fixed and random allows to more precisely dissociate the type of effects (null, constant or varying effects). For instance, the position 5-104 selected as random effect and varying over time in Marchadier et al. or Heuclin et al. is mostly identified as fixed effect (see table 2). Finally, comparing the *lppd* statistics

Table 2: *Selection of fixed effects β and scale parameters λ on arabidopsis thaliana dataset using HS-fC, HS-fHS and HS-fSS approaches. Alternative methods proposed by Marchadier et al. (2019) and Heuclin et al. (2020) are also indicated.*

	Chromosome	1	2	3	4	5	nb
Marchadier et al. (2019)		1-20	2-62	3-3, 3-91	4-45	5-76, 5-104	8
Heuclin et al. (2020)		1-1, 1-20, 1-79, 1-97, 1-110	2-62	3-1, 3-25, 3-91	4-45	5-33, 5-76, 5-104	13
Fixed effects	HS-fC		2-32, 2-62			5-104	3
	HS-fHS		2-32, 2-62			5-104	3
	HS-hSS		2-32, 2-62			5-104	3
Scale parameters	HS-fC	1-1, 1-20, 1-79, 1-97, 1-110	2-47, 2-62	3-91	4-45	5-33, 5-46, 5-76, 5-104	13
	HS-fHS	1-1, 1-20, 1-79, 1-97, 1-110		3-91	4-45	5-33, 5-46, 5-76	10
	HS-fSS	1-1, 1-20, 1-79, 1-97, 1-110		3-91	4-45	5-33, 5-46, 5-76	10

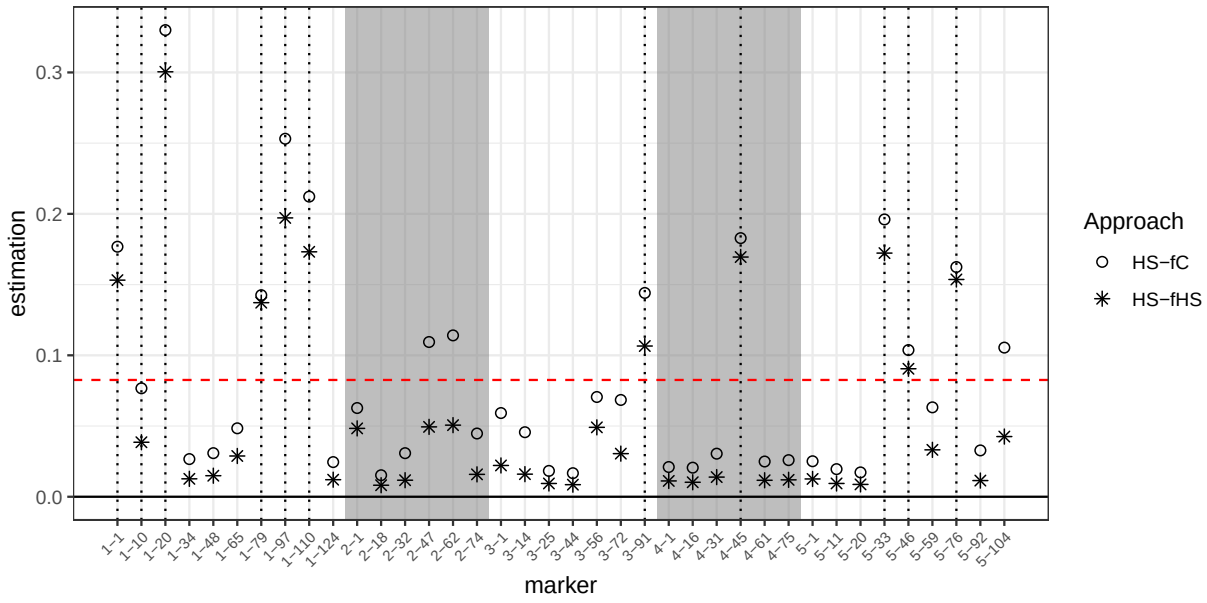


Figure 4: *Posterior median of standard deviation parameters λ_l on arabidopsis thaliana dataset. Bullets black and blue correspond to the HS-fHS and HS-fC approaches. Vertical red dotted lines correspond to the selected positions using the HS-fSS approach with posterior marginal probability of inclusion upper than a threshold of 50%. The horizontal red dashed line corresponds to a threshold of 0.083 which is the root of 0.1% of the response variance. The alternated white and gray areas delimit the 5 chromosomes.*

Table 3: *Log pointwise predictive density, considering either an unknown RIS correlation matrix with different fixed shrinkage parameters k, or an identity matrix.*

	$k=1$	$k=3$	$k=5$	$k=7$	$k=10$	Identity
HS-fC	-925	-923	-923	-923	-923	-925
HS-fHS	-923	-924	-925	-924	-925	-926
HS-fSS	-925	-925	-925	-925	-926	-927

using 10 cross-validation, favours the RIS model (-925 for the HS-fC approach with $R = I_{39}$) to the VCM model (-931).

5. Conclusion

In this paper, we show that the folded horseshoe prior should be promoted as a prior distribution for regularization in linear mixed models. Based on two real applications, we demonstrate that the folded horseshoe prior seems insensitive to high dimensional problems and leads to unbiased estimation even in low dimension. In the first example, where the number of parameters is close to the number of observations, the folded horseshoe prior shows advantages compared to the folded Cauchy and to the folded spike-and-slab priors. In particular, where the folded Cauchy prior does not allow to shrink parameters towards zero inducing a clear overfitting, the folded horseshoe prior performs well. Compared to the folded spike-and-slab prior, the folded horseshoe prior

presents similar effectiveness in terms of selection but a much greater computational efficiency. In the second application, where the number of observations is much greater than the number of parameters, no prior seems to take advantage. Such results observed in multivariate linear regression (van Erp et al., 2019) can then be extended to the linear mixed model framework. However, the folded horseshoe prior does not lead to biased estimations or under or over-fitting of models compared to the two other priors. Be that as it may, we recommend to use local-global priors.

We also propose a polar reparametrization of the model random effect correlation matrix. This approach has received little attention in the past few decades. While Pourahmadi and Wang were the first to develop a prior to generate high-dimensional random correlation matrix, Ghosh et al. were the first to infer, in a Bayesian framework, a correlation matrix in a longitudinal context. In this article, we show how this approach can be used to infer RIS correlation matrix. We also show that assuming independence or not can impact variance components selection. However, the number of parameters (angles) is equal to the number of elements of the sub-diagonal correlation matrix. Appropriate priors for the selection of angles such as considered by Ghosh et al. (2020) should be studied in combination with standard deviations shrinkage priors.

From a biological point of view, in the palm oil context, the folded horseshoe prior allows to identify positions which were segregated in a minor fraction of the population due to the unbalanced genetic design, while the frequentist stepwise selection approach considered by Tisné et al. (2015) does not. In the *Arabidobsis* context, as already noticed by Heuclin et al. (2020), we show that a longitudinal approach allows a better detection of relevant markers compared to an approach that analyzes a single time point as proposed by Marchadier et al. (2019). Both applications highlight that multivariate approaches increase the statistical power.

Acknowledgments:

F. Mortier and C. Trottier were supported by the GAMBAS project funded by the French National Research Agency (ANR-18-CE02-0025). M. Denis was fully supported by the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 840383. We thank all people from Cirad/PalmElit (France) who planned this trial. We acknowledge P.T. Socfin Indonesia (Indonesia) for planting, observing and collecting data, and authorizing use of the phenotypic data for this study.

References

Bickel, P. J., Li, B., Tsybakov, A. B., van de Geer, S. A., Yu, B., Valdés, T., Rivero, C., Fan, J., and van der Vaart, A. (2006). Regularization in statistics. *Test* **15,** 271–344.

Billotte, N., Jourjon, M.-F., Marseillac, N., Berger, A., Flori, A., Asmady, H., Adon, B., Singh, R., Nouy, B., Potier, F., et al. (2010). Qtl detection by multi-parent linkage mapping in oil palm (elaeis guineensis jacq.). *Theoretical and Applied Genetics* **120,** 1673–1687.

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

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H., and White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution* **24,** 127–135.

Bondell, H. D., Krishna, A., and Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics* **66,** 1069–1077.

Brown, H. and Prescott, R. (2014). *Applied Mixed Models in Medicine, Third Edition*. John Wiley Sons, Ltd.

Cadonna, A., Frühwirth-Schnatter, S., and Knaus, P. (2020). Triple the gamma—a unifying shrinkage prior for variance and variable selection in sparse state space and tvp models. *Econometrics* **8,**.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, pages 73–80.

Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics* **59,** 1069–1077.

Delattre, M. and Poursat, M.-A. (2020). An iterative algorithm for joint covariate and random effect selection in mixed effects models. *The International Journal of Biostatistics* **1,**.

Desboulets, L. D. D. (2018). A review on variable selection in regression analysis. *Econometrics* **6,** 45.

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. (2012). Variable selection in linear mixed effects models. *Annals of Statistics* **40,** 2043–2068.

Fridley, B. L. and Biernacka, J. M. (2011). Gene set analysis of snp data: benefits, challenges, and future directions. *European Journal of Human Genetics* **19,** 837–843.

Frühwirth-Schnatter, S. and Tüchler, R. (2008). Bayesian parsimonious covariance estimation for hierarchical linear mixed models. *Statistics and Computing* **18,** 1–13.

Frühwirth-Schnatter, S. and Wagner, H. (2011). Bayesian variable selection for random intercept modeling of gaussian and non-gaussian data.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis* **1,** 515–534.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and computing* **24,** 997–1016.

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

George, A. W., Visscher, P. M., and Haley, C. S. (2000). Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach. *Genetics* **156,** 2081–2092.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical*

Association **88,** 881–889.

Ghosh, R. P., Mallick, B., and Pourahmadi, M. (2020). Bayesian estimation of correlation matrices of longitudinal data. *Bayesian Analysis* .

Griffin, J. E., Brown, P. J., et al. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian analysis* **5,** 171–188.

Heuclin, B., Mortier, F., Trottier, C., and Denis, M. (2020). Bayesian varying coefficient model with selection: An application to functional mapping. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* .

Ibrahim, J. G., Zhu, H. T., Garcia, R. I., and Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics* **67,** 495–503.

Ives, A. R. and Helmus, M. H. (2011). Generalized linear mixed models for phylogenetic analyses of community structure. *Ecological Monograph* **81,** 511–525.

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

Lewandowski, D., Dorota Kurowick, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis* **100,** 1989–2001.

Li, Y., Wang, S., Song, P. X., Wang, N., Zhou, L., and Zhu, J. (2018). Doubly regularized estimation and selection in linear mixed-effects models for high-dimensional longitudinal data. *Stat Interface.* **11,** 721–737.

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

Loudet, O. (2018). Raw phenotypic data obtained on the arabidopsis rils with the phenoscope robots (marchadier, hanemian, tisnÃ© et al., 2019).

Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer.

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

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

Müller, S., Scealy, J. L., and Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science* **28,** 135–167.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association* **103,** 681–686.

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.

Piironen, J. and Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics* **11,** 5018–5051.

Pinheiro, J. C. and Bates, D. M. (1996). Unconstrained parametrizations for variance-covariance matrices. *Statistics and computing* **6,** 289–296.

Polson, N. G. and Scott, J. G. (2012). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Bayesian Analysis* **7,** 887–902.

Pourahmadi, M. and Wang, X. (2015). Distribution of random correlation matrices: Hyperspherical parameterization of the cholesky factor. *Statistics & Probability Letters* **106,** 5–12.

Rao, R. and Wu, Y. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika* **76,** 369–374.

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.

Seng, T.-Y., Ritter, E., Saad, S. H. M., Leao, L.-J., Singh, R. S. H., Zaman, F. Q., Tan, S.-G., Alwee, S. S. R. S., and Rao, V. (2016). Qtls for oil yield components in an elite oil palm (elaeis guineensis) cross. *Euphytica* **212,** 399–425.

Team, R. C. et al. (2013). R: A language and environment for statistical computing.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58,** 267–288.

Tisné, S., Denis, M., Cros, D., Pomiès, V., Riou, V., Syahputra, I., Omoré, A., Durand-Gasselin, T., Bouvet, J.-M., and Cochard, B. (2015). Mixed model approach for ibd-based qtl mapping in a complex oil palm pedigree. *BMC genomics* **16,** 1–12.

Ukoskit, K., Chanroj, V., Bhusudsawang, G., Pipatchartlearnwong, K., Tangphatsornruang, S., and Tragoonrung, S. (2014). Oil palm (elaeis guineensis jacq.) linkage map, and quantitative trait locus analysis for sex ratio and related traits. *Molecular breeding* **33,** 415–424.

Vaida, F. and Blanchard, S. (2005). Conditional akaike information for mixed-effects models. *Biometrika* **92,** 351–370.

van Eeuwijk, F. A., Boer, M., Totir, L. R., Bink, M., Wright, D., Winkler, C. R., Podlich, D., Boldman, K., Baumgarten, A., Smalley, M., et al. (2010). Mixed model approaches for the identification of qtls within a maize hybrid breeding program. *Theoretical and Applied Genetics* **120,** 429–440.

van Erp, S., Oberski, D. L., and Mulder, J. (2019). Shrinkage priors for bayesian penalized regression. *Journal of Mathematical Psychology* **89,** 31–50.

Supporting Information

Web appendix A, B and C, referenced in Section 1, 2 and 3, are available with this paper at the Biometrics website on Wiley Online Library. Algorithms for animal and RIS models are available in the R language (Team et al., 2013) on GitHub `https://github.com/Heuclin/variance_component_selection`. The *oil palm* dataset is available on request. For the *arabidopsis thaliana* dataset, 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`.

Received May 2021. *Revised* —-. *Accepted* —-.

Supporting Informatin for "Continuous shrinkage priors for fixed and random effects selection in linear mixed models: application to genetic mapping" by B. Heuclin, M. Denis, C. Trottier, S. Tisné and F. Mortier

May 12, 2021

Web Appendix A: Bayesian global-local shrinkage priors

Table 1: *Bayesian global-local shrinkage priors. p is the number of variables subject to selection.*

Prior	**Prior sur τ^2**	**Prior sur ω_j^2**	**R package**
NIG	$\tau^2 \sim \mathcal{IG}(s, r)$	$\omega_j^2 = 1$	BGLR (Pérez and de Los Campos, 2014) bayesreg
Student(d, v)	$\tau^2 = 1$	$\omega_j^2 = \mathcal{IG}(d/2, v\ d/2)$ $v \sim \mathcal{G}(s, r)$	BGLR (Pérez and de Los Campos, 2014)
NE (Park and Casella, 2008)	$\tau^2 \sim \mathcal{IG}(a, b)$	$\omega_j^2 \sim \mathcal{E}xp(1/2)$	BGLR (Pérez and de Los Campos, 2014) bayesreg (Makalic and Schmidt, 2016)
NEG (Griffin and Brown, 2007)	$\tau^2 = 1$	$\omega_j^2 \sim \mathcal{E}xp(z_j/2)$ $z_j \sim \mathcal{G}(a, b)$	
NGDP (Armagan, Dunson, and Lee, 2013)	$\tau^2 = 1$	$\omega_j^2 \sim \mathcal{E}xp(\lambda_j^2/2)$ $\lambda_j \sim \mathcal{G}(\alpha, \eta)$	
DL (Bhattacharya et al., 2015)	$\tau \sim \mathcal{G}(a\ p, 1/2)$	$\omega_j^2 = \psi_j \phi_j^2$ $\psi_j \sim \mathcal{E}xp(1/2)$ $\phi \sim \mathcal{D}ir(a, \ldots, a)$	dlbayes (Zhang and Li, 2018)
R2D2 (Zhang et al., 2020)	$(\tau^2)^2 \sim \mathcal{G}(b, 1)$	$\omega_j^2 = \psi_j (\sigma^2 \lambda_j / 2)^{1/2}$ $\psi_j \sim \mathcal{E}xp(1/2)$ $\lambda_j \vert \xi \sim \mathcal{G}(a, 1)$	
NTPB (Armagan, Clyde, and Dunson, 2011)	$\tau^2 \sim \pi(\tau^2)$	$\omega_j^2 \sim \mathcal{B}'(a, b)$	
NG (Griffin et al., 2010)	$\tau^2 = 1$	$\omega_j^2 \sim \mathcal{G}(\lambda, \gamma)$	
NGG (Griffin et al., 2017)	$\tau^2 = 1$	$\omega_j^2 \sim \mathcal{G}(\lambda, \gamma_j)$ $\gamma_j \sim \mathcal{G}(a, b)$	
HS (Carvalho et al., 2009) (Carvalho, Polson, and Scott, 2010)	$\tau \sim \mathcal{C}^+(0, 1)$	$\omega_j \sim \mathcal{C}^+(0, 1)$	bayesreg (Makalic and Schmidt, 2016) horseshoe (van der Pas et al., 2016) fastHorseshoe (Hahn, He, and Lopes, 2016)
HS+ (Bhadra et al., 2017)	$\tau \sim \mathcal{C}^+(0, 1)$	$\omega_j \sim \mathcal{C}^+(0, \eta_i)$ $\eta_i \sim \mathcal{C}^+(0, 1)$	bayesreg (Makalic and Schmidt, 2016)
NBP (Bai and Ghosh, 2019)	$\tau^2 = 1$	$\omega_j^2 \sim \mathcal{B}'(a_p, b)$ $a_p \sim \pi(a_p) \mathbb{1}_{[1/p < a_p < 1]}$	NormalBetaPrime (Bai and Ghosh, 2019)

Web Appendix B: Details of calculations

The different calculations present in the article involve Kronecker product ("$\otimes$"), Face-splitting product ("$\bullet$") and element-wise multiplication ("$\circ$"). Face-splitting product of two matrices A and B (with same number of rows) is a row-by-row Kronecker products of two matrices and can be expressed as (Eilers and Marx, 2003):

$$A \bullet B = (A \otimes \mathbb{1}'_{n_B}) \circ (\mathbb{1}'_{n_A} \otimes B), \tag{1}$$

where $\mathbb{1}_v$ is unit vector of length v with $v = n_A$ or n_B, the column number of matrices A and B respectively. Before to detail the different calculations, we remind some useful properties of the Kronecker product.

Let A, B, C and D be matrices.

1. Reparametrization:
$$(B' \otimes A)vec(C) = vec(ACB),$$
where $vec(C)$ denotes the vectorization of the matrix C, formed by stacking the columns of C into a single column vector.

2. The inversion of Kronecker product:
$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}.$$

3. The transposition of Kronecker product:
$$(A \otimes B)' = A' \otimes B'.$$

4. The mixed-product property:
If A, B, C and D are matrices of such size that one can form the matrix products AC and BD, then
$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$$

5. The orthonormal property: If W_1 and W_2 are two orthonormal matrices, then $W = W_1 \otimes W_2$ is also an orthonormal matrix.

6. Hadamard product:
$$(A \otimes B) \circ (C \otimes D) = (A \circ C) \otimes (B \circ D)$$
where $\circ$ is the Hadamard product.

Detail of the reparametrization of the general LMM (see equation (2) of the article):
Using the reparametrisation property 1 of the Kronecker product introduced above, a general LMM model (see equation (1) of the article) can be reformulated as:

$$\begin{aligned} y &= \mathbf{X}\beta + \mathbf{Z}\mathbf{\Lambda}u + \varepsilon \\ &= \mathbf{X}\beta + vec(\mathbf{Z}\mathbf{\Lambda}u) + \varepsilon \\ &= \mathbf{X}\beta + (u' \otimes \mathbf{Z})\, vec(\mathbf{\Lambda}) + \varepsilon \\ &= \mathbf{X}\beta + (u' \otimes \mathbf{Z})\, P\lambda + \varepsilon \\ &= [\mathbf{X}, (u' \otimes \mathbf{Z})\, P] \begin{bmatrix} \beta \\ \lambda \end{bmatrix} + \varepsilon \end{aligned}$$

where λ is the unique diagonal elements vector of λ and P is the matrix that transforms λ to $vec(\mathbf{\Lambda})$ (Ibrahim et al., 2011).

Detail of the calculation of the design matrix Z (see equation (10) of the article):
For the RIS model, the design matrix of the random effect $\mathbf{Z}$ is define by $(J \bullet X)$. In our specific context, the matrix J is the design matrix associated to the time factor and can be constructed as $J = I_{t_T} \otimes \mathbb{1}_n$. and we have $X = I_{t_T} \otimes X$. Thus, design matrix $\mathbf{Z}$ can also be simplified using the definition of the face-splitting product (see equation (1)) and property 6 previously introduced:

$$\begin{aligned}
\mathbf{Z} &= (J \bullet X) \\
&= (J \otimes \mathbb{1}'_{p+1}) \circ (\mathbb{1}'_{t_T} \otimes X) \\
&= (I_{t_T} \otimes \mathbb{1}_n \otimes \mathbb{1}'_n) \circ (\mathbb{1}'_{t_T} \otimes \mathbb{1}_{t_T} \otimes X) \\
&= (I_{t_T} \otimes \mathbb{1}_{n,p+1}) \circ (\mathbb{1}_{t_T,t_T} \otimes X) \\
&= (I_{t_T} \circ \mathbb{1}_{t_T,t_T}) \otimes (\mathbb{1}_{n,p+1} \circ X) \\
&= I_{t_T} \otimes X,
\end{aligned}$$

where $\mathbb{1}_v$ is a v unit vector and $\mathbb{1}_{v,v'}$ is a $v \times v'$ uni matrix, v and v' are equal to $p+1$, t_T or n.

Detail of the reparametrization of the RIS model (see equation (11) of the article):
Using the joint parametrization of LMM (see equation (1) of the article), specific context simplifications $\mathbf{X} = (\mathbb{1}_{t_T} \otimes X)$, $\mathbf{Z} = (I_{t_T} \otimes X)$ and $\mathbf{\Lambda} = (I_{t_T} \otimes \Lambda)$ in addition to the reparametrization property 1 and mixed-product property 4 of the Kroneker product introduced above, the RIS model (see equation (9) of the article) can be formulated as:

$$\begin{aligned}
y &= D\alpha + \mathbf{X}\beta + \mathbf{Z}\mathbf{\Lambda}u + \varepsilon \\
&= D\alpha + (\mathbb{1}_{t_T} \otimes X)\beta + (I_{t_T} \otimes X)(I_{t_T} \otimes \Lambda)u + \varepsilon \\
&= D\alpha + (\mathbb{1}_{t_T} \otimes X)\beta + (I_{t_T} \otimes X\Lambda)vec(U) + \varepsilon \\
&= D\alpha + (\mathbb{1}_{t_T} \otimes X)\beta + (X\Lambda U) + \varepsilon \\
&= D\alpha + (\mathbb{1}_{t_T} \otimes X)\beta + (U' \otimes X)vec(\Lambda) + \varepsilon \\
&= D\alpha + (\mathbb{1}_{t_T} \otimes X)\beta + (U' \otimes X)P\Lambda + \varepsilon \\
&= D\alpha + [\mathbb{1}_{t_T} \otimes X, (U' \otimes X)P] \begin{bmatrix} \beta \\ \lambda \end{bmatrix} + \varepsilon.
\end{aligned}$$

U is a $(p+1) \times t_T$ matrix of the collection of the t_T reparametrized vectors of random intercept and slopes associated to each time $U = \bigoplus_k^{t_T} u_k$ (U follows a matrix Gaussian distribution $\mathcal{MN}_{(p+1)\times t_T}(0, R, I_{t_T})$ and $D = (\mathbb{1}_{t_T} \otimes I_n)$ the design matrix associated to the individual random effects. $\varepsilon = (\varepsilon'_1, \ldots, \varepsilon'_{t_T})'$ is the concatenation of all residuals over time associated to all individuals where $\varepsilon_t = (\varepsilon_{1,t}, \ldots, \varepsilon_{n,t})'$ be a n vector of residuals associated to all individuals at time t. *varepsilon* is supposed to follow a Gaussian distribution centered on zero with covariance $\sigma^2\Gamma$ where $\Gamma = (\Gamma \otimes I_n)$.

Web Appendix C: Bayesian hierarchical models, Full conditional distributions and MCMC algorithms

Animal model

The Bayesian hierarchical model for the animal model is given by:

$$\begin{aligned} y|\mu,\lambda,U,\sigma^2 &\sim \mathcal{N}_n(\mu + U\lambda, \sigma^2 I_n) \\ \mu &\propto 1 \\ \lambda_l|\tau^2,\omega_l^2,\sigma^2 &\sim \mathcal{N}^f(0,\sigma^2\tau^2\omega_l^2) \\ \tau &\sim \mathcal{C}^+(0,1) \\ \omega_l &\sim \mathcal{C}^+(0,1) \\ \sigma^2 &\sim \mathcal{IG}(s,r) \end{aligned}$$

Half-Cauchy distribution can be represented as a scale mixture of of inverse-gamma distribution (Makalic and Schmidt, 2015) :

$$\tau \sim \mathcal{C}^+(0,1) \;\Leftrightarrow\; \tau^2|\xi_\tau \sim \mathcal{IG}(1/2, 1/\xi_\tau), \quad \xi_\tau \sim \mathcal{IG}(1/2,1/2),$$

and

$$\omega_l \sim \mathcal{C}^+(0,1) \;\Leftrightarrow\; \omega_l^2|\xi_{\omega_l} \sim \mathcal{IG}(1/2, 1/\xi_{\omega_l}),\; \xi_{\omega_l} \sim \mathcal{IG}(1/2,1/2).$$

This representation allows conjugate conditional posterior distribution.

