Improving heritability estimation by a variable selection approach in sparse high dimensional linear mixed models

Anna Bonnet

UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France.

E-mail: anna.bonnet@agroparistech.fr

Céline Lévy-Leduc

UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France.

Elisabeth Gassiat

Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, France.

Roberto Toro

Human Genetics and Cognitive Functions, Institut Pasteur, Paris, France.

Thomas Bourgeron

Human Genetics and Cognitive Functions, Institut Pasteur, Paris, France.

Summary. Motivated by applications in neuroanatomy, we propose a novel methodology to estimate heritability, which corresponds to the proportion of phenotypic variance that can be explained by genetic factors. Since the phenotypic variations may only be due to a small fraction of the available genetic information, we propose an estimator of heritability that can be used in sparse linear mixed models. Since the real genetic architecture is in general unknown in practice, our method allows the user to determine whether the genetic effects are very sparse: in that case, we propose a variable selection approach to recover the support of these genetic effects before estimating heritability. Otherwise, we use a classical maximum likelihood approach. We apply our method, implemented in the R package EstHer available on the CRAN, on neuroanatomical data from the project IMAGEN.

Keywords: Heritability, High Dimension, Linear Mixed Models, Variable Selection, Applications in neuroanatomy

1. Introduction

For many complex traits in human population, there exists 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). This gap has been called by Maher (2008) and Manolio et al. (2009) the "dark matter" of the genome or the "dark matter" of heritability. Various population studies have shown that up to 80% of the variability of neuroanatomical phenotypes such as the brain

volume could be explained by genetic factors, see for instance Stein et al. (2012). This result is very important since several psychiatric disorders are shown to be associated to neuroanatomical changes, for instance macrocephaly and autism Steen et al. (2006) or reduced hippocampus and schizophrenia Amaral et al. (2008). Estimating properly the impact of the genetic background on neuroanatomical changes is a crucial challenge in order to determine afterwards if this background can either be a risk factor or a protective factor from developing psychiatric disorders. The GWAS studies performed for instance by Stein et al. (2012) identified genetic variants involved in the neuroanatomical diversity, which contributes to understand the impact of genetic factors. However, in the course of these studies, it is shown that this approach only explains a small proportion of the phenotypic variance. In order to understand the nature of the genetic factors responsible for major variations of the brain volume, Toro et al. (2015) used linear mixed models (LMM) to consider the effects of all the common genetic diversity characterized by the Single Nucleotide Polymorphisms (SNPs). This approach had been suggested by Yang et al. (2011) to study the effects of the SNPs on the height variations. The model they considered is a LMM defined as follows:

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

where $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ is the vector of observations (phenotypes), $\mathbf{X}$ is a $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector containing the unknown linear effects of the predictors, $\mathbf{Z}$ is a $n \times N$ matrix - N being the number of SNPs - which contains the genetic information, $\mathbf{u}$ and $\mathbf{e}$ correspond to the random effects. More precisely, $\mathbf{Z}$ is a version of $\mathbf{W}$ with centered and normalized columns, where $\mathbf{W}$ is defined as follows: $W_{i,j} = 0$ (resp. 1, resp. 2) if the genotype of the ith individual at locus j is qq (resp. Qq, resp. QQ) where p_j denotes the frequency of the allele q at locus j. In (1), the vector $\mathbf{e}$ corresponds to the environment effects and the vector $\mathbf{u}$ corresponds to the genetic random effect, that is the j-th component of u is the effect of the j-th SNP on the phenotype. In the modeling of Yang et al. (2011), all the SNPs have an effect on the considered phenotype, that is

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

The covariance matrix of $\mathbf{Y}$ can thus be written as:

$$\mathrm{Var}(\mathbf{Y}) = N\sigma_u^{\star 2}\mathbf{R} + \sigma_e^{\star 2}\mathrm{Id}_{\mathbb{R}^n}\,, \text{ where } \mathbf{R} = \frac{\mathbf{Z}\mathbf{Z}'}{N}\,,$$

and the parameter $\eta^\star$ defined as

$$\eta^\star = \frac{N\sigma_u^{\star 2}}{N\sigma_u^{\star 2} + \sigma_e^{\star 2}} \tag{3}$$

is commonly called the heritability (Yang et al. (2011),Pirinen et al. (2013)), and corresponds to the proportion of phenotypic variance which is determined by all the SNPs.

Since all SNPs are not necessarily causal, it seems more realistic to extend the previous modeling by assuming that the genetic random effects can be sparse, that is only a proportion q of the components of $\mathbf{u}$ are non null:

$$u_i \overset{i.i.d.}{\sim} (1-q)\delta_0 + q\mathcal{N}(0, \sigma_u^{\star 2}), \text{ for all } 1 \leq i \leq N, \tag{4}$$

where q is in $(0, 1]$, and δ_0 is the point mass at 0. Then the definition of $\eta^\star$ has to be adjusted as follows:

$$\eta^\star = \frac{Nq\sigma_u^{\star 2}}{Nq\sigma_u^{\star 2} + \sigma_e^{\star 2}} \, . \tag{5}$$

It corresponds to the proportion of phenotypic variance which is due to a certain number of causal SNPs which are, obviously, unknown. Let us emphasize that, in most applications, the proportion q of causal SNPs is also unknown, and that it may happen that the scientist has no idea how small q is.

When $q = 1$, that is when considering the modeling (2), most proposed approaches to estimate the heritability derive from a likelihood methodology. We can quote for instance the REstricted Maximum Likelihood (REML) strategies, originally proposed by Patterson and Thompson (1971) and then developed in Searle et al. (1992). Several approximations of the REML algorithm have also been proposed, see for instance the software EMMA proposed by Pirinen et al. (2013) or the software GCTA (Yang et al. (2011),Yang et al. (2010)).

We proposed in Bonnet et al. (2015) another method based on a maximum likelihood strategy to estimate the heritability and implemented in the R package HiLMM. We proved in Bonnet et al. (2015) the following theoretical result: though the computation of the likelihood is based on the modeling assumption (2), the estimator is consistent (unbiased) under the less restrictive modeling assumption (4). We believe this consistency result remains true for the estimators produced using the algorithms REML, EMMA, GCTA. But we also proved that, when $q \neq 1$, the standard error is not the one computed by the softwares when $q = 1$ and may be very large. We obtained a theoretical formula for the asymptotic variance of the estimator (depending in particular on q) and conducted several numerical experiments to understand how this asymptotic variance gets larger depending on the various quantities, in particular with respect to q and the ratio n/N. We observed that this variance indeed gets larger when q gets smaller, so that the accuracy of the heritability estimator is slightly deteriorated when all SNPs are not causal. Thus, a first problem is to find a method able to produce an estimator with smaller standard error than those obtained using only likelihood strategies. Also, since this standard error depends on q, a second problem is to produce a confidence interval one could trust without knowing q.

The goal of this paper is to address both problems. The results we obtained in Bonnet et al. (2015) suggest the following. If we knew the set of causal SNPs, then, considering only this (small) subset in the genetic information matrix, we would obtain with HiLMM an estimator having a smaller standard error than when using all SNPs in the genetic information matrix. Thus, our new practical method contains a variable selection step.

Variable selection and signal detection in high dimensional linear models have been extensively studied in the past decade and there are many papers on this subject. Among them, we can quote Meinshausen and Bühlmann (2010) and Beinrucker et al. (2014) about variable selection and references therein. The case of high dimensional mixed models has received little attention. As far as variable selection methods in the random effects of LMM are concerned, we are only aware of the work of Fan and Li (2012) and Bondell et al. (2010). Let us mention that regarding the estimation of heritability with possible sparse effects, there is also the bayesian approach of Guan and Stephens (2011) and Zhou et al. (2013), which proposes an interesting estimator for the heritability but which is computationally very demanding. Notice that, in our framework, we are not far from the situation for which it is proved in Verzelen (2012) that the support cannot be fully recovered, which happens when $Nq\log(1/q) >> n$. The variable selection step we propose takes elements from both ultrahigh dimension methods (Fan and Lv (2008), Ji and Jin (2012), Meinshausen and Bühlmann (2010)) and classical variable selection techniques (Tibshirani (1996)).

The second step of our method is to apply HiLMM using the selected subset of causal SNPs produced by the first step. Finally, we propose a non parametric bootstrap procedure to get confidence intervals with prescribed coverage. The whole procedure requires only a few minutes of computation.

To conclude, we propose in this paper a very fast method to estimate the heritability and construct a confidence interval substantially smaller than without variable selection when the genetic effects are very sparse. We show indeed that this procedure is very efficient in very sparse scenarios but we also highlight that it can severely underestimate heritability when the number of causal genetic variants is high. Since the real genetic architecture is in general unknown in practice, we introduce an empirical criterion which allows the user to decide whether it is relevant to apply a variable selection based approach or not. Our method has also the advantage to return a list of SNPs possibly involved in the variations of a given quantitative feature. This set of SNPs can further be analyzed from a biological point of view.

The paper is organized as follows. Section 2 describes the data set which motivated our work. Section 3 provides the detailed description of the method, and Section 4 displays the results of the numerical study. They were obtained by using the R package EstHer that we developed and which is available from the Comprehensive R Archive Network (CRAN). The simulation results illustrate the performance of our method on simulations and show that it is very efficient from a statistical point of view. In Section 5, we provide an empirical criterion to help the user to decide whether it is relevant to apply a variable selection based approach or not. In Section 6, we propose a thorough comparison of our approach with other methods in terms of statistical and numerical performances. Finally, the results obtained on the brain data described in Section 2 can be found in Section 7. We also provide a discussion section at the end of the paper.

