Estimation de l'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. Pour de nombreux caractères complexes, il existe une grande différence entre la variation génétique expliquée par les études de population et celle expliquée par les variants spécifiques révélés grâce aux études d'association (GWAS). Nous proposons 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 est trop grande par rapport au nombre d'observations, nous ne pouvons fournir une estimation précise pour l'héritabilité. Malheureusement, la taille typique des données que nous étudions vérifie justement la condition $N >> n$, par exemple $n = 200$ et $N = 500000$.

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 méthode fournit également un intervalle de confiance grâce à une méthode de bootstrap non paramétrique adaptée à des observations corrélées.

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. Nous trouvons des résultats cohérents avec ceux trouvés avec des méthodes sans sélection de variable, mais nous avons des intervalles de confiance plus petits.

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. For many complex traits in human population, there is a huge gap between the genetic variance explained by population studies and the variance explained by specific variants found thanks to genome wide

association studies (GWAS). 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 of the random effects is too large compared to the number of observations, we cannot provide a precise estimation for the heritability. Unfortunately, typical datasets verify the condition $N >> n$ (for example, $n = 2000$ and $N = 500000$).

The next part of our work is to perform a variable selection method to reduce the size of the random effects and to improve the accuracy of the heritability estimation. We also provide confidence intervals by using an adapted non parametic bootstrap method which can deal with correlated observations.

We apply our method on brain data obtained by measuring the volume of several regions of the brain of approximately 2000 adolescents, who had also been genotyped. We compare the results to those obtained using methods without any selection, and we show that our method allows us to reduce the length of the confidence interval for the heritability.

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. Pour de nombreux caractères complexes, il existe une grande différence entre la variation génétique expliquée par les études de population et celle expliquée par les variants spécifiques révélés grâce aux études d'association (GWAS) [3].

2 Modèle

Pour estimer cette héritabilité manquante, Yang et al. [7] ont proposé d'utiliser un modèle mixte défini comme suit :

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

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 Z 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, Z est définie par :

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

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 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.

Z est donc une matrice dont le nombre de lignes n est égal au nombre d'individus (dans la plupart des cohortes, $n \approx 1000$) et le nombre de colonnes N est égal au nombre de SNPs pris en compte dans l'expérience, dans notre application $N \approx 300,000$.

Dans [7], **u** et **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

$$\eta^\star = \frac{N\sigma_u^{\star 2}}{N\sigma_u^{\star 2} + \sigma_e^{\star 2}}, \qquad (\star)$$

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 $(\star)$ 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. L'un de nos résultats est un théorème central limite dans le cas où le nombre d'observations n et la taille des effets aléatoires N tendent vers l'infini.

Théorème : Soit $Y = (Y_1, \ldots, Y_n)'$ qui vérifie le modèle (1) et supposons que les variables aléatoires $Z_{i,j}$ sont i.i.d. $\mathcal{N}(0,1)$. Alors pour tout $q \in (0,1]$, quand $n, N \to \infty$

et $n/N \to a > 0$, $\sqrt{n}(\hat{\eta} - \eta^\star)$ converge en distribution vers une loi normale centrée de variance

$$\tilde{\sigma}^2(a, \eta^\star, q) = \frac{2}{\hat{\sigma}^2(a, \eta^\star)} + 3\frac{a^2 \eta^{\star 2}}{\hat{\sigma}^4(a, \eta^\star)} \left(\frac{1}{q} - 1\right) S(a, \eta^\star)$$

où $S(a, \eta^\star) = \left[\int \frac{\lambda(\lambda-1)}{(\eta^\star(\lambda-1)+1)^2} d\mu_a(\lambda) - \int \frac{\lambda}{(\eta^\star(\lambda-1)+1)} d\mu_a(\lambda) \int \frac{\lambda-1}{(\eta^\star(\lambda-1)+1)} d\mu_a(\lambda)\right]^2$

et $\hat{\sigma}^2(a, \eta^\star) = \left\{\int (\frac{\lambda-1}{\eta(\lambda-1)+1})^2 d\mu_a(\lambda) - \left(\int \frac{\lambda-1}{\eta(\lambda-1)+1} d\mu_a(\lambda)\right)^2\right\}$

et μ_a est la mesure de Marchenko-Pastur [4].

Ce résultat ainsi que des simulations de Monte-Carlo mettent en évidence que lorsque la taille des effets aléatoires est trop grande par rapport au nombre d'observations, nous ne pouvons fournir une estimation précise pour l'héritabilité. Malheureusement, la taille typique des données que nous étudions vérifie justement la condition $N >> n$, par exemple $n = 2000$ et $N = 500000$.

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 est fondée sur plusieurs outils de sélection de variables :

• **Première étape : calcul des corrélations empiriques [2]**. Cette étape consiste à réduire le nombre de colonnes de Z qui sont pertinentes dans notre étude en essayant d'éliminer celles qui sont associées aux composantes nulles de u. La matrice Z restreinte aux colonnes les plus corrélées au phénotype observé est notée Z_{red}.

