Electronic Journal of Statistics
Vol. 12 (2018) 1662–1716
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1424

Heritability estimation in case-control studies

Anna Bonnet

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

AgroParisTech - UMR INRA MIA 518
16, Rue Claude Bernard - 75005 Paris
e-mail: anna.bonnet@agroparistech.fr

Abstract: In the field of genetics, 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, especially when the causes are plural and not clearly identified. Although the literature regarding heritability estimation for binary traits is less rich than for quantitative traits, several methods have been proposed to estimate the heritability of complex diseases. However, to the best of our knowledge, the existing methods are not supported by theoretical grounds. Moreover, most of the methodologies do not take into account a major specificity of the data coming from medical studies, which is the oversampling of the number of patients compared to controls. We propose in this paper to investigate the theoretical properties of the Phenotype Correlation Genotype Correlation (PCGC) regression developed by Golan, Lander and Rosset (2014), which is one of the major techniques used in statistical genetics and which is very efficient in practice, despite the oversampling of patients. Our main result is the proof of the consistency of this estimator, under several assumptions that we will state and discuss. We also provide a numerical study to compare two approximations leading to two heritability estimators.

Keywords and phrases: Case-control studies, heritability, high dimension, mixed models.

Received September 2017.

Contents

1 Introduction 1663
2 Model and definitions 1666
2.1 Liability model 1666
2.2 Case control study 1667
3 Heritability estimator 1668
3.1 Description of the PCGC regression 1668
3.2 Our method 1669
4 Consistency of the heritability estimator $\hat{\eta}^{(1)}$ 1671
5 Second order approximation of $\mathbb{E}[\mathbf{W}_i\mathbf{W}_j|\mathbf{Z}, \epsilon_i = \epsilon_j = 1]$ 1672

6 Numerical study 1673
6.1 Simulation process 1674
6.2 Results 1674
6.3 Normalization of $\mathbf{Z}$ in the study 1675
7 Discussion 1676
8 Proofs 1677
8.1 Summary of the proofs 1677
8.1.1 Short proof of Lemma 1 1677
8.1.2 Short proof of Lemma 2 1678
8.1.3 Short proof of Lemma 3 1679
8.1.4 Short proof of Lemma 4 1680
8.2 Taylor development of $\mathbb{E}[\mathbf{W}_i\mathbf{W}_j|\mathbf{Z}, \epsilon_i = \epsilon_j = 1]$ in Model (2.1) . 1680
8.3 Proof of Theorem 1 1684
8.3.1 Properties of $\mathbf{Z}$ 1684
8.3.2 Proof of Lemma 1 1684
8.3.3 Proof of Lemma 2 1687
8.3.4 Proof of Lemma 3 1691
8.3.5 Proof of Lemma 4 1694
8.4 Second order approximation of $\mathbb{E}[\mathbf{W}_i\mathbf{W}_j|\mathbf{Z}, \epsilon_i = \epsilon_j = 1]$ 1705
A Appendix 1709
A.1 Proof of Equation (2.7) 1709
A.2 Proof of Equation (3.2) 1710
A.3 Proof of Equation (3.6) 1711
A.4 Proof of Proposition 1 1712
Acknowledgments 1715
References 1715

1. Introduction

In the field of genetics, 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 challenge for diseases that are suspected to have a strong genetic component but whose causes are often vague and multiple. Indeed, determining a high value of heritability 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, none of them has been validated theoretically. Golan, Lander and Rosset (2014) developed a method, called phenotype correlation genotype correlation (PCGC) regression, 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 the PCGC method.

Let us first recall the main existing methods to estimate the 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

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

where $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_m)'$ is the vector of observations of a phenotype of interest, $\mathbf{X}$ is a $m \times p$ 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 $\sigma_u^{\star 2}\mathrm{Id}_{\mathbb{R}^N}$ and $\sigma_e^{\star 2}\mathrm{Id}_{\mathbb{R}^m}$ respectively. Moreover, $\mathbf{Z}$ is a $m \times N$ matrix which contains the genetic information. They proposed to estimate the parameter

$$\eta^\star = \frac{N\sigma_u^{\star 2}}{N\sigma_u^{\star 2} + \sigma_e^{\star 2}}, \tag{1.2}$$

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 and Thompson (1971), Searle, Casella and McCulloch (1992), Yang et al. (2011), Pirinen, Donnelly and Spencer (2013), Zhou and Stephens (2012).

From a theoretical point of view, Bonnet, Gassiat and Levy-Leduc (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 $\eta^\star$.

The previous model and the corresponding methods obviously do not apply when considering non continuous traits. However, the quantitative and the binary cases can be related by assuming the existence of an underlying Gaussian variable linked to the binary phenotype. More precisely, it consists of assuming that the observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ are distributed according to the following Generalized Linear Mixed Model (GLMM):

$$\mathbf{Y}_i \sim \mathcal{B}(p_i) \tag{1.3}$$

with $p_i = g^{-1}(\mathbf{G}_i)$ where g is a link function and $\mathbf{G}_i$ a Gaussian variable.

A particular case, which is very often used to define heritability of binary traits, is when g is a probit link function.

This model 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:

$$\mathbf{Y}_i = \mathbb{1}_{\{\mathbf{l}_i > t\}}, \tag{1.4}$$

with $\mathbf{l}_i$ defined by

$$\mathbf{l} = \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{1.5}$$

and $\mathbf{l} = (\mathbf{l}_1, \ldots, \mathbf{l}_n)$, $\mathbf{u} \sim \mathcal{N}(0, \sigma_u^{\star 2} \mathrm{Id}_{\mathbb{R}^N})$ and $\mathbf{e} \sim \mathcal{N}(0, \sigma_e^{\star 2} \mathrm{Id}_{\mathbb{R}^n})$, like in classical LMM defined in Equation (1.1). The heritability is then intuitively defined "at the liability scale", which means that it is the heritability of the unobserved continuous trait $\mathbf{l}$, and it is given by the same expression (1.2) as for quantitative traits.

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,

$$K = \mathbb{P}(\mathbf{Y}_i = 1) = \mathbb{P}(\mathbf{l}_i > t). \tag{1.6}$$

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

Several methods were established to estimate heritability in Model (1.3): among them we can quote the MCMC method of Hadfield (2010) and the penalized quasi-likelihood approach of Breslow and Clayton (1993). The theoretical properties of these estimators have not been demonstrated and their numerical performances can be found in the comparative study of de Villemereuil, Gimenez and Doligez (2013). 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, Lander and Rosset (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 true heritability is high.

However, all the aforementioned methods raise two main concerns: first, they have no theoretical validation. Second, they do not take into account an essential element of case-control studies: 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, which had been noticed for instance by Lee et al. (2011) but had never been properly addressed, is handled by the PCGC approach of Golan, Lander and Rosset (2014), who proposed a moment based method to estimate the heritability. The ground of their methodology was to compute an approximate quantity of the expectation E 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. This approach will be further described in Section 3.1.

Since the PCGC method presented very good numerical results but was not supported by theoretical grounds, we propose in this paper to investigate the theoretical properties of their method. Our main result is to show that the least squares estimator obtained with the first order approximation of E provides a consistent estimator of $\eta^\star$. We also propose a simulation study to compare

the numerical performances of the estimators obtained with first and second order approximations of E. We show in particular that the computational times associated to the second order estimator are substantially larger with no obvious improvement from the statistical point of view.

The model we study and the main definitions are given in Section 2. Section 3 contains the first order approximation of the expectation E with the corresponding estimator of $\eta^\star$ and Section 4 presents our consistency result for this estimator. The second order approximation of E is given in Section 5 and the numerical comparison of the two estimators can be found in Section 6. In Section 7, we discuss the results and potential perspectives. Finally, the proofs are given in Section 8.

2. Model and definitions

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 affected (then, individual i is called a case) and $\mathbf{Y}_i = 0$ if the individual i is unaffected (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

$$\mathbf{Y}_i = \mathbb{1}_{\{\mathbf{l}_i > t\}}, \tag{2.1}$$

where t is a given threshold, related to the prevalence K by (1.6), and the $\mathbf{l}_i$'s are defined as

$$\mathbf{l} = \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{2.2}$$

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 $\mathbf{e} \sim \mathcal{N}(0, \sigma_e^{\star 2}\mathrm{Id}_{\mathbb{R}^m})$. The vector $\mathbf{u}$ corresponds to the genetic effects and $\mathbf{e}$ to the environmental effects. Moreover, $\mathbf{Z}$ is a $m \times N$ 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 $A = (A_{i,k})_{1 \le i \le m,\, 1 \le k \le N}$ by

$$\mathbf{Z}_{i,k} = \frac{A_{i,k} - \overline{A}_k}{s_k},\ i = 1, \dots, m,\ k = 1, \dots, N\ , \tag{2.3}$$

where

$$\overline{A}_k = \frac{1}{m}\sum_{i=1}^{m} A_{i,k},\ s_k^2 = \frac{1}{m}\sum_{i=1}^{m}(A_{i,k} - \overline{A}_k)^2,\ k = 1, \dots, N\ . \tag{2.4}$$

In (2.3) and (2.4) the $A_{i,k}$'s are such that for each k in $\{1, \dots, N\}$ the $(A_{i,k})_{1 \le i \le m}$ are independent and identically distributed random variables and such that the columns of A are independent.

In practice, the matrix A contains, for all the individuals in the study, the genetic information described by the Single Nucleotide Polymorphisms (SNPs).

More precisely, at each SNP, the genotype can be either qq, qQ or QQ, q being the less frequent (or minor) allele.

Then 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 4. However, note that the assumption of independence between the columns of A is quite strong, since in particular it does not take into account the linkage disequilibrium, that is precisely the correlation betweens genetic variants. To the best of our knowledge, the other theoretical works regarding estimation of heritability (Jiang et al. (2016) and Bonnet, Gassiat and Levy-Leduc (2015)) also neglect these correlations, even in the Gaussian scenario, which shows the difficulty of getting rid of this assumption.

With the definition (2.3), 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:

$$\eta^\star = \frac{N\sigma_u^{\star 2}}{N\sigma_u^{\star 2} + \sigma_e^{\star 2}}. \tag{2.5}$$

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 (1.6), we estimate directly $t/\sigma^\star$.

Note that in Model (2.2), we consider a particular case of linear mixed model where there is no fixed effects. In the PCGC method, Golan, Lander and Rosset (2014) propose a solution to handle covariates that we did not study here, but it would be interesting to investigate as well the theoretical properties of such an approach. This point is further discussed in Section 7.

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. We recall that m corresponds to the total size of the population and we define n the number of individuals in the study. Each individual of the population will either be selected or not for the study with a probability depending on their status (case or control).

More precisely, if $p_{control}$ denotes the probability for a control to be selected in the study, we can define the corresponding variable $U_i \sim \mathcal{B}(p_{control})$ which is equal to 1 if individual i is part of the study. Similarly we define the probability p_{case} for a case to be selected for the study and the corresponding variable $V_i \sim \mathcal{B}(p_{case})$. We will increase the size population until we obtain n individuals in the study. 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 $U_1, \ldots, U_m, V_1, \ldots, V_m$ are independent and independent of $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ 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

$$p_{case} = 1, \tag{2.6}$$

it implies that

$$p_{control} = \frac{K(1-P)}{P(1-K)}. \tag{2.7}$$

The proof of (2.7) is given in Appendix A.1. Equation (2.6) means that all cases are accepted in the study and it is usually called a "full ascertainment" assumption (see for instance Golan, Lander and Rosset (2014)).

