Heritability Estimation in High-dimensional Mixed Models : Theory and Applications.

Anna Bonnet

Université Paris-Saclay

École doctorale de mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription : Université Paris-Sud

Établissement d'accueil : AgroParisTech

Laboratoire d'accueil : Mathématiques et informatique appliquées, UMR 518 INRA

THÈSE DE DOCTORAT ÈS MATHÉMATIQUES

Spécialité : Mathématiques appliquées

Anna BONNET

Heritability Estimation in High-dimensional Mixed Models: Theory and Applications

Date de soutenance : 5 décembre 2016

Après avis des rapporteurs : Lee Dicker (Rutgers University)
Nicolai Meinshausen (ETH Zurich)

Jury de soutenance :
Christophe Ambroise (Université Evry-Val d'Essonne) Invité
Thomas Bourgeron (Université Paris-Diderot) Examinateur
Lee Dicker (Rutgers University) Rapporteur
Elisabeth Gassiat (Université Paris-Sud) Directrice de thèse
Christophe Giraud (Université Paris-Sud) Président
Céline Lévy-Leduc (AgroParisTech) Co-directrice de thèse
Nicolas Verzelen (SupAgro) Examinateur

NNT : 2016SACLS498

Contents

Chapter 1

Introduction

Content

1.1 Biological Context

All biological traits are influenced by both genetic and environmental factors. Quantifying these two contributions for a particular trait is a fundamental and challenging question in biology. The concept of heritability refers to the part of the variability of an observed trait (or phenotype) which can be attributed to genetic causes. Several missconceptions regarding heritability are due to the use of the term in the common language, which differs from the technical definition in the genetic field. For instance, a frequent missconception would be that heritability is the proportion of a phenotype that is transmitted to the next generation. Firstly, genes are passed on from parents to offspring but phenotypes are not. Secondly, if half of the genetic effects are indeed transmitted from each parent, this particular half is specific to each offspring. Visscher et al. (2008) gathered these frequent questions and mistakes regarding heritability. The concept of heritability as it is used in the field of genetics is presented in the following section.

1.1.1 Definition of heritability

As elegantly explained by Visscher et al. (2008), we consider the simple modeling where a phenotype of interest is the result of genetic and environmental effects considered as independent:

$$\text{Phenotype (P)} = \text{Genotype (G)} + \text{Environment (E)}.$$

The variance of the observable phenotypes (σ_P^2) can then be expressed as a sum of unobserved underlying variances (σ_G^2 and σ_E^2):

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2.$$

Heritability (H^2) is defined as a ratio of variances and expresses the proportion of the phenotypic variance that can be attributed to genetic factors:

$$H^2 = \frac{\sigma_G^2}{\sigma_P^2}.$$

The genetic variability may be partitioned into variances coming from different sources, in particular the variance σ_A^2 of additive genetic effects. Such additive effects are characterized by the impact of single nucleotide polymorphisms (SNPs), which are DNA sequence differences at the positions of the genome where there exists considerable variability in the population. These positions are actually not frequent compared to the totality of the genome: the human genome is indeed composed of approximately 3 billions of base pairs, a very large fraction of which are identical for all humans. In the sequel, we will consider the "narrow sense heritability" which is the proportion of variability explained only by additive genetic effects, defined by

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2}.$$

Since the access to the genotype of thousands of individuals has been made possible by the spectacular decreased cost of DNA sequencing, the heritability of quantitative traits and pathologies has become widely studied. Yang et al. (2010) estimated for instance that around 45 % of human height was explained only by the most frequent SNPs.

1.1.2 Heritability in human genetics

We will here motivate the estimation of heritability for human traits. It is indeed a step toward the understanding of complex diseases, which have often multiple causes. We refer in particular to diseases which are not caused by a single affected gene but are nevertheless suspected to have a strong genetic component, probably split among different genes.
For instance, the causes of psychiatric disorders, such as autism or schizophrenia, remain vague. A genetic component has been suggested by the results of monozygotic and dizygotic twin

studies (monozygotic twins have identical genomes while dizygotic twins share around 50% of their genomes). These studies show that if one twin is affected by autistic disorders, the other one is also affected in 82 to 92 % of cases for monozygotic twins (Bailey et al. (1995)), or in 20% of cases for dizygotic twins (Hallmayer et al., 2011). Moreover, for a family which already has an autistic child, the risk of having another one is evaluated at 20 % against 1% in the general population. These studies describe autism as the psychiatric disease with the most important genetic component. However, the severity of the autistic traits (language and interaction disorders, intellectual disabilities...) can be very different for two patients with similar causes, for instance the same mutation. It would thus seem, as it is also the case for other genetic diseases, that the genetic background modulates the effect of a causal mutation and renders an individual more or less sensitive to developing autistic traits. Furthermore, all studies show that despite their identical genetic patrimony, the concordance of symptoms of monozygotic twins is never total, which confirms an epigenetic and/or environmental component. However, quantifying these different possible causes and potential interactions between them remains a challenging issue.

Determining a significant genetic component of a disease also constitutes a strong argument to refute some popular beliefs about causes of some illnesses. For instance, an important wave of anti-vaccine movement has been fueled by a presumed connection between the hepatitis B vaccine and multiple sclerosis. Similarly, the measles vaccine has been accused to cause autism (Uno et al., 2012). Although no link has ever been demonstrated (Poland & Jacobson, 2001), the consequences of the fact that many parents refuse to vaccinate their children remains a major public health issue. Indeed, a recent study (Uno et al. (2012)) showed that more than 25% of parents in the US refused to vaccinate their children against mortal diseases like measles. Regarding other proposed causes of autistic disorders, the "refrigerator mother theory" was developed by the psychiatrist Leo Kanner who claimed to observe a "genuine lack of maternal warmth" among his patients' mothers. Even though this theory has since been discarded, the mothers of autistic patients have suffered severe and unwarranted accusations for several decades.

1.1.3 Heritability in vegetal and animal genetics

In the field of vegetal and animal genetics, heritability estimation is the first step to the selection of traits of interest, generally related to the yield of a valuable resource. We can mention the examples of the optimization of the yield of milk, Visscher & Goddard (1995) or wheat Eid (2009). The goal in Eid (2009) is to determine strongly heritable traits related to the yield and then to obtain an optimal genotype. This genotype was even selected to be the most resistant to extreme environmental conditions like water deprivation, which is currently a fundamental issue.

If this kind of practice is generally accepted in animal genetics, it creates a controversy on possible consequences of heritability estimations of human traits. Several studies estimated the IQ heritability (Toro et al., 2015) and Davies et al. (2011) even announced that "Genome-wide association studies establish that human intelligence is highly heritable and polygenic". The controversy about IQ heritability is discussed in Visscher et al. (2008), who enumerates reasons for the polemic nature of this issue. These include the very controversial definition of IQ as a measure of intelligence as well as historical abuses related to eugenics. We will not further discuss this controversy here, we just mention it to illustrate a frequent issue when dealing with heritability of human features.

Having briefly argued for the general interest of estimating heritability, we will now present the statistical modeling used to provide these estimations.

1.2 Heritability estimations in high dimensional linear mixed models

1.2.1 State of the art

Linear Mixed Models (LMMs) have been widely used in several fields, especially in medicine and genetics. Yang et al. (2010) proposed to estimate the heritability of human height using a classical LMM defined as follows:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{1.1}$$

where $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)'$ is the vector of observations of a phenotype of interest, $\mathbf{X}$ is a $n \times p$ matrix of predictors (or fixed effects), $\boldsymbol{\beta}$ is a $p \times 1$ vector containing the unknown linear effects of the predictors, and $\mathbf{u}$ and $\mathbf{e}$ correspond to the Gaussian random effects with variances respectively equal to $\sigma_u^{\star 2}$ and $\sigma_e^{\star 2}$.
Moreover, $\mathbf{Z}$ is a $n \times N$ matrix which contains the genetic information. More precisely, the $Z_{i,j}$'s are normalized random variables in the following sense: they are defined from a matrix $\mathbf{W} = (W_{i,j})_{1 \leq i \leq n,\, 1 \leq j \leq N}$ by

$$Z_{i,j} = \frac{W_{i,j} - \overline{W}_j}{s_j},\ i = 1, \ldots, n,\ j = 1, \ldots, N\ , \tag{1.2}$$

where

$$\overline{W}_j = \frac{1}{n}\sum_{i=1}^{n} W_{i,j},\ s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(W_{i,j} - \overline{W}_j)^2,\ j = 1, \ldots, N\ . \tag{1.3}$$

In (1.2) and (1.3) the $W_{i,j}$'s are such that for each j in $\{1, \ldots, N\}$ the $(W_{i,j})_{1 \leq i \leq n}$ are independent and identically distributed random variables and such that the columns of $\mathbf{W}$ are independent. In genetic applications, the matrix $\mathbf{W}$ contains all the genetic information about all the individuals in the study.
With this definition the columns of $\mathbf{Z}$ are empirically centered and have an empirical variance equal to 1.

The LMM appears to be an intuitive modeling to describe the biological concept of heritability as a ratio of genetic and phenotypic variances. Yang et al. (2010) and Pirinen et al. (2013) proposed to estimate the parameter

$$\eta^\star = \frac{N\sigma_u^{\star 2}}{N\sigma_u^{\star 2} + \sigma_e^{\star 2}}, \tag{1.4}$$

commonly considered as the mathematical definition for heritability since it determines how the variance is shared between $\mathbf{u}$ and $\mathbf{e}$.

In Model (1.1), the log-likelihood conditionnaly to $\mathbf{Z}$ is given by:

$$L(\boldsymbol{\beta}, \sigma_u^2, \sigma_e^2) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|\mathbf{Z}\mathbf{Z}'\sigma_u^2 + \sigma_e^2 \mathrm{Id}_{\mathbb{R}^n}| - \frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Z}\mathbf{Z}'\sigma_u^2 + \sigma_e^2 \mathrm{Id}_{\mathbb{R}^n})^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}). \tag{1.5}$$

Searle et al. (1992) gathered plenty of optimization techniques to estimate the parameters $\boldsymbol{\beta}$, $\sigma_u^{2\star}$ and $\sigma_e^{2\star}$, among which we can quote Henderson equations or iterative methods like Fisher-Scoring and Newton-Raphson.
A natural idea to estimate heritability is to estimate the variance parameters $\sigma_u^{\star 2}$ and $\sigma_e^{\star 2}$ in order to obtain an estimator as the ratio:

$$N\hat{\sigma_u}^2/(N\hat{\sigma_u}^2 + \hat{\sigma_e}^2).$$

Pirinen et al. (2013) noticed that the model defined in (1.1) could be reparameterized with β, $\eta^\star$ and $\sigma^{\star 2} = N\sigma_u^{\star 2} + \sigma_e^{\star 2}$ as new parameters. More precisely,

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

where $\mathbf{R} = \mathbf{Z}\mathbf{Z}'/N$.
Let $\mathbf{U}$ be the orthogonal matrix ($\mathbf{U}'\mathbf{U} = \mathbf{U}\mathbf{U}' = \mathrm{Id}_{\mathbb{R}^n}$) such that $\mathbf{U}\mathbf{R}\mathbf{U}' = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix having its diagonal entries equal to $\lambda_1, \ldots, \lambda_n$. Hence, $\widetilde{\mathbf{Y}} = \mathbf{U}'\mathbf{Y}$ is a zero-mean Gaussian vector having a covariance matrix equal to $\mathrm{diag}(\eta^\star\sigma^{\star 2}\lambda_1 + (1-\eta^\star)\sigma^{\star 2}, \ldots, \eta^\star\sigma^{\star 2}\lambda_n + (1-\eta^\star)\sigma^{\star 2})$, where the λ_i's are the eigenvalues of $\mathbf{R}$. Let us also denote $\widetilde{\mathbf{X}} = \mathbf{U}'\mathbf{X}$. Finally they computed and maximized the log-likelihood:

$$L_n(\boldsymbol{\beta}, \sigma^2, \eta) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\log(\eta(\lambda_i - 1) + 1) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{(\widetilde{\mathbf{Y}}_i - \widetilde{\mathbf{X}}\beta)^2}{\eta(\lambda_i - 1) + 1} - \frac{n}{2}\log(2\pi), \tag{1.6}$$

where $\widetilde{\mathbf{Y}} = (\widetilde{\mathbf{Y}}_1, ..., \widetilde{\mathbf{Y}}_n)$.
The aforementioned approaches raise two main concerns: firstly, they all have been validated in the framework where N is fixed and n goes to infinity. Indeed, using classical results of the LMM, we can obtain properties of consistency and asymptotic normality for the maximum likelihood estimator of heritability. However, since in practice the number N of SNPs is widely greater than the number of individuals n, it would be more appropriate to validate these methods in the framework where n and N go to infinity, with n/N going to $a \in (0, +\infty)$.
Moreover, they all have been developed in a non sparse Gaussian framework, which would imply that all the available genetic information would impact the observed phenotype. This unlikely hypothesis has been discussed in particular by Jiang et al. (2014), who studied the potiential error caused by non impacting SNPs in the model when considering a maximum likelihood approach from both theoretical and numerical points of view.
We have only mentioned heritability estimation in linear mixed models, but there exist other ways to define and estimate heritability. Indeed, important theoretical results on heritability estimation have been proven in the framework where n and N go to infinity, with n/N going to $a \in (0, +\infty)$, in the linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1.7}$$

where the random component comes from the residual vector $\boldsymbol{\varepsilon}$ which is assumed to be a zero-mean Gaussian vector with variance σ_ϵ^2 and from the "SNP matrix" $\mathbf{X}$ which columns are assumed to be independent and identically distributed Gaussian variables. The heritability in this model is defined as the ratio

$$\eta^\star = \frac{||\boldsymbol{\beta}||_2^2}{\sigma_\epsilon^2 + ||\boldsymbol{\beta}||_2^2}. \tag{1.8}$$

An advantage of this model is that there is no assumption on the distribution of $\boldsymbol{\beta}$, in particular on its sparsity. However, strong assumptions are required on the stucture of the matrix $\mathbf{X}$. Several methods were proposed to estimate heritability in Model (1.7). Dicker (2014) proposed a method-of-moments estimator which is asymptotically normal when $n, N \to +\infty$ and $n/N \to a \in (0, +\infty)$. Janson et al. (2015) developed the Eigenprism procedure to build accurate confidence intervals for the heritability in finite sample size and also studied the asymptotic behavior of their estimator when $n, N \to +\infty$ and $n/N \to a \in (0, +\infty)$. Dicker & Erdogdu (2016) studied the properties of the maximum likelihood estimator and conducted a numerical comparison of the aforementioned methods which showed that the maximum likelihood estimator had a smaller empirical variance than the two others. Dicker & Erdogdu (2016) showed the consistency and the asymptotic normality of the maximum likelihood estimator and computed as well an explicit form of the asymptotic variance.
In the same model, Verzelen & Gassiat (2016) studied the optimality of different procedures depending on the sparsity. Indeed, Verzelen & Gassiat (2016) compared the performances of an approach with variable selection (Gauss-LASSO estimator) and without selection (dense estimator) in different sparsity regimes. They computed for each range of sparsity values the minimax risk and proposed an adaptive estimator which achieves the minimax risk in all sparsity regimes.

1.2.2 Contribution

Our first contribution was to propose an estimator for heritability in the context where n and N go to infinity, with n/N goes to $a \in (0, +\infty)$ and to establish its theoretical properties. This work is developed in Chapter 2 of this manuscript and has been published in the Electronic Journal of Statistics. We studied a model as the one defined in (1.1) except that we assumed that the random effects could be sparse, that is that only a proportion q of the components of $\mathbf{u}$ were non-zero:

$$u_i \overset{i.i.d.}{\sim} (1-q)\delta_0 + q\mathcal{N}(0, \sigma_u^{\star 2}) \text{ , for all } 1 \leq i \leq N \text{ and } \mathbf{e} \sim \mathcal{N}\left(0, \sigma_e^{\star 2}\mathrm{Id}_{\mathbb{R}^n}\right), \tag{1.9}$$

where $\mathrm{Id}_{\mathbb{R}^n}$ denotes the $n \times n$ identity matrix, q is in $(0, 1]$, and δ_0 is the point mass at 0.
Up to considering the projection of $\mathbf{Y}$ onto the orthogonal of the image of $\mathbf{X}$ and for notational simplicity, we studied the following model

$$\mathbf{Y} = \mathbf{Z}\mathbf{u} + \mathbf{e}\ . \tag{1.10}$$

Moreover, since in our case we are only interested in estimating $\eta^\star$, we plugged in L_n defined in (1.6) an estimator of $\sigma^{\star 2}$, that is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{\widetilde{Y}_i^{\,2}}{\eta(\lambda_i - 1) + 1}.$$

We implemented an estimator of $\eta^\star$ as the maximizer of this likelihood function depending only on parameter η:

$$L_n(\eta) = -\log\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\widetilde{Y}_i^2}{\eta(\lambda_i-1)+1}\right) - \frac{1}{n}\sum_{i=1}^{n}\log\left(\eta(\lambda_i-1)+1\right)\ , \tag{1.11}$$

We obtained two main results in the framework where n and N go to infinity, with n/N going to $a \in (0,+\infty)$: first, we proved that our estimator was $\sqrt{n}$-consistent despite the presence of null components in the random effects. This result was obtained under mild assumptions on the matrix $\mathbf{W}$ and for any unknown sparsity q.
Then we established a central limit theorem under the additional assumption that for all i and j, $\mathbf{Z}_{i,j}$ were Gaussian variables with zero mean and unit variance. We computed a closed-form expression for the asymptotic variance, given by

$$\tau^2(a,\eta^\star,q) = \frac{2}{\gamma^2(a,\eta^\star)} + 3\frac{a^2\eta^{\star 2}}{\gamma^4(a,\eta^\star)}\left(\frac{1}{q}-1\right)S(a,\eta^\star) \tag{1.12}$$

where

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

and

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

In the previous expression $\mathrm{d}\mu_a(\lambda)$ is the density of Marchenko-Pastur, which is the distribution of the eigenvalues of $\mathbf{ZZ}'/N$. This distribution obtained by Marchenko & Pastur (1968) was a key element to establish the proof of our results. We implemented this approach in the R package HiLMM, which is available on the CRAN.
We also conducted a simulation study with finite sample size corresponding to realistic practical studies. We showed that although the asymptotic variance defined in (1.12) was theoretically depending on the sparsity q, its influence was barely noticeable in practice. However, the asymptotic variance was shown to be very sensitive to the parameter $a = n/N$: more precisely, when the number of observations is very small compared to the size of the random effects (which is often the case in genetic studies), the variance of the heritability estimator increases substantially. This numerical result motivated the idea of developing a variable selection approach in order to reduce the size of the random effects and to improve the accuracy of heritability estimation.

1.3 Variable selection in the random effects of a high dimensional sparse linear mixed model

Motivated by the numerical performance of our estimator described in the previous section, it appeared to be a good idea to include a variable selection step in our method. The aim of this variable selection step is to recover the support of the random effects, which means in practice that we want to find the SNPs involved in the phenotypic variations. We would then consider only the matrix of SNPs reduced to these relevant SNPs and estimate the heritability with

smaller standard error than we would have obtained with the whole matrix of SNPs. Let us first present the existing methods and results regarding variable selection in the random effects of sparse linear mixed models.

1.3.1 State of the art

Although the case of linear mixed models has received less attention than the linear model, there exist several methods to perform variable selection in linear mixed models.

Several works focus on selecting variable in the fixed effects of sparse LMMs, as for instance Schelldorfer et al. (2011). For a complete review of these methods, we refer the reader to the work of Müller et al. (2013). Regarding selection in the random effects, we are only aware of the work of Fan & Li (2012) and Bondell et al. (2010). Bondell et al. (2010) proposed indeed a method to select jointly fixed effects and random effects based on a EM algorithm. Fan & Li (2012) proposed a penalized criterion with a particular penalty named SCAD (Smoothly Clipped Absolute Deviation) which combines $L1$ and $L2$ penalties. Both methods can be computationally very demanding in high dimension: on the one hand, the EM algorithm, on the other hand, the cross validation to choose the two regularization parameters.

Variable selection in such high dimensional frameworks as those we are interested in can be very tricky, as proven by Verzelen (2012) who studied the case of the random linear model defined in Equation (1.7). Verzelen (2012) indeed established that if the condition $Nq\log(1/q) >> n$ holds, namely when the number of causal SNPs (that is the number of non null components in the random effects) is larger than the number of individuals, the support cannot be fully recovered.

Regarding heritability estimation, the idea of introducing a variable selection step beforehand was already proposed by Guan & Stephens (2011) in a Bayesian framework. Guan & Stephens (2011) proposed indeed an approach, named BVSR (Bayesian Variable Selection Regression), that is very efficient to estimate heritability in a very sparse framework but which is biased when the number of causal SNPs is high. Zhou et al. (2013) then proposed a practical approach, called BSLMM (Bayesian Sparse Linear Mixed Model) defined as an hybrid estimator between BVSR and a classical maximum likelihood approach (without selection). This hybrid estimator behaves closely to BVSR in very sparse frameworks and like the maximum likelihood estimator (no selection) otherwise. These numerical observations of Zhou et al. (2013) are consistent with the theoretical grounds established by Verzelen & Gassiat (2016) in the linear model and described in Section 1.2.1.

1.3.2 Contribution

Methodology

We proposed a practical variable selection method to improve the accuracy of heritability estimation. This work has been submitted for publication and is contained in Chapter 3 of this manuscript. Our method is implemented in the R package EstHer available on the CRAN.

Our approach has two main features: firstly, it is very efficient from a statistical point of view since it provides confidence intervals considerably smaller than those obtained with methods without variable selection. Secondly, its very low computational burden makes it usable on very

large data sets coming from quantitative genetics. Our method can handle ultra high dimension scenarios by using as a first step the Sure Independence Screening developed by (Ji & Jin, 2012). Then we apply a LASSO criterion (Tibshirani, 1996) combined with the stability selection (Meinshausen & Bühlmann, 2010). We also propose a methodology to compute confidence intervals based on a non parametric bootstrap approach and validated on synthetic data. In the course of the numerical study, we observed similar conclusions to those obtained by Zhou et al. (2013) in the Bayesian framework: in very sparse scenarios (namely, less than 200 causal SNPs out of 100 000), the estimator which includes a variable selection step is unbiased and its variance is substantially smaller than the variance of the ML estimator. However, when the number of causal SNPs is high, the selection step is not efficient and the corresponding estimator can severely underestimate the heritability. We developed a criterion based on the data in order to have an idea of the sparsity regime and whether we should apply a variable selection technique or not. We developed a hybrid estimator able to adapt according to the sparsity and we showed on synthetic data that this procedure allows us to reduce substantially the confidence intervals of the heritability estimations compared to a classical maximum likelihood estimator in very sparse scenarios. Otherwise, if the number of causal SNPs is too high, our hybrid estimator behaves like the maximum likelihood estimator, which was expected after introducing the decision criterion we proposed. The benefit of our method compared to the Bayesian approach developed by Zhou et al. (2013) lies mainly in substantially smaller computational times than for MCMC procedures, and also we do not have to deal with the settings of the different parameters in order to ensure the convergence of the algorithms.

Applications in human neuroanatomy and in animal genetics

We applied this method to two different datasets.

The first one comes from the European project IMAGEN, which is a study on teenagers' mental health. We estimated the heritability of the brain volume and the volumes of the different subcortical regions. Six phenotypes out of nine did not pass the criterion so we can suspect that a large number of SNPs are involved in their variations, and we obtained similar results to those obtained with a classical maximum likelihood approach, such as those obtained by Toro et al. (2015) who studied the same data thanks to the software GCTA developed by Yang et al. (2011). However, for the other three phenotypes, we obtained heritability estimations with very small standard errors as well as a list of potential causal SNPs, the relevance of which could be analyzed from a biological point of view. This application to neuroanatomical data is described after the description and the validation of our method on synthetic data in Chapter 3 of this manuscript.

The second application is the study of a trout species named *Salmo trutta*. This brown trout, which lives in fresh water, may or may not, during its life, decide to leave fresh water to migrate to the sea. This migration has a major impact in the trout conservation, and we aim to understand the reasons of this decision. It appears that growth during the freshwater phase could potentially be a critical factor determining the fate of individuals as brown trout (remained in fresh water) or sea trout; indeed, if a fish is growing fast in fresh water then there is no real need to go to the sea where survival rate is much lower. However, if it is struggling to grow in fresh water, then the benefit of going to sea and having better growth prospects might

compensate the higher predation risk. Hence we aim to investigate the proportion of genetics and environment effects in length variations, and if possible we would like to determine which SNPs and environmental variables are associated with growth patterns. The size of this data set was substantially smaller than in the previous application: we had indeed access to the length of 192 trouts, the genotype of which is described by 4069 SNPs. We noticed, according to the results of a numerical study, that our R package EstHer, which was dedicated to the analysis of very large datasets, was not efficient in this case. However, when removing the first step of our method, the Sure Independence Screening, which was specific to ultra high dimension frameworks, we obtained satisfactory results again. This application is described in Chapter 4.

1.4 Heritability estimation for binary traits

We are interested in the extension of the previous methods to the heritability estimation of a disease, where the observations are categorical (patient or control). We found in the literature different models used to define and estimate heritability for binary data.

1.4.1 Generalized Linear Mixed Model and Liability Model

An intuitive generalization of the previous work for estimating the heritability of a binary trait would be to consider the following Generalized Linear Mixed Model:

$$\mathbf{Y}_i \sim \mathcal{B}(q_i), \tag{1.13}$$

with $q_i = g(\mathbf{l}_i)$ where g is a link function and $\mathbf{l}_i$ is defined as

$$\mathbf{l} = \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{1.14}$$

with $\mathbf{u} \sim \mathcal{N}(0, \sigma_u^{\star 2})$ and $\mathbf{e} \sim \mathcal{N}(0, \sigma_e^{\star 2})$, as in the classical LMM defined in Section 1.2.
A classical choice of link function in the case of binary data is for instance

$$g(x) = \frac{\exp(x)}{1 + \exp(x)},$$

which ensures that $q_i \in (0, 1)$.
The heritability can then be defined at the liability scale, that is the heritability of the continuous variable $\mathbf{l}$ which is identical to the definition of heritability in the previous sections when considering a Gaussian phenotype:

$$\eta^\star = \frac{N\sigma_u^{\star 2}}{N\sigma_u^{\star 2} + \sigma_e^{\star 2}}. \tag{1.15}$$

Another modeling and definition for heritability of a binary trait was proposed by Falconer (1965), who assumed that the binary observations could be seen as an indicator function of a Gaussian variable exceeding a certain threshold t:

$$\mathbf{Y}_i = \mathbb{1}_{\{\mathbf{l}_i > t\}} \tag{1.16}$$

with $\mathbf{l}_i$ defined by the same expression (1.14) as in the previous model.

The unobserved Gaussian variable $\mathbf{l}$ is also called the liability in this modeling, which is usually called the "liability model" (Falconer (1965), Lee et al. (2011), Tenesa & Haley (2013)). The heritability is then also defined as the heritability at the liability scale as written in Equation (1.15).

1.4.2 Existing methods for heritability estimation in the liability model

In the literature specific to the heritability of binary phenotypes, we found methods for heritability estimation based on each of the previously described models. For the first modeling described in Equation (1.13), de Villemereuil et al. (2013) proposed to estimate the variance of the random effects $\sigma_u^{\star 2}$ by using MCMC methods developed by Hadfield (2010) and then to estimate the heritability as

$$\hat{\eta} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + 1 + 1},$$

where the first 1 in the denominator stands for the residual variance and the second 1 for the distribution-specific variance of a probit-link function (Nakagawa & Schielzeth, 2010). The residual variance is indeed set to 1 because the binary data do not provide enough information to infer both variances $\sigma_u^{\star 2}$ and $\sigma_e^{\star 2}$.
Since the expression of the likelihood is not possible to optimize directly, Breslow & Clayton (1993) proposed to maximize a penalized quasi-likelihood, using a Laplace approximation of the likelihood. This method has been shown to underestimate the variance parameters, for instance in the numerical comparative study performed by de Villemereuil et al. (2013).

Regarding the procedures based on the second modeling defined in Equations (1.16) and (1.14), Lee et al. (2011) proposed to use a maximum likelihood approach as if the binary traits were Gaussian, and then to apply a multiplicative factor to correct this approximation. Golan et al. (2014) showed that this heritability estimator was strongly biased in several realistic scenarios, in particular it was very sensitive to the prevalence of the disease (when the disease is rarer, the bias increases). The estimator also underestimates the heritability when the real heritability is high.
Weissbrod et al. (2015) presented a different methodology to estimate heritability also in the liability model. They proposed a maximum likelihood based strategy to rebuild the underlying liability before estimating the heritability. More precisely, they differentiated the likelihood with respect to $\mathbf{u}$ and $\mathbf{e}$ and maximized the obtained functions for chosen values of σ_u^2 and σ_e^2. It gave them an estimator $\hat{u}$ of $\mathbf{u}$ and $\hat{e}$ of $\mathbf{e}$ from which they obtained an estimator of the liability as $\hat{l} = \mathbf{Z}\hat{u} + \hat{e}$. Then they used this liability to provide another estimation of σ_u^2 and σ_e^2 and they repeated this procedure until convergence.
A key element of the method proposed by Golan et al. (2014) was to notice the particular shape of the data, which is the oversampling of cases in the case-control study. In a medical study, the number of patients is indeed similar to the number of controls even though the studied disease might be rare, which means that the proportion of cases in the study does not reflect the proportion of cases in the population. The method developed by Golan et al. (2014) takes

into account such oversampling of the cases and, as far as we are aware, it is the only method to do so. They proposed a method of moments to estimate heritability.

More precisely, Golan et al. (2014) considered a simplified version of Model (1.14), where the liability is given by

$$\mathbf{l} = \mathbf{g} + \mathbf{e},$$

where $\mathbf{g}$ is a genetic random effect, which can be correlated across individuals, and $\mathbf{e}$ is the environmental random effect, which is assumed to be independent of the genetic effect. Both effects are assumed to be Gaussian: $\mathbf{e}$ has a variance equal to $(1-\eta^\star)\mathrm{Id}_{\mathbb{R}^n}$ and $\mathbf{g}$ has a covariance matrix, the diagonal entries of which are equal to $\eta^\star$ and the non diagonal term (i,j) is equal to $\eta^\star \mathbf{G}_{i,j}$. For $1 \leq i \neq j \leq n$, the covariance matrix of $(\mathbf{l}_i, \mathbf{l}_j)$ is given by

$$\begin{pmatrix} 1 & \mathbf{G}_{i,j}\eta^\star \\ \mathbf{G}_{i,j}\eta^\star & 1 \end{pmatrix}.$$

They defined the variable

$$p_i = \frac{\mathbf{Y}_i - P}{\sqrt{P(1-P)}},$$

where P is the proportion of cases in the study and the event $\{S = 1\}$ happens if both individuals i and j are selected in the study.

The heritability estimator proposed by Golan et al. (2014) is a least square estimator obtained as the minimizer of

$$\sum_{i \neq j} \left(p_i p_j - \mathbb{E}(p_i p_j | S = 1)\right)^2.$$

Since $\mathbb{E}(p_i p_j | S = 1)$ has no explicit formula, Golan et al. (2014) suggested to take advantage of the fact that the correlations $\mathbf{G}_{i,j}$ were small and proposed an approximation based on Taylor developments around $\mathbf{G}_{i,j}$.

1.4.3 Contribution

Since the method of Golan et al. (2014) seemed very efficient according to their numerical results and since it is the only method we have seen which considered the specificity of the data in a case-control study, we decided to establish the theoretical properties of their estimator in the framework: n and N go to infinity, and n/N goes to a $\in (0, +\infty)$.
We considered the model defined in Equations (1.16) and (1.14), and we assumed that $\mathbf{Z}$ was a random matrix with centered and normalized columns as defined in Section 1.2.
In this model, the covariance matrix of $(\mathbf{l}_i, \mathbf{l}_j)$ can be written as

$$\Sigma^{(N)} = \begin{pmatrix} 1 + \eta^\star(\mathbf{G}_N(i,i) - 1) & \eta^\star \mathbf{G}_N(i,j) \\ \eta^\star \mathbf{G}_N(i,j) & 1 + \eta^\star(\mathbf{G}_N(i,i) - 1), \end{pmatrix}$$

where for all $1 \leq i, j \leq n$,

$$\mathbf{G}_N(i,j) = \frac{1}{N}\sum_{k=1}^{N}\mathbf{Z}_{i,k}\mathbf{Z}_{j,k}. \tag{1.17}$$

The main idea is to notice that the quantities $\mathbf{G}_N(i,j)$, $\mathbf{G}_N(i,i)-1$ and $\mathbf{G}_N(j,j)-1$ are small, which implies that the matrix $\Sigma^{(N)}$ is close to identity.
Inspired by the method developed by Golan et al. (2014), we proposed a first and second order approximation of $\mathbb{E}[p_i p_j | Z, S = 1]$ based on Taylor developments in $\mathbf{G}_N(i,j)$, $\mathbf{G}_N(i,i)-1$ and $\mathbf{G}_N(j,j)-1$.
Despite differences in the models we considered, we found the same first order approximation than obtained by Golan et al. (2014) but we have several differences in the second order approximation.
First, we investigated the theoretical properties of the estimator found with the first order approximation: we showed indeed that it was consistent under very mild assumptions on the matrix of SNPs.
Then, we compared the numerical performances of the estimators obtained from the first and second order approximations, from both a statistical and computational point of view. We highlighted that the computation of the estimator obtained with the second order approximation was slower and did not improve the results compared to the first order approximation estimator. These results are contained in Chapter 5 of this manuscript.

Chapter 2

Heritability estimation in high dimensional sparse linear mixed models

The content of this chapter is contained in the article: *A. Bonnet, E. Gassiat, C. Lévy-Leduc, "Heritability estimation in high dimensional sparse linear mixed models", Electronic Journal of Statistics, 9(2):2099-2129, 2015.*
The method which is presented is implemented in the `HiLMM` R package, available on the CRAN.

Content

Abstract

Motivated by applications in genetic fields, we propose to estimate the heritability in high-dimensional sparse linear mixed models. The heritability determines how the variance is shared

between the different random components of a linear mixed model. The main novelty of our approach is to consider that the random effects can be sparse, that is may contain null components, but we do not know either their proportion or their positions. The estimator that we consider is strongly inspired by the one proposed by Pirinen et al. (2013), and is based on a maximum likelihood approach. We also study the theoretical properties of our estimator, namely we establish that our estimator of the heritability is $\sqrt{n}$-consistent when both the number of observations n and the number of random effects N tend to infinity under mild assumptions. We also prove that our estimator of the heritability satisfies a central limit theorem which gives as a byproduct a confidence interval for the heritability. Some Monte-Carlo experiments are also conducted in order to show the finite sample performances of our estimator.

2.1 Introduction

Linear mixed models (LMMs) have been widely used in various fields such as agriculture, biology, medicine and genetics. In quantitative genetics, LMMs have been used for estimating the heritability of traits and breeding values as explained for instance by Lynch & Walsh (1998). In Genome Wide Association Studies (GWAS), which is the application field that inspired our work, Yang et al. (2011) suggested the use of linear mixed models to measure genotypes at a large number of single nucleotide polymorphisms (SNPs) in large samples of individuals in order to identify genetic variants that explain variations in phenotypes.

The model that we shall study in this paper is a LMM defined as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}\,, \tag{2.1}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ is the vector of observations, $\mathbf{X}$ is a $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector containing the unknown linear effects of the predictors, and $\mathbf{u}$ and $\mathbf{e}$ correspond to the random effects. Moreover, in (2.1), $\mathbf{Z}$ is a $n \times N$ random matrix which will be further described in Section 2.2.

We shall assume that the random effects can be sparse, that is only a proportion q of the components of $\mathbf{u}$ are non-zero:

$$u_i \overset{i.i.d.}{\sim} (1-q)\delta_0 + q\mathcal{N}(0, \sigma_u^{\star 2})\,, \text{for all } 1 \leq i \leq N \text{ and } \mathbf{e} \sim \mathcal{N}\left(0, \sigma_e^{\star 2}\mathrm{Id}_{\mathbb{R}^n}\right)\,, \tag{2.2}$$

where $\mathrm{Id}_{\mathbb{R}^n}$ denotes the $n \times n$ identity matrix, q is in $(0, 1]$, and δ_0 is the point mass at 0. Note that this corresponds to a more general situation than the usual assumption of (non-sparse) Gaussian random effects which is recovered when $q = 1$.

The use of linear mixed models to estimate heritability has been proposed by Yang et al. (2011) as an alternative to the regression models usually used in GWAS. The goal is to consider the joint effect of all SNPs on a phenotype, and the heritability corresponds to the proportion of phenotypic variance explained by all SNPs.

In the GWAS framework, $\mathbf{Z}$ is thus a matrix having a number of rows equal to the number of individuals in the experiment that is $n \approx 1000$ and a number of columns equal to the number of SNPs taken into account in the experiment, namely $N \approx 500,000$. This application motivated the framework that we chose where n and N tend to infinity.

The major difference between the framework of Yang et al. (2011) and ours is that they consider that the random effects are Gaussian while we consider a mixture model between a

point mass at 0 and a Gaussian distribution. With this modeling, we assume that all SNPs are not necessarily causal, that is that all SNPs do not explain a given phenotype.

Our main goal in this paper is to propose an estimator for the heritability in this possibly sparse framework and to establish its theoretical properties in the non standard theoretical context where n and N tend to infinity.

In this paper, we prove that using a strategy close to the one proposed by Pirinen et al. (2013), which has been devised in the case $q = 1$, provides consistent estimators even in the case where $q < 1$. Moreover, we prove that this estimator is $\sqrt{n}$-consistent in the following asymptotic framework: $n \to \infty$ and $N \to \infty$ such as $n/N \to a > 0$ and satisfies under mild assumptions a central limit theorem in both cases $q = 1$ and $q < 1$. It has to be noticed that the classical results that exist in linear mixed models are established only in the case where $q = 1$, n tends to infinity and N is constant.

The paper is organized as follows. Section 2.2 provides a detailed description of the model and the heritability estimator that we propose. Section 2.3 reviews existing methods for heritability estimation. Section 2.4 is dedicated to the theoretical properties of our estimator. The numerical results are presented in Section 2.5. They have been obtained thanks to the R package HiLMM that we have developed and which is available from the Comprehensive R Archive Network (CRAN). In Section 2.6, we provide some additional comments on our work as well as some prospects such as the estimation of the proportion q of non null components in the random effects. Finally, the proofs are given in Section 2.7.

2.2 Model and heritability estimator

2.2.1 Model

In the sequel, up to considering the projection of $\mathbf{Y}$ onto the orthogonal of the image of $\mathbf{X}$ and for notational simplicity, we shall focus on the following model

$$\mathbf{Y} = \mathbf{Z}\mathbf{u} + \mathbf{e}\,, \tag{2.3}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ is the vector of observations, $\mathbf{u}$ and $\mathbf{e}$ correspond to the random effects, which are defined in (2.2). Moreover, $\mathbf{Z}$ is a $n \times N$ random matrix such that the $Z_{i,j}$ are normalized random variables in the following sense: they are defined from a matrix $\mathbf{W} = (W_{i,j})_{1\leq i\leq n,\, 1\leq j\leq N}$ by

$$Z_{i,j} = \frac{W_{i,j} - \overline{W}_j}{s_j},\ i = 1, \ldots, n,\ j = 1, \ldots, N\,, \tag{2.4}$$

where

$$\overline{W}_j = \frac{1}{n}\sum_{i=1}^{n} W_{i,j},\ s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(W_{i,j} - \overline{W}_j)^2,\ j = 1, \ldots, N\,. \tag{2.5}$$

In (2.4) and (2.5) the $W_{i,j}$'s are such that for each j in $\{1, \ldots, N\}$ the $(W_{i,j})_{1\leq i\leq n}$ are independent and identically distributed random variables and such that the columns of $\mathbf{W}$ are independent. With this definition the columns of $\mathbf{Z}$ are empirically centered and normalized.

In genetic applications, the matrix $\mathbf{W}$ contains all the genetic information about all the individuals in the study. More precisely, for each j, the $(W_{i,j})_{1\leq i\leq n}$ are i.i.d binomial random

variables with parameters 2 and p_j. $W_{i,j} = 0$ (resp. 1, resp. 2) if the genotype of the ith individual at locus j is qq (resp. Qq, resp. QQ) where p_j is the frequency of Q allele at locus j.

In Model (2.1) with (2.4), (2.5), (2.2), one can observe that

$$\mathrm{Var}(\mathbf{Y}|\mathbf{Z}) = Nq\sigma_u^{\star 2}\mathbf{R} + \sigma_e^{\star 2}\mathrm{Id}_{\mathbb{R}^n} \ , \text{ where } \mathbf{R} = \frac{\mathbf{Z}\mathbf{Z}'}{N} \text{ and } q \text{ is defined in (2.2)} \ .$$

Inspired by Pirinen et al. (2013), Model (2.1) can be rewritten by using the following parameters:

$$\sigma^{\star 2} = Nq\sigma_u^{\star 2} + \sigma_e^{\star 2} \text{ and } \eta^\star = \frac{Nq\sigma_u^{\star 2}}{Nq\sigma_u^{\star 2} + \sigma_e^{\star 2}} \ . \tag{2.6}$$

Thus,

$$\mathrm{Var}(\mathbf{Y}|\mathbf{Z}) = \eta^\star\sigma^{\star 2}\mathbf{R} + (1-\eta^\star)\sigma^{\star 2}\mathrm{Id}_{\mathbb{R}^n} \ .$$

The parameter $\eta^\star$ which belongs to $[0,1]$ is commonly called the heritability in the case where $q = 1$, see for instance Yang et al. (2010), and determines how the variance is shared between $\mathbf{u}$ and $\mathbf{e}$ when all the components of $\mathbf{u}$ are non zero. We propose in (2.6) to extend this definition to the case where $\mathbf{u}$ may contain null components and q is in $(0,1]$. The parameter q actually corresponds to the proportion of non null components in $\mathbf{u}$ that is to the proportion of causal SNPs. Then, the heritability defined by $\eta^\star$ in (2.6) corresponds to the proportion of phenotypic variance explained by the causal variants.

2.2.2 Heritability estimator

In the case where $q = 1$, observe that

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

where $\eta^\star$ and $\sigma^\star$ are defined in (2.6).

Let $\mathbf{U}$ as the orthogonal matrix ($\mathbf{U}'\mathbf{U} = \mathbf{U}\mathbf{U}' = \mathrm{Id}_{\mathbb{R}^n}$) such that $\mathbf{U}\mathbf{R}\mathbf{U}' = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix having its diagonal entries equal to $\lambda_1, \ldots, \lambda_n$. Hence, in the case where $q = 1$ and conditionally to $\mathbf{Z}$, $\widetilde{\mathbf{Y}} = \mathbf{U}'\mathbf{Y}$ is a zero-mean Gaussian vector having a covariance matrix equal to $\mathrm{diag}(\eta^\star\sigma^{\star 2}\lambda_1 + (1-\eta^\star)\sigma^{\star 2}, \ldots, \eta^\star\sigma^{\star 2}\lambda_n + (1-\eta^\star)\sigma^{\star 2})$, where the λ_i's are the eigenvalues of $\mathbf{R}$.

The method proposed by Pirinen et al. (2013) consists in computing the log-likelihood

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

and to maximize this function of two variables by iterative optimization techniques. Since in our case we are only interested in estimating $\eta^\star$, we plugged an estimator of $\sigma^{\star 2}$ that is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \frac{\widetilde{Y}_i^{\,2}}{\eta(\lambda_i - 1) + 1}$$

in L_n. Thus, in the case $q = 1$, the maximum likelihood strategy would lead to estimate $\eta^\star$, assumed to be in $[0, 1-\delta]$ with $\delta > 0$, by $\hat{\eta}$ defined as a maximizer of

$$L_n(\eta) = -\log\left(\frac{1}{n}\sum_{i=1}^n \frac{\widetilde{Y}_i^2}{\eta(\lambda_i - 1) + 1}\right) - \frac{1}{n}\sum_{i=1}^n \log\left(\eta(\lambda_i - 1) + 1\right) \ , \tag{2.7}$$

where the $\widetilde{Y}_i$'s are the components of the vector $\widetilde{\mathbf{Y}} = \mathbf{U}'\mathbf{Y}$.

We shall establish in Theorem 2, which is proved in Section 2.7, that this strategy produces $\sqrt{n}$-consistent estimators of $\eta^\star$ in both cases: $q = 1$ and $q < 1$ and also that this estimator satisfies a central limit theorem which provides as a by-product confidence intervals for $\eta^\star$.

2.3 Existing methods for heritability estimation

Several approaches can be used for estimating the heritability in the case where $q = 1$ but to the best of our knowledge, no theoretical results concerning the estimation of the heritability or the estimation of $\sigma_u^\star$, $\sigma_e^\star$ have been established in the framework where both n and N tend to infinity. This is one of the contributions of our paper. Among these approaches, we can quote the REML (REstricted Maximum Likelihood) approach, originally proposed by Patterson & Thompson (1971) and described for instance in Searle et al. (1992), which consists in estimating first $\sigma_u^\star$ and $\sigma_e^\star$ and then to estimate $\eta^\star$ as the ratio $\hat{\eta} = N\hat{\sigma_u}^2/(N\hat{\sigma_u}^2 + \hat{\sigma_e}^2)$. However, this type of approach is based on iterative procedures that require expensive matrix operations. Hence, several approximations have been proposed such as the AI algorithm (Gilmour et al. (1995)) which is used for instance in the software GCTA (Genome-wide Complex Trait Analysis) described in Yang et al. (2011). Other approximations have also been proposed in the EMMA algorithm (Kang et al. (2008)). For further details on the different approximations that could be used we refer the reader to Pirinen et al. (2013). The latter paper proposes another methodology for estimating the heritability which consists in rewriting Model (2.1) with the parameters (2.6) and in using an eigenvalue decomposition of the matrix $\mathbf{R}$. Further details on their methodology are given hereafter. According to the numerical experiments conducted in Pirinen et al. (2013) their approach has the lowest computational burden among the available algorithms.

In the case of sparse high dimensional framework, most of the papers studied the case of linear models. Among them, we can quote: Meinshausen & Bühlmann (2010) and Beinrucker et al. (2014). The high dimensional linear mixed model where $\mathbf{u}$ is sparse, that is the case where $q < 1$, which is the most realistic case for the applications that we have in view, has received little attention. It has been studied according two directions: detection and estimation. Concerning the detection field in this framework, we are only aware of the work of Arias-Castro et al. (2011) in which a testing procedure for detecting a sparse vector in high dimensional linear sparse regression model is also proposed and compared with the one proposed by Ingster et al. (2010). As for the procedures dedicated to the heritability estimation, there exist, to the best of our knowledge, only three approaches: the approach of Yang et al. (2010) who propose to approximate the genetic correlation between every pairs of individuals across the set of causal SNPs by the genetic correlation across the set of all SNPs, the approach of Golan & Rosset (2011) who propose a methodology based on MCEM (Monte-Carlo expectation-maximization) developed by Wei & Tanner (1990) and the Bayesian approaches of Guan & Stephens (2011) and Zhou et al. (2013). However, as far as the estimation issue in the high dimensional linear mixed model is concerned, the authors of these papers did not establish the theoretical properties of their estimators in the framework where both n and N tend to infinity.

2.4 Theoretical results

Observe that another way of writing Model (2.3) with the parameters defined in (2.6) consists in writing

$$\mathbf{Y} = \frac{1}{\sqrt{N}}\mathbf{Z}\mathbf{t} + \sigma^\star\sqrt{1-\eta^\star}\boldsymbol{\varepsilon}\ , \tag{2.8}$$

where $\boldsymbol{\varepsilon}$ is a $n \times 1$ Gaussian vector having a covariance matrix equal to identity and $\mathbf{t} = (t_1, \ldots, t_N)'$ is a random vector such that

$$t_i = \frac{\sigma^\star\sqrt{\eta^\star}}{\sqrt{q}} w_i \pi_i\ ,$$

where the w_i's and the π_i's are independent, $\mathbf{w} = (w_1, \ldots, w_N)'$ is a Gaussian vector with a covariance matrix equal to identity and the π_i's are i.i.d Bernoulli random variables such that $\mathbb{P}(\pi_1 = 1) = q$.

The estimator $\hat{\eta}$ is defined as a maximizer of $L_n(\eta)$ for $\eta \in [0, 1-\delta]$ for some small $\delta > 0$, L_n being given in (2.7). We shall study the asymptotic properties of $\hat{\eta}$ as n and N tend to infinity in a comparable way, that is when $n/N \to a > 0$. To understand the asymptotic behavior of $\hat{\eta}$, we shall first prove its consistency, then use a Taylor expansion of the derivative of L_n around $\hat{\eta}$ in the usual way. The computations as can be seen in (2.7) involve empirical means of functions of the eigenvalues λ_i of $\mathbf{R} = \frac{\mathbf{Z}\mathbf{Z}'}{N}$. Using Theorem 1.1 of Bai & Zhou (2008), we shall prove the almost sure convergence of such empirical quantities under a weak assumption denoted by Assumption 1 as follows.

Assumption 1. Let $\mathbf{Z}$ and $\mathbf{W}$ be two matrices defined by (2.4) and (2.5). Recall that for each j in $\{1, \ldots, N\}$ the $(W_{i,j})_{1\leq i\leq n}$ are independent and identically distributed random variables and such that the columns of $\mathbf{W}$ are independent (but not necessarily identically distributed). Assume that the entries $W_{i,j}$ of $\mathbf{W}$ are uniformly bounded, and have variance uniformly lower bounded, that is: there exist $W_M < \infty$ and $\kappa > 0$ such that $0 \leq W_{i,j} \leq W_M$ and $\sigma_j^2 = \mathrm{Var}(W_{i,j}) \geq \kappa$, for all j.

The following lemma ensures that the result of Marchenko & Pastur (1968) which gives the empirical spectral distribution of sample covariance matrices $\mathbf{Z}\mathbf{Z}'/N$ holds even when the entries $Z_{i,j}$ of the matrix $\mathbf{Z}$ are not i.i.d. random variables but when $\mathbf{Z}$ is obtained by empirical standardization of a matrix $\mathbf{W}$ satisfying Assumption 1.

Lemma 1. Under Assumption 1, as $n, N \to \infty$ such that $n/N \to a > 0$, the empirical spectral distribution of $R_N = \mathbf{Z}\mathbf{Z}'/N$: $F^{R_N}(x) = n^{-1}\sum_{k=1}^n \mathbb{1}_{\{\lambda_k \leq x\}}$ tends almost surely to the Marchenko-Pastur distribution defined as the distribution function of μ_a where, for any measurable set A,

$$\mu_a(A) = \begin{cases} \left(1 - \frac{1}{a}\right)\mathbb{1}_{0\in A} + \nu_a(A) & \text{if } a > 1 \\ \nu_a(A) & \text{if } a \leq 1 \end{cases}$$

with

$$\mathrm{d}\nu_a(\lambda) = \frac{1}{2\pi}\frac{\sqrt{(a_+ - \lambda)(\lambda - a_-)}}{a\lambda}\mathbb{1}_{[a_-, a_+]}(x)\mathrm{d}x,\ a_\pm = (1 \pm \sqrt{a})^2\ . \tag{2.9}$$

In $F^{R_N}(x)$, the λ_k's denote the eigenvalues of R_N.

