Sélection bayésienne de variables pour modèle linéaire à coefficients dynamiques

Benjamin Heuclin [1] & Marie Denis [2] & Frédéric Mortier [3] & Catherine Trottier [1,4]

[1] *Institut Montpelliérain Alexander Grothendieck, CNRS, Université de Montpellier ; benjamin.heuclin@umontpellier.fr, catherine.trottier@umontpellier.fr.*
[2] *UMR AGAP, CIRAD; marie.denis@cirad.fr.*
[3] *UPR Forêts et Sociétés, CIRAD; frederic.mortier@cirad.fr.*
[4] *Université Paul Valéry Montpellier 3; catherine.trottier@univ-montp3.fr.*

Résumé.
Comment l'architecture génétique des caractères quantitatifs évolue-t-elle au cours du temps ? La réponse à cette question est cruciale pour de nombreux domaines d'application tels que la génétique humaine et la sélection végétale ou animale. Au cours des dernières décennies, des techniques de génotypage à haut débit ont été utilisées pour mieux comprendre les liens entre l'information génétique et les caractères phénotypiques. Récemment, des méthodes de phénotypage à haut débit ont également été utilisées pour fournir de grandes quantités d'information à l'échelle phénotypique. En particulier, ces méthodes permettent de mesurer les caractères dans le temps, et ce, pour un grand nombre d'individus. La combinaison de ces deux informations peut donner des indications sur l'évolution de l'architecture génétique au cours du temps. Toutefois, ces données soulèvent de nouveaux défis statistiques liés, entre autres, à la dimension élevée, aux dépendances temporelles et aux effets variant dans le temps. Dans ce travail, nous proposons un modèle linéaire dynamique bayésien permettant, en une seule étape, l'identification des marqueurs génétiques impliqués dans la variabilité des caractères phénotypiques et l'estimation de leurs effets dynamiques. Cette approche combine des priors de type spike-and-slab pour la sélection de variables avec de l'interpolation de type P-spline pour l'estimation fonctionnelle des effets dynamiques.

Mots-clés. Modèle linéaire dynamique, Sélection de variable, P-spline, Méthodes bayésiennes, Méthode MCMC, Biostatistique, Génétique quantitative.

Abstract.
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 dynamic linear 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. This approach combines spike-and-slab priors for variable selection with P-spline interpolation for functional estimation of the dynamic effects.

Keywords. Varying coefficient model, Variable selection, P-spline, Bayesian method, MCMC method, Biostatistics, Quantitative genetics.

1 Introduction

Les programmes d'amélioration génétique, que ce soit dans le domaine végétal ou animal, 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 l'utilisation de modèles statistiques. Les outils de génotypage haut débit permettent d'utiliser de 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. Plus récemment, des outils de phénotypage haut débit ont fait leur apparition, tel que le robot Phénoscope dans le domaine végétal sur *Arabidopsis Thaliana* (Marchadier et al., 2018). Ainsi, il est possible d'obtenir un suivi régulier de caractères phénotypiques des individus dans le temps. Toutefois, l'utilisation de données de phénotypage et de génotypage à haut débit soulève des questions méthodologiques nouvelles, en particulier, en lien avec l'estimation de coefficients dynamiques de régression ainsi que la sélection de variables dans un modèle linéaire dynamique.

L'introduction de paramètres qui varient dans le temps dans les modèles accroît considérablement le nombre de paramètres à estimer. Dans le contexte de la génétique d'association, des approches fonctionnelles ont été développées (Wu et al., 2003; Wang et al., 2014) faisant l'hypothèse que les effets au cours du temps des variables génétiques peuvent être décrits par des fonctions paramétriques telles que des courbes logistiques ou de croissances. Cela permet de réduire considérablement le nombre de paramètres à estimer mais présente des limites dans la modélisation des effets pour des caractères complexes. Afin d'y remédier, des approches fonctionnelles non-paramétriques ont été développées parmi lesquelles celle de Wang et al. (2008) qui proposent une base de B-splines pour estimer les fonctions, ou encore celle de Li et al. (2015), qui utilisent une base polynomiale de Legendre. Les B-splines sont des outils d'interpolation performants, mais présentent l'inconvénient d'être sensibles au choix des nœuds. Li and Sillanpää (2013) proposent d'utiliser une approche d'interpolation P-spline adaptée au contexte