3. Heritability estimator

3.1. Description of the PCGC regression

Golan, Lander and Rosset (2014) considered a version of Model (2.2), 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 V_g defined for all $1 \leq i, j \leq n$, as:

$$(V_g)_{i,j} = \begin{cases} \eta^\star \mathbf{G}_{i,j} & \text{if } i \neq j \\ \eta^\star & \text{if } i = j. \end{cases}$$

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 of the PCGC regression is a least squares estimator obtained by minimizing

$$\sum_{i \neq j} \left(\mathbf{W}_i \mathbf{W}_j - \mathbb{E}[\mathbf{W}_i \mathbf{W}_j | \epsilon_i = \epsilon_j = 1] \right)^2. \tag{3.1}$$

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, Lander and Rosset (2014) proposed to take advantage of the fact that the correlations $\mathbf{G}_{i,j}$ are typically small for $i \neq j$.

The ground of the method is to write

$$\begin{aligned} &\mathbb{E}[\mathbf{W}_i \mathbf{W}_j | \epsilon_i = \epsilon_j = 1] \\ &= \frac{\frac{1-P}{P}\mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 1) - \frac{K(1-P)}{P(1-K)}\mathbb{P}(\mathbf{Y}_i \neq \mathbf{Y}_j) + \frac{K^2(1-P)}{P(1-K)^2}\mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 0)}{\mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 1) + \left(\frac{K(1-P)}{P(1-K)}\right)^2 \mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 0) + \frac{K(1-P)}{P(1-K)}\mathbb{P}(\mathbf{Y}_i \neq \mathbf{Y}_j)} \end{aligned} \tag{3.2}$$

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 (3.2) can be found in Appendix A.2.

This approximation, plugged in the least squares criterion (3.1), led to the heritability estimator given by

$$\hat{\eta} = \left[\frac{\sum_{i \neq j} \mathbf{W}_i \mathbf{W}_j \mathbf{G}_{i,j}}{c \sum_{i \neq j} \mathbf{G}_{i,j}^2} \wedge 1 \right] \vee 0, \tag{3.3}$$

where

$$c = \phi(t)^2 \frac{P(1-P)}{K^2(1-K)^2}, \tag{3.4}$$

ϕ being the density of the standard Gaussian distribution. The proof of (3.3) and (3.4) can be found in the supplementary material from Golan, Lander and Rosset (2014).

3.2. Our method

In Model defined in (2.1) and (2.2), 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 $1 \leq i, j \leq n$,

$$\mathbf{G}_N(i,j) = \frac{1}{N}\sum_{k=1}^{N} \mathbf{Z}_{i,k}\mathbf{Z}_{j,k}. \tag{3.5}$$

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, Lander and Rosset (2014). A key element is to notice that $\Sigma^{(N)}$ is close to the $n \times n$ identity matrix, more precisely

$$\Sigma^{(N)} = \begin{pmatrix} 1 + \eta^\star \frac{A_N(i)}{\sqrt{N}} & \eta^\star \frac{B_N(i,j)}{\sqrt{N}} \\ \eta^\star \frac{B_N(i,j)}{\sqrt{N}} & 1 + \eta^\star \frac{A_N(j)}{\sqrt{N}} \end{pmatrix} \tag{3.6}$$

where $A_N(i) = O_p(1)$, $A_N(j) = O_p(1)$ and $B_N(i,j) = O_p(1)$. The proof of (3.6) can be found in Appendix A.3.

Then, following the idea of Golan, Lander and Rosset (2014), we propose to approximate

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

defined in Equation (3.2) 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 8.2.

A first order approximation of $\mathbb{E}[\mathbf{W}_i\mathbf{W}_j|\mathbf{Z}, \epsilon_i = \epsilon_j = 1]$, plugged in (3.1), leads to the same estimator $\hat{\eta}^{(1)}$ as the one obtained with the PCGC regression. Indeed, we obtain

$$\hat{\eta}^{(1)} = \left[\frac{\sum_{i \neq j} \mathbf{W}_i\mathbf{W}_j\mathbf{G}_N(i,j)}{c\sum_{i \neq j} \mathbf{G}_N(i,j)^2} \wedge 1 \right] \vee 0, \tag{3.7}$$

where $c = \phi(t)^2 \frac{P(1-P)}{K^2(1-K)^2}$.

In Section 5, we consider the second order approximation, which is different from the one devised by Golan, Lander and Rosset (2014).

Remark 1. Note that in practice, we cannot access directly $\mathbf{G}_N(i,j)$ defined in (3.5), since the matrix $\mathbf{Z}$ should be centered and normalized in the whole population. This is obviously a limitation, but we propose to show with a numerical study that replacing $\mathbf{Z}$ by $\tilde{\mathbf{Z}}$ which has been centered and normalized in the study has a small influence on the heritability estimates. The results are displayed in Section 6.3.

Remark 2. The main difficulty to study the theoretical properties of $\hat{\eta}^{(1)}$ is due to the approximation

$$\mathbb{E}[\mathbf{W}_i\mathbf{W}_j|\mathbf{Z}, \epsilon_i = \epsilon_j = 1] \simeq c\eta^\star G_N(i,j),$$

neglecting a remainder term which depends on $A_N(i)$, $A_N(j)$ and $B_N(i,j)$ defined in (3.6), which means that it varies for each pair of individuals i and j.

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

In this section, we consider the heritability estimator $\hat{\eta}^{(1)}$ defined in Equation (3.7).

Assumption 1. 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 (2.3) and σ_k^2 is the variance of $A_{i,k}$.

Assumption 2. **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 1. *Let* $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ *satisfy Model* (1.4) *with* A *satisfying Assumptions 1 and 2, and* $\hat{\eta}^{(1)}$ *the estimator of* $\eta^\star$ *defined in Equation* (3.7)*. Then, as* $n, N \to \infty$ *such that* $n/N \to a \in (0, +\infty)$*,*

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

Note that we focus on the case where both the number n of individuals and the number N of genetic variants tend to infinity, which is the same framework chosen for instance by Jiang et al. (2016) and Bonnet, Gassiat and Levy-Leduc (2015). In practice, these values are obviously finite upper bounded, for instance by the length of the human genome for N.

The proof of Theorem 1 relies on the following lemmas.

Lemma 1. *When* n *and* N *tend to infinity and* n/N *tends to* a*,*

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

We will then have to focus on

$$\frac{1}{n}\sum_{i\neq j} \mathbf{W}_i\mathbf{W}_j\mathbf{G}_N(i,j) = \left[\frac{1}{n}\sum_{i\neq j}(\mathbf{W}_i\mathbf{W}_j - \mathbb{E}[\mathbf{W}_i\mathbf{W}_j|\mathbf{Z}, \epsilon_i = \epsilon_j = 1])\mathbf{G}_N(i,j) + \frac{1}{n}\sum_{i\neq j}\mathbb{E}[\mathbf{W}_i\mathbf{W}_j|\mathbf{Z}, \epsilon_i = \epsilon_j = 1]\mathbf{G}_N(i,j)\right]. \quad (4.1)$$

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

$$\epsilon_N = \frac{1}{N^{\frac{1}{2}-\gamma}} \tag{4.2}$$

with γ a positive number such that $\gamma < 1/10$.

The choice of ϵ_N is crucial, since it has to be large enough so that on the one hand, the probability not to be in event E_N is very small (Lemma 2). On the other hand, ϵ_N must be small enough to verify Lemmas 3 and 4, which ensure respectively that if E_N holds, the first term converges to 0 and the second term of (4.1) will converge to $\eta^\star$ (up to a constant).

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 2. *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 2, $\hat{\eta}^{(1)} \mathbb{1}_{E_N^c}$ 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 3. *When n and N tend to infinity and n/N tends 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 (3.4).

Lemma 4. *When n and N tend to infinity and n/N tends 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 3 and 4 achieve the proof of Theorem 1.

Section 8.1 contains a short version of the proofs, while the detailed proofs of Lemmas 1, 2, 3 and 4 are given in Section 8.3.

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:

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

The proof of this computation is detailed in Section 8.4.

The minimizer in η of the quantity

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

is the root of a third order polynomial and could be found thanks to a closed-form formula. Since the expressions are quite complex, we propose here, for the sake of simplicity, to 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, Lander and Rosset (2014).

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 3 and 5. Since Golan, Lander and Rosset (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.

6.1. Simulation process

In this simulation study, we generated data sets with $n = 200$, $N = 10000$ that is smaller than the size of typical data sets ($n \simeq 5000$, $N \simeq 500000$ for instance). The reason is purely computational, since we have to generate data for $m \simeq n/K$ individuals in the population in order to have n individuals in the study. However, we choose the values of n and N such that the classical scenario where $N >> n$ is respected. 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 size of the population can vary from one sample to another. In practice, we generate new individuals in the population until we have n individuals in the study.
- We generate the SNPs matrix A, the columns of which are independent binomial variables with parameters 2 and p_j, p_j being uniformly generated between 0.1 and 0.5 (it represents the probability of appearance of the less frequent SNP). The matrix $\mathbf{Z}$ is then obtained by centering and normalizing A in the whole population. Notice that we use $\mathbf{Z}$ to generate the data, but since we would not access to the whole matrix in practice, we will use for the estimations the matrix that we denote $\tilde{\mathbf{Z}}$, which is centered and normalized in the study. This point is further discussed in Section 6.3
- We generate the Gaussian random effects $\mathbf{u}$ and $\mathbf{e}$ with respective variances $\sigma_u^{\star 2} = \eta^\star/N$ and $\sigma_e^{\star 2} = 1 - \eta^\star$.
- We generate liabilities, from which we generate binary observations in order to have a prevalence equal to K in the general population.
- For each individual, we determine 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 (2.7).

6.2. Results

Figure 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.

Table 1 and Figure 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 increases substantially and the "slower" estimator needs up to 13500 seconds, that is almost 4 hours, to compute an estimation of $\eta^\star$.

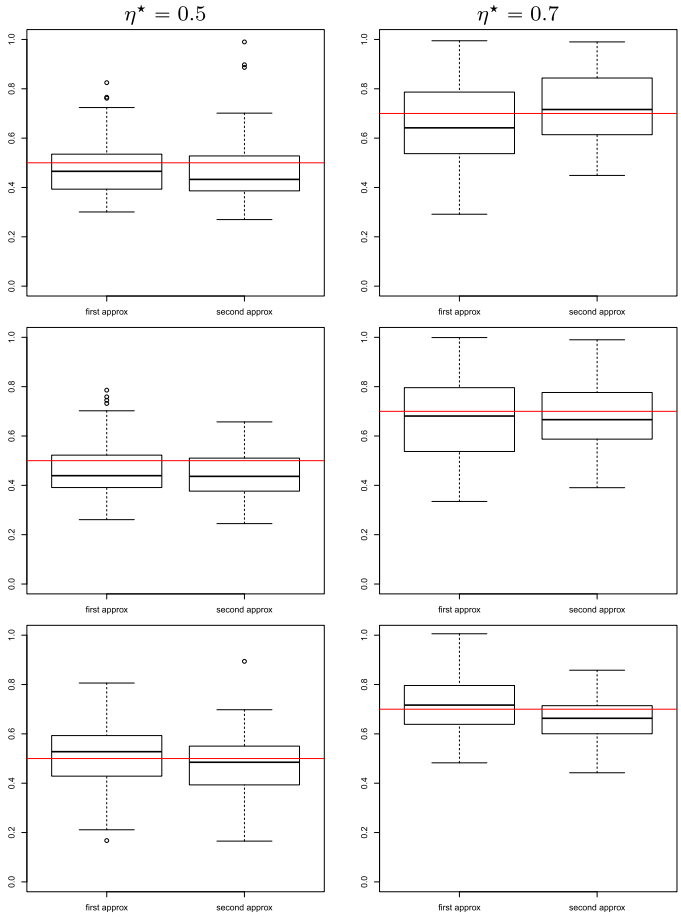

FIG 1. *Boxplots for* $\hat{\eta}^{(1)}$ *("first approx") and* $\hat{\eta}^{(2)}$ *("second approx") for different values of* $\eta^\star$*:* 0.5 *(left),* 0.7 *(right) and different values of the prevalence* K*:* 0.005 *(top),* 0.01 *(middle),* 0.1 *(bottom). The sample size is* $n = 200$ *and* $N = 10000$*. Each boxplot is generated from* 100 *replications.*

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 $\hat{\eta}^{(1)}$.