Our first main result is the $\sqrt{n}$-consistency of the estimator $\hat{\eta}$.

Theorem 1. *Let* $\mathbf{Y} = (Y_1, \dots, Y_n)'$ *satisfy Model (2.8) with* $\eta^\star > 0$ *and the entries* $W_{i,j}$ *of* $\mathbf{W}$ *satisfy Assumption 1. Then, for all* q *in* $(0, 1]$*, as* $n, N \to \infty$ *such that* $n/N \to a \in (0, 1]$*,*

$$\sqrt{n}(\hat{\eta} - \eta^\star) = O_P(1).$$

Such a result is a theoretical cornerstone to legitimate the use of an estimator. However, statistical inference has to be based on confidence sets. The next step is thus to find the asymptotic distribution of $\sqrt{n}(\hat{\eta} - \eta^\star)$. Define for any $\eta \in [0, 1]$ and $\lambda \geq 0$

$$g(\eta, \lambda) = \frac{\lambda - 1}{\eta(\lambda - 1) + 1} \, .$$

Define also

$$\gamma_n^2 = \left\{ \frac{1}{n} \sum_{i=1}^{n} g(\hat{\eta}, \lambda_i)^2 - \left(\frac{1}{n} \sum_{i=1}^{n} g(\hat{\eta}, \lambda_i) \right)^2 \right\}$$

and

$$\gamma^2(a, \eta^\star) = \left\{ \int g(\eta, \lambda)^2 \mathrm{d}\mu_a(\lambda) - \left(\int g(\eta, \lambda) \mathrm{d}\mu_a(\lambda) \right)^2 \right\} \, . \tag{2.10}$$

We are now ready to state our second main result about the asymptotic distribution of $\sqrt{n}(\hat{\eta} - \eta^\star)$. For general q, the result only holds when the entries of $\mathbf{Z}$, that is the random variables $Z_{i,j}$ are i.i.d. standard Gaussian. Indeed, as may be seen when computing the variances, we need to be able to find the asymptotic behavior of empirical means of functions of the eigenvalues together with the eigenvectors of the matrix $\mathbf{R} = \mathbf{Z}\mathbf{Z}'/N$.

Theorem 2. *Let* $\mathbf{Y} = (Y_1, \dots, Y_n)'$ *satisfy Model (2.8) with* $\eta^\star > 0$ *and assume that the random variables* $Z_{i,j}$ *are i.i.d.* $\mathcal{N}(0, 1)$*. Then for any* $q \in (0, 1]$*, as* $n, N \to \infty$ *such that* $n/N \to a > 0$*,*

$$\sqrt{n}(\hat{\eta} - \eta^\star)$$

converges in distribution to a centered Gaussian random variable with variance

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

where

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

In the case where $q = 1$, the result holds in the general situation described in Assumption 1, and allows us to propose confidence sets with precise asymptotic confidence level.

Theorem 3. *Let* $\mathbf{Y} = (Y_1, \dots, Y_n)'$ *satisfy Model (2.8) with* $q = 1$ *and with* $\eta^\star > 0$*. Assume also that the entries* $W_{i,j}$ *of* $\mathbf{W}$ *satisfy Assumption 1 then, as* $n, N \to \infty$ *such that* $n/N \to a > 0$*,*

$$\gamma_n \sqrt{\frac{n}{2}} \, (\hat{\eta} - \eta^\star)$$

converges in distribution to $\mathcal{N}(0, 1)$*.*

Let us now give some additional comments on the previous results. Firstly, it has to be noticed that none of the limiting variance depends on $\sigma^\star$. Secondly, Theorem 2 is proved here only in the case where the $Z_{i,j}$ are i.i.d. Gaussian. This is because we used several times that the matrix of eigenvectors of $\mathbf{Z}\mathbf{Z}'/N$ is independent of the eigenvalues, and uniformly distributed on the set of orthonormal matrices. We think that the result of Theorem 2 is also valid when the $Z_{i,j}$ are defined from the $W_{i,j}$ satisfying Assumption 1, as suggested by the numerical results obtained in Section 2.5. To prove it requires new results in an active research topic of the random matrix theory field. We can observe in the expression of $\tau^2(a, \eta^\star)$ given in Theorem 2 that the presence of q is counterbalanced by the presence of a^2. This will be confirmed by the results obtained in the numerical results given in Section 2.5. Finally, we can observe that $2/(n\gamma_n^2)$ corresponds to the usual inverse of the Fisher information associated to η. This result is classical in the case where N is fixed and n tends to infinity but did not exist in the framework where both n and N tend to infinity even if it was already used in biological applied papers for deriving standard errors and confidence intervals. Theorem 3 proves that this result still holds even in the case where both n and N tend to infinity.

To the best of our knowledge, the effect of the presence of null components in the random effects has never been taken into account for computing the asymptotic variance of an estimator of the heritability. This is the contribution of Theorem 2. This theorem shows that the asymptotic variance contains an additional term which increases its value in the case $q < 1$ with respect to the case $q = 1$. It is shown in Section 3.3 how the computation of the asymptotic variance can be altered if this additional term is neglected. In practical situations, computing the standard error given by Theorem 2 requires the knowledge of q which is in general unknown. However, if an estimation of q is available for any practical reasons, the result of Theorem 2 can be used for computing confidence intervals and standard errors, see Section 2.6 for further details.

2.5 Numerical experiments

In this section, we first explain how to implement our method and then we illustrate the theoretical results of Section 2.4 on finite sample size observations for both cases: $q = 1$ and $q < 1$. We also compare the results obtained with our approach to those obtained by the GCTA software described in Yang et al. (2010) and Yang et al. (2011) which is a reference in quantitative genetics.

2.5.1 Implementation

In order to obtain $\hat{\eta}$, we used a Newton-Raphson approach which is based on the following recursion: starting from an initial value $\eta^{(0)}$,

$$\eta^{(k+1)} = \eta^{(k)} - \frac{L_n'(\eta^{(k)})}{L_n''(\eta^{(k)})} \ , \ k \geq 1 \ ,$$

where L_n' and L_n'' denote the first and second derivatives of L_n defined in (2.7), respectively. The closed form expression of these quantities are given in (2.13) and (2.25), respectively. In practice, this approach converges after at most 20 iterations and is not very sensitive to the initialization, namely to the value of $\eta^{(0)}$. However, in particular cases, the value of the initialization can

have an influence on the estimation of $\eta^\star$. This is the case, for instance, when the real value $\eta^\star$ is close to 1. In these situations, our algorithm can provide an estimation bigger than 1 and we constrained our method to return a value equal to 0.99. Figure 2.1 shows the estimations obtained on 100 replications when $a = 0.1$ and $\eta^\star = 0.8$. From this figure, we can see that the estimation of $\eta^\star$ does not depend in general on the initialization, except in some cases. Moreover, the best choice for $\eta^{(0)}$ is not constant from one replication to another. In order to limit the effect of the initialization, our algorithm uses several values for $\eta^{(0)}$ and whenever the estimations differ, it keeps the estimation which is the farthest away from the boundaries.

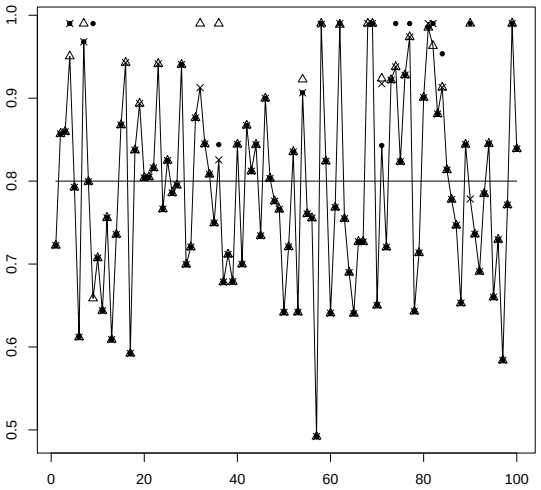


Figure 2.1 – Estimation of $\hat{\eta}$ obtained in the case $a = 0.1$ and $\eta^\star = 0.8$ for different values of initialization: $\eta^{(0)} = 0.1$ (dots), $\eta^{(0)} = 0.5$ (triangles) and $\eta^{(0)} = 0.9$ (crosses). The plain line displays the estimations obtained with our method to select the best initialization value and the x-axis is the replication number.

2.5.2 Results in Model (2.2) when $q = 1$

We shall first consider the performance of the estimator $\hat{\eta}$ when $q = 1$ for $\eta^\star$ in $\{0.3, 0.5, 0.7\}$, $n = 1000$, $\sigma_u^\star = 0.1$ and for a in $\{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1\}$, where $a = n/N$. We generated 500 data sets according to Model (2.1) using these parameters and $\mathbf{Z}$ as defined in (2.4) where the $W_{i,j}$ are binomial random variables with parameters 2 and p_j. In our experiments the p_j's are uniformly drawn in $[0.1, 0.5]$. The corresponding boxplots of $\hat{\eta}$ are displayed in Figure 2.2. We can see from this figure that our approach provides unbiased estimators of $\eta^\star$ and that the smaller the a the larger the empirical variance.

In order to illustrate the central limit theorem given in Theorem 3, we first display in Figure 2.3 the histograms of $\gamma_n (n/2)^{1/2} (\hat{\eta} - \eta^\star)$ along with the p.d.f of a standard Gaussian random variable for $\eta^\star = 0.5$ and different values of a. We can see that the Gaussian p.d.f fits well the data in all the considered cases. We also display in Figure 2.4 the values of $n^{-1/2}\sqrt{2\gamma_n^{-2}}$ and

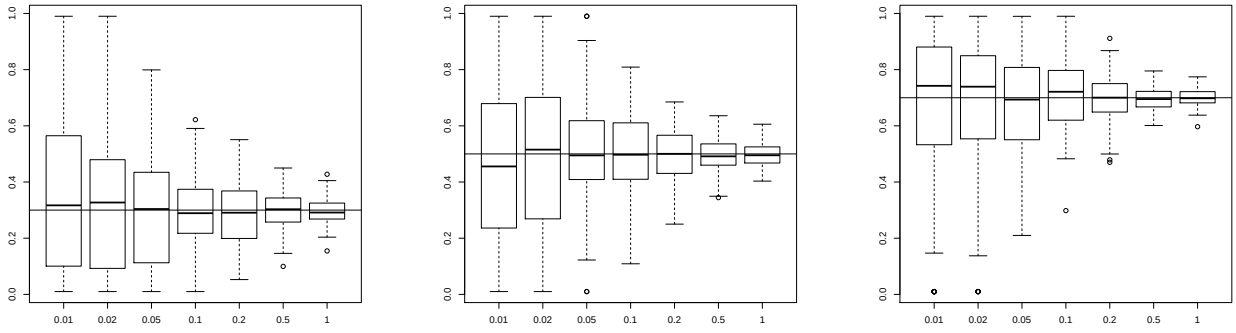


Figure 2.2 – Boxplots of $\hat{\eta}$ for different values of a, for $\eta^\star = 0.3$ (left), $\eta^\star = 0.5$ (middle) and $\eta^\star = 0.7$ (right). The horizontal line corresponds to the true value of $\eta^\star$. The whiskers of each boxplot correspond to the first and third quartiles.

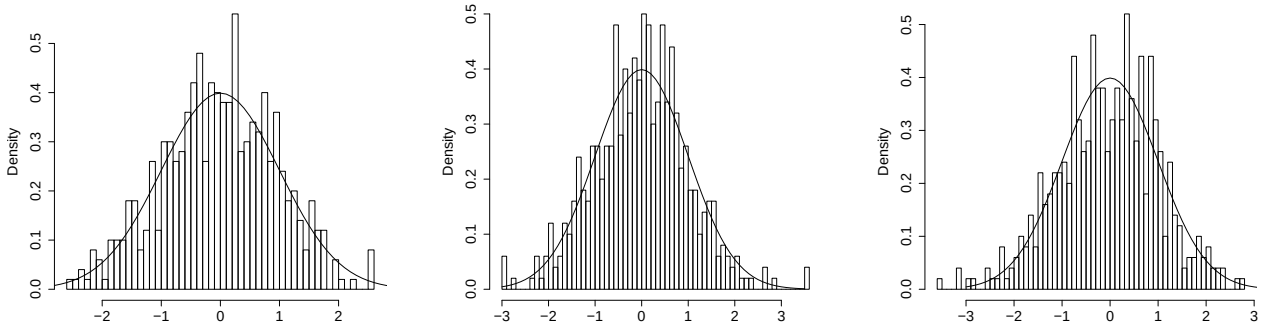


Figure 2.3 – Histograms of $\gamma_n (n/2)^{1/2} (\hat{\eta} - \eta^\star)$ for $\eta^\star = 0.5$ and $a = 0.05$ (left), $a = 0.1$ (middle), $a = 0.5$ (right) and the p.d.f of a standard Gaussian random variable in plain line.

the empirical standard deviation of $(\hat{\eta} - \eta^\star)$ averaged over all the experiments. As shown in Theorem 3, we also observe empirically that both quantities are very close.

In practice, the value of $\gamma_n^{-1}(n/2)^{-1/2}$ can be used for deriving confidence intervals for $\eta^\star$. As we can see from Figure 2.4, our approach leads to very accurate confidence intervals for a larger than 0.1 even in finite sample size cases.

Let us now compare our results with those obtained with the software GCTA. As we can see from Figure 2.5 which displays the boxplots of $\hat{\eta}$ for different values of a when $\eta^\star = 0.7$, the results found by our approach and GCTA are very close. In both cases, we observe that when a is close to 1 the estimations of $\eta^\star$ are very accurate but when a is small the standard error becomes very high.

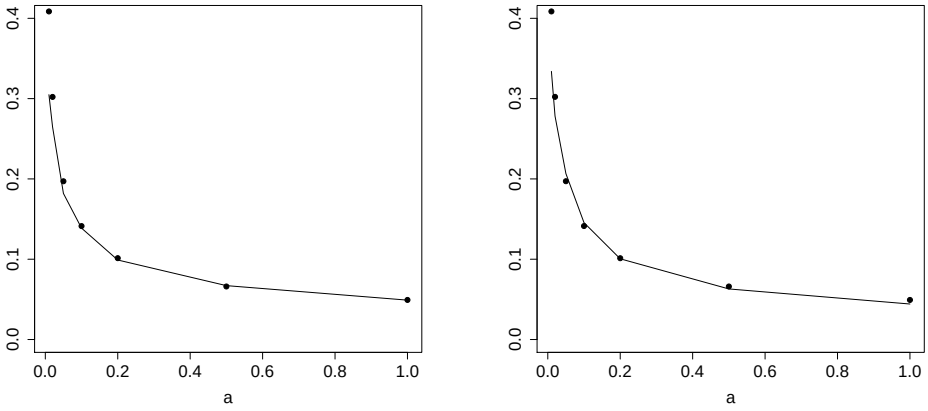


Figure 2.4 – Values of $n^{-1/2}\sqrt{2\gamma_n^{-2}}$ ("•") and the empirical standard deviation of $(\hat{\eta}-\eta^\star)$ (plain line) for several values of $\eta^\star$ (0.3 (left), 0.5 (right)).

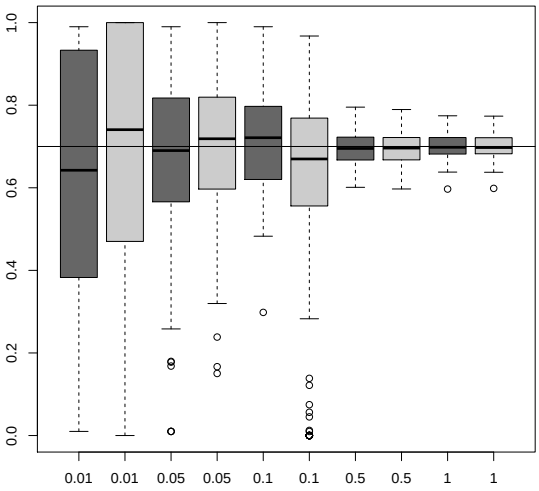


Figure 2.5 – Boxplots of $\hat{\eta}$ for different values of a, using our method (dark gray) and GCTA (light gray). The whiskers of each boxplot are the first and third quartiles.

2.5.3 Results in Model (2.2) when $q < 1$

This section is dedicated to the study of the performance of $\hat{\eta}$ when $q < 1$. We generated 500 data sets according to Model (2.1) for $\eta^\star = 0.7$, $a \in \{0.05, 0.1, 0.5, 1\}$, different values of q and $\mathbf{Z}$ defined in (2.4) where the $W_{i,j}$ are binomial random variables with parameters 2 and p_j. In our experiments the p_j's are uniformly drawn in $[0.1, 0.5]$.

Figure 2.6 displays the boxplots of $\hat{\eta}$ for these parameters. We can see from this figure that for small values of a, the estimators of $\eta^\star$ have the same behavior for $q = 1$ and $q < 1$. However, when $a = 1$ or $a = 0.5$, we can see from this figure that the presence of null components strongly alter the performance of the estimator of $\eta^\star$. Since in typical GWAS experiments, $a = 0.01$ or even smaller, the results of Figure 2.6 could lead to conclude that considering the case $q < 1$ is not necessary for such values of the parameter a. However, as already noticed from Figure 2.2, the variance of $\hat{\eta}$ is very large for small values of a, hence considering the presence of null components and proposing a strategy for selecting only the non null components of $\mathbf{u}$ could be one way to increase a and thus to diminish the variance of $\hat{\eta}$.

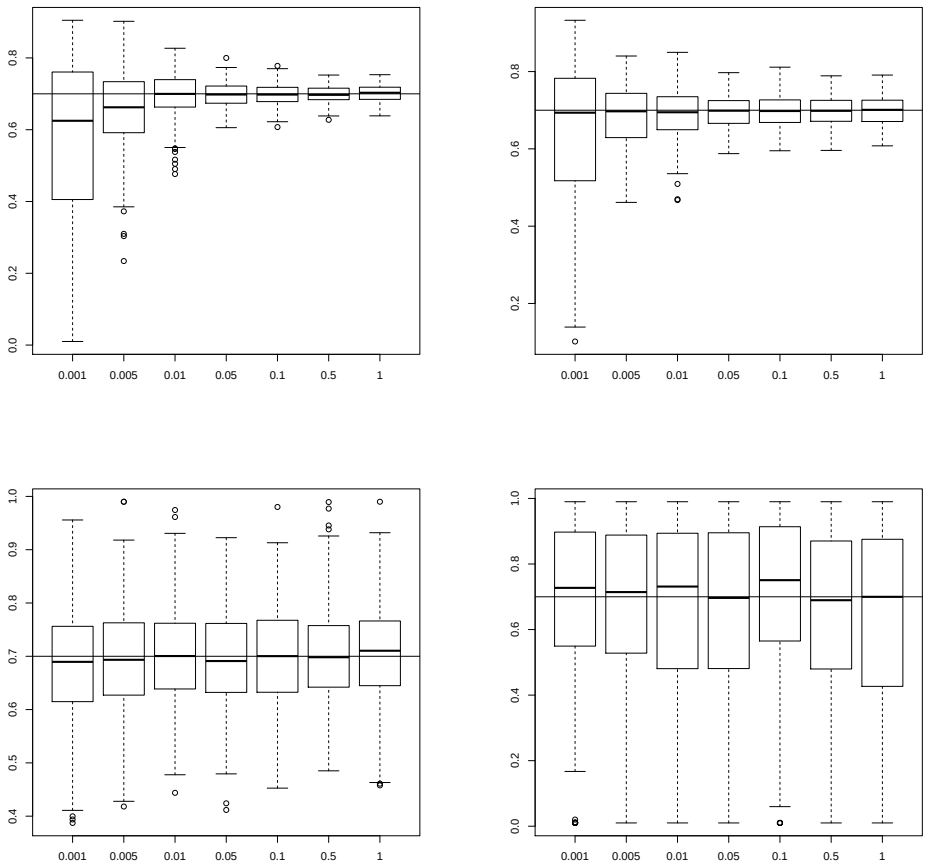


Figure 2.6 – Boxplots of $\hat{\eta}$ for different values of q, with $\eta^\star = 0.7$ and $a = 1$ (top left), $a = 0.5$ (top right), $a = 0.1$ (bottom left) and $a = 0.01$ (bottom right). The boxplots are based on 500 replications. The whiskers of each boxplot are the fist and third quartile.

In order to illustrate the central limit theorem given in Theorem 2, we first display in Figure 2.7 the histograms of $\tau_n^{-1} n^{1/2} (\hat{\eta} - \eta^\star)$ along with the p.d.f of a standard Gaussian random variable for $\eta^\star = 0.7$, two values of q: $q = 0.01$ and $q = 0.1$ and $a = 0.5$ (top) and two values of a: $a = 0.2$ and $a = 0.5$ with $q = 0.5$ (bottom). Here, τ_n is the empirical version of $\tau(a, \eta^\star, q)$

where γ is replaced by γ_n and $S(a, \eta^\star)$ is replaced by its empirical version with the eigenvalues of $\mathbf{R}$. When a is large ($a = 0.5$), we can see that the higher q the better the Gaussian p.d.f fits the histograms.

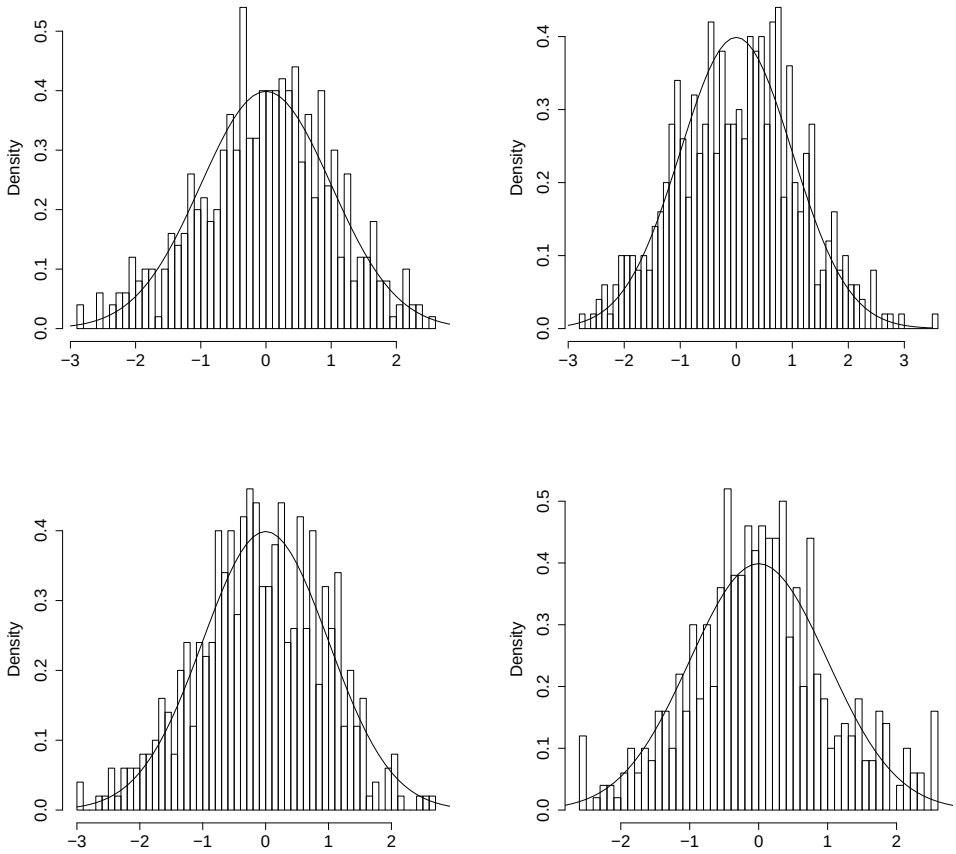


Figure 2.7 – Histograms of $\tau_n^{-1} n^{1/2} (\hat{\eta} - \eta^\star)$ for $a = 0.5$ and $q = 0.5$ (top left), $a = 0.1$ and $q = 0.1$ (top right), and for $a = 0.1$ and $q = 0.01$ (bottom left), $a = 0.05$ and $q = 0.1$ (bottom right).

We also display in Figure 2.8 the values of $n^{-1/2}\tau_n$ and the empirical standard deviation of $(\hat{\eta} - \eta^\star)$ averaged over all the experiments for $\eta^\star = 0.7$ and $q = 0.5$. As shown in Theorem 2, we observe empirically that both quantities are very close. We also display in this figure the value of $n^{-1/2}\tau_n$ with $q = 1$ which boils down to consider the asymptotic standard deviation found in the non sparse model. We can see from this figure that neglecting the term depending on q leads to underestimate the asymptotic variance of $\hat{\eta}$ and that this difference is all the more striking that a is close to 1.

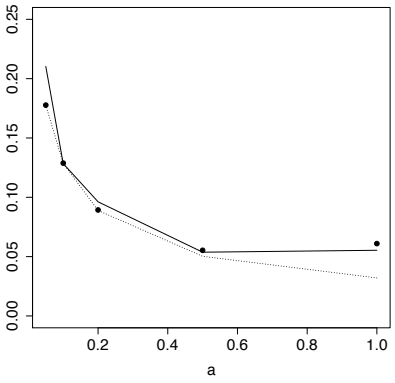


Figure 2.8 – Values of $n^{-1/2}\tau_n$ with the real value of q ($q = 0.5$) ("•"), $q = 1$ (dotted line) and the empirical standard deviation of $(\hat{\eta} - \eta^\star)$ (plain line) for $\eta^\star = 0.7$.

2.6 Discussion

In the course of this study, we have proposed a methodology for estimating the heritability in high dimensional linear mixed models. This methodology has two main features. Firstly, the theoretical performances of our estimator are established in a non standard theoretical framework where n and N tend to infinity and where the components of the random effect part can be equal to zero. Secondly, the computational burden of our approach is very low which makes its use possible on real data coming from GWAS experiments.

As a byproduct of the central limit theorem that we establish for $\eta^\star$ we can derive a confidence interval for the heritability. However, the confidence intervals depend on q which is the proportion of non null components in $\mathbf{u}$ and which is general unknown. For estimating q, several strategies can be considered. One could, for instance, use a GWAS approach to compute the p-values of the correlation tests of each SNP with the observations $\mathbf{Y}$ and then keep only the most significant ones. Such a practical approach can be used for providing a lower bound for q. A refinement of this approach has been proposed by Toro et al. (2015) who observed, through numerical studies, that for a fixed value of the heritability, the minimal p-value is all the more low that the number of causal SNPs is small. Hence, performing a GWAS approach on a given data set allows them to have an idea of the number of SNPs which are likely to be causal. One could also propose another practical method based on a variable selection technique. Such an approach has already been proposed by Fan & Li (2012) in the context of sparse linear mixed models. However, the framework in which their theoretical results are derived is different from the one that is considered in our paper. We are currently working on a paper Bonnet et al. (2016) which presents a variable selection method which is adapted to our framework and which could be used for estimating the proportion q of non null components in the random effects.

Moreover, we did not take into account the linkage disequilibrium issue which would require to extend our results to the case where the columns of the random matrix are correlated. This question will be the subject of a future work.

2.7 Proofs

Let us write the singular value decomposition (SVD) of the $n \times N$ matrix $\mathbf{Z}/\sqrt{N}$ as

$$\frac{1}{\sqrt{N}}\mathbf{Z} = \mathbf{U}\left(\sqrt{\mathbf{D}}\ \mathbf{0}\right)\mathbf{V}'$$

where $\mathbf{U}$ (already introduced in Section 2.1) is a $n \times n$ orthonormal matrix, $\mathbf{V}$ is a $N \times N$ orthonormal matrix and $\sqrt{\mathbf{D}}$ is a $n \times n$ diagonal matrix having its diagonal entries equal to $\sqrt{\lambda_i}$, the λ_i's being the eigenvalues of $\mathbf{R} = \mathbf{Z}\mathbf{Z}'/N$ previously defined. Thus, (2.8) rewrites as

$$\widetilde{\mathbf{Y}} = \mathbf{U}'\mathbf{Y} = \left(\sqrt{\mathbf{D}}\ \mathbf{0}\right)\mathbf{V}'\mathbf{t} + \sigma^\star\sqrt{1-\eta^\star}\,\widetilde{\boldsymbol{\varepsilon}}\,, \tag{2.11}$$

where $\widetilde{\boldsymbol{\varepsilon}} = \mathbf{U}'\boldsymbol{\varepsilon}$ is a $n \times 1$ centered Gaussian vector having a covariance matrix equal to identity.

We shall use repeatedly the following lemma which is proved in Section 2.7.4.

Lemma 2. Let $\widetilde{\mathbf{Y}}$ be defined by (2.11) and $\mathbf{H}$ be a $n \times n$ diagonal matrix, then

$$\mathrm{Var}\left(\widetilde{\mathbf{Y}}'\mathbf{H}\widetilde{\mathbf{Y}}|\mathbf{Z}\right) = 2\sigma^{\star 4}\,\mathrm{Tr}\left[\mathbf{H}^2\left\{(1-\eta^\star)\mathrm{Id}_{\mathbb{R}^n} + \eta^\star\mathbf{D}\right\}^2\right] + 3\sigma^{\star 4}\eta^{\star 2}\left(\frac{1}{q}-1\right)\sum_{1\le i\le N} M_{ii}^2\,,$$

where

$$\mathbf{M} = \mathbf{V}\begin{pmatrix}\mathbf{D}\mathbf{H} & 0\\ 0 & 0\end{pmatrix}\mathbf{V}'\,,$$

and

$$\mathrm{Var}\left(\widetilde{\mathbf{Y}}'\mathbf{H}\widetilde{\mathbf{Y}}|\mathbf{Z}\right) \le 2\sigma^{\star 4}\,\mathrm{Tr}\left[\mathbf{H}^2\left\{(1-\eta^\star)\mathrm{Id}_{\mathbb{R}^n} + \eta^\star\mathbf{D}\right\}^2\right] + 3\sigma^{\star 4}\eta^{\star 2}\left(\frac{1}{q}-1\right)\mathrm{Tr}[\mathbf{D}^2\mathbf{H}^2].$$

Another useful lemma will be the following.

Lemma 3. Under Assumption 1, let $h : \mathbb{R}^+ \to \mathbb{R}^+$ be such that there exist $\alpha > 0$ and C such that for all n,

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n h(\lambda_i)^{1+\alpha}\right) \le C.$$

Then

$$\frac{1}{n}\sum_{i=1}^n h(\lambda_i) = \int h(\lambda)d\nu_a(\lambda) + o_p(1).$$

The proof of this lemma follows from the application of Lemma 1 to the bounded function $h\mathbb{1}_{h\le M}$, and the Markov inequality applied to the empirical mean of $h\mathbb{1}_{h>M}$.

Lemma 4. Under Assumption 1 let $n, N \to \infty$ be such that $n/N \to a > 0$. Then there exists C such that for all n,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \lambda_i^2\right] \le C.$$

To prove the lemma, notice that $\sum_{i=1}^n \lambda_i^2 = \mathrm{Tr}[\mathbf{Z}\mathbf{Z}'/N^2]$. But

$$\begin{aligned}\mathbb{E}\left(\mathrm{Tr}\left[(\mathbf{Z}\mathbf{Z}')^2\right]\right) &= \sum_{k\neq k'}\sum_{i,j}\mathbb{E}(Z_{i,k}Z_{j,k})\mathbb{E}(Z_{i,k'}Z_{j,k'}) + \sum_k\sum_i \mathbb{E}(Z_{i,k}^2)\\ &= nN(N-1) + N(N-1)n(n-1)\left(\frac{1}{n-1}\right)^2 + n^2N\end{aligned}$$

where the values of the involved expectations may be found in the proof of Lemma 1 in Section 2.7.4. We thus have

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \lambda_i^2\right] \leq 2 + \frac{n}{N}$$

which ends the proof.

2.7.1 Proof of Theorem 1

The first step is to prove the consistency of $\hat{\eta}$. We shall first prove that $L_n(\eta)$ converges uniformly for $\eta \in [0, 1-\delta]$ in probability to $L(\eta)$ given by

$$L(\eta) = -2\log\sigma^\star - \log\int\left[\frac{\eta^\star(\lambda-1)+1}{\eta(\lambda-1)+1}\right]d\mu_a(\lambda) - \int\log\left(\eta(\lambda-1)+1\right)d\mu_a(\lambda).$$

Using Lemma 2 with $\mathbf{H}$ with diagonal entries $1/(\eta(\lambda_i-1)+1)$, we get that

$$\begin{aligned}\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^n \frac{\widetilde{Y}_i^2}{\eta(\lambda_i-1)+1}|\mathbf{Z}\right] &\leq \frac{\sigma^{\star 4}}{n^2}\sum_{i=1}^n\left[2\left(\frac{\eta^\star(\lambda_i-1)+1}{\eta(\lambda_i-1)+1}\right)^2\right.\\ &\qquad\left. + 3\left(\frac{1}{q}-1\right)\left(\frac{\eta^\star\lambda_i}{\eta(\lambda_i-1)+1}\right)^2\right]\\ &\leq \sigma^{\star 4}\left(2+3\left(\frac{1}{q}-1\right)\right)\frac{1}{n^2}\sum_{i=1}^n\left(\frac{\lambda_i+1}{\delta}\right)^2\end{aligned}$$

since $\eta \in [0, 1-\delta]$. Now, using Lemma 4 we get that

$$\frac{1}{n^2}\sum_{i=1}^n\left(\frac{\lambda_i+1}{\delta}\right)^2 = o_P(1)$$

which leads to

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n\frac{\widetilde{Y}_i^2}{\eta(\lambda_i-1)+1} &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\frac{\widetilde{Y}_i^2}{\eta(\lambda_i-1)+1}|\mathbf{Z}\right] + o_p(1)\\ &= \sigma^{\star 2}\frac{1}{n}\sum_{i=1}^n\frac{\eta^\star(\lambda_i-1)+1}{\eta(\lambda_i-1)+1} + o_P(1).\end{aligned}$$

Now, using Lemma 3 we easily get that $\frac{1}{n}\sum_{i=1}^n \frac{\eta^\star(\lambda_i-1)+1}{\eta(\lambda_i-1)+1}$ converges in probability to $\int[\frac{\eta^\star(\lambda-1)+1}{\eta(\lambda-1)+1}]d\mu_a(\lambda)$ and $\frac{1}{n}\sum_{i=1}^n \log[(\eta(\lambda_i-1)+1)]$ converges in probability to $\int\log(\eta(\lambda-1)+1)d\mu_a(\lambda)$ so that $L_n(\eta) = L(\eta) + o_P(1)$.

In order to prove the uniform convergence of L_n to L in probability on $[0, 1-\delta]$, we shall use the following lemma which is proved in section 2.7.4.

Lemma 5. Assume that for any $\eta \in [0, 1-\delta]$, $L_n(\eta)$ converges in probability to $L(\eta)$ and that

$$\sup_{\eta\in[0,1-\delta]} \left|L_n'(\eta)\right| = O_P(1), \text{ as } n \text{ tends to infinity,} \tag{2.12}$$

then

$$\sup_{\eta\in[0,1-\delta]} |L_n(\eta) - L(\eta)| = o_P(1), \text{ as } n \text{ tends to infinity.}$$

Let us now prove that $\sup_{\eta\in[0,1-\delta]} |L_n'(\eta)| = O_P(1)$. Note that

$$L_n'(\eta) = \left(\frac{1}{n}\sum_{i=1}^n \frac{\widetilde{Y}_i^2(\lambda_i-1)}{\{\eta(\lambda_i-1)+1\}^2}\right)\left(\frac{1}{n}\sum_{i=1}^n \frac{\widetilde{Y}_i^2}{\eta(\lambda_i-1)+1}\right)^{-1} - \frac{1}{n}\sum_{i=1}^n \frac{\lambda_i-1}{\eta(\lambda_i-1)+1}. \tag{2.13}$$

A study of $\eta \mapsto \left(\frac{1}{n}\sum_{i=1}^n \frac{\widetilde{Y}_i^2(\lambda_i-1)}{\{\eta(\lambda_i-1)+1\}^2}\right)\left(\frac{1}{n}\sum_{i=1}^n \frac{\widetilde{Y}_i^2}{\eta(\lambda_i-1)+1}\right)^{-1}$ shows that it is decreasing and that it takes negative values for $\eta \in [0, 1-\delta]$, so that its absolute value is maximum for $\eta = 1-\delta$. Thus

$$\begin{aligned}\sup_{\eta\in[0,1-\delta]} \left|L_n'(\eta)\right| &\leq \frac{1}{\delta}\left(\frac{1}{n}\sum_{i=1}^n \widetilde{Y}_i^2|\lambda_i-1|\right)\left(\frac{1}{n}\sum_{i=1}^n \widetilde{Y}_i^2\right)^{-1} + \frac{1}{n\delta}\sum_{i=1}^n |\lambda_i-1| \\ &\leq \frac{2}{\delta} + \frac{1}{\delta}\left(\frac{1}{n}\sum_{i=1}^n \widetilde{Y}_i^2\lambda_i\right)\left(\frac{1}{n}\sum_{i=1}^n \widetilde{Y}_i^2\right)^{-1} + \frac{1}{n\delta}\sum_{i=1}^n \lambda_i.\end{aligned}$$

By Lemma 2 with $\mathbf{H} = \mathrm{Id}$, we get

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \widetilde{Y}_i^2 = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \widetilde{Y}_i^2|\mathbf{Z}\right] + o_p(1) &= \frac{\sigma^{2\star}}{n}\sum_{i=1}^n [\eta^\star(\lambda_i-1)+1)] + o_p(1) \\ &= \sigma^{2\star}\int (\eta(\lambda-1)+1)d\mu_a(\lambda) + o_p(1),\end{aligned}$$

where the last equality comes from Lemma 3. In the same way, we get by using Lemma 2 with $\mathbf{H}$ having its diagonal entries equal to λ_i and Lemma 3 that

$$\frac{1}{n}\sum_{i=1}^n \widetilde{Y}_i^2\lambda_i = \sigma^{2\star}\int \lambda(\eta(\lambda-1)+1)d\mu_a(\lambda) + o_p(1) = O_P(1).$$

Finally, we get from Lemma 3 that

$$\frac{1}{n}\sum_{i=1}^n \lambda_i = \int \lambda d\mu_a(\lambda) + o_p(1) = O_P(1)$$

which ends the proof of (2.12). By Lemma 5, we thus have proved that

$$\sup_{\eta\in[0,1-\delta]} |L_n(\eta) - L(\eta)| = o_P(1). \tag{2.14}$$

Now, using Jensen's inequality, we easily get that for all $\eta \in [0,1]$, $L(\eta) \leq L(\eta^\star)$, with equality if and only if $\eta = \eta^\star$. This together with (2.14) gives

$$\hat{\eta} = \eta^\star + o_P(1). \tag{2.15}$$

The next step is to prove that $\sqrt{n}(\hat{\eta} - \eta^\star) = O_P(1)$. Let us first note that $\hat{\eta}$ satisfies the following equation:

$$\sqrt{n}(\hat{\eta} - \eta^\star) = -\frac{\sqrt{n}L_n'(\eta^\star)}{L_n''(\widetilde{\eta})}, \quad \widetilde{\eta} \in (\hat{\eta}, \eta^\star)\ . \tag{2.16}$$

We first prove the asymptotic convergence of $L_n''(\widetilde{\eta})$.

Lemma 6. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ satisfy Model (2.8) with $\eta^\star > 0$ and the entries $W_{i,j}$ of $\mathbf{W}$ satisfy Assumption 1. Then, for all q in $(0,1]$, as $n, N \to \infty$ such that $n/N \to a \in (0,1]$, for any random variable $\widetilde{\eta}$ such that $\widetilde{\eta} \in (\hat{\eta}, \eta^\star)$,

$$L_n''(\widetilde{\eta}) = -{\sigma^\star}^2 \gamma^2(a, \eta^\star) + o_P(1).$$

Lemma 6 is proved in Section 2.7.4.

Let us now focus on the properties of $L_n'(\eta^\star)$. Using the following notation

$$U_i = \frac{\widetilde{Y}_i}{\sqrt{\eta^\star(\lambda_i - 1) + 1}}\ , \tag{2.17}$$

we see that $\sqrt{n}L_n'(\eta^\star)$ can be rewritten as follows:

$$\begin{aligned}
&\left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^n \left({U_i}^2 - \frac{1}{n}\sum_{j=1}^n U_j^2 \right) g(\eta^\star, \lambda_i) \right\} \left(\frac{1}{n}\sum_{i=1}^n U_i^2 \right)^{-1} \\
&= \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[\left({U_i}^2 - 1 \right) + \left(1 - \frac{1}{n}\sum_{j=1}^n U_j^2 \right) \right] g(\eta^\star, \lambda_i) \right\} \left(\frac{1}{n}\sum_{i=1}^n U_i^2 \right)^{-1} \\
&= \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^n \left({U_i}^2 - 1 \right) g(\eta^\star, \lambda_i) \right\} \left(\frac{1}{n}\sum_{i=1}^n U_i^2 \right)^{-1} \\
&\qquad - \left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^n \left({U_j}^2 - 1 \right) \right\} \left\{ \frac{1}{n}\sum_{i=1}^n g(\eta^\star, \lambda_i) \right\} \left(\frac{1}{n}\sum_{i=1}^n U_i^2 \right)^{-1},
\end{aligned}$$

where

$$g(\eta, \lambda) = \frac{\lambda - 1}{\eta(\lambda - 1) + 1}\ .$$

But using Lemma 2 and Lemma 3 we get

$$\mathrm{Var}\left[n^{-1/2} \sum_{j=1}^n ({U_j}^2 - 1) | \mathbf{Z} \right] = O_P(1)$$

Moreover, by Lemma 3, $n^{-1}\sum_{i=1}^n g(\eta^\star, \lambda_i)$ converges in probability to $\int g(\eta^\star, \lambda) \mathrm{d}\mu_a(\lambda)$. Thus,

$$\sqrt{n}L_n'(\eta^\star) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left({U_i}^2 - 1 \right) \left(g(\eta^\star, \lambda_i) - \int g(\eta^\star, \lambda) \mathrm{d}\mu_a(\lambda) \right) + o_P(1), \text{ as } n \to \infty\ . \tag{2.18}$$

Using again Lemma 2 and Lemma 3 we obtain

$$\sqrt{n}L_n'(\eta^\star) = O_P(1).$$

This, together with Lemma 6 and (2.16) ends the proof of Theorem 1.

2.7.2 Proof of Theorem 2

Notice first that all previous results may be used, replacing Assumption 1 by the assumption that the $Z_{i,j}$ are i.i.d. standard Gaussian. Indeed, in this case, Lemma 1 reduces to the original result of Marchenko & Pastur (1968), Lemma 3 only involves Lemma 1 and truncation arguments, and the computations leading to Lemma 4 still hold. Thus, Theorem 1 and Lemma 6 also still hold.

Let us now prove that $\sqrt{n}L_n'(\eta^\star)$ converges in distribution to a centered Gaussian. Define $\mathbf{H}$ the diagonal $n \times n$ matrix with diagonal entries

$$H_i = \frac{1}{\eta^\star(\lambda_i - 1) + 1}\left[g(\eta^\star, \lambda_i) - \int g(\eta^\star, \lambda)\mathrm{d}\mu_a(\lambda)\right].$$

Define

$$\mathbf{L}_n = \frac{1}{\sqrt{n}}\widetilde{\mathbf{Y}}'\mathbf{H}\widetilde{\mathbf{Y}}.$$

Then using (2.18) and Lemma 3 we have

$$\sqrt{n}L_n'(\eta^\star) = \mathbf{L}_n - \mathbb{E}[\mathbf{L}_n|\mathbf{Z}] + o_P(1).$$

Now using Lemma 2 we get that setting $\gamma_n^2 = \mathrm{Var}\,[\mathbf{L}_n|\mathbf{Z}]$,

$$\begin{aligned}\gamma_n^2 &= 2\sigma^{\star 4}\frac{1}{n}\,\mathrm{Tr}\left[\mathbf{H}^2\left((1-\eta)^\star Id_{\mathbb{R}^n} + \eta^\star\mathbf{D}\right)^2\right] + 3\sigma^{\star 4}\eta^{\star 2}\left(\frac{1}{q} - 1\right)\frac{1}{n}\sum_{i=1}^N M_{i,i}^2 \\ &= 2\sigma^{\star 4}\frac{1}{n}\sum_{i=1}^n\left(g(\eta^\star, \lambda_i) - \int g(\eta^\star, \lambda)\mathrm{d}\mu_a(\lambda)\right)^2 \\ &\qquad + 3\sigma^{\star 4}\eta^{\star 2}\left(\frac{1}{q} - 1\right)\frac{1}{n}\sum_{i=1}^n\sum_{k,l=1}^n \lambda_k\lambda_l H_k H_l V_{i,k}^2 V_{i,l}^2.\end{aligned}$$

The first term in this sum converges as $n, N \to \infty$ to $2\sigma^{\star 4}\gamma^2(a, \eta^\star)$.
Under the assumption that the $Z_{i,j}$ are i.i.d. standard Gaussian, the matrix of eigenvectors $\mathbf{V}$ is Haar distributed on the orthonormal matrices, and is independent of $(\lambda_i)_{1\le i\le n}$, see Bai & Silverstein (2010) chapter 10. Conditionally to the eigenvalues $(\lambda_i)_{1\le i\le n}$, we thus get that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n\sum_{k,l=1}^n \lambda_k\lambda_l H_k H_l V_{i,k}^2 V_{i,l}^2|\mathbf{D}\right] = \left(\frac{1}{N}\sum_{k=1}^n \lambda_k H_k\right)^2(1 + o(1))$$

converges to

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

and

$$\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{k,l=1}^{n}\lambda_k\lambda_l H_k H_l V_{i,k}^2 V_{i,l}^2|\mathbf{D}\right] = o_P(1)$$

so that

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

Denote Δ the diagonal $N\times N$-matrix with diagonal entries $\Delta_i = \frac{\sigma^\star\sqrt{\eta^\star}}{\sqrt{q}}\pi_i$. Let us now write

$$\mathbf{L}_n - \mathbb{E}(\mathbf{L}_n|\mathbf{Z}) = \mathbf{L}_n - \mathbb{E}\left[\mathbf{L}_n|\Delta,\mathbf{Z}\right] + \mathbb{E}\left[\mathbf{L}_n|\Delta,\mathbf{Z}\right] - \mathbb{E}\left[\mathbf{L}_n|\mathbf{Z}\right].$$

We first have

$$\mathbb{E}\left[\mathbf{L}_n|\Delta,\mathbf{Z}\right] - E\left[\mathbf{L}_n|\mathbf{Z}\right] = \sigma^{\star 2}\eta^\star\frac{1}{\sqrt{n}}\sum_{i=1}^{N}\left(\frac{\pi_i^2}{q}-1\right)M_{i,i}$$

whose variance, conditionally to $\mathbf{Z}$ is

$$s_{n,1}^2 = \sigma^{\star 4}\eta^{\star 2}\left(\frac{1}{q}-1\right)\frac{1}{n}\sum_{i=1}^{N}M_{i,i}^2.$$

In the same way as for γ_n^2 we get that

$$s_{n,1}^2 = \sigma^{\star 4}\eta^{\star 2}\left(\frac{1}{q}-1\right)S(a,\eta^\star) + o_P(1).$$

Let

$$\xi_i = \left(\frac{\pi_i^2}{q}-1\right)M_{i,i} = \left(\frac{\pi_i^2}{q}-1\right)\sum_{k=1}^{n}\frac{\lambda_k(\lambda_k-1)}{(\eta^\star(\lambda_k-1)+1)^2}V_{i,k}^2.$$

Since $\eta^\star > 0$, the function $\lambda \mapsto \frac{\lambda(\lambda-1)}{(\eta^\star(\lambda-1)+1)^2}$ is bounded, and $\sum_{k=1}^{n}V_{i,k}^2 \leq \sum_{k=1}^{N}V_{i,k}^2 = 1$. Also, the variables $\left(\frac{\pi_i^2}{q}-1\right)$ are uniformly bounded by $1/q$. Thus

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\xi_i^2\mathbb{1}_{|\xi_i|\geq cn}|\mathbf{Z}\right] = 0$$

for large enough n. Then, by Lindeberg's Theorem, conditionally to $\mathbf{Z}$,

$$\frac{1}{s_{n,1}}\left(\mathbb{E}\left[\mathbf{L}_n|\Delta,\mathbf{Z}\right] - \mathbb{E}\left[\mathbf{L}_n|\mathbf{Z}\right]\right)$$

converges in distribution to $\mathcal{N}(0,1)$.
Let us now set

$$s_{n,2}^2 = \gamma_n^2 - s_{n,1}^2$$

and notice that $s_{n,2}^2$ converges to

$$2\sigma^{\star 4}\gamma^2(a,\eta^\star) + 2\sigma^{\star 4}\eta^{\star 2}\left(\frac{1}{q}-1\right)S(a,\eta^\star).$$

We shall prove that, conditionally to $\mathbf{Z}$ and Δ, $(\mathbf{L}_n - \mathbb{E}(\mathbf{L}_n|\Delta,\mathbf{Z}))/s_{n,2}$ converges in distribution to $\mathcal{N}(0,1)$, and thus also unconditionally. Write

$$\mathbf{L}_n = \frac{1}{\sqrt{n}} \begin{pmatrix} \mathbf{w}' & \boldsymbol{\varepsilon}' \end{pmatrix} B \begin{pmatrix} \mathbf{w} \\ \boldsymbol{\varepsilon} \end{pmatrix}$$

where B is the $(N+n)\times(N+n)$-matrix

$$B = \begin{pmatrix} \Delta & 0 \\ 0 & \sigma^\star(1-\eta^\star)^{\frac{1}{2}} Id_{\mathbb{R}^n} \end{pmatrix} \begin{pmatrix} \mathbf{V} \begin{pmatrix} \mathbf{DH} & 0 \\ 0 & 0 \end{pmatrix} \mathbf{V}' & \tilde{\mathbf{V}}\sqrt{\mathbf{D}}\mathbf{H} \\ \mathbf{H}\sqrt{\mathbf{D}}\tilde{\mathbf{V}}' & \mathbf{H} \end{pmatrix} \begin{pmatrix} \Delta & 0 \\ 0 & \sigma^\star(1-\eta^\star)^{\frac{1}{2}} Id_{\mathbb{R}^n} \end{pmatrix}.$$

Here, $\tilde{\mathbf{V}}$ is the $N\times n$-matrix which consists of the first n columns of $\mathbf{V}$. Let ϕ be the characteristic function of $(\mathbf{L}_n - \mathbb{E}(\mathbf{L}_n|\Delta,\mathbf{Z}))/s_{n,2}$ conditionally to $\mathbf{Z}$ and Δ. Notice first that if b_j, $j = 1,\dots,n+N$ are the eigenvalues of B, we may write

$$\mathbf{L}_n - \mathbb{E}\left[\mathbf{L}_n|\Delta,\mathbf{Z}\right] = \frac{1}{\sqrt{n}} \sum_{j=1}^{N+n} b_j(e_j^2 - 1).$$

for random variables e_j i.i.d. standard Gaussian. Thus

$$\phi\left(t\right) = \prod_{j=1}^{N+n} \left[\left(1 - 2i\frac{tb_j}{s_{n,2}\sqrt{n}}\right)^{-1/2} \exp\left(-i\frac{tb_j}{s_{n,2}\sqrt{n}}\right)\right]$$

and Taylor expansion leads to

$$\begin{aligned} \log\phi\left(t\right) &= \sum_{j=1}^{N+n} \left[-\frac{1}{2}\log\left(1 - 2i\frac{tb_j}{s_{n,2}\sqrt{n}}\right) - i\frac{tb_j}{s_{n,2}\sqrt{n}}\right] \\ &= -t^2\frac{1}{ns_{n,2}^2}\sum_{j=1}^{N+n} b_j^2 + O\left[\frac{1}{n\sqrt{n}s_{n,2}^3}\sum_{j=1}^{N+n} b_j^3\right]. \end{aligned}$$

We shall now prove that $\frac{1}{ns_{n,2}^2}\sum_{j=1}^{N+n} b_j^2$ converges to $1/2$. Tedious computations give

$$\begin{aligned} \sum_{j=1}^{N+n} b_j^2 &= \mathrm{Tr}(B^2) \\ &= \mathrm{Tr}(\Delta M \Delta^2 M \Delta) + \sigma^{\star 4}(1-\eta^{\star 2}\,\mathrm{Tr}(\mathbf{H}^2) + 2\sigma^{\star 2}(1-\eta^\star)\,\mathrm{Tr}[\Delta^2\tilde{\mathbf{V}}\mathbf{D}\mathbf{H}^2\tilde{\mathbf{V}}']. \end{aligned}$$

Using the distribution of $\mathbf{V}$ and its independence on $\mathbf{D}$ we get

$$\begin{aligned} \mathbb{E}\left[\sum_{j=1}^{N+n} b_j^2|\mathbf{D}\right] &= 2\sigma^{\star 4}\,\mathrm{Tr}\left[\mathbf{H}^2\left((1-\eta)^\star Id_{\mathbb{R}^n} + \eta^\star\mathbf{D}\right)^2\right] \\ &\quad + 2\sigma^{\star 4}\eta^{\star 2}\left(\frac{1}{q} - 1\right)\left(\frac{1}{N}\sum_{k=1}^{n}\lambda_k H_k\right)^2 (1+o(1)) \end{aligned}$$

so that

$$\mathbb{E}\left[\frac{1}{n}\sum_{j=1}^{N+n} b_j^2|\mathbf{D}\right] = 2\sigma^{\star 4}\gamma^2(a,\eta^\star) + 2\sigma^{\star 4}\eta^{\star 2}\left(\frac{1}{q}-1\right)S(a,\eta^\star) + o_P(1).$$

Moreover, tedious computations again give

$$\mathrm{Var}\left[\frac{1}{n}\sum_{j=1}^{N+n} b_j^2|\mathbf{D}\right] = o_P(1),$$

and we obtain that

$$\frac{1}{ns_{n,2}^2}\sum_{j=1}^{N+n} b_j^2 = \frac{1}{2} + o_P(1).$$

We shall now prove that $\frac{1}{n\sqrt{n}s_{n,2}^3}\sum_{j=1}^{N+n} b_j^3 = o_P(1)$. To do so, it is enough to prove that $\max_j |b_j| = o_P(\sqrt{n})$. Notice that for any normed vector $A = (A_1, A_2)$ in $\mathbb{R}^{N+n}$ where $A_1 \in \mathbb{R}^N$ and $A_2 \in \mathbb{R}^n$,

$$\max_j |b_j| \leq A'BA.$$

Now,

$$A'BA = A_1'(\Delta M\Delta)A_1 + 2\sigma^\star\sqrt{1-\eta^\star}A_1'(\Delta\tilde{\mathbf{V}}\sqrt{\mathbf{D}}\mathbf{H})A_2 + \sigma^{\star 2}(1-\eta^\star)A_2'\mathbf{H}A_2.$$

First, since $\eta^\star > 0$, all entries of $\mathbf{H}$ and $\mathbf{D}$ and $\mathbf{HD}$ are uniformly bounded and so are all entries of Δ. We thus get $A_2'\mathbf{H}A_2 = O(1)$ and $A_1'(\Delta\tilde{\mathbf{V}}\sqrt{\mathbf{D}}\mathbf{H})A_2 = O(1)$. Then, using the distribution of $\mathbf{V}$ and its independence on $\mathbf{D}$ we get

$$\mathbb{E}\left[A_1'(\Delta M\Delta)A_1|\mathbf{D}\right] = O\left(\frac{1}{N}\sum_{i=1}^{n}\lambda_i H_i\right)$$

and

$$\mathrm{Var}\left[A_1'(\Delta M\Delta)A_1|\mathbf{D}\right] = o_P(1),$$

so that $A'BA = O_P(1)$. We have thus proved that $\max_j |b_j| = O_P(1) = o_P(\sqrt{n})$.