2. Description of the data

We worked on data sets provided by the European project IMAGEN (Schumann et al., 2010), which is a major study on mental health and risk taking behaviour in teenagers. The research program includes questionnaires, interviews, behaviour tests, neuroimaging of the brain and genetic analyses. We will focus here on the genetic information collected on approximately 2000 teenagers as well as measurements of the volume of several features: the intracranial brain volume (icv), the thalamus (th), the caudate nucleus (ca), the amygdala (amy), the globus pallidus (pa), the putamen (pu), the hippocampus (hip), the nucleus accubens (acc) and the total brain volume (bv). Figure 1, which comes from Toro et al. (2015), is a schematic representation of these different areas of the brain. The data set contains $n = 2087$ individuals and $N = 273926$ SNPs, as well as a set of fixed effects, which in our case are the age (between 12 and 17), the gender and the city of residency (London, Nottingham, Dublin, Dresden, Berlin, Hamburg, Mannheim and Paris).

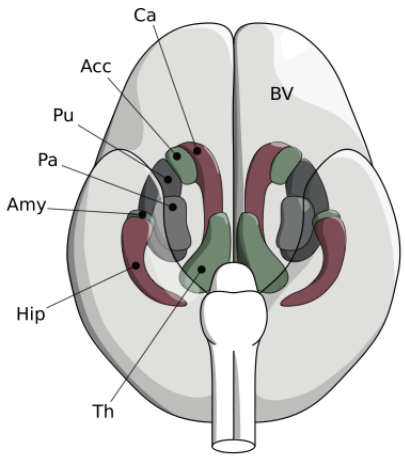


Fig. 1: Different regions of the brain (this figure is taken from Toro et al. (2015)).

In the following, our goal will thus be to provide a method to estimate the heritability of these neuroanatomical features.

3. Description of the method

The method that we propose can be split into two main parts: the first one consists in a variable selection approach and the second one provides an estimation of the heritability and the associated 95% confidence interval which is computed by using non parametric bootstrap.

At the beginning of this section we shall consider the case where there is no fixed effects, that is

$$\mathbf{Y} = \mathbf{Zu} + \mathbf{e} \tag{6}$$

but we explain at the end of this section how to deal with fixed effects. Let us first describe our variable selection method which consists of three steps.

3.1. Variable selection

Inspired by the ideas of Fan and Lv (2008), we do not directly apply a Lasso type approach since we are in an ultra-high dimension framework. Hence, we start our variable selection stage by the SIS (Sure Independence Screening) approach, as suggested by Fan and Lv (2008), in order to select the components of $\mathbf{u}$ which are the most correlated to the response $\mathbf{Y}$ and then we apply a Lasso criterion which depends on a regularization parameter λ. This regularization parameter is usually chosen by cross validation but here we decided to use the stability selection approach devised by Meinshausen and Bühlmann (2010) which provided better results in our framework.

Step 1: Empirical correlation computation

The first step consists in reducing the number of relevant columns of $\mathbf{Z}$ by trying to remove those associated to null components in the vector $\mathbf{u}$. For this, we use the SIS (Sure Independence Screening) approach proposed by Fan and Lv (2008) and improved by Ji and Jin (2012) in the ultra-high dimensional framework. More precisely, we compute for each column j of $\mathbf{Z}$:

$$C_j = \left| \sum Y_i Z_{i,j} \right|,$$

and we only keep the $N_{\max}$ columns of $\mathbf{Z}$ having the largest C_j. In practice, we choose the conservative value $N_{\max} = n$, inspired by the comments of Fan and Lv (2008) on the choice of $N_{\max}$.

In the sequel, we denote by $\mathbf{Z}_{\text{red}}$ the matrix containing these n relevant columns. This first step is essential for our method. Indeed, on the one hand, it substantially decreases the computational burden of our approach and on the other hand, it reduces the size of the data and thus makes classical variable selection tools efficient.

Step 2: LASSO criterion and stability selection

In order to refine the set of columns (or components of $\mathbf{u}$) selected in the first step and to remove the remaining null components in the vector $\mathbf{u}$, we apply a Lasso criterion originally devised by Tibshirani (1996) which has been used in many different contexts and has been thouroughly theoretically studied. It consists in minimizing with respect to u the following criterion:

$$\text{Crit}_\lambda(u) = \|\mathbf{Y} - \mathbf{Z}_{\text{red}} u\|_2^2 + \lambda \|u\|_1 \,, \tag{7}$$

which depends on the parameter λ and where $\|x\|_2^2 = \sum_{i=1}^p x_i^2$ and $\|x\|_1 = \sum_{i=1}^p |x_i|$ for $x = (x_1, \ldots, x_p)$. The choice of the regularization parameter λ is crucial since its value may strongly affect the selected variables set. Different approaches have been

proposed for choosing this parameter such as cross-validation which is implemented for instance in the `glmnet` R package. Here we shall use the following strategy based on the stability selection proposed by Meinshausen and Bühlmann (2010).

The vector of observations $\mathbf{Y}$ is randomly split into several subsamples of size $n/2$. For each subsample, we apply the LASSO criterion for a fixed parameter λ and the selected variables are stored. Then, for a given threshold, we keep in the final set of selected variables only the variables appearing a number of times larger than this threshold. In practice, we generated 50 subsamples of $\mathbf{Y}$ and we chose the parameter λ as the smallest value of the regularization path. As explained in Meinshausen and Bühlmann (2010), such a choice of λ ensures that some overfitting occurs and hence that the set of selected variables is large enough to include the true variables with high probability.

The matrix $\mathbf{Z}$ containing only the final set of selected columns will be denoted by $\mathbf{Z}_{\text{final}}$ in the following, where N_{final} denotes its number of columns.

The threshold has to be chosen carefully: keeping too many columns in $\mathbf{Z}_{\text{final}}$ could indeed lead to overestimating the heritability and, on the contrary, removing too many columns of $\mathbf{Z}$ could lead to underestimating the heritability. In the "small q" situations where it is relevant to use a variable selection approach a range of thresholds in which the heritability estimation is stable will appear as suggested by Meinshausen and Bühlmann (2010). In practice, we simulate observations $\mathbf{Y}$ satisfying (6), by using the matrix $\mathbf{Z}$, for different values of q and for different values $\eta^\star$ and we observe that this stability region for the threshold appear for small values of q. This procedure is further illustrated in Section 4.

3.2. Heritability estimation and confidence interval

3.2.1. Heritability estimation

For estimating the heritability, we used the approach that we proposed in Bonnet et al. (2015). It is based on a maximum likelihood strategy and was implemented in the R package HiLMM. Let us recall how this method works.

In the case where $q = 1$, which corresponds to the non sparse case,

$$\mathbf{Y} \sim \mathcal{N}\left(0, \eta^\star \sigma^{\star 2}\mathbf{R} + (1-\eta^\star)\sigma^{\star 2}\mathrm{Id}_{\mathbb{R}^n}\right),$$

with $\sigma^{\star 2} = N\sigma_u^{\star 2} + \sigma_e^{\star 2}$ and $\mathbf{R} = \mathbf{Z}_{\text{final}}\mathbf{Z}'_{\text{final}}/N_{\text{final}}$, where $\mathbf{Z}_{\text{final}}$ denotes the matrix $\mathbf{Z}$ in which the columns selected in the variable selection step described in Section 3.1 are kept.

Let $\mathbf{U}$ be defined as follows: $\mathbf{U}'\mathbf{U} = \mathbf{U}\mathbf{U}' = \mathrm{Id}_{\mathbb{R}^n}$ and $\mathbf{U}\mathbf{R}\mathbf{U}' = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, where the last quantity denotes the diagonal matrix having its diagonal entries equal to $\lambda_1, \ldots, \lambda_n$. Hence, in the case where $q = 1$,

$$\widetilde{\mathbf{Y}} = \mathbf{U}'\mathbf{Y} \sim \mathcal{N}(0, \Gamma)$$
$$\text{with } \Gamma = \mathrm{diag}(\eta^\star\sigma^{\star 2}\lambda_1 + (1-\eta^\star)\sigma^{\star 2}, \ldots, \eta^\star\sigma^{\star 2}\lambda_n + (1-\eta^\star)\sigma^{\star 2}), \quad (8)$$

where the λ_i's are the eigenvalues of $\mathbf{R}$.

We propose to define $\hat{\eta}$ as a maximizer of the log-likelihood

$$L_n(\eta) = -\log\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\widetilde{Y}_i^2}{\eta(\lambda_i-1)+1}\right) - \frac{1}{n}\sum_{i=1}^{n}\log\left(\eta(\lambda_i-1)+1\right) \ , \qquad (9)$$

where the $\widetilde{Y}_i$'s are the components of the vector $\widetilde{\mathbf{Y}} = \mathbf{U}'\mathbf{Y}$.
We now explain how to obtain accurate confidence intervals for the heritability by using a non parametric bootstrap approach.

