Contents

Chapter 1

Manuscript

Title: Statistical power in GWAS revisited: sample size, genetic relatedness, and gene-by-environment interactions

Abstract: Genome-wide association studies (GWAS) have identified thousands of genetic variants associated with complex diseases and heavily rely on increasing the sample size. Recent analyses of biobank-scale genetic data suggest: (i) inclusion of genetically related individuals empowers GWAS [**?**]; (ii) the wealth of collected environmental exposures has potential to uncover gene-by-environmental interactions [**?**]. However, quantification of GWAS power — the non-centrality parameter (NCP) of association test, which is proportional to the sample size (n) and the variance explained by genetic variant (q^2) — holds only for unrelated individuals. Here, we first expanded it by incorporating individual relationships by linear mixed model. We next studied gene-by-environment interactions, where interaction effect on trait is tested in the presence of marginal genetic effect. In result, the derived formulas have a range of implications. For testing marginal genetic effect, one can quickly assess the power in studies involving related individuals. Because of the potential gain in power for testing gene-by-environment interaction, the formula for interactions will allow optimization of the study design of related individuals.

1.1 Introduction

1.2 Methods and Analytical Derivations

1.2.1 Linear mixed model for genetic association study

We consider the following linear mixed model to study the impact of relatedness among individuals on modeling a continuous phenotype y.

$$y = X\beta + \sum_{k=1}^{m} r_k + e \tag{1.1}$$

where n is the number of individuals, p is the number of covariates or fixed effects, m is the number of structured random effects apart from the residuals errors, y is a phenotype vector of length n, X is a matrix of covariates of size $n \times p$, β is a vector of fixed effects of length p. The vectors of random effects r_k and e are mutually uncorrelated and multivariate normally distributed as $\mathcal{N}(0, \sigma_k^2 R_k)$ and $\mathcal{N}(0, \sigma_r^2 I)$. The variance-covariance matrices are parametrized with scalar parameters (referred as variance components) and constant matrices of size $n \times n$ that express relationships among n individuals. The first m random effects r_k are referred here as structured, whereas the last component e is simply the residual errors which are independent and identically distributed.

Thus, the phenotype follows a multivariate normal distribution (MVN) and Equation (1.1) can be rewritten:

$$y \sim \mathcal{N}(X\beta, V) = \mathcal{N}(X\beta, \sum_{k=1}^{m} \sigma_k^2 R_k + \sigma_r^2 I) \tag{1.2}$$

An assosiation test for a given variable in the matrix X consists in constructing the score test statistic based on the estimates of effect size and its variance, $Z = \hat{\beta}_x / \sqrt{var(\hat{\beta}_x)}$. The score follows the standard normal distribution $Z \sim \mathcal{N}(0,1)$, and the χ^2 test with non-centrality parameter $NCP = Z^2 = \hat{\beta^2}_x / var(\hat{\beta}_x)$ quantifies the statistical power.

We further consider several parameterizations of the model in Equation (1.2) that depend on (i) whether marginal genetic or gene-environment interaction effect is under testing; (ii) whether structured random effects are included or only the residual errors are present. The detailed derivation of the formulas presented next is given in Supplementary Material, Section 2.2.

We introduce common assumptions and notations before going further. We assume that all vectors of the phenotype (y) and covariates (columns in the matrix X) are centered. The phenotype vector is additionally standardized ($var(y) = 1$). The genotype vector x_g is considered as a realization of a vector of random variables $\mathcal{X}_g$, which is a genotype in n unrelated individuals with a minor allele frequency p. We denote the distribution $\mathcal{X}_g \sim (\mu_g, \Sigma_g) = (p1_n, \delta_g^2 K) = (p1_n, 2p(1-p)K)$, where K is the kinship matrix of size $n \times n$ (it can be the identity matrix I for genetically unrelated individuals) and 1_n is a vector of n ones.

1.2.2 Testing marginal genetic effect

The genetic effect on phenotype in unrelated individuals is evaluated under the standard linear model $y \sim \mathcal{N}(\mu x_0 + \beta_g x_g, \sigma_r^2 I)$, where $x_0 = 1_n$ is a vector of n ones, μ is a mean of the phenotypic values, x_g is a vector of length of the genotypic values, β_g is the effect size of the genotype. The NCP parameter of the test is well known to be proportional to the sample size and the variance captured by the genotype (see also Section 2.2.1).

$$NCP_{unrel} \approx \hat{\beta}_g^2 \delta_g^2 n = \hat{\beta}_g^2 2p(1-p)n \tag{1.3}$$

When the individuals are genetically related and/or the covariance of the phenotype is modeled using structured relationship matrices among individuals, the following linear mixed model is stated, $y \sim \mathcal{N}(\mu x_0 + \beta_g x_g, \sum_{k=1}^{m} \sigma_k^2 R_k + \sigma_r^2 I)$. The initial step in solving a linear mixed model is to estimate random effects parameters (σ_k^2 and σ_r^2) by restricted maximum likelihood (REML) or other optimization technique [**?**]. Once the estimate of the variance-covariance matrix is found, $\hat{V} = \sum \hat{\sigma}_i^2 R_i + \hat{\sigma}_r^2 I$, the generalized least squares (GLS) for fixed effects are applied in the following matrix form, $\hat{\beta} = (X^T \hat{V}^{-1} X)^{-1} X \hat{V}^{-1} y$.