Thus $\phi(t)$ converges in probability for all t to $\exp -\frac{t^2}{2}$ and the convergence may be strengthened by contradiction to an a.s. convergence, so that conditionally to $\mathbf{Z}$ and Δ, $(\mathbf{L}_n - \mathbb{E}(\mathbf{L}_n|\Delta,\mathbf{Z}))/s_{n,2}$ converges in distribution to $\mathcal{N}(0,1)$.

Now, conditionally to $\mathbf{Z}$ and Δ, $(\mathbf{L}_n - E(\mathbf{L}_n|\Delta,Z))/s_{n,2}$ converges in distribution to a Gaussian random variable independent of Δ. Thus conditionally to $\mathbf{Z}$, $\mathbf{L}_n - E\left[\mathbf{L}_n|\Delta,Z\right]$ and $E\left[\mathbf{L}_n|\Delta,Z\right] - E\left[\mathbf{L}_n|Z\right]$ converge in distribution to independent Gaussian variables, so that their sum converges in distribution to a centered Gaussian with variance the sum of the variances, namely the limit of γ_n^2, and Theorem 2 is proved.

2.7.3 Proof of Theorem 3

Using Lemma 6 and (2.16) , there remains to prove that $\sqrt{n}L'_n(\eta^\star)$ converges in distribution to $\mathcal{N}(0, 2\sigma^{\star 4}\gamma^2(a,\eta^\star))$ and that γ_n^2 converges in probability to $\gamma^2(a,\eta^\star)$.
Notice first that when $q=1$, $(U_1,\dots,U_n)|\mathbf{Z}$ is a centered Gaussian vector with a covariance matrix equal to $\sigma^{\star 2}$ times the identity matrix. We shall prove that conditionally to $\mathbf{Z}$, $\sqrt{n}L'_n(\eta^\star)$ converges in distribution to $\mathcal{N}(0, 2\sigma^{\star 4}\gamma^2(a,\eta^\star))$ so that the result still holds unconditionally. Using (2.18), it is only needed to prove it for $\frac{1}{\sqrt{n}}\sum_{i=1}^n \left({U_i}^2-1\right)\left(g(\eta^\star,\lambda_i)-\int g(\eta^\star,\lambda)\mathrm{d}\mu_a(\lambda)\right)$. Now, conditionally to $\mathbf{Z}$, the variance of

$$\sum_{i=1}^n \left({U_i}^2-1\right)\left(g(\eta^\star,\lambda_i)-\int g(\eta^\star,\lambda)\mathrm{d}\mu_a(\lambda)\right)$$

is

$$\gamma_n^2 = \frac{2\sigma^{\star 4}}{n}\sum_{i=1}^n \left(g(\eta^\star,\lambda_i)-\int g(\eta^\star,\lambda)\mathrm{d}\mu_a(\lambda)\right)^2.$$

Since $\eta^\star>0$, $g(\eta^\star,\lambda)$ is a bounded function of λ, and using Lemma 3,

$$\gamma_n^2 = 2\sigma^{\star 4}\gamma^2(a,\eta^\star)) + o_P(1).$$

Also, setting $\xi_i = \left({U_i}^2-1\right)\left(g(\eta^\star,\lambda_i)-\int g(\eta^\star,\lambda)\mathrm{d}\mu_a(\lambda)\right)$ and C an upper bound of $|g(\eta^\star,\lambda)|$, we get that for any $c>0$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\xi_i^2 \mathbb{1}_{|\xi_i|\geq cn}|\mathbf{Z}\right] &\leq 4C^2\sigma^{\star 4}\mathbb{E}\left[\left({U_1}^2-1\right)^2 \mathbb{1}_{2C|{U_1}^2-1|\geq cn}|Z\right]\\ &= 4C^2\sigma^{\star 4}\mathbb{E}\left[\left({U_1}^2-1\right)^2 \mathbb{1}_{2C|{U_1}^2-1|\geq cn}\right] = o(1),\end{aligned}$$

where the first equality comes from the fact that the distribution of $(U_1,\dots,U_n)|\mathbf{Z}$ does not depend on $\mathbf{Z}$ and is thus also the distribution of $(U_1,\dots,U_n)$. Then, using Lindeberg's Theorem, conditionally to $\mathbf{Z}$, $\sqrt{n}L'_n(\eta^\star)$ converges in distribution to $\mathcal{N}(0, 2\sigma^{\star 4}\gamma^2(a,\eta^\star))$ and thus also unconditionally.
The fact that γ_n^2 converges in probability to $\gamma^2(a,\eta^\star)$ is a straightforward consequence of Taylor expansion, the fact that $g(\eta^\star,\lambda)$ and its derivative with respect to η in the neighborhood of $\eta^\star$ are bounded functions of λ, and Slutzky's Lemma.

2.7.4 Proofs of technical lemmas

Proof of Lemma 1

As a byproduct of Theorem 1.1, Corollary 1.1 and Remark 1.1 of Bai & Zhou (2008), we use the following result to prove Lemma 1.

Theorem (Bai and Zhou (2008)). Let $\mathbf{Z}$ be a matrix of size $n\times N$ which columns, denoted by $Z_1,\dots,Z_N$, are independent and let us denote $\bar{\mathbf{Z}} = \frac{1}{N}\sum_{k=1}^N Z_k$. Let us also recall that $\mathbf{R} = \mathbf{Z}\mathbf{Z}'/N$ and $F^{\mathbf{R}}$ is its empirical spectral distribution defined by $F^{\mathbf{R}}(x) = \frac{1}{n}\sum_{k=1}^n 1_{\{\lambda_k\geq x\}}$, where $\lambda_1,\dots,\lambda_n$ are the eigenvalues of $\mathbf{R}$. As $N\to\infty$, assume the following:

1. $T=(t_{i,j})$ is a matrix such that $\mathbb{E}(\bar{Z}_{i,j}Z_{m,j}) = t_{m,i}$ for all j .

2. $\frac{1}{N}\max_{i\neq m}\mathbb{E}(\bar{Z}_{i,j}Z_{m,j})^2 \to 0$ uniformly in $j \leq N$.

3. $\frac{1}{N^2}\sum_{\Lambda}\left(\mathbb{E}(\bar{Z}_{i,j}Z_{m,j} - t_{m,i})(Z_{i',j}\bar{Z}_{m',j} - t_{i',m'})\right)^2 \to 0$ uniformly in $j \leq N$, with $\Lambda = \{(i,m,i',m') : 1 \leq i,m,i',m' \leq n\}\backslash\{(i,m,i',m') : i = i' \neq m = m' \text{ or } i = m' \neq i' = m\}$.

4. $\frac{n}{N} \to a \in (0,+\infty)$.

5. The norm of T is uniformly bounded and F^T tends to a degenerate distribution with mass at $1/a$.

Then, with probability 1, F^R converges to the Marchenko-Pastur distribution defined in (2.9).

Observe that for all $j = 1,\dots,N$,

$$\sum_{i=1}^{n} Z_{i,j} = 0 \tag{2.19}$$

and

$$\sum_{i=1}^{n} Z_{i,j}^2 = n. \tag{2.20}$$

Moreover, for each j, the random variables $(Z_{i,j})_{1\leq i\leq n}$ are exchangeable. Thus, we deduce from (2.20) that for all $i = 1,\dots,n$ and $j = 1,\dots,N$, $\mathbb{E}(Z_{i,j}^2) = 1$. Hence, by (2.19), we get that

$$0 = \left(\sum_{i=1}^{n} Z_{i,j}\right)^2 = \sum_{i=1}^{n} Z_{i,j}^2 + \sum_{1\leq i\neq m\leq n} Z_{i,j}Z_{m,j}\ ,$$

which, by (2.20), implies that for all $j = 1,\dots,N$ and $i \neq m = 1,\dots,n$,

$$\mathbb{E}(Z_{i,j}Z_{m,j}) = -\frac{n}{n(n-1)} = -\frac{1}{n-1}\ . \tag{2.21}$$

Thus, the matrix $\mathbf{T} = \mathbf{T}_n$ defined in Theorem (Bai and Zhou (2008)) is equal to $\mathbf{T} = n/(n-1)\mathrm{Id}_{\mathbb{R}^n} - \mathbf{J}_n/(n-1)$, where $\mathbf{J}_n$ is a $n \times n$ matrix having all its entries equal to 1. Hence the eigenvalues of $\mathbf{T}$ are 0 with multiplicity 1 and $n/(n-1)$ with multiplicity $(n-1)$, which gives Condition 5. of Theorem (Bai and Zhou (2008)).

Let us then check Condition 2. of Theorem (Bai and Zhou (2008)). Observe that, for $i \neq m$, $\mathbb{E}[(Z_{i,j}Z_{m,j} - t_{m,i})^2] = \mathbb{E}(Z_{i,j}^2Z_{m,j}^2) - t_{m,i}^2$. By (2.20), for all $j = 1,\dots,N$,

$$n^2 = \left(\sum_{i=1}^{n} Z_{i,j}^2\right)^2 = \sum_{i=1}^{n} Z_{i,j}^4 + \sum_{1\leq i\neq m\leq n} Z_{i,j}^2Z_{m,j}^2\ .$$

Since the $(Z_{i,j})_{1\leq i\leq n}$ are exchangeable for each $j = 1,\dots,N$, we get that for all $j = 1,\dots,N$,

$$n = \mathbb{E}[Z_{1,j}^4] + (n-1)\mathbb{E}[Z_{1,j}^2Z_{2,j}^2]\ .$$

Thus, for all $j = 1,\dots,N$, $\mathbb{E}[Z_{1,j}^2Z_{2,j}^2] \leq n/(n-1)$, which with the definition of the $t_{m,i}$'s gives the result.

Let us now check Condition 3. of Theorem (Bai and Zhou (2008)). Since the random variables $(Z_{i,j})_{1\leq i\leq n}$ are exchangeable, it is enough to prove that, uniformly in k,

(i) $\mathbb{E}[Z_{1,k}^4] = o(\sqrt{n})$,

(ii) $\mathbb{E}[Z_{1,k}^2 Z_{2,k}^2] - 1 = o(1)$,

(iii) $\mathbb{E}[Z_{1,k}^3 Z_{2,k}] = o(1)$,

(iv) $\sqrt{n}\mathbb{E}[Z_{1,k}^2 Z_{2,k} Z_{3,k}] = o(1)$,

(v) $n\mathbb{E}[Z_{1,k} Z_{2,k} Z_{3,k} Z_{4,k}] = o(1)$, as $n \to \infty$.

Observe that (i) implies (ii). Using (2.19), by expanding $0 = (\sum_{i=1}^n Z_{i,k})^2 \left(\sum_{i=1}^n Z_{i,k}^2\right)$ and taking the expectation, we get that (i) and (iii) imply (iv). By expanding $0 = (\sum_{i=1}^n Z_{i,k})^4$, which comes from (2.19), and by taking the expectation, (i) and (iii) imply (v). Hence, it is enough to prove (i) and (iii) to conclude the proof of Lemma 1.

Let us first prove (i). By the definition of $Z_{1,k}$ given in (2.4), we get that for all k, $Z_{1,k}^2 \leq n$. Hence,

$$Z_{1,k}^2 \leq \frac{(W_{1,k} - \overline{W}_k)^2}{2\sigma_k^2} \mathbb{1}_{\{s_k^2 \geq \sigma_k^2/2\}} + n\mathbb{1}_{\{s_k^2 < \sigma_k^2/2\}} ,$$

and, by the assumptions on the $W_{i,k}$'s and on the σ_k's,

$$\mathbb{E}(Z_{1,k}^4) \leq \frac{W_M^2}{\kappa^2/2} + n^2 \mathbb{P}(\sigma_k^2 - s_k^2 > \sigma_k^2/2) .$$

Theorem A of (Serfling, 1980, p. 201) implies that the second term of the previous inequality tends to zero as n tends to infinity uniformly in k, which concludes the proof of (i).

Let us now prove (iii). Using (2.19), we get $Z_{1,k}^3 (\sum_{i=1}^n Z_{i,k}) = 0$. By expanding this equation and taking the expectation, we obtain that $\mathbb{E}(Z_{1,k}^4) + \sum_{i=2}^n \mathbb{E}(Z_{1,k}^3 Z_{i,k}) = 0$. Since the $(Z_{i,k})_{1 \leq i \leq n}$ are exchangeable: $\mathbb{E}(Z_{1,k}^3 Z_{2,k}) = -\mathbb{E}(Z_{1,k}^4)/(n-1) = o(n^{-1/2})$, where the last equality comes from (i).

Proof of Lemma 2

Using (2.11) and the independence assumptions, we get

$$\begin{aligned}
&\mathrm{Var}(\widetilde{\mathbf{Y}}'\mathbf{H}\widetilde{\mathbf{Y}}|\mathbf{Z}) \\
&\quad = \mathrm{Var}\left[\mathbf{v}'\mathbf{V}\begin{pmatrix}\mathbf{DH} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{V}'\mathbf{v} + 2\sigma^\star\sqrt{1-\eta^\star}\mathbf{v}'\mathbf{V}\begin{pmatrix}\sqrt{\mathbf{D}}\\ \mathbf{0}\end{pmatrix}\mathbf{H}\widetilde{\boldsymbol{\varepsilon}} + \sigma^{\star 2}(1-\eta^\star)\widetilde{\boldsymbol{\varepsilon}}'\mathbf{H}\widetilde{\boldsymbol{\varepsilon}}|\mathbf{Z}\right] \\
&\quad = \mathrm{Var}\left[\mathbf{v}'\mathbf{M}\mathbf{v}|\mathbf{Z}\right] + 4\sigma^{\star 2}(1-\eta^\star)\,\mathrm{Var}\left[\mathbf{v}'\mathbf{V}\begin{pmatrix}\sqrt{\mathbf{D}}\\ \mathbf{0}\end{pmatrix}\mathbf{H}\widetilde{\boldsymbol{\varepsilon}}|\mathbf{Z}\right] + 2\sigma^{\star 4}(1-\eta^\star)^2\,\mathrm{Tr}(\mathbf{H}^2) , \quad (2.22)
\end{aligned}$$

where $\mathbf{M} = \mathbf{V}\begin{pmatrix}\mathbf{DH} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{V}'$. Using the independence assumptions, we get that

$$\begin{aligned}
4\sigma^{\star 2}(1-\eta^\star)\,\mathrm{Var}\left[\mathbf{v}'\mathbf{V}\begin{pmatrix}\sqrt{\mathbf{D}}\\ \mathbf{0}\end{pmatrix}\mathbf{H}\widetilde{\boldsymbol{\varepsilon}}|\mathbf{Z}\right] &= 4\sigma^{\star 4}\eta^\star(1-\eta^\star)\,\mathrm{Tr}(\mathbf{BB}') \\
&= 4\sigma^{\star 4}\eta^\star(1-\eta^\star)\,\mathrm{Tr}(\mathbf{DH}^2) , \quad (2.23)
\end{aligned}$$

where $\mathbf{B} = \mathbf{V}\begin{pmatrix}\sqrt{\mathbf{D}}\\ \mathbf{0}\end{pmatrix}\mathbf{H}$. Moreover, $\mathbb{E}(\mathbf{v}'\mathbf{M}\mathbf{v}|\mathbf{Z}) = \sigma^{\star 2}\eta^\star \operatorname{Tr}(\mathbf{D}^2\mathbf{H}^2)$ and

$$\begin{aligned}\mathbb{E}\left[(\mathbf{v}'\mathbf{M}\mathbf{v})^2|\mathbf{Z}\right] &= \frac{\sigma^{\star 4}\eta^{\star 2}}{q^2}\left[2q^2\sum_{1\le i\neq j\le N} M_{ij}^2 + q^2\sum_{1\le i\neq i'\le N} M_{ii}M_{i'i'} + 3q\sum_{1\le i\le N} M_{ii}^2\right]\\ &= \sigma^{\star 4}\eta^{\star 2}\left[2\operatorname{Tr}(\mathbf{M}^2) - 2\sum_{1\le i\le N} M_{ii}^2 + \operatorname{Tr}(\mathbf{M})^2 - \sum_{1\le i\le N} M_{ii}^2 + \frac{3}{q}\sum_{1\le i\le N} M_{ii}^2\right]\\ &= \sigma^{\star 4}\eta^{\star 2}\left[2\operatorname{Tr}(\mathbf{D}^2\mathbf{H}^2) + \operatorname{Tr}(\mathbf{M})^2 + 3\left(\frac{1}{q}-1\right)\sum_{1\le i\le N} M_{ii}^2\right].\end{aligned}$$

Thus,

$$\operatorname{Var}\left[\mathbf{v}'\mathbf{M}\mathbf{v}\middle|\,\mathbf{Z}\right] = \sigma^{\star 4}\eta^{\star 2}\left[2\operatorname{Tr}(\mathbf{D}^2\mathbf{H}^2) + 3\left(\frac{1}{q}-1\right)\sum_{1\le i\le N} M_{ii}^2\right]. \tag{2.24}$$

The proof of the equality in Lemma 2 follows from (2.22), (2.23) and (2.24). The proof of the inequality in Lemma 2 follows now from

$$\sum_{1\le i\le N} M_{ii}^2 \le \sum_{1\le i,j\le N} M_{ij}^2 = \operatorname{Tr}[\mathbf{D}^2\mathbf{H}^2].$$

Proof of Lemma 5

Let $\epsilon > 0$ and let $\{\eta_1 < \cdots < \eta_{K(\epsilon)}\}$ be a grid of $[0, 1-\delta]$ such that $|\eta_j - \eta_{j+1}| < \epsilon$ for all $j \in \{0, \ldots, K_\epsilon\}$ then

$$\begin{aligned}\sup_{\eta\in[0,1-\delta]} |L_n(\eta) - L(\eta)| &\le \sup_{j\in\{1,\ldots,K_\epsilon\}}\left[\sup_{\eta'\in[\eta_j,\eta_{j+1}]} |L_n(\eta') - L_n(\eta_j)| + |L_n(\eta_j) - L(\eta_j)|\right.\\ &\quad\left. + \sup_{\eta'\in[\eta_j,\eta_{j+1}]} |L(\eta_j) - L(\eta')|\right]\\ &\le \epsilon \sup_{\eta\in[0,1-\delta]} |L_n'(\eta)| + \sup_{j\in\{1,\ldots,K_\epsilon\}} |L_n(\eta_j) - L(\eta_j)| + \omega(\epsilon),\end{aligned}$$

where $\omega(\epsilon)$ is the modulus of continuity of L, which is continuous on the compact $[0, 1-\delta]$ and thus uniformly continuous on this compact. Since $\sup_{\eta\in[0,1-\delta]} |L_n'(\eta)| = O_P(1)$ then, for every $\beta > 0$, there exists C such that for all n, $\mathbb{P}(\sup_{\eta\in[0,1-\delta]} |L_n'(\eta)| \ge C) \le \beta$. Let $\alpha > 0$ and let us

consider the ϵ-grid such that $\epsilon \leq \alpha/3C$ and $\omega(\epsilon) \leq \alpha/3$, thus we get that

$$\begin{aligned}
&\mathbb{P}(\sup_{\eta\in[0,1-\delta]} |L_n(\eta) - L(\eta)| \geq \alpha) \\
&\leq \mathbb{P}(\sup_{\eta\in[0,1-\delta]} \left|L_n'(\eta)\right| \geq C) + \mathbb{P}(\sup_{j\in\{1,\dots,K_\epsilon\}} |L_n(\eta_j) - L(\eta_j)| \geq \alpha - C\epsilon - \omega(\epsilon)) \\
&\leq \mathbb{P}(\sup_{\eta\in[0,1-\delta]} \left|L_n'(\eta)\right| \geq C) + \mathbb{P}(\sup_{j\in\{1,\dots,K_\epsilon\}} |L_n(\eta_j) - L(\eta_j)| \geq \frac{\alpha}{3}) \\
&\leq \mathbb{P}(\sup_{\eta\in[0,1-\delta]} \left|L_n'(\eta)\right| \geq C) + \sum_{j=1}^{K_\epsilon} \mathbb{P}(|L_n(\eta_j) - L(\eta_j)| \geq \frac{\alpha}{3}),
\end{aligned}$$

which concludes the proof of Lemma 5 since each term tends to zero as n tends to infinity.

Proof of Lemma 6

The second derivative of L_n is given by

$$\begin{aligned}
L_n''(\eta) = &\left(-\frac{2}{n}\sum_{i=1}^n \frac{\widetilde{Y}_i^2(\lambda_i-1)^2}{\{\eta(\lambda_i-1)+1\}^3}\right)\left(\frac{1}{n}\sum_{i=1}^n \frac{\widetilde{Y}_i^2}{\{\eta(\lambda_i-1)+1\}}\right)^{-1} \\
&+ \left(\frac{1}{n}\sum_{i=1}^n \frac{\widetilde{Y}_i^2(\lambda_i-1)}{\{\eta(\lambda_i-1)+1\}^2}\right)^2 \left(\frac{1}{n}\sum_{i=1}^n \frac{\widetilde{Y}_i^2}{\{\eta(\lambda_i-1)+1\}}\right)^{-2} \\
&+ \frac{1}{n}\sum_{i=1}^n \frac{(\lambda_i-1)^2}{\{\eta(\lambda_i-1)+1\}^2}\,.
\end{aligned} \tag{2.25}$$

In particular for $\eta = \eta^\star$, we have

$$\frac{1}{n}\sum_{i=1}^n \frac{\widetilde{Y}_i^2}{\{\eta^\star(\lambda_i-1)+1\}} = 1 + o_P(1),$$

and using as previously Lemma 2, Lemma 3 and the fact that all functions of λ involved in the empirical means are bounded since $\eta^\star > 0$, we get

$$\begin{aligned}
\frac{2}{n}\sum_{i=1}^n \frac{\widetilde{Y}_i^2(\lambda_i-1)^2}{\{\eta(\lambda_i-1)+1\}^3} &= \frac{2\sigma^{\star 2}}{n}\sum_{i=1}^n \frac{(\lambda_i-1)^2}{\{\eta(\lambda_i-1)+1\}^2} + o_P(1) \\
&= 2\sigma^{\star 2}\int \frac{(\lambda-1)^2}{\{\eta(\lambda-1)+1\}^2} d\mu_a(\lambda) + o_P(1)
\end{aligned}$$

and

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \frac{\widetilde{Y}_i^2(\lambda_i-1)}{\{\eta(\lambda_i-1)+1\}^2} &= \frac{\sigma^{\star 2}}{n}\sum_{i=1}^n \frac{(\lambda_i-1)}{\{\eta(\lambda_i-1)+1\}} + o_P(1) \\
&= \sigma^{\star 2}\int \frac{(\lambda-1)}{\{\eta(\lambda-1)+1\}} d\mu_a(\lambda) + o_P(1)
\end{aligned}$$

leading to

$$L_n''(\eta) = -\sigma^{\star 2}\gamma^2(a,\eta^\star) + o_P(1).$$

Using Slutzky's Lemma and $\hat{\eta} = \eta^\star + o_P(1)$, there just remains to prove that for small enough $\alpha > 0$,

$$\sup_{|\eta - \eta^\star| \leq \alpha} |L_n''(\eta) - L_n''(\eta)| = O_p(\alpha).$$

But this comes easily from

$$\sup_{|\eta - \eta^\star| \leq \alpha} |L_n''(\eta) - L_n''(\eta)| \leq \alpha \sup_{|\eta - \eta^\star|} |L_n^{(3)}(\eta)|$$

where $L_n^{(3)}(\eta)$ is the third derivative of $L_n(\eta)$, and a similar handling of empirical means as before. Indeed, all functions of λ involved are bounded as soon as α is such that $\eta^\star \geq 2\alpha$.

Chapter 3

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

The content of this chapter is contained in the article sumbitted for publication:
A. Bonnet, C. Lévy-Leduc, E. Gassiat, R. Toro, and T. Bourgeron. Improving heritability by a variable selection approach in sparse high dimensional linear mixed models, 2016, http://arxiv.org/abs/1507.06245v3.
The method which is presented is implemented in the `EstHer` R package, available on the CRAN.

Content

Abstract

Motivated by applications in neuroanatomy, we propose a novel methodology for estimating the heritability which corresponds to the proportion of phenotypic variance which can be explained by genetic factors. Estimating this quantity for neuroanatomical features is a fundamental challenge in psychiatric disease research. Since the phenotypic variations may only be due to a small fraction of the available genetic information, we propose an estimator of the heritability that can be used in high dimensional sparse linear mixed models. Our method consists of three steps. Firstly, a variable selection stage is performed in order to recover the support of the genetic effects – also called causal variants – that is to find the genetic effects which really explain the phenotypic variations. Secondly, we propose a maximum likelihood strategy for estimating the heritability which only takes into account the causal genetic effects found in the first step. Thirdly, we compute the standard error and the 95% confidence interval associated to our heritability estimator thanks to a nonparametric bootstrap approach. Our contribution consists in providing an estimation of the heritability with standard errors substantially smaller than methods without variable selection when the genetic effects are very sparse. Since the real genetic architecture is in general unknown in practice, we also propose an empirical criterion which allows the user to decide whether it is relevant to apply a variable selection based approach or not. We illustrate the performance of our methodology, implemented in the R package EstHer, on synthetic and real neuroanatomic data coming from the Imagen project. We also show that our approach has a very low computational burden and is very efficient from a statistical point of view.

3.1 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:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}\,, \tag{3.1}$$

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 the genetic information matrix, $\mathbf{u}$ 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 (3.1), 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

$$\mathbf{u} \sim \mathcal{N}\left(0, \sigma_u^{\star 2}\mathrm{Id}_{\mathbb{R}^n}\right) \text{ and } \mathbf{e} \sim \mathcal{N}\left(0, \sigma_e^{\star 2}\mathrm{Id}_{\mathbb{R}^n}\right). \tag{3.2}$$

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

$$\eta^\star = \frac{N\sigma_u^{\star 2}}{N\sigma_u^{\star 2} + \sigma_e^{\star 2}} \tag{3.3}$$

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:

$$u_i \overset{i.i.d.}{\sim} (1-q)\delta_0 + q\mathcal{N}(0, \sigma_u^{\star 2}), \text{ for all } 1 \leq i \leq N, \tag{3.4}$$

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:

$$\eta^\star = \frac{Nq\sigma_u^{\star 2}}{Nq\sigma_u^{\star 2} + \sigma_e^{\star 2}}\ . \tag{3.5}$$

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 (3.2), 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 & 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 (3.2), the estimator is consistent (unbiased) under the less restrictive modeling assumption (3.4). 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 & 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 & 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 & 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 & Lv (2008), Ji & Jin (2012), Meinshausen & 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. Since the real genetic architecture is in general unknown in practice, we also propose 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 3.2 describes the data set which motivated our

work. Section 3.3 provides the detailed description of the method, and Section 3.4 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 3.5, 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 3.6, 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 3.2 can be found in Section 3.7. We also provide a discussion section at the end of the paper.

3.2 Description of the data

We worked on data sets provided by the European project Imagen, 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 3.1, 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).

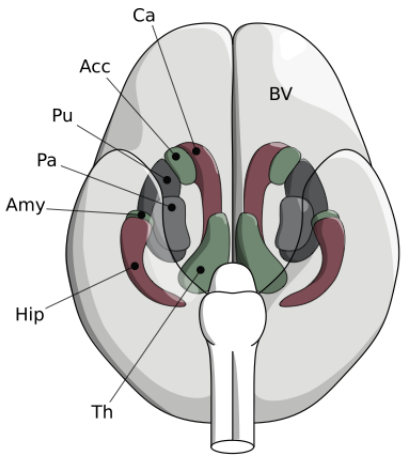


Figure 3.1 – Different regions of the brain (this figure is taken from Toro et al. (2015)).

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

3.3 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

$$\mathbf{Y} = \mathbf{Zu} + \mathbf{e} \tag{3.6}$$

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.

3.3.1 Variable selection

Inspired by the ideas of Fan & 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 & 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 & 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 & Lv (2008) and improved by Ji & 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 & 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:

$$\text{Crit}_\lambda(u) = \|\mathbf{Y} - \mathbf{Z}_{\text{red}} u\|_2^2 + \lambda \|u\|_1 \,, \tag{3.7}$$

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 & 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 & 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 & Bühlmann (2010). In practice, we simulate observations $\mathbf{Y}$ satisfying (3.6), 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 3.4.

3.3.2 Heritability estimation and confidence interval

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 3.3.1 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 $q = 1$,

$$\widetilde{\mathbf{Y}} = \mathbf{U}'\mathbf{Y} \sim \mathcal{N}(0, \Gamma) \text{ with } \Gamma = \mathrm{diag}(\eta^\star \sigma^{\star 2}\lambda_1 + (1 - \eta^\star)\sigma^{\star 2}, \ldots, \eta^\star \sigma^{\star 2}\lambda_n + (1 - \eta^\star)\sigma^{\star 2}), \quad (3.8)$$

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

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

$$L_n(\eta) = -\log\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\widetilde{Y}_i^2}{\eta(\lambda_i-1)+1}\right) - \frac{1}{n}\sum_{i=1}^{n}\log\left(\eta(\lambda_i-1)+1\right) \ , \tag{3.9}$$

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.

Bootstrap confidence interval

We used 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 (3.8) and $\hat{\Gamma}$ has the same structure as Γ defined in (3.8) 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.

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.

3.3.3 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 $\text{Im}(\mathbf{X})$, 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 $\mathbf{A}\mathbf{A}' = \mathbf{P}_{\mathbf{X}}$, $\mathbf{A}'\mathbf{A} = \text{Id}_{\mathbb{R}^{n-d}}$ and $\mathbf{P}_{\mathbf{X}} = \text{Id}_{\mathbb{R}^n} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. This procedure was for instance used by Fan & Li (2012).

3.4 Numerical study

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

3.4.1 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 (3.4) and then we compute a vector of observations such that $\mathbf{Y} = \mathbf{Zu} + \mathbf{e}$.

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 3.4.2 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 3.4.2 when fixed effects are taken into account.

3.4.2 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.

Choice of the threshold

In order to determine the threshold, we apply the procedure described in Section 3.3.1 and 3.3.2. Figure 3.2 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. We will use this value in the following numerical study. However, the way of choosing the threshold will be further discussed, especially in the section dedicated to the study of the genetic data.

Confidence intervals

We use the non parametric boostrap approach described in Section 3.3 in order to compute the confidence intervals associated to the estimations of the heritability. Table 3.1 shows that the 95% confidence intervals obtained by bootstrap and the empirical confidence intervals are very similar. The empirical confidence intervals are computed as follows: the different estimations

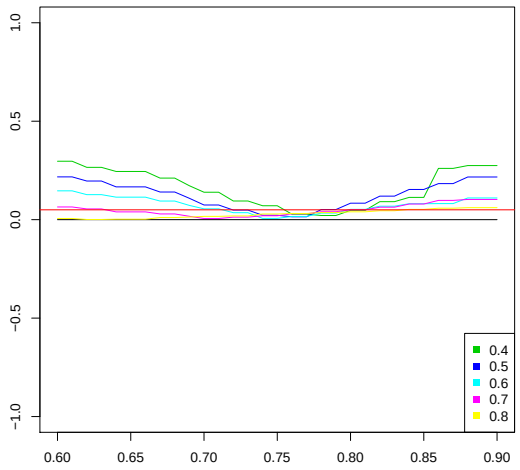


Figure 3.2 – Absolute difference between $\eta^\star$ and $\hat{\eta}$ for thresholds from 0.6 to 0.9 and for $q = 10^{-3}$ (100 causal SNPs).

Table 3.1 – 95 % confidence intervals for $\hat{\eta}$ obtained empirically and by our Bootstrap method.

$\eta^\star$	0.4	0.5	0.6	0.7
Bootstrap	[0.353 ; 0.503]	[0.413 ; 0.565]	[0.494 ; 0.654]	[0.596 ; 0.738]
Empirical	[0.391 ; 0.470]	[0.449 ; 0.542]	[0.496 ; 0.645]	[0.618 ; 0.720]

of $\eta^\star$ obtained along the different replications are ordered, the $\lfloor 0.975 \times M \rfloor$ largest and the $\lfloor 0.025 \times M \rfloor$ smallest values correspond to the upper (resp. lower) bound of the 95% empirical confidence interval. Here, $\lfloor x \rfloor$ denotes the integer part of x and M is the number of replications. From Table 3.1, we can see that the empirical confidence intervals are included in the bootstrap intervals, which means that our approach provides conservative intervals.

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). The results are displayed in Figure 3.3 and in Table 3.2. 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.

Additional fixed effects

We generated some synthetic data according to the process described in Section 3.4.1 but we added a matrix of fixed effects containing two colums. Figure 3.4 (a) displays the corresponding

Table 3.2 – 95 % confidence intervals for $\hat{\eta}$ obtained by our approach, GCTA, 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]

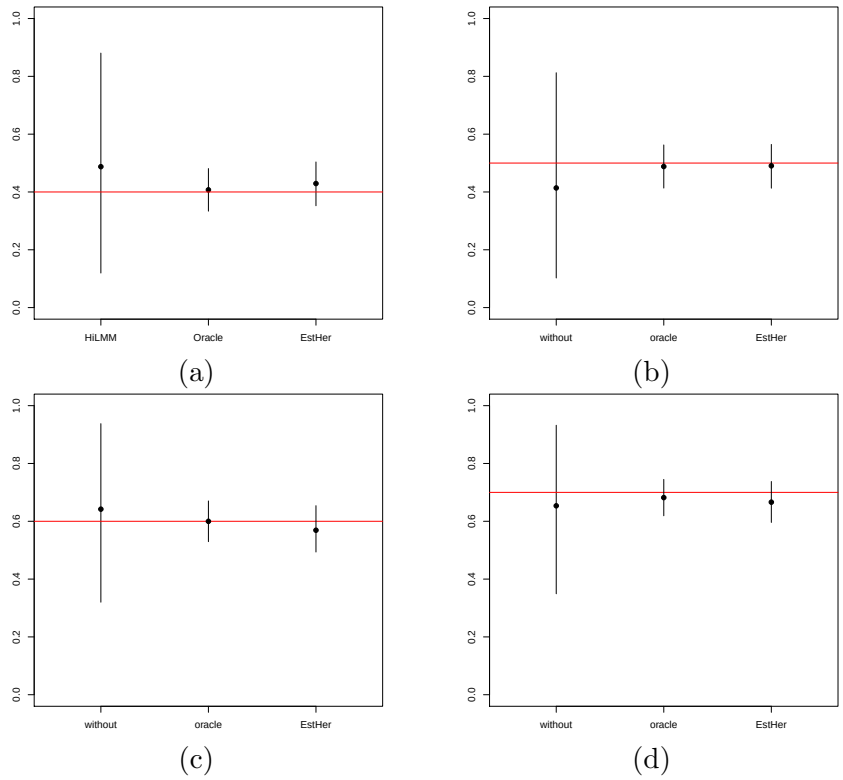


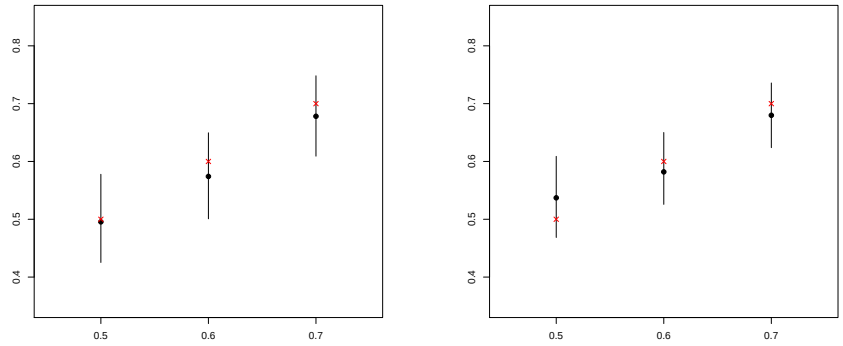
Figure 3.3 – Estimation of the heritability and the corresponding 95% confidence intervals when $q = 10^{-3}$, and for different values of $\eta^\star$: (a) $\eta^\star = 0.4$, (b) $\eta^\star = 0.5$, (c) $\eta^\star = 0.6$, (d) $\eta^\star = 0.7$. The means of the heritability estimators (displayed 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 $\eta^\star$.

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

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 $\mathbf{Z}$. Indeed, in the previous numerical study, we generated a matrix $\mathbf{W}$ with independent columns. The matrix $\mathbf{W}$ 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 3.4 (b). We can see that they are similar to those obtained previously in Figure 3.3, which means that our method does not seem to be sensitive to the presence of correlation between the columns of $\mathbf{W}$.



(a) (b)

Figure 3.4 – 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 $\mathbf{Z}$ 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 $\eta^\star$.

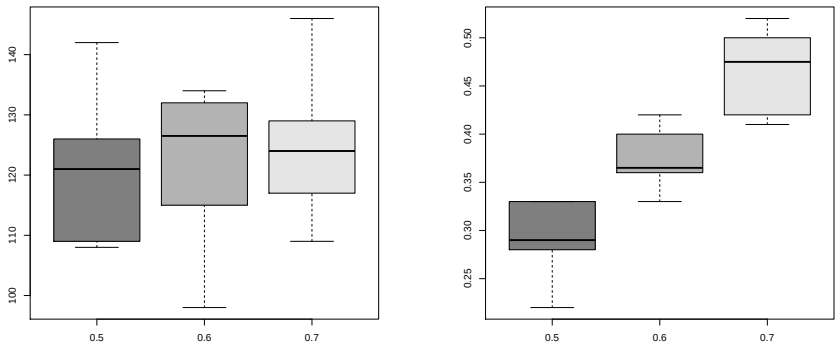
Computational times

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 $\times$ 2.3 GHz.

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 3.5, 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. Nevertheless, even in the worst case, that is $\eta^\star = 0.5$, Figure 3.6 shows that even if we keep only 30% of the real non null components, we select the most active ones.

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



(a) (b)

Figure 3.5 – (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.

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.

3.4.3 Results when the number of causal SNPs is high

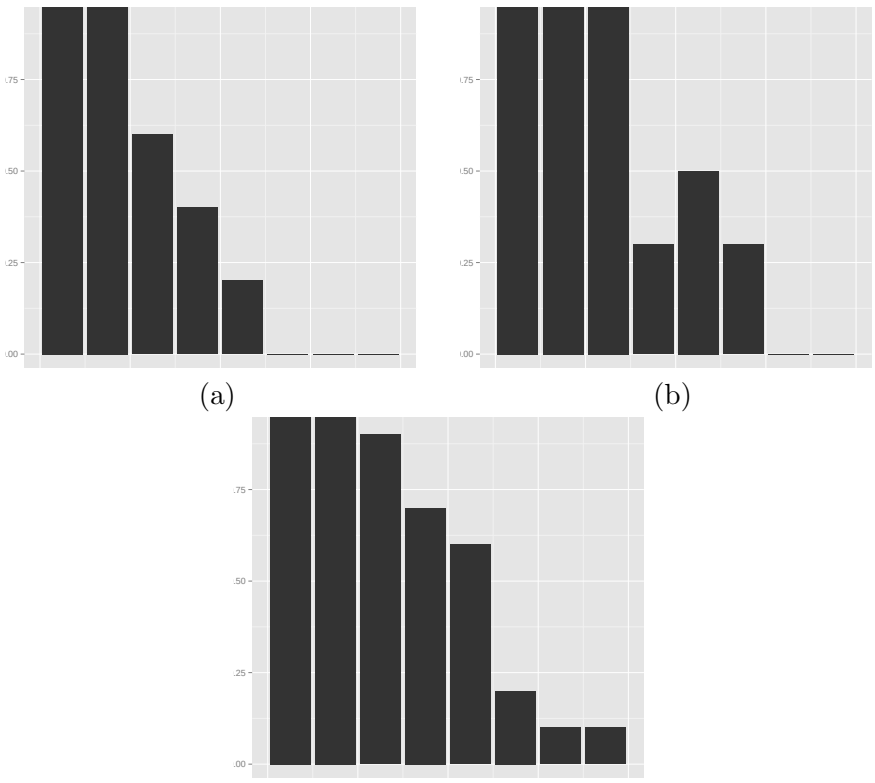
In subsection 3.4.2 we show the performance of our method in the case where the proportion of causal SNPs q is small, that is around 10^{-3}. In this subsection, we focus on a more polygenic scenario, that includes the cases where thousands of SNPs or ten of thousands of SNPs are causal.

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 3.7 that, in this case, EstHer provides unbiased estimations with a small variance.

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 3.8. Moreover, we can see from Figure 3.9 that there is no threshold choice that can provide accurate estimations of heritability for all values of $\eta^\star$.



(a) (b)

(c)

Figure 3.6 – 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).

3.5 A criterion to decide whether we should apply EstHer or HiLMM

On the one hand, we observed that applying HiLMM provides unbiased estimations of the heritability, no matter the number of causal SNPs. However, the main drawback of this estimator is its very large variance. On the other hand, if the number of causal SNPs is not too high, EstHer provides unbiased estimations of the heritability with standard errors substantially smaller than HiLMM. However, if the number of causal SNPs is high, EstHer underestimates the heritability. These observations are similar to those made by Zhou et al. (2013), who built an hybrid estimator able to deal with both sparse and non sparse scenario, to which we will compare our approach in Section 3.6. Therefore, we propose hereafter a rule to decide whether it is better to apply EstHer or HiLMM. We can see from Figure 3.2 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 3.9), 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

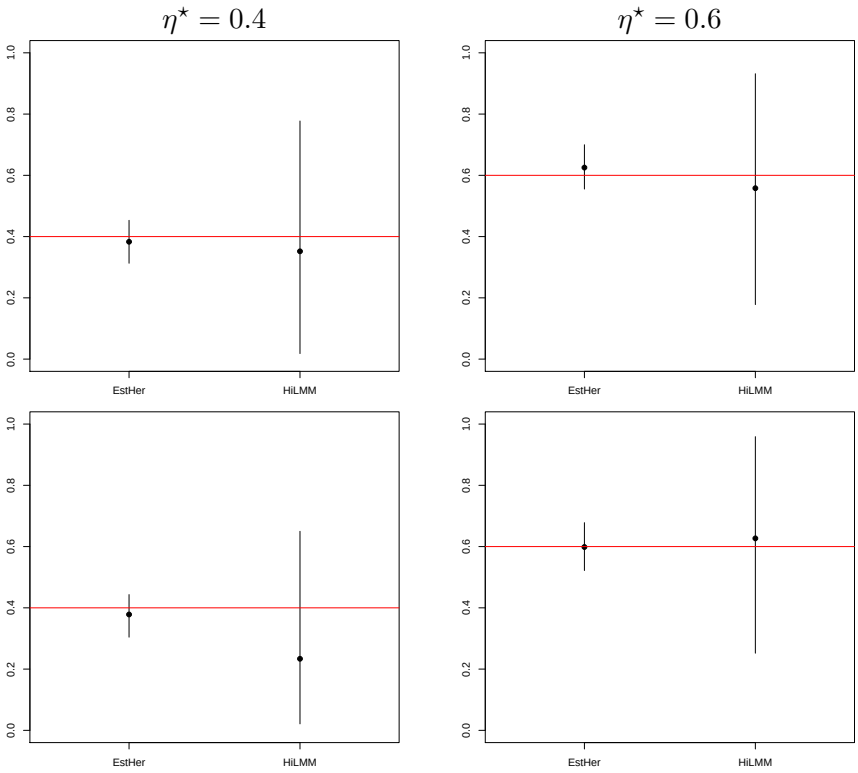


Figure 3.7 – Results of HiLMM and EstHer when there are a few causal SNPs with moderate effects and a lot of SNPs with small effects. The proportion of each is 100 out of 1000 (up) and 100 out of 10000 (bottom), with $\eta^\star = 0.4$ and 0.6.

thresholds for which the confidence intervals overlap. Figure 3.10 confirms the stability around the best threshold for different values of $\eta^\star$ and Table 3.3 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. The results obtained by using this criterion are displayed in Figure 3.11.

3.6 Results after applying the decision criterion and comparison to other methods

Table 3.3 – 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

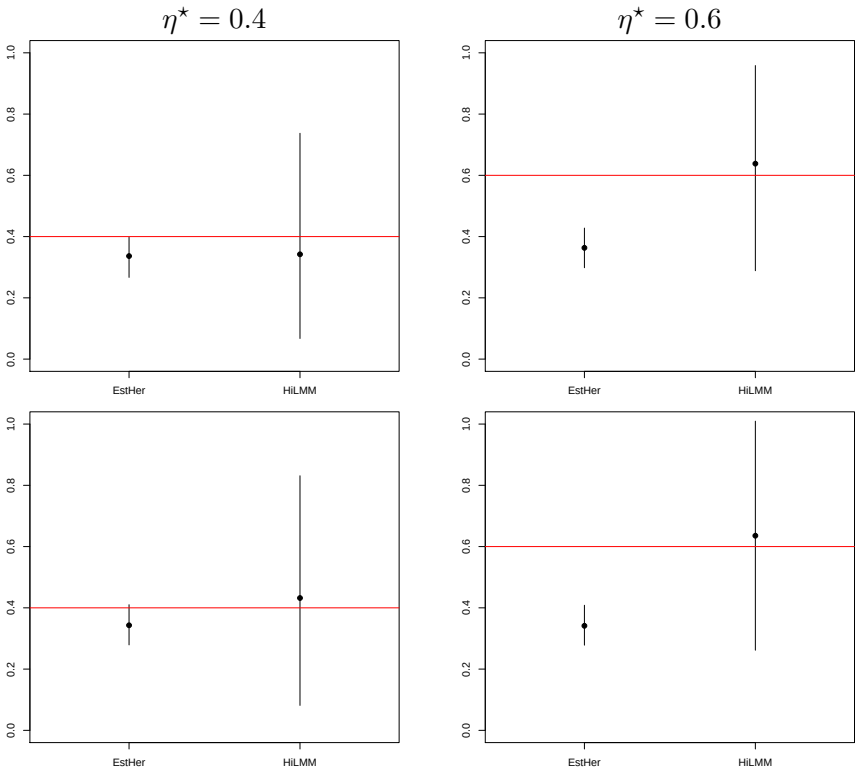


Figure 3.8 – Results of HiLMM and EstHer for 1000 (up) and 10000 (bottom) causal SNPs and for $\eta^\star = 0.4$ and 0.6.