3.2.2. Bootstrap confidence interval

For one vector of observations $\mathbf{Y}$, we derive a confidence interval for heritability by using the following procedure:

- Step 1: We estimate $\eta^\star$ and $\sigma^{\star 2}$ by using our approach described in the previous subsection. The corresponding estimators are denoted $\hat{\eta}$ and $\hat{\sigma}$.
- Step 2: We compute $\mathbf{Y}_{\text{new}} = \hat{\Gamma}^{-1/2}\widetilde{\mathbf{Y}}$, where $\widetilde{\mathbf{Y}}$ is defined in (8) and $\hat{\Gamma}$ has the same structure as Γ defined in (8) except that $\eta^\star$ and $\sigma^\star$ are replaced by their estimators $\hat{\eta}$ and $\hat{\sigma}$, respectively.
- Step 3: We create K vectors $(\mathbf{Y}_{\text{new},i})_{1\leq i\leq K}$ from $\mathbf{Y}_{\text{new}}$ by randomly choosing each of its components among those of $\mathbf{Y}_{\text{new}}$.
- Step 4: We then build K new vectors $(\widetilde{\mathbf{Y}}_{\text{samp},i})_{1\leq i\leq K}$ as follows: $\widetilde{\mathbf{Y}}_{\text{samp},i} = \hat{\Gamma}\mathbf{Y}_{\text{new},i}$. For each of them we estimate the heritability. We thus obtain a vector of heritability estimators $(\hat{\eta}_1, ..., \hat{\eta}_K)$.
- Step 5: For obtaining a 95% bootstrap confidence interval, we order these values of $\hat{\eta}_k$ and keep the ones corresponding to the $\lfloor 0.975 \times K \rfloor$ largest and the $\lfloor 0.025 \times K \rfloor$ smallest, where $\lfloor x \rfloor$ denotes the integer part of x. These values define the upper and lower bounds of the 95% bootstrap confidence interval for the heritability $\eta^\star$, respectively.

A bootstrap estimator of the variance can be obtained by computing the empirical variance estimator of the $\hat{\eta}_k$'s. In practice, we chose $K = 80$ replications.
Notice that this method should provide 95 % confidence intervals. We will verify on synthetic data if the actual probability for the true value $\eta^\star$ to be in the computed confidence interval is indeed greater than 95%. If not, we can choose a level α smaller than 0.05 to increase the coverage probability of our bootstrap method. This numerical guarantee of our method is addressed in Section 4.2.2. In the sequel, we will call 95% bootstrap confidence interval the interval computed with our bootstrap method with the level $\alpha = 0.05$.

In Step 2 of the previous algorithm, we should be in the non sparse case $q = 1$ thanks to the variable selection stage. Hence, the covariance matrix of $\mathbf{Y}_{\text{new}}$ should be close to identity.

Observe that our resampling technique is close to the one proposed by Abney (2015) for building permutation tests in linear mixed models.

3.3. Additional fixed effects

The method described above does not take into account the presence of fixed effects. For dealing with such effects we propose to use the following method, which mainly consists in projecting the observations onto the orthogonal of $\mathrm{Im}(\mathbf{X})$, the image of $\mathbf{X}$, to get rid of the fixed effects. In practice, instead of considering $\mathbf{Y}$ and $\mathbf{Z}$ we consider $\tilde{\mathbf{Y}} = \mathbf{A}'\mathbf{Y}$ and $\tilde{\mathbf{Z}} = \mathbf{A}'\mathbf{Z}$, where A is a $n \times (n-d)$ matrix (d being the rank of the fixed effects matrix), such that $\mathbf{A}\mathbf{A}' = \mathbf{P}_{\mathbf{X}}$, $\mathbf{A}'\mathbf{A} = \mathrm{Id}_{\mathbb{R}^{n-d}}$ and $\mathbf{P}_{\mathbf{X}} = \mathrm{Id}_{\mathbb{R}^n} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. This procedure was for instance used by Fan and Li (2012).

4. Numerical study

We present in this section the numerical results obtained with our approach which is implemented in the R package EstHer.

4.1. Simulation process

Since in genetic applications, the number n of individuals is very small with respect to the number N of SNPs, we chose $n = 2000$ and $N = 100000$ in our numerical study. We also set $\sigma_u^{\star 2} = 1$, we shall consider different values for q and we shall change the value of $\sigma_e^\star$ in order to have the following values for $\eta^\star$: 0.4, 0.5, 0.6 and 0.7. We generate a matrix $\mathbf{W}$ such that its columns W_j are independent binomial random variables of parameters n and p_j, where p_j is randomly chosen in $[0.1, 0.5]$. We compute $\mathbf{Z}$ by centering and empirically normalizing the matrix $\mathbf{W}$. The random effects are generated according to Equation (4) and then we compute a vector of observations such that $\mathbf{Y} = \mathbf{Z}\mathbf{u} + \mathbf{e}$.

We can make two important comments about the previous simulation process. Firstly, we generated a matrix $\mathbf{W}$ with independent columns, that is we assume that the SNPs are not correlated. Since this assumption may not be very realistic in practice, we provide in Section 4.2.5 some additional simulations where the generated matrix $\mathbf{W}$ has been replaced by the real matrix $\mathbf{W}$ coming from the IMAGEN project. Secondly, we did not include fixed effects but we show some results in Section 4.2.4 when fixed effects are taken into account.

4.2. Results in very sparse scenarios

In this section, we shall focus on the performances of our method in a very sparse scenario, that is 100 causal SNPs out of 100,000. We will describe all the results in terms of heritability estimation, support recovery and computational times in this particular case, then we will study other sparsity scenarios.

4.2.1. Choice of the threshold

In order to determine the threshold, we apply the procedure described in Section 3.1 and 3.2.1. Figure 2 displays the mean of the absolute value of the difference between $\eta^\star$ and the estimated value $\hat{\eta}$ for different thresholds and for different values of $\eta^\star$

obtained from 10 replications. We can see from this figure that in the case where the number of causal SNPs is relatively small: 100, that is $q = 10^{-3}$, our estimation procedure provides relevant estimations of the heritability for a range of thresholds around 0.75. Moreover, the optimal threshold leading to the smallest gap between $\hat{\eta}$ for different values of $\eta^\star$ is 0.76, so we will use this value in the following numerical study. Notice that our choice for the threshold calibration depends on the purpose of the variable selection. Indeed, if our goal was to recover the support and therefore to control the number of false positives, we would have chosen a threshold closer to 1. Here, having false positives is not our main concern, but heritability estimation is, that is why we calibrate the threshold to minimize the error we commit when estimating $\eta^\star$. The way of choosing the threshold and its impact on the heritability estimation will be further discussed in Sections 5 and 7.1.

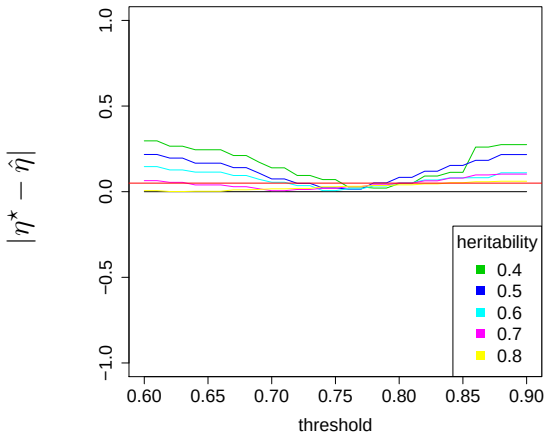


Fig. 2: Absolute difference between $\eta^\star$ and $\hat{\eta}$ for thresholds from 0.6 to 0.9 and for $q = 10^{-3}$ (100 causal SNPs). The results are obtained as a mean value for 10 simulations for each value of heritability and threshold.

4.2.2. Confidence intervals

We use the nonparametric bootstrap approach described in Section 3 with different levels α from 0.02 to 0.05 in order to compute the confidence intervals associated to the heritability estimations. For these different levels, we validate our procedure by comparing on the one hand the bootstrap confidence intervals and the empirical confidence intervals (Table 1) and on the other hand the coverage probabilities associated to the bootstrap approach (Table 3). The empirical confidence intervals are computed as follows: the different estimations of $\eta^\star$ obtained along the different replications are ordered, the $\lfloor (1-\alpha) \times M \rfloor$ largest and the $\lfloor \alpha \times M \rfloor$ smallest values correspond to the upper (resp. lower) bound of the $1-\alpha$ empirical confidence interval. Here, $\lfloor x \rfloor$ denotes the integer part of x and M is the number of replications. From Table 1,we can see that the empirical confidence intervals are generally

Table 1: Confidence intervals for $\hat{\eta}$ obtained empirically and by our Bootstrap method for different levels α from 0.02 to 0.05.

$\eta^\star$	0.4	0.5	0.6	0.7
Bootstrap $\alpha = 0.05$	[0.355 ; 0.505]	[0.420 ; 0.577]	[0.506 ; 0.659]	[0.600 ; 0.737]
Bootstrap $\alpha = 0.03$	[0.348 ; 0.516]	[0.413 ; 0.587]	[0.496 ; 0.666]	[0.593 ; 0.745]
Bootstrap $\alpha = 0.02$	[0.335 ; 0.536]	[0.400 ; 0.607]	[0.480 ; 0.681]	[0.579 ; 0.760]
Empirical	[0.373 ; 0.497]	[0.429 ; 0.568]	[0.501 ; 0.642]	[0.603 ; 0.730]

Table 2: Confidence intervals for $\hat{\eta}$ obtained by our approach with bootstrap ($\alpha = 0.05$), the oracle approach and the approach without selection ("without").