bayésien (Eilers and Marx, 1996; Lang and Brezger, 2004) permettant de s'affranchir du choix de la position des nœuds. Pour sélectionner les variables, (Wang et al., 2008) utilisent une procédure de vraisemblance pénalisée (SCAD). Dans un cadre bayésien, Li et al. (2015) proposent d'utiliser un prior visant à reproduire une sélection de type groupe Lasso. L'inconvenient de ce genre de prior est qu'ils peuvent produire des estimations biaisées, nécessitant un ajustement en deux étapes. Li and Sillanpää (2013) procèdent à une sélection de modèles *a posteriori* de type forward.

Nous proposons, dans un contexte bayésien, de combiner l'interpolation P-spline des effets dynamiques de régression avec une sélection de type spike-and-slab (George and McCulloch, 1993) sur les coefficients des P-splines. Cette approche permet, en une seule étape, d'estimer des courbes complexes d'effets dynamiques de régression et de sélectionner les variables génétiques significatives. Un algorithme de Monte Carlo par chaînes de Markov (MCMC) de type Gibbs a été implémenté pour inférer les paramètres du modèle.

2 Modèle statistique

Soit $Y_i = (y_i(t_1), \ldots, y_i(T))'$ le vecteur de longueur T des observations d'un caractère phénotypique de l'individu i sur T temps. Y_i peut se décomposer comme suit :

$$\begin{pmatrix} y_i(t_1) \\ \vdots \\ y_i(T) \end{pmatrix} = \begin{pmatrix} \mu(t_1) \\ \vdots \\ \mu(T) \end{pmatrix} + \begin{pmatrix} \beta_1(t_1) & \ldots & \beta_q(t_1) \\ \vdots & & \vdots \\ \beta_1(T) & \ldots & \beta_q(T) \end{pmatrix} \begin{pmatrix} X_{i,1} \\ \vdots \\ X_{i,q} \end{pmatrix} + \begin{pmatrix} \varepsilon_i(t_1) \\ \vdots \\ \varepsilon_i(T) \end{pmatrix} \quad (1)$$

avec $\mu = (\mu(t_1), \ldots, \mu(T))'$ le vecteur de moyenne générale de longueur T, $X_i = (X_{i,1}, \ldots, X_{i,q})'$ le vecteur de longueur q des marqueurs moléculaires de l'individu i, β la matrice de dimension (T x q) des coefficients de régression. La $t^{\text{ème}}$ ligne de β donne les coefficients de régression associé à chaque marqueur moléculaire au temps t et la $j^{\text{ème}}$ colonne donne l'évolution au cours du temps de l'effet du $j^{\text{ème}}$ marqueur moléculaire. $\varepsilon_i = (\varepsilon_i(t_1), \ldots, \varepsilon_i(T))'$ le vecteur de longueur T des erreurs résiduelles. Nous supposons que les mesures répétées d'un même individu sont corrélées, ainsi ε_i suit une loi gaussienne $N_T(0, \Sigma = \sigma^2\Gamma)$. La matrice de corrélation Γ est supposée être celle d'un modèle auto-régressif d'ordre 1. Cette modélisation suppose une variance constante et une corrélation décroissante en fonction de la distance temporelle entre deux mesures $\rho^{|t_i - t_j|}$, $0 < \rho < 1$, $i, j = 1, \ldots, T$.

Chaque vecteur de coefficients dynamiques associé à une variable est alors interpolé par un polynôme P-spline introduit par Eilers and Marx (1996). L'approche P-spline consiste à interpoler un polynôme sur un ensemble de points en utilisant une base B-spline (De Boor et al., 1978; Dierckx, 1995) tout en pénalisant la dérivée seconde de la

fonction pour obtenir une fonction lisse.