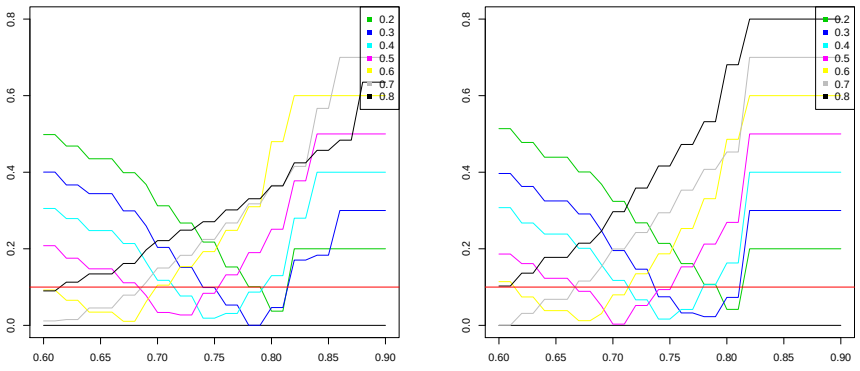


Figure 3.9 – Absolute difference $|\eta^\star - \hat{\eta}|$ for thresholds from 0.6 to 0.9 and for 1000 (left) and 10000 (right) causal SNPs.

3.6.1 Statistical performances

In this section we show the results obtained after applying the criterion described in Section 3.5. We compare these results to those obtained using HiLMM, but also with the software GEMMA described in Zhou & 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

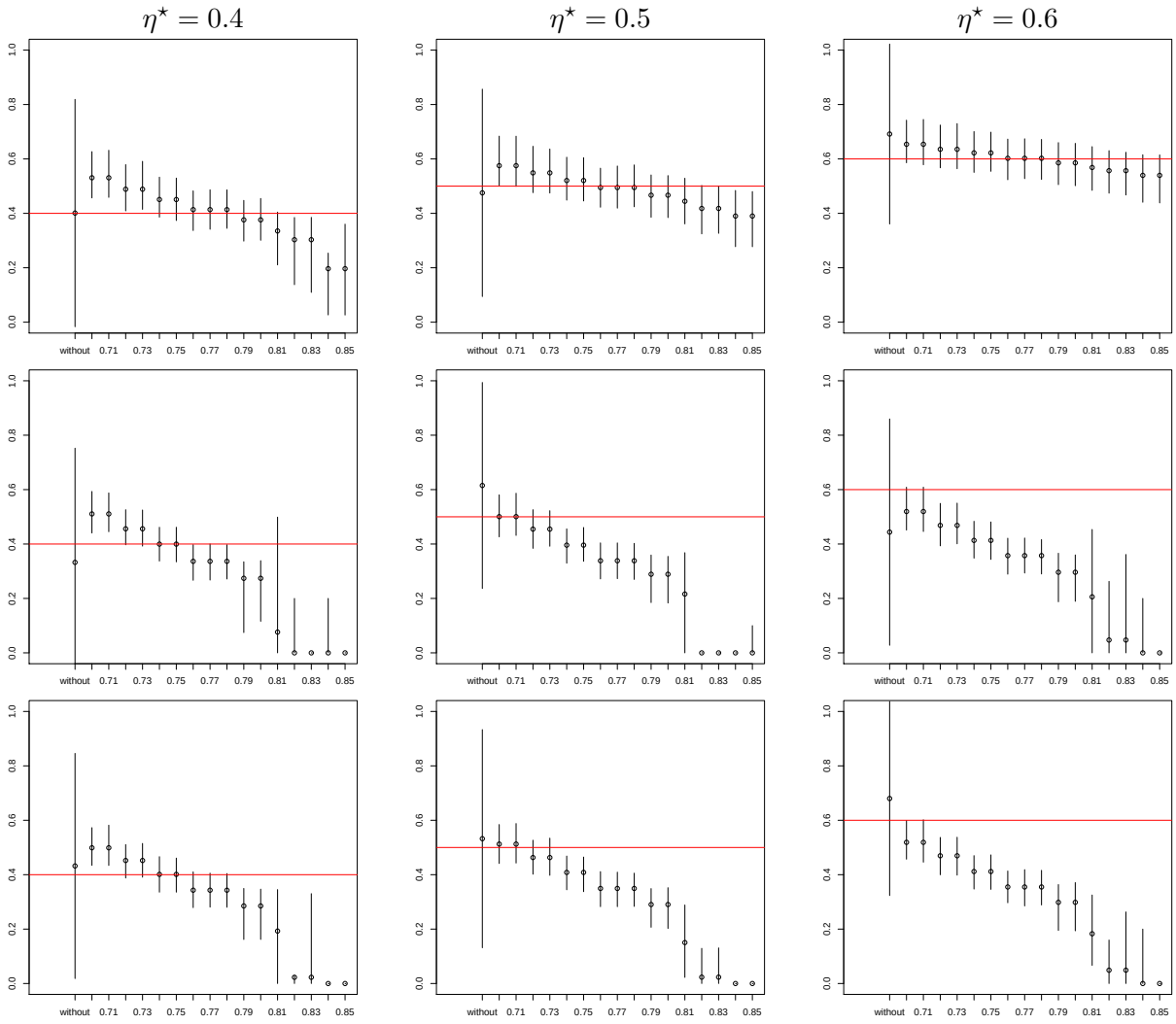


Figure 3.10 – Estimation of the heritability with 95% confidence intervals for $\eta^\star$ from 0.4 to 0.6 (from left to right), and from 100, 1000 and 10000 causal SNPs from top to bottom. Each graph shows the heritability estimations with 95% confidence intervals computed with HiLMM ("without") and for thresholds between 0.7 and 0.85.

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 3.11 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 3.11 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.

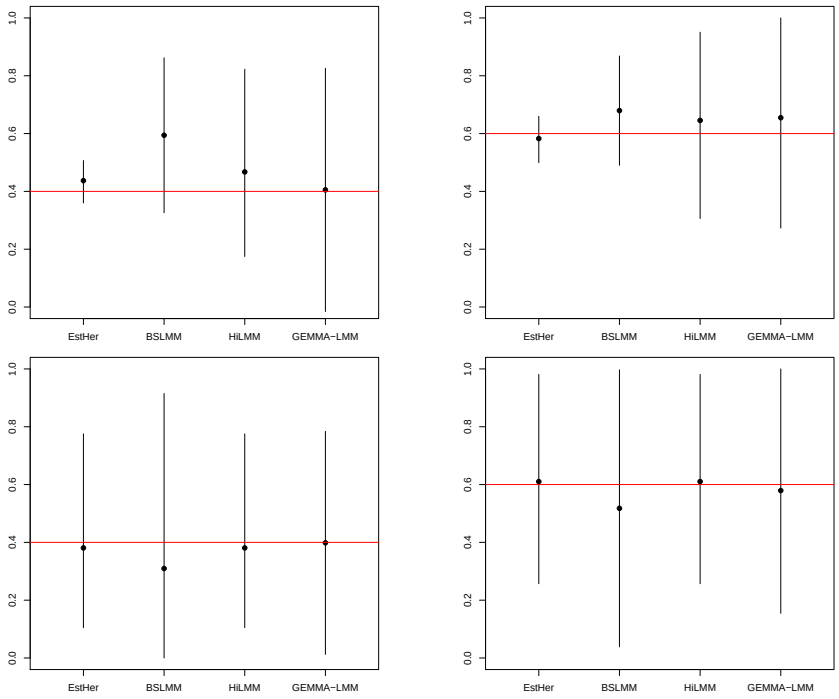


Figure 3.11 – Estimations of $\hat{\eta}$ with 95 % confidence intervals obtained using EstHer, BSLMM, HiLMM and GEMMA-LMM with 100 causal SNPs (top) and 10,000 causal SNPs (bottom). The results are obtained with 10 replications.

3.6.2 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 3.12. We chose this number of thresholds since we applied the criterion defined in Section 3.5. 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.

3.7 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 3.2.

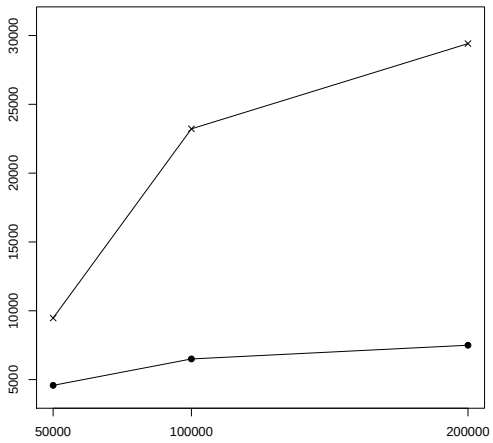
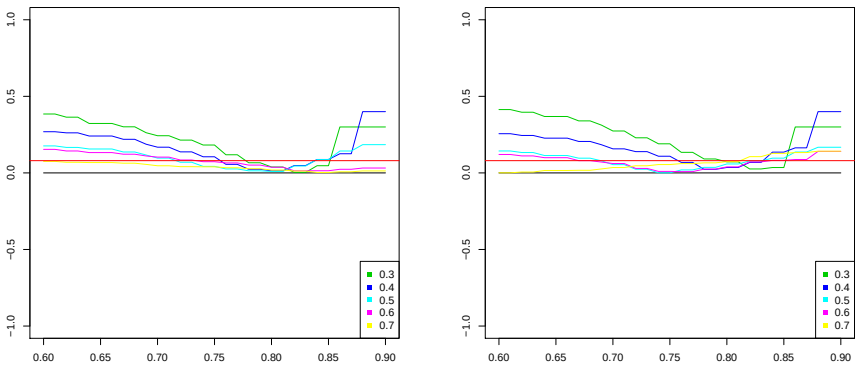


Figure 3.12 – Times (in seconds) to compute one heritability estimation with BSLMM (crosses) and EstHer (dots) by using 16 thresholds for $n = 2000$ and different values of N from $50,000$ to $200,000$.

3.7.1 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 3.4.2: for several values of $\eta^\star$ and several thresholds, we display the absolute value of $\eta^\star - \hat{\eta}$, see Figure 3.13. The only difference with Section 3.4.2 is that we generated the observations by using the matrix $\mathbf{W}$ coming from the IMAGEN data set. According to Figure 3.13, 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)

Figure 3.13 – 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 3.4 – 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

3.7.2 Application of the decision criterion

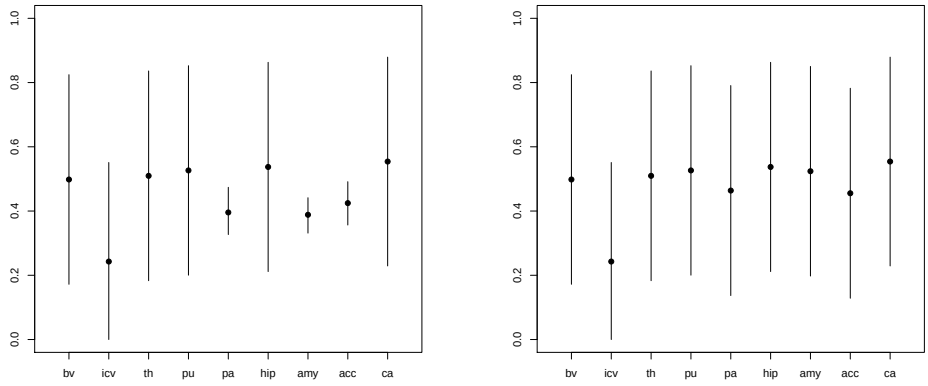
Since we determined in the previous section that the optimal threshold is 0.79, we apply EstHer for thresholds around this value, that is from 0.7 to 0.85. We then count the number of overlapping confidence intervals, as explained in Section 3.5. The results are displayed in Table 3.4. We observe from 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.

3.7.3 Results

Figure 3.14 (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 3.14 (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 3.14 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.

3.8 Conclusion

We propose in this paper a practical method to estimate the heritability in sparse linear mixed models using variable selection tools, as well as confidence interval obtained thanks to a non parametric bootstrap approach. Our approach is implemented in the R package EstHer which is available from the Comprehensive R Archive Network (CRAN) and from the web page of the first author. In the course of this study, we showed that our approach has two main features



(a) (b)

Figure 3.14 – (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.

which makes it very attractive. Firstly, it is very efficient from a statistical point of view since it provides confidence intervals considerably smaller than those obtained with methods without variable selection. Secondly, its very low computational burden makes its use feasible on very large data sets coming from quantitative genetics.

Moreover, we observed that the statistical performance of the EstHer approach are all the more impressive that the level of sparsity is high that is when q is small. For this reason, we also proposed an empirical criterion which allows the user to decide whether it is better to apply an approach that takes into account the sparsity and starts with a variable selection stage, namely EstHer, or an approach which ignores the potential sparsity in the observations, namely HiLMM.

Chapter 4

Application: heritability estimation of the size of juvenile trouts

The work presented in this chapter comes from a collaboration with Niklas Tysklind, who is a researcher in the UMR EcoFoG, INRA in Kourou. This work will be soon submitted to an international journal of ecology.

Content

4.1 Data and motivation

We have a dataset containing the size of 142 juvenile trouts, the genotype of which is described by 3917 SNPs. We also have several possible explaining variables: the river they lived in, the month the trouts have been captured, and several features regarding the environmental conditions: latitude, longitude, alkalinity, hardness, catchment area (it gives an idea of the potential amount of environment available for the sea trout), shreve, strahler (both are complexity values of the river in terms of how many affluents and branches each river has). Finally we have an additional

information called the "mean +0 and > 0+" juvenile densities, which are included as a proxy for the effect of competition between the juvenile trouts.

Our goal to estimate the heritability of the trouts size, that is the proportion of variability of the size which can be explained by genetic factors.

4.2 Model and method

4.2.1 Model

We propose to use the following sparse linear mixed model:

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

where

- $\mathbf{Y}$ is a vector of length 142 which contains the observations of the size of the trouts.
- $\mathbf{X}$ is a matrix associated to the fixed effects: it contains the explaining variables (river, month...).
- $\boldsymbol{\beta}$ corresponds to the coefficients associated to $\mathbf{X}$.
- $\mathbf{Z}$ is a $n \times N$ matrix where $Z_{i,j}$ is the value of the SNP j of the individual i.
- $\mathbf{u}$ and $\mathbf{e}$ are the random effects. We assume that

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

and since we do not know the SNPs which may have an effect on the observations, we shall assume that

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

where $\mathbf{u} = (u_1, \ldots, u_n)$.

4.2.2 Method

The method that we used is described in Chapter 3. However, the first step of the procedure, called the Sure Independance Screening (SIS), is specific to ultra high dimensional datasets. According to numerical results obtained with and without the SIS step, which are not presented here, we propose to skip the first step of our method.

The results of this approach are shown in the following simulation study.

4.3 Simulation Study

We propose a simulation study in the following framework: $n = 200$, $N = 4000$, which are close to the size of the trouts dataset that we want to analyze. We assume that the number of causal SNPs varies from 4 ($q = 10^{-3}$) to 400 ($q = 0.1$).

4.3.1 Results obtained without selection

The estimations obtained using our method HiLMM, described in Chapter 2, are shown in Figure 4.1. The large empirical variance of this estimator motivates an additional variable selection step before estimating heritability.

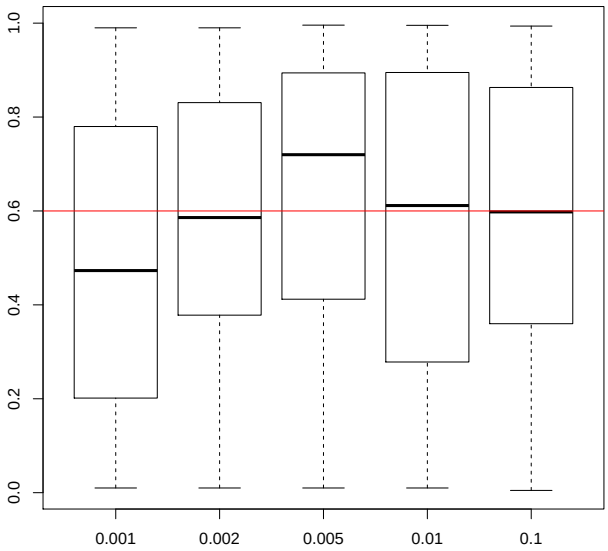


Figure 4.1 – Heritability estimations obtained from 100 replications without selection for $\eta^\star = 0.6$ and different values of q from 0.001 to 0.1.

4.3.2 Calibration of the threshold in the stability selection

As explained in the detailed procedure in Chapter 3, our method requires to determine the optimal threshold in the stability selection. For this purpose, we apply the procedure described in Section 4.2.2. Figure 4.2 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.

When the number of causal SNPs is small (namely, smaller than 20), applying our selection method with a threshold around 0.35 seems to provide accurate estimations for different values of the heritability $\eta^\star$. Otherwise, there is no common threshold that provides accurate estimations for all values of the heritability $\eta^\star$. In this scenario, we shall estimate directly heritability without performing a variable selection approach.

4.3.3 Results regarding heritability estimation

We observe firstly from Figure 4.3 that when the number of causal SNPs is small (namely, smaller than 20), applying our selection method substantially improves the heritability estimations compared to the method without selection. Secondly, when the number of causal SNPs is high, both methods have similar performances.

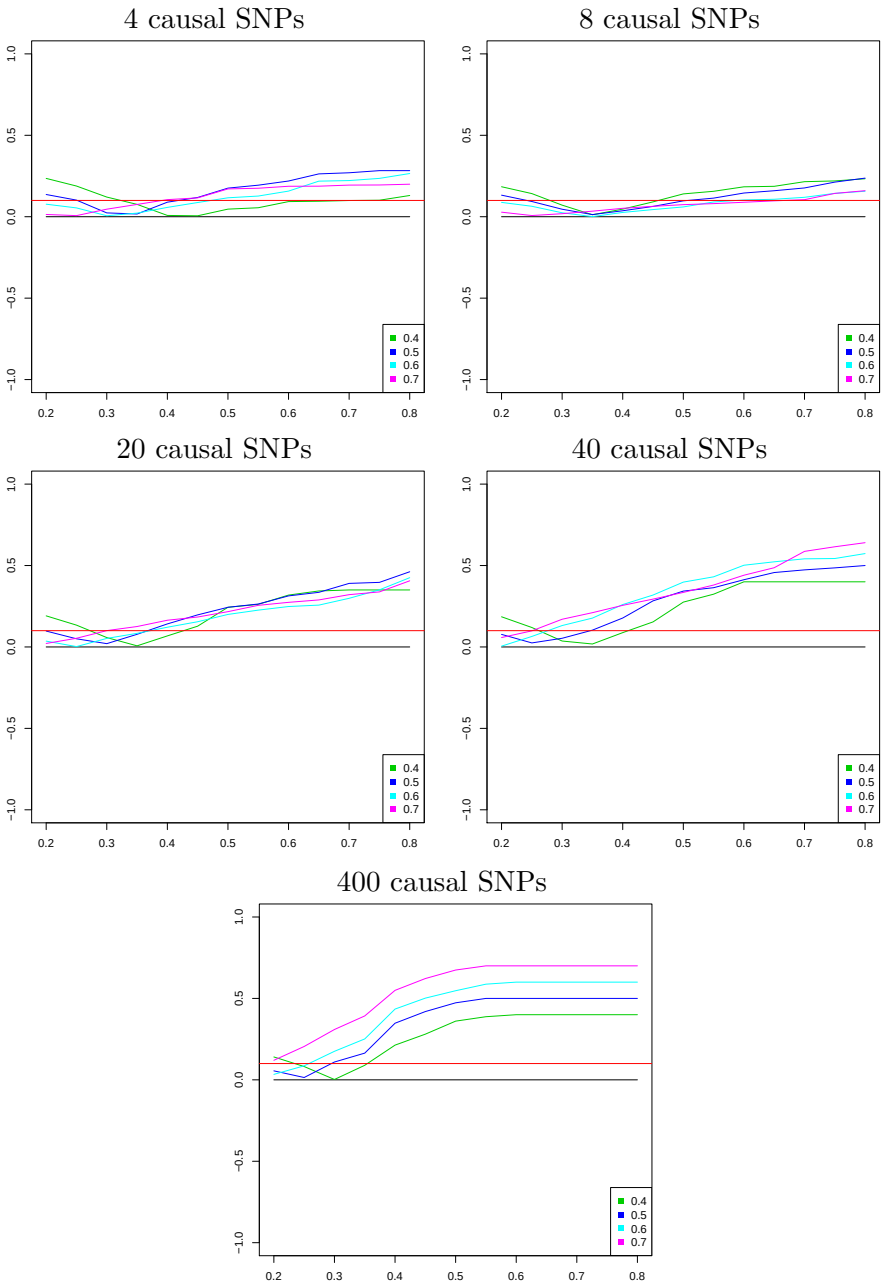


Figure 4.2 – Absolute difference $|\eta^\star - \hat{\eta}|$ for thresholds from 0.2 to 0.8 and for different numbers of causal SNPs, for different values of $\eta^\star$: 0.4 (green), 0.5 (dark blue), 0.6 (blue) and 0.7 (purple).

4.3.4 Support recovery when the selection is applied

When the number of causal SNPs is small, our criterion leads us to apply the variable selection approach. We focus on the support that we captured with our method: Figure 4.4 shows the

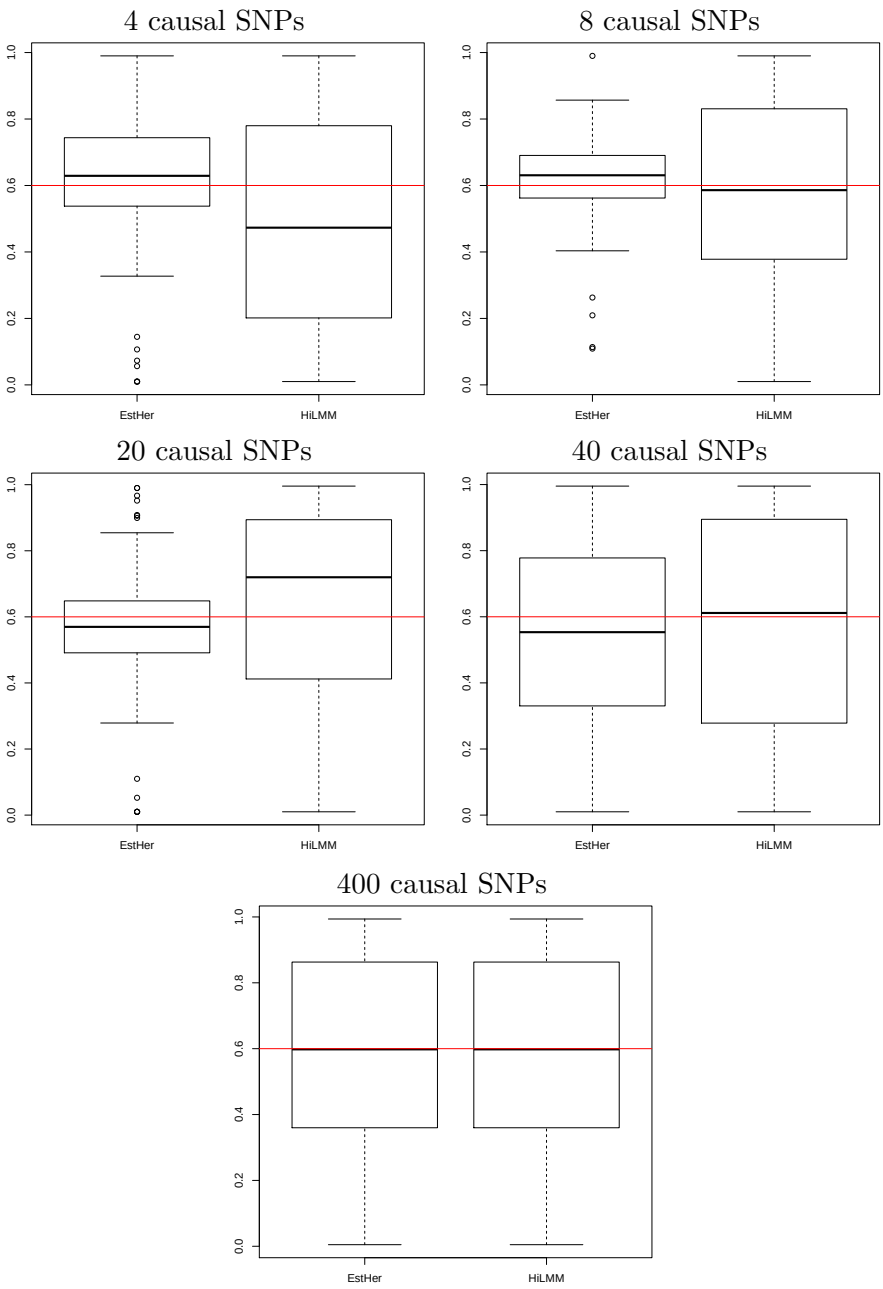


Figure 4.3 – Estimated values of $\hat{\eta}$ obtained with our variable selection method EstHer with a threshold of 0.35 and without selection (HiLMM) for different numbers of causal SNPs.

proportion of the recovered support and the number of variables selected by our method. We notice that choosing a threshold around 0.35 seems to be a relevant trade-off between a large true positive rate and a small false positive rate of non zero recovered components. We recover indeed at least 30% of the support and up to 60% when the number of causal SNPs is very small

(namely, 4 or 8 SNPs).

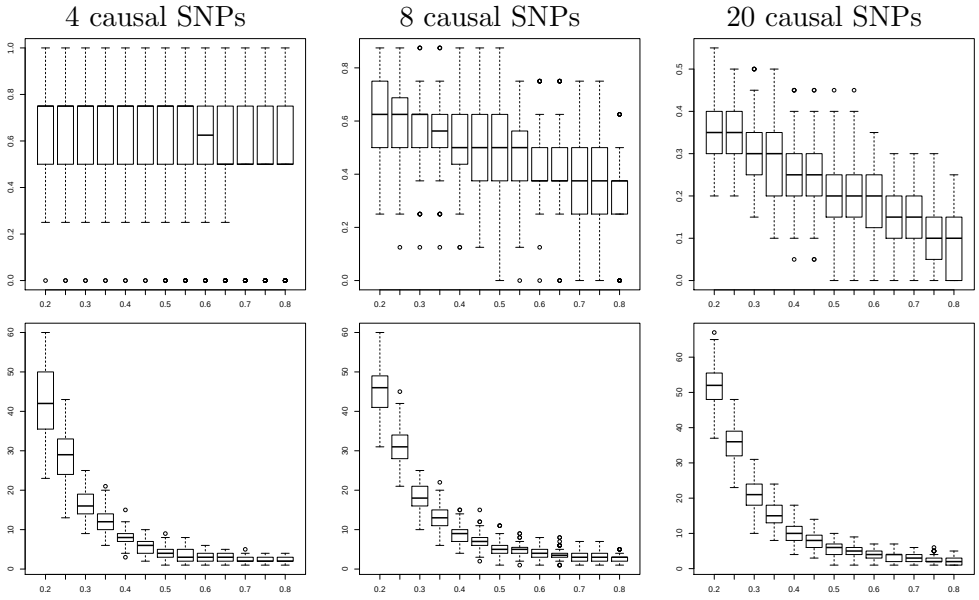


Figure 4.4 – Proportion of the support recovered by our selection method Esther (top) and number of selected variables (bottom) for different numbers of causal SNPs and for different thresholds.

4.4 Results on the juvenile size dataset

4.4.1 Results obtained without taking fixed effects into account

First we propose to estimate the heritability without taking any fixed effect into account. Figure 4.5 shows firstly that the heritability of the trouts size is very high (80%) despite the large confidence interval (40-100%). Moreover, we can see that EstHer and HiLMM provide the same estimation, which suggests that the number of causal SNPs must be high.

4.4.2 Including fixed effects

We estimated the heritability when considering as a fixed effect the river where the trouts were captured in. The results are shown in Figure 4.6: including the river as a fixed effect provides a higher heritability estimation (0.93 instead of 0.79) and more accurate (the length of the confidence interval is substantially smaller). It means that the river and the SNPs explain a very large proportion of the size variations. Including the month as a fixed effect does not seem to have a relevant effect: a possible explanation is that the month effect was already taken into account in the river effect, since all trouts of the same river have been captured the same month.

Then, we considered phylogeographic groups instead of considering the specific river: the 30 rivers are contained in 9 phylogeographic groups with similar environmental features. The

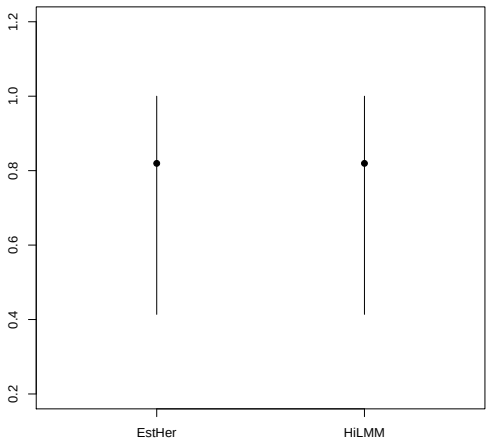


Figure 4.5 – Estimated value of the heritability with 95 % confidence intervals, using EstHer and HiLMM.

results using these phylogeographic groups as a fixed effect are displayed in Figure 4.6. It seems that these groups bring less information than the rivers, but if we join the groups with the month effect, we have similar results than those obtained with either the river or the combination river and month. This could mean that the information we lost when considering the groups instead of the rivers was contained in the month effect.

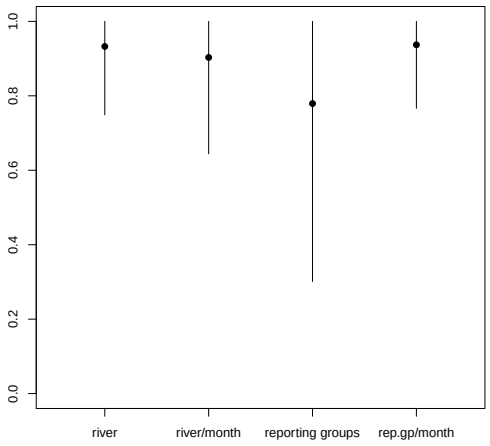


Figure 4.6 – Estimated value of the heritability using different fixed effects, from left to right: river, river/month, reporting groups, reporting groups and month.

Then, we wonder if we can determine the specific features of the river that can explain the increasing proportion of variability explained when knowing the river. Figure 4.7 shows the results of considering different groups of fixed effects. The groups are the following:

- Group 1: longitude, latitude, catchment area, shreve, strahler

- Group 2: longitude, latitude, catchment area, shreve, strahler, alkalinity, hardness
- Group 3 : longitude, latitude, catchment area, shreve, strahler, mean 0+ and mean > 0+

These groups were chosen according to the availability of the data: indeed, in the first group we have all the information for the 142 trouts. Alkalinity and hardness were missing for 12 individuals, and mean 0+ and mean > 0+ for 16 individuals (unfortunately, not the same than alkalinity and hardness).

The results are displayed in Figure 4.7. We see that Group 1 explains a proportion of variability bigger than no fixed effects but smaller than the river. Same result when we add mean 0+ and mean > 0+. The confidence interval increases when considering the additional effect of alkalinity and hardness. However, this could possibly be the effect of removing individuals from the study. This assumption is consistent with Figure 4.8, which shows the results obtained with the variables of Group 1 but without the individuals that we had to remove when we studied Group 2, compared to the results obtained with Group 2. Similarly, Figure 4.9 shows the results obtained with the variables of Group 1 but without the individuals that we had to remove when we studied Group 3, compared to the results obtained with Group 3. We see a substantial difference in the estimations, which suggests that mean 0+ and mean > 0+ could be relevant variables.

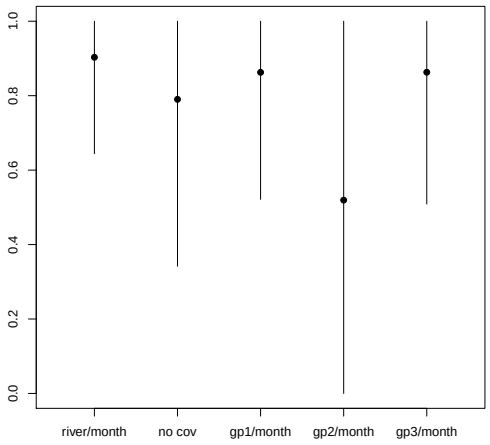


Figure 4.7 – Estimated value of the heritability with different covariates from left to right: river/month, no covariates, group 1/month, group 2/month, group 3/month

4.5 Conclusion

- The length variations have an important genetic component (≥40%).
- This genetic component might be due to large number of causal SNPs (at least, more than 20 according to the simulation study), which makes impossible the recovery of these causal SNPs.
- Including fixed effects can refine the heritability estimations: the river seems indeed to be a key factor in the length variation. Once the river has been taken into account, more

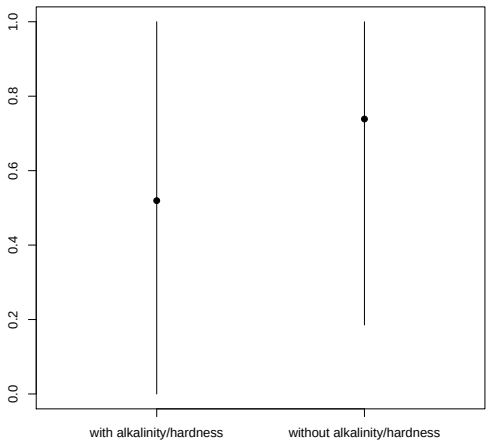


Figure 4.8 – Estimated value of the heritability with and without alkaninity and hardness as covariates, but removing 12 individuals from the study.

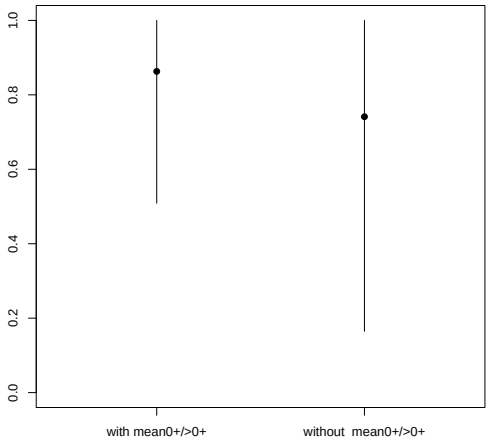


Figure 4.9 – Estimated value of the heritability with and without mean 0+ and mean $> 0+$ as fixed effects, but removing 16 individuals from the study.

than 75% of the remaining variation is due to genetic effects. However, it is important to notice that we cannot separate entirely the effects of the month and the river.

- The reporting groups alone do not improve the heritability estimations, but associated to the month effect, they explained as much variability as either the river, or the combined effect river/month.
- In the environmental factors, the combined effect of catching area, longitude, latitude, shreve, strahler explain more than no fixed effects, but less than the river.

Chapter 5

Heritability estimation in case-control studies

The work contained in this chapter will be soon submitted to a journal of statistics.

Content

Abstract

In the genetic field, the concept of heritability refers to the proportion of variations of a biological trait or disease that can be explained by genetic factors. Quantifying the heritability of a disease is a fundamental challenge in human genetics. Although the litterature regarding heritability estimation for binary traits is less rich than for quantitative traits, several methods have been proposed to estimate the heritability of diseases. However, to the best of our knowledge,

the existing methods raise at least one of the two concerns (generally both): either they have not been validated theoretically or they are severely affected by the oversampling of the number of patients compared to controls in a medical study. We propose in this paper to investigate the theoretical properties of the method developed by Golan et al. (2014), which is very efficient in practice, despite the oversampling of patients. Our main result is the proof of the consistency of this estimator. We also provide a numerical study to compare two approximations leading to two heritability estimators.

5.1 Introduction

In the genetic field, the concept of heritability refers to the proportion of variations of a biological trait or disease that can be explained by genetic factors. Quantifying the heritability is a major information for diseases that are suspected to have a strong genetic component but which causes are often vague and multiple. Indeed, determining a high value of heritability for a disease is a powerful argument in favor of further research for genetic causes, but it also opens the possibility of predicting a risk of illness based on the genetic background.
There exist several methods to estimate the heritability of quantitative traits, which we will describe hereafter, with interesting theoretical and practical properties. Regarding binary traits, such as the presence or absence of a disease, a few methodologies have been proposed, but as far as we know, most of them have not been validated theoretically. Golan et al. (2014) developed a method to estimate heritability of binary traits that they compared to recent methodologies and which was shown to be very efficient in practice. The aim of this paper is to investigate the theoretical properties of Golan et al. (2014)'s method.
Let us first recall the main existing methods to estimate heritability of quantitative traits, which will be strongly linked to the methods used for binary traits. Linear Mixed Models (LMMs) have been widely used for estimating the heritability of quantitative traits. Indeed, Yang et al. (2010) proposed for instance to estimate the heritability of human height by using a classical LMM defined by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{5.1}$$

where $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)'$ is the vector of observations of a phenotype of interest, $\mathbf{X}$ is a $n \times p$ matrix of predictors (or fixed effects), $\boldsymbol{\beta}$ is a $p \times 1$ vector containing the unknown linear effects of the predictors, and $\mathbf{u}$ and $\mathbf{e}$ correspond respectively to the genetic and the environmental random effects. We assume that $\mathbf{u}$ and $\mathbf{e}$ are Gaussian random effects with variances $\sigma_u^{\star 2}$ and $\sigma_e^{\star 2}$ respectively. Moreover, $\mathbf{Z}$ is a $n \times N$ matrix which contains the genetic information. They proposed to estimate the parameter

$$\eta^\star = \frac{N\sigma_u^{\star 2}}{N\sigma_u^{\star 2} + \sigma_e^{\star 2}}, \tag{5.2}$$

commonly considered as the mathematical definition for heritability since it determines how the variance is shared between $\mathbf{u}$ and $\mathbf{e}$.

Several methods were developed to estimate the parameter $\eta^\star$, see Patterson & Thompson (1971), Searle et al. (1992), Yang et al. (2011), Pirinen et al. (2013), Zhou & Stephens (2012). From a theoretical point of view, Bonnet et al. (2015) showed the asymptotic normality of the maximum likelihood estimator of $\eta^\star$ as well as a central limit theorem leading to confidence

intervals for $\eta^\star$.

All these methods obviously cannot be applied directly for binary traits, but there exist extensions by assuming an underlying Gaussian variable linked to the binary phenotype. There are two different modelings which connect binary phenotypes to a continuous quantity called the liability.
The first one consists in assuming that the observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are distributed according to the following Generalized Linear Mixed Model (GLMM):

$$\mathbf{Y}_i \sim \mathcal{B}(p_i) \tag{5.3}$$

with $p_i = g(\mathbf{l}_i)$ where g is a link function and $\mathbf{l}_i$ is defined as

$$\mathbf{l} = \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{5.4}$$

with $\mathbf{l} = (\mathbf{l}_1, \ldots, \mathbf{l}_n)$, $\mathbf{u} \sim \mathcal{N}(0, \sigma_u^{\star 2})$ and $\mathbf{e} \sim \mathcal{N}(0, \sigma_e^{\star 2})$, like in classical LMM defined in Equation (5.1). The heritability is then defined "at the liability scale", which means for the continuous variable $\mathbf{l}$, and is given by the same expression (5.2) as for quantitative traits.
Several methods were established to estimate heritability in Model (5.3): among them we can quote the MCMC method of Hadfield (2010) and the penalized quasi-likelihood approach of Breslow & Clayton (1993). The numerical performance of these methods can be found in the comparative study of de Villemereuil et al. (2013).

Another modeling and definition for the heritability of a binary trait, which is more frequently used than the previous one, was proposed by Falconer (1965), who assumed that the binary observations could be seen as an indicator function of a Gaussian variable exceeding a given threshold t:

$$\mathbf{Y}_i = \mathbb{1}_{\{\mathbf{l}_i > t\}}, \tag{5.5}$$

with $\mathbf{l}_i$ defined by the same expression (5.4) than in Model (5.3). Observe that the threshold t is directly linked to the prevalence of the disease in the population, that is the proportion K of the population which is affected by the disease. Indeed,

$$K = \mathbb{P}(\mathbf{Y}_i = 1) = \mathbb{P}(\mathbf{l}_i > t). \tag{5.6}$$

The unobserved Gaussian variable $\mathbf{l} = (\mathbf{l}_1, \ldots, \mathbf{l}_n)$ is also called the liability in this modeling, which is usually called the "liability model" (Falconer (1965), Lee et al. (2011), Tenesa & Haley (2013)) and has been shown to be a reasonable modeling for complex diseases, for instance by Purcell et al. (2009). The heritability is then also defined as the heritability at the liability scale as in Equation (5.2).

Regarding the procedures based on the second modeling defined in Equations (5.5) and (5.4), Lee et al. (2011) proposed to use a maximum likelihood approach as if the binary traits were Gaussian, and then to apply a multiplicative factor to correct this approximation. Golan et al. (2014) showed that this heritability estimator was strongly biased in several realistic scenarios, in particular it was very sensitive to the prevalence of the disease, which is the proportion of the

population that is affected by the disease (when the disease is rarer, the bias increases). The estimator also underestimates the heritability when the true heritability is high.
Weissbrod et al. (2015) introduced a maximum likelihood based strategy to rebuild the underlying liability before estimating the heritability.
However, all these methods do not take into account an essential element of case-control studies: indeed, in a medical study, the number of patients is similar to the number of controls even though the studied disease might be rare, which means that the proportion of cases in the study does not reflect the proportion of cases in the population. This oversampling of the cases has been noticed and handled by the approach of Golan et al. (2014), who proposed a moment based method to estimate the heritability. They computed an approximate quantity of the expectation of $\mathbf{W}_i\mathbf{W}_j$, for two individuals i and j, $\mathbf{W}_i$ being a centered and normalized version of the binary data $\mathbf{Y}_i$, and conditionally to the fact that individuals i and j are in the study.

Since the method of Golan et al. (2014) presented very good numerical results but was not supported by theoretical grounds, we propose in this paper a method which is strongly inspired from the one proposed by Golan et al. (2014): we obtained two approximations (depending on the order of the approximation) of the expectation E of $\mathbf{W}_i\mathbf{W}_j$ conditionally to the fact that both individuals are in the study. We show that the first order approximation provides an estimator with good theoretical properties: indeed, we prove that it is a consistent estimator of $\eta^\star$. We also propose a simulation study to compare the numerical performances of the estimators obtained with both approximations.
The model we study and the main definitions are given in Section 5.2. Section 5.3 contains the first order approximation of the expectation E with the corresponding estimator of $\eta^\star$ and Section 5.4 presents our consistency result for this estimator. The second order approximation of E is given in Section 5.5 and the numerical comparison of the two estimators can be found in Section 5.6. In Section 5.7, we discuss the results and potential perspectives. Finally, the proofs are given in Section 5.8.

5.2 Model and definitions

5.2.1 Liability model

Let us denote K the prevalence of a disease in a population, that is the proportion of the population affected by the disease. Let $\mathbf{Y}_i$ be the random variable such that $\mathbf{Y}_i = 1$ if the individual i is ill (then, individual i is called a case) and $\mathbf{Y}_i = 0$ if the individual i is healthy (then individual i is called a control). We assume that the $\mathbf{Y}_i$'s are linked to unobserved variables $\mathbf{l}_i$ as follows

$$\mathbf{Y}_i = \mathbb{1}_{\{\mathbf{l}_i > t\}}, \tag{5.7}$$

where t is a given threshold, related to the prevalence K by (5.6), and the $\mathbf{l}_i$'s are defined as

$$\mathbf{l} = \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{5.8}$$

where $\mathbf{l} = (\mathbf{l}_1, \dots, \mathbf{l}_n)$, $\mathbf{u}$ and $\mathbf{e}$ are random effects such that $\mathbf{u} \sim \mathcal{N}(0, \sigma_u^{\star 2}\mathrm{Id}_{\mathbb{R}^N})$ and $\mathbf{e} \sim \mathcal{N}(0, \sigma_e^{\star 2}\mathrm{Id}_{\mathbb{R}^n})$. The vector $\mathbf{u}$ corresponds to the genetic effects and $\mathbf{e}$ to the environmental effects. Moreover, $\mathbf{Z}$ is a $n \times N$ random matrix which contains the genetic information, and

which is such that the $\mathbf{Z}_{i,k}$ are normalized random variables in the following sense: they are defined from a matrix $A = (A_{i,k})_{1\leq i\leq n,\, 1\leq k\leq N}$ by

$$\mathbf{Z}_{i,k} = \frac{A_{i,k} - \overline{A}_k}{s_k},\ i = 1, \ldots, n,\ k = 1, \ldots, N\ , \tag{5.9}$$

where

$$\overline{A}_k = \frac{1}{n}\sum_{i=1}^{n} A_{i,k},\ s_k^2 = \frac{1}{n}\sum_{i=1}^{n}(A_{i,k} - \overline{A}_k)^2,\ k = 1, \ldots, N\ . \tag{5.10}$$

In (5.9) and (5.10) the $A_{i,k}$'s are such that for each k in $\{1, \ldots, N\}$ the $(A_{i,k})_{1\leq i\leq n}$ are independent and identically distributed random variables and such that the columns of A are independent.
In practice, the matrix A contains the genetic information about all the individuals in the study. More precisely, for each k, $A_{i,k} = 0$ (resp. 1, resp. 2) if the genotype of the ith individual at locus k is qq (resp. Qq, resp. QQ). In this paper, we consider a more general case with mild assumptions on the distribution of the random variables $A_{i,k}$, which are described in Section 5.4.
With the definition (5.9), the columns of $\mathbf{Z}$ are empirically centered and normalized, and one can observe that

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

The heritability at the liability scale, which is the parameter we want to estimate, is defined as the ratio of variances:

$$\eta^\star = \frac{N\sigma_u^{\star 2}}{N\sigma_u^{\star 2} + \sigma_e^{\star 2}}. \tag{5.11}$$

The variance of $\mathbf{l}$ conditionally to $\mathbf{Z}$ can then be rewritten with respect to $\eta^\star$ and $\sigma^{\star 2} = N\sigma_u^{\star 2} + \sigma_e^{\star 2}$ as:

$$\mathrm{Var}(\mathbf{l}|\mathbf{Z}) = \eta^\star\sigma^{\star 2}\mathbf{R} + (1-\eta^\star)\sigma^{\star 2}\mathrm{Id}_{\mathbb{R}^n}\ .$$

We will assume in the sequel without loss of generality that $\sigma^{\star 2} = 1$. Indeed, if $\sigma^{\star 2} \neq 1$, we can consider the variable $\mathbf{l}'_i = \frac{\mathbf{l}_i}{\sigma^\star}$ and then, instead of estimating t from the prevalence K with the relationship (5.6), we estimate directly $t/\sigma^\star$.

5.2.2 Case control study

Since the prevalence P in the study can be very different from the prevalence K in the general population (the cases are substantially oversampled in a case-control study), it is essential to consider that the observations that we have access to depend on the probabilities for both cases and controls to be selected in the study. Indeed, if $p_{control}$ denotes the probability for a control to be selected in the study, we can define the corresponding variable $U_i \sim \mathcal{B}(0, p_{control})$ which is equal to 1 if individual i is a control who is selected in the study. Similarly we define the probability p_{case} for a case to be selected for the sudy and the corresponding variable $V_i \sim \mathcal{B}(0, p_{case})$. Then for any individual i, we define the variable ϵ_i by

$$\epsilon_i = V_i\mathbf{Y}_i + U_i(1 - \mathbf{Y}_i),$$

which is equal to 1 if individual i belongs to the study and 0 if not. We assume that the variables $U_1, \dots, U_n, V_1, \dots, V_n$ are independent and independent of $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ and $\mathbf{Z}$.
Since we do not observe $\mathbf{Y}_i$ for the whole population but only for the individuals who belong to the study, we will work with the variables $\mathbf{W}_i$ defined by

$$\mathbf{W}_i = \frac{\mathbf{Y}_i - P}{\sqrt{P(1-P)}}\epsilon_i,$$

which are centered versions of $\mathbf{Y}_i$ in the study and are non-zero only if individual i belongs to the study.
The probabilities p_{case} and $p_{control}$ are chosen such that the prevalence in the study is equal to P. Indeed, if we assume that

$$p_{case} = 1, \tag{5.12}$$

it implies that

$$p_{control} = \frac{K(1-P)}{P(1-K)}. \tag{5.13}$$

The proof of (5.13) is given in Appendix 5.A. Equation (5.12) means that all cases are accepted in the study and it is usually called a "full ascertainment" assumption (see for instance Golan et al. (2014)).

5.3 Heritability estimator

5.3.1 Method of Golan et al. (2014)

Golan et al. (2014) considered a simplified version of Model (5.4), where the liability is given by

$$\mathbf{l} = \mathbf{g} + \mathbf{e},$$

where $\mathbf{g}$ is a genetic random effect, which can be correlated across individuals, and $\mathbf{e}$ is the environmental random effect, which is assumed to be independent of the genetic effect. Both effects are assumed to be Gaussian: $\mathbf{e}$ has a variance equal to $(1-\eta^\star)\mathrm{Id}_{\mathbb{R}^n}$ and $\mathbf{g}$ has a covariance matrix, the diagonal entries of which are equal to $\eta^\star$ and the non diagonal term (i,j) is equal to $\eta^\star \mathbf{G}_{i,j}$. The covariance matrix of $(\mathbf{l}_i, \mathbf{l}_j)$ is given by

$$\Sigma = \begin{pmatrix} 1 & \eta^\star \mathbf{G}_{i,j} \\ \eta^\star \mathbf{G}_{i,j} & 1 \end{pmatrix}.$$

The heritability estimator proposed by Golan et al. (2014) is a least square estimator obtained by minimizing

$$\sum_{i \neq j} \left(\mathbf{W}_i \mathbf{W}_j - \mathbb{E}[\mathbf{W}_i \mathbf{W}_j | \epsilon_i = \epsilon_j = 1]\right)^2. \tag{5.14}$$

Since the expression of $\mathbb{E}[\mathbf{W}_i \mathbf{W}_j | \epsilon_i = \epsilon_j = 1]$ has no explicit formula as we shall see hereafter,

Golan et al. (2014) proposed to take advantage of the fact that the correlations $\mathbf{G}_{i,j}$ are small for $i \neq j$.
The ground of the method is to write

$$\mathbb{E}[\mathbf{W}_i\mathbf{W}_j|\epsilon_i=\epsilon_j=1] = \frac{\frac{1-P}{P}\mathbb{P}(\mathbf{Y}_i=\mathbf{Y}_j=1) - \frac{K(1-P)}{P(1-K)}\mathbb{P}(\mathbf{Y}_i\neq\mathbf{Y}_j) + \frac{K^2(1-P)}{P(1-K)^2}\mathbb{P}(\mathbf{Y}_i=\mathbf{Y}_j=0)}{\mathbb{P}(\mathbf{Y}_i=\mathbf{Y}_j=1) + \left(\frac{K(1-P)}{P(1-K)}\right)^2\mathbb{P}(\mathbf{Y}_i=\mathbf{Y}_j=0) + \frac{K(1-P)}{P(1-K)}\mathbb{P}(\mathbf{Y}_i\neq\mathbf{Y}_j)} \tag{5.15}$$

and to propose approximations of $\mathbb{P}(\mathbf{Y}_i \neq \mathbf{Y}_j)$, $\mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 0)$ and $\mathbb{P}(\mathbf{Y}_i \neq \mathbf{Y}_j)$ thanks to Taylor developments around the quantity $\mathbf{G}_{i,j}$. The computations leading to (5.15) can be found in Appendix 5.B.
This approximation, plugged in the least square criterion (5.14), led to the heritability estimator given by

$$\hat{\eta} = \left[\frac{\sum_{i\neq j}\mathbf{W}_i\mathbf{W}_j\mathbf{G}_{i,j}}{c\sum_{i\neq j}\mathbf{G}_{i,j}^2} \wedge 1\right] \vee 0, \tag{5.16}$$

where

$$c = \phi(t)^2\frac{P(1-P)}{K^2(1-K)^2}, \tag{5.17}$$

ϕ being the density of the standard Gaussian distribution.