6.3. Normalization of $\mathbf{Z}$ *in the study*

We propose to study the impact of performing normalization described in Equations (2.3) and (2.4) in the study and not in the whole population. Since for synthetic data we have access to the complete matrix $\mathbf{Z}$, we propose to compare our results to those we would obtain when considering the reduction of $\mathbf{Z}$ to

TABLE 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

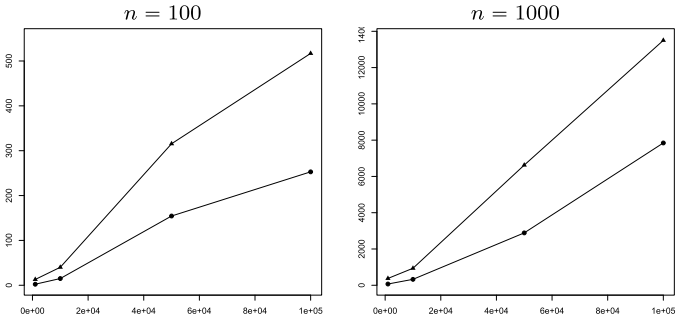

FIG 2. *Time in seconds to compute an estimation of $\eta^\star$ obtained with $\hat{\eta}^{(1)}$ and $\hat{\eta}^{(2)}$ for $n = 100$ (left) and $n = 1000$ (right) and for different values of N (from 1000 to 10^5).*

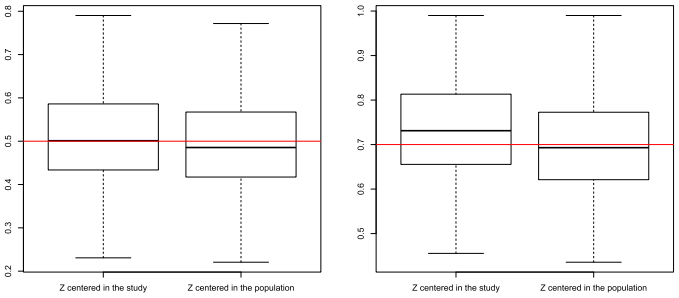

FIG 3. *Comparison of heritability estimates obtained when centering* **Z** *in the study or in the whole population, for two values of $\eta^\star$: 0.5 (left) and 0.7 (left) and for $K = 0.1$.*

the individuals of the study. The results are displayed in Figure 3, and we can see the minor changes obtained between the manners of centering the genetic information matrix.

7. Discussion

In this paper, we propose theoretical grounds to support the heritability estimator in case-control studies developed by Golan, Lander and Rosset (2014). We prove indeed its consistency in the framework where both the number n of individuals and the number N of SNPs tend to infinity, when the ratio n/N tends to

a constant a. This consistency result was obtained under several assumptions, the necessity of which it would be interesting to question. For instance, removing the Gaussianity assumption on the distribution of the random effects could allow to take into account possible sparsity and remains a very challenging issue.

The independence of the columns of the SNP matrix is also a very strong assumption, which neglects the linkage disequilibrium (LD) between alleles. Going beyond this assumption seems challenging from the theoretical point of view, even for quantitative traits (Jiang et al., 2016; Bonnet, Gassiat and Levy-Leduc, 2015). Indeed, independence is required to prove consistency, to determine the order of magnitude of different quantities but also to be able to apply large deviation results that are essential to prove Lemma 2. For quantitative traits, LD has been shown to result in an overestimated contribution of variants in strong LD (Speed et al., 2012). Despite theoretical limitations, several efficient filtering procedures (Patterson, Price and Reich, 2006; Speed et al., 2012) were proposed to modify the kinship matrix G before estimating heritability.

Another sensible question is the closeness to the asymptotic results on finite samples. One key ingredient could be a careful calibration of ϵ_N defined in (4.2) of Theorem 1. This quantity is indeed constrained by a lower bound to ensure that Lemma 2 holds and an upper bound coming from Lemmas 3 and 4. Determining the optimal balance for ϵ_N should lead to a lower bound of the convergence rate of the estimation procedure. The numerical results also confirm the good performance of the PCGC method on finite samples, and in particular the similar results obtained with first and second order approximations suggest that the remainder term is indeed negligible compared to the main term.

We would also like to extend our theoretical grounds to a more general model that includes fixed effects, and for instance investigate the properties of the PCGC regression in this scenario. Finally, it would also be interesting to complete this work with theoretical results which could allow the user to compute accurate confidence intervals, similarly to existing results for quantitative traits.

8. Proofs

8.1. Summary of the proofs

Since the proof of Theorem 1 is quite long and requires heavy computations, we propose in this section a short version of the main arguments that we used to prove Lemmas 1, 2, 3 and 4.

8.1.1. Short proof of Lemma 1

Let us write

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

Then let us show that

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

and

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

We will prove these two results by computing the expectation and variance and both terms, the order of magnitude of which we will evaluate thanks to the properties on $\mathbf{Z}$ that are given in Proposition 1 of Section 8.

8.1.2. Short proof of Lemma 2

We will show that $\mathbb{P}(E_N^c)$ can be upperbounded by a sum of two terms of theform

$$Cn^{\alpha}\exp(-\beta_N),$$

with C and α being positive constants and β_N going to infinity when N tends to infinity.

These two terms come from the first upper bound

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

We will rewrite each term with the $A_{i,k}$ instead of $\mathbf{Z}_{i,k}$ so that we can use Assumption 1. We need for instance to upper bound the probability that the difference between empirical mean and theoretical mean exceeds a certain value, that is:

$$\mathbb{P}(\bar{A}_k - m_k \geq \sqrt{\delta}).$$

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

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

Then, we use Assumption **1.2** to upper bound the right term and we obtain that, for all positive values of λ,

$$\mathbb{P}(n(\bar{A}_k-m_k)\geq n\sqrt{\delta}) \leq C\exp\left\{-n\sqrt{\delta}\lambda+nd\lambda^2\right\}. \tag{8.1}$$

The right term of (8.1) is minimum when

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

which implies in particular that

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

Similarly we will upper bound all terms using on the one hand Chernoff inequality and on the other hand, one or several assumptions from Assumptions 1 and 2. The detailed computations are given in Section 8.3.3.

8.1.3. Short proof of Lemma 3

According to the results of Section 8.3.2, we have

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

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 compute an explicit form of the remainder term $R_N(i,j)$ and then 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

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

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 $\gamma<1/10$.

8.1.4. Short proof of Lemma 4

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

$$\frac{1}{n^2}\sum_{\substack{i_1\neq i_2\\ i_3\neq i_4}} \mathbb{E}[(\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}\mathbf{W}_{i_3}\mathbf{W}_{i_4}|\mathbf{Z}] - \mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}|\mathbf{Z}]\mathbb{E}[\mathbf{W}_{i_3}\mathbf{W}_{i_4}|\mathbf{Z}])\mathbf{G}_N(i_1,i_2) \times \mathbf{G}_N(i_3,i_4)\mathbb{1}_{E_N}] \to 0 \tag{8.2}$$

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 (8.21).

-If card($\{i_1,i_2,i_3,i_4\}$)=2, since the sum in Equation (8.2) has only $n(n-1)$ terms, the upper bound of $\mathbf{G}_N(i_1,i_2)\mathbf{G}_N(i_3,i_4)$ on the event E_N will be enough to obtain the convergence to 0.

-If card($\{i_1,i_2,i_3,i_4\}$)=3, we will first prove that $\mathbb{E}[\mathbf{W}_{i_1}^2\mathbf{W}_{i_2}\mathbf{W}_{i_3}|\mathbf{Z}]$ has no term of order less than $1/\sqrt{N}$, that is no constant term.

Then the other terms can be upper bounded on E_N by terms of the form ϵ_N^k, with k large enough to compensate the $n(n-1)(n-2)$ terms of the sum.

-If card($\{i_1,i_2,i_3,i_4\}$)=4, each term can be handled using one of the following arguments:

- The order of the term is high enough so that it can be upper bounded on E_N by terms of the form ϵ_N^k, with k large enough to compensate the $n(n-1)(n-2)(n-2)$ terms of the sum.
- The term is equal to 0 (we will propose a detailed computation to verify this).
- The terms are equal in $\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}\mathbf{W}_{i_3}\mathbf{W}_{i_4}|\mathbf{Z}]$ and $\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}|\mathbf{Z}]\mathbb{E}[\mathbf{W}_{i_3}\mathbf{W}_{i_4}|\mathbf{Z}]$.

This achieves the proof of Equation (8.2).

8.2. Taylor development of $\mathbb{E}[\mathbf{W}_i\mathbf{W}_j|\mathbf{Z},\epsilon_i=\epsilon_j=1]$ in Model (2.1)

According to Equation (3.2), 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 (3.6), which will be demonstrated in Appendix A.3, that is

$$\Sigma^{(N)} = \begin{pmatrix} 1+\eta^\star\frac{A_N(i)}{\sqrt{N}} & \eta^\star\frac{B_N(i,j)}{\sqrt{N}} \\ \eta^\star\frac{B_N(i,j)}{\sqrt{N}} & 1+\eta^\star\frac{A_N(j)}{\sqrt{N}} \end{pmatrix}, \tag{8.3}$$

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

We have

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

Using a first order Taylor development around $\frac{A_N(i)}{\sqrt{N}}$, $\frac{A_N(j)}{\sqrt{N}}$ and $\frac{B_N(i,j)}{\sqrt{N}}$,

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

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,

$$\alpha_N = -(A_N(i)A_N(j) - B_N(i,j)^2)\frac{\eta^{\star 2}}{N}$$

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

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,

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

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,

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

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

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

with

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

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,

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

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

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

This remainder and its order will be carefully studied in Section 8.3.4.

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

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

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

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 (3.2) leads to:

$$\frac{\eta^\star}{\sqrt{N}}B_N(i,j)\phi(t)^2\frac{(1-P)}{P(1-K)^2} + r_N, \tag{8.4}$$

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

$$\mathbb{E}[\mathbf{W}_i\mathbf{W}_j|\mathbf{Z},\epsilon_i=\epsilon_j=1] = \frac{\frac{\eta^\star}{\sqrt{N}}B_N(i,j)\phi(t)^2\frac{(1-P)}{P(1-K)^2}+r_N}{\frac{K^2}{P^2}+\tilde{r}_N} \tag{8.5}$$

$$= \eta^\star \mathbf{G}_N(i,j)\phi(t)^2\frac{P(1-P)}{K^2(1-K)^2}+R_N(i,j) \tag{8.6}$$

where

$$R_N(i,j) = \left(\frac{\eta^\star}{\sqrt{N}}B_N(i,j)\phi(t)^2\frac{(1-P)}{P(1-K)^2}+r_N\right)\tilde{r}_N+\frac{K^2}{P^2}r_N. \tag{8.7}$$

8.3. Proof of Theorem 1

8.3.1. Properties of $\mathbf{Z}$

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 A.4.