In Section 2.2.2 we derived the estimate of β_g and its variance, $\hat{\beta}_g = (\tilde{x}_g^T \hat{V}^{-1} \tilde{x}_g)^{-1} \tilde{x}_g^T \hat{V}^{-1} \tilde{y}$ and $var(\hat{\beta_g}) = 1/(\tilde{x}_g^T \hat{V}^{-1} \tilde{x}_g)$, respectively. We further approximated the term $\tilde{x}_g^T \hat{V}^{-1} \tilde{x}_g$ using the expected mean of a quadratic form of the random variable $\tilde{\mathcal{X}}_g$ and the transformation matrix $\hat{V}^{-1}$ (see Equation (2.1)). The NCP parameter of the test has the following form, where tr denote the trace operator.

$$NCP_{rel} \approx \hat{\beta}_g^2 tr(\hat{V}^{-1} \Sigma_g) = \hat{\beta}_g^2 \delta_g^2 tr(\hat{V}^{-1} K) = \hat{\beta}_g^2 2p(1-p) tr(\hat{V}^{-1} K) \tag{1.4}$$

The effective size multiplier, defined as $NCP_{rel} / NCP_{unrel} = tr(\hat{V}^{-1} K)/n$, gives a quantitative assessment of gain or loss in power when comparing the study design of related and unrelated individuals.

We further expand Equation (1.4) for two specific cases of (i) related individuals in families; (ii) unrelated individuals under the the infinite-testimal model. We also make use of the connection between the trace operator and eigen-value decomposition (Section 2.1).

For individuals in families and the model $y \sim \mathcal{N}(\mu x_0 + \beta_g x_g, \sigma_k^2 K + \sigma_r^2 I))$, we have an updated formula of NCP_{ref}.

$$\begin{aligned} NCP_{fam} &= \hat{\beta}_g^2 2p(1-p) tr((\hat{\sigma}_k^2 K + \hat{\sigma}_r^2 I)^{-1} K) \\ &= \hat{\beta}_g^2 2p(1-p) tr((\hat{\sigma}_k^2 I + \hat{\sigma}_r^2 K^{-1})^{-1}) \\ &= \hat{\beta}_g^2 2p(1-p) \sum_{i=1}^{n} (\hat{\sigma}_k^2 + \hat{\sigma}_r^2 (\lambda_K^i)^{-1})^{-1} \end{aligned} \tag{1.5}$$

When modeling the polygenic effect in unrelated individuals using the genetic relationship matrix (GRM) (denoted as M in equations) $y \sim \mathcal{N}(\mu x_0 + \beta_g x_g, \sigma_m^2 M + \sigma_r^2 I))$, we rewrite NCP_{ref} as following.

$$\begin{aligned} NCP_{unrel+grm} &= \hat{\beta}_g^2 2p(1-p) tr((\hat{\sigma}_k^2 M + \hat{\sigma}_r^2 I)^{-1}) \\ &= \hat{\beta}_g^2 2p(1-p) \sum_{i=1}^{n} (\hat{\sigma}_m^2 \lambda_M^i + \hat{\sigma}_r^2)^{-1} \end{aligned} \tag{1.6}$$

1.2.3 Testing gene-environment interaction effect

The gene-environment interaction effect on phenotype in unrelated individuals is evaluated under the standard linear model $y \sim \mathcal{N}(\mu x_0 + \beta_g x_g + \beta_e x_e + \beta_{ge} x_{ge}, \sigma_r^2 I)$, where $x_0 = 1_n$ is a vector of n ones, μ is a mean of the phenotypic values, x_g is a vector of length of the genotypic values, β_g is the effect size of the genotype, x_e is a environment exposure vector of length n, β_e is the effect size of the environment exposure, x_{ge} is a vector of length n of gene-environment interaction, β_{ge} is the interaction effect size.

The coding scheme of the genotypic and environmental variables to study gene-environment interaction under linear model is important and has been reviewed elsewhere **[?]**. Here, we work with centered variables $\tilde{x}_g$ and $\tilde{x}_e$ and define the interaction variable $\tilde{x}_{ge}$ by (i) element-wise multiplication of the two variables, (ii) centering the resulted product. Once the covariates are centered as describe above, the effect sizes and their standard errors can be estimated independently from other covariates if we assume that the two random variables of genotype and environmental exposure are generated independently **[?**, Appendix C]. We also note that different coding schemes give different estimates of effect sizes, but the test statistic for gene-environment interaction (NCP) is the same **[?**, Appendix B].