$\eta^\star$	0.4	0.5	0.6	0.7
EstHer	[0.353 ; 0.503]	[0.413 ; 0.565]	[0.494 ; 0.654]	[0.596 ; 0.738]
Oracle	[0.362 ; 0.472]	[0.414 ; 0.563]	[0.529 ; 0.670]	[0.619 ; 0.745]
without	[0.120 ; 0.880]	[0.102 ; 0.812]	[0.320 ; 0.938]	[0.349 ; 0.932]

included in the bootstrap intervals for all levels α from 0.02 to 0.05 (all scenarios except when $\eta^\star = 0.6$ and $\alpha = 0.05$). We also observe that the differences between the results with different levels are quite small, so that choosing $\alpha = 0.02$ increases slightly the length of the confidence intervals but it allows us to obtain coverage probabilities greater that 97 % (see Table 3) for all values of $\eta^\star$.

4.2.3. Comparison between the methods with and without selection

Our results are compared to those obtained if we do not perform the selection before the estimation, that is with the method implemented in HiLMM ("without"), but also with an approach which assumes the position of the non null components to be known (oracle). We will also compare our procedure to other existing methods that are widely used in genetic studies. This comparison is provided in Section 6.1. The results are displayed in Figure 3 and in Table 2. In this table, the confidence intervals displayed for the lines "Oracle" and "without" are obtained by using the asymptotic variance derived in Bonnet et al. (2015) which corresponds to the classical inverse of the Fisher information in the case $q = 1$. We observe that our method without the selection step provides similar results, that is almost no bias but a very large variance due to the framework $N \gg n$. Our method EstHer considerably reduces the variance compared to this method and exhibits performances close to those of the oracle approach which, contrary to our approach, knows the position of the non null components.

4.2.4. Additional fixed effects

We generated some synthetic data according to the process described in Section 4.1 but we added a matrix of fixed effects containing two colums. Figure 4 (a) displays

Table 3: Coverage probability obtained by our bootstrap method, with $\alpha = 0.05$, $\alpha = 0.03$ and $\alpha = 0.02$, obtained from 500 replications.

$\eta^\star$	0.4	0.5	0.6	0.7
$\alpha = 0.05$	0.90	0.96	0.93	0.87
$\alpha = 0.03$	0.95	0.99	0.95	0.90
$\alpha = 0.02$	0.97	0.99	0.98	0.97

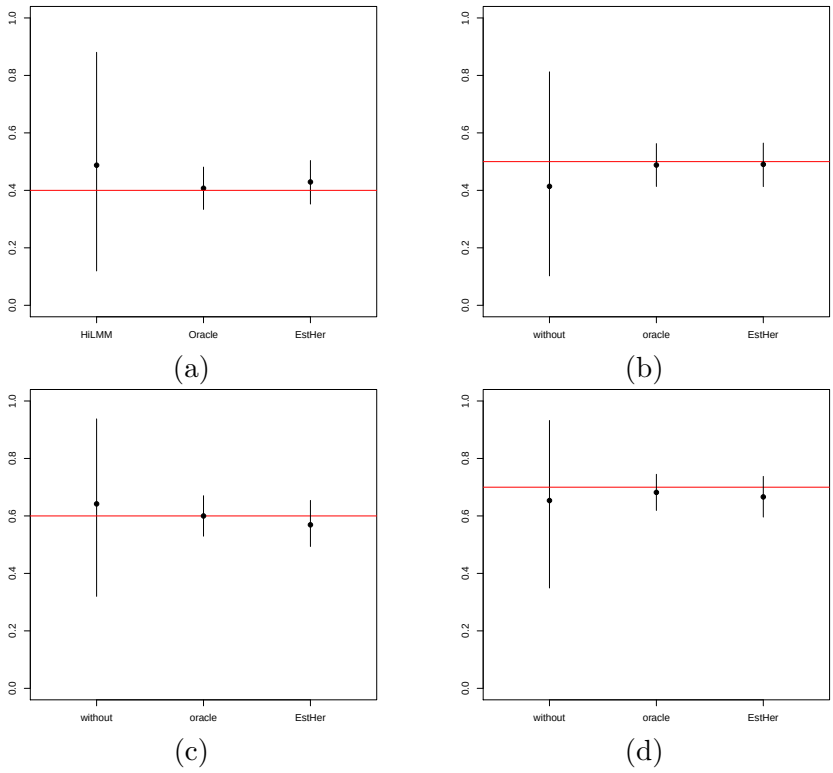
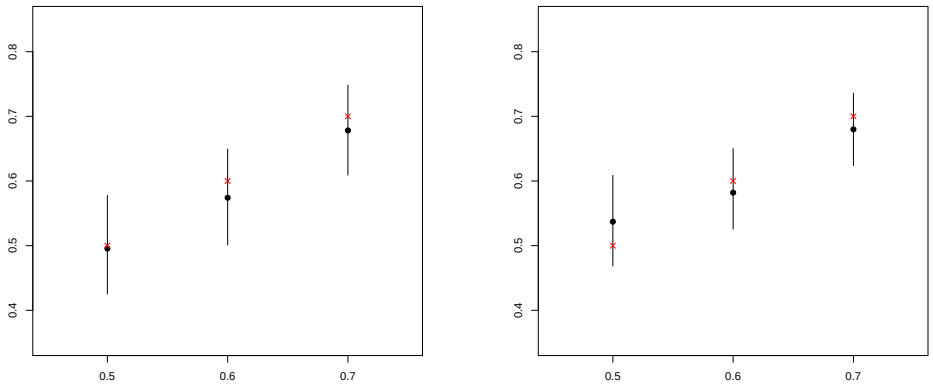


Fig. 3: Estimation of the heritability and the corresponding 95% confidence intervals when $q = 10^{-3}$, and for different values of $\eta^\star$: (a) $\eta^\star = 0.4$, (b) $\eta^\star = 0.5$, (c) $\eta^\star = 0.6$, (d) $\eta^\star = 0.7$. The means of the heritability estimators (displayed with black dots), the means of the lower and upper bounds of the 95% confidence intervals are obtained from 20 replicated data sets for the different methods: without selection ("without"), "oracle" which knows the position of the null components and EstHer. The horizontal gray line corresponds to the value of $\eta^\star$.

the corresponding results which show that the presence of fixed effects does not alter the heritability estimation.

4.2.5. Simulations with the matrix **W** *of the IMAGEN data set*
We conducted some additional simulations in order to see the impact of the linkage disequilibrium, that is the possible correlations between the columns of **Z**. Indeed, in the previous numerical study, we generated a matrix **W** with independent columns. The matrix **W** that we use now to generate the observations is the one from our genetic data set, except that we truncated it in order to have $n = 2000$ and $N = 100000$. The results of this additional study are presented in Figure 4 (b). We can see that they are similar to those obtained previously in Figure 3, which means that our method does not seem to be sensitive to the presence of correlation between the columns of **W**.



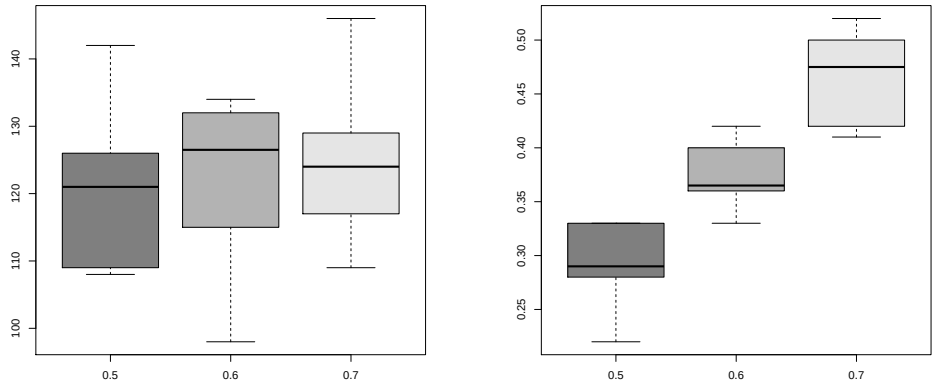
(a) (b)

Fig. 4: Estimated value of the heritability with 95 % confidence intervals. The results are displayed for several values of $\eta^\star$: 0.5, 0.6 and 0.7. (a) The data sets were generated including fixed effects. (b) The matrix **Z** used to generate data sets comes from the IMAGEN data. The black dots correspond to the mean of $\hat{\eta}$ over 10 replications and the crosses are the real value of $\eta^\star$.

4.2.6. Computational times
The implementation that we propose in the R package EstHer is very efficient since it only takes 45 seconds for estimating the heritability and 300 additional seconds to compute the associated 95% confidence interval. These results have been obtained with a computer having the following configuration: RAM 32 GB, CPU 4×2.3 GHz.

4.2.7. Recovering the support
When the number of causal SNPs is reasonably small, our variable selection method is efficient to estimate the heritability and we wonder if it is reliable as well to recover

the support of the random effects. In Figure 5, we see the proportion of support estimated by our method when there are 100 causal SNPs: our method selects around 130 components. We then focus on the proportion of the real support which has been captured by our method: we see that it may change according to $\eta^\star$. Indeed, the higher $\eta^\star$, the higher this proportion. Figure 5 then shows that EstHer is not designed to recover the support of the random effects. However, interestingly, even if we recover only a fraction of the support, we are able to estimate correctly the heritability. This is partly explained by the results displayed in Figure 6 where we can see that even in the worst case where we keep only 30% of the real non null components, we select the most active ones.