Full conditional distributions are given by:

- For the intercept:
$$\mu|. \sim \mathcal{N}(\mathbb{1}_n'(y - U\lambda)/n, \sigma^2/n)$$
- For the random effects, at each iteration, $q = 135$ by $n = 144$ random effects have to be updated. Simple approach leads to the use of the following full conditional distribution for $l = 1,\ldots,q$:
$$u_l|. \sim \mathcal{N}_n\left(\Sigma_{u_l}\frac{\lambda_l}{\sigma^2}\left(y - \mu - \sum_{k\neq l} u_k\lambda_k\right), \Sigma_{u_l}\right), \tag{2}$$
where the dense variance matrix $\Sigma_{u_l} = (\lambda_l^2/\sigma^2 I_n + A_l^{-1})^{-1}$ have to be updated for all random effects at each iteration. However, within this matrix, only parameters λ_l and σ^2 will be modified at each iteration and linear algebra property of the SVD decomposition of known definite positive matrix A_l can be used to optimize the inverse matrix operation:
$$\begin{aligned} \Sigma_{u_l} &= (\lambda_l^2/\sigma^2 I_n + A_l^{-1})^{-1} \\ &= (\lambda_l^2/\sigma^2 I_n + (W_l D_l W_l')^{-1})^{-1} \\ &= (\lambda_l^2/\sigma^2 I_n + W_l D_l^{-1} W_l')^{-1} \\ &= (W_l(\lambda_l^2/\sigma^2 I_n + D_l^{-1})W_l')^{-1} \\ &= W_l(\lambda_l^2/\sigma^2 I_n + D_l^{-1})^{-1} W_l', \end{aligned}$$
where W_l is an orthogonal matrix of singular vectors and D_l is a diagonal matrix of singular values such as $A_l = W_l D_l W_l'$. The matrix $(\lambda_l^2/\sigma^2 I_n + D^{-1})^{-1}$ can be calculate easily because it is simply the inverse of diagonal matrix and the Cholesky decomposition of Σ_{u_l} can be computed easily:

$$L_{u_l} = W_l(\lambda_l^2/\sigma^2 I_n + D_l^{-1})^{-\frac{1}{2}}.$$

Then, an efficient sampling scheme can be constructed to sample u_l from its full conditional distribution (see the Gibbs sampler algorithm bellow).

- Standard deviation parameters λ_l follow an *a priori* conditionally folded normal distribution. However, dealing directly with the full conditional distribution of λ_l is not a simple task:

$$P(\lambda_l|.) \propto P(y|.)P(\lambda_l), \tag{3}$$

where $P(y|.)$ is a Gaussian distribution and $P(\lambda_l)$ is a folded Gaussian distribution. Instead, we propose the following reparametrization $u_l\lambda_l = v_l\xi_l$ such as $\lambda_l = |\xi_l|$ with $\xi_l \sim \mathcal{N}(0, \sigma^2\tau^2\omega_l^2)$. This reparametrization, inspired by the work of Gelman (2006), allows to deal with known full conditional distribution instead of the unkwnown full conditional distribution (3):

$$\xi_l|. \sim \mathcal{N}\left(\Sigma_{\xi_l}\frac{sign(\xi_l)u_l}{\sigma^2}(y - \mu - \sum_{j\neq l} u_j\lambda_j), \Sigma_{\xi_l}\right). \tag{4}$$

with

$$\Sigma_{\xi_l} = \left(\frac{u_l'u_l}{\sigma^2} + \frac{1}{\sigma^2\tau^2\omega_l^2}\right)^{-1}$$

Then λ_l is can be updated by sampling ξ from its full conditional distribution (4) and then compute $\lambda_l = |\xi_l|$.

- For the global variance parameter:

$$\tau^2|. \sim \mathcal{IG}\left(\frac{1+q}{2}, \frac{1}{\nu_\tau} + \frac{1}{2\sigma^2}\sum_{k=1}^{q}\frac{\lambda_l^2}{\omega_l^2}\right),$$

$$\nu_\tau|. \sim \mathcal{IG}\left(1, 1 + \frac{1}{\tau^2}\right)$$

- For local variance parameters:

$$\omega_l^2|. \sim \mathcal{IG}\left(1, \frac{1}{\nu_{\omega_l}} + \frac{\lambda_l^2}{2\sigma^2\tau^2}\right),$$

$$\nu_{\omega_l}|. \sim \mathcal{IG}\left(1, 1 + \frac{1}{\omega_l^2}\right)$$

- For the residual variance

$$\sigma^2|. \sim \mathcal{IG}\left(s + \frac{n}{2} + \frac{q}{2}, r + \frac{1}{2}\mathbb{1}_n'(y - \mu - U\lambda) + \frac{1}{2\tau^2}\sum_{k=1}^{q}\frac{\lambda_k^2}{\omega_l^2}\right)$$

All full conditional distributions are known and an efficient sampling scheme through a Gibbs sampler algorithm (Gilks, Richardson, and Spiegelhalter, 1995) can be constructed:

Gibbs sampler algorithm for the animal model:

Choose starting values for all parameters μ, ξ_l, τ^2, ω_l^2 and σ^2, set $\lambda_l = |\xi_l|$, $j = 1, \ldots, q$ and compute W_l and D_l such as $A_l = W_lD_lW_l'$, the SVD decomposition of A_l for $l = 1, \ldots, q$. Repeat the following steps:

(a) Sample μ from Gaussian distribution $\mu|. \sim \mathcal{N}(\mathbb{1}_n'(y - U\lambda)/n, \sigma^2/n)$.

(b) For $l = 1, \ldots, q$, sample u_l from multivariate Gaussian distribution $p(u_l|.)$ (see equation (2)) using the following sampler scheme:

1. Calculate the lower triangular matrix L_{u_l} of the Cholesky decomposition of Σ_{u_l}:

$$L_{u_l} = W_l(\lambda_l^2/\sigma^2 I_n + D_l^{-1})^{-\frac{1}{2}}$$

2. Sample z_l from a Gaussian distribution $\mathcal{N}_n(0, I_n)$
3. Compute $u_l = L_{u_l}\left(z_l + L'_{u_l}\frac{\lambda_l}{\sigma^2}\left(y - \mu - \sum_{k\neq l} u_k\lambda_k\right)\right)$.

(c) For $l = 1, \ldots, q$:

1. Sample ξ from Gaussian distribution:

$$\xi_l|. \sim \mathcal{N}\left(\Sigma_{\xi_l}\frac{sign(\xi_l)u_l}{\sigma^2}(y - \mu - \sum_{j\neq l} u_j\lambda_j), \Sigma_{\xi_l}\right),$$

2. Set $\lambda_i = |\xi_l|$.

(d) Sample τ^2 from inverse gamma distribution $\tau^2|. \sim \mathcal{IG}\left(\frac{1+q}{2}, \frac{1}{\nu_\tau} + \frac{1}{2\sigma^2}\sum_{k=1}^q \frac{\lambda_l^2}{\omega_l^2}\right)$.

(e) Sample ν_τ from inverse gamma distribution $\nu_\tau|. \sim \mathcal{IG}\left(1, 1 + \frac{1}{\tau^2}\right)$.

(f) For $l = 1, \ldots, q$, sample ω_l^2 from inverse gamma distribution $\omega_l^2|. \sim \mathcal{IG}\left(1, \frac{1}{\nu_{\omega_l}} + \frac{\lambda_l^2}{2\sigma^2\tau^2}\right)$.

(g) For $l = 1, \ldots, q$, sample ν_{ω_l} from inverse gamma distribution $\nu_{\omega_l}|. \sim \mathcal{IG}\left(1, 1 + \frac{1}{\omega_l^2}\right)$.

(h) Sample σ^2 from inverse gamma distribution:

$$\sigma^2|. \sim \mathcal{IG}\left(s + \frac{n}{2} + \frac{q}{2}, r + \frac{1}{2}\mathbb{1}_n'(y - \mu - U\lambda) + \frac{1}{2\tau^2}\sum_{k=1}^q \frac{\lambda_k^2}{\omega_l^2}\right).$$

RIS model

Bayesian hierarchical model for the RIS model is given by:

$$
\begin{aligned}
y|v,\beta,u &\sim \mathcal{N}_{nt_T}(D\alpha + (\mathbb{1}_{t_T}\otimes X)\beta + (U'\otimes X)P\lambda,\ \sigma^2(\Gamma\otimes I_n))\\
\alpha|\sigma^2_\alpha &\sim \mathcal{N}_n(0,\sigma^2_\alpha I_n), \quad \sigma^2_\alpha \sim \mathcal{IG}(s_\alpha, r_\alpha)\\
\beta|\omega^2_\beta,\tau^2_\beta,\sigma^2 &\sim \mathcal{N}_{p+1}(0,\sigma^2 V_\beta), \quad V_\beta = diag(\omega^2_{\beta_l}\tau^2_\beta, l=0,\ldots,p)\\
\tau^2_\beta|\upsilon_\beta &\sim \mathcal{IG}(1/2, 1/\upsilon_\beta), \quad \upsilon_\beta \sim \mathcal{IG}(1/2,1)\\
\omega^2_{\beta_l}|\nu_{\beta_l} &\sim \mathcal{IG}(1/2, 1/\nu_{\beta_l}), \quad \nu_{\beta_l} \sim \mathcal{IG}(1/2,1), \quad l=0,\ldots,p\\
\lambda_l|\omega^2_{\lambda_l},\tau^2_\lambda,\sigma^2 &\sim \mathcal{N}^f(0,\sigma^2\tau^2_\lambda\omega^2_{\lambda_l}), \quad l=0,\ldots,p\\
\tau^2_\lambda|\upsilon_\lambda &\sim \mathcal{IG}(1/2, 1/\upsilon_\lambda), \quad \upsilon_\lambda \sim \mathcal{IG}(1/2,1)\\
\omega^2_{\lambda_l}|\nu_{\lambda_l} &\sim \mathcal{IG}(1/2, 1/\nu_{\lambda_l}), \quad \nu_{\lambda_l} \sim \mathcal{IG}(1/2,1), \quad l=0,\ldots,p\\
u|R &\sim \mathcal{N}_{t_T(p+1)}(0,(I_{t_T}\otimes R))\\
p(R|k) &= \prod_{j=1}^{q-1}\left[\frac{\Gamma(\frac{2k+j}{2}+1)}{\sqrt{\pi}\Gamma(\frac{2k+j+1}{2})}\right]^j [det(R)]^k\\
k &\sim \mathcal{U}(1,100)\\
\rho &\sim \mathcal{U}(0,1) \quad \sigma^2 \sim \mathcal{IG}(s_{\sigma^2}, r_{\sigma^2}).
\end{aligned}
$$

Bayesian inference of the RIS hierarchical model using horseshoe prior presented is achieve using Markov chain Monte Carlo (MCMC) algorithm sampling. It implies to calculate the full conditional distributions of each parameters. Reminder: $\mathbf{X} = \mathbb{1}_{t_T}\otimes X$, $\mathbf{Z} = I_{t_T}\otimes X$.

- The full conditional distribution of the individual random effect α is given by:

$$\alpha|. \sim \mathcal{N}_n\left(\Sigma_\alpha D'\frac{\Gamma^{-1}}{\sigma^2}(y-\mathbf{X}\beta - Z(U'\otimes I_{p+1})P\lambda), \Sigma_\alpha\right), \tag{5}$$

  with

$$\Sigma_\alpha = \left(\frac{D'\Gamma^{-1}D}{\sigma^2} + \frac{I_n}{\sigma^2_\alpha}\right)^{-1}.$$

  However, in our context, we have $D = (\mathbb{1}_{t_T}\otimes I_n)$ and $\Omega = (\Omega\otimes I_n)$ and product $D'(\Gamma^{-1}\otimes I_n)D$ leads to a diagonal matrix $(\mathbb{1}'_{t_T}\Gamma^{-1}\mathbb{1}_{t_T})I_n$ where $\mathbb{1}'_{t_T}\Gamma^{-1}\mathbb{1}_{t_T} = 2(1-\rho) + (t_T-2)(1-\rho)^2$ with ρ the autoregressive decay parameter. Thus, Cholesky decomposition of Σ_α can be computed easily and an effcieint sample scheme can be considered (see the Metropolis within Gibss algorithm bellow).

- The full condition distribution of the individual random effect variance σ^2_α:

$$\sigma^2_\alpha|. \sim \mathcal{IG}\left(s_\alpha + \frac{n}{2}, r_\alpha + \frac{\alpha'\alpha}{2}\right)$$

- The full condition distribution of fixed effects β is:

$$\beta|. \sim \mathcal{N}_{p+1}\left(\Sigma_\beta\frac{(\mathbb{1}'_{t_T}\Gamma^{-1}\otimes X')}{\sigma^2}(y - D\alpha - (U'\otimes X)P\lambda), \Sigma_\beta\right),$$

with

$$\Sigma_\beta = \sigma^2 \left((\mathbb{1}'_{t_T} \Gamma^{-1} \mathbb{1}_{t_T}) X'X + V_\beta^{-1} \right)^{-1}.$$

- Full conditional prior distributions for τ_β^2 and υ_β:

$$\begin{aligned} \tau_\beta^2|. &\sim \mathcal{IG}\left(\frac{1+p}{2},\ \frac{1}{\upsilon_\beta} + \frac{1}{2\sigma^2} \sum_{l=0}^{p} \beta_l^2 \omega_{\beta_l}^2 \right) \\ \upsilon_\beta|. &\sim \mathcal{IG}(1,\ 1 + 1/\tau_\beta^2) \end{aligned}$$

- Full conditional prior distributions for ω_{β_l} and ν_{β_l}:

$$\begin{aligned} \omega_{\beta_l}^2|. &\sim \mathcal{IG}\left(1,\ \frac{1}{\nu_{\beta_l}} + \frac{\beta_l^2}{2\sigma^2 \tau_\beta^2} \right) \\ \nu_{\beta_l}|. &\sim \mathcal{IG}(1,\ 1 + 1/\omega_{\beta_l}^2) \end{aligned}$$

- Full conditional distributions of standard deviations λ_j are proportional to a non-central multivariate folded-Gaussian distribution and have not a closed form implying not simple sample step. To overcome this difficulty, we propose the following reparametrisation of the model (11) of the article:

$$y = D\alpha + (\mathbb{1}_{t_T} \otimes X)\beta + ((SU)' \otimes X) P \xi + \varepsilon$$

where ξ is an unconstrained scale parameter such as $\lambda = |\xi|$ and S is a $(p+1) \times (p+1)$ diagonal matrix of sign of ξ elements ($\lambda = S\xi$). ξ follows then an *a priori* Gaussian distribution $\mathcal{N}_{p+1}(0, \sigma^2 V_\lambda$ where V_λ is a $(p+1) \times (p+1)$ diagonal matrix with elements given by $\tau_\lambda^2 \omega_{\lambda_l}^2$, for $l = 0, \ldots, p$.

Then, the full conditional distribution of ξ can be calculated easily:

$$\xi|. \sim \mathcal{N}_{p+1} \left(\frac{\Sigma_\xi}{\sigma^2} P' (SU\Gamma^{-1} \otimes X')(y - D\alpha - (\mathbb{1}_{t_T} \otimes X)\beta), \Sigma_\xi \right), \tag{6}$$

with

$$\Sigma_\xi = \sigma^2 (P'[(SU\Gamma^{-1}U'S) \otimes X'X]P + V_\lambda^{-1})^{-1}.$$

Thus, λ_l can be updated by sampling ξ_l from its full conditional distribution (6) and then compute $\lambda_l = |\xi_l|$.

- Full conditional prior distributions for τ_λ^2 and υ_λ:

$$\begin{aligned} \tau_\lambda^2|. &\sim \mathcal{IG}\left(\frac{2+p}{2},\ \frac{1}{\upsilon_\lambda} + \frac{1}{2\sigma^2} \sum_{l=0}^{p} {\lambda_l}^2 \omega_{\lambda_l}^2 \right) \\ \upsilon_\lambda|. &\sim \mathcal{IG}(1,\ 1 + 1/\tau_\lambda^2) \end{aligned}$$

- Full conditional prior distributions for ω_{λ_l} and ν_{λ_l}:

$$\begin{aligned} \omega_{\lambda_l}^2|. &\sim \mathcal{IG}\left(1,\ \frac{1}{\nu_{\lambda_l}} + \frac{{\lambda_l}^2}{2\sigma^2 \tau_\lambda^2} \right) \\ \nu_{\lambda_l}|. &\sim \mathcal{IG}(1,\ 1 + 1/\omega_{\lambda_l}^2) \end{aligned}$$

- For the random effect u:
  Using the joint parametrization of model (9) of the article:

$$y = D\alpha + (\mathbb{1}_{t_T} \otimes X)\beta + (I_{t_T} \otimes X\Lambda)u + \varepsilon,$$

  the full condition distribution of the time random effect u may be easily calculated:

$$u|. \sim \mathcal{N}_{t_T(p+1)}\left(\frac{\Sigma_u}{\sigma^2}(\Gamma^{-1} \otimes \Lambda X')(y - D\alpha - (I_{t_T} \otimes X)\beta), \Sigma_u\right), \tag{7}$$

  with

$$\Sigma_u = \left(\frac{1}{\sigma^2}(\Gamma^{-1} \otimes \Lambda' X' X \Lambda) + (I_{t_T} \otimes R^{-1})\right)^{-1}.$$

  However this distribution imply the inversion of a $(t_T(p+1) \times t_T(p+1))$ dense matrix and may be time consuming in the MCMC sampler (complexity is of $O((t_T(p+1))^3)$). To improve this step, we reparametrize equation (11) of the article by dropping out the correlation matrix of the time random effect $u = (I_{t_T} \otimes B)z$ where $R = BB'$ with $z \sim \mathcal{N}_{t_T(p+1)}(0, I_{t_T(p+1)})$. Posterior density of z is given by:

$$z|. \sim \mathcal{N}_{t_T(p+1)}\left(\frac{\Sigma_z}{\sigma^2}(\Gamma^{-1} \otimes B'\Lambda X')(y - D\alpha - (I_{t_T} \otimes X)\beta), \Sigma_z\right)$$

  with

$$\Sigma_z = \left(\frac{1}{\sigma^2}(\Gamma^{-1} \otimes B'\Lambda X' X \Lambda B) + I_{t_T(p+1)}\right)^{-1}.$$

  Now, as describing in section above, we can observe that this matrix is the inverse of the sum of a dense matrix and a diagonal matrix. Thus, the Cholseky decomposition of the covariance matrix Σ_z can be computed efficiently using the SVD decomposition and property of the Kronecker product. Let $W_{\Gamma^{-1}} D_{\Gamma^{-1}} W'_{\Gamma^{-1}}$ be the SVD decomposition of Γ^{-1} and $W_B D_B W'_B$ be the SVD decomposition of $B'\Lambda X' X \Lambda B$ where $W_{\Gamma^{-1}}$ and W_B are orthonormal matrices of singular vectors and $D_{\Gamma^{-1}}$ and D_B are diagonal matrices of singular values. Then, we obtain the following relation:

$$\begin{aligned}
\Sigma_z &= \left(\tfrac{1}{\sigma^2}(\Gamma^{-1} \otimes B'\Lambda X' X \Lambda B) + I_{t_T(p+1)}\right)^{-1} \\
&= \left((W_{\Gamma^{-1}} \otimes W_B)\left(\tfrac{1}{\sigma^2} D_{\Gamma^{-1}} \otimes D_B\right)(W'_{\Gamma^{-1}} \otimes W'_B) + I_{t_T(p+1)}\right)^{-1} \\
&= \left((W_{\Gamma^{-1}} \otimes W_B)\left(\tfrac{1}{\sigma^2} D_{\Gamma^{-1}} \otimes D_B + I_{t_T(p+1)}\right)(W'_{\Gamma^{-1}} \otimes W'_B)\right)^{-1} \\
&= (W_{\Gamma^{-1}} \otimes W_B)\left(\tfrac{1}{\sigma^2} D_{\Gamma^{-1}} \otimes D_B + I_{t_T(p+1)}\right)^{-1}(W'_{\Gamma^{-1}} \otimes W'_B).
\end{aligned}$$

  Note that $(W_{\Gamma^{-1}} \otimes W_B)$ is also an orthonormal matrix. Now, the SVD decomposition of the covariance matrix Σ_z is given by:

$$L_z = (W_{\Gamma^{-1}} \otimes W_B)\left(\frac{1}{\sigma^2} D_{\Gamma^{-1}} \otimes D_B + I_{t_T(p+1)}\right)^{-\frac{1}{2}}$$

  and an efficient sampling scheme can be considered (see the Metropolis within Gibss algorithm bellow). Thus, using this transformation, the most difficult task to computed the Cholesky decomposition of the covariance matrix Σ_z is SVD decompositions of Γ^{-1} and $B'\Lambda X' X \Lambda B$ where the complexity if of $O(t_T^3)$ and $O((p+1)^3)$ respectively.

- The full condition distribution of the angle parameter $\theta_{i,j}$:

$$\mathbf{R}|. \propto exp\left\{-\frac{1}{2}\sum_{t=1}^{t_T} u'_t \mathbf{R} u_t\right\} |\mathbf{R}|^{-t_T/2}$$

- The full condition distribution of the residual autoregressive parameter ρ:

$$\rho|. \sim \mathbb{1}_{(0<\rho<1)}|\Gamma|^{-\frac{n}{2}} exp\Bigg\{-\frac{1}{2\sigma^2}(y-D\alpha-(\mathbb{1}_{t_T}\otimes X)\beta-(U'\otimes X)P\lambda)'(\Gamma^{-1}\otimes I_n)(y-D\alpha-(\mathbb{1}_{t_T}\otimes X)\beta-(U'\otimes X)P\lambda)\Bigg\}$$

- The full condition distribution of the residual variance σ^2:

$$\sigma^2|. \sim \mathcal{IG}\left(s_{\sigma^2}+\frac{nt_T+2(p+1)}{2}, r_{\sigma^2}+\frac{1}{2}||y-D\alpha-(\mathbb{1}_{t_T}\otimes X)\beta-(U'\otimes X)P\lambda||_2^2+\frac{1}{2\tau_\beta^2}\sum_{j=0}^{p}\beta_j^2\omega_{\beta_j}^2+\frac{1}{2\tau_\lambda^2}\sum_{j=0}^{p}\lambda_j^2\omega_{\lambda_j}^2\right)$$

Full posterior conditional distributions of matrix R and parameter ρ have not a close form, thus, we use a Metropolis within Gibbs sampler algorithm:

Metropolis within Gibbs sampler algorithm for RIS model:

Choose starting values for all parameters α, σ_α^2, β, τ_β^2, υ_β, $\omega_{\beta_j}^2$, ν_{β_j}, λ, τ_λ^2, υ_λ, $\omega_{\lambda_j}^2$, ν_{λ_j}, u, R, ρ and σ^2 and repeat the following steps:

(a) Sample α from the multivariate Gaussian distribution $p(\alpha|.)$ given in (5):

1. Compute $L = \left(\frac{(\mathbb{1}'_{t_T}\Gamma^{-1}\mathbb{1}_{t_T})I_n}{\sigma^2}+\frac{I_n}{\sigma_\alpha^2}\right)^{-\frac{1}{2}}$, the lower triangular matrix of the Cholesky decomposition of Σ_α where $\mathbb{1}'_{t_T}\Gamma^{-1}\mathbb{1}_{t_T} = 2(1-\rho)+(t_T-2)(1-\rho)^2$
2. Sample a from a Gaussian distribution $\mathcal{N}_n(0, I_n)$
3. Compute $\alpha = La + LL'\frac{(\mathbb{1}'_{t_T}\Gamma^{-1}\otimes I_n)}{\sigma^2}(y-(\mathbb{1}_{t_T}\otimes X)\beta-(U'\otimes X)P\lambda)$

(b) Sample σ_α^2 from inverse-gamma distribution $\sigma_\alpha^2|. \sim \mathcal{IG}\left(s_\alpha+\frac{n}{2}, r_\alpha+\frac{\alpha'\alpha}{2}\right)$

(c) Sample β from multivariate Gaussian distribution:

$$\beta|. \sim \mathcal{N}_{p+1}\left(\Sigma_\beta\frac{(\mathbb{1}'_{t_T}\Gamma^{-1}\otimes X')}{\sigma^2}(y-(D\alpha-(U'\otimes X)P\lambda), \Sigma_\beta\right)$$

(d) Sample τ_β^2 from inverse-gamma distribution $\tau_\beta^2|. \sim \mathcal{IG}\left(\frac{1+p}{2},\ \frac{1}{\upsilon_\beta}+\frac{1}{2\sigma^2}\sum_{l=0}^{p}\beta_l^2\omega_{\beta_l}^2\right)$.

(e) Sample υ_β from inverse-gamma distribution $\upsilon_\beta|. \sim \mathcal{IG}(1,\ 1+1/\tau_\beta^2)$.

(f) For $l = 0, \ldots, p$, sample $\omega_{\beta_l}^2$ from inverse-gamma distribution $\omega_{\beta_l}^2|. \sim \mathcal{IG}\left(1,\ \frac{1}{\nu_{\beta_l}}+\frac{\beta_l^2}{2\sigma^2\tau_\beta^2}\right)$.

(g) For $l = 0, \ldots, p$, sample ν_{β_l} from inverse-gamma distribution $\nu_{\beta_l}|. \sim \mathcal{IG}(1,\ 1+1/\omega_{\beta_l}^2)$.

