Une approche de sélection de variables pour améliorer l'estimation d'héritabilité dans les modèles linéaires mixtes parcimonieux

Anna Bonnet [1] & Elisabeth Gassiat [2] & Céline Lévy-Leduc [3]

[1] *AgroParisTech/INRA, UMR 518 MIA, F-75005, Paris anna.bonnet@agroparistech.fr*
[2] *Laboratoire de Mathématique d'Orsay, Université Paris-Sud, F-91405, Orsay elisabeth.gassiat@math.u-psud.fr*
[3] *AgroParisTech/INRA, UMR 518 MIA, F-75005, Paris celine.levy-leduc@agroparistech.fr*

Résumé.
L'héritabilité d'un caractère biologique est définie comme la part de sa variation au sein d'une population qui est causée par des facteurs génétiques. Nous proposons dans un premier temps un estimateur de l'héritabilité dans les modèles linéaires mixtes parcimonieux en grande dimension, dont nous avons étudié les propriétés théoriques. Nous mettons en évidence que lorsque la taille des effets aléatoires N est trop grande par rapport au nombre d'observations n, nous ne pouvons fournir une estimation précise pour l'héritabilité.

La deuxième partie de notre travail a été de proposer une méthode de sélection de variables afin de réduire la taille des effets aléatoires, dans le but d'améliorer la précision de l'estimation de l'héritabilité. Néanmoins, nous montrons sur des simulations que ce type d'approche fonctionne uniquement lorsque le nombre composantes non nulles dans les effets aléatoires, c'est-à-dire le nombre de variants génétiques qui influencent la variation phénotypique, est assez faible. Nous avons ensuite établi un critère empirique pour déterminer les cas où il était possible de faire de la sélection de variables.

Nous avons appliqué notre méthode sur des données d'imagerie médicale sur le cerveau.

Mots-clés. Grande dimension, Héritabilité, Modèles linéaires mixtes, Sélection de variable

Abstract.
The heritability is defined for any biological quantitative feature as the proportion of its variation which can be explained by genetic factors. Firstly, we propose an estimator for the heritability in high dimensional sparse linear mixed models and we study its theoretical properties. We highlight that in the case where the size N of the random effects is too large compared to the number n of observations, we cannot provide a precise estimation for the heritability.

The next part of our work consists in proposing a variable selection method to reduce the size of the random effects, which improves the accuracy of the heritability estimation. However, we show on simulations that this kind of approach only works when the number of non null components in the random effects, that is the genetic variants which have an impact on the phenotypic variation, is small enough. We devised an empirical criterion to determine whether it is possible to apply our variable selection approach.

We applied our method on data obtained by measuring the volume of several regions of the brain.

Keywords. Heritability, High dimension, Linear mixed models, Variable selection

1 Introduction

L'héritabilité d'un caractère biologique est définie comme la part de sa variation au sein d'une population qui est causée par des facteurs génétiques. Notre objectif est de proposer une méthode pour estimer cette héritabilité.

2 Modèle

Pour estimer l'héritabilité de la taille, Yang et al. [4] ont proposé d'utiliser un modèle mixte défini comme suit :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \ , \tag{1}$$

où $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ est le vecteur des observations (phénotypes), $\mathbf{X}$ est une matrice $n \times p$ de prédicteurs, $\boldsymbol{\beta}$ est un vecteur $p \times 1$ qui contient les effets inconnus des prédicteurs. La matrice $\mathbf{Z}$, de taille $n \times N$, contient l'information génétique de tous les individus aux positions où il existe des variations au sein de la population mais où la copie la moins fréquente n'est pas trop rare : ces positions sont appelées des SNPs (Single Nucleotide Polymorphism). Plus précisément, $\mathbf{Z}$ est définie par :

$$\mathbf{Z}_{i,j} = \frac{W_{i,j} - \overline{W}_j}{s_j}, \ i = 1, \ldots, n, \ j = 1, \ldots, N \ , \tag{2}$$

où

$$\overline{W}_j = \frac{1}{n}\sum_{i=1}^{n} W_{i,j}, \ s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(W_{i,j} - \overline{W}_j)^2, \ j = 1, \ldots, N \ , \tag{3}$$

et la matrice $\mathbf{W}$ est telle que $W_{i,j} = 0$ (resp. 1, resp. 2) si le génotype du $i$eme individu au locus j est qq (resp. Qq, resp. QQ) où p_j est la fréquence de l'allèle Q au locus j.

Dans [4], $\mathbf{u}$ et $\mathbf{e}$ correspondent aux effets aléatoires, plus précisément la $i$ème composante de u donne l'effet du $i$ème SNP et e correspond aux effets environnementaux. Dans ce modèle,

$$\mathbf{u} \sim \mathcal{N}(0, \sigma_u^{\star 2}\mathrm{Id}_{\mathbb{R}^N}) \ \text{ et } \mathbf{e} \sim \mathcal{N}(0, \sigma_e^{\star 2}\mathrm{Id}_{\mathbb{R}^n}) \ .$$

où $\mathrm{Id}_{\mathbb{R}^n}$ est la matrice identité de taille $n \times n$.