(a) (b)

Fig. 5: (a) Boxplots of the length of the set of selected variables with EstHer for 40 repetitions. The real number of non null components is 100. (b) Boxplots of the proportion of the real non null components captured in the set of selected variables.

The ability of recovering the support in linear models has been studied by Verzelen (2012) in ultra high dimensional cases. The author shows that with a non null probability, the support cannot be estimated under some numerical conditions on the parameters q, N and n (namely if there are considerably more variables N than observations n, and if the number of non null components qN is relatively high). In this simulation study, even when we consider small values of q (for instance $q = 10^{-3}$, that is 100 causal SNPs), we are not far from to the ultra high dimensional framework described in Verzelen (2012), which can explain the difficulties to recover the full support.

4.3. Results when the number of causal SNPs is high

In subsection 4.2 we show the performance of our method in the case where the proportion of causal SNPs q is small, that is around 10^{-3}. In this subsection, we focus on a more polygenic scenario, that includes the cases where thousands of SNPs

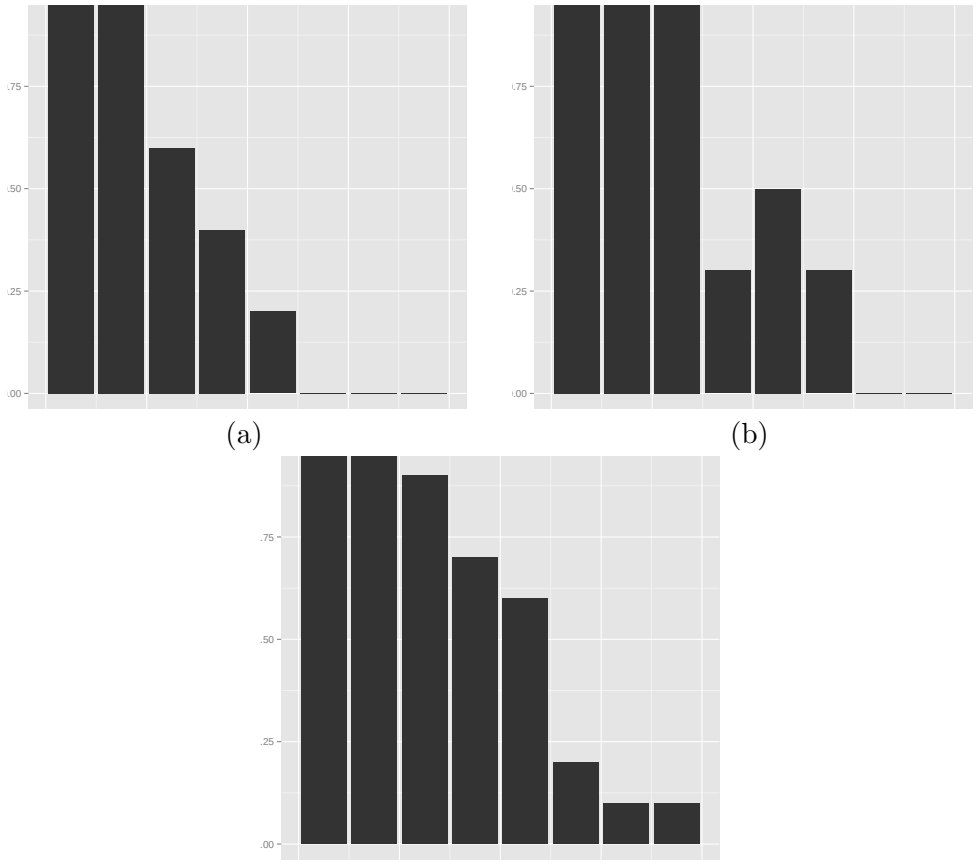


Fig. 6: Barplots of the proportion of components found by our method as function of the most efficient variables. For example, the first bar is the proportion of the 10 % higher components that we captured with our selection method. The histograms are displayed for several values of $\eta^\star$: 0.5 (a), 0.6 (b), 0.7 (c).

or ten of thousands of SNPs are causal.

4.3.1. Results when there are SNPs with moderate and weak effects

We first focus on the statistical performance of EstHer when there are a lot of SNPs (1000 or 10000) with small effects (for example, that explain 5% of the phenotypic variations), and a small number (around 100) with moderate effects. We can see from Figure 7 that, in this case, EstHer provides heritability estimates with tight confidence intervals which contain the true value of heritability .

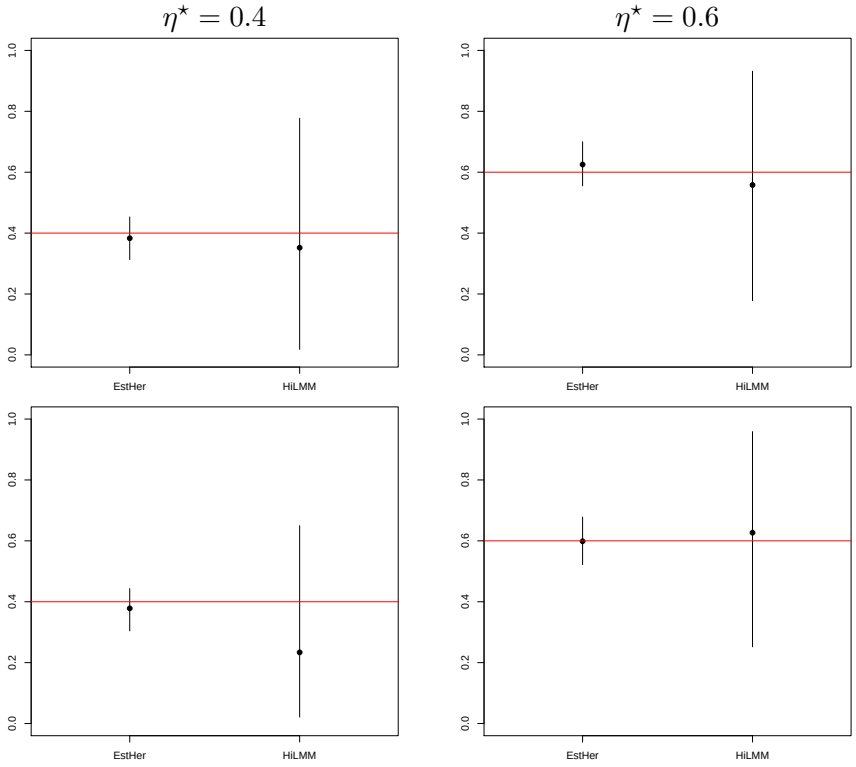


Fig. 7: Results of HiLMM and EstHer when there are a few causal SNPs with moderate effects and a lot of SNPs with small effects. The proportion of each is 100 out of 1000 (up) and 100 out of 10000 (bottom), with $\eta^\star = 0.4$ and 0.6.

4.3.2. Results when all SNPs have moderate effects

If all causal SNPs have moderate effects and if the number of these causal SNPs is high, namely greater than 1000, EstHer underestimates the heritability. These results are displayed in Figure 8. Moreover, we can see from Figure 9 that there is no threshold choice that can provide accurate estimations of heritability for all values of $\eta^\star$. This is a serious limitation to our variable selection approach: the number of causal SNPs can be up to 10000 for many complex traits. We will see how we can handle these very polygenic scenarios in Section 5.

5. A criterion to decide whether we should apply EstHer or HiLMM

In the previous numerical study, we distinguished two very different scenarios. First, when the number of causal SNPs is small (Section 4.2) or when a large fraction of heritability is explained by a small number of SNPs (Section 4.3.1), our variable selection approach EstHer is completely relevant before estimating heritability since it substantially reduces the length of the confidence intervals while not deteriorating the probability of the true heritability being in this confidence interval. Second, if

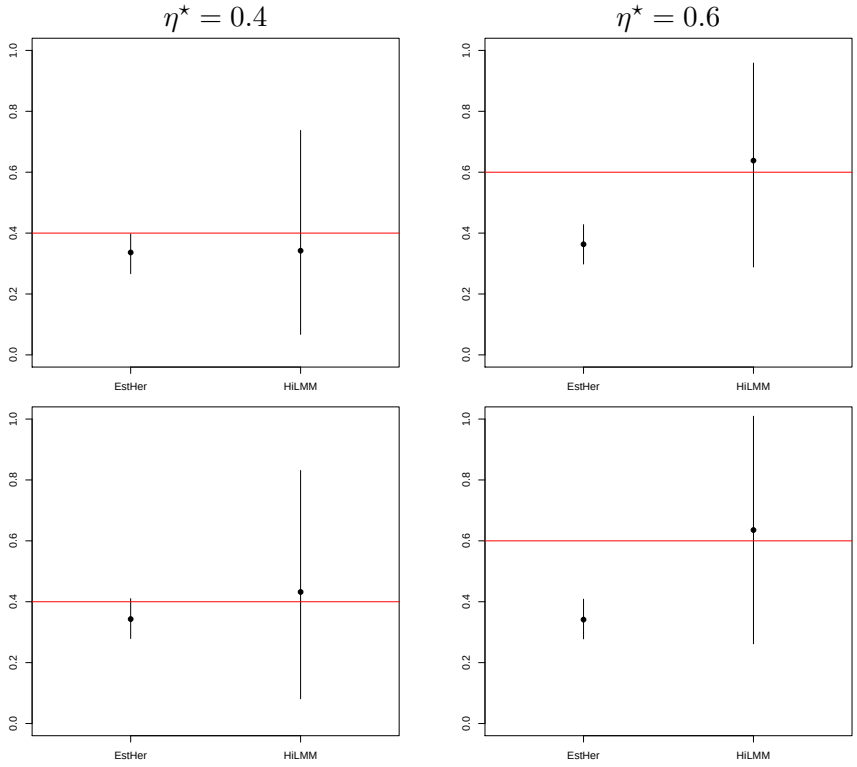


Fig. 8: Results of HiLMM and EstHer for 1000 (up) and 10000 (bottom) causal SNPs and for $\eta^\star = 0.4$ and 0.6.