(h) Sample λ from full conditional distribution $p(\lambda|.)$:

1. Calculate $S = diag\{sign(\xi_l), l = 0, \ldots, p\}$
2. Sample ξ from multivariate Gaussian distribution:

$$\xi|. \sim \mathcal{N}_{p+1}\left(\frac{\Sigma_\xi}{\sigma^2}P'(SU\Gamma^{-1}\otimes X')(y-D\alpha-(\mathbb{1}_{t_T}\otimes X)\beta), \Sigma_\xi\right)$$

3. Set $\lambda = |\xi|$ (or $\lambda = S\xi$) .

(i) Sample τ_λ^2 from inverse-gamma distribution $\tau_\lambda^2|. \sim \mathcal{IG}\left(\frac{2+p}{2},\ \frac{1}{\upsilon_\lambda} + \frac{1}{2\sigma^2}\sum_{l=0}^{p} {\lambda_l}^2\omega_{\lambda_l}^2\right)$.

(j) Sample υ_λ from inverse-gamma distribution $\upsilon_\lambda|. \sim \mathcal{IG}(1,\ 1+1/\tau_\lambda^2)$.

(k) For $l = 0, \ldots, p$, sample $\omega_{\lambda_l}^2$ from inverse-gamma distribution $\omega_{\lambda_l}^2|. \sim \mathcal{IG}\left(1,\ \frac{1}{\nu_{\lambda_l}} + \frac{{\lambda_l}^2}{2\sigma^2\tau_\lambda^2}\right)$.

(l) For $l = 0, \ldots, p$, sample ν_{λ_l} from inverse-gamma distribution $\nu_{\lambda_l}|. \sim \mathcal{IG}(1,\ 1+1/\omega_{\lambda_l}^2)$.

(m) Sample u from multivariate Gaussian posterior given in (7):

1. Calculate $W_{\Gamma^{-1}}$ and $D_{\Gamma^{-1}}$ such as $\Gamma^{-1} = W_{\Gamma^{-1}} D_{\Gamma^{-1}} W'_{\Gamma^{-1}}$, the SVD decomposition of Γ^{-1}
2. Calculate W_B and D_B such as $B'\Lambda X' X \Lambda B = W_B D_B W'_B$, the SVD decomposition of $B'\Lambda X' X \Lambda B$
3. Calculate the lower triangular matrix L_z of the Cholesky decomposition of Σ_z:

$$L_z = (W_{\Gamma^{-1}} \otimes W_B)\left(\frac{1}{\sigma^2} D_{\Gamma^{-1}} \otimes D_B + I_{t_T(p+1)}\right)^{-\frac{1}{2}}$$

4. Sample $\tilde{z}$ from a Gaussian distribution $\mathcal{N}_{t_T(p+1)}(0, I_{t_T(p+1)})$
5. Compute $z = L_z(\tilde{z} + L'_z \frac{1}{\sigma^2}(\Gamma^{-1} \otimes B'\Lambda X')(y - D\alpha - (I_{t_T} \otimes X)\beta))$
6. Compute $u = (I_{t_T} \otimes B)z$.

(n) Sample R using a random walk Metropolis-Hastings step:
For $j = 0, \ldots, p-1$, $i = j+1, \ldots, p$:
Let θ^{old} be the current angle matrix θ and R^{old} be the associated correlation matrix.

1. Let $\theta^* = \theta^{old}$.
2. Sample $\theta^*_{[i,j]}$ from truncated Gaussian distribution in $(0, \pi)$ centering on the current value $q(\theta^*_{[i,j]}|\theta^{old}_{[i,j]}) = \mathcal{N}^{(0,\pi)}(\theta^{old}_{[i,j]}, \epsilon_\theta)$.
3. Calculate correlation matrix R^* from new angle matrix θ^*.
4. Calculate densities $p(R^*|.)$ and $p(R^{old}|.)$ from full conditional posterior distribution:

$$p(R|.) \propto exp\left\{ -\frac{1}{2}\sum_{t=1}^{t_T} u'_t R u_t \right\} |R|^{-t_T/2}$$

.

5. Calculate densities $q(\theta^{old}_{[i,j]}|\theta^*_{[i,j]})$ and $q(\theta^*_{[i,j]}|\theta^{old}_{[i,j]})$ from proposal truncated Gaussian distribution.
6. Calculate acceptance ratio
$r(\theta^{old}, \theta^*) = min\left\{1, \frac{p(R^*|.)}{p(R^{old}|.)} \frac{q(\theta^{old}_{[i,j]}|\theta^*_{[i,j]})}{q(\theta^*_{[i,j]}|\theta^{old}_{[i,j]})}\right\}$.
7. Sample z from uniform distribution $\mathcal{U}(0,1)$
8. If $z \leq r(\theta^{old}, \theta^*)$ then $\theta = \theta^*$, $R = R^*$ else $\theta = \theta^{old}$, $R = R^{old}$.

(o) Sample ρ using a random walk Metropolis-Hastings step:
Let ρ^{old} be the current value of ρ

1. Sample ρ^* from truncated Gaussian distribution in $(-1, 1)$ centering on the current value $q(\rho^*|\rho^i) = \mathcal{N}^{(-1,1)}(\rho^{old}, \epsilon_\rho)$.

2. Calculate densities $p(\rho^*|.)$ and $p(\rho^{old}|.)$ from full conditional posterior distribution:

$$\rho|. \sim \mathbb{1}_{(0<\rho<1)}|\Gamma|^{-\frac{n}{2}} exp\Bigg\{-\frac{1}{2\sigma^2}(y-D\alpha-(\mathbb{1}_{t_T}\otimes X)\beta-(U'\otimes X)P\lambda)'(\Gamma^{-1}\otimes I_n)(y-D\alpha-(\mathbb{1}_{t_T}\otimes X)\beta-(U'\otimes X)P\lambda)\Bigg\}$$

3. Calculate densities $q(\rho^{old}|\rho^*)$ and $q(\rho^*|\rho^{old})$ from proposal truncated Gaussian distribution.
4. Calculate acceptance ratio $r(\rho^{old}, \rho^*) = min\left\{1, \frac{p(\rho^*|.)}{p(\rho^{old}|.)}\frac{q(\rho^{old}|\rho^*)}{q(\rho^*|\rho^{old})}\right\}$
5. Sample z from uniform distribution $\mathcal{U}(0,1)$
6. If $z \leq r(\rho^{old}, \rho^*)$ then $\rho = \rho^*$ else $\rho = \rho^{old}$

(p) Sample σ^2 from inverse-gamma distribution:

$$\sigma^2|. \sim \mathcal{IG}\Bigg(s_{\sigma^2}+\frac{nt_T+2(p+1)}{2}, r_{\sigma^2}+\frac{1}{2}||y-D\alpha-(\mathbb{1}_{t_T}\otimes X)\beta-(U'\otimes X)P\lambda||_2^2+\frac{1}{2\tau_\beta^2}\sum_{j=0}^{p}\beta_j^2\omega_{\beta_j}^2+\frac{1}{2\tau_\lambda^2}\sum_{j=0}^{p}\lambda_j^2\omega_{\lambda_j}^2\Bigg)$$

References

Armagan, A., Clyde, M., and Dunson, D. (2011). Generalized beta mixtures of gaussians. *Advances in neural information processing systems* **24,** 523–531.

Armagan, A., Dunson, D. B., and Lee, J. (2013). Generalized double pareto shrinkage. *Statistica Sinica* **23,** 119.

Bai, R. and Ghosh, M. (2019). Large-scale multiple hypothesis testing with the normal-beta prime prior. *Statistics* **53,** 1210–1233.

Bhadra, A., Datta, J., Polson, N. G., Willard, B., et al. (2017). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis* **12,** 1105–1131.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). Dirichlet–laplace priors for optimal shrinkage. *Journal of the American Statistical Association* **110,** 1479–1490.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, pages 73–80.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97,** 465–480.

Eilers, P. H. C. and Marx, B. D. (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and intelligent laboratory systems* **66,** 159–174.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis* **1,** 515–534.

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

Griffin, J., Brown, P., et al. (2017). Hierarchical shrinkage priors for regression models. *Bayesian Analysis* **12,** 135–159.

Griffin, J. E. and Brown, P. J. (2007). Bayesian adaptive lassos with non-convex penalization.

Griffin, J. E., Brown, P. J., et al. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian analysis* **5,** 171–188.

Hahn, P. R., He, J., and Lopes, H. (2016). Elliptical slice sampling for bayesian shrinkage regression with applications to causal inference. *U RL http://faculty. chicagobooth. edu/richard. hahn/research. html* .

Ibrahim, J. G., Zhu, H. T., Garcia, R. I., and Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics* **67,** 495–503.

Makalic, E. and Schmidt, D. F. (2015). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters* **23,** 179–182.

Makalic, E. and Schmidt, D. F. (2016). High-dimensional bayesian regularised regression with the bayesreg package. *arXiv preprint arXiv:1611.06649* .

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association* **103,** 681–686.

Pérez, P. and de Los Campos, G. (2014). Genome-wide regression and prediction with the bglr statistical package. *Genetics* **198,** 483–495.

van der Pas, S., Scott, J., Chakraborty, A., and Bhattacharya, A. (2016). horseshoe: Implementation of the horseshoe prior. *R package version 0.1. 0* .

Zhang, S. and Li, M. (2018). "dlbayes" available at cran, r package for implementing the dirichlet-laplace shrinkage prior in bayesian linear regression and variable selection.

Zhang, Y. D., Naughton, B. P., Bondell, H. D., and Reich, B. J. (2020). Bayesian regression using a prior on the model fit: The r2-d2 shrinkage prior. *Journal of the American Statistical Association* pages 1–37.

III

Sélection de groupe de paramètres ordonnés dans un modèle à coefficients variants

Sommaire

3.1 Introduction 49

3.2 Article : Bayesian varying coefficient model with selection: An application to functional mapping 50

3.1 Introduction

L'architecture génétique d'un caractère d'intérêt évolue au cours des différents stades de développement d'un individu. En agronomie par exemple, le rendement moyen d'une espèce peut ne pas être le même en fonction de l'âge des individus. De plus, face à certains changements de conditions environnementales, des mécanismes génétiques peuvent se déclencher pour que les individus puissent s'adapter et survivre. Mieux comprendre cette évolution de l'architecture génétique au cours du temps est alors crucial en sélection végétale ou animale. Les technologies actuelles, telles que les techniques génomiques à haut débit, ont été utilisées pour mieux comprendre les liens entre l'information génétique et les traits quantitatifs. Récemment, des méthodes de phénotypage à haut débit ont fait leur apparition, permettant de mesurer des caractères phénotypiques sur un grand nombre d'individus ainsi que des facteurs environnementaux au cours du temps. Certaines technologies permettent même de contrôler les conditions environnementales en serre et ainsi de simuler différents scénarios de stress environnemental. La combinaison de ces deux informations donne alors accès à des données permettant d'étudier l'évolution de l'architecture génétique au cours du temps. Cependant, de telles données soulèvent de nouveaux défis statistiques liés,

entre autres, aux dimensions élevées, aux dépendances temporelles et aux effets évoluant dans le temps. Dans ce chapitre, nous proposons un modèle à coefficients variants afin de modéliser un caractère phénotypique mesuré à différents temps sur plusieurs individus par des variables environnementales ainsi que des marqueurs moléculaires. L'effet de chaque variable environnementale est modélisé de manière non-linéaire à l'aide d'une interpolation P-spline. Les marqueurs, eux, ont des effets linéaires évoluant dans le temps. Nous proposons une modélisation non-paramétrique de ces évolutions de manière, soit fonctionnelle au travers d'une interpolation P-spline, soit non-fonctionnelle en estimant directement les effets au cours du temps pour chaque marqueur. La première approche produit des estimations plus lisses et permet de réduire le nombre de paramètres à estimer, tandis que la seconde permet d'obtenir des estimations plus flexibles avec potentiellement des changements abruptes. Ces deux approches font intervenir un groupe de paramètres ordonnés pour chaque marqueur. L'identification des marqueurs pertinents passe alors pas la mise à zéro ou non de groupe de paramètres dépendants. Dans cet objectif, nous proposons un prior *group spike-and-slab* (cf section 1.3.1) combiné à une distribution multivariée gaussienne de type marche aléatoire (cf section 1.3.2). Ce prior permet de tenir compte de l'ordre naturel des paramètres lié à l'estimation des effets au cours du temps de chaque marqueur (coefficients P-spline ou effets dans le temps directement), et suppose que les effets entre les différents marqueurs sont indépendants. Ainsi, pour chaque marqueur sélectionné, ce prior permet d'estimer un profil d'effet homogène au cours du temps et non nul.
Nous montrons sur simulation que les deux approches que nous proposons dépassent les approches alternatives existantes en terme de sélection et d'estimation. Enfin nous appliquons notre approche sur des données d'*arabidopsis thaliana* qui ont motivé ces développements.

Notons que ces mêmes données ont aussi été analysées au travers d'un modèle à intercept et pentes aléatoires (cf chapitre 2), où les effets à chaque pas de temps des marqueurs sont modélisés par des réalisations d'effets aléatoires. Contrairement au prior que nous proposons dans ce chapitre, cette modélisation suppose *a priori* une indépendance temporelle entre les pentes aléatoires d'un même marqueur. Cependant, à chaque pas de temps, elle suppose une corrélation inconnue entre les pentes aléatoires des différents marqueurs. D'un point de vue bayésien, ces deux modélisations sont donc proches, si ce n'est qu'elles structurent de manière différente et complémentaire, l'évolution au cours du temps des effets des marqueurs.

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

L'article a été accepté dans la revue *Journal of the Royal Statistical Society: Series C* (DOI: 10.1111/rssc.12447)

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

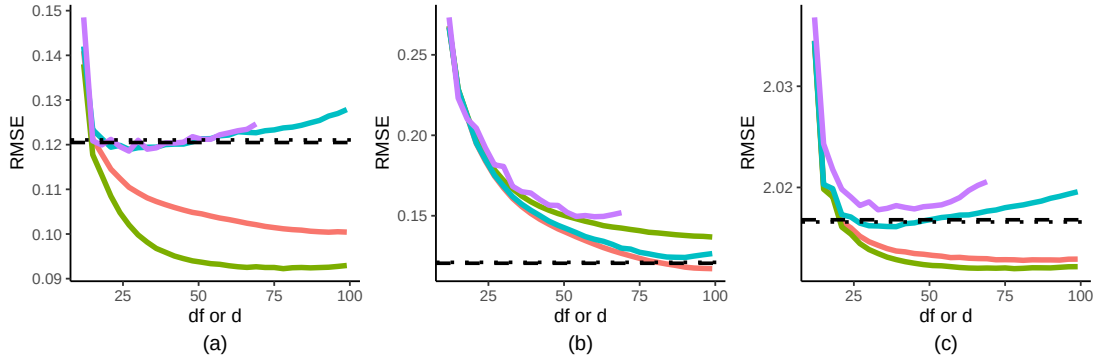


Fig. 1. Panel (a) presents the mean of RMSEs for functional estimation of the smooth effects $\beta_1(t)$, $\beta_2(t)$ and $\beta_3(t)$ for varying number of df and d. Panel (b) presents the RMSE for functional estimation of the piecewise constant effect $\beta_4(t)$ for varying number of df and d. 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

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

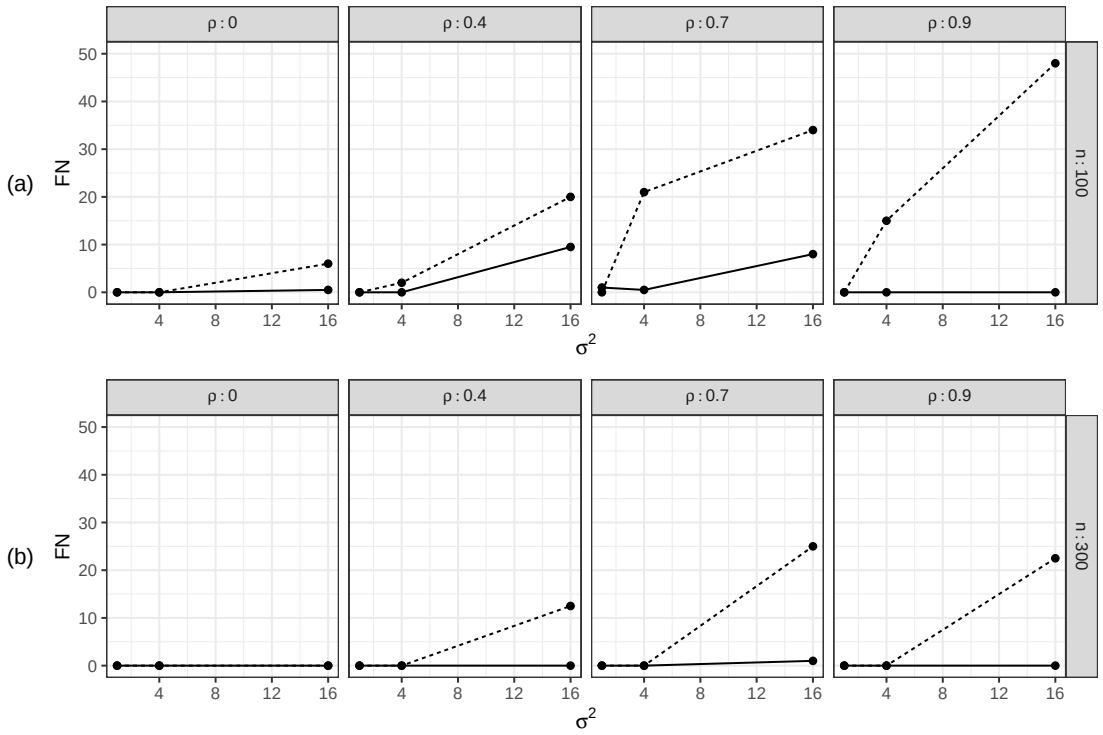
$RMSE_y$ and $RMSE_{B^tX}$ are very similar regardless of the number of individuals and markers (see Table 2). This suggests that even when the model selection fails, the global estimation remains acceptable. However, $RMSE_{B^tX_1}$ and $RMSE_{B^tX_0}$ clearly differ between the two cases ($n = 300,\ J = 500$ 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

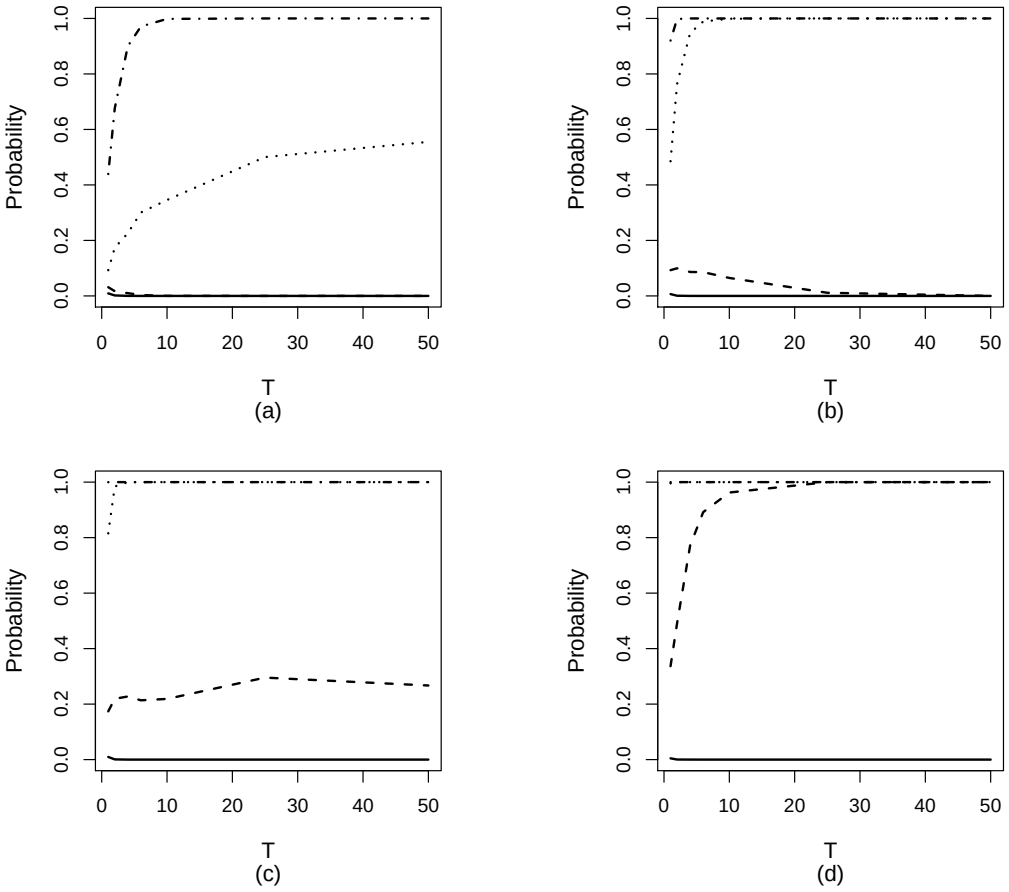


Fig. 3. Marginal probabilities of inclusion for each effect as a function of the number of time points T. Dotted-dashed line, dotted line, dashed line and solid line correspond to effects equal to $0.3, 0.2, 0.1$ and 0 respectively. Figures a, b, c and d are based on 100, 300, 500 and 1000 individuals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

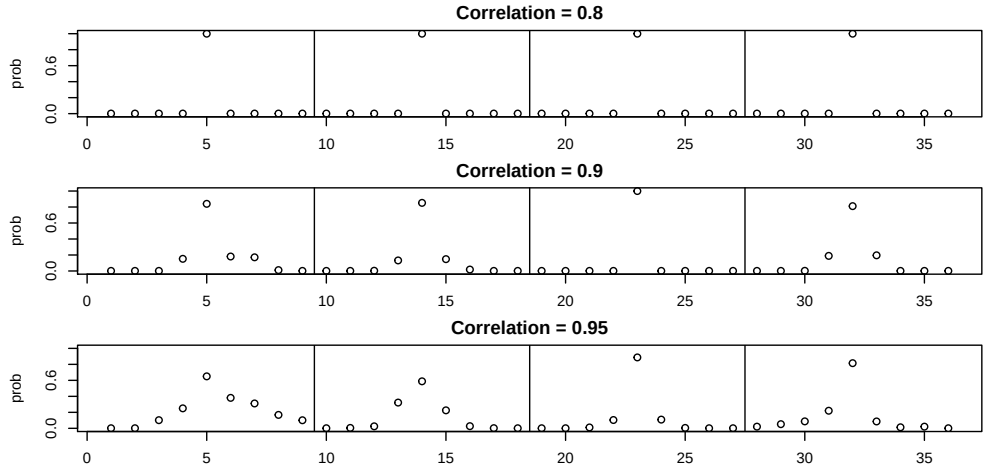


Fig. 4. Marginal probabilities of inclusion for each effect associated to correlated markers within four independent groups.

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

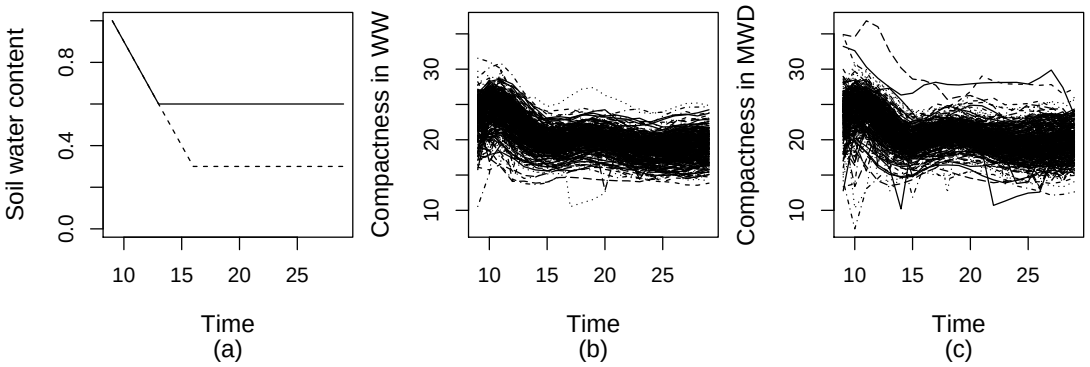


Fig. 5. Panel (a) presents the soil water content under the well-watered (WW) condition in solid line and the moderate water deficit (MWD) conditions in dashed line over time. Panel (b) presents compactness trait observations for the 358 individuals under the WW condition over 21 days. Panel (c) presents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

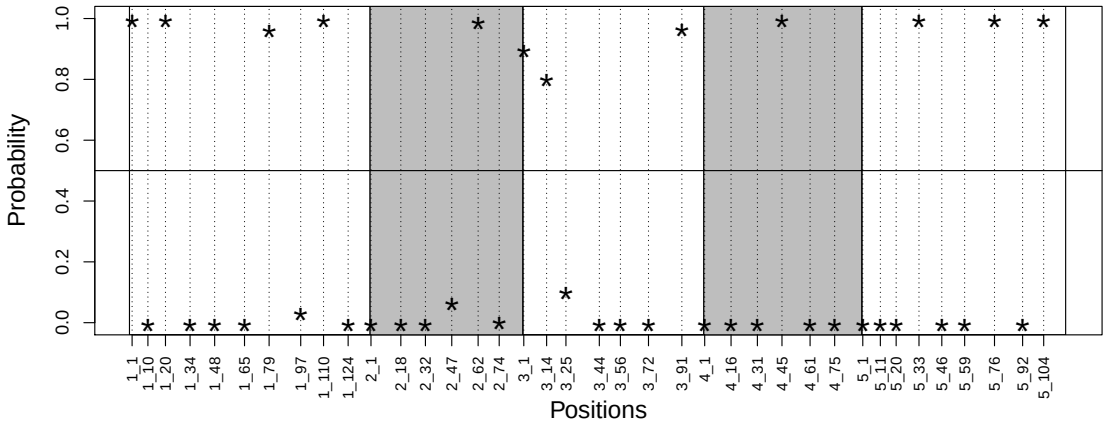


Fig. 6. Marginal posterior inclusion probabilities for the 38 markers in the genetic data $X_{0.7}$ using the PS_2 method. The alternation of white and gray area delimites the 5 chromosomes. A line at 0.5 representing a threshold at 0.5 is plotted.