$$\begin{pmatrix} \beta_j(t_1) \\ \vdots \\ \beta_j(T) \end{pmatrix} = \sum_{k=1}^{v} B_k b_{k,j} = B b_j \tag{2}$$

où v est le nombre de degrés de liberté de la base correspondant à la somme du nombre de nœuds et du degré des polynômes utilisés entre chaque nœud, B est la matrice de base des B-splines, b_j le vecteur des coefficients des B-splines associés à la $j^{\text{ème}}$ variable et $b = (b_1, \ldots, b_q)$ est la matrice des collections des vecteurs de coefficients de B-splines. On peut également réécrire le vecteur de moyenne générale : $(\mu(t_1), \ldots, \mu(T))' = Bm$ avec $m = (m_1, \ldots, m_v)'$. L'équation (1) peut donc se réécrire sous la forme :

$$Y_i = Bm + BbX_i + \varepsilon_i, \qquad \varepsilon_i \sim N_T(0, \sigma^2\Gamma) \tag{3}$$

L'approche bayésienne de la modélisation P-spline (Lang and Brezger, 2004) consiste à supposer une loi gaussienne sur les coefficient polynomiaux :

$$b_j \sim N_v(0, \sigma^2\tau^2 K^{-1}), \qquad K = \begin{pmatrix} 1 & -2 & 1 & & & & \\ -2 & 5 & -4 & \ddots & & & \\ 1 & -4 & 6 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 6 & -4 & 1 \\ & & & \ddots & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{pmatrix} \tag{4}$$

avec K la matrice associée à l'opérateur des différences finies permettant de pénaliser la dérivée seconde des fonctions.

Pour identifier les variables influençant la variabilité du caractère phénotypique, il est nécessaire de mettre en œuvre une sélection. Pour cela, nous combinons le prior de type P-spline sur b_j avec un prior de type zero-inflated group spike-and-slab. Le prior spike-and-slab a été introduit par George and McCulloch (1993) et est couramment rencontré dans le contexte de la sélection de variable bayésienne. Pour chaque variable, ce prior consiste à introduire une variable indicatrice latente γ_j de loi Bernoulli de paramètre π, et à décomposer la loi *a priori* de l'effet de la variable comme un mélange de deux lois : une diffuse si $\gamma_j = 1$ (partie slab) et une resserrée autour de zéro si $\gamma_j = 1$ (partie spike). Le prior zero-inflated spike-and-slab (George and McCulloch, 1997) consiste à placer une masse de Dirac sur la partie spike permettant d'estimer certains effets strictement égaux à zéro. Ce prior présente plusieurs avantages. Il permet, à chaque itération d'un algorithme MCMC, de ne considérer qu'un sous-groupe des variables sélectionnées et ainsi d'éviter des problèmes de sur-ajustement. Il permet également d'avoir accès aux probabilités *a*

posteriori marginales et jointes de sélection des variables. Xu et al. (2015) l'ont étendu pour faire de la sélection par groupe. Dans notre contexte, nous considérons le vecteur des effets polynomiaux b_j comme un groupe, nous utilisons alors un prior zero-inflated group spike-and-slab avec, pour loi sur la partie slab, la loi gaussienne multivariée (4).

Finalement, nous obtenons le modèle hiérarchique suivant :

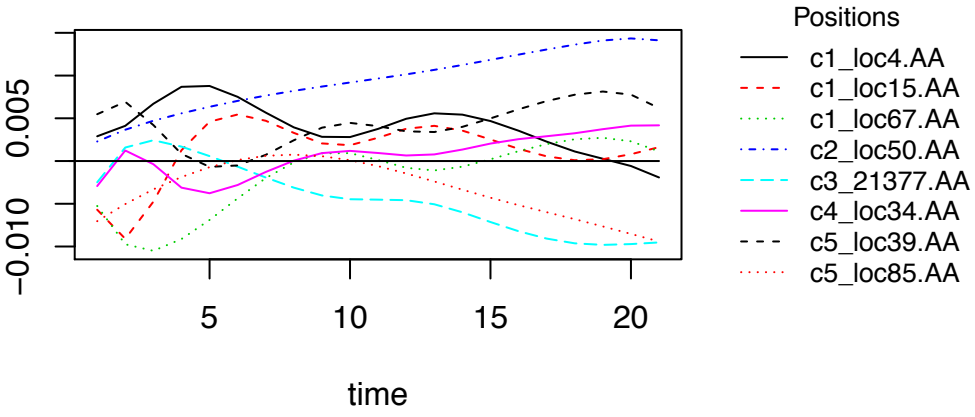
$$\begin{aligned}
&Y_i|m, b, \rho, \sigma^2 \sim N_T(Bm + BbX_i, \sigma^2\Gamma)\\
&m \sim N_v(0, \tau_0^2 K^{-1})\\
&b_j|\gamma_j, \tau_j^2, \sigma^2 \sim \gamma_j N_v(0, \sigma^2\tau^2 K^{-1}) + (1-\gamma_j)\delta_v(0), \quad j = 1, \ldots, q\\
&\gamma_j \sim Ber(\pi), \quad j = 1, \ldots, q\\
&\rho \sim U_{[-1,1]},\\
&\sigma^2 \sim IG(s_{\sigma^2}, r_{\sigma^2})
\end{aligned} \tag{5}$$