5.3.2 Our method

In Model (5.7) that we consider, the variance matrix $\Sigma^{(N)}$ of $(\mathbf{l}_i, \mathbf{l}_j)$ conditionally to $\mathbf{Z}$ can be written as

$$\Sigma^{(N)} = \begin{pmatrix} 1+\eta^\star(\mathbf{G}_N(i,i)-1) & \eta^\star\mathbf{G}_N(i,j) \\ \eta^\star\mathbf{G}_N(i,j) & 1+\eta^\star(\mathbf{G}_N(j,j)-1) \end{pmatrix},$$

where for all $1 \leq i, j \leq n$,

$$\mathbf{G}_N(i,j) = \frac{1}{N}\sum_{k=1}^{N}\mathbf{Z}_{i,k}\mathbf{Z}_{j,k}. \tag{5.18}$$

Note that in the model we consider, $\mathbf{G}_N(i,j)$ is a random variable, which is not the case of the quantity $\mathbf{G}_{i,j}$ in the model studied by Golan et al. (2014). A key element is to notice that $\Sigma^{(N)}$ is close to the $n \times n$ identity matrix, more precisely

$$\Sigma^{(N)} = \begin{pmatrix} 1+\eta^\star\frac{A_N(i)}{\sqrt{N}} & \eta^\star\frac{B_N(i,j)}{\sqrt{N}} \\ \eta^\star\frac{B_N(i,j)}{\sqrt{N}} & 1+\eta^\star\frac{A_N(j)}{\sqrt{N}} \end{pmatrix} \tag{5.19}$$

where $A_N(i) = O_p(1)$, $A_N(j) = O_p(1)$ and $B_N(i,j) = O_p(1)$. The proof of (5.19) can be found in Appendix 5.C.
Then, following the idea of Golan et al. (2014), we propose to approximate

$$\mathbb{E}[\mathbf{W}_i\mathbf{W}_j|\mathbf{Z}, \epsilon_i = \epsilon_j = 1]$$

defined in Equation (5.15) thanks to Taylor developments around $\frac{A_N(i)}{\sqrt{N}}$, $\frac{A_N(j)}{\sqrt{N}}$ and $\frac{B_N(i,j)}{\sqrt{N}}$.
The detailed computations are devised in Section 5.8.1.
A first order approximation of $\mathbb{E}[\mathbf{W}_i\mathbf{W}_j|\mathbf{Z}, \epsilon_i = \epsilon_j = 1]$, plugged in (5.14), leads to the same estimator $\hat{\eta}^{(1)}$ as the one proposed by Golan et al. (2014). Indeed, we obtain

$$\hat{\eta}^{(1)} = \left[\frac{\sum_{i\neq j}\mathbf{W}_i\mathbf{W}_j\mathbf{G}_N(i,j)}{c\sum_{i\neq j}\mathbf{G}_N(i,j)^2} \wedge 1\right] \vee 0, \tag{5.20}$$

where $c = \phi(t)^2\frac{P(1-P)}{K^2(1-K)^2}$.
In Section 5.5, we consider the second order approximation, which is different from the one devised by Golan et al. (2014).

5.4 Consistency of the heritability estimator $\hat{\eta}^{(1)}$

In this section, we consider the heritability estimator $\hat{\eta}^{(1)}$ defined in Equation (5.20).

Assumption 2. There exist $d > 0$, $C > 0$ and a neighborhood V_0 of 0 such that for all λ in V_0

1.1 $\mathbb{E}[\exp\left(\lambda(A_{i,k} - \mathbb{E}[A_{i,k}])^2 - \sigma_k^2\right)] \leq C\exp(d\lambda^2)$

1.2 $\mathbb{E}[\exp\left(\lambda(A_{i,k} - \mathbb{E}[A_{i,k}])\right)] \leq C\exp(d\lambda^2)$

1.3 $\mathbb{E}[\exp\left(\lambda(A_{i,k} - \mathbb{E}[A_{i,k}])(A_{j,k} - \mathbb{E}[A_{i,k}])\right)] \leq C\exp(d\lambda^2)$

for all $i \neq j$ and for all k, where the $A_{i,k}$'s are defined in (5.9) and σ_k^2 is the variance of $A_{i,k}$.
Assumption 3.

2.1 $\inf_{k=1..N}\sigma_k^2 = \delta_{min} > 0$

2.2 $\sup_{k=1..N}\sigma_k^2 = \delta_{max} < +\infty$

Theorem 4. *Let* $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ *satisfy Model* (5.5) *with* A *satisfying Assumptions 2 and 3, and* $\hat{\eta}^{(1)}$ *the estimator of* $\eta^\star$ *defined in Equation* (5.20). *Then, as* $n, N \to \infty$ *such that* $n/N \to a \in (0, +\infty)$,

$$\hat{\eta}^{(1)} = \eta^\star + o_p(1).$$

The proof of Theorem 4 relies on the following lemmas.

Lemma 7. When n and N go to infinity and n/N goes to a,

$$\frac{1}{n}\sum_{i\neq j}\mathbf{G}_N(i,j)^2 \text{converges in probability to } a.$$

We will then have to focus on

$$\frac{1}{n}\sum_{i\neq j}\mathbf{W}_i\mathbf{W}_j\mathbf{G}_N(i,j) = \left[\frac{1}{n}\sum_{i\neq j}(\mathbf{W}_i\mathbf{W}_j - \mathbb{E}[\mathbf{W}_i\mathbf{W}_j|\mathbf{Z},\epsilon_i=\epsilon_j=1])\mathbf{G}_N(i,j) + \frac{1}{n}\sum_{i\neq j}\mathbb{E}[\mathbf{W}_i\mathbf{W}_j|\mathbf{Z},\epsilon_i=\epsilon_j=1]\mathbf{G}_N(i,j)\right]. \tag{5.21}$$

Let E_N be the following event

$$E_N = \left\{\sup_i|\mathbf{G}_N(i,i)-1|\leq\epsilon_N \text{ and } \sup_{i\neq j}|\mathbf{G}_N(i,j)|\leq\epsilon_N\right\},$$

where $\epsilon_N = \frac{1}{N^{\frac{1}{2}-\gamma}}$ with γ a positive number such that $\gamma < 1/10$.
Let us denote E_N^c the complement of the event E_N. We consider the following decomposition

$$\hat{\eta}^{(1)} = \hat{\eta}^{(1)}\mathbb{1}_{E_N} + \hat{\eta}^{(1)}\mathbb{1}_{E_N^c}.$$

Lemma 8. For all values of q, the probability of E_N^c satisfies $\mathbb{P}(E_N^c) = O(\frac{1}{N^q})$ when $N\to+\infty$.

Using the result of Lemma 8, $\hat{\eta}^{(1)}\mathbb{1}_{E_N^c}$ converges in probability to 0 since

$$\mathbb{E}[|\hat{\eta}^{(1)}\mathbb{1}_{E_N^c}|] \leq \mathbb{P}(E_N^c) = O\left(\frac{1}{N^q}\right).$$

Lemma 9. When n and N go to infinity and n/N goes to $a\in(0,+\infty)$,

$$\frac{1}{n}\sum_{i\neq j}\mathbb{E}[\mathbf{W}_i\mathbf{W}_j|\mathbf{Z},\epsilon_i=\epsilon_j=1]\mathbf{G}_N(i,j)\mathbb{1}_{E_N}$$

converges in probability to $ac\eta^\star$, where c is defined in Equation (5.17).

Lemma 10. When n and N go to infinity and n/N goes to $a\in(0,+\infty)$,

$$\frac{1}{n}\sum_{i\neq j}(\mathbf{W}_i\mathbf{W}_j - \mathbb{E}[\mathbf{W}_i\mathbf{W}_j|\mathbf{Z},\epsilon_i=\epsilon_j=1])\mathbf{G}_N(i,j)\mathbb{1}_{E_N}$$

converges in probability to 0.

The results of Lemmas 9 and 10 achieve the proof of Theorem 4.
The proof of Lemmas 7, 8, 9 and 10 are given in Section 5.8.2.

5.5 Second order approximation of $\mathbb{E}[\mathbf{W}_i\mathbf{W}_j|\mathbf{Z}, \epsilon_i = \epsilon_j = 1]$

The purpose of this section is to study the behaviour of the heritability estimator obtained thanks to a second order approximation of $\mathbb{E}[\mathbf{W}_i\mathbf{W}_j|\mathbf{Z}, \epsilon_i = \epsilon_j = 1]$.
Instead of computing the approximation till order $1/\sqrt{N}$, we compute the approximation till order $1/N$ and we obtain:

$$\begin{aligned}\mathbb{E}[\mathbf{W}_i\mathbf{W}_j|\mathbf{Z}, \epsilon_i = \epsilon_j = 1] = &\frac{\eta^\star}{\sqrt{N}}\frac{P(1-P)}{K^2(1-K)^2}\phi(t)^2 B_N(i,j) + \frac{t^2}{4}\frac{\eta^{\star 2}}{N}A_N(i)A_N(j)\frac{P(1-P)}{K^2(1-K)^2} \\ &+ \frac{\eta^{\star 2}}{N}\frac{P(1-P)}{K^2(1-K)^2}\phi(t)^2 B_N(i,j)^2\left[\frac{t^2}{2} - \frac{(P-K)^2}{K^2(1-K)^2}\right] \\ &+ \frac{\eta^{\star 2}}{2N}\frac{P(1-P)}{K^2(1-K)^2}\phi(t)^2 B_N(i,j)(A_N(i)+A_N(j))\left[t^2 - 1 - \frac{P-K}{K(1-K)}t\phi(t)\right] + O_p\left(\frac{1}{N^{\frac{3}{2}}}\right)\end{aligned}$$

The proof of this computation is detailed in Section 5.8.3.
Since the minimizer in η of the quantity

$$\begin{aligned}g(\eta) = \sum_{i\neq j}\Bigg(&\mathbf{W}_i\mathbf{W}_j - \frac{\eta}{\sqrt{N}}\frac{P(1-P)}{K^2(1-K)^2}\phi(t)^2 B_N(i,j) - \frac{t^2}{4}\frac{\eta^2}{N}A_N(i)A_N(j)\frac{P(1-P)}{K^2(1-K)^2} \\ &- \frac{\eta^2}{N}\frac{P(1-P)}{K^2(1-K)^2}\phi(t)^2 B_N(i,j)^2\left[\frac{t^2}{2} - \frac{(P-K)^2}{K^2(1-K)^2}\right] \\ &- \frac{\eta^2}{2N}\frac{P(1-P)}{K^2(1-K)^2}\phi(t)^2 B_N(i,j)(A_N(i)+A_N(j))\left[t^2 - 1 - \frac{P-K}{K(1-K)}t\phi(t)\right]\Bigg)^2\end{aligned}$$

has no explicit form, we use a Newton-Raphson approach to obtain the corresponding heritability estimator $\hat{\eta}^{(2)}$ of the second order approximation.
Note that the second order approximation, which depends on $B_N(i,j)$ but also on $A_N(i)$ and $A_N(j)$, is different from the one found by Golan et al. (2014).

5.6 Numerical study

In this section, we propose to study the numerical performance of the estimators $\hat{\eta}^{(1)}$ and $\hat{\eta}^{(2)}$ devised respectively in Sections 5.3 and 5.5. Since Golan et al. (2014) already compared the estimator $\hat{\eta}^{(1)}$ to the one proposed by Lee et al. (2011) and stated several arguments in favor of their estimator, we will focus on comparing our two estimators in terms of statistical and computational efficiency.

5.6.1 Simulation process

In this simulation study, we generated data sets with $n \simeq 200$, $N = 10000$ in order to respect the classical scenario where $N >> n$. The value of the prevalence in the population varies from 0.005 to 0.1. The observations were generated as follows.

- We set the parameters $\eta^\star$, K, $P = 1/2$ and the size of the general population, chosen very large. Notice that the number of individuals selected in the study varies from one sample to another. We chosed in practice a population size in order to have around 100 patients in the study.
- We generated the Gaussian random effects $\mathbf{e}$ and $\mathbf{u}$ with respective variances $\sigma_u^{\star 2} = \eta^\star / N$ and $\sigma_e^{\star 2} = 1 - \eta^\star$.
- We generated liabilities, from which we generated binary observations in order to have a certain prevalence K in the population, that is certain number of cases.
- For each individual, we determined those who stayed in the study: the cases are automatically selected (full ascertainment assumption) but each control is selected with probability $p_{control}$ computed in Equation (5.13).

5.6.2 Results

Figure 5.1 displays the estimations of $\eta^\star$ obtained with both estimators $\hat{\eta}^{(1)}$ and $\hat{\eta}^{(2)}$. First, we can notice that both estimators seem empirically unbiased. Second, we observe no obvious improvement of the performance of $\hat{\eta}^{(2)}$ compared to $\hat{\eta}^{(1)}$ in terms of empirical variance. Finally, we can also note that the estimations seem more accurate when the prevalence K is high, namely $K = 0.1$.
Table 5.1 and Figure 5.2 show the computational performance of both estimators. The computation of the estimator $\hat{\eta}^{(2)}$ obtained with the more refined approximation is obviously slower, but for small values of n (namely, $n = 100$), the time required to compute an estimation of $\eta^\star$ remains quite small (86 seconds, against 40 seconds for the other estimator). However, when n is larger, the computational time increase substantially and the "slower" estimator needs up to 13500 seconds, that is almost 4 hours, to compute an estimation of $\eta^\star$.
In conclusion, both estimators are empirically unbiased and since the computation of the estimator $\hat{\eta}^{(2)}$ is slower and does not improve the estimations of $\eta^\star$, we are satisfied with the first order approximation and the corresponding estimator $\hat{\eta}^{(1)}$.

Table 5.1 – Times in seconds to compute an estimation of $\eta^\star$ obtained with $\hat{\eta}^{(1)}$ and $\hat{\eta}^{(2)}$ for different values of n (100 and 1000) and N (from 1000 to 10^5).

n	N	1000	10000	50000	10^5
100	$\hat{\eta}^{(1)}$	0.478	2.390	28.595	40.528
	$\hat{\eta}^{(2)}$	3.148	7.127	56.761	86.156
1000	$\hat{\eta}^{(1)}$	69.047	327.240	2887.518	7845.281
	$\hat{\eta}^{(2)}$	376.363	936.845	6624.186	13500.510

5.7 Discussion

In this paper, we proposed theoretical grounds to support the heritability estimator in case-control studies developed by Golan et al. (2014). We proved indeed its consistency in the framework where both the number of individuals n and the number N of SNPs go to infinity,

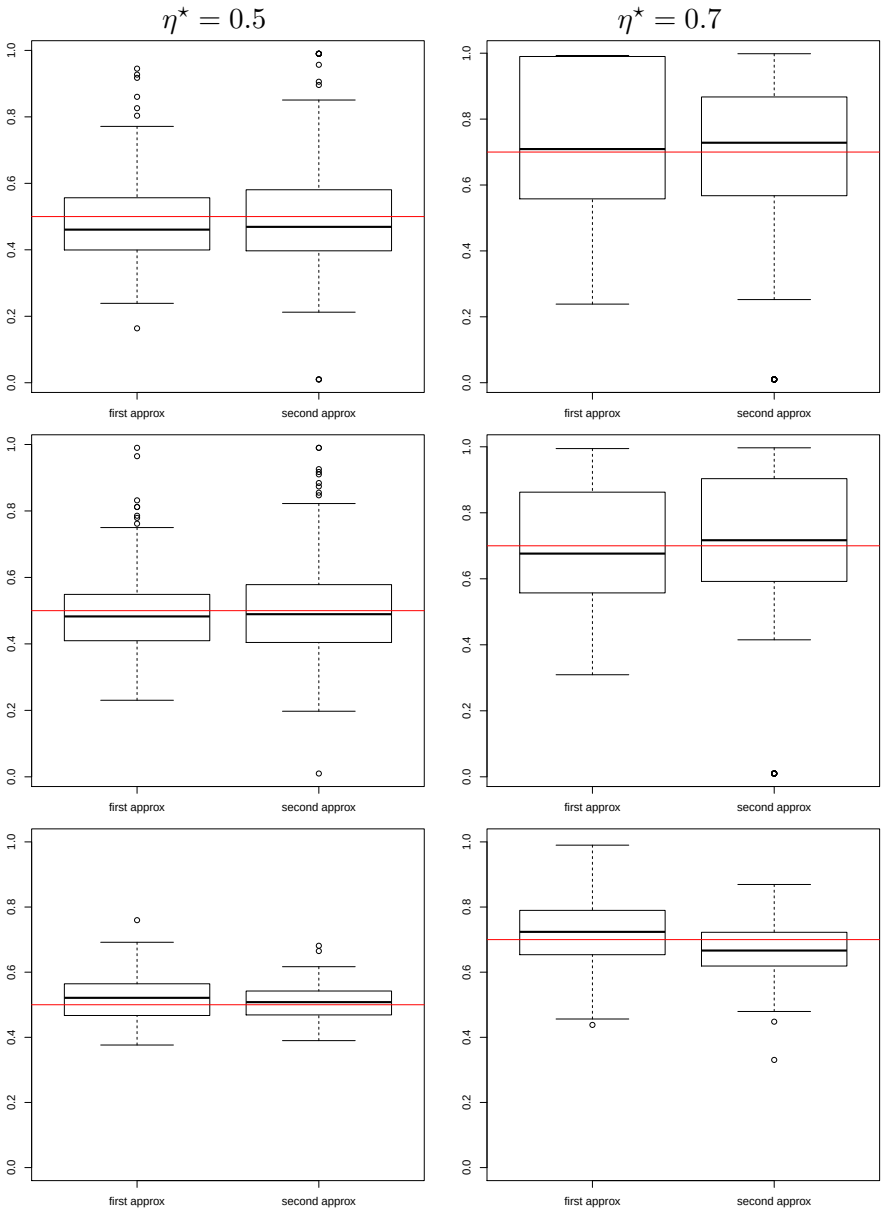


Figure 5.1 – Estimations obtained with $\hat{\eta}^{(1)}$ ("first approx") and $\hat{\eta}^{(2)}$ ("second approx") for different values of $\eta^\star$: 0.5 (left), 0.7 (right) and different values of the prevalence K: 0.005 (top), 0.01 (middle), 0.1 (bottom). The sample size is $n \simeq 200$ and $N = 10000$. Each boxplot is generated from 200 replications.

when the ratio n/N goes to a constant a.
It would be interesting to complete this work with theoretical results which could allow the user to compute accurate confidence intervals, similarly to existing results for quantitive traits. As it is often the case in genetic applications, the question of removing strong assumptions such as the Gaussianity of the random effects or the independence of the columns of the SNP matrix

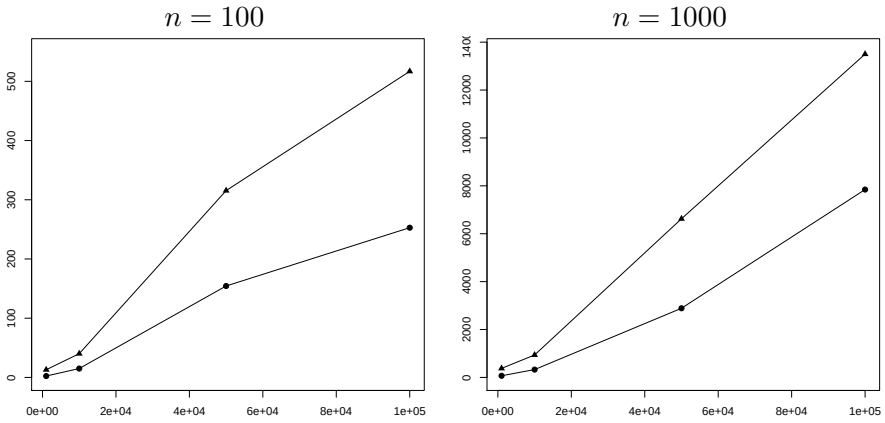


Figure 5.2 – Time in seconds to compute an estimation of $\eta^\star$ obtained with $\hat{\eta}^{(1)}$ (dots) and $\hat{\eta}^{(2)}$ (triangles) for $n = 100$ (left) and $n = 1000$ (right) and for different values of N (from 1000 to 10^5).

remains a challenging issue. Considering possible sparsity in the random effects would also be an interesting improvement and will be the subject of a future work.

5.8 Proofs

5.8.1 Taylor development of $\mathbb{E}[\mathbf{W}_i\mathbf{W}_j|\mathbf{Z}, \epsilon_i = \epsilon_j = 1]$ in Model (5.7)

According to Equation (5.15), we only need to compute approximations of $\mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 1|\mathbf{Z})$, $\mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 0|\mathbf{Z})$ and $\mathbb{P}(\mathbf{Y}_i \neq \mathbf{Y}_j|\mathbf{Z})$ to obtain an approximation of $\mathbb{E}(\mathbf{W}_i\mathbf{W}_j|\mathbf{Z}, \epsilon_i = \epsilon_j = 1)$.

$$\mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 1|\mathbf{Z}) = \int_t^\infty \int_t^\infty f(x,y)dxdy,$$

$$\mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 0|\mathbf{Z}) = \int_{-\infty}^t \int_{-\infty}^t f(x,y)dxdy$$

and

$$\mathbb{P}(\mathbf{Y}_i \neq \mathbf{Y}_j|\mathbf{Z}) = 2\int_{-\infty}^t \int_t^\infty f(x,y)dxdy,$$

with

$$f(x,y) = \frac{1}{2\pi}|\Sigma^{(N)}|^{-\frac{1}{2}} \exp\left\{-\frac{(x,y)\Sigma^{(N)-1}(x,y)^t}{2}\right\}.$$

where the matrix $\Sigma^{(N)}$ is the covariance matrix of $(\mathbf{l}_i, \mathbf{l}_j)$.

We will use the result of Equation (5.19), which will be demonstrated in Appendix 5.C, that is

$$\Sigma^{(N)} = \begin{pmatrix} 1+\eta^\star \frac{A_N(i)}{\sqrt{N}} & \eta^\star \frac{B_N(i,j)}{\sqrt{N}} \\ \eta^\star \frac{B_N(i,j)}{\sqrt{N}} & 1+\eta^\star \frac{A_N(j)}{\sqrt{N}} \end{pmatrix}, \tag{5.22}$$

where $A_N(i) = O_p(1)$, $A_N(j) = O_p(1)$ and $B_N(i,j) = O_p(1)$.
We have

$$\begin{aligned} f(x,y) &= \frac{1}{2\pi|\Sigma^{(N)}|^{-\frac{1}{2}}} \exp\left\{-\frac{1}{2|\Sigma^{(N)}|}\left[x^2(1+\frac{\eta^\star}{\sqrt{N}}A_N(j)) + y^2(1+\frac{\eta^\star}{\sqrt{N}}A_N(i)) - 2xy\frac{\eta^\star}{\sqrt{N}}B_N(i,j)\right]\right\} \\ &= \frac{1}{2\pi|\Sigma^{(N)}|^{-\frac{1}{2}}} \exp(-\frac{x^2}{2})\exp(-\frac{y^2}{2})\exp\left\{-\frac{x^2}{2}\left(\frac{1}{|\Sigma^{(N)}|}\left[1+\frac{\eta^\star}{\sqrt{N}}A_N(j)\right]-1\right)\right. \\ &\qquad \left. -\frac{y^2}{2}\left(\frac{1}{|\Sigma^{(N)}|}\left[1+\frac{\eta^\star}{\sqrt{N}}A_N(i)\right]-1\right) + \frac{1}{|\Sigma^{(N)}|}xy\frac{\eta^\star}{\sqrt{N}}B_N(i,j)\right\}. \end{aligned}$$

Using a first order Taylor development,

$$|\Sigma^{(N)}|^{-1} = 1 - (A_N(i)+A_N(j))\frac{\eta^\star}{\sqrt{N}} + \alpha_N$$

and

$$|\Sigma^{(N)}|^{-\frac{1}{2}} = 1 - \frac{1}{2}(A_N(i)+A_N(j))\frac{\eta^\star}{\sqrt{N}} + \beta_N,$$

where $\alpha_N = O_p(\frac{1}{N})$ and $\beta_N = O_p(\frac{1}{N})$.
More precisely,

$$\begin{aligned} \alpha_N &= -(A_N(i)A_N(j) - B_N(i,j)^2)\frac{\eta^{\star 2}}{N} \\ &\quad + \frac{1}{2}\left(-(A_N(i)+A_N(j))\frac{\eta^\star}{\sqrt{N}} - (A_N(i)A_N(j) - B_N(i,j)^2)\frac{\eta^{\star 2}}{N}\right)^2 \frac{1}{(1+\tilde{\alpha})^3}, \end{aligned}$$

with $|\tilde{\alpha}| \leq |(A_N(i)+A_N(j))\frac{\eta^\star}{\sqrt{N}} + (A_N(i)A_N(j) - B_N(i,j)^2)\frac{\eta^{\star 2}}{N}|$.
Similarly,

$$\begin{aligned} \beta_N &= -\frac{1}{2}(A_N(i)A_N(j) - B_N(i,j)^2)\frac{\eta^{\star 2}}{N} \\ &\quad + \frac{1}{2}\left(-\frac{1}{2}(A_N(i)+A_N(j))\frac{\eta^\star}{\sqrt{N}} - \frac{1}{2}(A_N(i)A_N(j) - B_N(i,j)^2)\frac{\eta^{\star 2}}{N}\right)^2 \frac{3}{4}\frac{1}{(1+\tilde{\beta})^{\frac{5}{2}}}, \end{aligned}$$

with $|\tilde{\beta}| \leq |\frac{1}{2}(A_N(i)+A_N(j))\frac{\eta^\star}{\sqrt{N}} + \frac{1}{2}(A_N(i)A_N(j) - B_N(i,j)^2)\frac{\eta^{\star 2}}{N}|$.

Then,

$$f(x,y) = \left(1 - \frac{1}{2}(A_N(i) + A_N(j))\frac{\eta^\star}{\sqrt{N}} + \beta_N\right)\phi(x)\phi(y)$$
$$\times \exp\left\{-\frac{x^2}{2}(-A_N(i)\frac{\eta^\star}{\sqrt{N}} + \gamma_N) - \frac{y^2}{2}(-A_N(j)\frac{\eta^\star}{\sqrt{N}} + \tilde{\gamma}_N) + xy\left(\frac{\eta^\star}{\sqrt{N}}B_N(i,j) + \tilde{\tilde{\gamma}}_N\right)\right\}$$

where $\gamma_N = -A_N(j)(A_N(i) + A_N(j))\frac{\eta^{\star 2}}{N} + \alpha_N(1 + A_N(j)\frac{\eta^\star}{\sqrt{N}}) = O_p\left(\frac{1}{N}\right)$,
$\tilde{\gamma_N} = -A_N(i)(A_N(i) + A_N(j))\frac{\eta^{\star 2}}{N} + \alpha_N(1 + A_N(i)\frac{\eta^\star}{\sqrt{N}}) = O_p\left(\frac{1}{N}\right)$ and
$\tilde{\tilde{\gamma}}_N = \frac{\eta^\star}{\sqrt{N}}B_N(i,j)\left(-(A_N(i) + A_N(j))\frac{\eta^\star}{\sqrt{N}} + \alpha_N\right) = O_p\left(\frac{1}{N}\right)$
A Taylor development of the exponential function leads to

$$f(x,y) = \left(1 - \frac{1}{2}(A_N(i) + A_N(j))\frac{\eta^\star}{\sqrt{N}} + \beta_N\right)\phi(x)\phi(y)$$
$$\times \left[1 + \frac{x^2}{2}\frac{\eta^\star}{\sqrt{N}}A_N(i) + \frac{y^2}{2}\frac{\eta^\star}{\sqrt{N}}A_N(j) + xy\frac{\eta^\star}{\sqrt{N}}B_N(i,j) + \nu_N(x)\right]$$

with

$$\nu_N(x) = -\frac{x^2}{2}\gamma_N - \frac{y^2}{2}\tilde{\gamma}_N + xy\tilde{\tilde{\gamma}}_N$$
$$+ \frac{1}{2}\left(\frac{x^2}{2}\frac{\eta^\star}{\sqrt{N}}A_N(i) + \frac{y^2}{2}\frac{\eta^\star}{\sqrt{N}}A_N(j) + xy\frac{\eta^\star}{\sqrt{N}}B_N(i,j) - \frac{x^2}{2}\gamma_N - \frac{y^2}{2}\tilde{\gamma}_N + xy\tilde{\tilde{\gamma}}_N\right)^2 \exp\tilde{u}$$

where $|\tilde{u}| \leq |\frac{x^2}{2}\frac{\eta^\star}{\sqrt{N}}A_N(i) + \frac{y^2}{2}\frac{\eta^\star}{\sqrt{N}}A_N(j) + xy\frac{\eta^\star}{\sqrt{N}}B_N(i,j) - \frac{x^2}{2}\gamma_N - \frac{y^2}{2}\tilde{\gamma_N} + xy\tilde{\tilde{\gamma}}_N|$.
Then,

$$\int_t^\infty\int_t^\infty f(x,y)dxdy = \left(1 - \frac{1}{2}(A_N(i) + A_N(j))\frac{\eta^\star}{\sqrt{N}} + \beta_N\right)$$
$$\left[K^2 + \frac{1}{2}\frac{\eta^\star}{\sqrt{N}}(A_N(j) + A_N(i))K(K + t\phi(t)) + B_N(i,j)\frac{\eta^\star}{\sqrt{N}}\phi(t)^2\right] + \mu_N$$
$$= K^2 + \frac{1}{2}(A_N(i) + A_N(j))\frac{\eta^\star}{\sqrt{N}}Kt\phi(t) + B_N(i,j)\frac{\eta^\star}{\sqrt{N}}\phi(t)^2 + \mu'_N$$

where $\mu_N = \left(1 - \frac{1}{2}(A_N(i) + A_N(j) + \beta_N)\frac{\eta^\star}{\sqrt{N}}\right)\int_t^\infty\int_t^\infty \phi(x)\phi(y)\nu_N(x)dxdy$

$$\text{and } \mu'_N = \mu_N + \beta_N\left(K^2 + \frac{1}{2}\frac{\eta^\star}{\sqrt{N}}(A_N(j) + A_N(i))K(K + t\phi(t)) + B_N(i,j)\frac{\eta^\star}{\sqrt{N}}\phi(t)^2\right)$$
$$- \frac{1}{2}(A_N(i) + A_N(j))\frac{\eta^{\star 2}}{N}B_N(i,j)\phi(t)^2.$$

This remainder and its order will be carefully studied in Section 5.8.2.

Similarly, we can compute $\mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 0|\mathbf{Z})$ and $\mathbb{P}(\mathbf{Y}_i \neq \mathbf{Y}_j|\mathbf{Z})$:

$$\int_{-\infty}^{t}\int_{-\infty}^{t} f(x,y)dxdy = (1-K)^2 - \frac{1}{2}(A_N(i)+A_N(j))\frac{\eta^\star}{\sqrt{N}}(1-K)t\phi(t) + B_N(i,j)\frac{\eta^\star}{\sqrt{N}}\phi(t)^2 + \tilde{\mu}_N$$

$$\int_{-\infty}^{t}\int_{t}^{\infty} f(x,y)dxdy + \int_{t}^{\infty}\int_{-\infty}^{t} f(x,y)dxdy = 2K(1-K) + (A_N(i)+A_N(j))\frac{\eta^\star}{\sqrt{N}}(1-2K)t\phi(t) - 2B_N(i,j)\frac{\eta^\star}{\sqrt{N}}\phi(t)^2 + \tilde{\tilde{\mu}}_N.$$

Replacing these terms in the expression of the numerator of $\mathbb{E}(\mathbf{W}_i\mathbf{W}_j|\mathbf{Z}, \epsilon_i = \epsilon_j = 1)$ given in equation (5.15) leads to:

$$\frac{\eta^\star}{\sqrt{N}}B_N(i,j)\phi(t)^2\frac{(1-P)}{P(1-K)^2} + r_N, \tag{5.23}$$

where r_N is a linear combination of μ'_N, $\tilde{\mu}_N$ and $\tilde{\tilde{\mu}}_N$.
Since there is no constant term in this numerator, we only need the development of order 0 of the denominator of $\mathbb{E}(\mathbf{W}_i\mathbf{W}_j|\mathbf{Z}, \epsilon_i = \epsilon_j = 1)$ to obtain the first order approximation of $\mathbb{E}(\mathbf{W}_i\mathbf{W}_j|\mathbf{Z}, \epsilon_i = \epsilon_j = 1)$.
We obtain that the denominator can be written as

$$\frac{K^2}{P^2} + \tilde{r}_N,$$

where $\tilde{r}_N$ is the sum of a term of order $\frac{1}{\sqrt{N}}$ and a linear combination of μ'_N, $\tilde{\mu}_N$ and $\tilde{\tilde{\mu}}_N$. Thus, we obtain that

$$\mathbb{E}(\mathbf{W}_i\mathbf{W}_j|\mathbf{Z}, \epsilon_i = \epsilon_j = 1) = \frac{\frac{\eta^\star}{\sqrt{N}}B_N(i,j)\phi(t)^2\frac{(1-P)}{P(1-K)^2} + r_N}{\frac{K^2}{P^2} + \tilde{r}_N} \tag{5.24}$$

$$= \eta^\star\mathbf{G}_N(i,j)\phi(t)^2\frac{P(1-P)}{K^2(1-K)^2} + R_N(i,j) \tag{5.25}$$

where

$$R_N(i,j) = \left(\frac{\eta^\star}{\sqrt{N}}B_N(i,j)\phi(t)^2\frac{(1-P)}{P(1-K)^2} + r_N\right)\tilde{r}_N + \frac{K^2}{P^2}r_N. \tag{5.26}$$

5.8.2 Proof of Theorem 4

Properties of Z

In the following proofs, we will use several properties of the matrix $\mathbf{Z}$, which are stated in Proposition 1.

Proposition 1. Uniformly in k,

(1) $\mathbb{E}(\mathbf{Z}_{1,k}\mathbf{Z}_{2,k}) = -\frac{1}{n-1}$.

(2) $\mathbb{E}(\mathbf{Z}_{1,k}^p) = O(1)$, for all p.

(3) $\mathbb{E}(\mathbf{Z}_{1,k}^2\mathbf{Z}_{2,k}^2) = 1 + o(1)$.

(4) $\mathbb{E}[\mathbf{Z}_{1,k}^3\mathbf{Z}_{2,k}] = O\left(\frac{1}{n}\right)$.

(5) $\mathbb{E}[\mathbf{Z}_{1,k}^2\mathbf{Z}_{2,k}\mathbf{Z}_{3,k}] = O\left(\frac{1}{n}\right)$.

(6) $\mathbb{E}[\mathbf{Z}_{1,k}\mathbf{Z}_{2,k}\mathbf{Z}_{3,k}\mathbf{Z}_{4,k}] = O\left(\frac{1}{n^2}\right)$.

(7) $\mathbb{E}[\mathbf{Z}_{1,k}^5\mathbf{Z}_{2,k}] = O\left(\frac{1}{n}\right)$.

(8) $\mathbb{E}[\mathbf{Z}_{1,k}^3\mathbf{Z}_{2,k}^3] = O(1)$.

(9) $\mathbb{E}[\mathbf{Z}_{1,k}^4\mathbf{Z}_{2,k}^2] = O(1)$.

(10) $\mathbb{E}[\mathbf{Z}_{1,k}^4\mathbf{Z}_{2,k}\mathbf{Z}_{3,k}] = O(\frac{1}{n})$.

(11) $\mathbb{E}[\mathbf{Z}_{1,k}^3\mathbf{Z}_{2,k}^2\mathbf{Z}_{3,k}] = O(\frac{1}{n})$.

(12) $\mathbb{E}[\mathbf{Z}_{1,k}^3\mathbf{Z}_{2,k}\mathbf{Z}_{3,k}\mathbf{Z}_{4,k}] = O(\frac{1}{n^2})$.

The proof of Proposition 1 is given in Appendix 5.D.

Proof of Lemma 7

Let us prove that, when n and N go to infinity and n/N goes to a,

$$\frac{1}{n}\sum_{i\neq j}\mathbf{G}_N(i,j)^2 \xrightarrow{P} a,$$

where $\xrightarrow{P}$ denotes the convergence in probability.

$$\mathbf{G}_N(i,j)^2 = \frac{1}{N^2}\sum_{k=1}^{N}\mathbf{Z}_{i,k}^2\mathbf{Z}_{j,k}^2 + \frac{1}{N^2}\sum_{k\neq l}\mathbf{Z}_{i,k}\mathbf{Z}_{j,k}\mathbf{Z}_{i,l}\mathbf{Z}_{j,l}$$

Since $\mathbf{Z}_{i,k}$ and $\mathbf{Z}_{j,l}$ are independent for any i and j when $k \neq l$, we will always consider separately the cases where $k = l$ from the cases where $k \neq l$.
Indeed, let us show that

$$\frac{1}{n}\frac{1}{N^2}\sum_{i\neq j}\sum_{k=1}^{N}\mathbf{Z}_{i,k}^2\mathbf{Z}_{j,k}^2 \xrightarrow{P} a \tag{5.27}$$

and

$$\frac{1}{n}\frac{1}{N^2}\sum_{i\neq j}\sum_{k\neq l}\mathbf{Z}_{i,k}\mathbf{Z}_{j,k}\mathbf{Z}_{i,l}\mathbf{Z}_{j,l} \xrightarrow{P} 0. \tag{5.28}$$

Note that

$$\begin{aligned}\mathbb{E}(\frac{1}{n}\frac{1}{N^2}\sum_{i\neq j}\sum_{k=1}^{N}\mathbf{Z}_{i,k}^2\mathbf{Z}_{j,k}^2) &= \frac{1}{n}\frac{1}{N^2}\sum_{i\neq j}\sum_{k=1}^{N}\mathbb{E}(\mathbf{Z}_{i,k}^2\mathbf{Z}_{j,k}^2)\\ &= \frac{n-1}{N}(1+o(1)) \text{ by (3) of Proposition 1}\\ &= a+o(1)\end{aligned}$$

Moreover,

$$\mathrm{Var}(\frac{1}{n}\frac{1}{N^2}\sum_{i\neq j}\sum_{k=1}^{N}\mathbf{Z}_{i,k}^2\mathbf{Z}_{j,k}^2) = \frac{1}{n^2}\frac{1}{N^4}\sum_{k=1}^{N}\sum_{i_1\neq j_1}\sum_{i_2\neq j_2}\mathbb{E}(\mathbf{Z}_{i_1,k}^2\mathbf{Z}_{j_1,k}^2\mathbf{Z}_{i_2,k}^2\mathbf{Z}_{j_2,k}^2) - \frac{1}{n^2}\frac{1}{N^4}\sum_{k=1}^{N}\left(\sum_{i\neq j}\mathbb{E}(\mathbf{Z}_{i,k}^2\mathbf{Z}_{j,k}^2)\right)^2 \tag{5.29}$$

The second term of (5.29) can be rewritten as:

$$\begin{aligned}\frac{1}{n^2}\frac{1}{N^4}\sum_{k=1}^{N}\left(\sum_{i\neq j}\mathbb{E}(\mathbf{Z}_{i,k}^2\mathbf{Z}_{j,k}^2)\right)^2 &= \frac{Nn^2(n-1)^2}{n^2N^4}(1+o(1)) \text{ by (3) of Proposition 1}\\ &= O\left(\frac{1}{n}\right)\end{aligned}$$

$$\sum_{i_1\neq j_1}\sum_{i_2\neq j_2}\mathbb{E}(\mathbf{Z}_{i_1,k}^2\mathbf{Z}_{j_1,k}^2\mathbf{Z}_{i_2,k}^2\mathbf{Z}_{j_2,k}^2) \leq \mathbb{E}(\sum_{i_1,j_1,i_2,j_2}\mathbf{Z}_{i_1,k}^2\mathbf{Z}_{j_1,k}^2\mathbf{Z}_{i_2,k}^2\mathbf{Z}_{j_2,k}^2) = \mathbb{E}\left(\sum_{i=1}^{n}\mathbf{Z}_{i,k}^2\right)^4 = n^4$$

This last equality comes from the definition of $\mathbf{Z}$ as a centered and normalized variable given in Equation (5.9), which implies that for all k,

$$\sum_{i=1}^{n}\mathbf{Z}_{i,k}^2 = n.$$

Then,

$$\frac{1}{n^2}\frac{1}{N^4}\sum_{k=1}^{N}\sum_{i_1\neq j_1}\sum_{i_2\neq j_2}\mathbb{E}(\mathbf{Z}_{i_1,k}^2\mathbf{Z}_{j_1,k}^2\mathbf{Z}_{i_2,k}^2\mathbf{Z}_{j_2,k}^2) \leq \frac{n^4N}{n^2N^4} = O\left(\frac{1}{n}\right).$$

This proves (5.27).

$$\begin{aligned}\mathbb{E}(\frac{1}{n}\frac{1}{N^2}\sum_{i\neq j}\sum_{k\neq l}\mathbf{Z}_{i,k}\mathbf{Z}_{j,k}\mathbf{Z}_{i,l}\mathbf{Z}_{j,l}) &= \frac{1}{n}\frac{1}{N^2}\sum_{i\neq j}\sum_{k\neq l}\mathbb{E}(\mathbf{Z}_{i,k}\mathbf{Z}_{j,k})\mathbb{E}(\mathbf{Z}_{i,l}\mathbf{Z}_{j,l})\\ &= \frac{n(n-1)N(N-1)}{nN^2(n-1)^2} \text{ by (1) of Proposition 1}\\ &= O\left(\frac{1}{n}\right)\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(\frac{1}{n}\frac{1}{N^2}\sum_{i\neq j}\sum_{k\neq l}\mathbf{Z}_{i,k}\mathbf{Z}_{j,k}\mathbf{Z}_{i,l}\mathbf{Z}_{j,l}) &= \frac{1}{n^2}\frac{1}{N^4}\sum_{k\neq l}\sum_{i_1\neq j_1}\sum_{i_2\neq j_2}\mathbb{E}(\mathbf{Z}_{i_1,k}\mathbf{Z}_{i_2,k}\mathbf{Z}_{j_1,k}\mathbf{Z}_{j_2,k})\mathbb{E}(\mathbf{Z}_{i_1,l}\mathbf{Z}_{i_2,l}\mathbf{Z}_{j_1,l}\mathbf{Z}_{j_2,l})\\ &\quad - \frac{1}{n^2}\frac{1}{N^4}\sum_{k\neq l}\left(\sum_{i\neq j}\mathbb{E}(\mathbf{Z}_{i,k}\mathbf{Z}_{j,k})\mathbb{E}(\mathbf{Z}_{i,l}\mathbf{Z}_{j,l})\right)^2\end{aligned}$$

$$\begin{aligned}\frac{1}{n^2}\frac{1}{N^4}\sum_{k\neq l}\left(\sum_{i\neq j}\mathbb{E}(\mathbf{Z}_{i,k}\mathbf{Z}_{j,k})\mathbb{E}(\mathbf{Z}_{i,l}\mathbf{Z}_{j,l})\right)^2 &= \frac{N(N-1)n^2(n-1)^2}{n^2N^4(n-1)^4} \text{ by (1) of Proposition 1}\\ &= O\left(\frac{1}{n^4}\right)\end{aligned}$$

In the first term, $\{i_1, i_2, j_1, j_2\}$ can be of cardinal 2, 3 or 4 and counting the number of combinations gives the expression:

$$\begin{aligned}\sum_{i_1\neq j_1}\sum_{i_2\neq j_2}\mathbb{E}(\mathbf{Z}_{i_1,k}\mathbf{Z}_{i_2,k}\mathbf{Z}_{j_1,k}\mathbf{Z}_{j_2,k})\mathbb{E}(\mathbf{Z}_{i_1,l}\mathbf{Z}_{i_2,l}\mathbf{Z}_{j_1,l}\mathbf{Z}_{j_2,l}) &= 2\sum_{i\neq j}\mathbb{E}(\mathbf{Z}_{i,k}^2\mathbf{Z}_{j,k}^2)\mathbb{E}(\mathbf{Z}_{i,l}^2\mathbf{Z}_{j,l}^2)\\ &\quad + 4\sum_{i\neq j_1\neq j_2}\mathbb{E}(\mathbf{Z}_{i,k}^2\mathbf{Z}_{j_1,k}\mathbf{Z}_{j_2,k})\mathbb{E}(\mathbf{Z}_{i,l}^2\mathbf{Z}_{j_1,l}\mathbf{Z}_{j_2,l})\\ &\quad + \sum_{i_1\neq i_2\neq j_1\neq j_2}\mathbb{E}(\mathbf{Z}_{i_1,k}\mathbf{Z}_{i_2,k}\mathbf{Z}_{j_1,k}\mathbf{Z}_{j_2,k})\mathbb{E}(\mathbf{Z}_{i_1,l}\mathbf{Z}_{i_2,l}\mathbf{Z}_{j_1,l}\mathbf{Z}_{j_2,l})\\ &= 2n(n-1)(1+o(1)) + 4\frac{n(n-1)(n-2)}{n}o(1) + \frac{n(n-1)(n-2)(n-3)}{n^2}o(1) = O(n^2)\end{aligned}$$

This was obtained by using (3),(5) and (6) of Proposition 1.
Finally,

$$Var\left(\frac{1}{n}\frac{1}{N^2}\sum_{i\neq j}\sum_{k\neq l}\mathbf{Z}_{i,k}\mathbf{Z}_{j,k}\mathbf{Z}_{i,l}\mathbf{Z}_{j,l}\right) = O\left(\frac{1}{n^2}\right).$$

This completes the proof of (5.28).

Proof of Lemma 8

Note that

$$\mathbb{P}(E_N^c) \leq n \sup_i \mathbb{P}\left(|\sum_{k=1}^N (\mathbf{Z}_{i,k}^2 - 1)| \geq N\epsilon_N\right) + n(n-1) \sup_{i\neq j} \mathbb{P}\left(|\sum_{k=1}^N \mathbf{Z}_{i,k}\mathbf{Z}_{j,k}| \geq N\epsilon_N\right)$$
$$= n\mathbb{P}\left(|\sum_{k=1}^N (\mathbf{Z}_{1,k}^2 - 1)| \geq N\epsilon_N\right) + n(n-1)\mathbb{P}\left(|\sum_{k=1}^N \mathbf{Z}_{1,k}\mathbf{Z}_{2,k}| \geq N\epsilon_N\right).$$

Let δ be a positive real number such that $\sqrt{\delta}/2c \in V_0$ and $\delta \leq \frac{\delta_{min}}{4}$, where V_0 and δ_{min} are defined in Assumptions 2 and **2.1** respectively.

$$\mathbb{P}\left(|\sum_{k=1}^N (\mathbf{Z}_{i,k}^2 - 1)| \geq N\epsilon_N\right) \leq \mathbb{P}\left(\exists k, s_k^2 \leq \delta\right) + \mathbb{P}\left(|\sum_{k=1}^N (A_{i,k} - \bar{A}_k)^2 - s_k^2)| \geq N\delta\epsilon_N\right)$$

Note also that

$$\left\{\exists k, s_k^2 \leq \delta\right\} = \bigcup_{k=1}^N \left\{\sum_{i=1}^n (A_{i,k} - \bar{A}_k)^2 \leq n\delta\right\} = \bigcup_{k=1}^N \left\{\sum_{i=1}^N (A_{i,k} - m_k + m_k - \bar{A}_k)^2 \leq n\delta\right\}$$

where $m_k = \mathbb{E}[A_{i,k}]$.