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 $X_{0.95}$

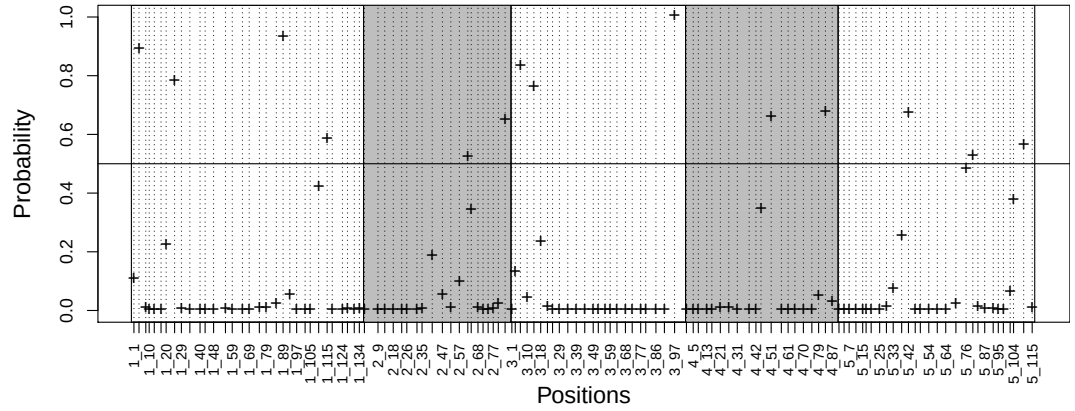


Fig. 7. Marginal posterior inclusion probabilities for the 125 markers of the genetic data $X_{0.95}$ using the PS_2 method. The alternation of white and gray area delimits the five chromosomes. A line at 0.5 representing a threshold at 0.5 is plotted.

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.

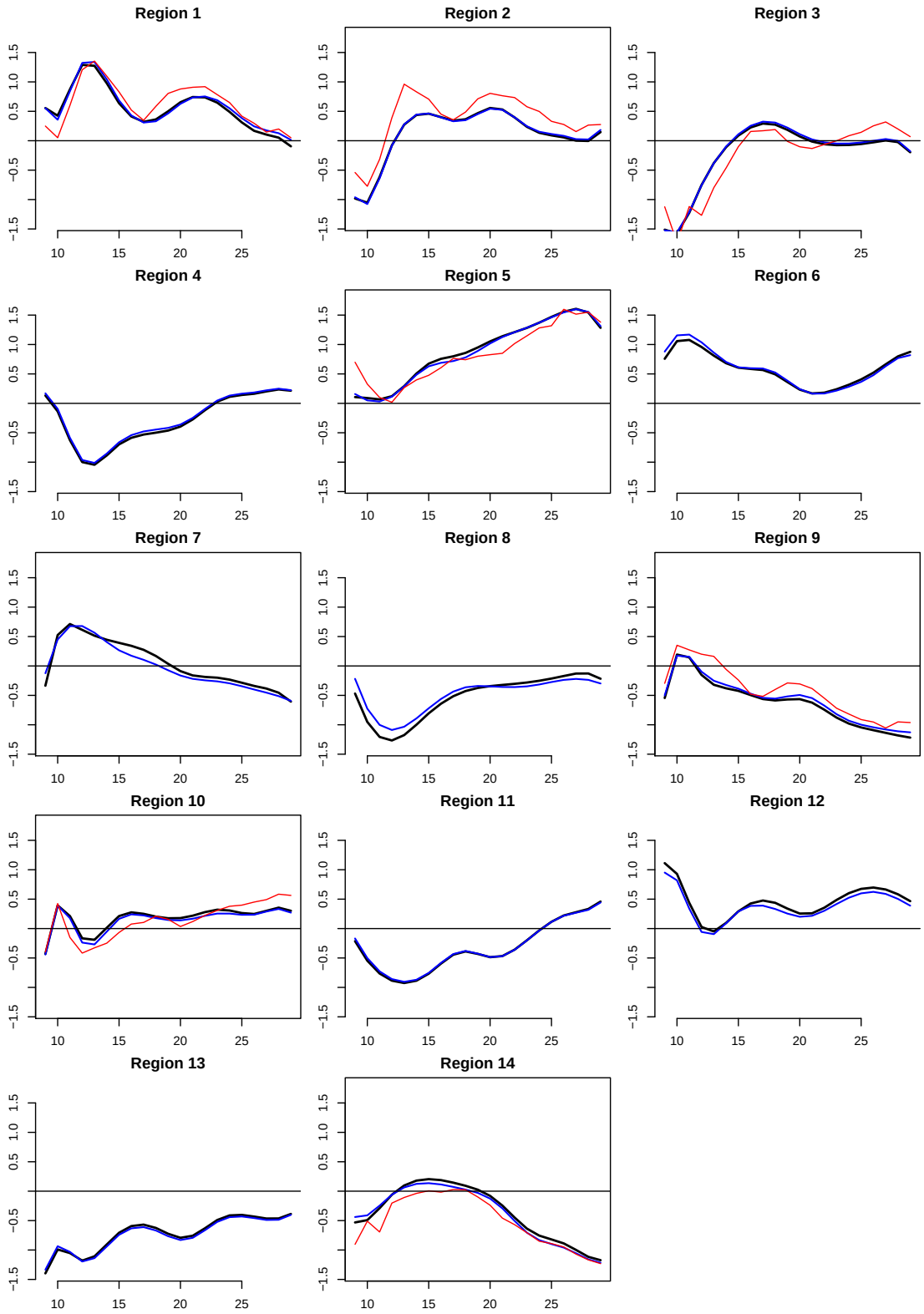


Fig. 8. Estimation of the effect for the marker which has the highest marginal posterior inclusion probability within each region in the $X_{0.95}$ subset. The blue, black, and red lines represent the estimation using the PS_1, PS_2, and RW_2 methods respectively. Plots with box are associated to markers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

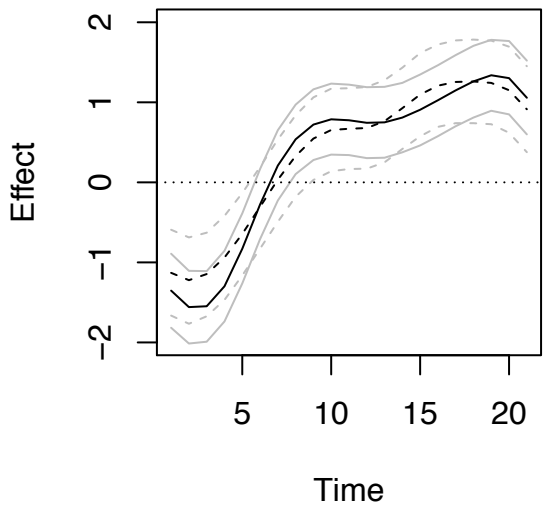


Fig. 9. Cumulative genetic effect of common markers selected in both conditions. The solid line represents the effect for the WW condition and the dashed line represents the effect of MWD conditions. Gray lines represent 95% credible intervals.

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.

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

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

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

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

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

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

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

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

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

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

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

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.

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

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

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

O'Sullivan, F. (1986) A statistical perspective on ill-posed in- verse problems (with 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 $\bar{y_i}=y_i-\alpha 1-\widetilde{B^t}\widetilde{m}-\sum_{l=1}^{L}\widetilde{B}^{e^l}\widetilde{a}_l-\sum_{j=1}^{J}x_{i,j}Zb_j$

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

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

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)

IV

Double sélection bayésienne de groupes de variables et de variables en présence d'une forte corrélation entre les variables au sein de chaque groupe

Sommaire

4.1 Introduction 85

4.2 Article : Bayesian sparse group selection with indexed regressors within groups: the group fused horseshoe prior 87

4.3 Application à l'identification de positions génétiques influant sur un caractère phénotypique 109

4.1 Introduction

Comprendre comment l'environnement influence les cultures est essentiel pour appréhender l'impact du changement climatique. Chez le palmier à huile par exemple, l'abscission des fruits doit se produire au bon moment et dans des conditions appropriées pour permettre une dispersion optimale des graines et ainsi assurer la survie de l'espèce. Cependant, il est connu que des facteurs environnementaux peuvent perturber ce processus de développement. Ces facteurs peuvent accélérer la chute du fruit avant la maturation des graines et des embryons, ou au contraire la retarder. Toutefois, une question subsiste : à quel stade du processus d'abscission ces facteurs influent ils ? Pour répondre à cette question, il est possible de mesurer différents facteurs environnementaux tout au long de ce processus comme proposé par Tisné et al. [2020]. Ainsi pour chaque indice d'abscission de fruit, des facteurs environnementaux sont enregistrés 180 jours avant la floraison et jusqu'à 180 jours après, par pas de temps de 3 jours,

aboutissant à 121 mesures. Pour chaque individu (fruit), l'ensemble des 121 mesures de chaque facteur forme alors un signal évoluant dans le temps propre à l'individu. Pour chaque facteur environnemental, la collection des signaux aboutit alors à une matrice de 121 variables ordonnées présentant une forte corrélation deux à deux. L'analyse d'un tel jeu de données peut être réalisé au travers d'un modèle de régression en considérant des groupes de variables avec une forte dépendance intra-groupe. Ce modèle peut également être vu comme un modèle de régression sur signaux étendue au cas de groupes de signaux. Le double objectif d'identification des facteurs environnementaux ayant un effet sur l'indice d'abscission, ainsi que l'identification des périodes de temps sur lesquelles ces facteurs influent, implique une double sélection de groupes et de variables sous la contrainte de la forte dépendance deux à deux intra-groupe.
Dans ce chapitre, nous proposons le prior *group fused horseshoe* permettant de généraliser le prior *global-local* au cas de la double sélection de groupes de paramètres et de paramètres ordonnés. Dans l'idée du prior *fused Lasso* introduit par Kyung et al. [2010], nous proposons de décomposer la matrice de précision associée à chaque groupe de coefficients lié à un facteur environnemental, comme la somme de deux matrices. La première est diagonale contenant des paramètres de variance locaux. Elle permet d'identifier les coefficients de régression non nuls à l'intérieur du groupe et donc d'identifier les périodes de temps où le facteur environnemental concerné influe sur l'indice d'abscission. La seconde matrice, non diagonale, permet de refléter la structure de dépendance présente à l'intérieur des groupes. Nous proposons de considérer une structure de type marche aléatoire avec des paramètres de variance locaux. Cette structure permet d'introduire un lien de continuité entre les coefficients deux à deux avec potentiellement des changements abrupts. Un paramètre de variance global ainsi qu'un paramètre de variance propre à chaque groupe viennent également compléter le prior. Le paramètre global à pour objectif de réduire l'ensemble des coefficients à zéro, tandis que le paramètre propre au groupe a pour objectif d'identifier les groupes pertinents, c'est à dire les facteurs ayant un effet sur l'indice d'abscission des fruits. Dans l'idée du prior *horseshoe* [Carvalho et al., 2009], nous considérons des distributions demi-Cauchy $\mathcal{C}^{+}(0,1)$ sur l'ensemble des paramètres de variance. Contrairement au prior proposé dans le chapitre précédent (cf chapitre 3), ce prior permet une sélection des paramètres à l'intérieur des groupes, aboutissant à une estimation de profils pouvant être non-nuls uniquement sur certaines périodes. De plus, la considération d'un paramètre de variance locale propre à chaque différence permet d'estimer des profils d'effets complexes (non-homogènes au cours du temps, changements abrupts). Nous montrons sur simulations que le prior proposé présente d'excellentes performances de sélection et d'estimation des effets, ou encore de prédiction, en comparaison avec d'autres priors tel que le *fused Lasso* [Kyung et al., 2010], *fused Student* [Rue and Held, 2005], ou encore des approches alternatives telles que la régression PLS et SPLS. Enfin nous appliquons le prior proposé sur les données réelles qui ont motivé ce développement. Quatre facteurs sur neuf sont alors identifiés par l'approche, avec des effets essentiellement sur les derniers stades du processus d'abscission.

Enfin, l'approche proposée ne se restreint pas uniquement à l'application l'ayant motivée, mais peut être mise en œuvre sur une large gamme d'applications. Dans le contexte de l'identification de positions génétiques influant sur un caractère phénotypique par exemple, nous montrons comment cette approche peut être mise en œuvre pour sélectionner les chromosomes et les marqueurs pertinents, tout en prenant en

compte la forte corrélation entre les marqueurs deux à deux. Elle permet d'identifier clairement les régions du génome impliquées dans la variabilité du caractère, et présente une réelle alternative aux approches classiquement utilisées telles que la régression Lasso ou Elastic-Net [Brault et al., 2020].

4.2 Article : Bayesian sparse group selection with indexed regressors within groups: the group fused horseshoe prior

L'article sera soumis à la revue *Journal of Agricultural, Biological and Environmental Statistics*.

Bayesian sparse group selection with indexed regressors within groups: the group fused horseshoe prior

B. Heuclin[1,2], J. Gibaud[1], F. Mortier[4,5], C. Trottier[1,6], S. Tisné[2,3], and M. Denis[2,3]

[1] IMAG, Univ Montpellier, CNRS, Montpellier, France,

[2] CIRAD, UMR AGAP Institut, F-34398 Montpellier, France

[3] UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, F-34398 Montpellier, France

[4] Forêts et Sociétés, Cirad, F-34398 Montpellier, France,

[5] Forêts et Sociétés, Univ Montpellier, Cirad, Montpellier, France,

[6] Univ Paul-Valéry Montpellier 3, Montpellier, France.

Abstract

The abscission process is strongly involved in a series of physiological events; an optimal execution of this process is of major importance for species survival. Environmental variables impact organ losses. Thus identifying their effects, as well as time periods involved in abscission process stages, is crucial to deal with climate changes. It raises two statistical challenges: selecting as a group each environmental variable impacting phenotypic variations, and selecting some variables within these groups while these variables are serially correlated. To answer both objectives we develop a group fused horseshoe prior. This prior encompasses main priors already proposed but present much better properties in terms of selection and estimation, allowing to identify either smooth or more complex effects. The fruit abscission of oil palm trees has motivated this development. This application based on an impressive experimental design in Benin Republic illustrates performances of the group fused horseshoe prior to select environmental variables as well as successive past environmental variations involved in bunch harvesting. Additional numerical simulations underline that this prior outperforms known methods.

Keywords: Bayesian variable selection, Fusion and Fused priors, Horseshoe prior, Structured variables.

1 Introduction

Understanding the impact of environmental variables on development and adaptation processes is crucial face to climate changes. The abscission consists in the shedding of various parts of organisms, such as leaves during autumn or flowers after fertilization. It is one of the most important adaptation process. This biological mechanism is highly sensitive to climate conditions and to their variations over the growing seasons and years, as illustrated by the well-known leaf senescence and abscission of deciduous trees, which was delayed in response to an increase in temperature between 1931 and 2010 in northern hemisphere (Gill et al., 2015). Environmental stresses may severely impact abscission processes due to complex regulations involving exogenous and endogenous signals (Sawicki et al., 2015). For instance, drought stress can induce activation and premature flower abscission in lupine (Wilmowicz et al., 2021) or tomato (Reichardt et al., 2020) and so impact crop productivity negatively.

In many contexts, while it is clear that environmental variables have consequences on organ losses, it is not yet clear which one, either exogenous or endogenous, are responsible of the responses observed and at which stages of the organ development or the abscission process the regulation occurs. The identification of the environmental variables and these stages appears to be important to gain insights into the impact of environmental conditions on adaptation abilities. In this paper, we aim to identify in the oil palm breeding context, the optimal period of bunch harvesting according to past environmental variations. In oil palm, the abscission time is critical because fruit bunches are harvested when the first fruits detach and fall to the ground. A premature abscission of fruits can lower the oil yield if the optimal maturity is not reached while too much abscission leads to extra work in order to collect detached fruits on the ground. A recent study have shown that environmental variables, such as temperature or solar radiation, alter the oil palm tree reproductive development by modulating the timing of fruit drop (Tisné et al., 2020). Identifying the relevant environmental variables and also the relevant time periods involved in the phenotypic variations of fruit abscission, raise at least two challenges both related to model regularization and variable selection. The first one is the selection of groups of variables (environmental variables). The second one is the selection of the variables within groups (time step). However, natural ordering of variables within groups can lead to potentially high correlation between consecutive variables. These dependencies have to be taken into account to avoid ill-conditioned problems and over-fitting, but also to better reflect reality and detect successive meaningful time periods.

Considerable attention has been paid in the last decades to variable and group selection. Developed methods are mainly related to penalized likelihood techniques in a frequentist context, or to the use of appropriate priors reflecting desired penalties in a Bayesian context. Among others the Lasso (Tibshirani, 1996), the SCAD (Fan and Li, 2001) or yet the Elastic-Net (Zou and Hastie, 2005) are classically used. Note that Elastic-Net is well adapted when variables are correlated. It is based on the combination of ℓ_1- and ℓ_2-norms in the penalization term, combining shrinkage properties from Lasso and regularization capacities from

Ridge regression (Hoerl and Kennard, 1970). In Bayesian multiple linear regression, the set of priors for variable selection has also been extensively developed, among others: the spike-and-slab prior (Mitchell and Beauchamp, 1988; George and McCulloch, 1993, 1997), the Bayesian Lasso prior (Park and Casella, 2008), the Elastic-Net prior (Kyung et al., 2010), the normal-gamma prior (Griffin et al., 2010) and the horseshoe prior (Carvalho et al., 2009; Piironen et al., 2017). Nevertheless, these methods do not take into account group structure within covariates. Lasso extensions to group selection have been developed in frequentist (Yuan and Lin, 2006) or Bayesian (Kyung et al., 2010; Liquet et al., 2017) contexts. To select sparse groups as well as variables within groups, Xu et al. (2015) proposed the sparse group Lasso prior. This approach mimics the frequentist sparse group Lasso penalty introduced by Simon et al. (2013). Xu et al. (2016) extended such a prior considering a horseshoe prior and a scale mixture of independent Gaussian distributions with three levels variance parameters: one global and common to all coefficients, one specific to each group and one for each coefficient.

The above methods do not allow to take into account serial correlations between successive variables within groups. These dependencies may lead to identifiability problems impacting the estimation task which aims to assign similar effects for two adjacent variables. To allow the integration of this information and to constrain estimation, in a linear regression context, the fusion and fused Lasso are introduced by Land and Friedman (1997) and Tibshirani et al. (2005). The fusion Lasso penalizes the ℓ_1-norm of successive differences of parameters, the fused Lasso combines the fusion Lasso with the usual Lasso penalization on individual coefficients. Kyung et al. (2010) proposed a Bayesian fused Lasso with a Laplace distribution on differences and also on individual coefficients in the linear regression context. However, various studies pointed out that the Bayesian Lasso prior does not shrink enough individual coefficients or differences towards zero, leading to biased (Carvalho et al., 2009; Polson and Scott, 2010) and smooth estimations without possible abrupt changes (Faulkner and Minin, 2018). To allow more flexibility and sparser estimations, other shrinkage priors, with stronger mass on zero and heavier tails, have been investigated. For instance, Rue and Held (2005) and Song and Cheng (2018) used a Student distribution on the differences, while Faulkner and Minin (2018) proposed the horseshoe prior in a Bayesian approach. These methodologies show good properties allowing the estimation of smooth functions but also to detect abrupt changes. However, all these approaches have been designed for only one group. Recently, Zhang et al. (2014) introduced the group spike-and-slab prior combined with the Bayesian fused Lasso. This method can suffer from low shrinkage properties of the Bayesian Lasso and leads to poor estimations when the number of covariates within groups is large. In this paper, to overcome such limitations, we propose the group fused horseshoe prior encompassing some previously developed priors and allowing a double selection as well as the estimation of either smooth or more complex effect profiles over time.

This paper is organized as follows. Section 2 is dedicated to the construction of the general group fused horseshoe prior in a linear regression context. Section 3 presents how and why

this class of prior is helpful to select environmental variables clearly involved in the abscission process, but also to delineate crucial environmental time periods implicated. Section 4 is finally dedicated, using simulated data, to evaluate the efficiency of the method derived from this prior, but also to compare its performances face to more classical or already known methods. In particular, we propose to compare our method to five alternative techniques: (i) the sparse partial least squares regression (Hoerl and Kennard, 1970; Kim et al., 2009), (ii) the Elastic-Net regression (Zou and Hastie, 2005), (iii) the bi-level selection using the component MCP penalty (Breheny and Huang, 2009), (iv) the composite MCP penalty combined with the Ridge penalty (Breheny and Huang, 2009) and (v) the group spike-and-slab prior combined with the Bayesian fused Lasso (Zhang et al., 2014).

2 Statistical model and prior construction

Let $\mathbf{y} = (y_1, \ldots, y_n)'$ be the $n-$response vector of a phenotypic trait and $\mathbf{X_g} = [\mathbf{x_{g1}}, \ldots, \mathbf{x_{gT}}]$ be the $(n \times T)$-matrix associated to the g^{th} environmental variable ($g = 1, \ldots, G$) and measured at T ordered time points ($t = 1 < 2 < \ldots < T$). The $\mathbf{X_g}$ matrix is thus considered as a group of ordered variables. In our application, $G = 9$ environmental variables are measured over $T = 121$ time points (one year measurements over a 3 days grid). We assume that the response variable is related to time varying environmental variables through a linear regression model:

$$\mathbf{y} = \mu + \sum_{g=1}^{G} \mathbf{X_g}\boldsymbol{\beta_g} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I_n) \tag{1}$$

where μ is an intercept, $\boldsymbol{\beta_g}$ a T-vector of regression coefficients associated to variables of group g and $\boldsymbol{\varepsilon}$ a n-vector of independent Gaussian residuals with zero mean and variance equal to σ^2.

In this paper, we propose to consider a prior combining the horseshoe (Carvalho et al., 2009; Polson and Scott, 2010) and the fused (Tibshirani et al., 2005; Zhang et al., 2014) priors to select groups and variables within groups, but also time periods involved in phenotypic variations. This prior takes advantage of the effectiveness of the fused prior to take into account serial correlation between variables, smoothing environmental signal, but also of shrinkage properties of the horseshoe prior to allow possible abrupt changes in the estimation. Hereafter, we will refer to this prior as the group fused horseshoe prior.
Our approach consists in modelling regression coefficients associated to each group g with the following multivariate Gaussian distribution:

$$\boldsymbol{\beta_g} | \boldsymbol{\upsilon_g}, \tau^2, \lambda_g^2, \boldsymbol{\omega_g}, \sigma^2 \sim \mathcal{N}_T(0, \sigma^2 {\boldsymbol{Q_g}}^{-1}), \quad g = 1, \ldots, G \tag{2}$$

where $\boldsymbol{Q_g}$ is a precision matrix equal to:

$$\boldsymbol{Q_g} = \left(\frac{1}{\tau^2 \lambda_g^2} {\boldsymbol{D_g}}^{(k)\prime} {\boldsymbol{\Omega_g}}^{-1} \boldsymbol{D_g^{(k)}} + {\boldsymbol{\Upsilon_g}}^{-1} \right). \tag{3}$$

$\boldsymbol{Q_g}$ is divided into two parts. The first part is related to the differences, the second one to the effect of each variable. τ^2 corresponds to the global variance parameter of independent groups, λ_g^2 the specific group global variance parameter, $\boldsymbol{D_g^{(k)}}$ a known $T \times (T-k)$-matrix associated to the finite differences operator of order k, and $\boldsymbol{\Omega_g} = diag(\omega_{g_1}^2, \ldots, \omega_{g_{T-k}}^2)$ the $(T-k) \times (T-k)$-diagonal matrix of local variance parameters. The first part of the precision matrix aims to control dependencies between adjacent regression coefficients, leading to more or less smooth coefficients profiles over time according to the k-th order difference. In the following, we will focus on the first-order differences to estimate piecewise-constant profile (Tibshirani et al., 2005) in order to identify relevant time periods. However, when the number of variables is larger than the number of observations, considering only this term, as proposed in trend filtering problems without group consideration (Faulkner and Minin, 2018), turns out to be insufficient to shrink coefficients towards zero (Tibshirani et al., 2005). To overcome such limitations, a $T \times T$-diagonal matrix of local variance parameters $\boldsymbol{\Upsilon_g} = diag(\upsilon_{g_1}^2, \ldots, \upsilon_{g_T}^2)$ is added to shrink each regression coefficient. This prior encompasses a broad class of local-global shrinkage priors classically used either in individual variable selection or in structured variables contexts (Table 2 in Appendix A.1).