Pour inférer les paramètres de ce modèle, un algorithme de Metropolis-Hastings within Gibbs a été implémenté.

3 Essais numériques et discussions

Nous procédons tout d'abord à des simulations afin d'évaluer les performances de sélection et d'estimation de l'approche que nous proposons. Pour cela différents jeux de données ont été simulés en faisant varier le rapport signal sur bruit et le rapport nombre d'observations sur nombre de variables. Nous comparons également cette approche avec celle de Li et al. (2015). Les résultats mettent en évidence les performances de l'approche proposée tant sur la qualité de la sélection que sur l'estimation. En particulier, ils soulignent l'intérêt de combiner une modélisation par P-splines pour interpoler précisément des courbes et de procéder à la sélection des variables pertinentes dans la procédure d'estimation. En effet, les erreurs quadratiques moyennes sont systématiquement plus faibles en utilisant cette nouvelle approche que celles obtenues par la méthode introduite par Li et al. (2015). Ces différences sont d'autant plus fortes lorsque q le nombre de marqueurs augmente.

L'approche développée a également été mise en œuvre sur des jeux de données réelles générées par le robot Phénoscope (https://phenoscope.versailles.inra.fr/) sur l'espèce Arabidopsis Thaliana (Marchadier et al., 2018). 300 individus suivis en laboratoire sur $T = 21$ jours ont été séquencés sur 92 marqueurs. Huit marqueurs sont identifiés et sélectionnés. Les effets associés sont présentés dans la figure ci-dessous :



Pour conclure et en guise de perspective, il serait intéressant d'étendre l'approche pour prendre en compte différents environnements afin d'étudier les interactions génotype-environnement.

References

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

Dierckx, P. (1995). Curve and Surface Fitting with Splines. Oxford University Press.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B -splines and penalties. Statistical Science, 11(2):89–121.

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. page 35.

Lang, S. and Brezger, A. (2004). Bayesian P-Splines. Journal of Computational and Graphical Statistics, 13(1):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(2):640–664.

Li, Z. and Sillanpää, M. J. (2013). A Bayesian Nonparametric Approach for Mapping Dynamic Quantitative Traits. Genetics, 194(4):997–1016.

Marchadier, E., Hanemian, M., Tisne, S., Bach, L., Bazakos, C., Gilbault, E., Haddadi, P., Virlouvet, L., and Loudet, O. (2018). The complex genetic architecture of shoot growth natural variation in Arabidopsis thaliana.

Wang, L., Li, H., and Huang, J. Z. (2008). Variable Selection in Nonparametric Varying-Coefficient Models for Analysis of Repeated Measurements. Journal of the American Statistical Association, 103(484):1556–1569.

Wang, Z., Pang, X., Wu, W., Wang, J., Wang, Z., and Wu, R. (2014). MODELING PHENOTYPIC PLASTICITY IN GROWTH TRAJECTORIES: A STATISTICAL FRAMEWORK: PERSPECTIVE. Evolution, 68(1):81–91.

Wu, R., Ma, C.-X., Zhao, W., and Casella, G. (2003). Functional mapping for quantitative trait loci governing growth rates: A parametric model. Physiological Genomics, 14(3):241–249.

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