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

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

où le paramètre λ est choisi selon la méthode "stability selection" [5]. Plus précisément, le vecteur d'observations Y est aléatoirement divisé en plusieurs sous-échantillons de taille $n/2$. Pour chaque sous-échantillon, la deuxième étape est réalisée. 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. Enfin, le choix du

seuil est réalisé grâce à des simulations : en effet, nous avons regardé les résultats obtenus pour l'estimation de $\hat{\eta}$ pour différents seuils qui varient de 0.2 à 0.8 (figure 1). On peut voir que pour les différentes valeurs de $\eta^\star$, le choix optimal du seuil est de 0.75, c'est donc le seuil que l'on utilisera pour cette gamme de paramètres n et N.

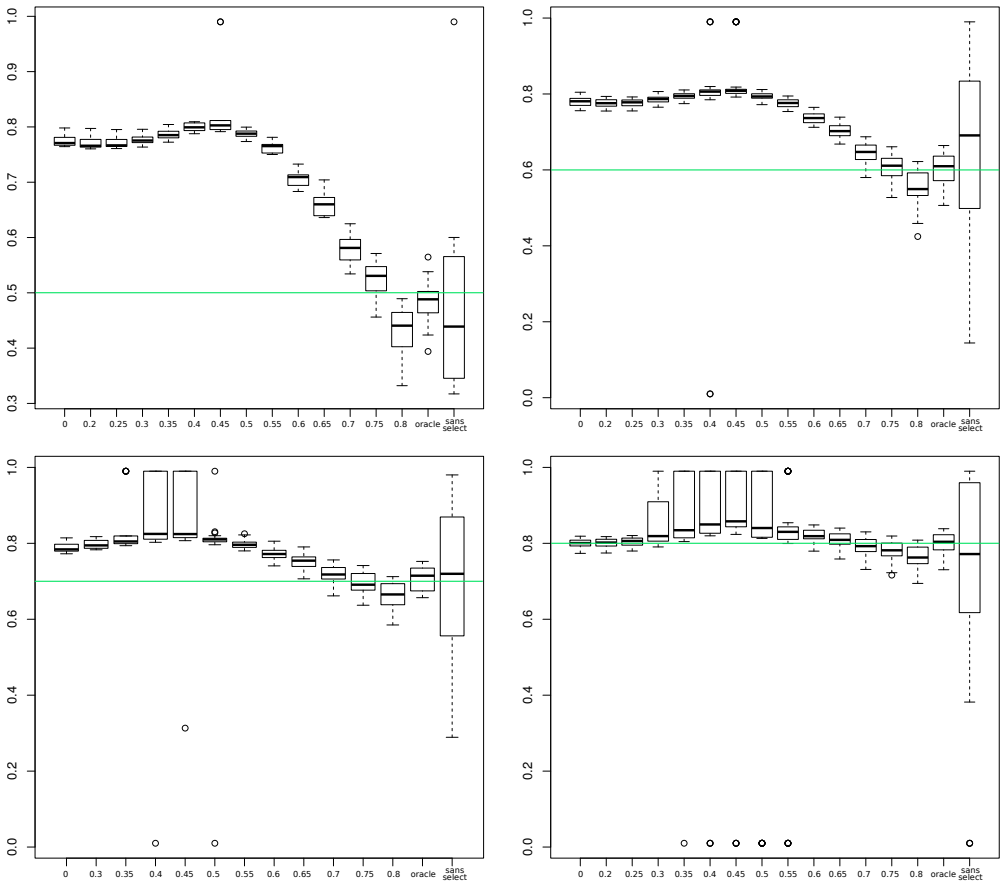

FIGURE 1 – Estimation de $\eta^\star$ pour différents choix de seuils de 0.2 à 0.8. Le premier boxplot correspond aux SNPs sélectionnés après la première étape, les deux derniers correspondent aux résultats d'un estimateur oracle connaissant les positions des composantes nulles de u et d'un estimateur sans sélection. Les simulations ont été réalisées pour $n = 2000$, $N = 10^5$, $q = 10^{-3}$ et diffférentes valeurs de $\eta^\star$: 0.5 (en haut à gauche), 0.6 (en haut à droite), 0.7 (en bas à gauche) and 0.8 (en bas à droite).

Notre méthode 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.

L'autre contribution de notre méthode est de fournir une liste de SNPs impliqués dans l'expression du phénotype étudié, correspondant aux indices des composantes non nulles de u.

5 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. Nous trouvons des résultats cohérents avec ceux trouvés par Toro et al [6] mais nous avons des intervalles de confiances plus petits.

Références

[1] Anna Bonnet, Elisabeth Gassiat, and Celine Levy-Leduc. Heritability estimation in high-dimensional sparse linear mixed models. *soumis*, 2014. http ://arxiv.org/abs/1404.3397.

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

[3] Brendan Maher. Personal genomes : The case of the missing heritability. *Nature*, 456(7218) :18–21, 2008.

[4] V.A. Marchenko and L.A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Math. USSR, Sb.*, 1 :457–483, 1968.

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

[6] Roberto Toro, Jean-Baptiste Pioline, and Guillaume Huguet. Genomic architecture of human neuroanatomical diversity. *Molecular Psychiatry)*, 2014.

[7] 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.