We complete the Bayesian formulation assuming that all unknown quantities in $\boldsymbol{Q_g}$ and the residual variance follow a half-Cauchy distribution, $\mathcal{C}^+(0,1)$. The intercept is assumed to be uniformly distributed on $\mathbb{R}$.
In this paper, we use the scale mixture representation of the half-Cauchy distribution (Makalic and Schmidt, 2015). More precisely, the half-Cauchy distribution can be expressed as a scale mixture of inverse-gamma distributions:

$$\sigma^2 \sim \mathcal{C}^+(0,1) \quad \Leftrightarrow \quad \sigma^2|a \sim \mathcal{IG}(1/2, 1/a), \quad a \sim \mathcal{IG}(1/2, 1/2).$$

This representation allows to develop an efficient MCMC algorithm. Full conditional distributions have a closed form (see appendix A.2). We propose a Gibbs sampler algorithm (Gilks et al., 1995). Code is available in the R language (Team et al., 2013) on GitHub: `https://github.com/Heuclin/GroupFusedHorseshoe`.

3 The abscission dataset

This application aims at identifying environmental variables and time periods affecting the oil palm fruit abscission process (Tisné et al., 2020). The dataset is provided by "le Centre de Recherches Agricoles-Plantes Pérennes (CRA-PP)" from the republic of Benin which carried out an experimental trial to follow-up, from 2014 until 2018, a self-pollinated population of 138 oil palm trees planted between 2000 and 2005. For each tree, the pollination date was recorded, and each pollinated bunch was monitored up to its harvest. 1,173 bunches were considered over multiple years, taking advantage of the climatic seasonality and the continuous fruit production of this species. We used the days from pollination to fruit drop (DFD) as the response variable. DFD is the classical harvest time indicator and its variation integrates different underlying abscission processes at different developmental stages.

Additionally, five climatic variables were recorded from 2014 until 2018: the maximum and minimum temperature (Tmax, Tmin, in °C), the relative air humidity (RH, in %), the rainfall (R, in mm), and the solar radiation (SR, in cal.cm^{-2}.d^{-1}). Five ecophysiological variables were calculated using climate and individual production data: three environmental variables including the maximum daily vapor pressure deficit (VPD), the fraction of transpirable soil water (FTSW), and two trophic variables: the supply–demand ratio (SD) and the daily reproductive demand (DRD) (see Tisné et al. (2020) for details of the calculations). These variables can have ponctual or cumulative effects, depending on the biological process or the developmental stage. Hence, for example, temperature can have ponctual effects, for instance the arrest of growing at low temperature, but also cumulative effects on developmental rates that led to the thermal time development. A three-day time grid, from −180 (individualization of the floral meristem) to +180 (ripe fruit) days after pollination, was used to calculate either the average values over three days (Tmax, Tmin, RH, VPD, FTSW, DRD, and SD) or the cumulative values over 15 days (R and SR) of each variable. Nine matrices of 1173 rows and 121 columns associated with each variable were then constructed. Within each matrix the i^{th} row corresponds to the i^{th} bunch analyzed and the t^{th} column corresponds to the value of the corresponding climatic/ecophysiological variable at time t for each bunch. All matrices have been scaled to obtain a similar order of magnitude. To apply our group fused horseshoe prior, 50 MCMC chains are initialized at random starting values, each with 50,000 iterations, a burn-in of 20,000 and a thining of 10. A group is considered selected if at least one regression effect within it has a credible interval that does not contain zero.

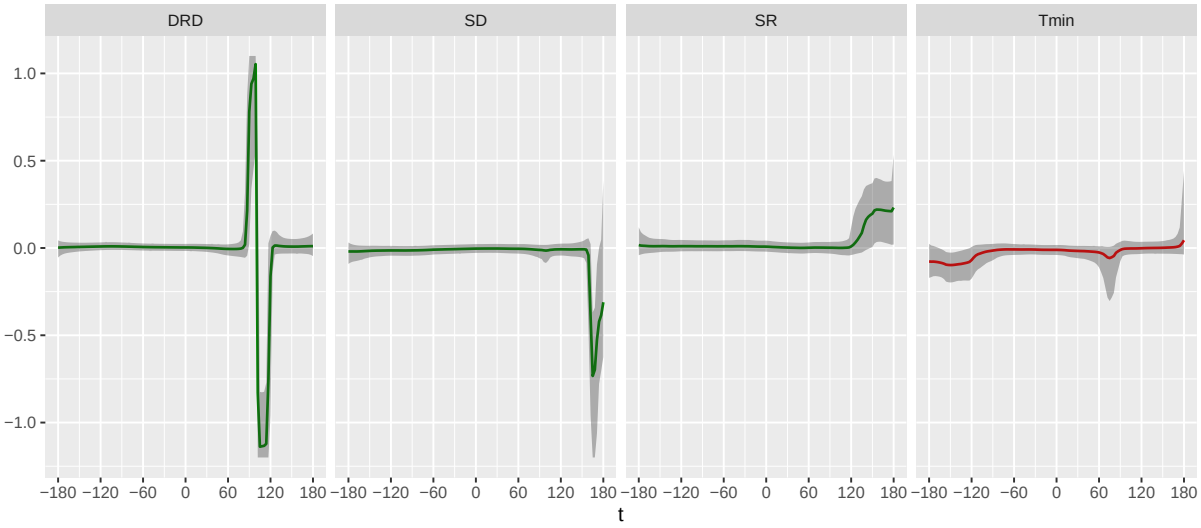


Figure 1: Non-zero coefficient profile estimation provided by the group fused horseshoe prior on the abscission dataset. Gray shadows represent the 95% credible interval. Colors represent the different categories of environmental variables, green is for photosynthesis variables (DRD, SR, SD) and red is for temperature variable.

Comparison with previous studies and biological interpretation

Estimated coefficient profiles provided by the group fused horseshoe prior are very clear and allow to identify relevant time periods of four variables (Tmin, SR, DRD, and SD). Two types of patterns are observed: with smooth effects for Tmin and SR and with punctual effects for DRD and SD. The Tmin variable has a negative effect during the inflorescence development from day -180 to -100, while the three other variables have an effect at the end of the fruit bunch development. SR, the solar radiation variable, displays a positive effect from day 120 to 180 at the final stage before the fruit drop. The DRD variable has punctual effets at days 99 and 105 after pollination with first a positive effect before an inversion of its effect direction at day 100. The SD factor has a negative effect that peaks at day 160. The striking pattern of DRD around day 100 after pollination observed in Tisné et al. (2020), is thus confirmed and corresponds to the "lag period" of the oil palm fruit bunch development between the cell division/expansion phase and the maturation phase (Tranbarger et al., 2011). The selection of the DRD variable at this key developmental stage suggests that the considered fruit bunch integrates current and future whole plant photosynthate demand due to concomitant developing bunches, and this, in order to modulate its maturation and abscission timing. Such carbohydrate based regulation is commonly found in fruit tree species and lead to the wave of abscission concerning fruitlets (Sawicki et al., 2015), the only difference being that the oil palm regulates ripe fruit abscission timing instead of dropping unripe fruits. In contrast with DRD and SD that have similar punctual patterns with those of Tisné et al. (2020) study, Tmin effect profile is different, showing a continuous moderate effect instead of many weak effects spread over the -180 to -100 period. The SR factor was not selected by Tisné et al. (2020) while it has a positive effect from day 120 to 180 using our prior. These discrepancies may be due to the cumulative nature of both Tmin and SR effects at their respective developmental stages. Hence, the Tmin effect at the early inflorescence developmental stages could be related to thermal time which is known to be associated with developmental rates. In the period identified, the differentiation of floral organs occurs (Adam et al., 2011) and variation in cumulative thermal time could modulate the developmental program and ultimately the fruit drop timing. Concerning the cumulative effect of the radiation, it was identified all over the final stage before fruit drop that corresponds to the fruit maturation with intensive lipid accumulation, which is highly related to photosynthate availability (Tranbarger et al., 2011). Our proposed prior, which has been designed to estimate smooth and flexible coefficient profile is then well suited to study the effect of cumulative effect variables in addition to the punctual effect variables that were identified consistently between both approaches.

4 Simulations

This section aims at investigating, through a simulation study, the performances of the group fused horseshoe prior we propose, denoted in the following group fused HS 3L. We intend to study the impact of the assumptions we made, such as considering sparsity of the coefficients

and of their differences, but also to show the benefits of using three levels variance parameters on coefficient differences. The first objective is achieved comparing the group fused HS 3L prior with a fusion version (the group fusion HS 3L). This latter prior consists in retaining only the first part of the precision matrix $\boldsymbol{Q_g}$ (cf equation 3) and thus applying penalties only on differences. The second objective is realized by fixing the group specific variance parameters λ_g^2 to one, leading to the group fusion/fused HS 2L/3L priors. We also compare our results with those obtained using several other variable selection approaches that handle the group structure and/or the correlation between variables. In particular, we focus on the sparse PLS regression (SPLSR) and the Elastic-Net regression (ENR), which do not consider the group structure information but handle the predictors' correlation. We also compare the proposed prior with the composite MCP penalty (cMCP), and (cMCP_Ridge) (Breheny and Huang, 2009). Those approaches perform a double selection (group and variable within group), and the latter also allows to handle the correlation between variables within groups but does not consider prior information on the structure between them. Note that we also try to run the Bayesian hierarchical structured variable selection approach introduced by Zhang et al. (2014). This approach allows a double selection of groups and structured parameters within groups by combining a group spike-and-slab with a fused Laplace prior on the slab part. Unfortunately, the results reveal a high sensitivity to the number of variables within groups leading to very poor results.

To mimic a real case study, we consider $n = 500$ observations with $G = 10$ groups. The number of variables within groups is set to $T = 150$, leading to a total number of variables equal to $1,500$. Within each group, variables are generated from a multivariate Gaussian distribution with zero mean and a covariance matrix defined by a first-order autoregressive structure with a parameter equal to 0.95. We consider that only groups one, three and five have non-null coefficients with profiles defined by the functions $sin(4t/T-2)+2exp(-30(4t/T-2)^2)$, $\mathbb{1}_{(0.1T\leq t)}$ and $\mathbb{1}_{(0.8T\leq t)}$ respectively, with $t = 1, \ldots, T$. The first function was initially presented by DiMatteo et al. (2001) and frequently used by others, including Faulkner and Minin (2018). Finally, we assume that the residuals are distributed from a normal distribution with 0 mean and variance equal to 2.

For the eight methods, results are based on 5 replicates of a 10-fold cross-validation (CV) procedure, leading to 50 repetitions. For the Bayesian approaches, it results in 50 MCMC chains randomly initialized. Each MCMC algorithm was run for $50,000$ iterations with a burn-in of $10,000$ samples and a thinning of 10. For the frequentist approaches, different R packages were used to run the four methods: the *spls* package (Chung et al., 2019) for SPLSR, the *glmnet* R package (Friedman et al., 2010) for ENR , and the *grpreg* R package (Breheny and Huang, 2009) for the cMCP and cMCP_Ridge approaches. For all of the four methods, penalty parameters were estimated using a 5-fold CV procedure.

To evaluate variable selection performances, we compute the rate (in percentage) of false positives (FPRv) and false negatives (FNRv), as well as the Matthews correlation coefficient

(MCCv) (Matthews, 1975). For the frequentist approaches, a variable is selected if its coefficient is different from zero. For the Bayesian ones, a variable is assumed to be selected if zero does not belong to its 95% credible interval. For all methods, a group is deemed relevant if it contains at least one selected variable. Note that 95% credible intervals are calculated on the estimations associated with all repetitions. The performance in terms of group selection is assessed using the same criterion, but at the group level (FPRg, FNRg, MCCg respectively). To access the estimation accuracy, we record the 95% credible interval width (CIW), the root mean squared error (RMSE) of the non-null coefficients (RMSE_1) and of the null coefficients (RMSE_0). We also compute the RMSE of the non-null coefficients associated with the varying function (RMSE_v) and with piecewise constant functions (RMSE_c). Finally, to investigate prediction performances, we compute the predictive RMSE (RMSE_P).

Approach	$\mathbf{RMSE_p}$	$\mathbf{RMSE_1}$	$\mathbf{RMSE}_v$	$\mathbf{RMSE}_c$	$\mathbf{RMSE_0}$	**CIW**
group fused HS 3L	1.28	0.044	0.064	0.002	0	0.040
group fused HS 2L	1.31	0.035	0.050	0.006	0	0.140
group fusion HS 3L	1.29	0.022	0.033	0.002	0	0.030
group fusion HS 2L	1.59	0.025	0.035	0.010	0.004	0.120
SPLSR	1.78	0.053	0.049	0.057	0.018	
ENR	1.58	0.058	0.060	0.056	0.008	
cMCP	6.04	0.409	0.379	0.435	0.012	
cMCP_Ridge	1.41	0.046	0.047	0.046	0.005	

Approach	$\mathbf{FPR_v}$	$\mathbf{FNR_v}$	$\mathbf{MCC_v}$	$\mathbf{FPR_g}$	$\mathbf{FNR_g}$	$\mathbf{MCC_g}$
group fused HS 3L	0	2	0.99	0	0	1
group fused HS 2L	0	2	0.99	0	0	1
group fusion HS 3L	0.3	1	0.99	7	0	0.92
group fusion HS 2L	0.7	1	0.98	42	0	0.61
SPLSR	7	1	0.86	3	0	0.96
ENR	6	1	0.87	74	0	0.42
cMCP	2	67	0.45	0	0	1
cMCP_Ridge	2	0	0.96	14	0	0.84

Table 1: RMSE_p, RMSE_1, RMSE_v, RMSE_c, RMSE_0, CIW, FPR_v, FNR_v, MCC_v, FPR_g, FNR_g and MCC_g criteria obtained using the different approaches on the simulated dataset and averaged over repetitions.

Performances of the group fused HS 3L prior and comparison with alternative sub-priors (group fused HS 2L and group fusion HS 2L and 3L)

In terms of variable selection, the group fused and fusion HS 3L and 2L priors give similar results with MCC_v values close to 1 (Table 1). However, we observe some differences: the fused priors miss a few relevant variables (2% of FNR_v), while the fusion priors tend to select a small number of false positives (0.3% and 0.7% of FPR_v respectively). Thus, by considering a double sparsity, fused priors encourage a strong shrinkage toward zero, whereas fusion priors, which assume only sparsity of the differences, may not shrink enough. Although these

differences do not impact the variable selection, this greatly affects the group selection, as displayed in Table 1 where MCC_g values are lower for fusion priors than for fused priors, which end up with values equal to 1. In fact, in the case of fusion priors, most false positive variables are detected in non relevant groups. This result, combined with the criterion chosen to select groups (one group is deemed if at least one variable within the group is selected) explains the poor group selection performances. In terms of estimation, fusion priors yield better $RMSE_1$ values along with narrower credible interval widths compared to fused priors. Nevertheless, the $RMSE_v$ and $RMSE_c$ values reveal that the estimation accuracy depends on the type of coefficient profiles. Using first-order differences, the fused priors encourage, as expected, piecewise constant estimations, while the fusion versions tend to provide smoother estimations involving lower $RMSE_v$, and higher $RMSE_c$ and $RMSE_0$. With regards to predictive performances, fused priors outperform fusion priors. This result may be partly explained by the fact that the double sparsity leads to less biased estimation of coefficients, especially for the null coefficients, hence avoiding over-fitting. These results point out the necessity to add coefficients' sparsity for a better regularization of the model. We also examine the results obtained with the 2L and 3L versions of each prior. The inclusion of group specific variance parameters conducts to a better estimation with less variability in the coefficient estimations (narrower CIW) and improves the quality of prediction. This also encourages piecewise constant profiles and sparse estimations which results in slightly higher $RMSE_v$, and lower $RMSE_c$ and $RMSE_0$ for the 3L versions.

All of these results support the interest of considering three levels variance parameters combined with double sparsity of coefficients and of their differences either for selection, estimation or prediction. The proposed prior provides a good trade-off between sparsity, which is required in many applications, and flexibility in the estimation of different types of profiles by encouraging piecewise constant estimations.

Comparison with alternative approaches

The group fused HS 3L prior outperforms all frequentist approaches considered in terms of estimation, selection and prediction (Table 1). With regards to the methods considering the correlation between variables but not the group structure (SPLSR and ENR approaches), Table 1 shows that both approaches have worse performances for identifying the relevant variables, with MCC_v values around 0.86. In the case of ENR, this affects greatly its ability to identify relevant groups. Concerning approaches performing a double selection, cMCP_Ridge outperforms cMCP in terms of prediction and variable selection with a $RMSE_p$ three times smaller and a MCC_v two times greater. This result is mainly due to the fact that cMCP tends to select only a few representatives among correlated variables, which leads to a high number of false negatives (67%). Also, we observe that cMCP_Ridge yields better results than the SPLSR and ENR methods and provides the best results after the proposed prior.

Through these comparisons, we emphasize the interest of handling both structures in and

within groups, but we also point out the necessity of considering the structure between covariates as done with the group fused HS 3L prior. For a better understanding, we display in Figure 2 in Appendix A.3 the estimated coefficients profiles averaged over the 50 repetitions along with the 95% confidence intervals for the eight approaches. While most approaches are able to fit different types of profiles, we notice a higher variability of estimations for frequentist approaches. Indeed, since the proposed prior takes into account the indexation structure information, smoothness over adjacent coefficients is encouraged and conducts to more stable results.

To go further, we also investigate more complicated simulations considering $T = 300$ variables within each group, leading to a total of $3,000$ variables, all other settings were unchanged. We defer results to the Appendix (Table 3 and Figure 3 in Appendix A.3). Globally, this scenario leads to the same conclusions than previously, but highlights more pronounced differences between approaches. Thus, the importance of integrating sparsity of the coefficients and their differences is clearly demonstrated with the systematic crash of the fusion versions. We also observe that a higher number of variables impacts greatly the quality of prediction of the SPLSR, ENR, and cMCP approaches with values of RMSE_p twice higher than under the scenario with $1,500$ variables. Finally, this simulation shows again that group fused HS 3L prior gives the best results followed by cMCP_Ridge approach.

5 Conclusion

We propose a group fused horseshoe prior, combining the horseshoe and fused priors, for selecting groups and variables within groups while taking into account the structure between variables. It extends the approach from Faulkner and Minin (2018) to the multi-group regression context. It also can be considered as an alternative to the fused Laplace prior in the case of a large number of variables within groups (Zhang et al., 2014). Finally, it encompasses some well known sub-priors as shown in Table 2 from Appendix A.1.

Through simulations, we demonstrate better performances of the proposed prior compared to alternative approaches commonly used to analyze grouped and/or correlated variables. By investigating scenarios with a different number of variables, we show the importance of adding local variance parameters to shrink more strongly coefficients toward zero. This helps for group selection and prevents over-fitting, especially when a high number of predictors is considered. We also evidence the benefit of considering three levels variance parameters on coefficient differences (global, group specific and local), which allows to address various situations with different levels of sparsity between and within groups, while ensuring an overall shrinkage. While the group specific variance parameters allow to consider a group structure, the local variance parameters provide, as demonstrated in Faulkner and Minin (2018), flexibility in the estimation of smooth and/or complex coefficient profiles with abrupt changes. In addition, we show the importance of integrating the correlation structure into models, which helps for model building and increases the detection power.

Finally, whatever the number of variables considered, we demonstrate the robustness of our approach, with more stable results and lower variability over repetitions than the other approaches. To sum up, we clearly point out the advantages of the group fused horseshoe prior prior to integrate the group structure and to obtain sparse estimations, and the benefits of the fused prior to consider the variable structure.

From a biological point of view, the proposed prior clearly identifies four environmental variables as well as periods at which they affect the oil palm abscission process. By providing flexibility in the estimation of regression coefficient profiles, we identify one supplementary environmental variable than the previous study, and improve the interpretability of the regression profiles. Moreover, by giving high predictive performances, the proposed prior may be an useful tool to assist the biologists in identifying the best time to harvest the bunches.

Our proposed prior may be directly applied on a broad type of applications such as in the near infrared spectroscopy context, which involves one group of ordered variables through a spectrum, or in the genetic mapping context, where markers may be viewed as groups of ordered variables at the chromosome level. Moreover, our prior may be used in various models as varying coefficient models (Heuclin et al., 2020), additive models (Scheipl et al., 2012) or TVP-VAR models (Koop and Korobilis, 2013; Bitto and Frühwirth-Schnatter, 2019), which involve (independent) groups of ordered regression coefficients. Finally, some extensions may be proposed. The first consists in extending the proposed model to analyze binary, binomial, or Poisson responses. The second aims at considering a multi-dimensional indexation instead of only one dimensional indexation (through time in our case). Indeed, in many applications, variables may be structured according to various factors. For instance, in the disease mapping context, observations are structured in time and space. This extension may be achieved by considering a Kronecker product between finite difference operator matrices associated to each dimension.

Acknowledgments: J. Gibaud, F. Mortier and C. Trottier were supported by the GAMBAS project funded by the French National Research Agency (ANR-18-CE02-0025). M. Denis was fully supported by the European Union's Horizon 2020 Research and Innovation program under grant agreement No 840383. We thank all people from Cirad/PalmElit (France) who planned this trial. We acknowledge "le Centre de Recherches Agricoles-Plantes Pérennes (CRA-PP)" (the republic of Benin) for planting, observing and collecting data, and authorizing use of the phenotypic data for this study.

References

Adam, H., Collin, M., Richaud, F., Beulé, T., Cros, D., Omoré, A., Nodichao, L., Nouy, B., and Tregear, J. W. (2011). Environmental regulation of sex determination in oil palm: current knowledge and insights from other species. *Annals of botany*, 108(8):1529–1537.

Bitto, A. and Frühwirth-Schnatter, S. (2019). Achieving shrinkage in a time-varying parameter model framework. *Journal of Econometrics*, 210(1):75–97.

Breheny, P. and Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and its interface*, 2(3):369.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, pages 73–80.

Chung, D., Chun, H., Todorov, M. V., and Imports, M. (2019). Package 'spls'.

DiMatteo, I., Genovese, C. R., and Kass, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika*, 88(4):1055–1071.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Faulkner, J. R. and Minin, V. N. (2018). Locally adaptive smoothing with markov random fields and shrinkage priors. *Bayesian analysis*, 13(1):225.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica sinica*, pages 339–373.

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

Gill, A. L., Gallinat, A. S., Sanders-DeMott, R., Rigden, A. J., Short Gianotti, D. J., Mantooth, J. A., and Templer, P. H. (2015). Changes in autumn senescence in northern hemisphere deciduous trees: a meta-analysis of autumn phenology studies. *Annals of botany*, 116(6):875–888.

Griffin, J. E., Brown, P. J., et al. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian analysis*, 5(1):171–188.

Heuclin, B., Mortier, F., Trottier, C., and Denis, M. (2020). Bayesian varying coefficient model with selection: An application to functional mapping. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Kim, S.-J., Koh, K., Boyd, S., and Gorinevsky, D. (2009). \ell_1 trend filtering. *SIAM review*, 51(2):339–360.

Koop, G. and Korobilis, D. (2013). Large time-varying parameter vars. *Journal of Econometrics*, 177(2):185–198.

Kyung, M., Gill, J., Ghosh, M., Casella, G., et al. (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5(2):369–411.

Land, S. R. and Friedman, J. H. (1997). Variable fusion: A new adaptive signal regression method. Technical report, Technical Report 656, Department of Statistics, Carnegie Mellon University.

Lang, S. and Brezger, A. (2004). Bayesian P-Splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212.

Liquet, B., Mengersen, K., Pettitt, A., Sutton, M., et al. (2017). Bayesian variable selection regression of multivariate responses for group data. *Bayesian Analysis*, 12(4):1039–1067.

Makalic, E. and Schmidt, D. F. (2015). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Piironen, J., Vehtari, A., et al. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2):5018–5051.

Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian statistics*, 9(501-538):105.

Reichardt, S., Piepho, H.-P., Stintzi, A., and Schaller, A. (2020). Peptide signaling for drought-induced tomato flower drop. *Science*, 367(6485):1482–1485.

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

Sawicki, M., Aït Barka, E., Clément, C., Vaillant-Gaveau, N., and Jacquard, C. (2015). Cross-talk between environmental stresses and plant metabolism during reproductive organ abscission. *Journal of Experimental Botany*, 66(7):1707–1719.

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(500):1518–1532.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245.

Song, Q. and Cheng, G. (2018). Bayesian fusion estimation via t shrinkage. *Sankhya A*, pages 1–33.

Team, R. C. et al. (2013). R: A language and environment for statistical computing.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

Tisné, S., Denis, M., Domonhédo, H., Pallas, B., Cazemajor, M., Tranbarger, T. J., and Morcillo, F. (2020). Environmental and trophic determinism of fruit abscission and outlook with climate change in tropical regions. *Plant-Environment Interactions*, 1(1):17–28.

Tranbarger, T. J., Dussert, S., Joët, T., Argout, X., Summo, M., Champion, A., Cros, D., Omore, A., Nouy, B., and Morcillo, F. (2011). Regulatory mechanisms underlying oil palm fruit mesocarp maturation, ripening, and functional specialization in lipid and carotenoid metabolism. *Plant physiology*, 156(2):564–584.

Wilmowicz, E., Kućko, A., Pokora, W., Kapusta, M., Jasieniecka-Gazarkiewicz, K., Tranbarger, T. J., Wolska, M., and Panek, K. (2021). Epip-evoked modifications of redox, lipid, and pectin homeostasis in the abscission zone of lupine flowers. *International journal of molecular sciences*, 22(6):3001.