Observe that

$$\left\{\sum_{i=1}^n (A_{i,k} - m_k + m_k - \bar{A}_k)^2 \leq n\delta\right\} \subset \left\{|\bar{A}_k - m_k| \geq \sqrt{\delta}\right\} \cup \left\{\sum_{i=1}^n (A_{i,k} - m_k)^2 \leq 4n\delta\right\}. \tag{5.30}$$

Let us show that

$$\mathbb{P}(|\bar{A}_k - m_k| \geq \sqrt{\delta}) \leq 2C \exp\left\{-\frac{n\delta}{4d}\right\}. \tag{5.31}$$

$$\mathbb{P}(|\bar{A}_k - m_k| \geq \sqrt{\delta}) = \mathbb{P}(\bar{A}_k - m_k \geq \sqrt{\delta}) + \mathbb{P}(\bar{A}_k - m_k \leq -\sqrt{\delta})$$

By Chernoff inequality, for all $\lambda \geq 0$,

$$\mathbb{P}(n(\bar{A}_k - m_k) \geq n\sqrt{\delta}) \leq \exp\left\{-n\sqrt{\delta}\lambda + \log\left(\mathbb{E}[\exp(n(\bar{A}_k - m_k))]\right)\right\}$$
$$= \exp\left\{-n\sqrt{\delta}\lambda + n\log\left(\mathbb{E}[\exp(A_{i,k} - m_k)]\right)\right\}$$

Then, by Assumption **1.2**, for all positive values of λ in V_0,

$$\mathbb{P}(n(\bar{A}_k - m_k) \geq n\sqrt{\delta}) \leq C \exp\left\{-n\sqrt{\delta}\lambda + nd\lambda^2\right\}. \tag{5.32}$$

The right term of (5.32) is maximum when

$$\lambda = \frac{\sqrt{\delta}}{2d},$$

which implies that

$$\mathbb{P}(\bar{A}_k - m_k \geq \sqrt{\delta}) \leq C \exp\left\{-\frac{n\delta}{4d}\right\}.$$

Similarly, for all negative values of λ in V_0,

$$\mathbb{P}(n(\bar{A}_k - m_k) \leq -n\sqrt{\delta}) \leq C \exp\left\{n\sqrt{\delta}\lambda + nd\lambda^2\right\}. \tag{5.33}$$

The right term of (5.33) is maximum when

$$\lambda = -\frac{\sqrt{\delta}}{2d},$$

which implies that

$$\mathbb{P}(\bar{A}_k - m_k \geq \sqrt{\delta}) \leq C \exp\left\{-\frac{n\delta}{4d}\right\},$$

which proves (5.31).

$$\mathbb{P}(\sum_{i=1}^{n}(A_{i,k} - m_k)^2 \leq 4n\delta) \leq \mathbb{P}(\sum_{i=1}^{n}[(A_{i,k} - m_k)^2 - \sigma_k^2] \leq n(4\delta - \delta_{min}))$$

Since $4\delta - \delta_{min} < 0$ by assumption on δ, we apply again Chernoff inequality, which gives us that:

$$\mathbb{P}(\sum_{i=1}^{n}[(A_{i,k} - m_k)^2 - \sigma_k^2] \leq n(4\delta - \delta_{min})) \leq C \exp\left\{-n\frac{(4\delta - \delta_{min})^2}{2d}\right\}$$

This result, combined with (5.31), proves that

$$\mathbb{P}\left(\exists k, s_k^2 \leq \delta\right) \leq 2NC \exp\left\{-\frac{n\delta}{4d}\right\} + NC \exp\left\{-n\frac{(4\delta - \delta_{min})^2}{2d}\right\} \tag{5.34}$$

Notice that

$$\left\{\left|\sum_{k=1}^{N}(A_{i,k} - \bar{A}_k)^2 - s_k^2\right| \geq N\delta\epsilon_N\right\} = \left\{\frac{1}{n}\left|\sum_{k=1}^{N}\sum_{l=1}^{n}(A_{i,k} - \bar{A}_k)^2 - (A_{l,k} - \bar{A}_k)^2\right| \geq N\delta\epsilon_N\right\}$$
$$\subset \left\{\left|\sum_{k=1}^{N}(A_{i,k} - m_k)^2 - \sigma_k^2\right| \geq \frac{N\delta\epsilon_N}{4}\right\} \cup \left\{\left|\sum_{k=1}^{N}\sum_{l=1}^{n}(A_{l,k} - m_k)^2 - \sigma_k^2\right| \geq \frac{nN\delta\epsilon_N}{4}\right\}$$
$$\cup \left\{\left|\sum_{k=1}^{N}(A_{i,k} - m_k)(m_k - \bar{A}_k)\right| \geq \frac{N\delta\epsilon_N}{8}\right\} \cup \left\{\left|\sum_{k=1}^{N}\sum_{l=1}^{n}(A_{l,k} - m_k)(m_k - \bar{A}_k)\right| \geq \frac{nN\delta\epsilon_N}{8}\right\}$$

Using Chernoff inequality and Assumption **1.1**, we can prove that

$$\mathbb{P}\left(\left|\sum_{k=1}^{N}(A_{i,k}-m_k)^2-\sigma_k^2\right|\geq\frac{N\delta\epsilon_N}{4}\right)\leq 2C\exp\left\{-\frac{N\delta^2\epsilon_N^2}{64d}\right\}$$

and

$$\mathbb{P}\left(\left|\sum_{k=1}^{N}\sum_{l=1}^{n}(A_{l,k}-m_k)^2-\sigma_k^2\right|\geq\frac{nN\delta\epsilon_N}{4}\right)\leq 2C\exp\left\{-\frac{Nn\delta^2\epsilon_N^2}{64d}\right\}$$

Moreover,

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{k=1}^{n}(A_{i,k}-m_k)(m_k-\bar{A}_k)\right|\geq\frac{N\delta\epsilon_N}{4}\right)&\leq\mathbb{P}\left(\sum_{k=1}^{n}(A_{i,k}-m_k)^2\geq Nn\frac{\delta\epsilon_N}{8}\right)\\&+\mathbb{P}\left(\left|\sum_{k=1}^{n}\sum_{l\neq i}(A_{i,k}-m_k)(m_k-A_{l,k})\right|\geq nN\frac{\delta\epsilon_N}{8}\right)\end{aligned}$$

Using Chernoff inequality and Assumption **1.3**, we obtain that

$$\mathbb{P}\left(\left|\sum_{k=1}^{n}\sum_{l\neq i}(A_{i,k}-m_k)(m_k-A_{l,k})\right|\geq nN\frac{\delta\epsilon_N}{8}\right)\leq 2C\exp\left\{-\frac{nN\delta^2\epsilon_N^2}{256d}\right\}$$

and with Assumption **1.1** we have

$$\mathbb{P}\left(\sum_{k=1}^{n}(A_{i,k}-m_k)^2\geq Nn\frac{\delta\epsilon_N}{8}\right)\leq C\exp\left\{-\frac{n^2N\delta^2\epsilon_N^2}{256d}+\frac{nN\delta\delta_{max}\epsilon_N}{16d}-\frac{N\delta_{max}}{4d}\right\},\tag{5.35}$$

with $n^2N\epsilon_N^2=a^2N^{2+2\gamma}$ and $nN\epsilon_N=aN^{\frac{3}{2}+\gamma}$ where $\gamma>0$, which implies that the main term in the exponential is $-\frac{n^2N\delta^2\epsilon_N^2}{256d}$.

Similarly, we can show that

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{k=1}^{N}\sum_{l=1}^{n}(A_{l,k}-m_k)(m_k-\bar{A}_k)\right|\geq\frac{nN\delta\epsilon_N}{8}\right)&\leq 2C\exp\left\{-\frac{n^2N\delta^2\epsilon_N^2}{256d}\right\}\\&+C\exp\left\{-\frac{n^3N\delta^2\epsilon_N^2}{256d}+\frac{n^2N\delta\delta_{max}\epsilon_N}{16d}-\frac{Nn\delta_{max}}{4d}\right\}.\end{aligned}$$

This concludes the proof that for all values of q,

$$\mathbb{P}\left(|\sum_{k=1}^{n}(\mathbf{Z}_{i,k}^2-1)|\geq N\epsilon_N\right)=O\left(\frac{1}{N^q}\right).$$

We use similar techniques to otain an upper bound for $\mathbb{P}\left(\left|\sum_{k=1}^{n}\mathbf{Z}_{i,k}\mathbf{Z}_{j,k}\right| \geq N\epsilon_N\right)$.

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{k=1}^{n}\mathbf{Z}_{i,k}\mathbf{Z}_{j,k}\right| \geq N\epsilon_N\right) &= \mathbb{P}\left(\left|\sum_{k=1}^{n}\frac{(A_{i,k}-\bar{A}_k)(A_{j,k}-\bar{A}_k)}{s_k^2}\right| \geq N\epsilon_N\right)\\ &\leq \mathbb{P}(\exists k, s_k^2 \leq \delta) + \mathbb{P}\left(\left|\sum_{k=1}^{n}(A_{i,k}-\bar{A}_k)(A_{j,k}-\bar{A}_k)\right| \geq N\delta\epsilon_N\right)\end{aligned}$$

Since we have already proved (5.34) and (5.35), we will conclude the proof by showing that

$$\mathbb{P}\left(\left|\sum_{k=1}^{n}(A_{i,k}-m_k)(A_{j,k}-m_k)\right| \geq N\frac{\delta\epsilon_N}{4}\right) \leq 2C\exp\left\{-\frac{N\delta\epsilon_N}{64d}\right\}, \tag{5.36}$$

and

$$\mathbb{P}\left(\sum_{k=1}^{n}(\bar{A}_k-m_k)^2 \geq N\frac{\delta\epsilon_N}{4}\right) \leq N^2C\exp\left\{-\frac{N\delta\epsilon_N}{16d}\right\}. \tag{5.37}$$

(5.36) is obtained using Assumption **1.3** and Chernoff inequality.

$$\begin{aligned}\mathbb{P}\left(\sum_{k=1}^{n}(\bar{A}_k-m_k)^2 \geq N\frac{\delta\epsilon_N}{4}\right) &\leq \mathbb{P}\left(\sup_k (m_k-\bar{A}_k)^2 \geq \frac{\delta\epsilon_N}{4}\right)\\ &\leq N\sup_k \mathbb{P}\left((m_k-\bar{A}_k)^2 \geq \frac{\delta\epsilon_N}{4}\right)\\ &\leq N^2C\exp\left\{-\frac{N\delta\epsilon_N}{16d}\right\},\end{aligned}$$

which proves (5.37) and achieves the proof of Lemma 8.

Proof of Lemma 9

According to the results of Section 5.8.2, we have

$$\begin{aligned}\frac{1}{n}\sum_{i\neq j}\mathbb{E}[\mathbf{W}_i\mathbf{W}_j|\mathbf{Z},\epsilon_i=\epsilon_j=1]\mathbf{G}_N(i,j)\mathbb{1}_{E_N} &= \frac{1}{n}\sum_{i\neq j}(c\eta^\star\mathbf{G}_N(i,j)+R_N(i,j))\mathbf{G}_N(i,j)\mathbb{1}_{E_N}\\ &= ac\eta^\star + \frac{1}{n}\sum_{i\neq j}R_N(i,j)\mathbf{G}_N(i,j)\mathbb{1}_{E_N} + o_p(1)\end{aligned}$$

Thus, we just need to prove that $\sum_{i\neq j}\mathbf{G}_N(i,j)R_N(i,j)\mathbb{1}_{E_N} = o_p(1)$.
We shall see that $R_N(i,j)\mathbb{1}_{E_N}$ may be upper bounded by a finite sum of terms of the form

$$|\mathbf{G}_N(i,j)|^{k_1}|\mathbf{G}_N(i,i)-1|^{k_2}|\mathbf{G}_N(j,j)-1|^{k_3}, \tag{5.38}$$

with k in $[\![2,22]\!]$ and $k_1+k_2+k_3=k$.

Thus, $\frac{1}{n}\sum_{i\neq j}R_N(i,j)\mathbf{G}_N(i,j)\mathbb{1}_{E_N}$ is upper bounded by a finite sum of terms of the form

$$\frac{1}{n}\sum_{i\neq j}|\mathbf{G}_N(i,j)|^{k_1+1}|\mathbf{G}_N(i,i)-1|^{k_2}|\mathbf{G}_N(j,j)-1|^{k_3}.$$

But

$$\begin{aligned}\frac{1}{n}\sum_{i\neq j}|\mathbf{G}_N(i,j)|^{k_1+1}|\mathbf{G}_N(i,i)-1|^{k_2}|\mathbf{G}_N(j,j)-1|^{k_3}\mathbb{1}_{E_N} &\leq \epsilon_N^{k_1+k_2+k_3+1}\frac{n(n-1)}{n}\\ &= O\left(\frac{1}{N^{\frac{1}{2}-3\gamma}}\right)\\ &= o(1),\end{aligned}$$

since $k_1+k_2+k_3+1\geq 3$ and $\gamma<1/10$.

This achieves the proof of Lemma 9.

Let us explain why Equation (5.38) holds.

We need to evaluate $|R_N(i,j)\mathbb{1}_{E_N}|$. Then, let us look at the previous remainders which compose $R_N(i,j)$, and we will provide upper bounds when E_N holds.

$$|\alpha_N| = |A_N(i)A_N(j)-B_N(i,j)^2|\frac{\eta^{\star 2}}{N}+\frac{1}{2}|(A_N(i)+A_N(j))\frac{\eta^\star}{\sqrt{N}}+(A_N(i)A_N(j)-B_N(i,j)^2)\frac{\eta^{\star 2}}{N}|^2\frac{1}{|1+\tilde{\alpha}|^3},$$

with $|\tilde{\alpha}|\leq|(A_N(i)+A_N(j))\frac{\eta^\star}{\sqrt{N}}+(A_N(i)A_N(j)-B_N(i,j)^2)\frac{\eta^{\star 2}}{N}|\leq 2\epsilon_N\eta^\star+2\epsilon_N^2\eta^{\star 2}$.

Similarly,

$$\begin{aligned}|\beta_N| = \frac{1}{2}|A_N(i)A_N(j)-B_N(i,j)^2|\frac{\eta^{\star 2}}{N}+\frac{1}{2}|\frac{1}{2}(A_N(i)+A_N(j))\frac{\eta^\star}{\sqrt{N}}\\ +\frac{1}{2}(A_N(i)A_N(j)-B_N(i,j)^2)\frac{\eta^{\star 2}}{N}|^2\frac{3}{4}\frac{1}{|1+\tilde{\beta}|^{\frac{5}{2}}},\end{aligned}$$

with $|\tilde{\beta}|\leq|\frac{1}{2}(A_N(i)+A_N(j))\frac{\eta^\star}{\sqrt{N}}+\frac{1}{2}(A_N(i)A_N(j)-B_N(i,j)^2)\frac{\eta^{\star 2}}{N}|\leq\epsilon_N\eta^\star+\epsilon_N^2\eta^{\star 2}$.

The remainders γ_N,$\tilde{\gamma_N}$ and $\tilde{\tilde{\gamma}}_N$ are only products of α_N, $A_N(i)$, $A_N(j)$ and $B_N(i,j)$.

$|\gamma_N|\leq|A_N(j)(A_N(i)+A_N(j))\frac{\eta^{\star 2}}{N}|+|\alpha_N(1+A_N(j)\frac{\eta^\star}{\sqrt{N}})|$,

$|\tilde{\gamma_N}|\leq|A_N(i)(A_N(i)+A_N(j))\frac{\eta^{\star 2}}{N}|+|\alpha_N(1+A_N(i)\frac{\eta^\star}{\sqrt{N}})|$ and

$|\tilde{\tilde{\gamma}}_N|\leq|\frac{\eta^\star}{\sqrt{N}}B_N(i,j)|\left(|(A_N(i)+A_N(j))\frac{\eta^\star}{\sqrt{N}}|+|\alpha_N|\right)$

The next remainder is μ_N, which is defined as

$$\mu_N=\left(1-\frac{1}{2}(A_N(i)+A_N(j))\frac{\eta^\star}{\sqrt{N}}\right)\int_t^\infty\int_t^\infty\phi(x)\phi(y)\nu_N(x,y)dxdy,$$

with

$$\nu_N(x,y) = -\frac{x^2}{2}\gamma_N - \frac{y^2}{2}\tilde{\gamma_N} + xy\tilde{\tilde{\gamma}}_N + \frac{1}{2}\left(\frac{x^2}{2}\frac{\eta^\star}{\sqrt{N}}A_N(i) + \frac{y^2}{2}\frac{\eta^\star}{\sqrt{N}}A_N(j) + xy\frac{\eta^\star}{\sqrt{N}}B_N(i,j) - \frac{x^2}{2}\gamma_N - \frac{y^2}{2}\tilde{\gamma_N} + xy\tilde{\tilde{\gamma}}_N\right)^2 \exp\tilde{u}$$

Integrating the first terms of $\nu_N(x,y)$ gives

$$\int_t^\infty\int_t^\infty \phi(x)\phi(y)\left(-\frac{x^2}{2}\gamma_N - \frac{y^2}{2}\tilde{\gamma_N} + xy\tilde{\tilde{\gamma}}_N\right)dxdy = -\frac{1}{2}K(t\phi(t)+K)(\gamma_N+\tilde{\gamma_N}) + \phi(t)^2\tilde{\tilde{\gamma}}_N.$$

.
Moreover, we have the upper bound

$$\exp\tilde{u} \leq \max(\exp\left\{\frac{x^2}{2}\frac{\eta^\star}{\sqrt{N}}A_N(i) + \frac{y^2}{2}\frac{\eta^\star}{\sqrt{N}}A_N(j) + xy\frac{\eta^\star}{\sqrt{N}}B_N(i,j) + \frac{x^2}{2}\gamma_N - \frac{y^2}{2}\tilde{\gamma_N} - xy\tilde{\tilde{\gamma}}_N\right\},1).$$

If $\max(\exp\left\{\frac{x^2}{2}\frac{\eta^\star}{\sqrt{N}}A_N(i) + \frac{y^2}{2}\frac{\eta^\star}{\sqrt{N}}A_N(j) + xy\frac{\eta^\star}{\sqrt{N}}B_N(i,j) - \frac{x^2}{2}\gamma_N - \frac{y^2}{2}\tilde{\gamma_N} + xy\tilde{\tilde{\gamma}}_N\right\},1) = 1,$

$$\begin{aligned}
&\int_t^\infty\int_t^\infty \phi(x)\phi(y)\nu_N(x,y)dxdy \\
&\leq \int_t^\infty\int_t^\infty \phi(x)\phi(y)\left(\frac{x^2}{2}\frac{\eta^\star}{\sqrt{N}}A_N(i) + \frac{y^2}{2}\frac{\eta^\star}{\sqrt{N}}A_N(j) + xy\frac{\eta^\star}{\sqrt{N}}B_N(i,j) - \frac{x^2}{2}\gamma_N - \frac{y^2}{2}\tilde{\gamma_N} + xy\tilde{\tilde{\gamma}}_N\right)^2 dxdy \\
&= \frac{1}{N}J, \qquad (5.39)
\end{aligned}$$

where
$J = \int_t^\infty\int_t^\infty \phi(x)\phi(y)\left(\frac{x^2}{2}\eta^\star A_N(i) + \frac{y^2}{2}\eta^\star A_N(j) + xy\eta^\star B_N(i,j) - \frac{x^2}{2}\frac{\gamma_N}{\sqrt{N}} - \frac{y^2}{2}\frac{\tilde{\gamma_N}}{\sqrt{N}} + xy\frac{\tilde{\tilde{\gamma}}_N}{\sqrt{N}}\right)^2 dxdy$
is finite.
Otherwise,

$$\exp(\tilde{u}) \leq \exp\left\{\frac{x^2}{2}(\epsilon_N\eta^\star + P_1(\epsilon_N)) + \frac{y^2}{2}(\epsilon_N\eta^\star + P_2(\epsilon_N)) + xy(\epsilon_N\eta^\star + P_3(\epsilon_N))\right\},$$

where P_1, P_2, P_3 are polynomial functions. This expression comes from upper bounding the terms $A_N(i)/N$, $A_N(j)/N$ and $B_N(i,j)/N$ by ϵ_N in γ_N, $\tilde{\gamma}_N$ and $\tilde{\tilde{\gamma}}_N$.
There exists N_0, such that for all $N \geq N_0$, $\epsilon_N\eta^\star + P_1(\epsilon_N) \leq \frac{1}{4}$, $\epsilon_N\eta^\star + P_2(\epsilon_N) \leq \frac{1}{4}$ and $\epsilon_N\eta^\star + P_3(\epsilon_N) \leq \frac{1}{4}$.

Then $\exp(\tilde{u}) \leq \exp\left\{\frac{x^2}{8} + \frac{y^2}{8} + \frac{xy}{4}\right\} \leq \exp\left\{\frac{x^2}{4} + \frac{y^2}{4}\right\}$.
Thus, similarly to the expression 5.39,

$$\begin{aligned}\int_t^\infty \int_t^\infty \phi(x)\phi(y) \left(\frac{x^2}{2}\frac{\eta^\star}{\sqrt{N}}A_N(i) + \frac{y^2}{2}\frac{\eta^\star}{\sqrt{N}}A_N(j) + xy\frac{\eta^\star}{\sqrt{N}}B_N(i,j) + \frac{x^2}{2}\gamma_N - \frac{y^2}{2}\tilde{\gamma_N} - xy\tilde{\gamma_N}\right)^2 \exp(\tilde{u})dxdy \\ \leq \frac{1}{2\pi}\int_t^\infty \int_t^\infty \exp(-\frac{x^2}{4})\exp(-\frac{y^2}{4})\left(\frac{x^2}{2}\frac{\eta^\star}{\sqrt{N}}A_N(i)\right. \\ \left. + \frac{y^2}{2}\frac{\eta^\star}{\sqrt{N}}A_N(j) + xy\frac{\eta^\star}{\sqrt{N}}B_N(i,j) + \frac{x^2}{2}\gamma_N - \frac{y^2}{2}\tilde{\gamma_N} - xy\tilde{\gamma_N}\right)^2 dxdy \\ \leq \frac{1}{N}J'\end{aligned}$$

where J' is finite.
Similarly to the computations made for α_N, β_N, γ_N, μ_N, all the remainder terms can be upper bounded by products of $A_N(i)/\sqrt{N}$, $A_N(j)/\sqrt{N}$ and $B_N(i,j)/\sqrt{N}$, which proves (5.38).

Proof of Lemma 10

In this section, all the expectations that we consider are conditionally to the presence of the observed individuals in the study, for instance $\{\epsilon_i = \epsilon_j = 1\}$ or $\{\epsilon_{i_1} = \epsilon_{i_2} = \epsilon_{i_3} = 1\}$. However, for the sake of simplicity, we will not always make explicit such conditioning.
Let us show that

$$\mathrm{Var}(\frac{1}{n}\sum_{i\neq j}(\mathbf{W}_i\mathbf{W}_j - \mathbb{E}[\mathbf{W}_i\mathbf{W}_j|\mathbf{Z}])\mathbf{G}_N(i,j)\mathbb{1}_{E_N}) \to 0,$$

that is

$$\frac{1}{n^2}\sum_{\substack{i_1\neq i_2\\ i_3\neq i_4}} \mathbb{E}\left[(\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}\mathbf{W}_{i_3}\mathbf{W}_{i4}|\mathbf{Z}] - \mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}|\mathbf{Z}]\mathbb{E}[\mathbf{W}_{i_3}\mathbf{W}_{i_4}|\mathbf{Z}])\mathbf{G}_N(i_1,i_2)\mathbf{G}_N(i_3,i_4)\mathbb{1}_{E_N}\right] \to 0 \tag{5.40}$$

For this purpose, we will separate three cases depending on the cardinal of the set $\{i_1, i_2, i_3, i_4\}$ in the sum of Equation (5.40).
-If card($\{i_1, i_2, i_3, i_4\}$)=2, the corresponding terms in (5.40) are equal to

$$\frac{1}{n^2}\sum_{i\neq j}\mathbb{E}\left[\mathbb{E}[(\mathbf{W}_i^2\mathbf{W}_j^2|\mathbf{Z}] - \mathbb{E}[\mathbf{W}_i\mathbf{W}_j|\mathbf{Z}]^2)\mathbf{G}_N(i,j)^2\mathbb{1}_{E_N}\right] \leq \frac{1}{n^2}\sum_{i\neq j}\mathbb{E}[(\alpha + \rho_N(i,j))\mathbf{G}_N(i,j)^2\mathbb{1}_{E_N}]$$

where α is a positive constant and $\rho_N(i,j)$ can be upper bounded by a finite product of $\mathbf{G}_N(i,j)$, $\mathbf{G}_N(i,i) - 1$ and $\mathbf{G}_N(j,j) - 1$, according to proof of Lemma 9. This result is obtained by using a similar decomposition of $\mathbb{E}[\mathbf{W}_i^2\mathbf{W}_j^2|\mathbf{Z}]$ than the one that we explicited for $\mathbb{E}[\mathbf{W}_i\mathbf{W}_j|\mathbf{Z}]$.

Since $\mathbb{E}[\mathbf{G}_N(i,j)^2\mathbb{1}_{E_N}] \leq \epsilon_N^2$ and all terms of $\rho_N(i,j)$ are upper bounded by a finite sum of ϵ_N^k, with k greater than 1, which all tend to 0, it is clear that

$$\frac{1}{n^2}\sum_{i\neq j}\mathbb{E}\left[\mathbb{E}[(\mathbf{W}_i^2\mathbf{W}_j^2|\mathbf{Z}] - \mathbb{E}[\mathbf{W}_i\mathbf{W}_j|Z]^2)\mathbf{G}_N(i,j)^2\mathbb{1}_{E_N}\right] \to 0.$$

- If card($\{i_1,i_2,i_3,i_4\}$)=3, the corresponding terms in (5.40) are equal to

$$\frac{1}{n^2}\sum_{i_1\neq i_2\neq i_3}\mathbb{E}\left[(\mathbb{E}[\mathbf{W}_{i_1}^2\mathbf{W}_{i_2}\mathbf{W}_{i_3}|\mathbf{Z}] - \mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}|Z]\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_3}|Z])\mathbf{G}_N(i_1,i_2)\mathbf{G}_N(i_1,i_3)\mathbb{1}_{E_N}\right]. \tag{5.41}$$

Since the sum of Equation (5.41) has $n(n-1)(n-2)$ terms, we have the refine the upper bound that we used in the case where the cardinal of $\{i_1,i_2,i_3,i_4\}$ was equal to 2. Indeed, we will use the following proposition:

Proposition 2. $\mathbb{E}[\mathbf{W}_{i_1}^2\mathbf{W}_{i_2}\mathbf{W}_{i_3}|\mathbf{Z}]$ has no term of order less than $1/\sqrt{N}$, that is no constant term.

Let us explain why Proposition 2 is enough to prove

$$\frac{1}{n^2}\sum_{i_1\neq i_2\neq i_3}\mathbb{E}\left[(\mathbb{E}[\mathbf{W}_{i_1}^2\mathbf{W}_{i_2}\mathbf{W}_{i_3}|\mathbf{Z}] - \mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}|Z]\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_3}|Z])\mathbf{G}_N(i_1,i_2)\mathbf{G}_N(i_1,i_3)\mathbb{1}_{E_N}\right] \to 0. \tag{5.42}$$

Let us first recall that, according to Lemma 9,

$$\begin{aligned}\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}|Z]\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_3}|Z] = {} & c^2\eta^{\star 2}\mathbf{G}_N(i_1,i_2)\mathbf{G}_N(i_1,i_3) + c\eta^\star\mathbf{G}_N(i_1,i_3)R_N(i_1,i_2)\\ & + c\eta^\star\mathbf{G}_N(i_1,i_2)R_N(i_1,i_3) + R_N(i_1,i_2)R_N(i_1,i_3),\end{aligned}$$

where, if E_N holds, all these terms are upper bounded by a finite sum of terms of the form ϵ_N^k, with $k \geq 2$.
Then,

$$\mathbb{E}\left[\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}|Z]\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_3}|Z]\mathbf{G}_N(i_1,i_2)\mathbf{G}_N(i_1,i_3)\mathbb{1}_{E_N}\right]$$

can be upper bounded by a finite sum of terms of the form ϵ_N^k, with $k \geq 4$.
Since

$$\frac{N(N-1)(N-2)\epsilon_N^4}{n^2} \to 0,$$

it shows that

$$\frac{1}{n^2}\sum_{i_1\neq i_2\neq i_3}\mathbb{E}\left[\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}|Z]\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_3}|Z])\mathbb{1}_{E_N}\right]\to 0.$$

Similarly, according to Proposition 2, each term of $\mathbb{E}\left[\mathbb{E}[\mathbf{W}_{i_1}^2\mathbf{W}_{i_2}\mathbf{W}_{i_3}|\mathbf{Z}]\mathbf{G}_N(i_1,i_2)\mathbf{G}_N(i_1,i_3)\mathbb{1}_{E_N}\right]$ can be upper bounded by a finite sum of ϵ_N^k, with $k\geq 3$.
Since

$$\frac{n(n-1)(n-2)\epsilon_N^3}{n^2}=O\left(\frac{1}{N^{1/2-3\gamma}}\right)\to 0,$$

it achieves the proof of (5.41).
- If card($\{i_1,i_2,i_3,i_4\}$)=4, let us first observe that

$$\frac{N(N-1)(N-2)(N-3)\epsilon_N^5}{n^2}\to 0,$$

which means that we shall only focus on the approximation of
$\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}\mathbf{W}_{i_3}\mathbf{W}_{i4}|\mathbf{Z}]-\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}|Z]\mathbb{E}[\mathbf{W}_{i_3}\mathbf{W}_{i_4}|Z]$ of order $1/N$.
Let us recall that

$$\mathbb{E}\left[\mathbf{W}_{i_1}\mathbf{W}_{i_2}|Z\right]\mathbb{E}[\mathbf{W}_{i_3}\mathbf{W}_{i_4}|Z]=c^2\eta^{\star 2}\mathbf{G}_N(i_1,i_2)\mathbf{G}_N(i_3,i_4)+R_N(i_1,i_2,i_3,i_4),$$

where

$$R_N(i_1,i_2,i_3,i_4)=c\eta^\star\mathbf{G}_N(i_1,i_2)R_N(i_3,i_4)+c\eta^\star\mathbf{G}_N(i_3,i_4)R_N(i_1,i_2)+R_N(i_1,i_2)R_N(i_3,i_4)$$

is a remainder, each term of which is upper bounded by a finite sum of terms of the form ϵ_N^k, with $k\geq 2$. In particular, it implies that

$$\mathbb{E}[\frac{N(N-1)(N-2)(N-3)}{n^2}R_N(i_1,i_2,i_3,i_4)\mathbf{G}_N(i_1,i_2)\mathbf{G}_N(i_3,i_4)]\to 0.$$

Thus, we need to prove that

$$\frac{1}{n^2}\sum_{i_1\neq i_2\neq i_3}\mathbb{E}\left[(\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}\mathbf{W}_{i_3}\mathbf{W}_{i_4}|\mathbf{Z}]-c^2\eta^{\star 2}\mathbf{G}_N(i_1,i_2)\mathbf{G}_N(i_3,i_4))\mathbf{G}_N(i_1,i_2)\mathbf{G}_N(i_3,i_4)\mathbb{1}_{E_N}\right]\to 0, \tag{5.43}$$

To do so, we shall prove first the following proposition:

Proposition 3. The terms of order less than or equal to $1/\sqrt{N}$ in $\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}\mathbf{W}_{i_3}\mathbf{W}_{i4}|\mathbf{Z}]$ are null.

The term of order exactly $1/N$ in $\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}\mathbf{W}_{i_3}\mathbf{W}_{i4}|\mathbf{Z}]$ contains all combinations of products of two terms between $\mathbf{G}_N(i_1,i_2)$, $\mathbf{G}_N(i_1,i_3)$, $\mathbf{G}_N(i_1,i_4)$, $\mathbf{G}_N(i_2,i_3)$, $\mathbf{G}_N(i_2,i_4)$, $\mathbf{G}_N(i_3,i_4)$, $\mathbf{G}_N(i_1,i_1)-1$, $\mathbf{G}_N(i_2,i_2)-1$, $\mathbf{G}_N(i_3,i_3)-1$ and $\mathbf{G}_N(i_4,i_4)-1$.
We will demonstrate the propositions:

Proposition 4. The term in $\mathbf{G}_N(i_1,i_2)\mathbf{G}_N(i_3,i_4)$ of $\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}\mathbf{W}_{i_3}\mathbf{W}_{i4}|\mathbf{Z}]$ is equal to $c^2\eta^{\star 2}\mathbf{G}_N(i_1,i_2)\mathbf{G}_N(i_3,i_4)$.

Proposition 5. For all terms $T_N(i_1,i_2,i_3,i_4)$ of order $1/N$ in $\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}\mathbf{W}_{i_3}\mathbf{W}_{i4}|\mathbf{Z}]$,

$$\frac{1}{n^2}\mathbb{E}[T_N(i_1,i_2,i_3,i_4)\mathbf{G}_N(i_1,i_2)\mathbf{G}_N(i_3,i_4)]\to 0,$$

except for the term in $\mathbf{G}_N(i_1,i_2)\mathbf{G}_N(i_3,i_4)$.

Propositions 3, 4 and 5 prove (5.43).
Let us prove now Propositions 2, 3, 5 and 4.

If card($\{i_1,i_2,i_3,i_4\}$)=3, conditionally to $\{\epsilon_{i_1}=\epsilon_{i_2}=\epsilon_{i_3}=1\}$, $\mathbf{W}_{i_1}^2\mathbf{W}_{i_2}\mathbf{W}_{i_3}$ can take several values:

- $\frac{(1-P)^2}{P^2}$ if $\mathbf{Y}_{i_1}=\mathbf{Y}_{i_2}=\mathbf{Y}_{i_3}=1$.
- $\frac{-(1-P)}{P}$ if $\mathbf{Y}_{i_1}=1$ and $\mathbf{Y}_{i_2}\neq\mathbf{Y}_{i_3}$.
- 1 if $\mathbf{Y}_{i_1}=1$ and $\mathbf{Y}_{i_2}=\mathbf{Y}_{i_3}=0$ or $\mathbf{Y}_{i_1}=0$ and $\mathbf{Y}_{i_2}=\mathbf{Y}_{i_3}=1$.
- $\frac{-P}{1-P}$ if $\mathbf{Y}_{i_1}=0$ and $\mathbf{Y}_{i_2}\neq\mathbf{Y}_{i_3}$.
- $\frac{P^2}{(1-P)^2}$ if $\mathbf{Y}_{i_1}=\mathbf{Y}_{i_2}=\mathbf{Y}_{i_3}=0$.

Since each case has a probability 1 and each control a probability $K(1-P)/P(1-K)$ to be in the study (these probabilities are given in Equation (5.12) and (5.13)),

$$\begin{aligned}
&\mathbb{E}[\mathbf{W}_{i_1}^2\mathbf{W}_{i_2}\mathbf{W}_{i_3}|\mathbf{Z},\epsilon_{i_1}=\epsilon_{i_2}=\epsilon_{i_3}=1]=\frac{1}{\mathbb{P}(\epsilon_{i_1}=\epsilon_{i_2}=\epsilon_{i_3}=1)}\\
&\times\left\{\frac{(1-P)^2}{P^2}\mathbb{P}(\mathbf{Y}_{i_1}=\mathbf{Y}_{i_2}=\mathbf{Y}_{i_3}=1|\mathbf{Z})-\frac{1-P}{P}\left(\frac{K(1-P)}{P(1-K)}\right)\mathbb{P}(\mathbf{Y}_{i_1}=1,\mathbf{Y}_{i_2}\neq\mathbf{Y}_{i_3}|\mathbf{Z})\right.\\
&+\left(\frac{K(1-P)}{P(1-K)}\right)\mathbb{P}(\mathbf{Y}_{i_1}=0,\mathbf{Y}_{i_2}=\mathbf{Y}_{i_3}=1|\mathbf{Z})+\left(\frac{K(1-P)}{P(1-K)}\right)^2\mathbb{P}(\mathbf{Y}_{i_1}=1,\mathbf{Y}_{i_2}=\mathbf{Y}_{i_3}=0|\mathbf{Z})\\
&-\frac{P}{1-P}\left(\frac{K(1-P)}{P(1-K)}\right)^2\mathbb{P}(\mathbf{Y}_{i_1}=0,\mathbf{Y}_{i_2}\neq\mathbf{Y}_{i_3}|\mathbf{Z})\\
&\left.+\frac{(1-P)^2}{P^2}\left(\frac{K(1-P)}{P(1-K)}\right)^3\mathbb{P}(\mathbf{Y}_{i_1}=\mathbf{Y}_{i_2}=\mathbf{Y}_{i_3}=0|\mathbf{Z})\right\}
\end{aligned}\tag{5.44}$$

The computations leading to (5.44) are very similar to those leading to (5.15), which are detailed

in Appendix 5.B.
The development of order 0 of $\mathbb{P}(\mathbf{Y}_{i_1} = \mathbf{Y}_{i_2} = \mathbf{Y}_{i_3} = 1|\mathbf{Z})$ is

$$\frac{1}{(2\pi)^{3/2}} \int_t^{+\infty} \int_t^{+\infty} \int_t^{+\infty} \phi(x)\phi(y)\phi(z)dxdydz = K^3 + O_p\left(\frac{1}{\sqrt{N}}\right).$$

Similarly,

$$\mathbb{P}(\mathbf{Y}_{i_1} = 1, \mathbf{Y}_{i_2} \neq \mathbf{Y}_{i_3}|\mathbf{Z}) = 2K^2(1-K) + O_p\left(\frac{1}{\sqrt{N}}\right)$$

$$\mathbb{P}(\mathbf{Y}_{i_1} = 0, \mathbf{Y}_{i_2} = \mathbf{Y}_{i_3} = 1|\mathbf{Z}) = K^2(1-K) + O_p\left(\frac{1}{\sqrt{N}}\right)$$

$$\mathbb{P}(\mathbf{Y}_{i_1} = 1, \mathbf{Y}_{i_2} = \mathbf{Y}_{i_3} = 0|\mathbf{Z}) = K(1-K)^2 + O_p\left(\frac{1}{\sqrt{N}}\right)$$

$$\mathbb{P}(\mathbf{Y}_{i_1} = 0, \mathbf{Y}_{i_2} \neq \mathbf{Y}_{i_3}|\mathbf{Z}) = 2K(1-K)^2 + O_p\left(\frac{1}{\sqrt{N}}\right)$$

$$\mathbb{P}(\mathbf{Y}_{i_1} = \mathbf{Y}_{i_2} = \mathbf{Y}_{i_3} = 0|\mathbf{Z}) = (1-K)^3 + O_p\left(\frac{1}{\sqrt{N}}\right)$$

Replacing all these expressions in (5.44) gives us that the approximation of order 0 is null, which achieves the proof of Proposition 2.
Let us prove now Proposition 3.
If card($\{i_1, i_2, i_3, i_4\}$)=4, let us compute the approximation of order $1/\sqrt{N}$ of $\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}\mathbf{W}_{i_3}\mathbf{W}_{i_4}|\mathbf{Z}]$.
Conditionally to $\{\epsilon_{i_1} = \epsilon_{i_2} = \epsilon_{i_3} = \epsilon_{i_4} = 1\}$, $\mathbf{W}_{i_1}\mathbf{W}_{i_2}\mathbf{W}_{i_3}\mathbf{W}_{i4}$ can take values:

- $\frac{(1-P)^2}{P^2}$ if all individuals are cases, that is $\mathbf{Y}_{i_1} = \mathbf{Y}_{i_2} = \mathbf{Y}_{i_3} = \mathbf{Y}_{i_4} = 1$.
- $\frac{-(1-P)}{P}$ if one individual is a control and the three others are cases.
- 1 if two individuals are controls and two are cases.
- $\frac{-P}{1-P}$ if one individual is a case and the three others are controls.
- $\frac{P^2}{(1-P)^2}$ if all individuals are controls.

$$\begin{aligned}
&\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}\mathbf{W}_{i_3}\mathbf{W}_{i_4}|\mathbf{Z}, \epsilon_{i_1} = \epsilon_{i_2} = \epsilon_{i_3} = \epsilon_{i_4} = 1] = \frac{1}{\mathbb{P}(\epsilon_{i_1} = \epsilon_{i_2} = \epsilon_{i_3} = \epsilon_{i_4} = 1)} \\
&\times \left\{ \frac{(1-P)^2}{P^2}\mathbb{P}(\text{"4 cases"}|\mathbf{Z}) - \frac{1-P}{P}\left(\frac{K(1-P)}{P(1-K)}\right)\mathbb{P}(\text{"3 cases, 1 control"}|\mathbf{Z}) \right. \\
&+ \left(\frac{K(1-P)}{P(1-K)}\right)^2 \mathbb{P}(\text{"2 cases, 2 controls"}|\mathbf{Z}) - \frac{1-P}{P}\left(\frac{K(1-P)}{P(1-K)}\right)^3 \mathbb{P}(\text{"3 controls, 1 case"}|\mathbf{Z}) \\
&\left. - + \frac{(1-P)^2}{P^2}\left(\frac{K(1-P)}{P(1-K)}\right)^4 \mathbb{P}(\text{"4 controls"}|\mathbf{Z}) \right\} \qquad (5.45)
\end{aligned}$$

The covariance matrix of $(\mathbf{l}_{i_1}, \mathbf{l}_{i_2}, \mathbf{l}_{i_3}, \mathbf{l}_{i_4})$ is

$$\Sigma = \begin{pmatrix} 1+\eta^\star(\mathbf{G}_N(i_1,i_1)-1) & \mathbf{G}_N(i_1,i_2) & \mathbf{G}_N(i_1,i_3) & \mathbf{G}_N(i_1,i_4) \\ \mathbf{G}_N(i_1,i_2) & 1+\eta^\star(\mathbf{G}_N(i_2,i_2)-1) & \mathbf{G}_N(i_2,i_3) & \mathbf{G}_N(i_2,i_4) \\ \mathbf{G}_N(i_1,i_3) & \mathbf{G}_N(i_2,i_3) & 1+\eta^\star(\mathbf{G}_N(i_3,i_3)-1) & \mathbf{G}_N(i_3,i_4) \\ \mathbf{G}_N(i_1,i_4) & \mathbf{G}_N(i_2,i_4) & \mathbf{G}_N(i_3,i_4) & 1+\eta^\star(\mathbf{G}_N(i_4,i_4)-1) \end{pmatrix}.$$

For the sake of clarity, let us denote $A_1 = \frac{1}{\sqrt{N}}\sum_{k=1}^{N}(\mathbf{Z}_{i_1,k}^2 - 1) = \sqrt{N}(\mathbf{G}_N(i_1,i_1)-1)$, and similarly we define A_2, A_3 and A_4.
Let us also denote $C_{1,2} = \sqrt{N}\mathbf{G}_N(i_1,i_2)$ and similarly, $C_{1,3},\ldots,C_{3,4}$.
Then, let us rewrite Σ as:

$$\Sigma = \begin{pmatrix} 1+\frac{\eta^\star}{\sqrt{N}}A_1 & \frac{C_{1,2}}{\sqrt{n}} & \frac{C_{1,3}}{\sqrt{N}} & \frac{C_{1,4}}{\sqrt{N}} \\ \frac{C_{1,2}}{\sqrt{N}} & 1+\frac{\eta^\star}{\sqrt{N}}A_2 & \frac{C_{2,3}}{\sqrt{N}} & \frac{C_{2,4}}{\sqrt{N}} \\ \frac{C_{1,3}}{\sqrt{N}} & \frac{C_{2,3}}{\sqrt{N}} & 1+\frac{\eta^\star}{\sqrt{N}}A_3 & \frac{C_{3,4}}{\sqrt{N}} \\ \frac{C_{1,4}}{\sqrt{N}} & \frac{C_{2,4}}{\sqrt{N}} & \frac{C_{3,4}}{\sqrt{N}} & 1+\frac{\eta^\star}{\sqrt{N}}A_4 \end{pmatrix}.$$

The approximation of order $1/\sqrt{n}$ of its inverse matrix is given by

$$\Sigma^{-1} \simeq |\Sigma|^{-1} \times \begin{pmatrix} 1+\frac{\eta^\star}{\sqrt{N}}(A_2+A_3+A_4) & -\frac{C_{1,2}}{\sqrt{N}} & -\frac{C_{1,3}}{\sqrt{N}} & -\frac{C_{1,4}}{\sqrt{N}} \\ -\frac{C_{1,2}}{\sqrt{N}} & 1+\frac{\eta^\star}{\sqrt{N}}(A_1+A_3+A_4) & -\frac{C_{2,3}}{\sqrt{N}} & -\frac{C_{2,4}}{\sqrt{N}} \\ -\frac{C_{1,3}}{\sqrt{N}} & -\frac{C_{2,3}}{\sqrt{N}} & 1+\frac{\eta^\star}{\sqrt{N}}(A_1+A_2+A_4) & -\frac{C_{3,4}}{\sqrt{N}} \\ -\frac{C_{1,4}}{\sqrt{N}} & -\frac{C_{2,4}}{\sqrt{N}} & -\frac{C_{3,4}}{\sqrt{N}} & 1+\frac{\eta^\star}{\sqrt{n}}(A_1+A_2+A_2) \end{pmatrix}$$

where $|\Sigma|^{-1} = 1 - \frac{\eta^\star}{\sqrt{N}}(A_1+A_2+A_3+A_4) + O_p\left(\frac{1}{N}\right)$.
Let us compute

$$\mathbb{P}(\mathbf{Y}_{i_1} = \mathbf{Y}_{i_2} = \mathbf{Y}_{i_3} = \mathbf{Y}_{i_4} = 1|\mathbf{Z}) = \int_t^{+\infty}\int_t^{+\infty}\int_t^{+\infty}\int_t^{+\infty} f(w,x,y,z)dwdxdydz,$$

where

$$
\begin{aligned}
f(w,x,y,z) &= \frac{1}{(2\pi)^2|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{w^2}{2|\Sigma|}(1+\frac{\eta^\star}{\sqrt{n}}(A_2+A_3+A_4))-\cdots-\frac{z^2}{2|\Sigma|}(1+\frac{\eta^\star}{\sqrt{N}}(A_1+A_2+A_3))\right.\\
&\left.+\frac{wx}{|\Sigma|}\frac{\eta^\star}{\sqrt{n}}C_{1,2}+\frac{wy}{|\Sigma|}\frac{\eta^\star}{\sqrt{n}}C_{1,3}+\cdots+\frac{yz}{|\Sigma|}\frac{\eta^\star}{\sqrt{N}}C_{3,4}+O_p\left(\frac{1}{N}\right)\right\}\\
&= \frac{1}{(2\pi)^2|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{w^2}{2}(1-\frac{\eta^\star}{\sqrt{N}}(A_1+A_2+A_3+A_4))(1+\frac{\eta^\star}{\sqrt{N}}(A_2+A_3+A_4))-\ldots\right.\\
&\qquad -\frac{z^2}{2}(1-\frac{\eta^\star}{\sqrt{N}}(A_1+A_2+A_3+A_4))(1+\frac{\eta^\star}{\sqrt{n}}(A_1+A_2+A_3))\\
&\qquad +wx\frac{\eta^\star}{\sqrt{N}}(1-\frac{\eta^\star}{\sqrt{N}}(A_1+A_2+A_3+A_4))C_{1,2}+\ldots\\
&\qquad \left.+yz\frac{\eta^\star}{\sqrt{N}}(1-\frac{\eta^\star}{\sqrt{N}}(A_1+A_2+A_3+A_4))C_{3,4}+O_p\left(\frac{1}{N}\right)\right\}\\
&= \frac{1}{|\Sigma|^{\frac{1}{2}}}\phi(w)\phi(x)\phi(y)\phi(z)\exp\left\{\frac{w^2}{2}\frac{\eta^\star}{\sqrt{N}}A_1+\cdots+\frac{z^2}{2}\frac{\eta^\star}{\sqrt{N}}A_4\right.\\
&\qquad \left.-wx\frac{\eta^\star}{\sqrt{N}}C_{1,2}-\cdots-yz\frac{\eta^\star}{\sqrt{N}}C_{3,4}+O_p\left(\frac{1}{N}\right)\right\}\\
&= \frac{1}{|\Sigma|^{\frac{1}{2}}}\phi(w)\phi(x)\phi(y)\phi(z)\left[1+\frac{w^2}{2}\frac{\eta^\star}{\sqrt{N}}A_1+\cdots+\frac{z^2}{2}\frac{\eta^\star}{\sqrt{N}}A_4-wx\frac{\eta^\star}{\sqrt{N}}C_{1,2}-\ldots\right.\\
&\qquad \left.-yz\frac{\eta^\star}{\sqrt{N}}C_{3,4}+O_p\left(\frac{1}{N}\right)\right] \qquad (5.46)
\end{aligned}
$$

Finally,

- $$\mathbb{P}(\mathbf{Y}_{i_1}=\mathbf{Y}_{i_2}=\mathbf{Y}_{i_3}=\mathbf{Y}_{i_4}=1|\mathbf{Z}) = \frac{1}{|\Sigma|^{\frac{1}{2}}}\left[K^4+\frac{K^3}{2}(t\phi(t)+K)\frac{\eta^\star}{\sqrt{n}}(A_1+A_2+A_3+A_4)\right.$$
$$\left.+K^2\phi(t)^2\frac{\eta^\star}{\sqrt{n}}(C_{1,2}+\cdots+C_{3,4})+O_p\left(\frac{1}{N}\right)\right]$$

Similarly, we compute

- $$\mathbb{P}(\text{“1 control, 3 cases”}) = \frac{1}{|\Sigma|^{\frac{1}{2}}}\left[4K^3(1-K)\right.$$
$$+\frac{K^2}{2}\left((3-4K)t\phi(t)+4K(1-K)\right)\frac{\eta^\star}{\sqrt{N}}(A_1+A_2+A_3+A_4)$$
$$\left.+2\phi(t)^2K(1-2K)\frac{\eta^\star}{\sqrt{N}}(C_{1,2}+\cdots+C_{3,4})+O_p\left(\frac{1}{N}\right)\right]$$

- $\mathbb{P}(\text{“2 controls, 2 cases”}) = \frac{1}{|\Sigma|^{\frac{1}{2}}}\Big[6K^2(1-K)^2$
$+\frac{3K(1-K)}{2}\left((1-2K)t\phi(t)+2K(1-K)\right)\frac{\eta^\star}{\sqrt{N}}(A_1+A_2+A_3+A_4)$
$+\phi(t)^2(6K^2-6K+1)\frac{\eta^\star}{\sqrt{N}}(C_{1,2}+\cdots+C_{3,4})+O_p\left(\frac{1}{N}\right)\Big]$

- $\mathbb{P}(\text{“3 controls, 1 case”}) = \frac{1}{|\Sigma|^{\frac{1}{2}}}\Big[4K(1-K)^3$
$+\frac{(1-K)^2}{2}\left((1-4K)t\phi(t)+4K(1-K)\right)\frac{\eta^\star}{\sqrt{n}}(A_1+A_2+A_3+A_4)$
$-2\phi(t)^2(1-K)(1-2K)\frac{\eta^\star}{\sqrt{N}}(C_{1,2}+\cdots+C_{3,4})+O_p\left(\frac{1}{N}\right)\Big]$

- $\mathbb{P}(\mathbf{Y}_{i_1}=\mathbf{Y}_{i_2}=\mathbf{Y}_{i_3}=\mathbf{Y}_{i_4}=0|\mathbf{Z}) = \frac{1}{|\Sigma|^{\frac{1}{2}}}\Big[(1-K)^4$
$+\frac{(1-K)^3}{2}(-t\phi(t)+1-K)\frac{\eta^\star}{\sqrt{n}}(A_1+A_2+A_3+A_4)$
$+(1-K)^2\phi(t)^2\frac{\eta^\star}{\sqrt{N}}(C_{1,2}+\cdots+C_{3,4})+O_p\left(\frac{1}{N}\right)\Big]$

Regrouping all the first terms in the expression of $\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}\mathbf{W}_{i_3}\mathbf{W}_{i_4}|\mathbf{Z}]$ given in (5.45) leads to

$$\begin{aligned}&\frac{1}{|\Sigma|^{\frac{1}{2}}}\left[\frac{(1-P)^2}{P^2}K^4-\frac{(1-P)}{P}\left(\frac{K(1-P)}{P(1-K)}\right)4K^3(1-K)+\left(\frac{K(1-P)}{P(1-K)}\right)^2 6K^2(1-K)^2\right.\\&\left.-\frac{P}{1-P}\left(\frac{K(1-P)}{P(1-K)}\right)^3 4K(1-K)^3+\frac{P^2}{(1-P)^2}\left(\frac{K(1-P)}{P(1-K)}\right)^4(1-K)^4\right]\\&=\frac{1}{(2\pi)^2|\Sigma|^{\frac{1}{2}}}\left(\frac{(1-P)^2K^4}{P^2}\right)[1-4+6-4+1]=0\end{aligned}$$

Similarly we regroup the terms in $\frac{\eta^\star}{\sqrt{N}}(A_1+A_2+A_3+A_4)$:

$$\begin{aligned}
&\frac{1}{|\Sigma|^{\frac{1}{2}}}\frac{\eta^\star}{\sqrt{N}}(A_1+A_2+A_3+A_4)\left[\frac{(1-P)^2}{P^2}\frac{K^3}{2}(t\phi(t)+K)\right.\\
&\quad-\frac{(1-P)}{P}\left(\frac{K(1-P)}{P(1-K)}\right)\frac{K^2}{2}\left((3-4K)t\phi(t)+4K(1-K)\right)\\
&\quad+\left(\frac{K(1-P)}{P(1-K)}\right)^2\frac{3K(1-K)}{2}\left((1-2K)t\phi(t)+2K(1-K)\right)\\
&\quad-\frac{P}{1-P}\left(\frac{K(1-P)}{P(1-K)}\right)^3\frac{(1-K)^2}{2}\left((1-4K)t\phi(t)+4K(1-K)\right)\\
&\quad\left.+\frac{P^2}{(1-P)^2}\left(\frac{K(1-P)}{P(1-K)}\right)^4\frac{(1-K)^3}{2}(-t\phi(t)+1-K)\right]\\
&\quad=\frac{1}{(2\pi)^2|\Sigma|^{\frac{1}{2}}}\left(\frac{(1-P)^2K^4}{2P^2}\right)[1-4+6-4+1]\\
&\quad+\frac{1}{(2\pi)^2|\Sigma|^{\frac{1}{2}}}\left(\frac{(1-P)^2K^3}{2P^2(1-K)}\right)[1-K-3+4K+3(1-2K)-1+4K-K]=0
\end{aligned}$$

Finally, we regroup all the terms in $\frac{\eta^\star}{\sqrt{n}}(C_{1,2}+\cdots+C_{3,4})$:

$$\begin{aligned}
&\frac{1}{|\Sigma|^{\frac{1}{2}}}\left(\frac{(1-P)^2K^2}{P^2(1-K)^2}\right)\phi(t)^2\left[(1-K)^2-2(1-K)(1-2K)+6K^2-6K+1+2K(1-2K)+K^2\right]\\
&\qquad\qquad=0
\end{aligned}$$

This proves Proposition 3.