8.3.2. Proof of Lemma 1

Let us prove that, when n and N tend to infinity and n/N tends 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

$$\frac{1}{n}\frac{1}{N^2}\sum_{i\neq j}\sum_{k=1}^{N}\mathbf{Z}_{i,k}^2\mathbf{Z}_{j,k}^2 \xrightarrow{P} a \tag{8.8}$$

and

$$\frac{1}{n}\frac{1}{N^2}\sum_{i\neq j}\sum_{k\neq l}\mathbf{Z}_{i,k}\mathbf{Z}_{j,k}\mathbf{Z}_{i,l}\mathbf{Z}_{j,l} \xrightarrow{P} 0. \tag{8.9}$$

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,

$$\begin{aligned}\mathrm{Var}(\frac{1}{n}\frac{1}{N^2}\sum_{i\neq j}\sum_{k=1}^{N}\mathbf{Z}_{i,k}^2\mathbf{Z}_{j,k}^2) = \frac{1}{n^2}\frac{1}{N^4}\sum_{k=1}^{N}\sum_{i_1\neq j_1}\sum_{i_2\neq j_2}\mathbb{E}[\mathbf{Z}_{i_1,k}^2\mathbf{Z}_{j_1,k}^2\mathbf{Z}_{i_2,k}^2\mathbf{Z}_{j_2,k}^2]\\ - \frac{1}{n^2}\frac{1}{N^4}\sum_{k=1}^{N}\left(\sum_{i\neq j}\mathbb{E}[\mathbf{Z}_{i,k}^2\mathbf{Z}_{j,k}^2]\right)^2\end{aligned} \tag{8.10}$$

The second term of (8.10) can be rewritten as:

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

$$\begin{aligned}&\sum_{i_1\neq j_1}\sum_{i_2\neq j_2}\mathbb{E}[\mathbf{Z}_{i_1,k}^2\mathbf{Z}_{j_1,k}^2\mathbf{Z}_{i_2,k}^2\mathbf{Z}_{j_2,k}^2]\\ &\leq \mathbb{E}[\sum_{i_1,j_1,i_2,j_2}\mathbf{Z}_{i_1,k}^2\mathbf{Z}_{j_1,k}^2\mathbf{Z}_{i_2,k}^2\mathbf{Z}_{j_2,k}^2] = \mathbb{E}\left[\sum_{i=1}^{n}\mathbf{Z}_{i,k}^2\right]^4 = n^4\end{aligned}$$

This last equality comes from the definition of $\mathbf{Z}$ as a centered and normalized variable given in Equation (2.3), which implies that for all k,

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

Then,

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

This proves (8.8).

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

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

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

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:

$$\begin{aligned}&\sum_{i_1\neq j_1}\sum_{i_2\neq j_2}\mathbb{E}[\mathbf{Z}_{i_1,k}\mathbf{Z}_{i_2,k}\mathbf{Z}_{j_1,k}\mathbf{Z}_{j_2,k}]\mathbb{E}[\mathbf{Z}_{i_1,l}\mathbf{Z}_{i_2,l}\mathbf{Z}_{j_1,l}\mathbf{Z}_{j_2,l}]\\ &= 2\sum_{i\neq j}\mathbb{E}[\mathbf{Z}_{i,k}^2\mathbf{Z}_{j,k}^2]\mathbb{E}[\mathbf{Z}_{i,l}^2\mathbf{Z}_{j,l}^2]\end{aligned}$$

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

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 (8.9).

8.3.3. Proof of Lemma 2

Note that

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

Let δ be a positive real number such that $\sqrt{\delta}/2c \in V_0$ and $\delta < \frac{\delta_{min}}{4}$, where V_0 and δ_{min} are defined in Assumptions 1 and **2.1** respectively.

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

Note also that

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

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

Observe that

$$\left\{\sum_{i=1}^{n}(A_{i,k} - m_k + m_k - \bar{A}_k)^2 \leq n\delta\right\} \subset \left\{|\bar{A}_k - m_k| \geq \sqrt{\delta}\right\} \cup \left\{\sum_{i=1}^{n}(A_{i,k} - m_k)^2 \leq 4n\delta\right\}. \tag{8.11}$$

Let us show that

$$\mathbb{P}(|\bar{A}_k - m_k| \geq \sqrt{\delta}) \leq 2C\exp\left\{-\frac{n\delta}{4d}\right\}. \tag{8.12}$$

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

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

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

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

$$\mathbb{P}(n(\bar{A}_k - m_k) \geq n\sqrt{\delta}) \leq C\exp\left\{-n\sqrt{\delta}\lambda + nd\lambda^2\right\}. \tag{8.13}$$

The right term of (8.13) is minimum when

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

which implies in particular 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,

$$\mathbb{P}(n(\bar{A}_k - m_k) \leq -n\sqrt{\delta}) \leq C\exp\left\{n\sqrt{\delta}\lambda + nd\lambda^2\right\}. \tag{8.14}$$

The right term of (8.14) is minimum when

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

which implies that

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

which proves (8.12).

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

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 (8.12), proves that

$$\mathbb{P}\left(\exists k, s_k^2 \leq \delta\right) \leq 2NC\exp\left\{-\frac{n\delta}{4d}\right\} + NC\exp\left\{-n\frac{(4\delta-\delta_{min})^2}{2d}\right\} \tag{8.15}$$

Notice that

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

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

$$\begin{aligned}\mathbb{P}\left(\sum_{k=1}^{n}(A_{i,k}-m_k)^2\geq Nn\frac{\delta\epsilon_N}{8}\right)\\ \leq C\exp\left\{-\frac{n^2N\delta^2\epsilon_N^2}{256d}+\frac{nN\delta\delta_{max}\epsilon_N}{16d}-\frac{N\delta_{max}}{4d}\right\},\end{aligned}\tag{8.16}$$

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

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

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

Since we have already proved (8.15) and (8.16), we will conclude the proof by showing that

$$\mathbb{P}\left(\left|\sum_{k=1}^{n}(A_{i,k}-m_k)(A_{j,k}-m_k)\right| \geq N\frac{\delta\epsilon_N}{4}\right) \leq 2C\exp\left\{-\frac{N\delta\epsilon_N}{64d}\right\}, \tag{8.17}$$

and

$$\mathbb{P}\left(\sum_{k=1}^{n}(\bar{A}_k-m_k)^2 \geq N\frac{\delta\epsilon_N}{4}\right) \leq N^2C\exp\left\{-\frac{N\delta\epsilon_N}{16d}\right\}. \tag{8.18}$$

(8.17) 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 (8.18) and achieves the proof of Lemma 2.

8.3.4. Proof of Lemma 3

According to the results of Section 8.3.2, we have

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

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

$$|\mathbf{G}_N(i,j)|^{k_1}|\mathbf{G}_N(i,i)-1|^{k_2}|\mathbf{G}_N(j,j)-1|^{k_3}, \tag{8.19}$$

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 3.

Let us explain why Equation (8.19) 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.

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

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

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,$\gamma_N^\sim$ and $\gamma_N^{\tilde{\sim}}$ are only products of α_N, $A_N(i)$, $A_N(j)$ and $B_N(i,j)$.

$$\begin{aligned}|\gamma_N| &\leq |A_N(j)(A_N(i)+A_N(j))\frac{\eta^{\star 2}}{N}| + |\alpha_N(1+A_N(j)\frac{\eta^\star}{\sqrt{N}})|,\\ |\gamma_N^\sim| &\leq |A_N(i)(A_N(i)+A_N(j))\frac{\eta^{\star 2}}{N}| + |\alpha_N(1+A_N(i)\frac{\eta^\star}{\sqrt{N}})| \text{ and}\end{aligned}$$

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

with

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

The integral of the first terms of $\nu_N(x,y)$ is

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

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

There are two possibilities, either

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

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

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

$$\int_t^\infty \int_t^\infty \phi(x)\phi(y)\left(\frac{x^2}{2}\frac{\eta^\star}{\sqrt{N}}A_N(i) + \frac{y^2}{2}\frac{\eta^\star}{\sqrt{N}}A_N(j) + xy\frac{\eta^\star}{\sqrt{N}}B_N(i,j) - \frac{x^2}{2}\gamma_N\right.$$

$$-\frac{y^2}{2}\tilde{\gamma_N}+xy\tilde{\tilde{\gamma}}_N\Big)^2 dxdy=\frac{1}{N}J, \tag{8.20}$$

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

Otherwise,

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

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 $\exp(\tilde{u})\leq\exp\Big\{\frac{x^2}{4}+\frac{y^2}{4}+\frac{xy}{4}\Big\}\leq\exp\Big\{\frac{3x^2}{8}+\frac{3y^2}{8}\Big\}$

Then similarly to the expression 8.20,

$$\begin{aligned}\int_t^\infty\int_t^\infty\phi(x)\phi(y)\Big(&\frac{x^2}{2}\frac{\eta^\star}{\sqrt{N}}A_N(i)+\frac{y^2}{2}\frac{\eta^\star}{\sqrt{N}}A_N(j)+xy\frac{\eta^\star}{\sqrt{N}}B_N(i,j)+\frac{x^2}{2}\gamma_N\\&-\frac{y^2}{2}\tilde{\gamma_N}-xy\tilde{\tilde{\gamma}}_N\Big)^2\exp(\tilde{u})dxdy\\&\leq\frac{1}{2\pi}\int_t^\infty\int_t^\infty\exp(-\frac{x^2}{8})\exp(-\frac{y^2}{8})\Big(\frac{x^2}{2}\frac{\eta^\star}{\sqrt{N}}A_N(i)\\&+\frac{y^2}{2}\frac{\eta^\star}{\sqrt{N}}A_N(j)+xy\frac{\eta^\star}{\sqrt{N}}B_N(i,j)+\frac{x^2}{2}\gamma_N\\&-\frac{y^2}{2}\tilde{\gamma_N}-xy\tilde{\tilde{\gamma}}_N\Big)^2dxdy\leq\frac{1}{N}J'\end{aligned}$$

where J' is finite.

Similarly to the computations made for α_N, β_N, γ_N, ν_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 (8.19).

8.3.5. Proof of Lemma 4

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

$$\frac{1}{n^2}\sum_{\substack{i_1\neq i_2\\ i_3\neq i_4}}\mathbb{E}[(\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}\mathbf{W}_{i_3}\mathbf{W}_{i_4}|\mathbf{Z}] - \mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}|\mathbf{Z}]\mathbb{E}[\mathbf{W}_{i_3}\mathbf{W}_{i_4}|\mathbf{Z}])\mathbf{G}_N(i_1,i_2)\mathbf{G}_N(i_3,i_4)\mathbb{1}_{E_N}] \to 0 \tag{8.21}$$

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 (8.21).

-If card($\{i_1,i_2,i_3,i_4\}$)=2, the corresponding terms in (8.21) 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 3. 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 (8.21) are equal to

$$\frac{1}{n^2}\sum_{i_1\neq i_2\neq i_3}\mathbb{E}[(\mathbb{E}[\mathbf{W}_{i_1}^2\mathbf{W}_{i_2}\mathbf{W}_{i_3}|\mathbf{Z}] - \mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}|Z]\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_3}|Z])\mathbf{G}_N(i_1,i_2)\mathbf{G}_N(i_1,i_3)\mathbb{1}_{E_N}]. \tag{8.22}$$

Since the sum of Equation (8.22) 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

$$\frac{1}{n^2}\sum_{i_1\neq i_2\neq i_3}\mathbb{E}[(\mathbb{E}[\mathbf{W}_{i_1}^2\mathbf{W}_{i_2}\mathbf{W}_{i_3}|\mathbf{Z}] - \mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}|\mathbf{Z}]\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_3}|\mathbf{Z}])\mathbf{G}_N(i_1,i_2)\mathbf{G}_N(i_1,i_3)\mathbb{1}_{E_N}] \to 0. \tag{8.23}$$

Let us first recall that, according to Lemma 3,

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

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,

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

can be upper bounded by a finite sum of 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

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

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 (8.22).