Xu, X., Ghosh, M., et al. (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Analysis*, 10(4):909–936.

Xu, Z., Schmidt, D. F., Makalic, E., Qian, G., and Hopper, J. L. (2016). Bayesian grouped horseshoe regression with application to additive models. In *Australasian Joint Conference on Artificial Intelligence*, pages 229–240. Springer.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhang, L., Baladandayuthapani, V., Mallick, B. K., Manyam, G. C., Thompson, P. A., Bondy, M. L., and Do, K.-A. (2014). Bayesian hierarchical structured variable selection methods with application to molecular inversion probe studies in breast cancer. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(4):595–620.

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

A Appendix

A.1 Known priors derived from the group fused horseshoe prior

	Prior	$\upsilon^2_{g_t}$	τ^2	λ^2_g	$\omega^2_{g_j}$
One group	Random walk Lang and Brezger (2004) Rue and Held (2005)	-	-	$\mathcal{IG}(s, r)$	1
	Fusion Laplace Kyung et al. (2010)	-	-	1	$\mathcal{E}xp(\lambda/2)$
	Fusion Student $t_{df}(s\sigma)$ Rue and Held (2005) Song and Cheng (2018)	-	-	1	$\mathcal{IG}(df/2, s^2df/2)$
	Fusion horseshoe Faulkner and Minin (2018)	-	-	$\mathcal{C}^+(0, 1)$	$\mathcal{C}^+(0, 1)$
	Fused Laplace Kyung et al. (2010)	$\mathcal{E}xp(\lambda_1/2)$	-	1	$\mathcal{E}xp(\lambda_2/2)$
	Group fused horseshoe	$\mathcal{C}^+(0, 1)$	$\mathcal{C}^+(0, 1)$	$\mathcal{C}^+(0, 1)$	$\mathcal{C}^+(0, 1)$

Table 2: Priors included in the group fused local-global prior (see equation 2). A dash indicates that the parameter is not considered in the prior, 1 indicates that the parameter is fixed to one.

A.2 Bayesian hierarchical model and inference

The Bayesian hierarchical model using our proposed group fused horseshoe prior is given by:

$$
\begin{aligned}
&\boldsymbol{y}|\mu, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}_n(\mu \mathbb{1} + \sum_{g=1}^{G} \boldsymbol{X_g}\boldsymbol{\beta_g},\ \sigma^2 I_n) \\
&\mu \sim \mathcal{U}_{(-\infty,\infty)} \\
&\boldsymbol{\beta_g}|\boldsymbol{v_g}, \tau^2, \lambda_g^2, \boldsymbol{\omega_g}, \sigma^2 \sim \mathcal{N}_T\left(0, \sigma^2\left(\frac{1}{\tau^2\lambda_g^2}\boldsymbol{D_g^{(k)\prime}}{\boldsymbol{\Omega_g}}^{-1}\boldsymbol{D_g^{(k)}} + {\boldsymbol{\Upsilon_g}}^{-1}\right)^{-1}\right) \\
&\tau^2|\xi \sim IG\left(\frac{1}{2}, \frac{1}{\xi}\right),\ \ \xi \sim IG\left(\frac{1}{2}, 1\right) \\
&\lambda_g^2|\psi_g \sim IG\left(\frac{1}{2}, \frac{1}{\psi_g}\right),\ \ \psi_g \sim IG\left(\frac{1}{2}, 1\right),\ \ g = 1, \ldots, G \\
&\omega_{g_j}^2|\phi_{g_j} \sim IG\left(\frac{1}{2}, \frac{1}{\phi_{g_j}}\right),\quad \phi_{g_j} \sim IG\left(\frac{1}{2}, 1\right),\quad g = 1, \ldots, G,\ \ j = 1, \ldots, T-k \\
&\upsilon_{g_t}^2|\eta_{g_t} \sim IG\left(\frac{1}{2}, \frac{1}{\eta_{g_t}}\right),\quad \eta_{g_t} \sim IG\left(\frac{1}{2}, 1\right),\quad g = 1, \ldots, G,\ \ t = 1, \ldots, T \\
&\sigma^2|a \sim IG\left(\frac{1}{2}, \frac{1}{a}\right),\quad a \sim IG\left(\frac{1}{2}, 1\right)
\end{aligned}
$$

where $\boldsymbol{\Upsilon_g} = \text{diag}(\upsilon_{g_1}^2, \ldots, \upsilon_{g_T}^2)$ and $\boldsymbol{\Omega_g} = \text{diag}(\omega_{g_1}^2, \ldots, \omega_{g_{T-k}}^2)$.

Full conditional distributions are given by:

$$\mu|. \sim \mathcal{N}\left(\frac{1}{n}\mathbb{1}'(\boldsymbol{y} - \sum_{g=1}^{G} \boldsymbol{X_g}\boldsymbol{\beta_g}), \frac{\sigma^2}{n}\right)$$

$$\boldsymbol{\beta_g}|. \sim \mathcal{N}_T\left(\boldsymbol{\Sigma_{b_g}}\frac{\boldsymbol{X_g}'}{\sigma^2}(\boldsymbol{y} - \mu\mathbb{1} - \sum_{\tilde{g}\neq g} \boldsymbol{X_{\tilde{g}}}\boldsymbol{\beta_{\tilde{g}}}), \boldsymbol{\Sigma_{b_g}} = \sigma^2\left(\boldsymbol{X_g}'\boldsymbol{X_g} + \frac{\boldsymbol{D_g^{(k)}}'\boldsymbol{\Omega_g}^{-1}\boldsymbol{D_g^{(k)}}}{\tau^2\lambda_g^2} + \boldsymbol{\Upsilon_g}^{-1}\right)^{-1}\right)$$

$$\tau^2|. \sim IG\left(\frac{1+T}{2}, \frac{1}{\xi} + \sum_{g=1}^{G}\frac{\boldsymbol{\beta_g}'\boldsymbol{D^{(k)}}'\boldsymbol{\Omega_g}^{-1}\boldsymbol{D^{(k)}}\boldsymbol{\beta_g}}{2\sigma^2\lambda_g^2}\right), \quad \xi|. \sim IG(1, 1 + 1/\tau^2)$$

$$\lambda_g^2|. \sim IG\left(\frac{1}{2} + \frac{T}{2}, \frac{1}{\psi_g} + \frac{\boldsymbol{\beta_g}'\boldsymbol{D_g^{(k)}}'\boldsymbol{\Omega_g}^{-1}\boldsymbol{D_g^{(k)}}\boldsymbol{\beta_g}}{2\sigma^2\tau^2}\right), \quad \psi_g|. \sim IG(1, 1 + 1/\lambda_g^2), \quad g = 1, \ldots, G$$

$$\omega_{g_j}^2|. \sim IG\left(1, \frac{1}{\phi_{g_j}} + \frac{((\boldsymbol{D_g^{(k)}}\boldsymbol{\beta_g})_{[j]})^2}{2\sigma^2\tau^2\lambda_g^2}\right), \quad \phi_{g_j}|. \sim IG(1, 1 + 1/\omega_{g_j}^2), \quad g = 1, \ldots, G, \quad j = 1, \ldots, T-k$$

$$\upsilon_{g_t}^2|. \sim IG\left(1, \frac{1}{\eta_{g_t}} + \frac{\beta_{g_t}'\beta_{g_t}}{2\sigma^2}\right), \quad \eta_{g_t}|. \sim IG(1, 1 + 1/\upsilon_{g_t}^2), \quad g = 1, \ldots, G, \quad t = 1, \ldots, T$$

$$\sigma^2|. \sim IG\left(\frac{1+T+n}{2}, \frac{1}{a} + \frac{1}{2}\sum_{g=1}^{G}\beta_g'\left(\boldsymbol{\Upsilon_g}^{-1} + \frac{\boldsymbol{D_g^{(k)}}'\boldsymbol{\Omega_g}^{-1}\boldsymbol{D_g^{(k)}}}{\tau^2\lambda_g^2}\right)\boldsymbol{\beta_g} + \frac{1}{2}||\boldsymbol{y} - \mu\mathbb{1} - \sum_{g=1}^{G}\boldsymbol{X_g}\boldsymbol{\beta_g}||_2^2\right)$$

$$a|. \sim IG(1, 1 + 1/\sigma^2)$$

All full conditional distributions have a closed form and an efficient Gibbs sampler algorithm can be constructed.

A.3 Complementary results for multi-group simulations

Approach	**RMSE$_\mathbf{p}$**	**RMSE$_1$**	**RMSE$_v$**	**RMSE$_c$**	**RMSE$_0$**	**CIW**
group fused HS 3L	1.43	0.040	0.058	0.004	0	0.05
group fused HS 2L	1.47	0.043	0.047 0.039	0	0.27	
group fusion HS 3L						
group fusion HS 2L						
SPLSR	3.41	0.063	0.056	0.070	0.017	
ENR	3.07	0.078	0.072	0.083	0.006	
cMCP	18.19	0.653	0.649	0.657	0.014	
cMCP_Ridge	1.70	0.042	0.040	0.044	0.007	
Approach	**FPR$_\mathbf{v}$**	**FNR$_\mathbf{v}$**	**MCC$_\mathbf{v}$**	**FPR$_\mathbf{g}$**	**FNR$_\mathbf{g}$**	**MCC$_\mathbf{g}$**
group fused HS 3L	0	2	0.99	0	0	1
group fused HS 2L	0	2	0.99	0	0	1
group fusion HS 3L						
group fusion HS 2L						
SPLSR	3	3	0.92	0	0	1
ENR	2	1	0.95	26	0	0.73
cMCP	0.7	84	0.32	0	13	0.91
cMCP_Ridge	2	0.2	0.96	14	0	0.83

Table 3: RMSE$_\mathrm{p}$, RMSE$_1$, RMSE$_0$, CIW, FPR$_\mathrm{v}$, FNR$_\mathrm{v}$, MCC$_\mathrm{v}$, FPR$_\mathrm{g}$, FNR$_\mathrm{g}$ and MCC$_\mathrm{g}$ criteria averaged over repetitions and obtained using the different approaches on the simulated dataset with $T = 300$ leading to 3000 variables.

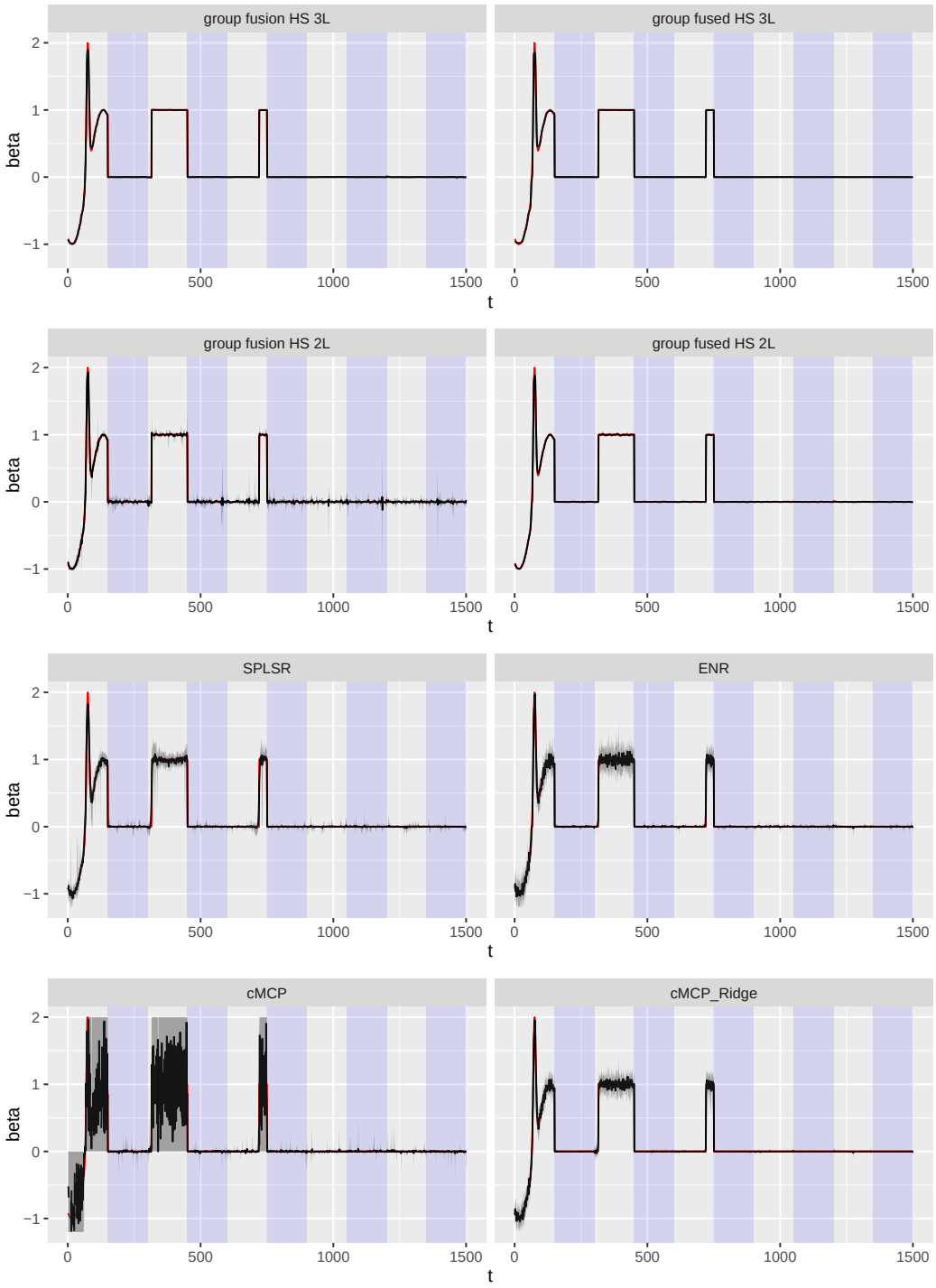


Figure 2: For the different approaches, coefficient profile estimations averaged over repetitions for the simulated dataset with $T = 150$ ($1,500$ variables). Gray shadows represent 95% confidence intervals computed over repetitions. True coefficient profiles are displayed in red. The alternation of gray and blue areas delimits the 10 groups.

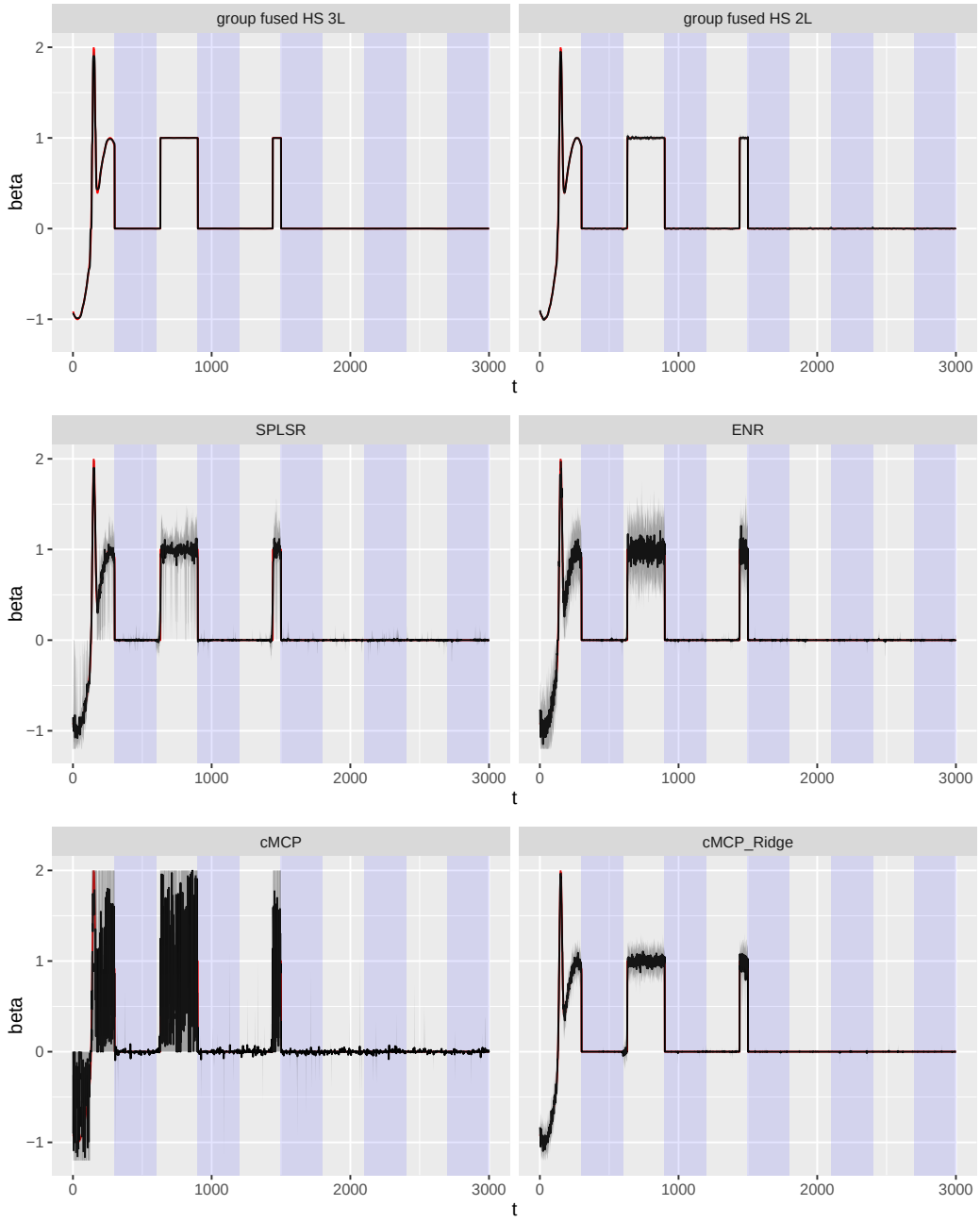


Figure 3: For the different approaches, coefficient profile estimations averaged over repetitions for the simulated dataset with $T = 300$ ($3,000$ variables). Gray shadows represent 95% confidence intervals computed over repetitions. True coefficient profiles are displayed in red. The alternation of gray and blue areas delimits the 10 groups.

4.3 Application à l'identification de positions génétiques influant sur un caractère phénotypique

Dans le contexte de l'identification de positions génétiques influant sur un caractère phénotypique, nous avons observé dans le chapitre précédent que la forte densité de marqueurs tout au long du génome implique une forte corrélation deux à deux intra chromosome (cf figure 4.1). Ainsi, les marqueurs contenus dans un chromosome peuvent être vus comme un groupe de variables explicatives ordonnées le long du génome. En se concentrant uniquement sur un temps donné, leurs effets peuvent être considérés comme un groupe de paramètres ordonnés. L'identification des marqueurs pertinents peut alors être vue comme une double sélection de groupes de paramètres et de paramètres indexés. L'approche développée dans ce chapitre peut ainsi être directement appliquée, permettant d'estimer un profil d'effet tout au long du génome, pouvant être nul sur certaines régions.

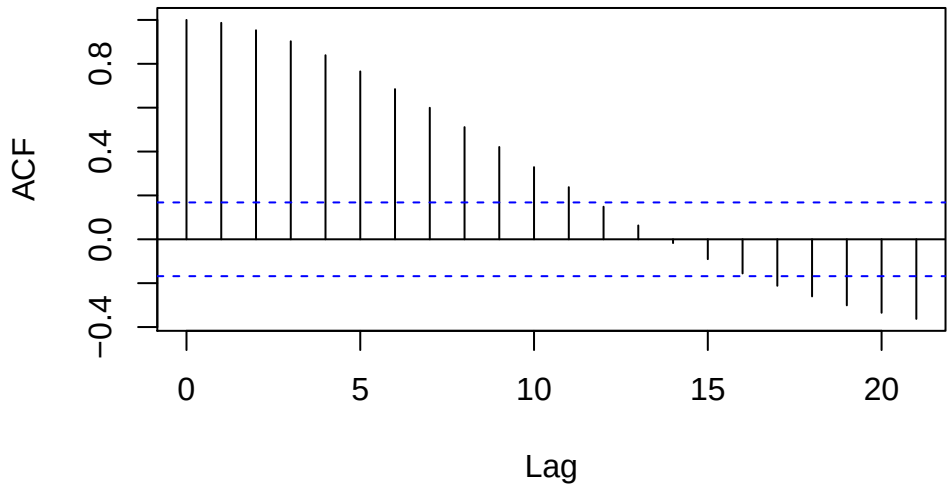


FIGURE 4.1 – *Graphique ACF des marqueurs du chromosome d'un individu sur le jeu de données complet arabidopsis thaliana analysé dans les chapitres 2 et 3.*

Nous présentons ici des résultats préliminaires de l'application des priors *group fused horseshoe* et *group fusion horseshoe* sur les données *arabidopsis thaliana* analysées dans les chapitres 2 et 3, en se focalisant sur le dernier pas de temps. Le jeu de données complet contient $p = 532$ marqueurs sur l'ensemble des 5 chromosomes. Nous nous restreignons à $n = 100$ individus choisis aléatoirement sur les 357 disponibles pour complexifier la situation (ratio nombre d'individus / nombre de variables plus faible). En comparaison, nous appliquons également les approches alternatives (fréquentistes) considérées dans l'article (SPLSR, EN, cMCP et cMCP_Ridge). 5 essais d'une validation croisée à 10 groupes ont été réalisés pour calculer le RMSE prédictif. Cela correspond donc à $5 \times 10 = 50$ répétitions. Concernant les approches bayésiennes, 50.000 itérations MCMC ont été réalisées. Une itération sur dix est gardée, après avoir sup-

Approche	RMSE_p
group fusion horseshoe	1.243
cMCP_Ridge	1.373
EN	1.380
group fused horseshoe	1.386
cMCP	1.394
SPLSR	1.412

TABLE 4.1 – *RMSE prédictif obtenue sur des données arabidopsis thaliana à l'aide des différentes approches.*

primé les 10.000 premières. Les statistiques PSRF de Gelman et al. [1992] pour chacun des paramètres sont inférieures à 1.1, indiquant la convergence des 50 chaînes MCMC. Concernant les approches fréquentistes, les paramètres de pénalisation ont été estimés par validation croisée à 5 groupes.

Les estimations moyennes ainsi que les intervalles de confiance obtenus sur l'ensemble des 50 répétitions et pour l'ensemble des méthodes sont présentés dans la figure 4.2. D'un point de vue prédictif (cf tableau 4.1), le *group fusion horseshoe* présente les meilleures performances. L'utilisation de ce prior permet, en considérant les positions associées aux coefficients de régression dont les intervalles de confiance ne contiennent pas zéro, d'identifier clairement 9 régions génomiques (1, 2, 2, 1 et 3 régions, sur les chromosomes 1 à 5 respectivement). Ces régions sont en accord avec celles déjà identifiées dans les chapitres 2 et 3. Ces régions sont également observées en utilisant les autres approches mais cela est nettement moins net. Le *group fused horseshoe* tend à forcer vers zéro l'effet des paramètres, et ainsi à potentiellement sous-estimer ces effets. Cela se traduit en particulier par des intervalles de confiance très étroits autour de zéro. En revanche, les approches fréquentistes conduisent à des résultats instables d'une répétition à l'autre (cf figure 4.3), aboutissant à de larges intervalles de confiance contenant quasiment systématiquement zéro lorsqu'on agrège les estimations sur l'ensemble des répétitions (cf figure 4.2). L'interprétation des résultats devient donc difficile.

Ces résultats confirment que l'hypothèse d'indexation des marqueurs le long du génome est pertinente et la prise en compte de cette structure de dépendance dans l'estimation des effets aide à l'identification des régions génomiques. L'approche que nous proposons semble prometteuse et présente une réelle alternative aux approches classiquement utilisées telles que les régressions Lasso ou Elastic-Net pour la détection de position comme pour la prédiction [Brault et al., 2020]. Ici, nous pouvons remarquer que la version *fusion* de notre prior donne de meilleures performances que la version *fused*. Cela peut s'expliquer par le fait que le nombre de variables est peu important. Cela va dans le sens de ce que nous avions déjà observé dans l'article précédent. Ainsi, des recherches et analyses doivent être approfondies pour mieux comprendre les performances des priors *group fusion/fused horseshoe* selon la dimension des données. Cependant, l'algorithme MCMC proposé présente un temps de calcul bien supérieur aux approches fréquentistes mises en œuvre (80 min pour 50.000 itérations vs moins d'une minute). Plus de recherche doit également être menée pour proposer une méthode d'inférence optimisée.