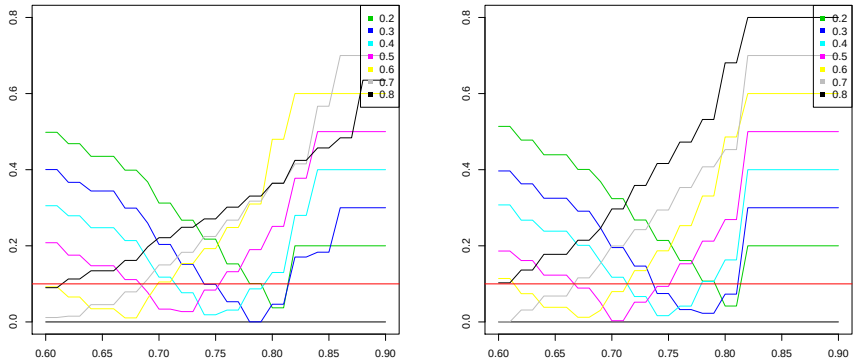


Fig. 9: Absolute difference $|\eta^\star - \hat{\eta}|$ for thresholds from 0.6 to 0.9 and for 1000 (left) and 10000 (right) causal SNPs.

there is a large number of causal SNPs, all with moderate effects, the selection approach introduces an important bias in the heritability estimates (Section 4.3.2).

These observations are similar to those made by Zhou et al. (2013), who built an

Table 4: Mean value of the number of overlapping confidence intervals for 16 thresholds from 0.7 to 0.85.

$\eta^\star$	100 causal SNPs	1000 causal SNPs	10000 causal SNPs
0.4	12.2	6.6	6.9
0.5	14.9	6.6	6.3
0.6	16	7.8	7.2

hybrid estimator able to deal with both sparse and non sparse scenario, to which we will compare our approach in Section 6.

Therefore, we propose hereafter a rule to decide whether it is better to apply our procedure with selection, EstHer or without selection, HiLMM. We can see from Figure 2 that when there are 100 causal SNPs, there is a large range of threshold values which provide an accurate estimation of $\eta^\star$, but when there are 1000 or 10000 causal SNPs, see Figure 9), the estimations are very different even for close thresholds. This observation gave us the idea of quantifying the stability of the estimations around the threshold that we determined as the optimal one. More precisely, for each threshold, we have an estimation of heritability with a 95% confidence interval, and we count the number of thresholds for which the confidence intervals overlap. Figure 10 confirms the stability around the best threshold for different values of $\eta^\star$ and Table 2 displays the number of ovelapping confidence intervals. We empirically determine the following criterion: if the mean number of thresholds is greater than 10 (over 16 tested thresholds), we apply EstHer, if not, we apply HiLMM.
Besides the ability of detecting the scenarios when the variable selection improves heritability estimation, this criterion has also the benefit of limiting the error that we might commit when choosing a threshold in the stability selection step. It guarantees indeed that we perform a variable selection only when the choice of the threshold has a small effect on heritability estimates. The results obtained by using this criterion are displayed in Figure 11.

6. Results after applying the decision criterion and comparison to other methods

6.1. Statistical performances

In this section we show the results obtained after applying the criterion described in Section 5 and we compare these results to existing methods. The most commonly used approach to estimate the heritability of complex traits is certainly GCTA (Yang et al., 2010), which is based on a restricted maximum likelihood maximization. Our method without selection, HiLMM, is, up to a few practical details, very close to GCTA, and has the same numerical performances, as we can see on Figure 5 from Bonnet et al. (2015), that is why we choose here to compare EstHer to HiLMM and not to GCTA. We also compare these results with the software GEMMA described

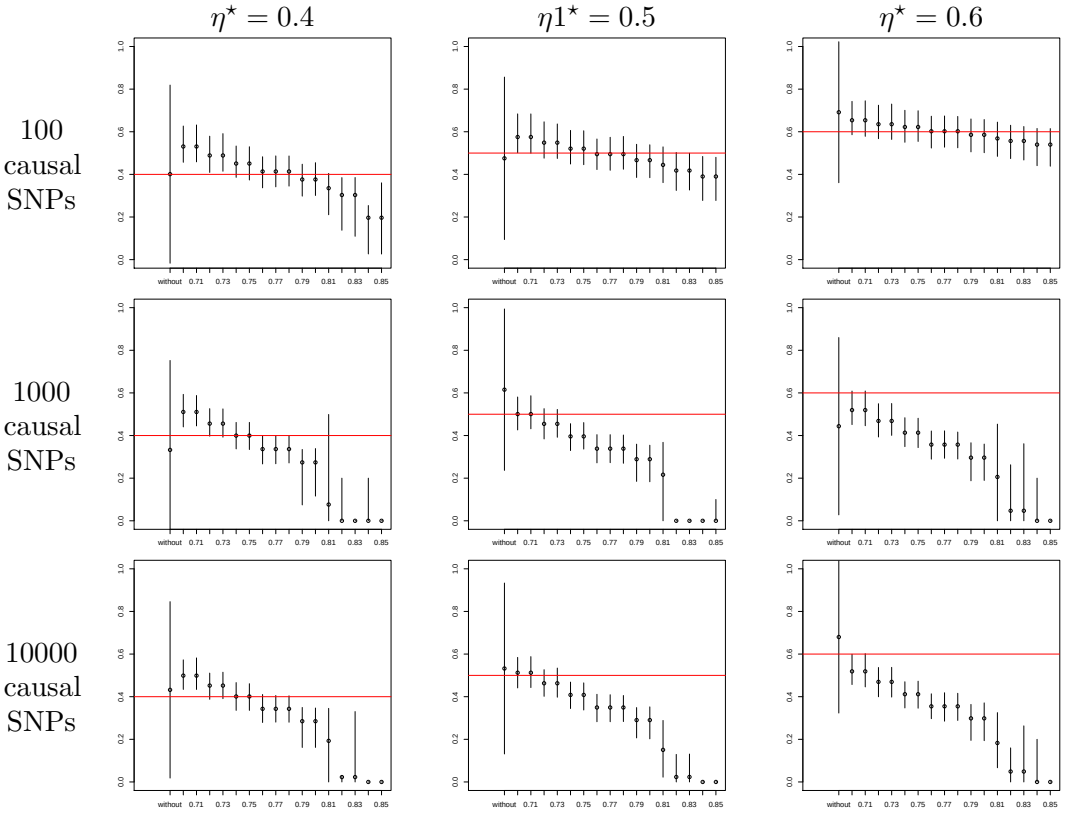


Fig. 10: Estimation of the heritability with 95% confidence intervals for $\eta^\star$ from 0.4 to 0.6 (from left to right), and from 100, 1000 and 10000 causal SNPs from top to bottom. Each graph shows the heritability estimations with 95% confidence intervals computed with HiLMM ("without") and for thresholds between 0.7 and 0.85.

in Zhou and Stephens (2012). GEMMA can fit both a non sparse linear mixed model (GEMMA-LMM) and a sparse linear mixed model if the BSLMM option is chosen denoted by BSLMM in the sequel. As explained in Zhou et al. (2013), BSLMM can deal with very sparse and also with very polygenic scenarios.

We can see from the bottom part of Figure 11 that, in very polygenic scenarios ($q = 0.1$, namely 10,000 causal SNPs), all the methods provide similar results: the four estimators are indeed empirically unbiased, but with a very large variance.

In sparse scenarios ($q = 10^{-3}$, namely 100 causal SNPs), we can see from the top part of Figure 11 that EstHer provides better results than HiLMM and GEMMA-LMM which exhibit similar statistical performances. In sparse scenarios, the variance of the BSLMM estimator is larger than the one provided by EstHer and smaller than the one provided by GEMMA-LMM and HiLMM. However, the performances of BSLMM could perhaps be improved by changing the MCMC parameters. Here, for computational time reasons, we used the default parameters that is 100,000 and

1,000,000 for the number of burn-in steps and the number of sampling, respectively.

Note that BSLMM averages the heritability estimates obtained for different prior distributions of the random effects, namely different values of sparsity q when we propose a binary criterion to decide if q is small enough to select variables or not. Although we are satisfied with the results of our current method, it could be interesting to inspire from BSLMM to develop a more "continuous" criterion to associate a non-binary weight on the estimations obtained with and without selection.