- 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

$$\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}\mathbf{W}_{i_3}\mathbf{W}_{i_4}|\mathbf{Z}] - \mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}|\mathbf{Z}]\mathbb{E}[\mathbf{W}_{i_3}\mathbf{W}_{i_4}|\mathbf{Z}]$$

of order $1/N$.

Let us recall that

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

where

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

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

$$\frac{1}{n^2} \sum_{i_1 \neq i_2 \neq i_3} \mathbb{E}[(\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}\mathbf{W}_{i_3}\mathbf{W}_{i_4}|\mathbf{Z}] - c^2\eta^{\star 2}\mathbf{G}_N(i_1,i_2)\mathbf{G}_N(i_3,i_4))\mathbf{G}_N(i_1,i_2)\mathbf{G}_N(i_3,i_4)\mathbb{1}_{E_N}] \to 0, \tag{8.24}$$

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*

$$\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}\mathbf{W}_{i_3}\mathbf{W}_{i_4}|\mathbf{Z}]$$

are null.

The term of order exactly $1/N$ in $\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}\mathbf{W}_{i_3}\mathbf{W}_{i_4}|\mathbf{Z}]$ 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* $\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}\mathbf{W}_{i_3}\mathbf{W}_{i_4}|\mathbf{Z}]$ *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*

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

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 (8.24).

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 (2.6) and (2.7)),

$$\mathbb{E}[\mathbf{W}_{i_1}^2\mathbf{W}_{i_2}\mathbf{W}_{i_3}|\mathbf{Z},\epsilon_{i_1}=\epsilon_{i_2}=\epsilon_{i_3}=1] = \frac{1}{\mathbb{P}(\epsilon_{i_1}=\epsilon_{i_2}=\epsilon_{i_3}=1)}$$

$$\times \left\{ \frac{(1-P)^2}{P^2} \mathbb{P}(\mathbf{Y}_{i_1} = \mathbf{Y}_{i_2} = \mathbf{Y}_{i_3} = 1|\mathbf{Z}) - \frac{1-P}{P} \left(\frac{K(1-P)}{P(1-K)} \right) \right.$$
$$\times \mathbb{P}(\mathbf{Y}_{i_1} = 1, \mathbf{Y}_{i_2} \neq \mathbf{Y}_{i_3}|\mathbf{Z})$$
$$+ \left(\frac{K(1-P)}{P(1-K)} \right) \mathbb{P}(\mathbf{Y}_{i_1} = 0, \mathbf{Y}_{i_2} = \mathbf{Y}_{i_3} = 1|\mathbf{Z}) + \left(\frac{K(1-P)}{P(1-K)} \right)^2$$
$$\times \mathbb{P}(\mathbf{Y}_{i_1} = 1, \mathbf{Y}_{i_2} = \mathbf{Y}_{i_3} = 0|\mathbf{Z})$$
$$- \frac{P}{1-P} \left(\frac{K(1-P)}{P(1-K)} \right)^2 \mathbb{P}(\mathbf{Y}_{i_1} = 0, \mathbf{Y}_{i_2} \neq \mathbf{Y}_{i_3}|\mathbf{Z}) + \frac{(1-P)^2}{P^2} \left(\frac{K(1-P)}{P(1-K)} \right)^3$$
$$\left. \times \mathbb{P}(\mathbf{Y}_{i_1} = \mathbf{Y}_{i_2} = \mathbf{Y}_{i_3} = 0|\mathbf{Z}) \right\}$$

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

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

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 $\mathbb{E}[\mathbf{W}_{i_1}^2 \mathbf{W}_{i_2} \mathbf{W}_{i_3}|\mathbf{Z}, \epsilon_{i_1} = \epsilon_{i_2} = \epsilon_{i_3} = 1]$ 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 $\{\epsilon_{i_1} = \epsilon_{i_2} = \epsilon_{i_3} = \epsilon_{i_4} = 1\}$, $\mathbf{W}_{i_1} \mathbf{W}_{i_2} \mathbf{W}_{i_3} \mathbf{W}_{i_4}$ 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.

The matrix of variance covariance 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

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

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

$$\mathbb{P}(\mathbf{Y}_{i_1} = \mathbf{Y}_{i_2} = \mathbf{Y}_{i_3} = \mathbf{Y}_{i_4} = 1 | \mathbf{Z}) = \frac{1}{|\Sigma|^{\frac{1}{2}}} \int_t^{+\infty} \int_t^{+\infty} \int_t^{+\infty} \int_t^{+\infty} f(w,x,y,z) dw dx dy dz,$$

where

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

Finally,

$$
\begin{aligned}
\bullet\ & \mathbb{P}(\mathbf{Y}_{i_1}=\mathbf{Y}_{i_2}=\mathbf{Y}_{i_3}=\mathbf{Y}_{i_4}=1|\mathbf{Z}) \\
& = \frac{1}{|\Sigma|^{\frac{1}{2}}}\Big[K^4+\frac{K^3}{2}(t\phi(t)+K)\frac{\eta^\star}{\sqrt{n}}(A_1+A_2+A_3+A_4) \\
& +K^2\phi(t)^2\frac{\eta^\star}{\sqrt{n}}(C_{1,2}+...+C_{3,4})\Big]
\end{aligned}
$$

Similarly, we compute

$$
\begin{aligned}
\bullet\ & \mathbb{P}(\text{“1 control, 3 cases”}|\mathbf{Z}) \\
& = \frac{1}{|\Sigma|^{\frac{1}{2}}}\Big[4K^3(1-K) \\
& +\frac{K^2}{2}\left((3-4K)t\phi(t)+4K(1-K)\right)\frac{\eta^\star}{\sqrt{N}}(A_1+A_2+A_3+A_4) \\
& +2\phi(t)^2K(1-2K)\frac{\eta^\star}{\sqrt{N}}(C_{1,2}+...+C_{3,4})\Big]
\end{aligned}
$$

- $\mathbb{P}(\text{“2 controls, 2 cases”}|\mathbf{Z})$
$$= \frac{1}{|\Sigma|^{\frac{1}{2}}}\left[6K^2(1-K)^2\right.$$
$$+ \frac{3K(1-K)}{2}\left((1-2K)t\phi(t)+2K(1-K)\right)\frac{\eta^\star}{\sqrt{N}}(A_1+A_2+A_3+A_4)$$
$$\left.+\phi(t)^2(6K^2-6K+1)\frac{\eta^\star}{\sqrt{N}}(C_{1,2}+...+C_{3,4})\right]$$

- $\mathbb{P}(\text{“3 controls, 1 case”}|\mathbf{Z})$
$$= \frac{1}{|\Sigma|^{\frac{1}{2}}}\left[4K(1-K)^3\right.$$
$$+ \frac{(1-K)^2}{2}\left((1-4K)t\phi(t)+4K(1-K)\right)\frac{\eta^\star}{\sqrt{n}}(A_1+A_2+A_3+A_4)$$
$$\left.-2\phi(t)^2(1-K)(1-2K)\frac{\eta^\star}{\sqrt{N}}(C_{1,2}+...+C_{3,4})\right]$$

- $\mathbb{P}(\mathbf{Y}_{i_1}=\mathbf{Y}_{i_2}=\mathbf{Y}_{i_3}=\mathbf{Y}_{i_4}=0|\mathbf{Z})$
$$= \frac{1}{|\Sigma|^{\frac{1}{2}}}\left[(1-K)^4+\frac{(1-K)^3}{2}(-t\phi(t)+1-K)\frac{\eta^\star}{\sqrt{n}}(A_1+A_2+A_3+A_4)\right.$$
$$\left.+(1-K)^2\phi(t)^2\frac{\eta^\star}{\sqrt{N}}(C_{1,2}+...+C_{3,4})\right]$$

Regrouping all the first terms in the expression of $\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}\mathbf{W}_{i_3}\mathbf{W}_{i_4}|\mathbf{Z}]$ gives

$$\frac{1}{|\Sigma|^{\frac{1}{2}}}\left[\frac{(1-P)^2}{P^2}K^4-\frac{(1-P)}{P}\left(\frac{K(1-P)}{P(1-K)}\right)4K^3(1-K)\right.$$
$$+\left(\frac{K(1-P)}{P(1-K)}\right)^2 6K^2(1-K)^2$$
$$\left.-\frac{P}{1-P}\left(\frac{K(1-P)}{P(1-K)}\right)^3 4K(1-K)^3+\frac{P^2}{(1-P)^2}\left(\frac{K(1-P)}{P(1-K)}\right)^4(1-K)^4\right]$$
$$= \frac{1}{|\Sigma|^{\frac{1}{2}}}\left(\frac{(1-P)^2K^4}{P^2}\right)[1-4+6-4+1]=0$$

Similarly we regroup the terms in $\frac{\eta^\star}{\sqrt{N}}(A_1+A_2+A_3+A_4)$:

$$\frac{1}{|\Sigma|^{\frac{1}{2}}}\frac{\eta^\star}{\sqrt{N}}(A_1+A_2+A_3+A_4)\left[\frac{(1-P)^2}{P^2}\frac{K^3}{2}(t\phi(t)+K)\right.$$
$$\left.-\frac{(1-P)}{P}\left(\frac{K(1-P)}{P(1-K)}\right)\frac{K^2}{2}\left((3-4K)t\phi(t)+4K(1-K)\right)\right.$$

$$
\begin{aligned}
&+\left(\frac{K(1-P)}{P(1-K)}\right)^2 \frac{3K(1-K)}{2}\left((1-2K)t\phi(t)+2K(1-K)\right)\\
&-\frac{P}{1-P}\left(\frac{K(1-P)}{P(1-K)}\right)^3 \frac{(1-K)^2}{2}\left((1-4K)t\phi(t)+4K(1-K)\right)\\
&+\frac{P^2}{(1-P)^2}\left(\frac{K(1-P)}{P(1-K)}\right)^4 \frac{(1-K)^3}{2}(-t\phi(t)+1-K)\Bigg]\\
&=\frac{1}{|\Sigma|^{\frac{1}{2}}}\left(\frac{(1-P)^2K^4}{2P^2}\right)[1-4+6-4+1]\\
&+\frac{1}{|\Sigma|^{\frac{1}{2}}}\left(\frac{(1-P)^2K^3}{2P^2(1-K)}\right)[1-K-3+4K+3(1-2K)-1+4K-K]=0
\end{aligned}
$$

Finally, we regroup all the terms in $\frac{\eta^\star}{\sqrt{n}}(C_{1,2}+...+C_{3,4})$:

$$
\begin{aligned}
&\frac{1}{|\Sigma|^{\frac{1}{2}}}\left(\frac{(1-P)^2K^2}{P^2(1-K)^2}\right)\phi(t)^2\\
&\quad\times\left[(1-K)^2-2(1-K)(1-2K)+6K^2-6K+1+2K(1-2K)+K^2\right]=0.
\end{aligned}
$$

This proves Proposition 3.

Let us prove Proposition 5.