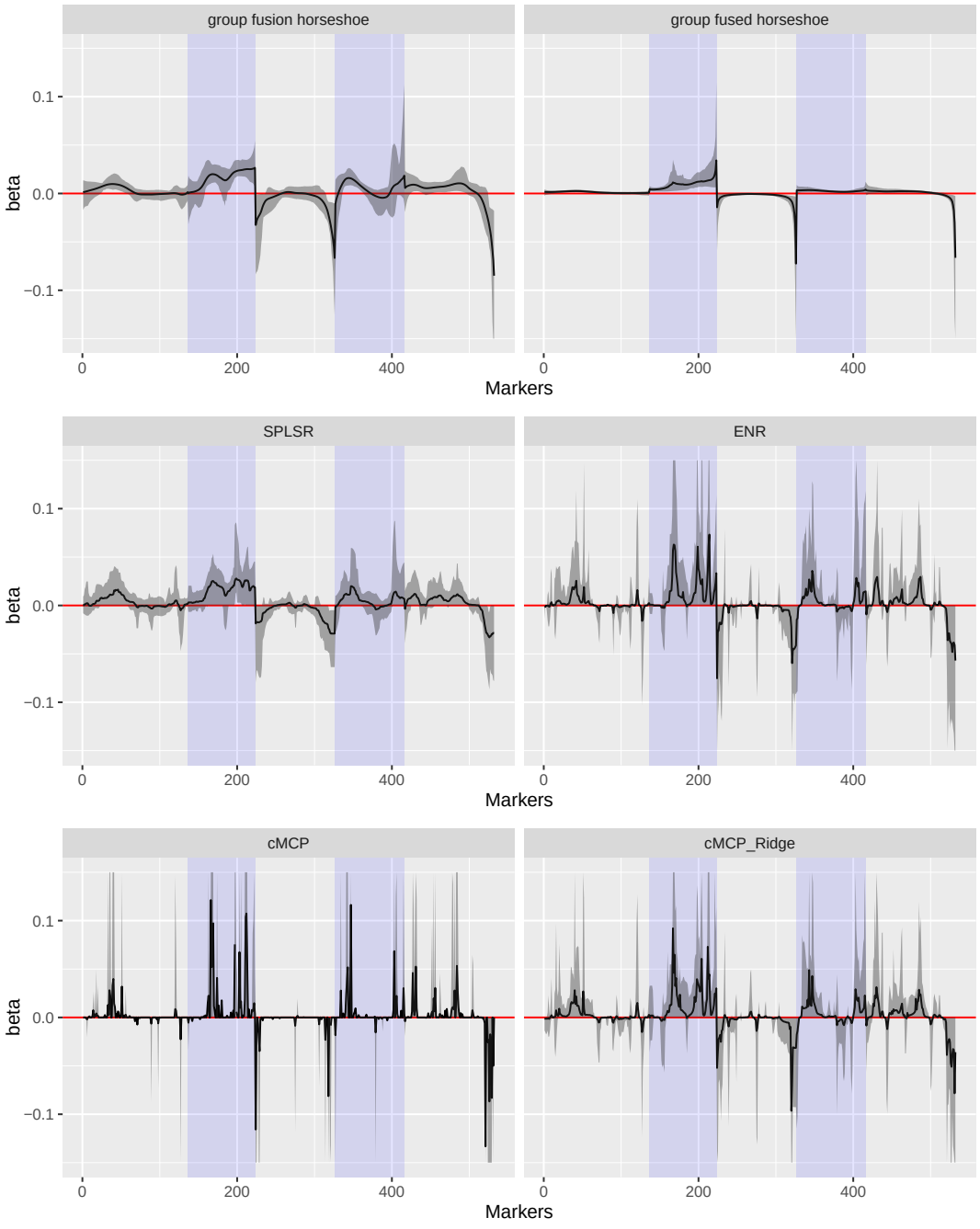


FIGURE 4.2 – *Profil d'effet des marqueurs le long du génome obtenu à partir des différentes approches sur 50 répétitions. L'ombre grise donne les intervalles de crédibilité à 95%. L'alternance de fonds gris et bleu délimite les différents chromosomes.*

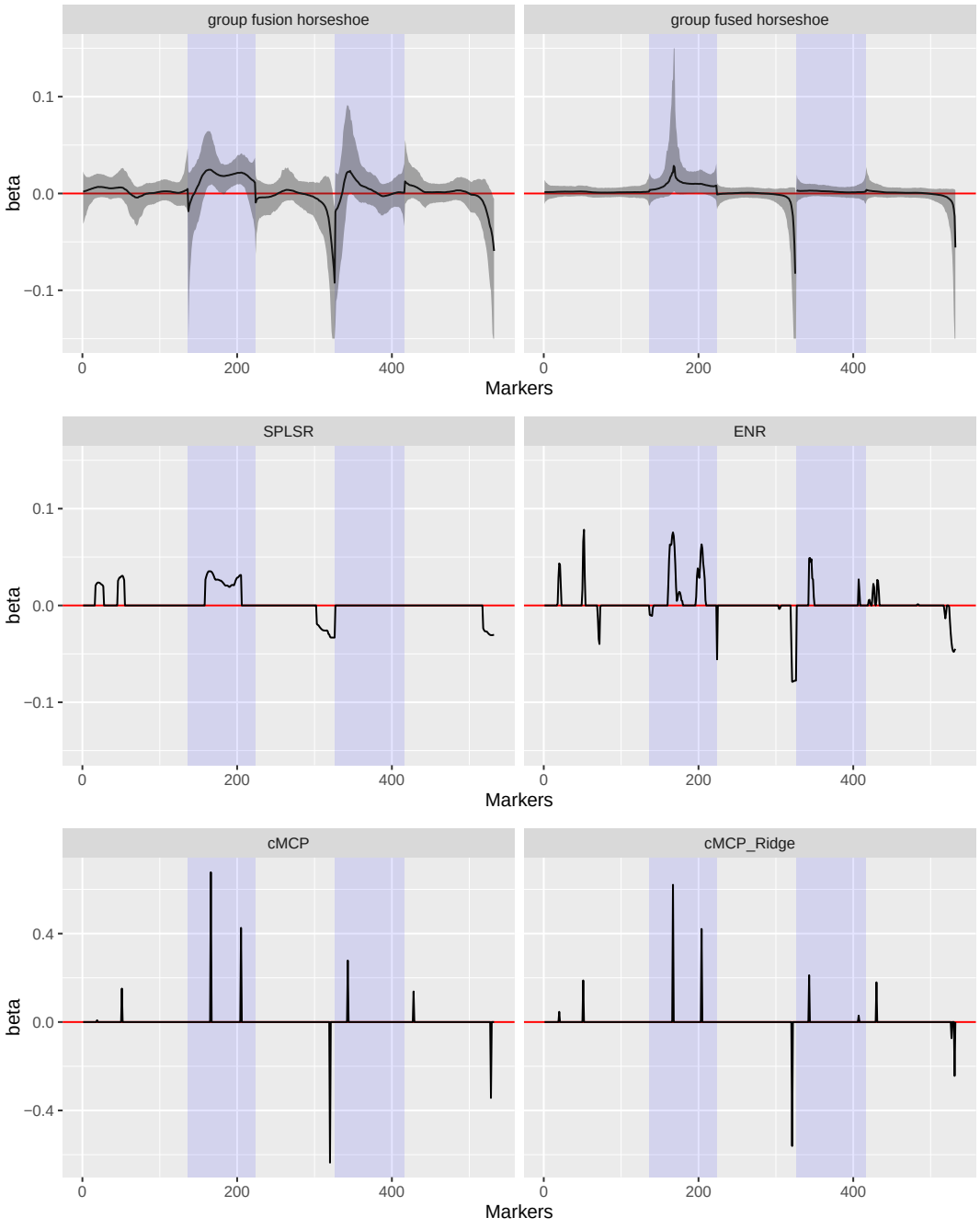


FIGURE 4.3 – *Profil d'effet des marqueurs le long du génome obtenu à partir des différentes approches sur une répétition. L'ombre grise donne les intervalles de crédibilité à 95%. L'alternance de fonds gris et bleu délimite les différents chromosomes.*

V

Conclusions et perspectives

Sommaire

5.1 Contributions par rapport aux objectifs initiaux 114

5.2 Perspectives 116

Ce travail de recherche a visé à proposer des approches innovantes pour répondre à des problématiques statistiques soulevées par des questionnements agronomiques et génétiques concrets.

Le génotypage haut débit a été largement mis à profit ces dernières années pour identifier les positions le long du génome impliquées dans la variabilité de caractères d'intérêt. Il permet d'avoir accès à l'information génétique au travers l'acquisition d'un grand nombre de marqueurs. Cependant, de récentes études ont montré les limites des méthodes d'analyse menées pour ce type de données (nombre de marqueurs bien supérieur au nombre d'individus, marqueurs peu informatifs, forte corrélation) [Bonhomme et al., 2019]. Ainsi, de nouvelles approches visant à résumer l'information génétique, parfois combinée à l'information pedigree, ont été explorées notamment dans le contexte de plans d'expérience multi-parentaux. Ces résumés se traduisent par un ensemble de matrices d'apparentement (IBDs) associées à différentes positions sur l'ensemble du génome. L'analyse globale intégrant l'ensemble des matrices d'apparentement fait intervenir un modèle linéaire mixte (LMM) intégrant un grand nombre d'effets aléatoires, et dont l'objectif est d'identifier les positions impliquées dans la variabilité du caractère étudié. Cette analyse soulève alors la question de l'identification des effets aléatoires pertinents, et donc la question de la sélection de composantes de la variance.

Toujours dans le contexte de la génétique quantitative, nous nous sommes intéressés à l'étude de l'évolution de l'architecture génétique d'un caractère phénotypique. Ce type d'étude est désormais rendu possible grâce aux technologies de phénotypage haut débit qui permettent d'acquérir des mesures répétées dans le temps des caractères phénotypiques, et ainsi de suivre leur évolution. L'analyse de telles données soulève

plusieurs questions, notamment celles de la prise en compte des dépendances temporelles, de l'identification des positions génétiques pertinentes, ou encore de l'estimation de leurs effets évoluant au cours du temps. Une telle analyse peut être réalisée, entre autres, grâce à un modèle à intercept et pentes aléatoires (RIS), ou encore au travers de modèles à coefficients variants (VCM). La première approche soulève la question de la sélection d'effets fixes et aléatoires dans les modèles linéaires mixtes (LMM), tandis que la deuxième soulève la question de la sélection de groupes de coefficients structurés (ordonnés) dans un VCM.

Enfin, nous nous sommes penchés sur l'étude de l'influence, au cours du temps, de facteurs environnementaux sur des processus biologiques tels que l'abscission des organes pour le palmier à huile. En particulier, la question de l'identification des périodes de temps dans le processus d'abscission sur lesquels les facteurs influent, a retenu notre attention. Pour cela, le suivi des facteurs tout au long du processus est nécessaire. Ainsi, un suivi au cours du temps de chaque facteur a été réalisé, conduisant à des groupes de variables ordonnées. Cependant, l'analyse de telles données soulève la question de la double sélection de groupes de variables (facteurs) puis de variables à l'intérieur des groupes (pas de temps), tout en tenant compte de la forte corrélation temporelle présente entre les variables.

5.1 Contributions par rapport aux objectifs initiaux

Dans un cadre bayésien, l'objectif de ce travail de recherche a été de proposer des priors innovants pour répondre à ces différentes problématiques. Ces priors ont pour but de simultanément sélectionner les variables pertinentes, estimer des effets pouvant évoluer au cours du temps et prendre en compte diverses structurations de dépendance présentes dans les données. L'utilisation de ces priors permet ainsi de se prémunir d'un sur-ajustement et mène à des estimations parcimonieuses pouvant être facilement interprétables d'un point de vue biologique.

Dans un premier temps (cf chapitre 2), nous nous sommes focalisés sur l'estimation parcimonieuse de modèles linéaires mixtes au travers de la sélection d'effets fixes et aléatoires. Dans l'objectif de réduire à zéro les écart-types associés aux effets aléatoires non-pertinents, nous avons proposé une version pliée du prior *horseshoe*. Nous comparons ce prior avec des versions pliées des priors *Cauchy* et *spike-and-slab*. Un prior *horseshoe* est systématiquement placé sur les effets fixes. La comparaison est réalisée à l'aide de deux applications réelles. La première porte sur la cartographie de QTL à l'aide de matrice d'apparentements et a pour objectif d'étudier le rendement du palmier à huile grâce à un modèle animal. La deuxième est l'étude de la dynamique évolutive au cours du temps de l'architecture génétique de la compacité des feuilles pour *arabidopsis thaliana*, impliquant un modèle à intercept et pentes aléatoires. Nous montrons que le prior *horseshoe* plié présente de bonnes performances, est robuste face au grand nombre d'effets aléatoires et permet des estimations sans biais contrairement au prior *Cauchy* plié. Il offre des performances similaires au prior *spike-and-slab* plié en termes de sélection, mais a une efficacité de calcul supérieure. Le prior *horseshoe* est donc à privilégier à la fois pour la sélection d'effets fixes et aléatoires.

Nous avons également proposé de mettre en œuvre une reparamétrisation polaire pour modéliser la matrice de corrélation des effets aléatoires pour le modèle RIS. Cette approche n'a reçu que peu d'attention jusqu'à ce jour et n'avait jamais été mise en œuvre dans ce contexte. Nous montrons que supposer une dépendance (inconnue) entre les pentes peut avoir un impact sur la sélection des composantes de la variance. De plus, l'estimation de cette matrice permet d'identifier les marqueurs ayant des profils d'effet similaires ou inversés au cours du temps.
D'un point de vue biologique, dans le contexte de l'application à la cartographie de QTL pour le *palmier à huile*, le prior *horseshoe* plié permet d'identifier les positions qui étaient en ségrégation dans une fraction mineure de la population, en raison d'un plan génétique déséquilibré, tandis que l'approche de sélection fréquentiste de type *stepwise* utilisée dans l'étude originelle [Tisné et al., 2015] ne le permettait pas. Dans le contexte de l'étude de la dynamique de l'évolution de l'architecture génétique pour *arabidobsis thaliana*, nous retrouvons des résultats proches de ceux trouvés au travers de l'utilisation d'un modèle à coefficients variants (cf chapitre 3). Les deux applications soulignent que les approches multivariées augmentent la puissance statistique comparativement à une approche uni-temporelle comme le propose Marchadier et al. [2019].

Dans un deuxième temps (cf chapitre 3), nous nous sommes focalisés sur l'estimation parcimonieuse de modèles à coefficients variants pour l'étude de la dynamique d'évolution de l'architecture génétique d'un caractère d'intérêt. Nous proposons de combiner le prior de sélection *group spike-and-slab* pour la sélection des marqueurs, avec soit une interpolation P-spline, soit une estimation directe des effets évoluant au cours du temps. Ces deux modélisations non-paramétriques font intervenir une pénalisation sur les différences, qui, dans le contexte bayésien, est réalisée par l'emploi d'un prior multivarié gaussien de type marche aléatoire comme loi diffuse du prior *group spike-and-slab*. Ainsi, après une reparamétrisation du modèle dans le cas de l'interpolation P-spline, ces deux approches aboutissent au même prior placé sur les paramètres liés à la modélisation non-paramétrique. Les bonnes performances de ces deux approches, en comparaison à des méthodes alternatives, sont démontrées sur simulations.
Les priors développés ont alors été mis en œuvre sur le jeu de données *arabidopsis thaliana* (suivi sur 21 pas de temps) dans l'objectif d'étudier l'évolution de l'architecture génétique de la compacité des feuilles. D'un point de vue pratique, nous montrons que l'approche longitudinale développée permet une meilleure détection des marqueurs pertinents, comparativement aux approches temps par temps menées jusqu'alors. Le prior *group spike-and-slab* combiné à une interpolation P-spline a permis de mettre en évidence sept nouvelles régions génomiques. Les profils d'effets au cours du temps des différents marqueurs sélectionnés, présentent une variabilité plus importante sur les premiers stades de développement, et sont plus lisses sur les derniers. L'approche a également été appliquée à un deuxième jeu de données issu d'une expérimentation où les plantes ont subi un stress hydrique. Cependant, peu de différences dans l'estimation et la sélection ont été observées, ne permettant pas d'identifier des mécanismes de réaction de type génotype $\times$ environnement. Des recherches peuvent être menées à ce sujet pour permettre aux méthodes proposées de prendre en compte simultanément différentes conditions environnementales. Nous reviendrons sur ce point dans la section "Perspectives".

Dans un troisième temps (cf chapitre 4), nous avons approfondi la question de la

double sélection de groupes de variables puis de variables à l'intérieur des groupes, tout en tenant compte de la forte corrélation entre les variables explicatives. Pour répondre à cet objectif, nous proposons le prior *group fused horseshoe*. Ce prior peut être vu comme la généralisation du prior *sparse group horseshoe* et du prior *fused Lasso*, permettant de réduire à zéro les groupes de coefficients, les coefficients ainsi que les différences des coefficients simultanément. Nous montrons sur des simulations que ce prior présente d'excellentes performances, tant pour l'identification des groupes pertinents que pour l'identification des variables. Il permet une flexibilité dans l'estimation des coefficients, et ainsi l'estimation de profils complexes ou au contraire très lisses. De plus, nous montrons qu'il est peu sensible au ratio nombre d'observations / nombre de variables, démontrant ainsi sa robustesse. Nous montrons également que notre approche surpasse clairement les approches classiquement utilisées pour l'analyse de données fortement corrélées telles que les approches PLSR et SPLSR. Ce résultat souligne l'importance de prendre en compte l'information de structuration des variables, lorsque celle-ci est connue, dans l'estimation de leurs effets.
L'application de notre prior au jeu de données portant sur l'abscission des fruits du *palmier à huile*, a permis d'identifier quatre facteurs environnementaux. Elle a aussi permis d'identifier clairement les différentes périodes de temps sur lesquelles ces facteurs influent. Ces périodes correspondent essentiellement aux derniers jours du processus d'abscission. Ces résultats sont validés par une bonne qualité de prédiction, meilleure que celles des approches alternatives.

5.2 Perspectives

Ces différents travaux ouvrent un certain nombre de perspectives que nous avons évoquées à différentes occasions dans le document. Nous reprenons et élargissons certaines de ces pistes ci-dessous.

Tout d'abord, dans les approches développées, nous avons considéré une variable réponse quantitative continue. Il est donc naturel d'envisager d'étendre ces approches au cas d'une variable binaire ou de comptage (Poisson) en se positionnant dans le contexte des modèles linéaires généralisés. Les enjeux applicatifs sont en effet nombreux et cela permettrait d'ouvrir plus largement l'étude de divers caractères phénotypiques tels que la présence d'infection ou le nombre d'arbres infectés dans chaque parcelle élémentaire. La prise en compte d'une variable réponse binaire peut être réalisée à l'aide d'un modèle probit. Nous avons déjà exploré cette approche et l'avons intégrée dans l'algorithme du modèle hiérarchique bayésien associé au prior *group fused horseshoe* (cf chapitre 4, approche non discutée dans l'article). Les algorithmes développés dans ce travail de recherche peuvent être facilement généralisés au cas du modèle probit, car sa fonction de lien est particulière et peut s'exprimer à l'aide de la fonction de répartition de la loi gaussienne. Plus de recherches sont nécessaires pour d'autres modèles binaires, ou encore pour des variables réponses multinomiales ou de comptage. Des pistes sont apportées par Gelman et al. [2013].

Nous avons proposé un prior *horseshoe* plié pour la sélection d'effet aléatoire dans le cadre d'un modèle animal et à intercept et pentes aléatoires (cf chapitre 2). Il serait

intéressant de développer un package pour pouvoir faire de la sélection d'effet fixe et aléatoire dans n'importe quel modèle linéaire mixte au travers de ce prior. Cela permettrait de valoriser ce travail en le rendant accessible à l'ensemble de la communauté scientifique.

En restant dans le cadre de la sélection d'effet aléatoire, et dans le contexte de la cartographie de QTL à l'aide de matrice d'apparentements, tout comme deux marqueurs adjacents peuvent partager quasiment la même information, deux matrices d'apparentement adjacentes peuvent présenter une forte similarité et donc mener à des problèmes de sur-ajustement. Il serait intéressant d'explorer la possibilité de fusionner les écarts-types des effets aléatoires associés à des matrices d'apparentement adjacentes dans l'idée du travail du chapitre 4. Un prior *fusion horseshoe* plié pourrait être réalisé au travers d'un prior multivarié gaussien plié en considérant une matrice de précision reflétant l'indexation des positions le long du génome :

$$\lambda|\tau^2, \omega^2, \sigma^2 \sim \mathcal{N}_q^+(0, \sigma^2\tau^2(D^{(k)}\Omega^{-1}D^{(k)})^{-1}),$$

où $D^{(k)}$ est la matrice de l'opérateur des différences finies d'ordre k et Ω une matrice diagonale contenant les paramètres locaux de variance.

Dans un cadre bayésien, la question de la réduction à zéro de paramètres de variance se retrouve également dans l'identification de groupes de paramètres pertinents dans un modèle linéaire. Nous retrouvons cette méthodologie notamment dans les modèles additifs [Scheipl et al., 2012], ou encore les modèles TVP-VAR [Koop and Korobilis, 2013, Bitto and Frühwirth-Schnatter, 2019, Cadonna et al., 2020]. Notre prior pourrait ainsi être mis en œuvre dans de nombreuses applications. En particulier, nous pourrions imaginer l'appliquer pour le modèle à coefficients variants proposé dans le chapitre 3. Le prior *group spike-and-slab* combiné au prior multivarié gaussien de type marche aléatoire serait alors remplacé par une combinaison d'un prior multivarié gaussien de type marche aléatoire, avec un prior *horseshoe* plié placé sur la racine carré du paramètre de variance associé aux différences :

$$\begin{aligned} b_j|\tau_{b_j}, \sigma^2 &\sim \mathcal{N}(0, \sigma^2(\tau_{b_j}D'D)^{-1}), \\ \sqrt{\frac{1}{\tau_{b_j}}}|\theta^2\omega_j^2 &\sim \mathcal{N}^+(0, \theta^2\omega_j^2), \\ \theta, \omega_j, &\sim \mathcal{C}^+(0, 1). \end{aligned}$$

Ce prior devrait permettre de meilleures performances en termes de temps de calcul comparativement au prior *group spike-and-slab*.

Par ailleurs, dans l'étude de l'évolution au cours du temps de l'architecture génétique de la compacité des feuilles chez *arabidopsis thaliana*, les priors que nous proposons dans les chapitres 2 et 3 permettent d'identifier si l'effet d'un marqueur évolue dans le temps. Cependant, elles ne permettent pas de déterminer si cette évolution est due à l'évolution physiologique de l'individu ou si c'est une réaction face à un changement des conditions environnementales, faisant alors partie d'un mécanisme de survie et d'adaptation de l'espèce. Pour répondre à cette question, deux pistes sont envisageables selon l'expérimentation.
La première, dans le cas où l'on souhaite comparer plusieurs blocs ayant subi des

conditions environnementales différentes (cf application *arabidopsis thaliana* chapitre 3), on pourrait considérer le facteur bloc comme facteur à effet fixe dans le modèle. Que ce soit dans une analyse par un modèle RIS ou par un VCM, cela reviendrait alors à considérer une triple interaction entre les marqueurs (fixes), le temps (fixe pour VCM et aléatoire pour RIS) et le bloc (fixe).
Dans la deuxième, pour une expérimentation au champ, tous les individus subissent le même environnement qui évolue au cours du temps. Dans le cas du VCM, on pourrait alors imaginer décomposer le vecteur d'effets (au cours du temps) de chaque marqueur comme la somme d'une fonction de chaque variable environnementale et d'une fonction du temps :

$$\beta_{jt} = f_1(env_{1_t}) + ... + f_q(env_{q_t}) + f_{temps}(t).$$

Cette décomposition, sous la forme d'un modèle additif, permettrait de déterminer si l'effet d'un marqueur (β_{jt}) est dû à une réaction face à une variation d'une variable environnementale ($f_1(env_{1_t}), \ldots, f_q(env_{q_t})$), ou simplement dû à l'évolution physiologique de l'individu ($f_{temps}(t)$). Cette approche implique un grand nombre de paramètres dont peu sont pertinents. Un travail de recherche est nécessaire pour pouvoir inférer un tel modèle tout en identifiant simultanément les marqueurs pertinents ainsi que les facteurs environnementaux qui expliquent l'évolution de leurs effets au cours du temps.

Toujours dans l'étude de l'évolution au cours du temps de l'architecture génétique (cf chapitres 2 et 3), nous avons analysé le jeu de données au travers de deux modélisations différentes. La première modélisation fait appel à un modèle à intercept et pentes aléatoires, où les effets à chaque pas de temps des marqueurs sont modélisés par des réalisations d'effets aléatoires. À un temps donné, les effets aléatoires sont supposés structurés entre les marqueurs, mais ils sont indépendants au cours du temps. La deuxième fait appel à un modèle à coefficients variants où l'évolution des effets de chaque marqueur est modélisée par un prior multivarié gaussien de type marche aléatoire. Ce prior permet d'introduire une structuration des effets au cours du temps, et suppose que les effets entre les différents marqueurs sont indépendants. D'un point de vue bayésien, ces deux modélisations sont donc proches, si ce n'est qu'elles décrivent de manière différente et complémentaire, l'évolution au cours du temps des effets des marqueurs. Ces deux modélisations pourraient être combinées pour offrir une structuration optimale des effets à la fois dans le temps et entre les marqueurs. Cela pourrait être réalisé en considérant un prior multivarié gaussien centré en zéro ayant pour matrice de précision un produit de Kronecker entre une matrice de différences finies pour décrire les effets dans le temps, et une matrice inconnue pour décrire les effets entre les marqueurs. Ce prior pourrait alors être considéré pour les effets aléatoires d'un modèle RIS, ou directement sur des coefficients de régression dans le cas du modèle à coefficients variants. Notons qu'une modélisation de type marche aléatoire pourrait également être considérée pour structurer les pentes entre les marqueurs lorsque ceux-ci présentent une forte corrélation deux à deux (cf chapitre 4).