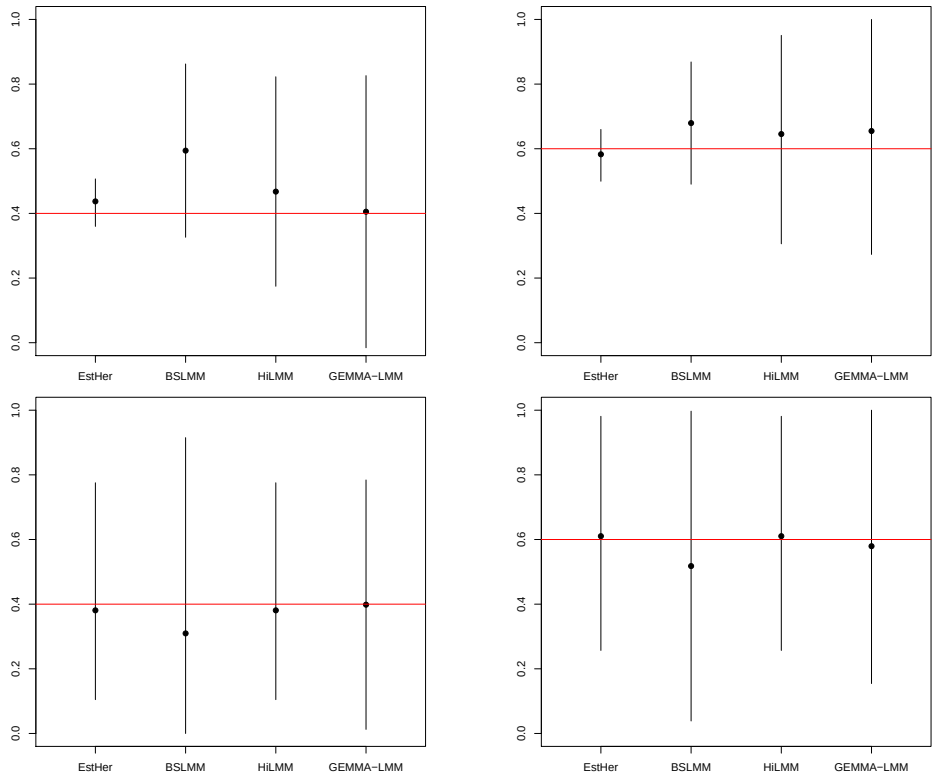


Fig. 11: Estimations of $\hat{\eta}$ with 95 % confidence intervals obtained using EstHer, BSLMM, HiLMM and GEMMA-LMM with 100 causal SNPs (top) and 10,000 causal SNPs (bottom). The results are obtained with 10 replications.

6.2. Computational times

The computational times in seconds for one estimation of the heritability with BSLMM and the heritability estimation for 16 thresholds as well as the associated confidence intervals with our method EstHer are displayed in Figure 12. We chose this number of thresholds since we applied the criterion defined in Section 5. It should be noticed that the computational times for EstHer could be reduced by diminishing the number of thresholds. For BSLMM we used the default parameters

for the number of burn-in steps and the number of sampling. We can see from this figure that the gap between EstHer and BSLMM is all the more important that N is large. Contrary to our approach, BSLMM seems to be very sensitive in terms of computational time to the value of N.

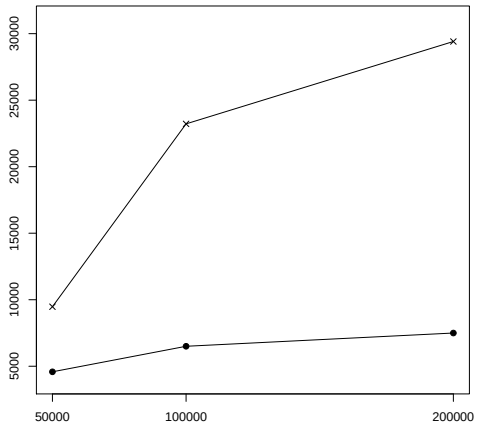


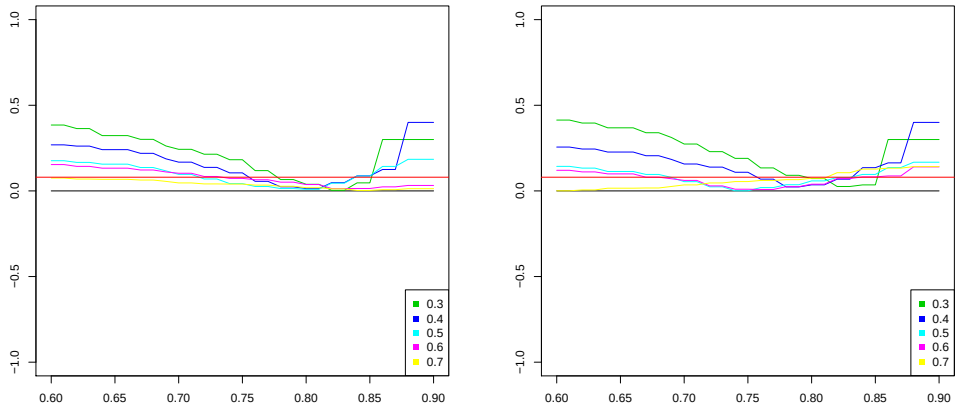
Fig. 12: Times (in seconds) to compute one heritability estimation with BSLMM (crosses) and EstHer (dots) by using 16 thresholds for $n = 2000$ and different values of N from $50,000$ to $200,000$.

7. Applications to genetic data

In this section, we applied our method to the neuroanatomic data coming from the Imagen project. In this data set, $n = 2087$ individuals and $N = 273926$ SNPs. For further details on this data set, we refer the reader to Section 2. This data has already been studied by Toro et al. (2015) to estimate heritability: it will obviously be a point of comparison for our analysis. Notice that they proposed to perform a principal component analysis to take into account a potential population structure in the data by including the first ten principal components as fixed effects. They showed that this procedure did not influence significantly the heritability estimates for this dataset.

7.1. Calibration of the threshold

We start by finding the threshold which is the most adapted to the IMAGEN data set. We use the same technique as the one described in Section 4.2.1: for several values of $\eta^\star$ and several thresholds, we display the absolute value of $\eta^\star - \hat{\eta}$, see Figure 13. The only difference with Section 4.2.1 is that we generated the observations by using the matrix $\mathbf{W}$ coming from the IMAGEN data set. According to Figure 13, we can find a reliable range of thresholds for estimating the heritability for all $\eta^\star$ from 0.4 to 0.7 when the number of causal SNPs is smaller than 100. This optimal threshold is equal to 0.79. We shall use this value in the sequel.



(a) (b)

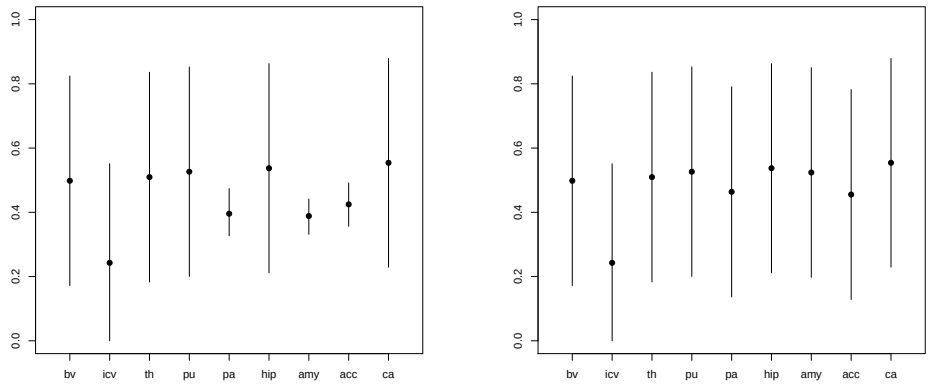
Fig. 13: Absolute value of the difference between $\eta^\star$ and $\hat{\eta}$ for thresholds from 0.6 to 0.9, and for different values of qN: (a) 50 causal SNPs, (b) 100 causal SNPs. Each difference has been computed as the mean of 10 replications.

Table 5: Mean value of the number of overlapping confidence intervals for 16 thresholds from 0.7 to 0.85.

Phenotype	Number of thresholds
Bv	7.19
Hip	7.5
Icv	7.37
Acc	9.94
Amy	9.88
Th	7.5
Ca	7.13
Pu	7.13
Pa	10.75

7.2. Application of the decision criterion

Since we determined in the previous section that the optimal threshold is 0.79, we apply EstHer for thresholds around this value, that is from 0.7 to 0.85. We then count the number of overlapping confidence intervals, as explained in Section 5. The results are displayed in Table 5. We observe from this table that the sensitivity to the choice of the threshold varies substantially from one phenotype to another. Hence, we choose to apply our EstHer approach to the most stable phenotypes with respect to our criterion, namely pa, amy and acc. For the other phenotypes we recommand to apply HiLMM or another similar approach such as GCTA or GEMMA-LMM.



(a) (b)

Fig. 14: (a) Heritability estimations of bv, icv, th, pu, pa, hip, amy, acc, and ca with 95% confidence intervals obtained using EstHer or HiLMM according to the outcome of our decision criterion. (b) Heritability estimations of bv, icv, th, pu, pa, hip, amy, acc and ca with 95% confidence intervals obtained using HiLMM.