L'héritabilité est définie par le ratio

$$\eta^\star = \frac{N\sigma_u^{\star 2}}{N\sigma_u^{\star 2} + \sigma_e^{\star 2}} \tag{4}$$

ce qui correspond bien à la part de variation phénotypique expliquée par les variations génétiques.

En tenant compte du fait que u peut contenir des composantes nulles, nous avons étudié le cas où

$$u_i \overset{i.i.d.}{\sim} (1-q)\delta_0 + q\mathcal{N}(0, \sigma_u^{\star 2}) \text{ , pour tout } 1 \leqslant i \leqslant N,$$

c'est-à-dire que seulement une proportion q inconnue des SNPs aurait un impact sur le phénotype. Nous adaptons alors la définition de l'héritabilité donnée par l'équation (4) comme suit :

$$\eta^\star = \frac{qN\sigma_u^{\star 2}}{qN\sigma_u^{\star 2} + \sigma_e^{\star 2}}.$$

3 Estimation de l'héritabilité

Nous avons proposé dans [1] un estimateur de l'héritabilité dans les modèles linéaires mixtes parcimonieux en grande dimension, dont nous avons étudié les propriétés théoriques. Nous avons montré que cet estimateur était consistant et nous avons pu calculer sa variance asymptotique.

Ce résultat ainsi que des simulations de Monte-Carlo (Figure 1) mettent en évidence que lorsque la taille N des effets aléatoires est trop grande par rapport au nombre n d'observations, nous ne pouvons fournir une estimation précise pour l'héritabilité. La taille typique des données que nous étudions vérifie justement la condition $N >> n$.

4 Sélection de variables

La deuxième partie de notre travail a été de proposer une méthode de sélection de variables afin de réduire la taille des effets aléatoires, dans le but d'améliorer la précision de l'estimation de l'héritabilité. Notre approche se décompose en plusieurs étapes :

• **Première étape : calcul des corrélations empiriques [2]**. Cette étape consiste à réduire le nombre de colonnes de $\mathbf{Z}$ qui sont pertinentes dans notre étude en essayant d'éliminer celles qui sont associées aux composantes nulles de u. Plus précisément, pour chaque colonne j, on calcule le coefficient de corrélation empirique $C_j = \left|\frac{1}{n}\sum Y_i Z_{i,j}\right|$ et on garde les colonnes ayant les plus grands C_j. La matrice $\mathbf{Z}$ restreinte aux colonnes les plus corrélées au phénotype observé est notée $\mathbf{Z}_{red}$.

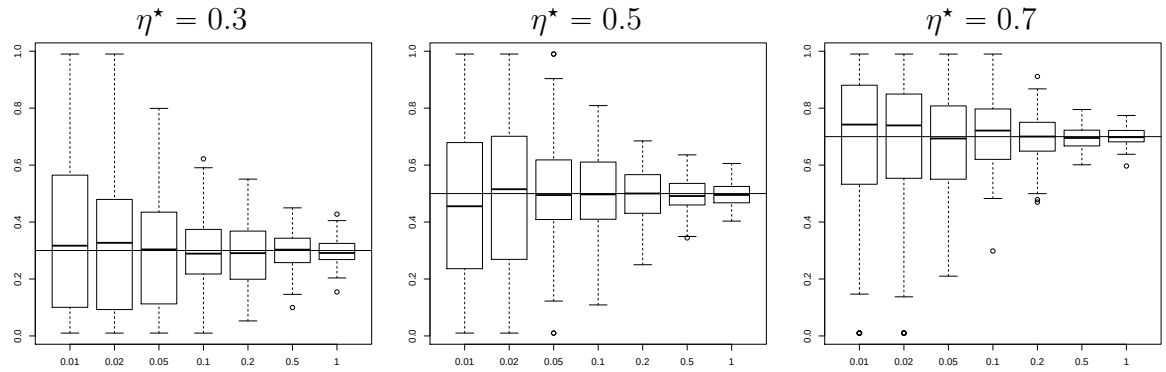

FIGURE 1 – Estimation de $\eta^\star$ pour différentes valeurs de $a = n/N$ en abscisse et de $\eta^\star$.

• **Deuxième étape : le critère LASSO**. Cette étape consiste à minimiser par rapport à u le critère suivant :

$$Crit_\lambda(u) = \|Y - Z_{red}u\|_2^2 + \lambda\|u\|_1$$

qui dépend d'un paramètre $\lambda > 0$. Pour choisir les composantes non nulles de $\mathbf{u}$ on procède de la façon suivante : le vecteur d'observations Y est aléatoirement divisé en plusieurs sous-échantillons de taille $n/2$. Pour chaque sous-échantillon, on minimise le critère LASSO avec λ choisi petit de sorte qu'on garde un nombre maximal de composantes de $\mathbf{u}$. Alors, pour un seuil fixé, nous gardons dans l'ensemble final de composantes sélectionnées uniquement les composantes qui sont apparues un nombre de fois supérieur à ce seuil. Cette technique, appelée stability selection, est proposée dans [3].