Let us prove Proposition 5.
Let us denote $f_2(w,x,y,z)$ the density function defined in (5.46) developed till order $1/N$.

$$
\begin{aligned}
f_2(w,x,y,z) = \frac{1}{|\Sigma|^{\frac{1}{2}}}\phi(w)\phi(x)\phi(y)\phi(z)\Bigg[1 &+ \frac{w^2}{2}(\frac{\eta^\star}{\sqrt{N}}A_1 - \frac{\eta^{\star 2}}{N}(A_1^2 + C_{1,2}^2 + C_{1,3}^2 + C_{1,2}^2) \\
&+ \cdots + \frac{z^2}{2}(\frac{\eta^\star}{\sqrt{N}}A_4 - \frac{\eta^{\star 2}}{N}(A_4^2 + C_{1,4}^2 + C_{2,4}^2 + C_{3,4}^2) \\
&+ wx(C_{1,2}\frac{\eta^\star}{\sqrt{N}} - \frac{\eta^{\star 2}}{N}[(A_1 + A_2)C_{1,2} + C_{1,3}C_{2,3} + C_{1,4}C_{2,4}]) + \dots \\
&+ yz(C_{3,4}\frac{\eta^\star}{\sqrt{N}} - \frac{\eta^{\star 2}}{N}[(A_3 + A_4)C_{3,4} + C_{1,3}C_{1,4} + C_{2,3}C_{2,4}]) \\
&+ \frac{w^4}{8}\frac{\eta^{\star 2}}{N}A_1^2 + \cdots + \frac{z^4}{8}\frac{\eta^{\star 2}}{N}A_4^2 + \frac{w^2x^2}{2}\frac{\eta^{\star 2}}{N}(C_{1,2}^2 + \frac{A_1A_2}{2}) + \dots \\
&+ \frac{y^2z^2}{2}\frac{\eta^{\star 2}}{N}(C_{3,4}^2 + \frac{A_3A_4}{2}) + \frac{w^3x}{2}\frac{\eta^{\star 2}}{N}A_1C_{1,2} + \dots \\
&+ \frac{z^3y}{2}\frac{\eta^{\star 2}}{N}A_4C_{3,4} + w^2xy\frac{\eta^{\star 2}}{N}[\frac{A_1C_{2,3}}{2} + C_{1,2}C_{1,3}] + \dots \\
&+ z^2xy\frac{\eta^{\star 2}}{N}[\frac{A_4C_{2,3}}{2} + C_{2,4}C_{3,4}] + wxyz\frac{\eta^{\star 2}}{N}(C_{1,2}C_{3,4} + C_{2,3}C_{1,4} + C_{1,3}C_{2,4})\Bigg]
\end{aligned} \tag{5.47}
$$

In order to prove Proposition 5, we will show that:

$$\frac{1}{n^2}\sum_{i_1\neq i_2\neq i_3\neq i_4}\mathbb{E}(A_1^2C_{1,2}C_{3,4}) \to 0 \tag{5.48}$$

$$\frac{1}{n^2}\sum_{i_1\neq i_2\neq i_3\neq i_4}\mathbb{E}(A_1A_2C_{1,2}C_{3,4}) \to 0 \tag{5.49}$$

$$\frac{1}{n^2}\sum_{i_1\neq i_2\neq i_3\neq i_4}\mathbb{E}(A_1C_{1,2}^2C_{3,4}) \to 0 \tag{5.50}$$

$$\frac{1}{n^2}\sum_{i_1\neq i_2\neq i_3\neq i_4}\mathbb{E}(A_1C_{1,2}C_{13}C_{3,4}) \to 0 \tag{5.51}$$

$$\frac{1}{n^2}\sum_{i_1\neq i_2\neq i_3\neq i_4}\mathbb{E}(C_{1,2}^2C_{2,3}C_{3,4}) \to 0 \tag{5.52}$$

$$\frac{1}{n^2}\sum_{i_1\neq i_2\neq i_3\neq i_4}\mathbb{E}(C_{1,2}C_{1,3}C_{2,4}C_{3,4}) \to 0 \tag{5.53}$$

$$\frac{1}{n^2}\sum_{i_1\neq i_2\neq i_3\neq i_4}\mathbb{E}(C_{1,2}^3C_{3,4}) \to 0 \tag{5.54}$$

We will develop the proof of Equation (5.49).
By exchangeability of the $(\mathbf{Z}_{i,k})_{1\leq i\leq n}$, we can write

$$\begin{aligned}\mathbb{E}[A_1A_2C_{1,2}C_{3,4}] &= \sum_{k,l,m,r}\mathbb{E}[(\mathbf{Z}_{1,k}^2-1)(\mathbf{Z}_{2,l}^2-1)\mathbf{Z}_{1,m}\mathbf{Z}_{2,m}\mathbf{Z}_{3,r}\mathbf{Z}_{4,r}]\\ &= \sum_{k,l,m,r}\mathbb{E}[\mathbf{Z}_{1,k}^2\mathbf{Z}_{2,l}^2\mathbf{Z}_{1,m}\mathbf{Z}_{2,m}\mathbf{Z}_{3,r}\mathbf{Z}_{4,r}] - 2N\sum_{k,m,r}\mathbb{E}[\mathbf{Z}_{1,k}^2\mathbf{Z}_{1,m}\mathbf{Z}_{2,m}\mathbf{Z}_{3,r}\mathbf{Z}_{4,r}]\\ &+ N^2\sum_{m,r}\mathbb{E}[\mathbf{Z}_{1,m}\mathbf{Z}_{2,m}\mathbf{Z}_{3,r}\mathbf{Z}_{4,r}] \qquad (5.55)\end{aligned}$$

We recall that since $\mathbf{Z}_{i,k}$ and $\mathbf{Z}_{j,l}$ are independent for any i and j when $k \neq l$, we will always consider separately the cases where $k = l$ from the cases $k \neq l$. Let us first focus on the last term of (5.55).

$$\begin{aligned}\sum_{m,r}\mathbb{E}[\mathbf{Z}_{1,m}\mathbf{Z}_{2,m}\mathbf{Z}_{3,r}\mathbf{Z}_{4,r}] &= \sum_{m=1}^{N}\mathbb{E}[\mathbf{Z}_{1,m}\mathbf{Z}_{2,m}\mathbf{Z}_{3,m}\mathbf{Z}_{4,m}] + \sum_{m\neq r}\mathbb{E}[\mathbf{Z}_{1,m}\mathbf{Z}_{2,m}]\mathbb{E}[\mathbf{Z}_{3,r}\mathbf{Z}_{4,r}]\\ &= N\times O\left(\frac{1}{n^2}\right) + N(N-1)\times\frac{1}{(n-1)^2}\end{aligned}$$

Then,
$\frac{1}{n^2}\frac{1}{N^4}\sum_{i_1\neq i_2\neq i_3\neq i_4}(N^2\sum_{m,r}\mathbb{E}[\mathbf{Z}_{1,m}\mathbf{Z}_{2,m}\mathbf{Z}_{3,r}\mathbf{Z}_{4,r}]) = \frac{(N-1)(n-2)(n-3)}{Nn(n-1)} + o(1)$
Now let us decompose the second term of (5.55) as:

$$\begin{aligned}&\sum_{k,m,r}\mathbb{E}[\mathbf{Z}_{1,k}^2\mathbf{Z}_{1,m}\mathbf{Z}_{2,m}\mathbf{Z}_{3,r}\mathbf{Z}_{4,r}] = \sum_{k=1}^{N}\mathbb{E}[\mathbf{Z}_{1,k}^3\mathbf{Z}_{2,k}\mathbf{Z}_{3,k}\mathbf{Z}_{4,k}] + \sum_{k\neq l}\mathbb{E}[\mathbf{Z}_{1,k}^3\mathbf{Z}_{2,k}]\mathbb{E}[\mathbf{Z}_{3,l}\mathbf{Z}_{4,l}]\\ &+\sum_{k\neq l}\mathbb{E}[\mathbf{Z}_{1,k}^2]\mathbb{E}[\mathbf{Z}_{1,l}\mathbf{Z}_{2,l}\mathbf{Z}_{3,l}\mathbf{Z}_{4,l}] + \sum_{k\neq l}\mathbb{E}[\mathbf{Z}_{1,k}^2\mathbf{Z}_{3,k}\mathbf{Z}_{3,k}]\mathbb{E}[\mathbf{Z}_{1,l}\mathbf{Z}_{2,l}] + \sum_{k\neq l\neq m}\mathbb{E}[\mathbf{Z}_{1,k}^2]\mathbb{E}[\mathbf{Z}_{1,l}\mathbf{Z}_{2,l}]\mathbb{E}[\mathbf{Z}_{3,m}\mathbf{Z}_{4,m}]\end{aligned}$$

Using the results given by Proposition 1, we obtain that

$$\frac{1}{n^2}\frac{1}{N^4}\left(-2N\sum_{k,m,r}\mathbb{E}[\mathbf{Z}_{1,k}^2\mathbf{Z}_{1,m}\mathbf{Z}_{2,m}\mathbf{Z}_{3,r}\mathbf{Z}_{4,r}]\right) = -\frac{2(N-1)(n-2)(n-3)}{Nn(n-1)} + o(1).$$

Similarly, we can prove that

$$\frac{1}{n^2}\frac{1}{N^4}\left(\sum_{k,l,m,r}\mathbb{E}[\mathbf{Z}_{1,k}^2\mathbf{Z}_{2,l}^2\mathbf{Z}_{1,m}\mathbf{Z}_{2,m}\mathbf{Z}_{3,r}\mathbf{Z}_{4,r}]\right) = \frac{(N-1)(n-2)(n-3)}{Nn(n-1)} + o(1),$$

by using the properties of Proposition 1 or similar relationships coming from other properties of $\mathbf{Z}$ that we have not detailed here.
Hence we have shown (5.49). The proofs of (5.48), (5.50), (5.51), (5.52), (5.53), (5.54) are very

similar to this proof.
It remains to prove Proposition 4.
According to the expression of $f_2(w,x,y,z)$ given in (5.47) and since

$$|\Sigma|^{-\frac{1}{2}} = 1 - \frac{\eta^\star}{2\sqrt{N}}(A_1+A_2+A_3+A_4) + \frac{\eta^{\star 2}}{4N}(A_1A_2+\cdots+A_3A_4) + \frac{3\eta^{\star 2}}{8N}(A_1^2+A_2^2+A_3^2+A_4^2) + \frac{\eta^{\star 2}}{2N}(C_{1,2}^2+\cdots+C_{3,4}^2) + O_p\left(\frac{1}{N^{\frac{3}{2}}}\right),$$

the only term in $C_{1,2}C_{3,4}$ of $\mathbb{P}(\mathbf{Y}_{i_1}=\mathbf{Y}_{i_2}=\mathbf{Y}_{i_3}=\mathbf{Y}_{i_4}=1|\mathbf{Z})$ is

$$\frac{1}{(2\pi)^2}\frac{\eta^{\star 2}}{N}\int_t^{+\infty}\int_t^{+\infty}\int_t^{+\infty}\int_t^{+\infty} wxyz\phi(w)\phi(x)\phi(y)\phi(z)C_{1,2}C_{3,4}dwdxdydz = \phi(t)^4C_{1,2}C_{3,4}\frac{\eta^{\star 2}}{N}.$$

The term in $C_{1,2}C_{3,4}$ of $\mathbb{P}(\text{“3 cases, 1 control”}|\mathbf{Z})$ is

$$-4\phi(t)^4C_{1,2}C_{3,4}\frac{\eta^{\star 2}}{N}.$$

The term in $C_{1,2}C_{3,4}$ of $\mathbb{P}(\text{“2 cases, 2 controls”}|\mathbf{Z})$ is

$$6\phi(t)^4C_{1,2}C_{3,4}\frac{\eta^{\star 2}}{N}.$$

The term in $C_{1,2}C_{3,4}$ of $\mathbb{P}(\text{“1 case, 3 controls”}|\mathbf{Z})$ is

$$-4\phi(t)^4C_{1,2}C_{3,4}\frac{\eta^{\star 2}}{N}.$$

The term in $C_{1,2}C_{3,4}$ of $\mathbb{P}(Y_{i_1}=Y_{i_2}=Y_{i_3}=Y_{i_4}=0|\mathbf{Z})$ is

$$\phi(t)^4C_{1,2}C_{3,4}\frac{\eta^{\star 2}}{N}.$$

It remains to compute the approximation of the denominator of $\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}\mathbf{W}_{i_3}\mathbf{W}_{i_4}|\mathbf{Z}]$ of order 0, that is

$$\begin{aligned}
&K^4 + 4K^3(1-K)\left(\frac{K(1-P)}{P(1-K)}\right) + 6K^2(1-K)^2\left(\frac{K(1-P)}{P(1-K)}\right)^2 \\
&\quad + 4K(1-K)^3\left(\frac{K(1-P)}{P(1-K)}\right)^3 + (1-K)^4\left(\frac{K(1-P)}{P(1-K)}\right)^4 \\
&= \frac{K^4}{P^4}\left[P^4 + 4P^3(1-P) + 6P^2(1-P)^2 + 4P(1-P)^3 + (1-P)^4\right] \\
&= \frac{K^4}{P^4}.
\end{aligned}$$

Finally, the term $C_{1,2}C_{3,4}$ in $\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}\mathbf{W}_{i_3}\mathbf{W}_{i4}|\mathbf{Z}]$ is

$$\begin{aligned}\phi(t)^4\frac{\eta^{\star 2}}{N}C_{1,2}C_{3,4}\Bigg[&\frac{(1-P)^2}{P^2}+2\frac{1-P}{P}\left(\frac{K(1-P)}{P(1-K)}\right)+6\left(\frac{K(1-P)}{P(1-K)}\right)^2\\&+2\frac{P}{1-P}\left(\frac{K(1-P)}{P(1-K)}\right)^3+\left(\frac{K(1-P)}{P(1-K)}\right)^4\Bigg]\times\frac{P^4}{K^4}\\&=\frac{P^2(1-P)^2}{K^4(1-K)^4}\phi(t)^4\frac{\eta^{\star 2}}{N}C_{1,2}C_{3,4},\end{aligned}$$

which is exactly the term in $C_{1,2}C_{3,4}$ of $\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}|Z]\mathbb{E}[\mathbf{W}_{i_3}\mathbf{W}_{i_4}|Z]$.
This proves Proposition 4.

5.8.3 Second order approximation of $\mathbb{E}[\mathbf{W}_i\mathbf{W}_j|\mathbf{Z},\epsilon_i=\epsilon_j=1]$

The density function f can still be written as

$$f(x,y)=\frac{1}{2\pi|\Sigma^{(N)}|^{-\frac{1}{2}}}\exp\left\{-\frac{1}{2|\Sigma^{(N)}|}\left[x^2(1+\frac{\eta^\star}{\sqrt{N}}A_N(j))+y^2(1+\frac{\eta^\star}{\sqrt{N}}A_N(i))-2xy\frac{B_N(i,j)}{\sqrt{N}}\right]\right\},$$

but with the explicit term of order $1/N$ in the expressions of $|\Sigma^{(N)}|^{-1}$ and $|\Sigma^{(N)}|^{-\frac{1}{2}}$:

$$\begin{aligned}|\Sigma^{(N)}|^{-1}&=1-\frac{\eta^\star}{\sqrt{N}}(A_N(i)+A_N(j))+\frac{\eta^{\star 2}}{N}\left(-A_N(i)A_N(j)+B_N(i,j)^2+(A_N(i)+A_N(j))^2\right)+O_p\left(\frac{1}{N^{\frac{3}{2}}}\right)\\&=1-\frac{\eta^\star}{\sqrt{N}}(A_N(i)+A_N(j))+\frac{\eta^{\star 2}}{N}\left(A_N(i)A_N(j)+A_N(i)^2+A_N(j)^2+B_N(i,j)^2\right)+O_p\left(\frac{1}{N^{\frac{3}{2}}}\right)\end{aligned}$$

and

$$\begin{aligned}|\Sigma^{(N)}|^{-\frac{1}{2}}&=1-\frac{\eta^\star}{2\sqrt{N}}(A_N(i)+A_N(j))+\frac{\eta^{\star 2}}{2N}\left(-A_N(i)A_N(j)+B_N(i,j)^2+\frac{3}{8}(A_N(i)+A_N(j))^2\right)\\&\quad+O_p\left(\frac{1}{N^{\frac{3}{2}}}\right).\end{aligned}$$

Thus,

$$\begin{aligned}
&\frac{1}{2\pi}\exp\left\{-\frac{1}{2|\Sigma^{(N)}|}\left[x^2(1+\frac{\eta^\star}{\sqrt{N}}A_N(j))+y^2(1+\frac{\eta^\star}{\sqrt{N}}A_N(i))-2xy\frac{B_N(i,j)}{\sqrt{N}}\right]\right\}\\
&=\phi(x)\phi(y)\exp\left\{-\frac{x^2}{2}(-A_N(i)\frac{\eta^\star}{\sqrt{N}}+\frac{\eta^{\star 2}}{N}(A_N(i)^2+B_N(i,j)^2)\right.\\
&\qquad -\frac{y^2}{2}(-A_N(j)\frac{\eta^\star}{\sqrt{N}}+\frac{\eta^{\star 2}}{N}(A_N(j)^2+B_N(i,j)^2))\\
&\qquad \left.+xy(\frac{\eta^\star}{\sqrt{N}}B_N(i,j)-\frac{\eta^{\star 2}}{N}B_N(i,j)(A_N(i)+A_N(j)))\right\}+O_p\left(\frac{1}{N^{\frac{3}{2}}}\right)\\
&=\phi(x)\phi(y)\left[1+\frac{x^2}{2}(A_N(i)\frac{\eta^\star}{\sqrt{N}}-\frac{\eta^{\star 2}}{N}(A_N(i)^2+B_N(i,j)^2)+\frac{y^2}{2}(A_N(j)\frac{\eta^\star}{\sqrt{N}}\right.\\
&\qquad -\frac{\eta^{\star 2}}{N}(A_N(j)^2+B_N(i,j)^2))+\frac{x^4}{8}\frac{\eta^{\star 2}}{N}A_N(i)^2+\frac{y^4}{8}\frac{\eta^{\star 2}}{N}A_N(j)^2\\
&\qquad \left.+xy(\frac{\eta^\star}{\sqrt{N}}B_N(i,j)-\frac{\eta^{\star 2}}{N}B_N(i,j)(A_N(i)+A_N(j)))+\frac{x^2y^2}{2}\frac{\eta^{\star 2}}{N}B_N(i,j)^2+O_p\left(\frac{1}{N^{\frac{3}{2}}}\right)\right]
\end{aligned}$$

with the last term obtained by developing the exponential function.
Since

$$\int_t^\infty\int_t^\infty x^4dxdy=t^3\phi(t)+3t\phi(t)+3K$$

and

$$\int_t^\infty\int_t^\infty x^2y^2dxdy=(t\phi(t)+K)^2$$

$$\begin{aligned}
&\frac{1}{2\pi}\int_t^\infty\int_t^\infty\exp\left\{-\frac{1}{2|\Sigma^{(N)}|}\left[x^2(1+\frac{\eta^\star}{\sqrt{N}}A_N(j))+y^2(1+\frac{\eta^\star}{\sqrt{N}}A_N(i))-2xy\frac{B_N(i,j)}{\sqrt{N}}\right]\right\}dxdy\\
&=K^2+\frac{K}{2}(t\phi(t)+K)\left[\frac{\eta^\star}{\sqrt{N}}(A_N(i)+A_N(j))-\frac{\eta^{\star 2}}{N}(A_N(i)^2+A_N(j)^2+2B_N(i,j)^2)\right]\\
&+\frac{K}{8}\frac{\eta^{\star 2}}{N}(t^3\phi(t)+3t\phi(t)+3K)(A_N(i)^2+A_N(j)^2)+\phi(t)^2\left[\frac{\eta^\star}{\sqrt{N}}B_N(i,j)-\frac{\eta^{\star 2}}{N}B_N(i,j)(A_N(i)+A_N(j))\right]\\
&+\frac{1}{2}(t\phi(t)+K)^2\frac{\eta^{\star 2}}{N}(B_N(i,j)^2+\frac{A_N(i)A_N(j)}{2}+\frac{\phi(t)^2}{2}\frac{\eta^{\star 2}}{N}(t^2+2)B_N(i,j)(A_N(i)+A_N(j))+O_p\left(\frac{1}{N^{\frac{3}{2}}}\right)
\end{aligned}$$

Multiplying by

$$|\Sigma^{(N)}|^{-\frac{1}{2}} = 1 - \frac{\eta^\star}{2\sqrt{N}}(A_N(i) + A_N(j)) + \frac{\eta^{\star 2}}{2N}\left(-A_N(i)A_N(j) + B_N(i,j)^2 + \frac{3}{4}(A_N(i) + A_N(j))^2\right) + O_p\left(\frac{1}{N^{\frac{3}{2}}}\right),$$

we obtain

$$\begin{aligned}\int_t^\infty \int_t^\infty f(x,y)dxdy = K^2 &+ \frac{K}{2}t\phi(t)\frac{\eta^\star}{\sqrt{N}}(A_N(i) + A_N(j)) + \phi(t)^2\frac{\eta^\star}{\sqrt{N}}B_N(i,j) + \frac{K}{8}\frac{\eta^{\star 2}}{N}(t^3\phi(t) - 3t\phi(t)) \\ &+ \frac{\eta^{\star 2}}{N}\frac{t^2\phi(t)^2}{4}A_N(i)A_N(j) + \frac{\eta^{\star 2}}{N}B_N(i,j)^2\frac{t^2}{2}\phi(t)^2 \\ &+ \frac{\eta^{\star 2}}{N}\frac{\phi(t)^2}{2}(t^2-1)B_N(i,j)(A_N(i) + A_N(j))\phi(t)^2 + O_p\left(\frac{1}{N^{\frac{3}{2}}}\right).\end{aligned}$$

Similarly,

$$\int_{-\infty}^t \int_{-\infty}^t x^4 dxdy = -t^3\phi(t) - 3t\phi(t) + 3(1-K)$$

and

$$\int_{-\infty}^t \int_{-\infty}^t x^2y^2 dxdy = (-t\phi(t) + 1 - K)^2.$$

Then we have

$$\begin{aligned}&\int_{-\infty}^t \int_{-\infty}^t \exp\left\{-\frac{1}{2|\Sigma^{(N)}|}\left[x^2(1 + \frac{\eta^\star}{\sqrt{N}}A_N(j)) + y^2(1 + \frac{\eta^\star}{\sqrt{N}}A_N(i)) - 2xy\frac{B_N(i,j)}{\sqrt{N}}\right]\right\}dxdy \\ &= (1-K)^2 + \frac{1-K}{2}(-t\phi(t) + 1 - K)\left[\frac{\eta^\star}{\sqrt{N}}(A_N(i) + A_N(j)) - \frac{\eta^{\star 2}}{N}(A_N(i)^2 + A_N(j)^2 + 2B_N(i,j)^2)\right] \\ &\quad + \frac{1-K}{8}\frac{\eta^{\star 2}}{N}\left(-t^3\phi(t) - 3t\phi(t) + 3(1-K)\right)(A_N(i)^2 + A_N(j)^2) \\ &\quad + \frac{\phi(t)^2}{2}\frac{\eta^{\star 2}}{N}(t^2+2)B_N(i,j)(A_N(i) + A_N(j))\end{aligned}$$

Multiplying by

$$|\Sigma^{(N)}|^{-\frac{1}{2}} = 1 - \frac{\eta^\star}{\sqrt{N}}(A_N(i) + A_N(j)) + \frac{\eta^{\star 2}}{2N}\left(-A_N(i)A_N(j) + B_N(i,j)^2 + \frac{3}{4}(A_N(i) + A_N(j))^2\right) + O_p\left(\frac{1}{N^{\frac{3}{2}}}\right),$$

we obtain that

$$\int_{-\infty}^{t}\int_{-\infty}^{t} f(x,y)dxdy = (1-K)^2 - \frac{1-K}{2}t\phi(t)\frac{\eta^\star}{\sqrt{N}}(A_N(i)+A_N(j)) + \phi(t)^2\frac{\eta^\star}{\sqrt{N}}B_N(i,j)$$
$$+ \frac{1-K}{8}\frac{\eta^{\star 2}}{N}(A_N(i)^2+A_N(j)^2)(-t^3\phi(t)+3t\phi(t)) + \frac{\eta^{\star 2}}{N}\frac{t^2\phi(t)^2}{4}A_N(i)A_N(j)$$
$$+ \frac{\eta^{\star 2}}{N}B_N(i,j)^2\frac{t^2}{2}\phi(t)^2 - \frac{\eta^{\star 2}}{N}B_N(i,j)(A_N(i)+A_N(j))\frac{\phi(t)^2}{2}(t^2-1) + O_p\left(\frac{1}{N^{\frac{3}{2}}}\right).$$

Finally, we compute similarly $\mathbb{P}(\mathbf{Y}_i \neq \mathbf{Y}_j|\mathbf{Z}) = \int_{-\infty}^{t}\int_{t}^{+\infty} f(x,y)dxdy + \int_{t}^{+\infty}\int_{-\infty}^{t} f(x,y)dxdy$. We obtain

$$\mathbb{P}(\mathbf{Y}_i \neq \mathbf{Y}_j|\mathbf{Z}) = 2K(1-K) - \frac{1-2K}{2}t\phi(t)\frac{\eta^\star}{\sqrt{N}}(A_N(i)+A_N(j)) - 2\phi(t)^2\frac{\eta^\star}{\sqrt{N}}B_N(i,j)$$
$$+ \frac{1-2K}{8}\frac{\eta^{\star 2}}{N}(A_N(i)^2+A_N(j)^2)(t^3\phi(t)-3t\phi(t)) - \frac{\eta^{\star 2}}{N}\frac{t^2\phi(t)^2}{2}A_N(i)A_N(j)$$
$$- \frac{\eta^{\star 2}}{N}B_N(i,j)^2t^2\phi(t)^2 + \frac{\eta^{\star 2}}{N}B_N(i,j)(A_N(i)+A_N(j))\phi(t)^2(-t^2+1) + O_p\left(\frac{1}{N^{\frac{3}{2}}}\right)$$

We replace the expressions of $\mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 1|\mathbf{Z})$, $\mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 0|\mathbf{Z})$ and $\mathbb{P}(\mathbf{Y}_i \neq \mathbf{Y}_j|\mathbf{Z})$ in the expression of $\mathbb{E}(\mathbf{W}_i\mathbf{W}_j|\mathbf{Z}, \epsilon_i = \epsilon_j = 1)$. Since we already computed the terms of order $\frac{1}{\sqrt{N}}$ for the numerator, it only remains the terms of order $\frac{1}{N}$.
Eventually, we find that the numerator can be writen as :

$$\frac{\eta^\star}{\sqrt{N}}\frac{1-P}{P(1-K)^2}\phi(t)^2B_N(i,j) + \frac{\eta^{\star 2}}{N}\frac{t^2\phi(t)^2}{4}A_N(i)A_N(j)\frac{1-P}{P(1-K)^2} + \frac{\eta^{\star 2}}{2N}B_N(i,j)^2\frac{1-P}{P(1-K)^2}t^2\phi(t)^2$$
$$+ \frac{\eta^{\star 2}}{N}\frac{\phi(t)^2}{2}\frac{1-P}{P(1-K)^2}(t^2-1)B_N(i,j)(A_N(i)+A_N(j)) + O_p\left(\frac{1}{N^{\frac{3}{2}}}\right)$$

Similarly, we compute the expression of the denominator (at order $\frac{1}{\sqrt{N}}$ since the main term of the numerator is of order $\frac{1}{\sqrt{N}}$). We obtain the following expression:

$$\frac{K^2}{P^2} + \frac{\eta^\star}{\sqrt{N}}\frac{t}{2}\phi(t)(A_N(i)+A_N(j))\frac{K(P-K)}{P^2(1-K)} + \frac{\eta^\star}{\sqrt{N}}\phi(t)^2B_N(i,j)\frac{(P-K)^2}{P^2(1-K)^2} + O_p\left(\frac{1}{N}\right)$$

$$
\begin{aligned}
&\mathbb{E}(\mathbf{W}_i\mathbf{W}_j|\mathbf{Z},\epsilon_i=\epsilon_j=1)\\
&=\frac{P^2}{K^2}\left[1-\frac{\eta^\star}{\sqrt{N}}\frac{t}{2}\phi(t)(A_N(i)+A_N(j))\frac{(P-K)}{K(1-K)}-\frac{\eta^\star}{\sqrt{N}}\phi(t)^2B_N(i,j)\frac{(P-K)^2}{K^2(1-K)^2}\right]\\
&\quad\times\left[\frac{\eta^\star}{\sqrt{N}}\frac{1-P}{P(1-K)^2}\phi(t)^2B_N(i,j)+\frac{\eta^{\star 2}}{N}\frac{t^2\phi(t)^2}{4}A_N(i)A_N(j)\frac{1-P}{P(1-K)^2}\right.\\
&\qquad\left.+\frac{\eta^{\star 2}}{2N}B_N(i,j)^2\frac{1-P}{P(1-K)^2}t^2\phi(t)^2-\frac{\eta^{\star 2}}{N}\frac{\phi(t)^2}{2}\frac{1-P}{P(1-K)^2}(t^2-1)B_N(i,j)(A_N(i)+A_N(j))\right]\\
&\quad+O_p\left(\frac{1}{N^{\frac{3}{2}}}\right)\\
&=\frac{\eta^\star}{\sqrt{N}}\frac{P(1-P)}{K^2(1-K)^2}\phi(t)^2B_N(i,j)+\frac{t^2}{4}\frac{\eta^{\star 2}}{N}A_N(i)A_N(j)\frac{P(1-P)}{K^2(1-K)^2}\\
&\quad+\frac{\eta^{\star 2}}{N}\frac{P(1-P)}{K^2(1-K)^2}\phi(t)^2B_N(i,j)^2\left[\frac{t^2}{2}-\frac{(P-K)^2}{K^2(1-K)^2}\right]\\
&\quad+\frac{\eta^{\star 2}}{2N}\frac{P(1-P)}{K^2(1-K)^2}\phi(t)^2B_N(i,j)(A_N(i)+A_N(j))\left[t^2-1-\frac{P-K}{K(1-K)}t\phi(t)\right]
\end{aligned}
$$

Appendix

5.A Proof of Equation (5.13)

By definition, the probabilities p_{case} and $p_{control}$ are linked to the variables ϵ_i as follows:

$$p_{case} = \mathbb{P}(\epsilon_i = 1 | \mathbf{Z}, \mathbf{Y}_i = 1)$$

and

$$p_{control} = \mathbb{P}(\epsilon_i = 1 | \mathbf{Z}, \mathbf{Y}_i = 0).$$

The ratio of the two following equations:

$$P = \mathbb{P}(\mathbf{Y}_i = 1 | \epsilon_i = 1) = \frac{\mathbb{P}(\mathbf{Y}_i = 1, \epsilon_i = 1)}{\mathbb{P}(\epsilon_i = 1)} = \frac{\mathbb{P}(\mathbf{Y}_i = 1, V_i = 1)}{\mathbb{P}(\epsilon_i = 1)} = \frac{K p_{case}}{\mathbb{P}(\epsilon_i = 1)}$$

and

$$1 - P = \mathbb{P}(\mathbf{Y}_i = 0 | \epsilon_i = 1) = \frac{\mathbb{P}(\mathbf{Y}_i = 0, \epsilon_i = 1)}{\mathbb{P}(\epsilon_i = 1)} = \frac{\mathbb{P}(\mathbf{Y}_i = 0, U_i = 1)}{\mathbb{P}(\epsilon_i = 1)} = \frac{(1 - K) p_{control}}{\mathbb{P}(\epsilon_i = 1)},$$

with the full ascertainment assumption given by (5.12) prove equation (5.13).

5.B Proof of Equation (5.15)

This equation was proved in Golan et al. (2014), we recall the proof here for the sake of completeness.
Conditionally to the event $\{\epsilon_i = \epsilon_j = 1\}$, the variable $\mathbf{W}_i \mathbf{W}_j$ can take the following values:

- $\frac{1-p}{p}$ if $\mathbf{Y}_i = \mathbf{Y}_j = 1$.
- $\frac{p}{1-p}$ if $\mathbf{Y}_i = \mathbf{Y}_j = 0$.
- -1 if $\mathbf{Y}_i \neq \mathbf{Y}_j$.

Let us write the expectaction of $\mathbf{W}_i \mathbf{W}_j$ conditionally to $\mathbf{Z}$ and conditionally to $\{\epsilon_i = \epsilon_j = 1\}$:

$$\begin{aligned}\mathbb{E}(\mathbf{W}_i \mathbf{W}_j | \mathbf{Z}, \epsilon_i = \epsilon_j = 1) = \frac{1 - P}{P} \mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 1 | \mathbf{Z}, \epsilon_i = \epsilon_j = 1) - \mathbb{P}(\mathbf{Y}_i \neq \mathbf{Y}_j | \mathbf{Z}, \epsilon_i = \epsilon_j = 1) \\ + \frac{P}{1 - P} \mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 0 | \mathbf{Z}, \epsilon_i = \epsilon_j = 1). \qquad (5.56)\end{aligned}$$

$$\mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 1|\mathbf{Z}, \epsilon_i = \epsilon_j = 1) = \frac{\mathbb{P}(\epsilon_i = \epsilon_j = 1|\mathbf{Y}_i = \mathbf{Y}_j = 1, \mathbf{Z})\mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 1|\mathbf{Z})}{\mathbb{P}(\epsilon_i = \epsilon_j = 1|\mathbf{Z})}$$
$$= \frac{\mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 1|\mathbf{Z})}{\mathbb{P}(\epsilon_i = \epsilon_j = 1|\mathbf{Z})}$$

under the full ascertainment assumption given by Equation (5.12).

Similarly, since we have seen in Equation (5.13) that a control has a probability $\frac{K(1-P)}{P(1-K)}$ to be selected in the study and since ϵ_i and ϵ_j are assumed to be independent conditionally to $\mathbf{Z}$, $\mathbf{Y}_i$ and $\mathbf{Y}_j$:

$$\mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 0|\mathbf{Z}, \epsilon_i = \epsilon_j = 1) = \frac{\mathbb{P}(\epsilon_i = \epsilon_j = 1|\mathbf{Y}_i = \mathbf{Y}_j = 0, \mathbf{Z})\mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 0|\mathbf{Z})}{\mathbb{P}(\epsilon_i = \epsilon_j = 1|\mathbf{Z})}$$
$$= \left(\frac{K(1-P)}{P(1-K)}\right)^2 \frac{\mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 1|\mathbf{Z})}{\mathbb{P}(\epsilon_i = \epsilon_j = 1|\mathbf{Z})}$$

and

$$\mathbb{P}(\mathbf{Y}_i \neq \mathbf{Y}_j|\mathbf{Z}, \epsilon_i = \epsilon_j = 1) = \frac{\mathbb{P}(\epsilon_i = \epsilon_j = 1|\mathbf{Y}_i \neq \mathbf{Y}_j, \mathbf{Z})\mathbb{P}(\mathbf{Y}_i \neq \mathbf{Y}_j|\mathbf{Z})}{\mathbb{P}(\epsilon_i = \epsilon_j = 1|\mathbf{Z})}$$
$$= \left(\frac{K(1-P)}{P(1-K)}\right) \frac{\mathbb{P}(\mathbf{Y}_i \neq \mathbf{Y}_j|\mathbf{Z})}{\mathbb{P}(\epsilon_i = \epsilon_j = 1|\mathbf{Z})}.$$

The probability that both individuals i and j are included in the study is equal to

$$\begin{aligned}\mathbb{P}(\epsilon_i = \epsilon_j = 1|\mathbf{Z}) &= \mathbb{P}(\epsilon_i = \epsilon_j = 1|\mathbf{Z}, \mathbf{Y}_i = \mathbf{Y}_j = 1)\mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 1|\mathbf{Z}) \\ &+ \mathbb{P}(\epsilon_i = \epsilon_j = 1|\mathbf{Z}, \mathbf{Y}_i = \mathbf{Y}_j = 0)\mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 0|\mathbf{Z}) \\ &+ \mathbb{P}(\epsilon_i = \epsilon_j = 1|\mathbf{Z}, \mathbf{Y}_i \neq \mathbf{Y}_j)\mathbb{P}(\mathbf{Y}_i \neq \mathbf{Y}_j|\mathbf{Z}) \\ &= \mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 1|\mathbf{Z}) + \left(\frac{K(1-P)}{P(1-K)}\right)^2 \mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 0|\mathbf{Z}) + \left(\frac{K(1-P)}{P(1-K)}\right) \mathbb{P}(\mathbf{Y}_i \neq \mathbf{Y}_j|\mathbf{Z}).\end{aligned}$$

If we combine all these computations and we plug them in the expression (5.56), we obtain (5.15).

5.C Proof of Equation (5.19)

Notice first that

$$\mathbf{G}_N(i,i) - 1 = \frac{1}{\sqrt{N}} \left(\frac{1}{\sqrt{N}} \sum_{k=1}^{N} (\mathbf{Z}_{i,k}^2 - 1) \right)$$

with

$$\text{Var}\left(\frac{1}{\sqrt{N}} \sum_{k=1}^{N} (\mathbf{Z}_{i,k}^2 - 1) \right) = \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}(\mathbf{Z}_{i,k}^4) - (\mathbb{E}(\mathbf{Z}_{i,k}^2))^2.$$

Moreover, since the variables $(\mathbf{Z}_{i,k})_{1 \leq i \leq n}$ are normalized according to Equation (5.9),

$$\sum_{i=1}^{N} \mathbf{Z}_{i,k}^2 = n.$$

By taking the expectation and since the variables $(\mathbf{Z}_{i,k})_{1\leq i\leq n}$ are exchangeable, we obtain that

$$\mathbb{E}[\mathbf{Z}_{i,k}^2] = 1. \tag{5.57}$$

Using (2) of Proposition 1 and Equation (5.57), we obtain that

$$\mathrm{Var}\left(\frac{1}{\sqrt{N}}\sum_{k=1}^{N}(\mathbf{Z}_{i,k}^2 - 1)\right)$$

is bounded and

$$\mathbf{G}_N(i,i) - 1 = \frac{1}{N}\sum_{k=1}^{N}(\mathbf{Z}_{i,k}^2 - 1) = O_p\left(\frac{1}{\sqrt{N}}\right).$$

Similarly,

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{\sqrt{N}}\sum_{k=1}^{N}\mathbf{Z}_{i,k}\mathbf{Z}_{j,k}\right) &= \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}(\mathbf{Z}_{i,k}^2\mathbf{Z}_{j,k}^2) - \mathbb{E}(\mathbf{Z}_{i,k}\mathbf{Z}_{j,k})^2 \\
&= \frac{1}{N}\sum_{k=1}^{N}\left(1 + o(1) - \frac{1}{(n-1)^2}\right) \text{ using (3) and (1) of Proposition 1} \\
&= 1 + o(1).
\end{aligned}$$

Then, $\frac{1}{N}\sum_{k=1}^{N}\mathbf{Z}_{i,k}\mathbf{Z}_{j,k} = O_p\left(\frac{1}{\sqrt{N}}\right)$.

Thus, we can write

$$\Sigma^{(N)} = \begin{pmatrix} 1 + \frac{A_N(i)}{\sqrt{N}}\eta^\star & \frac{B_N(i,j)}{\sqrt{N}}\eta^\star \\ \frac{B_N(i,j)}{\sqrt{N}}\eta^\star & 1 + \frac{A_N(j)}{\sqrt{N}}\eta^\star \end{pmatrix},$$

where $A_N(i) = \frac{1}{\sqrt{N}}\sum_{k=1}^{N}(\mathbf{Z}_{i,k}^2 - 1) = O_p(1)$ for all i,

and $B_N(i,j) = \frac{1}{\sqrt{N}}\sum_{k=1}^{N}\mathbf{Z}_{i,k}\mathbf{Z}_{j,k} = O_p(1)$ for all $i \neq j$.

5.D Proof of Proposition 1

Observe that for all $k = 1, \ldots, N$,

$$\sum_{i=1}^{n}\mathbf{Z}_{i,k} = 0 \tag{5.58}$$

and

$$\sum_{i=1}^{n} \mathbf{Z}_{i,k}^2 = n. \tag{5.59}$$

Moreover, for each k, the random variables $(\mathbf{Z}_{i,k})_{1\leq i\leq n}$ are exchangeable. Thus, we deduce from (5.59) that for all $i = 1, \ldots, n$ and $k = 1, \ldots, N$, $\mathbb{E}(\mathbf{Z}_{i,k}^2) = 1$. Hence, by (5.58), we get that

$$0 = \left(\sum_{i=1}^{n} \mathbf{Z}_{i,k}\right)^2 = \sum_{i=1}^{n} \mathbf{Z}_{i,k}^2 + \sum_{1\leq i\neq j\leq n} \mathbf{Z}_{i,k}\mathbf{Z}_{j,k} \ ,$$

which, by (5.59), implies that for all $k = 1, \ldots, N$ and $i \neq j = 1, \ldots, n$,

$$\mathbb{E}(\mathbf{Z}_{i,k}\mathbf{Z}_{j,k}) = -\frac{n}{n(n-1)} = -\frac{1}{n-1} \ , \tag{5.60}$$

that is (1).
The proof of (2) comes from the decomposition:

$$\begin{aligned} |\mathbf{Z}_{1,k}|^p &= |\mathbf{Z}_{1,k}|^p \mathbb{1}_{\{s_k^2 > \frac{\delta_{min}}{2}\}} + |\mathbf{Z}_{1,k}|^p \mathbb{1}_{\{s_k^2 \leq \frac{\delta_{min}}{2}\}} \\ &\leq \frac{|A_{1,k} - \bar{A}_k|^p}{\left(\frac{\delta_{min}}{2}\right)^p} + n^p \mathbb{1}_{\{s_k^2 \leq \frac{\delta_{min}}{2}\}} \end{aligned}$$

Assumption **1.2** implies that $\sup_k \mathbb{E}\left[|A_{1,k} - \bar{A}_k|^p\right] < +\infty$ and the upper bound for $\mathbb{P}(s_k^2 \leq \delta)$ of Equation (5.34) prove (2).
By (5.59), for all $k = 1, \ldots, N$,

$$n^2 = \left(\sum_{i=1}^{n} \mathbf{Z}_{i,k}^2\right)^2 = \sum_{i=1}^{n} \mathbf{Z}_{i,k}^4 + \sum_{1\leq i\neq j\leq n} \mathbf{Z}_{i,k}^2\mathbf{Z}_{j,k}^2$$

Since the $(\mathbf{Z}_{i,k})_{1\leq i\leq n}$ are exchangeable for each $k = 1, \ldots, N$, we get that for all $k = 1, \ldots, N$,

$$n = \mathbb{E}[\mathbf{Z}_{1,k}^4] + (n-1)\mathbb{E}[\mathbf{Z}_{1,k}^2\mathbf{Z}_{2,k}^2] \ ,$$

which gives us (3) by using (2).
If we take the expectation of

$$\mathbf{Z}_{1,k}^3 \sum_{i=1}^{n} \mathbf{Z}_{i,k} = 0,$$

we obtain

$$\mathbb{E}[\mathbf{Z}_{1,k}^4] + (n-1)\mathbb{E}[\mathbf{Z}_{1,k}^3\mathbf{Z}_{2,k}] = 0.$$

Then, (2) implies (4).
Similarly, since

$$\mathbf{Z}_{1,k}\mathbf{Z}_{2,k} \sum_{i=1}^{n} \mathbf{Z}_{i,k}^2 = n\mathbf{Z}_{1,k}\mathbf{Z}_{2,k},$$

we obtain that

$$2\mathbb{E}[\mathbf{Z}_{1,k}^3\mathbf{Z}_{2,k}] + (n-2)\mathbb{E}[\mathbf{Z}_{1,k}^2\mathbf{Z}_{2,k}\mathbf{Z}_{3,k}] = n\mathbb{E}[\mathbf{Z}_{1,k}\mathbf{Z}_{2,k}].$$

Then (1) and (4) imply (5).
Since

$$\mathbf{Z}_{1,k}\mathbf{Z}_{2,k}\mathbf{Z}_{3,k}\sum_{i=1}^{n}\mathbf{Z}_{i,k} = 0,$$

we obtain that

$$3\mathbb{E}[\mathbf{Z}_{1,k}^2\mathbf{Z}_{2,k}\mathbf{Z}_{3,k}] + (n-3)\mathbb{E}[\mathbf{Z}_{1,k}\mathbf{Z}_{2,k}\mathbf{Z}_{3,k}\mathbf{Z}_{4,k}] = 0.$$

Then, (5) implies (6).
Since

$$\mathbf{Z}_{1,k}^5\sum_{i=1}^{n}\mathbf{Z}_{i,k} = 0,$$

we obtain that
Then, (2) implies (7).
The proof of (8) is very similar to the proof of (2) but we use Assumption **1.3** which gives us that $\sup_k \mathbb{E}\left[|(A_{1,k}-\bar{A}_k)(A_{2,k}-\bar{A}_k)|^p\right] < +\infty$.
Since

$$\mathbf{Z}_{1,k}^4\sum_{i=1}^{n}\mathbf{Z}_{i,k}^2 = n\mathbf{Z}_{1,k}^4,$$

$$\mathbb{E}[\mathbf{Z}_{1,k}^6] + (n-1)\mathbb{E}[\mathbf{Z}_{1,k}^4\mathbf{Z}_{2,k}^2] = n\mathbb{E}[\mathbf{Z}_{1,k}^4].$$

Then (2) implies (9).
Similarly, since

$$\mathbf{Z}_{1,k}^4\mathbf{Z}_{2,k}\sum_{i=1}^{n}\mathbf{Z}_{i,k} = 0,$$

we obtain that

$$\mathbb{E}[\mathbf{Z}_{1,k}^5\mathbf{Z}_{2,k}] + \mathbb{E}[\mathbf{Z}_{1,k}^4\mathbf{Z}_{2,k}^2] + (n-2)\mathbb{E}[\mathbf{Z}_{1,k}^4\mathbf{Z}_{2,k}\mathbf{Z}_{3,k}] = 0.$$

Then, (7) and (9) imply (10).
Since

$$\mathbf{Z}_{1,k}^3\mathbf{Z}_{2,k}\sum_{i=1}^{n}\mathbf{Z}_{i,k}^2 = n\mathbf{Z}_{1,k}^3\mathbf{Z}_{2,k},$$

we obtain that

$$\mathbb{E}[\mathbf{Z}_{1,k}^5\mathbf{Z}_{2,k}] + \mathbb{E}[\mathbf{Z}_{1,k}^3\mathbf{Z}_{2,k}^3] + (n-2)\mathbb{E}[\mathbf{Z}_{1,k}^3\mathbf{Z}_{2,k}^2\mathbf{Z}_{3,k}] = n\mathbb{E}[\mathbf{Z}_{1,k}^3\mathbf{Z}_{2,k}].$$

Then, (7), (8) and (4) imply (11).
Finally, since $\mathbf{Z}_{1,k}^3 \mathbf{Z}_{2,k} (\sum_{i=1}^{n} \mathbf{Z}_{i,k})^2 = 0$,

$$\begin{aligned}&\mathbb{E}[\mathbf{Z}_{1,k}^5 \mathbf{Z}_{2,k}] + \mathbb{E}[\mathbf{Z}_{1,k}^3 \mathbf{Z}_{2,k}^3] + 2\mathbb{E}[\mathbf{Z}_{1,k}^4 \mathbf{Z}_{2,k}^2] + 2(n-2)\mathbb{E}[\mathbf{Z}_{1,k}^4 \mathbf{Z}_{2,k} \mathbf{Z}_{3,k}] \\ &+ 2(n-2)\mathbb{E}[\mathbf{Z}_{1,k}^3 \mathbf{Z}_{2,k}^2 \mathbf{Z}_{3,k}] + (n-2)^2 \mathbb{E}[\mathbf{Z}_{1,k}^3 \mathbf{Z}_{2,k} \mathbf{Z}_{3,k} \mathbf{Z}_{4,k}] = 0\end{aligned}$$

Then, (7), (8), (9), (10) and (11) imply (12).

Chapitre 6

Résumé en français

6.1 Contexte biologique

Tous les caractères biologiques, par exemple la taille ou le poids, sont influencés à la fois par des facteurs génétiques et environnementaux. Quantifier ces deux contributions pour un trait donné est une problématique difficile et fondamentale en biologie. Le concept d'héritabilité désigne la part de la variabilité d'un trait observé (ou phénotype) qui peut être attribuée à des causes génétiques.

Plusieurs erreurs concernant l'héritabilité sont dues à l'utilisation du terme dans le langage commun, mais avec un sens différent du terme technique utilisé en génétique. Par exemple, une idée reçue serait de penser que l'héritabilité définit la proportion d'un phénotype qui est transmis des parents aux enfants. Tout d'abord, ce ne sont pas les phénotypes qui sont transmis d'une génération à l'autre mais les gènes. De plus, si la moitié des effets génétiques sont effectivement transmis par chaque parent, cette moitié est spécifique à chaque enfant.Visscher et al. (2008) a regroupé ces fréquentes erreurs et idées reçues concernant l'héritabilité. Nous allons à présent détailler le concept d'héritabilité tel qu'il est utilisé dans le domaine de la génétique.