The main term of the second order approximation of $f(w,x,y,z)$ can be written as:

$$
\begin{aligned}
f_2(w,x,y,z)=\phi(w)\phi(x)\phi(y)\phi(z)\Bigg[&1+\frac{w^2}{2}\Big(\frac{\eta^\star}{\sqrt{N}}A_1-\frac{\eta^{\star 2}}{N}(A_1^2+C_{1,2}^2+C_{1,3}^2\\
&+C_{1,4}^2)+\cdots+\frac{z^2}{2}\Big(\frac{\eta^\star}{\sqrt{N}}A_4-\frac{\eta^{\star 2}}{N}(A_4^2+C_{1,4}^2+C_{2,4}^2+C_{3,4}^2)\\
&+wx\Big(C_{1,2}\frac{\eta^\star}{\sqrt{N}}-\frac{\eta^{\star 2}}{N}[(A_1+A_2)C_{1,2}+C_{1,3}C_{2,3}+C_{1,4}C_{2,4}]\Big)+\dots\\
&+yz\Big(C_{3,4}\frac{\eta^\star}{\sqrt{N}}-\frac{\eta^{\star 2}}{N}[(A_3+A_4)C_{3,4}+C_{1,3}C_{1,4}+C_{2,3}C_{2,4}]\Big)\\
&+\frac{w^4}{8}\frac{\eta^{\star 2}}{N}A_1^2+\cdots+\frac{z^4}{8}\frac{\eta^{\star 2}}{N}A_4^2+\frac{w^2x^2}{2}\frac{\eta^{\star 2}}{N}\Big(C_{1,2}^2+\frac{A_1A_2}{2}\Big)+\dots\\
&+\frac{y^2z^2}{2}\frac{\eta^{\star 2}}{N}\Big(C_{3,4}^2+\frac{A_3A_4}{2}\Big)+\frac{w^3x}{2}\frac{\eta^{\star 2}}{N}A_1C_{1,2}+\dots\\
&+\frac{z^3y}{2}\frac{\eta^{\star 2}}{N}A_4C_{3,4}+w^2xy\frac{\eta^{\star 2}}{N}\Big[\frac{A_1C_{2,3}}{2}+C_{1,2}C_{1,3}\Big]+\dots\\
&+z^2xy\frac{\eta^{\star 2}}{N}\Big[\frac{A_4C_{2,3}}{2}+C_{2,4}C_{3,4}\Big]+wxyz\frac{\eta^{\star 2}}{N}(C_{1,2}C_{3,4}+C_{2,3}C_{1,4}\\
&+C_{1,3}C_{2,4})\Bigg].
\end{aligned}
\tag{8.25}
$$

In order to prove Proposition 5, we will show that:

$$\frac{1}{n^2}\sum_{i_1\neq i_2\neq i_3\neq i_4}\mathbb{E}[A_1^2C_{1,2}C_{3,4}]\to 0 \tag{8.26}$$

$$\frac{1}{n^2}\sum_{i_1\neq i_2\neq i_3\neq i_4}\mathbb{E}[A_1A_2C_{1,2}C_{3,4}]\to 0 \tag{8.27}$$

$$\frac{1}{n^2}\sum_{i_1\neq i_2\neq i_3\neq i_4}\mathbb{E}[A_1C_{1,2}^2C_{3,4}]\to 0 \tag{8.28}$$

$$\frac{1}{n^2}\sum_{i_1\neq i_2\neq i_3\neq i_4}\mathbb{E}[A_1C_{1,2}C_{13}C_{3,4}]\to 0 \tag{8.29}$$

$$\frac{1}{n^2}\sum_{i_1\neq i_2\neq i_3\neq i_4}\mathbb{E}[C_{1,2}^2C_{2,3}C_{3,4}]\to 0 \tag{8.30}$$

$$\frac{1}{n^2}\sum_{i_1\neq i_2\neq i_3\neq i_4}\mathbb{E}[C_{1,2}C_{1,3}C_{2,4}C_{3,4}]\to 0 \tag{8.31}$$

$$\frac{1}{n^2}\sum_{i_1\neq i_2\neq i_3\neq i_4}\mathbb{E}[C_{1,2}^3C_{3,4}]\to 0 \tag{8.32}$$

We will develop the proof of Equation (8.27).

By exchangeability of the $(\mathbf{Z}_{i,k})_{1\leq i\leq n}$, we can write

$$\mathbb{E}[A_1A_2C_{1,2}C_{3,4}] = \sum_{k,l,m,r}\mathbb{E}[(\mathbf{Z}_{1,k}^2-1)(\mathbf{Z}_{2,l}^2-1)\mathbf{Z}_{1,m}\mathbf{Z}_{2,m}\mathbf{Z}_{3,r}\mathbf{Z}_{4,r}] \tag{8.33}$$

$$= \sum_{k,l,m,r}\mathbb{E}[\mathbf{Z}_{1,k}^2\mathbf{Z}_{2,l}^2\mathbf{Z}_{1,m}\mathbf{Z}_{2,m}\mathbf{Z}_{3,r}\mathbf{Z}_{4,r}]$$

$$-2N\sum_{k,m,r}\mathbb{E}[\mathbf{Z}_{1,k}^2\mathbf{Z}_{1,m}\mathbf{Z}_{2,m}\mathbf{Z}_{3,r}\mathbf{Z}_{4,r}] \tag{8.34}$$

$$+N^2\sum_{m,r}\mathbb{E}[\mathbf{Z}_{1,m}\mathbf{Z}_{2,m}\mathbf{Z}_{3,r}\mathbf{Z}_{4,r}]. \tag{8.35}$$

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 (8.35).

$$\begin{aligned}\sum_{m,r}\mathbb{E}[\mathbf{Z}_{1,m}\mathbf{Z}_{2,m}\mathbf{Z}_{3,r}\mathbf{Z}_{4,r}] &= \sum_{m=1}^{N}\mathbb{E}[\mathbf{Z}_{1,m}\mathbf{Z}_{2,m}\mathbf{Z}_{3,m}\mathbf{Z}_{4,m}]\\ &\quad+\sum_{m\neq r}\mathbb{E}[\mathbf{Z}_{1,m}\mathbf{Z}_{2,m}]\mathbb{E}[\mathbf{Z}_{3,r}\mathbf{Z}_{4,r}]\\ &= N\times o\left(\frac{1}{n}\right)+N(N-1)\times\frac{1}{(n-1)^2}\end{aligned}$$

Then, $\frac{1}{n^2}\frac{1}{N^4}\sum_{i_1\neq i_2\neq i_3\neq i_4}(N^2\sum_{m,r}\mathbb{E}[\mathbf{Z}_{1,m}\mathbf{Z}_{2,m}\mathbf{Z}_{3,r}\mathbf{Z}_{4,r}]) = \frac{(N-1)^2(N-2)(N-3)}{n^2(n-1)^2}+o(1)$

Now let us decompose the second term of (8.35) as:

$$\sum_{k,m,r}\mathbb{E}[\mathbf{Z}_{1,k}^2\mathbf{Z}_{1,m}\mathbf{Z}_{2,m}\mathbf{Z}_{3,r}\mathbf{Z}_{4,r}] = \sum_{k=1}^{N}\mathbb{E}[\mathbf{Z}_{1,k}^3\mathbf{Z}_{2,k}\mathbf{Z}_{3,k}\mathbf{Z}_{4,k}]$$
$$+\sum_{k\neq l}\mathbb{E}[\mathbf{Z}_{1,k}^3\mathbf{Z}_{2,k}]\mathbb{E}[\mathbf{Z}_{3,l}\mathbf{Z}_{4,l}] + \sum_{k\neq l}\mathbb{E}[\mathbf{Z}_{1,k}^2]\mathbb{E}[\mathbf{Z}_{1,l}\mathbf{Z}_{2,l}\mathbf{Z}_{3,l}\mathbf{Z}_{4,l}]$$
$$+\sum_{k\neq l}\mathbb{E}[\mathbf{Z}_{1,k}^2\mathbf{Z}_{3,k}\mathbf{Z}_{3,k}]\mathbb{E}[\mathbf{Z}_{1,l}\mathbf{Z}_{2,l}] + \sum_{k\neq l\neq m}\mathbb{E}[\mathbf{Z}_{1,k}^2]\mathbb{E}[\mathbf{Z}_{1,l}\mathbf{Z}_{2,l}]\mathbb{E}[\mathbf{Z}_{3,m}\mathbf{Z}_{4,m}].$$

Using the results given by Proposition 1, we obtain that

$$\frac{1}{n^2}\frac{1}{N^4}\left(-2N\sum_{k,m,r}\mathbb{E}[\mathbf{Z}_{1,k}^2\mathbf{Z}_{1,m}\mathbf{Z}_{2,m}\mathbf{Z}_{3,r}\mathbf{Z}_{4,r}]\right)$$
$$= -\frac{2(N-1)^2(N-2)(N-3)}{n^2(n-1)^2}+o(1).$$

Similarly, we can prove that

$$\frac{1}{n^2}\frac{1}{N^4}\left(\sum_{k,l,m,r}\mathbb{E}[\mathbf{Z}_{1,k}^2\mathbf{Z}_{2,l}^2\mathbf{Z}_{1,m}\mathbf{Z}_{2,m}\mathbf{Z}_{3,r}\mathbf{Z}_{4,r}]\right)$$
$$= \frac{(N-1)^2(N-2)(N-3)}{n^2(n-1)^2}+o(1),$$

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 (8.27). The proofs of (8.26), (8.28), (8.29), (8.30), (8.31), (8.32) 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 (8.25) and since

$$|\Sigma|^{-\frac{1}{2}} = 1-\frac{\eta^\star}{2\sqrt{N}}(A_1+A_2+A_3+A_4)+\frac{\eta^{\star 2}}{4N}(A_1A_2+\ldots+A_3A_4)$$
$$+\frac{3\eta^{\star 2}}{8N}(A_1^2+A_2^2+A_3^2+A_4^2)+\frac{\eta^{\star 2}}{2N}(C_{1,2}^2+\ldots+C_{3,4}^2)+O_p\left(\frac{1}{N^{\frac{3}{2}}}\right),$$

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

$$\frac{1}{(2\pi)^2}\frac{\eta^{\star 2}}{N}\int_t^{+\infty}\int_t^{+\infty}\int_t^{+\infty}\int_t^{+\infty}wxyzC_{1,2}C_{3,4}dwdxdydz = \phi(t)^4C_{1,2}C_{3,4}\frac{\eta^{\star 2}}{N}.$$

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

$$6\phi(t)^4 C_{1,2}C_{3,4}\frac{\eta^{\star 2}}{N}.$$

The term in $C_{1,2}C_{3,4}$ of $\mathbb{P}(\text{“1 case, 3 controls”}|\mathbf{Z})$ is

$$-4\phi(t)^4 C_{1,2}C_{3,4}\frac{\eta^{\star 2}}{N}.$$

The term in $C_{1,2}C_{3,4}$ of $\mathbb{P}(\mathbf{Y}_{i_1}=\mathbf{Y}_{i_2}=\mathbf{Y}_{i_3}=\mathbf{Y}_{i_4}=0|\mathbf{Z})$ is

$$\phi(t)^4 C_{1,2}C_{3,4}\frac{\eta^{\star 2}}{N}.$$

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

$$\begin{aligned}
&K^4 + 4K^3(1-K)\left(\frac{K(1-P)}{P(1-K)}\right) + 6K^2(1-K)^2\left(\frac{K(1-P)}{P(1-K)}\right)^2 \\
&\quad + 4K(1-K)^3\left(\frac{K(1-P)}{P(1-K)}\right)^3 + (1-K)^4\left(\frac{K(1-P)}{P(1-K)}\right)^4 \\
&\quad = \frac{K^4}{P^4}\left[P^4 + 4P^3(1-P) + 6P^2(1-P)^2 + 4P(1-P)^3 + (1-P)^4\right] \\
&\quad = \frac{K^4}{P^4}.
\end{aligned}$$

Finally, the term $C_{1,2}C_{3,4}$ in $\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}\mathbf{W}_{i_3}\mathbf{W}_{i_4}|\mathbf{Z}]$ 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 $\mathbb{E}[\mathbf{W}_{i_1}\mathbf{W}_{i_2}|\mathbf{Z}]\mathbb{E}[\mathbf{W}_{i_3}\mathbf{W}_{i_4}|\mathbf{Z}]$.