Notre méthode est implémentée dans le package R EstHer. Elle fournit également un intervalle de confiance pour l'héritabilité grâce une méthode de bootstrap non paramétrique adaptée à des observations corrélées. Des simulations de Monte-Carlo sont réalisées pour montrer à la fois les performances de notre estimateur et la qualité des intervalles de confiance.

5 Performances des méthodes avec et sans sélection

On distingue deux cas en fonction du nombre de composantes de $\mathbf{u}$ non nulles, ce qui correspond au nombre de SNPs qui sont causaux. En effet, si ce nombre de SNPs causaux est relativement faible (par exemple, 100 SNPs), alors la sélection fonctionne et notre estimateur de l'héritabilité est, d'après les simulations, non biaisé et avec une variance

très faible. Par contre si le nombre de SNPs causaux est élevé, alors EstHer sous-estime l'héritabilité. Ces résultats sont montrés Figure 2. On voit également sur cette figure que la méthode sans sélection, n'est pas affectée par le nombre de SNPs causaux.

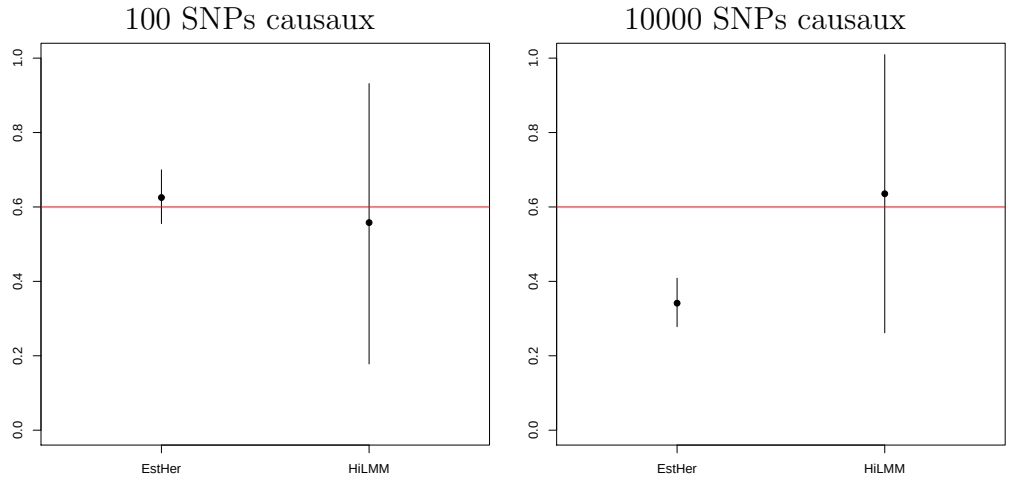

FIGURE 2 – Estimation de $\eta^\star$ en utilisant EstHer (sélection de variable) et HiLMM (estimation directe sans sélection), avec intervalles de confiance à 95%. A gauche le cas où 100 SNPs sont causaux, à droite 10000 SNPs causaux. La ligne rouge représente la vraie valeur de $\eta^\star$, c'est-à-dire 0.6.

Nous souhaitons pouvoir distinguer les situations "nombre de SNPs causaux faible" et "nombre de SNPs causaux élevé" pour pouvoir appliquer EstHer dans le premier cas et ainsi obtenir un estimateur non biaisé avec une variance faible, et appliquer HiLMM dans le deuxième cas pour avoir un estimateur non biaisé malgré sa grande variance. Nous avons donc déterminé un critère empirique qui permet de savoir si le nombre de SNPs causaux est suffisamment faible pour que l'on puisse appliquer EstHer ou non. Ce critère est fondé sur l'observation suivante : lorsque le nombre de SNPs causaux est faible, le choix du seuil dans la stability selection a peu d'influence sur les SNPs sélectionnés et donc sur l'héritabilité estimée. Au contraire, lorsque le nombre de SNPs causaux est élevé, une légère variation du seuil entraîne une différence importante dans l'estimation de l'héritabilité. Nous avons donc quantifié les variations de l'héritabilité estimée pour différents seuils afin de proposer un critère pour appliquer ou non notre méthode de sélection de variables.

6 Application

Nous avons appliqué notre méthode sur des données sur le cerveau : il s'agit d'environ 2000 adolescents qui ont été génotypés et dont le volume des différentes régions du cerveau a été mesuré grâce à des IRM.

Références

[1] Anna Bonnet, Elisabeth Gassiat, and Celine Levy-Leduc. Heritability estimation in high-dimensional sparse linear mixed models. *Electronic journal of Statistics*, 2015.

[2] Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *JRSS : Series B (Statistical Methodology)*, 2008.

[3] N. Meinshausen and P. Bühlmann. Stability selection. *JRSS : Series B (Statistical Methodology)*, 2010.

[4] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael E Goddard, and Peter M Visscher. Common snps explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7) :565–569, 2010.