7.3. Results

Figure 14 (a) shows the heritability estimation with 95 % confidence intervals for all phenotypes, using either EstHer or HiLMM according to the outcome of our decision criterion. Figure 14 (b) shows the results obtained by using HiLMM, namely without any variable selection step. We compare our results with the ones obtained by Toro et al. (2015) who estimated the heritability of the same phenotypes by using the software GCTA. On the one hand, we can see from Figure 14 that in the cases where EstHer is used the confidence intervals given by our methodology are substantially smaller and included in those provided by either HiLMM or Toro et al. (2015). On the other hand, when HiLMM is used our results are on a par with those obtained by Toro et al. (2015). Moreover, our approach provides a list of SNPs which may contribute to the variations of a given phenotype and which could be further analyzed from a biological point of view in order to identify new biological pathways.

8. Discussion

We show in this paper that the genetic architecture, that is the number of causal genetic variants and the intensity of their effects, has a strong impact on heritability estimation. Indeed, we compare approaches that include or not a variable selection step before estimating heritability and we show that the optimal method depends on the sparsity setting. More precisely, including variable selection reduces substantially the variance of the estimator when the random effects are very sparse but introduces a bias when the trait is actually very polygenic. However, this ge-

netic architecture is generally unknown in practice, which increases the difficulty of choosing an appropriate method to estimate heritability. A safe choice consists in always applying a maximum likelihood approach (GCTA, HiLMM...) which ensures that the heritability estimator is unbiased, but with a very large variance in a typical scenario where the number of SNPs is very large compared to the number of individuals. Therefore, we propose here a criterion to determine which of the two procedures (with or without selection) is more appropriate to the observations, in order to reduce the variance of the estimator when it is possible without introducing a bias. Besides its efficient statistical performance, we also propose a method with a very low computational burden, which makes its use attractive on very large data sets coming from quantitative genetics.
The choice of the optimal estimator for heritability in a linear model has been handled from a theoretical point of view in Verzelen and Gassiat (2017); it confirms that the estimator that achieves the minimax risk is an adaptive estimator that includes a variable selection approach only in very sparse scenarios. The theoretical calibration of the threshold in the stability selection step is a very difficult issue but would also be of great interest to understand the impact of genetic architecture on heritability estimation.

Acknowledgments

The authors would like to thank Nicolai Meinshausen and Nicolas Verzelen for fruitful discussions and the IMAGEN consortium for providing the data.

References

Abney, M. (2015). Permutation testing in the presence of polygenic variation. *Genetic Epidemiology 39*(4), 249–258.

Amaral, D. G., C. M. Schumann, and C. W. Nordahl (2008). Neuroanatomy of autism. *Trends in Neurosciences 31*(3), 137 – 145.

Beinrucker, A., U. Dogan, and G. Blanchard (2014). Extensions of Stability Selection using subsamples of observations and covariates. arXiv:1407.4916v1.

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

Bonnet, A., E. Gassiat, and C. Levy-Leduc (2015). Heritability estimation in high-dimensional sparse linear mixed models. *Electronic Journal of Statistics 9*(2), 2099–2129.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(5), 849–911.

Fan, Y. and R. Li (2012). Variable selection in mixed effects models. *Annals of Statistics 40*(4), 2043–2068.

Guan, Y. and M. Stephens (2011, 09). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics 5*(3), 1780–1815.

Ji, P. and J. Jin (2012). UPS delivers optimal phase diagram in high-dimensional variable selection. *Annals of Statistics 40*(1), 73–103.

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

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher (2009). Finding the missing heritability of complex diseases. *Nature 461*(7265), 747–753.

Meinshausen, N. and P. Bühlmann (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(4), 417–473.

Patterson, H. and R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika 58*, 545–554.

Pirinen, M., P. Donnelly, and C. C. A. Spencer (2013). Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *The Annals of Applied Statistics 7*(1), 369–390.

Schumann, G., E. Loth, T. Banaschewski, A. Barbot, G. Barker, C. Büchel, P. Conrod, J. Dalley, H. Flor, J. Gallinat, et al. (2010). The imagen study: reinforcement-related behaviour in normal brain function and psychopathology. *Molecular psychiatry 15*(12), 1128–1139.

Searle, S., G. Casella, and C. McCulloch (1992). *Variance Components*. Wiley Series in Probability and Statistics. Wiley.

Steen, R. G., C. Mull, R. McClure, R. M. Hamer, and J. A. Lieberman (2006). Brain volume in first-episode schizophrenia. *The British Journal of Psychiatry 188*(6), 510–518.

Stein, J. L., S. E. Medland, A. A. Vasquez, D. P. Hibar, R. E. Senstad, A. M. Winkler, R. Toro, K. Appel, R. Bartecek, O. Bergmann, M. Bernard, A. A. Brown, D. M. Cannon, M. M. Chakravarty, A. Christoforou, M. Domin, O. Grimm, M. Hollinshead, A. J. Holmes, G. Homuth, J.-J. Hottenga, C. Langan, L. M. Lopez, N. K. Hansell, K. S. Hwang, S. Kim, G. Laje, P. H. Lee, X. Liu, E. Loth, A. Lourdusamy, M. Mattingsdal, S. Mohnke, S. M. Maniega, K. Nho, A. C. Nugent, C. O'Brien, M. Papmeyer, B. Putz, A. Ramasamy, J. Rasmussen, M. Rijpkema, S. L. Risacher, J. C. Roddey, E. J. Rose, M. Ryten, L. Shen, E. Sprooten, E. Strengman, A. Teumer, D. Trabzuni, J. Turner, K. van Eijk, T. G. M. van

Erp, M.-J. van Tol, K. Wittfeld, C. Wolf, S. Woudstra, A. Aleman, S. Alhusaini, L. Almasy, E. B. Binder, D. G. Brohawn, R. M. Cantor, M. A. Carless, A. Corvin, M. Czisch, J. E. Curran, G. Davies, M. A. A. de Almeida, N. Delanty, C. Depondt, R. Duggirala, T. D. Dyer, S. Erk, J. Fagerness, P. T. Fox, N. B. Freimer, M. Gill, H. H. H. Goring, D. J. Hagler, D. Hoehn, F. Holsboer, M. Hoogman, N. Hosten, N. Jahanshad, M. P. Johnson, D. Kasperaviciute, J. W. Kent, P. Kochunov, J. L. Lancaster, S. M. Lawrie, D. C. Liewald, R. Mandl, M. Matarin, M. Mattheisen, E. Meisenzahl, I. Melle, E. K. Moses, T. W. Muhleisen, M. Nauck, M. M. Nothen, R. L. Olvera, M. Pandolfo, G. B. Pike, R. Puls, I. Reinvang, M. E. Renteria, M. Rietschel, J. L. Roffman, N. A. Royle, D. Rujescu, J. Savitz, H. G. Schnack, K. Schnell, N. Seiferth, C. Smith, V. M. Steen, M. C. Valdes Hernandez, M. Van den Heuvel, N. J. van der Wee, N. E. M. Van Haren, J. A. Veltman, H. Volzke, R. Walker, L. T. Westlye, C. D. Whelan, I. Agartz, D. I. Boomsma, G. L. Cavalleri, A. M. Dale, S. Djurovic, W. C. Drevets, P. Hagoort, J. Hall, A. Heinz, C. R. Jack, T. M. Foroud, S. Le Hellard, F. Macciardi, G. W. Montgomery, J. B. Poline, D. J. Porteous, S. M. Sisodiya, J. M. Starr, J. Sussmann, A. W. Toga, D. J. Veltman, H. Walter, M. W. Weiner, J. C. Bis, M. A. Ikram, A. V. Smith, V. Gudnason, C. Tzourio, M. W. Vernooij, L. J. Launer, C. DeCarli, and S. Seshadri (2012). Identification of common variants associated with human hippocampal and intracranial volumes. *Nat Genet 44*(5), 552–561.

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

Toro, R., J.-B. Poline, G. Huguet, E. Loth, V. Frouin, T. Banaschewski, G. J. Barker, A. Bokde, C. Büchel, F. Carvalho, P. Conrod, M. Fauth-Bühler, H. Flor, J. Gallinat, H. Garavan, P. Gowloan, A. Heinz, B. Ittermann, C. Lawrence, H. Lemaître, K. Mann, F. Nees, T. Paus, Z. Pausova, M. Rietschel, T. Robbins, M. Smolka, A. Ströhle, G. Schumann, and T. Bourgeron (2015). Genomic architecture of human neuroanatomical diversity. *Molecular Psychiatry 20*(8), 1011–1016.

Verzelen, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electronic Journal of Statistics 6*, 38–90.

Verzelen, N. and E. Gassiat (2017). Adaptive estimation of high-dimensional signal-to-noise ratios. http://arxiv.org/abs/1602.08006, Bernoulli, to appear.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher (2010). Common snps explain a large proportion of the heritability for human height. *Nature Genetics 42*(7), 565–569.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher (2011). GCTA: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics 88*(1), 76 – 82.

Zhou, X., P. Carbonetto, and M. Stephens (2013). Polygenic modeling with bayesian sparse linear mixed models. *PLoS genetics 9*(2), e1003264.

Zhou, X. and M. Stephens (2012). Genome-wide efficient mixed model analysis for association studies. *Nature Genetics 44*, 821–824.