6.1.1 Définition de l'héritabilité

En reprenant les explications élégantes de Visscher et al. (2008), nous considérons la modélisation simple où un phénotype d'intérêt est décrit comme le résultat d'effets génétiques et environnementaux, considérés comme indépendants :

$$\text{Phénotype (P)} = \text{Génotype (G)} + \text{Environnement (E)}.$$

La variance phénotypique (σ_P^2) peut être décomposée comme la somme des variances non observées (σ_G^2 et σ_E^2) :

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2.$$

L'héritabilité (H^2) est définie mathématiquement comme un rapport de variances et exprime la proportion de la variance phénotypique qui peut être attribuée à des facteurs génétiques :

$$H^2 = \frac{\sigma_G^2}{\sigma_P^2}.$$

La variabilité génétique peut être partitionnée selon différentes sources, en particulier, la variance σ_A^2 des effets génétiques additifs. Ces effets additifs sont caractérisés par l'impact des polymorphismes de nucléotide simple (SNPs), qui sont des différences de la séquence d'ADN aux positions du génome auxquelles il existe une variabilité relativement importante dans la population. Ces positions sont en réalité peu fréquentes par rapport à l'intégralité du génome qui contient environ 3 milliards de paires de bases, parmi lesquelles la grande majorité est identique chez tous les êtres humains.

Dans la suite, nous allons considérer l'héritabilité "au sens strict", c'est-à-dire la proportion de variabilité expliquée seulement par les effets génétiques additifs, définie par l'expression suivante :

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2}.$$

Comme l'accès au génotype de milliers d'individus a été rendu possible par la spectaculaire diminution des coûts du séquençage de l'ADN, l'héritabilité de traits quantitatifs ainsi que de pathologies est devenu un sujet très étudié. Yang et al. (2010) a par exemple estimé que, chez les humains, environ 45 % des variations de la taille sont expliquées par les SNPs les plus fréquents.

6.1.2 L'héritabilité en génétique humaine

Nous présentons à présent l'intérêt d'estimer l'héritabilité de caractères humains. C'est en effet un premier pas vers la compréhension de maladies complexes, qui ont souvent des causes multiples. Nous faisons référence en particulier à des maladies qui ne sont pas causées par un seul gène affecté mais dont on soupçonne néanmoins qu'elles ont une composante génétique importante, probablement répartie sur plusieurs gènes. Par exemple, les causes de certaines maladies psychiatriques comme l'autisme et la schizophrénie sont encore vagues aujourd'hui. Une composante génétique est suggérée par les résultats des études de jumeaux monozygotes et dizygotes (les jumeaux monozygotes ont des génomes identiques alors que les jumeaux dizygotes ont environ 50% de leurs génomes en commun). Ces études montrent que si l'un des jumeaux souffre de troubles autistiques, l'autre également dans 82 à 92 % des cas pour les jumeaux monozygotes (Bailey et al. (1995)) contre 20 % des cas chez les jumeaux dizygotes (Hallmayer et al., 2011). Ces études décrivent l'autisme comme la maladie psychiatrique avec la plus importante composante génétique. Cependant, la sévérité des traits autistiques (troubles des interactions et du langage, handicap mental...) peuvent varier fortement chez deux patients qui présentent des causes similaires, par exemple la même mutation. Ces observations laissent penser, comme c'est le cas pour d'autres maladies génétiques, que le fond génétique module les effets d'une mutation et rend certains individus plus ou moins sensibles au développement de traits autistiques.

De plus, toutes ces études montrent que, malgré des patrimoines génétiques identiques, la concordance des symptômes chez les jumeaux monozygotes n'est jamais totale, ce qui confirme une composante épigénétique et/ou environnementale. Néanmoins, quantifier ces possibles causes et leurs interactions potentielles reste une question difficile.

La mise en évidence d'une composante génétique importante d'une maladie est également un argument puissant pour réfuter des croyances populaires quant aux causes de certaines maladies. Par exemple, une vague de mouvements anti-vaccins a été déclenchée par un lien présumé entre le vaccin contre l'hépatite B et la sclérose en plaques. De même, le vaccin contre la rougeole a été accusé de provoquer des cas d'autisme (Uno et al., 2012).

Bien qu'aucun lien n'ait jamais été démontré (Poland & Jacobson, 2001), les conséquences du refus de nombreux parents de vacciner leurs enfants reste un problème de santé publique majeur. En effet, une étude récente (Uno et al., 2012) a montré que plus de 25 % des parents américains avaient refusé de vacciner leurs enfants contre des maladies mortelles comme la rougeole. Quant à d'autres causes présumées de l'autisme, la théorie de la "mère réfrigérateur" a été formulée par le psychiatre Leo Kanner dans les années 1940 car il avait observé un manque d'affection maternelle chez les mères de ses patients. Bien que cette théorie ait largement été discréditée depuis, les mères de patients autistes ont souffert d'accusations sévères et injustifiées pendant plusieurs décennies.

6.1.3 L'héritabilité en génétique animale et végétale

En génétique animale ou végétale, l'estimation de l'héritabilité est souvent la première étape de la sélection de traits d'intérêt, généralement liés au rendement d'une ressource précieuse. Nous pouvons par exemple citer les exemples d'optimisation de production du lait (Visscher & Goddard, 1995) ou du blé (Eid, 2009). Le but de Eid (2009) est de déterminer des traits héritables liés au rendement du blé et ensuite d'obtenir un génotype optimal. Un tel génotype peut même être choisi pour être le plus résistant possible à des conditions environnementales extrêmes, comme l'absence d'eau, ce qui est actuellement un problème fondamental.

Si ces pratiques sont généralement bien acceptées dans le cadre de la génétique animale, elles ouvrent une controverse quant à de possibles conséquences de l'estimation de l'héritabilité de traits humains. En effet, plusieurs études ont estimé l'héritabilité du QI, par exemple (Toro et al., 2015), et Davies et al. (2011) a même annoncé que l'intelligence humaine était fortement héritable. La controverse relative à l'héritabilité du QI est discutée dans Visscher et al. (2008), qui énumère des raisons à cette polémique. Ces raisons vont de la très controversée définition du QI en tant que mesure de l'intelligence jusqu'à des abus historiques liés à l'eugénisme. Nous ne discuterons pas plus en détail cette polémique ici, mais nous la mentionnant pour illustrer des fréquentes réactions quand on travaille sur l'héritabilité de traits humains.

Après avoir brièvement argumenté en faveur de l'intérêt général d'estimer l'héritabilité, nous allons à présent introduire les modèles statistiques utilisés pour fournir ces estimations.

6.2 Estimation de l'héritabilité dans les modèles linéaires mixtes

6.2.1 Etat de l'art

Les modèles linéaires mixtes (LMMs) sont utilisés dans de nombreux domaines, en particulier en médecine et en génétique. Yang et al. (2010) ont notamment proposé d'estimer l'héritabilité de la taille humaine en utilisant un modèle linéaire mixte classique défini comme suit :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{6.1}$$

où $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ est le vecteur d'observations d'un phénotype d'intérêt, $\mathbf{X}$ est une matrice de taille $n \times p$ de prédicteurs, $\boldsymbol{\beta}$ est un vecteur de taille $p \times 1$ qui contient l'effet inconnu des prédicteurs (aussi appelés effets fixes), et $\mathbf{u}$ et $\mathbf{e}$ correspondent à des effets aléatoires gaussiens de variances respectives $\sigma_u^{\star 2}$ et $\sigma_e^{\star 2}$.

De plus, $\mathbf{Z}$ est une matrice de taille $n \times N$ qui contient l'information génétique. Plus précisément, les $Z_{i,j}$ sont variables aléatoires normalisées au sens où elles sont définies à partir d'une matrice $\mathbf{W} = (W_{i,j})_{1 \leq i \leq n,\, 1 \leq j \leq N}$ par la relation

$$Z_{i,j} = \frac{W_{i,j} - \overline{W}_j}{s_j},\ i = 1, \ldots, n,\ j = 1, \ldots, N\ , \tag{6.2}$$

où

$$\overline{W}_j = \frac{1}{n}\sum_{i=1}^{n} W_{i,j},\ s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(W_{i,j} - \overline{W}_j)^2,\ j = 1, \ldots, N\ . \tag{6.3}$$

Dans les équations (6.2) et (6.3), les $W_{i,j}$ sont de telle sorte que pour chaque j dans $\{1, \ldots, N\}$ les $(W_{i,j})_{1 \leq i \leq n}$ sont des variables aléatoires indépendantes et identiquement distribuées et les colonnes de $\mathbf{W}$ sont indépendantes.

Dans les applications en génétique, la matrice $\mathbf{W}$ contient l'information génétique des tous les individus étudiés. Plus précisément, pour chaque j, les $(W_{i,j})_{1 \leq i \leq n}$ sont des variables binomiales i.i.d de paramètres 2 et p_j : $W_{i,j} = 0$ (resp. 1, resp. 2) si le génotype de l'individu i au locus j est qq (resp. Qq, resp. QQ) où p_j est la fréquence de l'allèle Q au locus j.

Avec cette définition, les colonnes de $\mathbf{Z}$ sont empiriquement centrées et de variance empirique égale à 1.

Le modèle linéaire mixte apparaît comme un choix intuitif de modèle pour décrire le concept d'héritabilité en tant que ratio des variances génétique et phénotypique. Yang et al. (2010) et Pirinen et al. (2013) proposent d'estimer le paramètre

$$\eta^\star = \frac{N\sigma_u^{\star 2}}{N\sigma_u^{\star 2} + \sigma_e^{\star 2}}, \tag{6.4}$$

communément considéré comme la définition mathématique de l'héritabilité puisqu'il détermine comment les variances sont réparties entre $\mathbf{u}$ et $\mathbf{e}$.

Dans le modèle (6.1), la log-vraisemblance conditionnellement à $\mathbf{Z}$ est donnée par la formule :

$$L(\beta, \sigma_u^2, \sigma_e^2) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|\mathbf{Z}\mathbf{Z}'\sigma_u^2 + \sigma_e^2 \mathrm{Id}_{\mathbb{R}^n}| - \frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Z}\mathbf{Z}'\sigma_u^2 + \sigma_e^2 \mathrm{Id}_{\mathbb{R}^n})^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}). \tag{6.5}$$

Searle et al. (1992) ont regroupé de nombreuses techniques d'optimisation pour estimer les paramètres β, $\sigma_u^{\star 2}$ et $\sigma_e^{\star 2}$, parmi lesquelles nous pouvons citer les équations d'Henderson ou bien des méthodes itératives comme Fisher-Scoring et Newton-Raphson.

Une idée naturelle pour estimer l'héritabilité est d'estimer les paramètres de variances $\sigma_u^{\star 2}$ et $\sigma_e^{\star 2}$ afin d'obtenir un estimateur de $\eta^\star$ en tant que ratio :

$$N\hat{\sigma_u}^2/(N\hat{\sigma_u}^2 + \hat{\sigma_e}^2).$$

Pirinen et al. (2013) ont remarqué que le modèle défini par l'équation (6.1) peut être reparamétrisé en fonction de β, $\eta^\star$ et $\sigma^{\star 2} = N\sigma_u^{\star 2} + \sigma_e^{\star 2}$. Plus précisément,

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

où $\mathbf{R} = \mathbf{Z}\mathbf{Z}'/N$.

Si l'on note $\mathbf{U}$ la matrice orthogonale ($\mathbf{U}'\mathbf{U} = \mathbf{U}\mathbf{U}' = \mathrm{Id}_{\mathbb{R}^n}$) telle que $\mathbf{U}\mathbf{R}\mathbf{U}' = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ est une matrice diagonale dont les entrées diagonales sont égales à $\lambda_1, \ldots, \lambda_n$. Alors $\widetilde{\mathbf{Y}} = \mathbf{U}'\mathbf{Y}$ est un vecteur gaussien centré de matrice de variance covariance $\mathrm{diag}(\eta^\star\sigma^{\star 2}\lambda_1 + (1-\eta^\star)\sigma^{\star 2}, \ldots, \eta^\star\sigma^{\star 2}\lambda_n + (1-\eta^\star)\sigma^{\star 2})$, où les λ_i sont les valeurs propres de $\mathbf{R}$. Notons également $\widetilde{\mathbf{X}} = \mathbf{U}'\mathbf{X}$. Finalement, Pirinen et al. (2013) ont maximisé la log-vraisemblance suivante :

$$L_n(\beta, \sigma^2, \eta) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\log(\eta(\lambda_i - 1) + 1) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{(\widetilde{\mathbf{Y}}_i - \widetilde{\mathbf{X}}\beta)^2}{\eta(\lambda_i - 1) + 1} - \frac{n}{2}\log(2\pi), \quad (6.6)$$

où $\widetilde{\mathbf{Y}} = (\widetilde{\mathbf{Y}}_1, ..., \widetilde{\mathbf{Y}}_n)$.

Les méthodes présentées précédemment ont deux possibles faiblesses dans les applications qui nous intéressent. Tout d'abord, elles ont été validées théoriquement uniquement dans le cadre où N est fixé et n tend vers l'infini. En effet, des résultats classiques du modèle linéaire mixte permettent d'obtenir des résultats de consistance et de normalité asymptotiques pour l'estimateur du maximum de vraisemblance. Cependant, puisqu'en pratique le nombre N de SNPs est largement supérieur au nombre n d'individus, il serait plus adapté de valider ces méthodes dans le cadre où n et N tendent vers l'infini, avec n/N qui tend vers une constante a dans $(0, +\infty)$.

De plus, toutes ces méthodes ont été développées dans le cas d'effets aléatoires gaussiens, ce qui impliquerait que toute l'information génétique disponible ait un impact sur le phénotype observé. Cette hypothèse, qui semble assez improbable, a été discutée en particulier par Jiang et al. (2014), qui ont étudié les conséquences d'une modélisation qui ne prend pas en compte le fait que les effets de certains SNPs (possiblement nombreux) soient nuls.

Nous avons pour l'instant mentionné uniquement les méthodes d'estimation de l'héritabilité dans les modèles linéaires mixtes, mais il existe d'autres façons de définir et d'estimer l'héritabilité. En effet, d'importants résultats théoriques ont été prouvés dans le cas du modèle linéaire

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (6.7)$$

où la composante aléatoire vient d'une part de l'erreur résiduelle $\boldsymbol{\varepsilon}$, supposée gaussienne centrée et de variance σ_ϵ^2 mais aussi de la matrice de SNPs $\mathbf{X}$ dont les colonnes sont supposées gaussiennes centrées réduites indépendantes. Dans ce modèle, l'héritabilité est définie comme le ratio

$$\eta^\star = \frac{||\boldsymbol{\beta}||_2^2}{\sigma_\epsilon^2 + ||\boldsymbol{\beta}||_2^2}. \quad (6.8)$$

Un avantage de ce modèle est qu'il n'y a pas d'hypothèse sur la distribution de $\boldsymbol{\beta}$, en particulier sur sa parcimonie. Néanmoins, des hypothèses fortes sont requises sur la structure de la matrice $\mathbf{X}$. Plusieurs méthodes ont été proposées pour estimer l'héritabilité dans le modèle (6.7). Dicker (2014) a proposé un estimateur de moments qui est asymptotiquement normal quand $n, N \to a \in (0, +\infty)$. Janson et al. (2015) ont développé la procédure Eigenprism pour

construire des intervalles de confiance pour $\eta^\star$ pour des échantillons finis et dont ils ont également étudié les propriétés asymptotiques quand $n, N \to a \in (0, +\infty)$. Dicker & Erdogdu (2016) ont étudié les propriétés de l'estimateur du maximum de vraisemblance et ont mené une étude de simulation qui compare les méthodes précédemment citées et qui a montré que l'estimateur du maximum de vraisemblance a une variance empirique plus petite que les deux autres approches. Dicker & Erdogdu (2016) ont également montré la consistance et la normalité asymptotique de l'estimateur du maximum de vraisemblance et ont calculé une formule explicite de la variance asymptotique.

Toujours dans le modèle linéaire, Verzelen & Gassiat (2016) ont étudié l'optimalité de différentes procédures selon la parcimonie des effets aléatoires. En effet, Verzelen & Gassiat (2016) ont comparé les performances d'une approche avec sélection (estimateur Gauss-LASSO) et sans sélection (estimateur dense) dans différents régimes de parcimonie. Ils ont distingué des gammes de valeurs pour la parcimonie et pour chacune, ont évalué le risque minimax : ils ont montré que la meilleure procédure était un estimateur adaptatif qui se comporte comme l'estimateur Gauss-LASSO dans les cas très parcimonieux et comme l'estimateur dense dans les cas peu parcimonieux

6.2.2 Contribution

Notre première contribution a été de proposer un estimateur de l'héritabilité dans le contexte où n et N tendent vers l'infini, avec n/N qui tend vers a dans $(0, +\infty)$, et d'en établir les propriétés théoriques. Ce travail est développé dans le chapitre 2 de ce manuscrit et a fait l'objet d'un article publié dans Electronic Journal of Statistics. Nous avons étudié le modèle défini par (6.1) sauf que nous avons supposé que les effets aléatoires pouvaient être parcimonieux, c'est-à-dire que seulement une proportion q des composantes de $\mathbf{u}$ étaient non nulles :

$$u_i \overset{i.i.d.}{\sim} (1-q)\delta_0 + q\mathcal{N}(0, \sigma_u^{\star 2}) \text{ , pour tous } 1 \leq i \leq N \text{ et } \mathbf{e} \sim \mathcal{N}\left(0, \sigma_e^{\star 2}\mathrm{Id}_{\mathbb{R}^n}\right), \tag{6.9}$$

où $\mathrm{Id}_{\mathbb{R}^n}$ est la matrice identité de taille $n \times n$, q est dans $(0, 1]$, et δ_0 la masse de Dirac en 0.

Quitte à considérer la projection de $\mathbf{Y}$ sur l'orthogonal de l'image de $\mathbf{X}$ et pour simplifier les calculs, nous avons étudié le modèle suivant

$$\mathbf{Y} = \mathbf{Z}\mathbf{u} + \mathbf{e} \, . \tag{6.10}$$

De plus, comme dans notre cas nous nous intéressons uniquement à l'estimation de $\eta^\star$, nous avons remplacé $\sigma^{\star 2}$ par son estimateur

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{\widetilde{Y}_i^{\,2}}{\eta(\lambda_i - 1) + 1}.$$

Nous avons implémenté un estimateur de $\eta^\star$ comme le maximiseur de la fonction de vraisemblance dépendant uniquement du paramètre η :

$$L_n(\eta) = -\log\left(\frac{1}{n}\sum_{i=1}^{n} \frac{\widetilde{Y}_i^2}{\eta(\lambda_i - 1) + 1}\right) - \frac{1}{n}\sum_{i=1}^{n} \log\left(\eta(\lambda_i - 1) + 1\right) \, , \tag{6.11}$$

Nous avons obtenu deux résultats principaux dans le cadre où n et N tendent vers l'infini, avec n/N qui tend vers a dans $(0, 1)$: premièrement, nous avons prouvé que notre estimateur était $\sqrt{n}$-consistant malgré la présence de composantes nulles dans les effets aléatoires. Ce résultat a été obtenu sous de faibles hypothèses sur la matrice $\mathbf{W}$ et pour toute valeur de parcimonie q.

Nous avons ensuite établi un théorème de la limite centrale sous l'hypothèse supplémentaire que pour tous i et j, les $Z_{i,j}$ étaient des variables gaussiennes centrées réduites et indépendantes. Nous avons calculé une formule explicite de la variance asymptotique, donnée par

$$\tau^2(a, \eta^\star, q) = \frac{2}{\gamma^2(a, \eta^\star)} + 3\frac{a^2\eta^{\star 2}}{\gamma^4(a, \eta^\star)}\left(\frac{1}{q} - 1\right) S(a, \eta^\star) \tag{6.12}$$

où

$$\gamma^2(a, \eta^\star) = \left\{ \int g(\eta, \lambda)^2 \mathrm{d}\mu_a(\lambda) - \left(\int g(\eta, \lambda) \mathrm{d}\mu_a(\lambda) \right)^2 \right\}$$

et

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

Dans l'expression précédente, $\mathrm{d}\mu_a(\lambda)$ est la densité de Marchenko-Pastur, qui est la distribution des valeurs propres de $\mathbf{ZZ}'/N$. Cette distribution, obtenue par Marchenko & Pastur (1968) a été un élément clef pour établir les preuves de nos résultats. Nous avons implémenté notre approche dans le package R HiLMM, disponible sur le CRAN.

Nous avons également mené une étude de simulation sur des échantillons finis dont les tailles correspondent à des cas réalistes en pratiques. Nous avons montré que, bien que la variance asymptotique définie par la formule (6.12) dépende en théorie de la parcimonie q, son influence était à peine remarquable en pratique. Par contre, nous avons observé que la variance de notre estimateur était influencée par la paramètre $a = n/N$. Plus précisément, quand le nombre d'observations est très petit comparé à la taille des effets aléatoires (ce qui est très souvent le cas dans les études en génétique), la variance de l'estimateur était très grande. Ce constat numérique a motivé l'idée de développer une méthode de sélection de variables afin de réduire la taille des effets aléatoires et d'améliorer la précision des estimations de l'héritabilité.

6.3 Sélection de variables dans les effets aléatoires des modèles linéaires mixtes

Suite aux performances numériques de notre estimateur de l'héritabilité présentées dans la section précédente, nous avons décidé d'inclure une étape de sélection de variables à notre méthode. Le but de cette étape serait de retrouver le support des effets aléatoires, qui correspond en pratique aux SNPs impliqués dans les variations du phénotype observé. Nous considérerions ensuite la matrice de SNPs réduite à ces SNPs sélectionnés et nous estimerions l'héritabilité avec un erreur standard plus petite que celle nous aurions obtenue en utilisant la matrice de SNPs entière. Nous allons dans un premier temps présenter les méthodes et les résultats qui existent sur la sélection de variables dans les effets fixes des modèles linéaires mixtes parcimonieux.

6.3.1 Etat de l'art

Bien que la littérature concernant la sélection de variable dans les modèles linéaires mixtes soit moins riche que celle sur les modèles linéaires, plusieurs méthodes ont été proposées pour sélectionner des variables dans la partie des effets fixes et des effets aléatoires.

Pour ce qui est de la sélection dans les effets fixes, nous renvoyons le lecteur à la revue proposée par Müller et al. (2013). Concernant la sélection dans les effets aléatoires, nous avons connaissance du travail de Fan & Li (2012) and Bondell et al. (2010). Bondell et al. (2010) ont proposé une méthode basée sur un algorithme EM pour sélectionner conjointement des effets fixes et aléatoires. Fan & Li (2012) ont utilisé un critère pénalisé avec une pénalité particulière, appelée SCAD (Smoothly Clipped Absolute Deviation), qui combine des pénalités ℓ_1 et ℓ_2. Ces deux méthodes peuvent demander des temps de calcul important en grande dimension, pour l'une à cause de l'algorithme EM et pour l'autre à cause de la validation croisée nécessaire au choix des deux paramètre de régularisation.

La sélection de variable dans des cadres de très grande dimension comme ceux auxquels nous nous intéressons est un problème difficile, comme montré par Verzelen (2012) qui a étudié le cas du modèle linéaire défini par l'équation (6.7). Verzelen (2012) a en effet établi que si la condition $Nq\log(1/q) >> n$ était vérifiée, ce qui est le cas notamment si le nombre de SNPs causaux (nombre de composantes non nulles dans les effets aléatoires) est grand par rapport au nombre d'individus, le support ne peut pas être entièrement retrouvé.

Concernant l'estimation de l'héritabilité, l'idée d'introduire une étape de sélection de variable au préalable a été proposée par Guan & Stephens (2011) dans le cadre Bayésien. Guan & Stephens (2011) ont en effet proposé une approche, appelée BVSR (Bayesian Variable Selection Regression), qui est très efficace dans des cas où les effets aléatoires sont très parcimonieux mais qui est biaisé quand le nombre de SNPs causaux est élevé. Zhou et al. (2013) ont ensuite proposé une approche, appelée BSLMM (Bayesian Sparse Linear Mixed Model), définie comme un estimateur hybride qui a un comportement proche de BVSR dans les cas très parcimonieux et comme l'estimateur du maximum de vraisemblance (sans sélection) sinon.

Ces résultats numériques obtenus par Zhou et al. (2013) rejoignent les résultats théoriques obtenus par Verzelen & Gassiat (2016) avec leur procédure adaptative optimale dans le cas du modèle linéaire décrit dans la section 6.2.1.

6.3.2 Contribution

Méthodologie

Nous avons proposé une méthode de sélection de variables dans le but d'améliorer la précision de l'estimation de l'héritabilité. Ce travail a fait l'objet d'un article soumis à la publication et qui est décrit dans le chapitre 3 de ce manuscrit. Notre méthode est implémentée dans le package R EstHer, disponible sur le CRAN.

Notre approche présente deux caractéristiques principales : premièrement, elle est très efficace d'un point de vue statistique car elle permet de réduire considérablement la taille des intervalles de confiance par rapport à des méthodes sans sélection. Deuxièmement, elle est très rapide, ce qui la rend facilement utilisable sur des jeux de données très grands qui sont fréquents dans les études génétiques. Notre méthode peut gérer des cas de très grande dimension grâce à une étape

de Sure Independence Screening développée par (Ji & Jin, 2012). Nous appliquons ensuite un critère LASSO (Tibshirani, 1996) combiné avec la méthode de "stability selection" (Meinshausen & Bühlmann, 2010). Nous proposons également une méthode de bootstrap non paramétrique pour calculer des intervalles de confiance, qui ont ensuite été validés sur des données simulées.

Au cours de notre étude numérique, nous avons observé des conclusions similaires à celles de Zhou et al. (2013) dans le cadre Bayésien : dans des cas très parcimonieux (par exemple, 200 SNPs causaux sur 100000), l'estimateur qui inclut une étape de sélection de variable est non biaisé et sa variance empiriquement est considérablement plus faible que celle de l'estimateur du maximum de vraisemblance. Cependant, quand le nombre de SNPs causaux est élevé, inclure une étape de sélection de variables conduit à sous-estimer fortement l'héritabilité. Nous avons développé un critère à appliquer sur les données pour avoir une idée de la parcimonie et pour décider s'il est judicieux ou non d'appliquer une technique de sélection de variables avant d'estimer l'héritabilité.

Nous avons donc construit un estimateur hybride capable de s'adapter au régime de parcimonie et nous avons montré sur des données simulées que cette procédure permet de réduire considérablement les intervalles de confiance dans les cas très parcimonieux par rapport à l'estimateur du maximum de vraisemblance.

L'avantage de notre méthode par rapport à celle de Zhou et al. (2013) est surtout d'avoir un temps de calcul beaucoup plus faible et aussi de ne pas avoir à régler les différents paramètres nécessaires à la convergence des algorithmes MCMC.

Applications en neuro-anatomie et en génétique animale

Nous avons appliqué notre méthode à deux jeux de données.

Le premier provient du projet européen IMAGEN, qui est une étude sur la santé mentale des adolescents. Nous avons estimé l'héritabilité du volume du cerveau et des volumes des différentes régions sous-corticales. Six des neuf phénotypes n'ont pas passé le critère de sélection que nous avons proposé, nous pouvons donc penser qu'un grand nombre de SNPs sont impliqués dans leurs variations. Nous avons obtenus avec notre méthode des résultats similaires à ceux obtenus avec des résultats obtenus avec des approches classiques de maximum de vraisemblance, comme ceux obtenus par Toro et al. (2015) qui ont étudié les mêmes données à l'aide du logiciel GCTA développé par Yang et al. (2011). Cependant, pour les trois autres phénotypes, nous avons obtenu des estimations de l'héritabilité avec de très petites erreurs standard mais également une liste de SNPs potentiellement impliqués dans les variations et dont la pertinence pourrait être étudiée d'un point de vue biologique. Cette application est décrite dans le Chapitre 3 de ce manuscrit, après la présentation de notre méthode et la validation de celle-ci sur des données simulées.

La deuxième application est l'étude d'une espèce de truites appelée *Salmo trutta*. Cette truite d'eau douce peut ou non, durant sa vie, décider de quitter l'eau douce pour migrer vers la mer. Cette migration a un impact majeur sur la conservation de la truite, et nous voudrions comprendre les raisons de cette décision. Il apparaît que la croissance durant la phase en eau douce pourrait être un facteur déterminant sur la décision des truites : en effet, si un poisson a une bonne croissance en eau douce, il n'a pas d'intérêt à aller en mer où les chances de survie sont plus faibles. Néanmoins, si la truite n'arrive pas à grandir en eau douce, le bénéfice de rejoindre la mer où les chances de grandir sont plus importantes peut compenser les risques associés à ce changement. Nous souhaiterions alors rechercher la proportion de facteurs génétiques et

environnementaux impliqués dans les variations de la taille des truites, et si possible déterminer quels SNPs et quelles variables environnementales sont associées à la croissance des truites. Le jeu de données contient la taille de 192 truites dont le génotype est décrit par 4069 SNPs. Cette application est décrite dans le chapitre 4.

6.4 Estimation de l'héritabilité pour des traits binaires

Nous nous intéressons à l'extension des méthodes précédentes à l'estimation de l'héritabilité d'une maladie, auquel cas les observations sont catégorielles (patient ou contrôle). Nous avons trouvé dans la littérature différents modèles utilisés pour estimer l'héritabilité de traits binaires.

6.4.1 Modèle linéaire mixte généralisé et "liability model".

Une généralisation intuitive des travaux précédents à l'estimation de l'héritabilité de traits binaires serait de considérer le modèle linéaire généralisé suivant :

$$\mathbf{Y}_i \sim \mathcal{B}(q_i), \tag{6.13}$$

avec $q_i = g(\mathbf{l}_i)$ où g est une fonction de lien et $\mathbf{l}_i$ est défini par

$$\mathbf{l} = \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{6.14}$$

avec $\mathbf{u} \sim \mathcal{N}(0, \sigma_u^{\star 2})$ et $\mathbf{e} \sim \mathcal{N}(0, \sigma_e^{\star 2})$, comme dans le LMM classique défini à la Section 6.2.
Un choix classique de fonction de lien dans le cas de données binaires est par exemple

$$g(x) = \frac{\exp(x)}{1 + \exp(x)},$$

ce qui garantit que $q_i \in (0, 1)$.
L'héritabilité peut alors être définie pour la variable continue $\mathbf{l}$, et donc la définition est identique à celle que l'on a considérée précédemment pour des phénotypes gaussiens :

$$\eta^\star = \frac{N\sigma_u^{\star 2}}{N\sigma_u^{\star 2} + \sigma_e^{\star 2}}. \tag{6.15}$$

Un autre modèle, plus ancien et plus fréquemment utilisé, a été proposé par Falconer (1965), qui a fait l'hypothèse que les observations binaires pouvaient être vues comme des indicatrices qu'une certaine variable continue et non observée dépasse un seuil t :

$$\mathbf{Y}_i = \mathbb{1}_{\{\mathbf{l}_i > t\}} \tag{6.16}$$

avec $\mathbf{l}_i$ définie par la même expression (6.14) que dans le modèle précédent.

Ce modèle est usuellement appelé le "liability model" (Falconer (1965), Lee et al. (2011), Tenesa & Haley (2013)). L'héritabilité est alors également définie à l'échelle continue, avec la même définition que celle donnée dans l'équation (6.15).

6.4.2 Méthodes existantes

Dans la littérature spécifique à l'estimation de l'héritabilité de phénotypes binaires, nous avons trouvé des méthodes associées à chacun des modèles décrits précédemment. Pour le premier modèle (6.13), de Villemereuil et al. (2013) ont proposé d'estimer la variance $\sigma_u^{\star 2}$ des effets aléatoires en utilisant des méthodes MCMC développées par Hadfield (2010), puis d'estimer l'héritabilité par le ratio

$$\hat{\eta} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + 1 + 1},$$

où le premier 1 du dénominateur représente la variance résiduelle et le deuxième 1 la distribution spécifique à la fonction de lien probit (Nakagawa & Schielzeth, 2010). La variance résiduelle est en effet fixée à 1 car les données binaires ne permettent pas de fournir suffisamment d'information pour inférer les deux variances $\sigma_u^{\star 2}$ et $\sigma_e^{\star 2}$.
Puisque l'expression de la vraisemblance n'est pas possible à optimiser directement, Breslow & Clayton (1993) ont proposé de maximiser une quasi-vraisemblance pénalisée, en utilisant une approximation de Laplace de la vraisemblance. Il n'existe pas de résultats théoriques concernant ces méthodes mais leurs performances numériques ont été étudiées et comparées par de Villemereuil et al. (2013).

Concernant les procédures qui s'appuient sur le second modèle défini par les équations (6.14) et (6.16), Lee et al. (2011) ont proposé d'utiliser une approche de maximum de vraisemblance comme si les traits binaires étaient gaussiens, puis d'appliquer un facteur multiplicatif pour corriger cette approximation. Golan et al. (2014) ont montré que cet estimateur était biaisé dans plusieurs scenarios réalistes, en particulier qu'il était très sensible à la prévalence de la maladie (quand la maladie est rare, le biais croît). Cet estimateur sous-estime également l'héritabilité quand celle-ci est élevée.
Weissbrod et al. (2015) ont proposé une approche du maximum de vraisemblance pour reconstruire la variable gaussienne non observée $\mathbf{l}$ avant d'estimer l'héritabilité.
A notre connaissance, aucune des méthodes précédemment présentées ne possède de validation théorique. De plus, elles ne prennent pas en compte un élément essentiel des données de cas-contrôles qui proviennent des études médicales. En effet, bien que la maladie étudiée puisse être rare, dans une étude médicale, le nombre de patients est à peu près égal au nombre de contrôles, ce qui veut dire que la proportion de patients dans l'étude peut être très différente de la proportion de patients dans la population générale. La méthode proposée par Golan et al. (2014) est une méthode de moments qui prend en compte ce sur-échantillonnage des patients et c'est, à notre connaissance, la seule méthode dans ce cas.
Plus précisément, Golan et al. (2014) ont considéré une version simplifiée du modèle (6.14), où la variable continue $\mathbf{l}$ est définie par

$$\mathbf{l} = \mathbf{g} + \mathbf{e},$$

où $\mathbf{g}$ est un effet génétique aléatoire, dont les composantes sont corrélées, et $\mathbf{e}$ est un effet aléatoire environnemental, que l'on suppose indépendant des effets génétiques. Les deux effets sont supposés gaussiens : $\mathbf{e}$ a une variance égale à $(1 - \eta^\star)\mathrm{Id}_{\mathbb{R}^n}$ et $\mathbf{g}$ a une matrice de variance covariance dont les entrées diagonales sont égales à $\eta^\star$ et le terme non diagonal (i, j) est égal à

$\eta^\star \mathbf{G}_{i,j}$. Pour $1 \leq i \neq j \leq n$, la matrice de variance covariance de $(\mathbf{l}_i, \mathbf{l}_j)$ est donnée par

$$\begin{pmatrix} 1 & \mathbf{G}_{i,j}\eta^\star \\ \mathbf{G}_{i,j}\eta^\star & 1 \end{pmatrix}.$$

Ils ont ensuite défini la variable

$$p_i = \frac{\mathbf{Y}_i - P}{\sqrt{P(1-P)}},$$

où P est la proportion de cas dans l'étude et l'événement $\{S = 1\}$ est réalisé si les individus i et j sont sélectionnés dans l'étude.

L'estimateur de l'héritabilité proposé par Golan et al. (2014) est un estimateur des moindres carrés obtenu en minimisant le critère

$$\sum_{i \neq j} \left(p_i p_j - \mathbb{E}(p_i p_j | S = 1)\right)^2.$$

Puisque $\mathbb{E}(p_i p_j | S = 1)$ n'a pas d'expression que l'on peut calculer explicitement, Golan et al. (2014) ont eu l'idée d'utiliser le fait que les corrélations $\mathbf{G}_{i,j}$ soient petites et ont proposé une approximation grâce à des développements de Taylor autour de la quantité $\mathbf{G}_{i,j}$.

6.4.3 Contribution

Comme la méthode de Golan et al. (2014) semblait très efficace en pratique et puisque c'est la seule méthode que nous ayons vue qui prenne en compte la spécificité des données de cas-contrôles, nous avons cherché à établir les propriétés théoriques de leur estimateur dans le cas où n et N tendent vers l'infini, et n/N tend vers a $\in (0, +\infty)$.
Nous avons considéré le modèle défini par les équations (6.16) et (6.14), et nous avons supposé que $\mathbf{Z}$ était une matrice aléatoire dont les colonnes sont centrées et réduites comme nous l'avions défini à la Section 6.2.
Dans ce modèle, la matrice de variance covariance de $(\mathbf{l}_i, \mathbf{l}_j)$ s'écrit :

$$\Sigma^{(N)} = \begin{pmatrix} 1 + \eta^\star(\mathbf{G}_N(i,i) - 1) & \eta^\star \mathbf{G}_N(i,j) \\ \eta^\star \mathbf{G}_N(i,j) & 1 + \eta^\star(\mathbf{G}_N(i,i) - 1), \end{pmatrix}$$

où, pour tous $1 \leq i, j \leq n$,

$$\mathbf{G}_N(i,j) = \frac{1}{N}\sum_{k=1}^{N} \mathbf{Z}_{i,k}\mathbf{Z}_{j,k}. \tag{6.17}$$

L'idée principale est de remarquer que les quantités $\mathbf{G}_N(i,j)$, $\mathbf{G}_N(i,i) - 1$ et $\mathbf{G}_N(j,j) - 1$ sont petites, ce qui implique que la matrice $\Sigma^{(N)}$ est proche de l'identité.
Inspirés par la méthode de Golan et al. (2014), nous avons proposé des approximations au premier et au second ordre de $\mathbb{E}[p_i p_j | \mathbf{Z}, S = 1]$ grâce à des développement de Taylor autour de $\mathbf{G}_N(i,j)$, $\mathbf{G}_N(i,i) - 1$ et $\mathbf{G}_N(j,j) - 1$, qui sont ici des variables aléatoires.

Malgré quelques différences entre les modèles considérés, nous trouvons la même approximation au premier ordre que Golan et al. (2014) mais nous avons des différences entre les deux approximations du deuxième ordre.
Nous avons, dans un premier temps, recherché les propriétés théoriques de l'estimateur obtenu avec l'approximation du premier ordre : nous avons montré que c'était un estimateur consistant de l'héritabilité sous des hypothèses faibles sur la matrice de SNPs.
Ensuite, nous avons comparé les performances numériques des estimateurs obtenus avec les deux approximations, à la fois d'un point de vue statistique et computationnel.
Ces résultats sont décrits dans le chapitre 5 de ce manuscrit.

Bibliography

M. Abney. Permutation testing in the presence of polygenic variation. *Genetic Epidemiology*, 39(4):249–258, 2015. ISSN 1098-2272. doi: 10.1002/gepi.21893.

D. G. Amaral, C. M. Schumann, and C. W. Nordahl. Neuroanatomy of autism. *Trends in Neurosciences*, 31(3):137 – 145, 2008. ISSN 0166-2236. doi: http://dx.doi.org/10.1016/j.tins.2007.12.005.

E. Arias-Castro, E. J. Candès, and Y. Plan. Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *The Annals of Statistics*, 39(5):2533–2556, 2011.

Z. Bai and J. W. Silverstein. *Spectral analysis of large dimensional random matrices*. Springer Series in Statistics. Springer, New York, second edition, 2010. ISBN 978-1-4419-0660-1. doi: 10.1007/978-1-4419-0661-8. URL `http://dx.doi.org/10.1007/978-1-4419-0661-8`.

Z. Bai and W. Zhou. Large sample covariance matrices without independence structures in columns. *Statistica Sinica*, 18:425–442, 2008.

A. Bailey, A. Le Couteur, I. Gottesman, P. Bolton, and E. Simonoff. Autism as a strongly genetic disorder: evidence from a british twin study. *Psychological Medicine*, 1995.

A. Beinrucker, U. Dogan, and G. Blanchard. Extensions of stability selection using subsamples of observations and covariates, 2014. arXiv:1407.4916v1.

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

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

A. Bonnet, C. Lévy-Leduc, E. Gassiat, R. Toro, and T. Bourgeron. Improving heritability by a variable selection approach in sparse high dimensional linear mixed models, 2016. Submitted.

N. Breslow and D. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25, 1993. ISSN 01621459. URL `http://www.jstor.org/stable/2290687`.

G. Davies, A. Tenesa, A. Payton, J. Yang, S. E. Harris, D. Liewald, X. Ke, S. Le Hellard, A. Christoforou, M. Luciano, et al. Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Molecular psychiatry*, 16(10):996–1005, 2011.

P. de Villemereuil, O. Gimenez, and B. Doligez. Comparing parent–offspring regression with frequentist and Bayesian animal models to estimate heritability in wild populations: a simulation study for Gaussian and binary traits. *Methods in Ecology and Evolution*, 4(3):260–275, 2013. ISSN 2041-210X. doi: 10.1111/2041-210X.12011. URL `http://devillemereuil.legtux.org/publis/deVillemereuiletal.-2013-Comparingparent\OT1\textendashoffspringregressionwithfrequen.pdf`.

L. H. Dicker. Variance estimation in high-dimensional linear models. *Biometrika*, 101(2):269–284, 2014.

L. H. Dicker and M. A. Erdogdu. Maximum likelihood for variance estimation in high-dimensional linear models. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 159–167, 2016.

M. Eid. Estimation of heritability and genetic advance of yield traits in wheat (Triticum aestivum L.) under drought condition. *International Journal of Genetics and Molecular Biology*, 1(7):115–120, 2009. URL `http://www.academicjournals.org/article/article1379515024{_}Eid..pdf`.

D. S. Falconer. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics*, 29(1):51–76, 1965. ISSN 1469-1809. doi: 10.1111/j.1469-1809.1965.tb00500.x. URL `http://dx.doi.org/10.1111/j.1469-1809.1965.tb00500.x`.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2008.00674.x. URL `http://dx.doi.org/10.1111/j.1467-9868.2008.00674.x`.

Y. Fan and R. Li. Variable selection in mixed effects models. *Annals of Statistics*, 2012.

A. Gilmour, R. Thompson, and B. Cullis. Average information reml: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, 51(4):1440–1450, 1995.

D. Golan and S. Rosset. Accurate estimation of heritability in genome wide studies using random effects models. *Bioinformatics [ISMB/ECCB]*, 27(13):317–323, 2011.

D. Golan, E. S. Lander, and S. Rosset. Measuring missing heritability: Inferring the contribution of common variants. *Proceedings of the National Academy of Sciences*, 111(49):E5272–E5281, 2014.

Y. Guan and M. Stephens. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*, 5(3):1780–1815, 09 2011. doi: 10.1214/11-AOAS455. URL `http://dx.doi.org/10.1214/11-AOAS455`.

J. D. Hadfield. MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, 33(2):1–22, 2010. URL `http://www.jstatsoft.org/v33/i02/`.

J. Hallmayer, S. Cleveland, A. Torres, J. Phillips, and B. Cohen. Genetic heritability and shared environmental factors among twin pairs with autism. *Arch Gen Psychiatry 68: 1095-102*, 2011.

Y. I. Ingster, A. B. Tsybakov, and N. Verzelen. Detection boundary in sparse regression. *Electronic Journal of Statistics*, 4:1476–1526, 2010.

L. Janson, R. F. Barber, and E. Candes. Eigenprism: Inference for high-dimensional signal-to-noise ratios. *arXiv preprint arXiv:1505.02097*, 2015.

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

J. Jiang, C. Li, D. Paul, C. Yang, and H. Zhao. High-dimensional genome-wide association study and misspecified mixed model analysis. *arXiv preprint arXiv:1404.2355*, 2014.

H. M. Kang, N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178:1709–1723, 2008.

S. H. Lee, N. R. Wray, M. E. Goddard, and P. M. Visscher. Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics*, 88 (3):294–305, 2011.

M. Lynch and B. Walsh. *Genetics and Analysis of Quantitative Traits*. Sunderland, M, 1998.

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

T. A. Manolio, 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. Finding the missing heritability of complex diseases. *Nature*, 461(7265): 747–753, 2009.

V. Marchenko and L. Pastur. Distribution of eigenvalues for some sets of random matrices. *Math. USSR, Sb.*, 1:457–483, 1968. ISSN 0025-5734. doi: 10.1070/SM1967v001n04ABEH001994.

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

S. Müller, J. L. Scealy, and A. H. Welsh. Model selection in linear mixed models. *Statist. Sci.*, 28(2):135–167, 05 2013. doi: 10.1214/12-STS410. URL `http://dx.doi.org/10.1214/12-STS410`.

S. Nakagawa and H. Schielzeth. Repeatability for gaussian and non-gaussian data: a practical guide for biologists. *Biological Reviews*, 85(4):935–956, 2010.

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

M. Pirinen, P. Donnelly, and C. C. A. Spencer. 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, 2013. doi: 10.1214/12-AOAS586. URL `http://dx.doi.org/10.1214/12-AOAS586`.

G. A. Poland and R. M. Jacobson. Understanding those who do not understand: a brief review of the anti-vaccine movement. *Vaccine*, 19(17):2440–2445, 2001.

S. Purcell, N. Wray, J. Stone, P. Visscher, M. O'Donovan, P. Sullivan, P. Sklar, International Schizophrenia Consortium, and B. Pickard. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–752, 8 2009. ISSN 0028-0836. doi: 10.1038/nature08185.

J. Schelldorfer, P. Bühlmann, G. DE, and S. VAN. Estimation for high-dimensional linear mixed-effects models using ℓ_1-penalization. *Scandinavian Journal of Statistics*, 38(2):197–214, 2011.

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

R. Serfling. *Approximation theorems of mathematical statistics*. Wiley series in probability and mathematical statistics. Wiley, New York, NY, 1980.

R. G. Steen, C. Mull, R. McClure, R. M. Hamer, and J. A. Lieberman. Brain volume in first-episode schizophrenia. *The British Journal of Psychiatry*, 188(6):510–518, 2006. ISSN 0007-1250. doi: 10.1192/bjp.188.6.510.

J. L. Stein, 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. Identification of common variants associated with human hippocampal and intracranial volumes. *Nat Genet*, 44(5):552–561, 2012. ISSN 1061-4036. doi: 10.1038/ng.2250.

A. Tenesa and C. Haley. The heritability of human disease: estimation, uses and abuses. *Nature Reviews Genetics*, 14(2):139–49, 2013. ISSN 1471-0056. doi: 10.1038/nrg3377.

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

R. Toro, 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. Genomic architecture of human neuroanatomical diversity. *Molecular Psychiatry*, 20(8):1011–1016, 2015. doi: 10.1038//mp.2014.99.

Y. Uno, T. Uchiyama, M. Kurosawa, B. Aleksic, and N. Ozaki. The combined measles, mumps, and rubella vaccines and the total number of vaccines are not associated with development of autism spectrum disorder: the first case–control study in asia. *Vaccine*, 30(28):4292–4298, 2012.

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

N. Verzelen and E. Gassiat. Adaptive estimation of high-dimensional signal-to-noise ratios. 2016. http://arxiv.org/abs/1602.08006.

P. Visscher and M. Goddard. Genetic parameters for milk yield, survival, workability, and type traits for australian dairy cattle. *Journal of Dairy Science*, 78(1):205–220, 1995.

P. Visscher, W. Hill, and N. Wray. Heritability in the genomics era - concepts and misconceptions. *Nature Reviews Genetics*, 9(4):255–266, 4 2008. ISSN 1471-0056. doi: 10.1038/nrg2322.

G. C. G. Wei and M. A. Tanner. A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704, 1990.

O. Weissbrod, C. Lippert, D. Geiger, and D. Heckerman. Accurate liability estimation improves power in ascertained case-control studies. *Nature methods*, 12(4):332–334, 2015.

J. Yang, 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. Common snps explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569, 2010. doi: 10.1038/ng.608.

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

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

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

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Estimation de l'héritabilité dans les modèles mixtes en grande dimension : théorie et applications

Mots clés : Grande dimension, Héritabilité, Modèles mixtes, Sélection de variables.

Résumé : Nous nous intéressons à des méthodes statistiques pour estimer l'héritabilité d'un caractère biologique, qui correspond à la part des variations de ce caractère qui peut être attribuée à des facteurs génétiques. Nous proposons dans un premier temps d'étudier l'héritabilité de traits biologiques continus à l'aide de modèles linéaires mixtes parcimonieux en grande dimension. Nous avons recherché les propriétés théoriques de l'estimateur du maximum de vraisemblance de l'héritabilité : nous avons montré que cet estimateur était consistant et vérifiait un théorème central limite avec une variance asymptotique que nous avons calculée explicitement. Ce résultat, appuyé par des simulations numériques sur des échantillons finis, nous a permis de constater que la variance de notre estimateur était très fortement influencée par le ratio entre le nombre d'observations et la taille des effets génétiques. Plus précisément, quand le nombre d'observations est faible comparé à la taille des effets génétiques (ce qui est très souvent le cas dans les études génétiques), la variance de l'estimateur était très grande. Ce constat a motivé le développement d'une méthode de sélection de variables afin de ne garder que les variants génétiques les plus impliqués dans les variations phénotypiques et d'améliorer la précision des estimations de l'héritabilité.
La dernière partie de cette thèse est consacrée à l'estimation d'héritabilité de données binaires, dans le but d'étudier la part de facteurs génétiques impliqués dans des maladies complexes. Nous proposons d'étudier les propriétés théoriques de la méthode développée par Golan et al. (2014) pour des données de cas-contrôles et très efficace en pratique. Nous montrons notamment la consistance de l'estimateur de l'héritabilité proposé par Golan et al. (2014).

Title: Heritability estimation in high-dimensional mixed models : theory and applications

Keywords : High Dimension, Heritability, Mixed Models, Variable Selection.

Abstract : We study statistical methods to estimate the heritability of a biological trait, which is the proportion of variations of this trait that can be explained by genetic factors. First, we propose to study the heritability of quantitative traits using high-dimensional sparse linear mixed models. We investigate the theoretical properties of the maximum likelihood estimator for the heritability and we show that it is a consistent estimator and that it satisfies a central limit theorem with a closed-form expression for the asymptotic variance. This result, supported by an extended numerical study, shows that the variance of our estimator is strongly affected by the ratio between the number of observations and the size of the random genetic effects. More precisely, when the number of observations is small compared to the size of the genetic effects (which is often the case in genetic studies), the variance of our estimator is very large. This motivated the development of a variable selection method in order to capture the genetic variants which are involved the most in the phenotypic variations and provide more accurate heritability estimations. We propose then a variable selection method adapted to high dimensional settings and we show that, depending on the number of genetic variants actually involved in the phenotypic variations, called causal variants, it was a good idea to include or not a variable selection step before estimating heritability.
The last part of this thesis is dedicated to heritability estimation for binary data, in order to study the proportion of genetic factors involved in complex diseases. We propose to study the theoretical properties of the method developed by Golan et al. (2014) for case-control data, which is very efficient in practice. Our main result is the proof of the consistency of their heritability estimator.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France