Le prior *group spike-and-slab* combiné à une distribution multivariée gaussienne de type marche aléatoire que nous proposons dans le chapitre 3, et le prior *group fused horseshoe* que nous proposons dans le chapitre 4, permettent tous les deux de sélectionner des groupes de coefficients ordonnés et d'estimer un profil de coefficients évoluant au cours du temps. Le premier considère une distribution gaussienne de type marche

aléatoire sur chaque groupe de coefficients, avec un seul paramètre de variance spécifique à chaque groupe. Ce prior ne permet pas la sélection de coefficients à l'intérieur des groupes. Il est alors plus adapté pour estimer des coefficients ayant un profil non nul sur l'ensemble des pas de temps, avec un niveau de lissage homogène. Le second prior considère lui aussi une distribution multivariée gaussienne sur chaque groupe de paramètres avec, quant à lui, une matrice de précision plus évoluée permettant de réduire à zéro simultanément les coefficients et leurs différences. De plus, trois niveaux de paramètres de variance (global, spécifique aux groupes et local) sont considérés pour une régularisation optimale du modèle. Ce prior est alors plus adapté pour estimer des coefficients potentiellement nuls sur certaines périodes, et ayant un profil pouvant être à la fois très lisse ou au contraire complexe. Il permet d'obtenir des profils non-homogènes au cours du temps, présentant potentiellement des sauts ou encore un profil constant par morceaux. Les simulations ainsi que l'application sur les données réelles ont montré que ce prior est performant et qu'il permet de s'adapter à diverses situations. Ainsi, même si sa mise en œuvre est plus compliquée, nous recommandons l'utilisation de ce prior dans les modèles à coefficients variants (et les modèles qui en découlent : modèle additif, modèle TVP-VAR, ...). Enfin, ce prior peut être directement utilisé dans un large éventail d'applications. Par exemple, il peut être appliqué en chimiométrie pour la calibration de spectroscopie proche infrarouge, impliquant un groupe de variables ordonnées à travers un spectre. Ou encore, comme nous l'avons montré dans le chapitre 4, ce prior peut être directement utilisé pour des applications de cartographie génétique (architecture génétique) où les marqueurs d'un même chromosome peuvent être considérés comme un groupe de régresseurs ordonnés le long du génome, présentant une forte corrélation deux à deux.

Le prior *group fused horseshoe*, développé dans le chapitre 4, permet de prendre en compte une indexation unidimensionnelle (au cours du temps pour le jeu de données *abscission* ou le long du génome dans le cas du jeu de données *arabidopsis thaliana*). Une perspective intéressante serait d'appliquer ces approches à des données génétiques dont les dimensions sont plus proches de la réalité : quelques centaines d'individus avec quelques milliers de marqueurs voir plus.
Ce travail peut aussi être étendu à une indexation multidimensionnelle. Par exemple, dans le cas de l'étude de l'évolution au cours du temps de l'architecture génétique d'un caractère, la prise en compte simultanée de l'indexation des marqueurs le long du génome et de l'indexation des effets au cours du temps, pourrait être réalisée par une telle extension. Cette extension serait définie par un produit de Kronecker entre des matrices de structuration associées à chaque dimension :

$$Q_g = \left(\Upsilon_g^{-1} + \frac{1}{\tau^2 \xi_g^2} (D_{1_g}^{(k)\prime} \Omega_{1_g}^{-1} D_{1_g}^{(k)}) \otimes (D_{2_g}^{(k)\prime} \Omega_{2_g}^{-1} D_{2_g}^{(k)}) \right).$$

où $D_{1_g}^{(k)}$ et $D_{2_g}^{(k)}$ sont les matrices des différences finies associées au groupe g (le groupe de marqueurs sur le $g^{\text{ème}}$ chromosome) et respectivement à la position sur le génome (1) ou au temps (2). Ω_{1_g} et Ω_{2_g} sont des matrices diagonales de paramètres locaux de variance du groupe g et des dimensions 1 et 2 respectivement.

Enfin, au cours de ce travail, nous avons montré encore une fois la flexibilité des modèles hiérarchiques bayésiens, qui permettent une modélisation fine des données. Cependant, l'inférence de tels modèles au travers d'algorithmes d'échantillonnage MCMC

(échantillonneurs de *Gibbs*, *Metropolis-Hastings* ou encore *Metropolis within Gibbs*) est complexe à réaliser. En effet, pour chaque modèle hiérarchique, la mise en œuvre de tels algorithmes nécessite un certain nombre de calculs préalables, ainsi qu'un temps conséquent de programmation. Certains logiciels/librairies R ont fait leur apparition tel que BUGS [Lunn et al., 2009], Nimble [de Valpine et al.], STAN [Carpenter et al., 2017] ou encore INLA (*integrated nested Laplace approximation*, Rue et al. [2009]). Ils permettent à l'utilisateur d'inférer facilement un modèle hiérarchique bayésien sans avoir à construire l'algorithme d'inférence. Cependant, ils restent limités dans le panel des modèles possibles, et ne permettent pas d'inférer les modèles que nous avons proposés. De plus, les temps de calcul de ces algorithmes MCMC explosent avec le nombre d'observations ou de variables, et cela, même au travers d'une implémentation optimisée dans le langage C++. Pour certains modèles, la librairie INLA propose une alternative d'estimation rapide par approximation de Laplace. Ces dernières années, certains auteurs ont exploré la parallélisation des algorithmes MCMC sur carte graphique (GPU). Terenin et al. [2019] proposent ainsi un modèle probit avec un prior *horseshoe* sur les effets fixes, inféré grâce à un algorithme de Gibbs parallélisé sur GPU. Leur approche permet de gagner en temps de calcul considérablement, passant d'un ordre de temps en O(jours) à un ordre en O(minutes). Cependant, ce genre d'approches nécessite un niveau de connaissance élevé en programmation sur GPU et est donc peu accessible. L'approche EM [Dempster et al., 1977] présente également une alternative intéressante aux algorithmes MCMC pour l'inférence de modèles hiérarchiques. Ročková and George [2014], Rockova and McAlinn [2021] démontrent que cette classe d'algorithme permet d'inférer des modèles complexes tout en présentant des temps de calcul raisonnables. Cependant, leur mise en œuvre reste délicate. Finalement, une solution serait de contribuer au développement de la librairie INLA pour étendre la gamme de modèles et de priors.

Bibliographie

A. Armagan, M. Clyde, and D. Dunson. Generalized beta mixtures of gaussians. *Advances in neural information processing systems*, 24:523–531, 2011. 12, 13

A. Armagan, D. B. Dunson, and J. Lee. Generalized double pareto shrinkage. *Statistica Sinica*, 23(1):119, 2013. 12

R. Bai and M. Ghosh. Large-scale multiple hypothesis testing with the normal-beta prime prior. *Statistics*, 53(6):1210–1233, 2019. 12

J. Bartholomé, A. Mabiala, R. Burlett, D. Bert, J.-C. Leplé, C. Plomion, and J.-M. Gion. The pulse of the tree is under genetic control: eucalyptus as a case study. *The Plant Journal*, 103(1):338–356, 2020. 3

A. Bhadra, J. Datta, N. G. Polson, B. Willard, et al. The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis*, 12(4):1105–1131, 2017. 12, 13

A. Bhattacharya, D. Pati, N. S. Pillai, and D. B. Dunson. Dirichlet–laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490, 2015. 12, 13

P. Bickel, B. Li, A. Tsybakov, S. van de Geer, B. Yu, T. Valdés, C. Rivero, J. Fan, and A. van der Vaart. Regularization in statistics. *Test*, 15:271–344, 2006. 9

A. Bitto and S. Frühwirth-Schnatter. Achieving shrinkage in a time-varying parameter model framework. *Journal of Econometrics*, 210(1):75–97, 2019. 19, 117

H. Bondell, A. Krishna, and S. Ghosh. Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66(4):1069–1077, 2010. 18

M. Bonhomme, M. I. Fariello, H. Navier, A. Hajri, Y. Badis, H. Miteul, D. A. Samac, B. Dumas, A. Baranger, C. Jacquet, et al. A local score approach improves gwas resolution and detects minor qtl: application to medicago truncatula quantitative disease resistance to multiple aphanomyces euteiches isolates. *Heredity*, 123(4):517–531, 2019. 113

C. Brault, A. Doligez, L. Le Cunff, A. Coupel-Ledru, T. Simonneau, P. This, and T. Flutre. Harnessing multivariate, penalized regression methods for genomic prediction and qtl detection to cope with climate change affecting grapevine. *bioRxiv*, 2020. 87, 110

A. Cadonna, S. Frühwirth-Schnatter, and P. Knaus. Triple the gamma—a unifying shrinkage prior for variance and variable selection in sparse state space and tvp models. *Econometrics*, 8(2):20, 2020. 19, 117

B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: a probabilistic programming language. *Grantee Submission*, 76(1):1–32, 2017. 120

C. M. Carvalho, N. G. Polson, and J. G. Scott. Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, pages 73–80, 2009. 12, 13, 86

C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010. 12, 13

Z. Chen and D. B. Dunson. Random effects selection in linear mixed models. *Biometrics*, 59(4):762–769, 2003. 18, 22

C. De Boor, C. De Boor, E.-U. Mathématicien, C. De Boor, and C. De Boor. *A practical guide to splines*, volume 27. springer-verlag New York, 1978. 8

P. de Valpine, D. Turek, C. Paciorek, C. Anderson-Bergman, D. Temple Lang, and R. Bodik. Programming with models: writing statistical algorithms for general model structures with NIMBLE. 120

M. Delattre and M.-A. Poursat. An iterative algorithm for joint covariate and random effect selection in mixed effects models. *The International Journal of Biostatistics*, 1 (ahead-of-print), 2020. 18

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. 120

P. Dierckx. *Curve and surface fitting with splines*. Oxford University Press, 1995. 8

P. H. Eilers and B. D. Marx. Flexible smoothing with b-splines and penalties. *Statistical science*, pages 89–102, 1996. 8

R. L. Eubank. *Spline smoothing and nonparametric regression*, volume 90. M. Dekker New York, 1988. 8

L. Fahrmeir, T. Kneib, et al. Bayesian smoothing and regression for longitudinal, spatial and event history data. *OUP Catalogue*, 2011. 6

J. Fan and R. Li. Variable selection in linear mixed effects models. *Annals of Statistics*, 40:2043–2068, 2012. 18

J. R. Faulkner and V. N. Minin. Locally adaptive smoothing with markov random fields and shrinkage priors. *Bayesian analysis*, 13(1):225, 2018. 6, 17

G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs. *Longitudinal data analysis*. CRC press, 2008. 8

J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010. 15

S. Frühwirth-Schnatter and R. Tüchler. Bayesian parsimonious covariance estimation for hierarchical linear mixed models. *Statistics and Computing*, 18(1):1–13, 2008. 19

S. Frühwirth-Schnatter and H. Wagner. Stochastic model specification search for gaussian and partial non-gaussian state space models. *Journal of Econometrics*, 154(1): 85–100, 2010. 6

S. Frühwirth-Schnatter and H. Wagner. Bayesian variable selection for random intercept modeling of gaussian and non-gaussian data. 10 2011. doi: 10.1093/acprof:oso/9780199694587.003.0006. 19

A. Gelman, D. B. Rubin, et al. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992. 110

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. CRC press, 2013. 116

A. Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1(3):515–534, 2006. 15, 19, 22

A. W. George, P. M. Visscher, and C. S. Haley. Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach. *Genetics*, 156(4):2081–2092, 2000. 3

E. George and R. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993. 10

E. George and R. McCulloch. Approaches for Bayesian variable selection. *Statistica sinica*, pages 339–373, 1997. 10

J. Geweke. Variable selection and model comparison in regression. *In Bayesian Statistics 5*, 1996. 10

J. Griffin, P. Brown, et al. Hierarchical shrinkage priors for regression models. *Bayesian Analysis*, 12(1):135–159, 2017. 12

J. E. Griffin and P. J. Brown. Bayesian adaptive lassos with non-convex penalization. 2007. 12

J. E. Griffin, P. J. Brown, et al. Inference with normal-gamma prior distributions in regression problems. *Bayesian analysis*, 5(1):171–188, 2010. 12, 19

P. R. Hahn, J. He, and H. Lopes. Elliptical slice sampling for bayesian shrinkage regression with applications to causal inference. *U RL http://faculty. chicagobooth. edu/-richard. hahn/research. html*, 2016. 12

T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779, 1993. 6

T. J. Hastie and R. J. Tibshirani. *Generalized additive models*, volume 43. CRC press, 1990. 7

R. J. Hodrick and E. C. Prescott. Postwar us business cycles: an empirical investigation. *Journal of Money, credit, and Banking*, pages 1–16, 1997. 6

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. 11

J. Ibrahim, H. Zhu, R. Garcia, and R. Guo. Fixed and random effects selection in mixed effects models. *Biometrics*, 67:495–503, 2011. 18

H. Ishwaran, J. S. Rao, et al. Spike and slab variable selection: frequentist and bayesian strategies. *Annals of statistics*, 33(2):730–773, 2005. 11, 15

S.-J. Kim, K. Koh, S. Boyd, and D. Gorinevsky. \ell_1 trend filtering. *SIAM review*, 51 (2):339–360, 2009. 6

S. K. Kinney and D. B. Dunson. Fixed and random effects selection in linear and logistic models. *Biometrics*, 63(3):690–698, 2007. 18, 22

G. Koop and D. Korobilis. Large time-varying parameter vars. *Journal of Econometrics*, 177(2):185–198, 2013. 117

G. Korontzis, M. Malosetti, C. Zheng, C. Maliepaard, H. A. Mulder, P. Lindhout, R. F. Veerkamp, and F. A. van Eeuwijk. Qtl detection in a pedigreed breeding population of diploid potato. *Euphytica*, 216(9):1–14, 2020. 3

L. Kuo and B. Mallick. Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81, 1998. 10

M. Kyung, J. Gill, M. Ghosh, G. Casella, et al. Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5(2):369–411, 2010. 11, 15, 17, 18, 86

N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982. 4

S. Land and J. Friedman. Variable fusion: a new method of adaptive signal regression. *Tehnical Report*, 1996. 7, 17

S. Lang and A. Brezger. Bayesian p-splines. *Journal of computational and graphical statistics*, 13(1):183–212, 2004. 8

Y. Li, S. Wang, P. Song, N. Wang, L. Zhou, and J. Zhu. Doubly regularized estimation and selection in linear mixed-effects models for high-dimensional longitudinal data. *Stat Interface.*, 11:721–737, 2018. 18

Z. Li and M. J. Sillanpää. A bayesian nonparametric approach for mapping dynamic quantitative traits. *Genetics*, 194(4):997–1016, 2013. 6

Z. Li and M. J. Sillanpää. Dynamic quantitative trait locus analysis of plant phenomic data. *Trends in plant science*, 20(12):822–833, 2015. 7

B. Liquet, K. Mengersen, A. Pettitt, M. Sutton, et al. Bayesian variable selection regression of multivariate responses for group data. *Bayesian Analysis*, 12(4):1039–1067, 2017. 16

Z.-H. Lu, H. Zhu, R. C. Knickmeyer, P. F. Sullivan, S. N. Williams, F. Zou, and A. D. N. Initiative. Multiple snp set analysis for genome-wide association studies through bayesian latent variable selection. *Genetic epidemiology*, 39(8):664–677, 2015. 3, 11

D. Lunn, D. Spiegelhalter, A. Thomas, and N. Best. The bugs project: Evolution, critique and future directions. *Statistics in medicine*, 28(25):3049–3067, 2009. 120

M. Lynch, B. Walsh, et al. *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, MA, 1998. 2

C.-X. Ma, G. Casella, and R. Wu. Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. *Genetics*, 161(4):1751–1762, 2002. 6

E. Makalic and D. F. Schmidt. High-dimensional bayesian regularised regression with the bayesreg package. *arXiv preprint arXiv:1611.06649*, 2016. 12

G. Malsiner-Walli and H. Wagner. Comparing spike and slab priors for bayesian variable selection. *arXiv preprint arXiv:1812.07259*, 2018. 10

E. Marchadier, M. Hanemian, S. Tisne, L. Bach, C. Bazakos, E. Gilbault, P. Haddadi, L. Virlouvet, and O. Loudet. The complex genetic architecture of shoot growth natural variation in arabidopsis thaliana. *PLoS genetics*, 15(4):e1007954, 2019. 3, 115

B. D. Marx and P. H. Eilers. Generalized linear regression on sampled signals and curves: a p-spline approach. *Technometrics*, 41(1):1–13, 1999. 7

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032, 1988. 9

R. A. Mrode. *Linear models for the prediction of animal breeding values*. Cabi, 2014. 2, 3

S. Müller, J. Scealy, and A. Welsh. Model selection in linear mixed models. *Statistical Science*, 28:135–167, 2013. 18

T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008. 11, 12

M.-E. Perez, L. R. Pericchi, I. C. Ramírez, et al. The scaled beta2 distribution as a robust prior for scales. *Bayesian Analysis*, 12(3):615–637, 2017. 19

P. Pérez and G. de Los Campos. Genome-wide regression and prediction with the bglr statistical package. *Genetics*, 198(2):483–495, 2014. 11, 12

J. C. Pinheiro and D. M. Bates. Unconstrained parametrizations for variance-covariance matrices. *Statistics and computing*, 6(3):289–296, 1996. 22

N. G. Polson and J. G. Scott. Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian statistics*, 9(501-538):105, 2010. 11, 12

N. G. Polson, J. G. Scott, et al. On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902, 2012. 19

M. Pourahmadi and X. Wang. Distribution of random correlation matrices: Hyperspherical parameterization of the cholesky factor. *Statistics & Probability Letters*, 106: 5–12, 2015. 22

R. Rao and Y. Wu. A strongly consistent procedure for model selection in a regression problem. *Biometrika*, 76(2):369–374, 1989. 18

C. E. Rasmussen and C. Williams. Gaussian processes for machine learning, vol. 1. *MIT press*, 39:40–43, 2006. 8, 16

V. Ročková and E. I. George. Emvs: The em approach to bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846, 2014. 120

V. Ročková and E. I. George. The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444, 2018. 10, 17

V. Rockova and K. McAlinn. Dynamic variable selection with spike-and-slab process priors. *Bayesian Analysis*, 16(1):233–269, 2021. 120

H. Rue and L. Held. *Gaussian Markov random fields: theory and applications*. CRC press, 2005. 8, 17, 86

H. Rue, S. Martino, and N. Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009. 120

F. Scheipl, L. Fahrmeir, and T. Kneib. Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association*, 107(500):1518–1532, 2012. 15, 17, 18, 117

J. G. Scott and J. O. Berger. Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, pages 2587–2619, 2010. 10

M. Smith and R. Kohn. Nonparametric regression using bayesian variable selection. *Journal of Econometrics*, 75(2):317–343, 1996. 10

C. Sun, P. Madsen, U. Nielsen, Y. Zhang, M. Lund, and G. Su. Comparison between a sire model and an animal model for genetic evaluation of fertility traits in danish holstein population. *Journal of Dairy Science*, 92(8):4063–4071, 2009. 2

A. Terenin, S. Dong, and D. Draper. Gpu-accelerated gibbs sampling: a case study of the horseshoe probit model. *Statistics and Computing*, 29(2):301–310, 2019. 120

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. 11

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005. 7, 17

S. Tisné, M. Denis, D. Cros, V. Pomiès, V. Riou, I. Syahputra, A. Omoré, T. Durand-Gasselin, J.-M. Bouvet, and B. Cochard. Mixed model approach for ibd-based qtl mapping in a complex oil palm pedigree. *BMC genomics*, 16(1):1–12, 2015. 3, 115

S. Tisné, M. Denis, H. Domonhedo, B. Pallas, M. Cazemajor, T. J. Tranbarger, and F. Morcillo. Environmental and trophic determinism of fruit abscission and outlook with climate change in tropical regions. *Plant-Environment Interactions*, 1(1):17–28, 2020. 85

F. Vaida and S. Blanchard. Conditional akaike information for mixed-effects models. *Biometrika*, 92(2):351–370, 2005. 18

R. van de Schoot, S. Depaoli, R. King, B. Kramer, K. Märtens, M. G. Tadesse, M. Vannucci, A. Gelman, D. Veen, J. Willemsen, et al. Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1):1–26, 2021. 10

S. van der Pas, J. Scott, A. Chakraborty, and A. Bhattacharya. horseshoe: Implementation of the horseshoe prior. *R package version 0.1. 0*, 2016. 12

F. A. van Eeuwijk, M. Boer, L. R. Totir, M. Bink, D. Wright, C. R. Winkler, D. Podlich, K. Boldman, A. Baumgarten, M. Smalley, et al. Mixed model approaches for the identification of qtls within a maize hybrid breeding program. *Theoretical and Applied Genetics*, 120(2):429–440, 2010. 3

J. Vanhatalo, Z. Li, and M. J. Sillanpää. A gaussian process model and bayesian variable selection for mapping function-valued quantitative traits with incomplete phenotypic data. *Bioinformatics*, 35(19):3684–3692, 2019. 16

G. Wahba. *Spline models for observational data*. SIAM, 1990. 8

M. West, P. J. Harrison, and H. S. Migon. Dynamic generalized linear models and bayesian forecasting. *Journal of the American Statistical Association*, 80(389):73–83, 1985. 6

A. J. Wilson, D. Reale, M. N. Clements, M. M. Morrissey, E. Postma, C. A. Walling, L. E. Kruuk, and D. H. Nussey. An ecologist's guide to the animal model. *Journal of Animal Ecology*, 79(1):13–26, 2010. 2

X. Xu, M. Ghosh, et al. Bayesian variable selection and estimation for group lasso. *Bayesian Analysis*, 10(4):909–936, 2015. 15, 16

X. Yang, N. N. Narisetty, et al. Consistent group selection with bayesian high dimensional modeling. *Bayesian Analysis*, 15(3):909–935, 2020. 16

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1): 49–67, 2006. 14, 15

S. Zhang and M. Li. "dlbayes" available at cran, r package for implementing the dirichlet-laplace shrinkage prior in bayesian linear regression and variable selection. 2018. 12

W. Zhang, C. Leng, and C. Y. Tang. A joint modelling approach for longitudinal studies. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 219–238, 2015. 18

Y. D. Zhang, B. P. Naughton, H. D. Bondell, and B. J. Reich. Bayesian regression using a prior on the model fit: The r2-d2 shrinkage prior. *Journal of the American Statistical Association*, (just-accepted):1–37, 2020. 12, 13

Résumé : En agronomie et plus spécifiquement en amélioration génétique, le génotypage haut débit a été largement mis à profit, depuis maintenant plus de 20 ans, pour accéder à une information génétique toujours plus riche et abondante. Celle-ci a permis d'identifier les positions le long du génome impliquées dans la variabilité de caractères d'intérêt. Plus récemment, les méthodes de phénotypage haut débit ont fait leur apparition. Elles donnent accès au suivi de l'évolution de plusieurs caractères phénotypiques au cours du temps. Ces données, longitudinales, permettent d'étudier finement la dynamique évolutive de ces caractères tout en identifiant les facteurs environnementaux qui influencent leur variabilité selon les stades de développement.
Cependant, l'analyse de telles données soulève plusieurs défis statistiques. Cette thèse propose des développements méthodologiques afin de prendre en compte les dépendances entre observations et entre variables, de sélectionner les variables génétiques ou environnementales pertinentes, ou encore d'estimer des effets qui évoluent au cours du temps. Le cadre bayésien est un formalisme statistique élégant pour répondre à ces différentes problématiques notamment au travers de la construction de lois *a priori*.
Nous étudions et comparons différentes lois *a priori* pour simultanément inférer et sélectionner les effets fixes et/ou aléatoires quand ceux-ci peuvent être nombreux. Nous considérons différents cadres de modélisation statistique classiquement utilisés pour l'analyse de données longitudinales. En particulier, nous nous focaliserons sur les modèles à coefficients variants, les modèles linéaires mixtes ou encore la régression sur signal.
Ce travail a été motivé par différentes applications pratiques portant sur l'évolution temporelle de l'architecture génétique, la détection de QTL ou l'impact des variations climatiques sur la variabilité phénotypique. Trois jeux de données, issus de contextes agronomiques variés, sont utilisés pour illustrer ces nouvelles approches.

Mots clés : Données corrélées, Données longitudinales, Modèle linéaire mixte, Modèle à coefficients variants, Sélection de variables à effets fixes et aléatoires.

Abstract: In agronomy, and more specifically in genetic breeding, high throughput genotyping has been widely used for more than 20 years to access increasingly rich and abundant genetic information. This has allowed the identification of positions along the genome involved in the variability of traits of interest. More recently, high throughput phenotyping methods have been developed. They give access to the monitoring of the evolution of several phenotypic traits over time. These longitudinal data allow a fine study of the dynamics of these traits, while identifying the environmental factors that influence their variability according to developmental stages.
However, the analysis of such data raises several statistical challenges. This thesis proposes methodological developments in order to take into account the dependencies between observations and between variables, to select relevant genetic or environmental variables, or to estimate effects that evolve over time. The Bayesian framework is an elegant statistical formalism to address these different issues, especially through the construction of priors. We study and compare different priors to simultaneously infer and select fixed and/or random effects when they can be numerous. We consider different statistical modeling frameworks classically used for longitudinal data analysis. In particular, we focus on linear mixed models, varying coefficient models or signal regression.
This work was motivated by various practical applications concerning the QTL detection, the temporal evolution of genetic architecture or the impact of climatic variations on phenotypic variability. Three datasets, from various agronomical contexts, are used to illustrate these new approaches.

Keywords: Correlated data, Longitudinal data, Linear mixed model, Varying coefficients model, Fixed and random variables selection.