This proves Proposition 4.

8.4. 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

$$f(x,y) = \frac{1}{2\pi|\Sigma^{(N)}|^{-\frac{1}{2}}}\exp\left\{-\frac{1}{2|\Sigma^{(N)}|}\left[x^2(1+\frac{\eta^\star}{\sqrt{N}}A_N(j))\right.\right.$$

$$+ y^2(1 + \frac{\eta^\star}{\sqrt{N}} A_N(i)) - 2xy \frac{B_N(i,j)}{\sqrt{N}} \Big] \Big\},$$

but with the explicit term of order $1/N$ in the expressions of $|\Sigma^{(N)}|^{-1}$ and $|\Sigma^{(N)}|^{-\frac{1}{2}}$:

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

and

$$\begin{aligned}
|\Sigma^{(N)}|^{-\frac{1}{2}} &= 1 - \frac{\eta^\star}{2\sqrt{N}}(A_N(i) + A_N(j)) \\
&\quad + \frac{\eta^{\star 2}}{2N}\left(-A_N(i)A_N(j) + B_N(i,j)^2 + \frac{3}{8}(A_N(i) + A_N(j))^2\right) \\
&\quad + O_p\left(\frac{1}{N^{\frac{3}{2}}}\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\exp\left\{-\frac{1}{2|\Sigma^{(N)}|}\left[x^2(1 + \frac{\eta^\star}{\sqrt{N}}A_N(j)) + y^2(1 + \frac{\eta^\star}{\sqrt{N}}A_N(i)) - 2xy\frac{B_N(i,j)}{\sqrt{N}}\right]\right\} \\
&= \phi(x)\phi(y)\exp\left\{-\frac{x^2}{2}(-A_N(i)\frac{\eta^\star}{\sqrt{N}} + \frac{\eta^{\star 2}}{N}(A_N(i)^2 + B_N(i,j)^2)\right. \\
&- \frac{y^2}{2}(-A_N(j)\frac{\eta^\star}{\sqrt{N}} + \frac{\eta^{\star 2}}{N}(A_N(j)^2 + B_N(i,j)^2)) + xy(\frac{\eta^\star}{\sqrt{N}}B_N(i,j) \\
&\left. - \frac{\eta^{\star 2}}{N}B_N(i,j)(A_N(i) + A_N(j)))\right\} + O_p\left(\frac{1}{N^{\frac{3}{2}}}\right) \\
&= \phi(x)\phi(y)\left[1 + \frac{x^2}{2}(A_N(i)\frac{\eta^\star}{\sqrt{N}} - \frac{\eta^{\star 2}}{N}(A_N(i)^2 + B_N(i,j)^2)\right. \\
&+ \frac{y^2}{2}(A_N(j)\frac{\eta^\star}{\sqrt{N}} - \frac{\eta^{\star 2}}{N}(A_N(j)^2 + B_N(i,j)^2)) + \frac{x^4}{8}\frac{\eta^{\star 2}}{N}A_N(i)^2 \\
&+ \frac{y^4}{8}\frac{\eta^{\star 2}}{N}A_N(j)^2 + xy(\frac{\eta^\star}{\sqrt{N}}B_N(i,j) - \frac{\eta^{\star 2}}{N}B_N(i,j)(A_N(i) + A_N(j))) \\
&\left. + \frac{x^2y^2}{2}\frac{\eta^{\star 2}}{N}B_N(i,j)^2 + O_p\left(\frac{1}{N^{\frac{3}{2}}}\right)\right]
\end{aligned}$$

with the last term obtained by developing the exponential function.

Since

$$\int_t^\infty \int_t^\infty x^4 dxdy = t^3\phi(t) + 3t\phi(t) + 3K$$

and

$$\int_t^\infty \int_t^\infty x^2y^2 dxdy = (t\phi(t) + K)^2,$$

we have:

$$\begin{aligned}
&\int_t^\infty \int_t^\infty \exp\Big\{ -\frac{1}{2|\Sigma^{(N)}|}\Big[x^2(1+\frac{\eta^\star}{\sqrt{N}}A_N(j)) + y^2(1+\frac{\eta^\star}{\sqrt{N}}A_N(i)) \\
&\quad - 2xy\frac{B_N(i,j)}{\sqrt{N}}\Big]\Big\}dxdy = K^2 + \frac{K}{2}(t\phi(t)+K)\Big[\frac{\eta^\star}{\sqrt{N}}(A_N(i)+A_N(j)) \\
&\quad - \frac{\eta^{\star 2}}{N}(A_N(i)^2 + A_N(j)^2 + 2B_N(i,j)^2)\Big] \\
&\quad + \frac{K}{8}\frac{\eta^{\star 2}}{N}(t^3\phi(t) + 3t\phi(t) + 3K)(A_N(i)^2 + A_N(j)^2) \\
&\quad + \phi(t)^2\Big[\frac{\eta^\star}{\sqrt{N}}B_N(i,j) - \frac{\eta^{\star 2}}{N}B_N(i,j)(A_N(i)+A_N(j))\Big] \\
&\quad + \frac{1}{2}(t\phi(t)+K)^2\frac{\eta^{\star 2}}{N}(B_N(i,j)^2 + \frac{A_N(i)A_N(j)}{2} \\
&\quad + \frac{\phi(t)^2}{2}\frac{\eta^{\star 2}}{N}(t^2+2)B_N(i,j)(A_N(i)+A_N(j)) + O_p\Big(\frac{1}{N^{\frac{3}{2}}}\Big)
\end{aligned}$$

Multiplying by

$$\begin{aligned}
|\Sigma^{(N)}|^{-\frac{1}{2}} = 1 - \frac{\eta^\star}{2\sqrt{N}}(A_N(i)+A_N(j)) + \frac{\eta^{\star 2}}{2N}\big(-A_N(i)A_N(j) + B_N(i,j)^2 \\
+\frac{3}{4}(A_N(i)+A_N(j))^2\Big) + O_p\left(\frac{1}{N^{\frac{3}{2}}}\right),
\end{aligned}$$

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

$$\int_{-\infty}^{t}\int_{-\infty}^{t} x^2y^2dxdy = (-t\phi(t)+1-K)^2.$$
$$\int_{-\infty}^{t}\int_{-\infty}^{t} \exp\Big\{ -\frac{1}{2|\Sigma^{(N)}|}\Big[x^2(1+\frac{\eta^\star}{\sqrt{N}}A_N(j))+y^2(1+\frac{\eta^\star}{\sqrt{N}}A_N(i))$$
$$-2xy\frac{B_N(i,j)}{\sqrt{N}}\Big]\Big\}dxdy = (1-K)^2+\frac{1-K}{2}(-t\phi(t)+1-K)\Big[\frac{\eta^\star}{\sqrt{N}}(A_N(i)$$
$$+A_N(j))-\frac{\eta^{\star 2}}{N}(A_N(i)^2+A_N(j)^2+2B_N(i,j)^2)\Big]$$
$$+\frac{1-K}{8}\frac{\eta^{\star 2}}{N}\Big(-t^3\phi(t)-3t\phi(t)+3(1-K)\Big)(A_N(i)^2+A_N(j)^2)$$
$$+\frac{\phi(t)^2}{2}\frac{\eta^{\star 2}}{N}(t^2+2)B_N(i,j)(A_N(i)+A_N(j))$$

Multiplying by

$$|\Sigma^{(N)}|^{-\frac{1}{2}} = 1-\frac{\eta^\star}{\sqrt{N}}(A_N(i)+A_N(j))+\frac{\eta^{\star 2}}{2N}\Big(-A_N(i)A_N(j)+B_N(i,j)^2$$
$$+\frac{3}{4}(A_N(i)+A_N(j))^2\Big)+O_p\Big(\frac{1}{N^{\frac{3}{2}}}\Big),$$
$$\int_{-\infty}^{t}\int_{-\infty}^{t} f(x,y)dxdy = (1-K)^2-\frac{1-K}{2}t\phi(t)\frac{\eta^\star}{\sqrt{N}}(A_N(i)+A_N(j))$$
$$+\phi(t)^2\frac{\eta^\star}{\sqrt{N}}B_N(i,j)+\frac{1-K}{8}\frac{\eta^{\star 2}}{N}(A_N(i)^2+A_N(j)^2)(-t^3\phi(t)+3t\phi(t))$$
$$+\frac{\eta^{\star 2}}{N}\frac{t^2\phi(t)^2}{4}A_N(i)A_N(j)+\frac{\eta^{\star 2}}{N}B_N(i,j)^2\frac{t^2}{2}\phi(t)^2$$
$$-\frac{\eta^{\star 2}}{N}B_N(i,j)(A_N(i)+A_N(j))\frac{\phi(t)^2}{2}(t^2-1)+O_p\Big(\frac{1}{N^{\frac{3}{2}}}\Big).$$

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:

$$\begin{aligned}
&\frac{\eta^\star}{\sqrt{N}}\frac{1-P}{P(1-K)^2}\phi(t)^2 B_N(i,j) \\
&+ \frac{\eta^{\star 2}}{N}\frac{t^2\phi(t)^2}{4}A_N(i)A_N(j)\frac{1-P}{P(1-K)^2} + \frac{\eta^{\star 2}}{2N}B_N(i,j)^2\frac{1-P}{P(1-K)^2}t^2\phi(t)^2 \\
&+ \frac{\eta^{\star 2}}{N}\frac{\phi(t)^2}{2}\frac{1-P}{P(1-K)^2}(t^2-1)B_N(i,j)(A_N(i)+A_N(j)) + O_p\left(\frac{1}{N^{\frac{3}{2}}}\right).
\end{aligned}$$

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:

$$\begin{aligned}
&\frac{K^2}{P^2} + \frac{\eta^\star}{\sqrt{N}}\frac{t}{2}\phi(t)(A_N(i)+A_N(j))\frac{K(P-K)}{P^2(1-K)} + \frac{\eta^\star}{\sqrt{N}}\phi(t)^2 B_N(i,j)\frac{(P-K)^2}{P^2(1-K)^2} \\
&\quad + O_p\left(\frac{1}{N}\right).
\end{aligned}$$

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

Appendix A: Appendix

A.1. Proof of Equation (2.7)

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

$$\begin{aligned} 1 - P = \mathbb{P}(\mathbf{Y}_i = 0 | \epsilon_i = 1) &= \frac{\mathbb{P}(\mathbf{Y}_i = 0, \epsilon_i = 1)}{\mathbb{P}(\epsilon_i = 1)} = \frac{\mathbb{P}(\mathbf{Y}_i = 0, U_i = 1)}{\mathbb{P}(\epsilon_i = 1)} \\ &= \frac{(1 - K) p_{control}}{\mathbb{P}(\epsilon_i = 1)}, \end{aligned}$$

with the full ascertainment assumption given by (2.6) prove equation (2.7).

A.2. Proof of Equation (3.2)

This equation was proved in Golan, Lander and Rosset (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 $\{\epsilon_i = \epsilon_j = 1\}$:

$$\begin{aligned} \mathbb{E}[\mathbf{W}_i \mathbf{W}_j | \mathbf{Z}, \epsilon_i = \epsilon_j = 1] = \frac{1 - P}{P} &\mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 1 | \mathbf{Z}, \epsilon_i = \epsilon_j = 1) \\ &- \mathbb{P}(\mathbf{Y}_i \neq \mathbf{Y}_j | \mathbf{Z}, \epsilon_i = \epsilon_j = 1) \\ &+ \frac{P}{1 - P} \mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 0 | \mathbf{Z}, \epsilon_i = \epsilon_j = 1). \end{aligned} \tag{A.1}$$

$$\begin{aligned} &\mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 1 | \mathbf{Z}, \epsilon_i = \epsilon_j = 1) \\ &\quad = \frac{\mathbb{P}(\epsilon_i = \epsilon_j = 1 | \mathbf{Y}_i = \mathbf{Y}_j = 1, \mathbf{Z}) \mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 1 | \mathbf{Z})}{\mathbb{P}(\epsilon_i = \epsilon_j = 1 | \mathbf{Z})} \\ &\quad = \frac{\mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 1 | \mathbf{Z})}{\mathbb{P}(\epsilon_i = \epsilon_j = 1 | \mathbf{Z})} \end{aligned}$$

under the full ascertainment assumption given by Equation (2.6).

Similarly, since we have seen in Equation (2.7) 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 $\mathbf{Y}_j$:

$$\begin{aligned}
&\mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 0 | \mathbf{Z}, \epsilon_i = \epsilon_j = 1) \\
&\quad = \frac{\mathbb{P}(\epsilon_i = \epsilon_j = 1 | \mathbf{Y}_i = \mathbf{Y}_j = 0, \mathbf{Z}) \mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 0 | \mathbf{Z})}{\mathbb{P}(\epsilon_i = \epsilon_j = 1 | \mathbf{Z})} \\
&\quad = \left(\frac{K(1-P)}{P(1-K)} \right)^2 \frac{\mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 1 | \mathbf{Z})}{\mathbb{P}(\epsilon_i = \epsilon_j = 1 | \mathbf{Z})}
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{P}(\mathbf{Y}_i \neq \mathbf{Y}_j | \mathbf{Z}, \epsilon_i = \epsilon_j = 1) &= \frac{\mathbb{P}(\epsilon_i = \epsilon_j = 1 | \mathbf{Y}_i \neq \mathbf{Y}_j, \mathbf{Z}) \mathbb{P}(\mathbf{Y}_i \neq \mathbf{Y}_j | \mathbf{Z})}{\mathbb{P}(\epsilon_i = \epsilon_j = 1 | \mathbf{Z})} \\
&= \left(\frac{K(1-P)}{P(1-K)} \right) \frac{\mathbb{P}(\mathbf{Y}_i \neq \mathbf{Y}_j | \mathbf{Z})}{\mathbb{P}(\epsilon_i = \epsilon_j = 1 | \mathbf{Z})}.
\end{aligned}$$

The probability that both individuals i and j are included in the study is equal to

$$\begin{aligned}
\mathbb{P}(\epsilon_i = \epsilon_j = 1 | \mathbf{Z}) &= \mathbb{P}(\epsilon_i = \epsilon_j = 1 | \mathbf{Z}, \mathbf{Y}_i = \mathbf{Y}_j = 1) \mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 1 | \mathbf{Z}) \\
&\quad + \mathbb{P}(\epsilon_i = \epsilon_j = 1 | \mathbf{Z}, \mathbf{Y}_i = \mathbf{Y}_j = 0) \mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 0 | \mathbf{Z}) \\
&\quad + \mathbb{P}(\epsilon_i = \epsilon_j = 1 | \mathbf{Z}, \mathbf{Y}_i \neq \mathbf{Y}_j) \mathbb{P}(\mathbf{Y}_i \neq \mathbf{Y}_j | \mathbf{Z}) \\
&\quad = \mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 1 | \mathbf{Z}) + \left(\frac{K(1-P)}{P(1-K)} \right)^2 \mathbb{P}(\mathbf{Y}_i = \mathbf{Y}_j = 0 | \mathbf{Z}) \\
&\quad + \left(\frac{K(1-P)}{P(1-K)} \right) \mathbb{P}(\mathbf{Y}_i \neq \mathbf{Y}_j | \mathbf{Z}).
\end{aligned}$$

If we combine all these computations and we plug them in the expression (A.1), we obtain (3.2).

A.3. Proof of Equation (3.6)

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

$$\mathrm{Var} \left(\frac{1}{\sqrt{N}} \sum_{k=1}^{N} (\mathbf{Z}_{i,k}^2 - 1) \right) = \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}[\mathbf{Z}_{i,k}^4] - (\mathbb{E}[\mathbf{Z}_{i,k}^2])^2.$$

Moreover, since the variables $(\mathbf{Z}_{i,k})_{1 \leq i \leq n}$ are normalized according to Equation (2.3),

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

By taking the expectation and since the variables $(\mathbf{Z}_{i,k})_{1\leq i\leq n}$ are exchangeable, we obtain that

$$\mathbb{E}[\mathbf{Z}_{i,k}^2] = 1. \tag{A.2}$$

Using (2) of Proposition 1 and Equation (A.2), 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,

$$\begin{aligned}\mathrm{Var}\left(\frac{1}{\sqrt{N}}\sum_{k=1}^{N}\mathbf{Z}_{i,k}\mathbf{Z}_{j,k}\right) &= \frac{1}{N}\sum_{k=1}^{N}\mathbb{E}(\mathbf{Z}_{i,k}^2\mathbf{Z}_{j,k}^2) - \mathbb{E}(\mathbf{Z}_{i,k}\mathbf{Z}_{j,k})^2 \\ &= \frac{1}{N}\sum_{k=1}^{N}\left(1 + o(1) - \frac{1}{(n-1)^2}\right) \\ &\quad \text{using (3) and (1) of Proposition 1} \\ &= 1 + o(1).\end{aligned}$$

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 $i \neq j$.

A.4. Proof of Proposition 1

Observe that for all $k = 1, \ldots, N$,

$$\sum_{i=1}^{n}\mathbf{Z}_{i,k} = 0 \tag{A.3}$$

and

$$\sum_{i=1}^{n}\mathbf{Z}_{i,k}^2 = n. \tag{A.4}$$

Moreover, for each k, the random variables $(\mathbf{Z}_{i,k})_{1\le i\le n}$ are exchangeable. Thus, we deduce from (A.4) that for all $i=1,\dots,n$ and $k=1,\dots,N$, $\mathbb{E}(\mathbf{Z}_{i,k}^2)=1$. Hence, by (A.3), we get that

$$0=\left(\sum_{i=1}^{n}\mathbf{Z}_{i,k}\right)^2=\sum_{i=1}^{n}\mathbf{Z}_{i,k}^2+\sum_{1\le i\ne j\le n}\mathbf{Z}_{i,k}\mathbf{Z}_{j,k}\ ,$$

which, by (A.4), implies that for all $k=1,\dots,N$ and $i\ne j=1,\dots,n$,

$$\mathbb{E}(\mathbf{Z}_{i,k}\mathbf{Z}_{j,k})=-\frac{n}{n(n-1)}=-\frac{1}{n-1}\ , \tag{A.5}$$

that is (1).

The proof of (2) comes from the decomposition:

$$\begin{aligned}|\mathbf{Z}_{1,k}|^p&=|\mathbf{Z}_{1,k}|^p\mathbb{1}_{\{s_k^2>\frac{\delta_{min}}{2}\}}+|\mathbf{Z}_{1,k}|^p\mathbb{1}_{\{s_k^2\le\frac{\delta_{min}}{2}\}}\\&\le\frac{|A_{1,k}-\bar{A}_k|^p}{\left(\frac{\delta_{min}}{2}\right)^p}+n^p\mathbb{1}_{\{s_k^2\le\frac{\delta_{min}}{2}\}}\end{aligned}$$

Assumption **1.2** implies that $\sup_k \mathbb{E}\left[|A_{1,k}-\bar{A}_k|^p\right]<+\infty$ and the upper bound for $\mathbb{P}(s_k^2\le\delta)$ of Equation (8.15) prove (2).

By (A.4), for all $k=1,\dots,N$,

$$n^2=\left(\sum_{i=1}^{n}\mathbf{Z}_{i,k}^2\right)^2=\sum_{i=1}^{n}\mathbf{Z}_{i,k}^4+\sum_{1\le i\ne j\le n}\mathbf{Z}_{i,k}^2\mathbf{Z}_{j,k}^2$$

Since the $(\mathbf{Z}_{i,k})_{1\le i\le n}$ are exchangeable for each $k=1,\dots,N$, we get that for all $k=1,\dots,N$,

$$n=\mathbb{E}[\mathbf{Z}_{1,k}^4]+(n-1)\mathbb{E}[\mathbf{Z}_{1,k}^2\mathbf{Z}_{2,k}^2]\ ,$$

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

$$\mathbb{E}[\mathbf{Z}_{1,k}^6]+(n-1)\mathbb{E}[\mathbf{Z}_{1,k}^5\mathbf{Z}_{2,k}]=0.$$

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).

Acknowledgments

The author is very grateful to Elisabeth Gassiat for her insightful comments and valuable discussions.

References

Bonnet, A., Gassiat, E. and Levy-Leduc, C. (2015). Heritability estimation in high-dimensional sparse linear mixed models. *Electronic Journal of Statistics* **9** 2099–2129. MR3400534

Breslow, N. E. and Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association* **88** 9-25. MR1397972

de Villemereuil, P., Gimenez, O. and Doligez, B. (2013). 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** 260–275.

Falconer, D. S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics* **29** 51–76.

Golan, D., Lander, E. S. and Rosset, S. (2014). Measuring missing heritability: Inferring the contribution of common variants. *Proceedings of the National Academy of Sciences* **111** E5272–E5281.

Hadfield, J. D. (2010). MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software* **33** 1–22.

Jiang, J., Li, C., Paul, D., Yang, C. and Zhao, H. (2016). On high-dimensional misspecified mixed model analysis in genome-wide association study. *Ann. Statist.* **44** 2127–2160. MR3546446

Lee, S. H., Wray, N. R., Goddard, M. E. and Visscher, P. M. (2011). Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics* **88** 294–305.

Patterson, N., Price, A. L. and Reich, D. (2006). Population Structure and Eigenanalysis. *PLoS Genetics* **2** 997–1004.

Patterson, H. D. and Thompson, R. (1971). Recovery of Inter-block Information when Block Sizes are Unequal. *Biometrika* **58** 545-554. MR0319325

Pirinen, M., Donnelly, P. and Spencer, C. C. A. (2013). Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *The Annals of Applied Statistics* **7** 369–390. MR3086423

Purcell, S., Wray, N., Stone, J., Visscher, P., O'Donovan, M., Sullivan, P., Sklar, P., International Schizophrenia Consortium and Pickard, B. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460** 748–752.

Searle, S. R., Casella, G. and McCulloch, C. E. (1992). *Variance Components. Wiley Series in Probability and Statistics*. Wiley, New Jersey. MR1190470

Speed, D., Hemani, G., Johnson, M. and Balding, D. (2012). Improved Heritability Estimation from Genome-wide SNPs. *The American Journal of Human Genetics* **91** 1011–1021.

Tenesa, A. and Haley, C. (2013). The heritability of human disease: estimation, uses and abuses. *Nature Reviews Genetics* **14** 139–49.

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

Yang, J., Lee, S. H., Goddard, M. E. and Visscher, P. M. (2011). GCTA: A Tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics* **88** 76–82.

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