We first need to introduce a matrix E related to the environment exposure (centered) vector $\tilde{x}_e$: E is the diagonal matrix with values equal to those observed in $\tilde{x}_e$, i.e. $diag(E) = \tilde{x}_e$. When the environmental exposure is binary and the observed frequency of exposure is f, then we denote the matrix as E_b. Then the values on diagonal of the matrix E_b are either $-f$ or $1-f$.

Then the NCP parameter of the test has the following form (see Section 2.2.5).

$$NCP_{unrel}^i \approx \hat{\beta}_{ge}^2 \delta_g^2 tr(E_b^2) = \hat{\beta}_{ge}^2 2p(1-p) f(1-f) n \tag{1.7}$$

We again use the linear mixed model $y \sim \mathcal{N}(\mu x_0 + \beta_g x_g + \beta_e x_e + \beta_{ge} x_{ge}, \sum_{k=1}^{m} \sigma_k^2 R_k + \sigma_r^2 I)$, when the individuals are genetically related and/or the covariance of the phenotype exists.

In addition to the matrix E defined previously, we introduce a matrix D, which value at row i and column j is equal to the product of two diagonal entries i and j of E, i.e. $D_{i,j} = E_{i,i} E_{j,j}$. When the environmental exposure is binary, we denote the matrix D as D_b with the values equal to either f^2, $(1-f)^2$ or $f(1-f)$.

In Section 2.2.5 we derived the estimate of β_{ge} and its variance for linear mixed model, $\hat{\beta}_{ge} = (\tilde{x}_{ge}^T \hat{V}^{-1} \tilde{x}_{ge})^{-1} \tilde{x}_{ge}^T \hat{V}^{-1} \tilde{y}$ and $var(\hat{\beta_{ge}}) = 1/(\tilde{x}_{ge}^T \hat{V}^{-1} \tilde{x}_{ge})$, respectively. As $\tilde{x}_{ge}$ is a realization of a random variable $\tilde{\mathcal{X}}_{ge} = E\tilde{\mathcal{X}}_g$, we showed that $var(\tilde{\mathcal{X}}_{ge}) = \delta_g^2 D \cdot K$.

Here, we introduce a special kinship matrix K_D "masked" by the (observed) environmental exposure though the matrix D (the operator $\cdot$ denotes the Hadamard product, i.e. the element-wise multiplication).

$$K_D = D \cdot K \tag{1.8}$$

In Section 2.2.5 we further approximated the term $\tilde{x}_{ge}^T \hat{V}^{-1} \tilde{x}_{ge}$ by applying the expression for the mean of a quadratic form of the random variable $\tilde{\mathcal{X}}_{\}]}$ and the transformation matrix $\hat{V}^{-1}$ (see Equation (2.1)). The NCP parameter of the test has the following form.

$$NCP_{rel}^{i} \approx \hat{\beta}_{ge}^2 \delta_g^2 tr(\hat{V}^{-1} K_D) = \hat{\beta}_{ge}^2 2p(1-p) tr(\hat{V}^{-1} K_D) \tag{1.9}$$

The K_D matrix is equal to the E^2 matrix in the case of genetically unrelated individuals ($K = I$), and the two formulas (1.9) and (1.7) become the same. We also note that the variance of the environmental exposure is contained within the matrices K_D and E^2, although it is possible to similarly define the scaled matrices.

TODO: $V = \sigma_k^2 K + \sigma_i^2 K_i + \sigma_r^2 I$ **[?]**

1.3 Results

1.3.1 Analytical results for testing marginal genetic effect

Table 1.1: Analytical comparison of study designs to detect marginal genetic association. Study designs differ in individual relationships that informs modeling of outcome (y) and distribution of genotype under association test (x_g). Study designs under comparison include: unrelated individuals; related individuals in families; unrelated individuals with a grouping factor such as house-hold (not related to a variable under test). Notation: $\tilde{x_g}$, mean-centered genotype vector x_g; $\delta_g^2 = 2p(1-p)$, the variance of genotype random variable with the minor allele frequency p; K, the additive kinship matrix for family-based study design; NCP, the non-centrality parameter of the test; $\hat{V}$, the estimated variance-covariance matrix of y.

Study design	$V = Var(y)$	$\Sigma_g = Var(x_g)$	NCP
Unrelated	$\sigma_r^2 I$	$\delta_g^2 I$	$\hat{\beta}_g^2 (\tilde{x_g}^T \tilde{x_g}) \approx \hat{\beta}_g^2 \delta_g^2 n$
Families	$\sigma_k^2 K + \sigma_r^2 I$	$\delta_g^2 K$	$\hat{\beta}_g^2 (\tilde{x_g}^T \hat{V}^{-1} \tilde{x_g}) \approx \hat{\beta}_g^2 \delta_g^2 \; tr(\hat{V}^{-1} K)$
Unrelated + Grouping	$\sigma_h^2 H + \sigma_r^2 I$	$\delta_g^2 I$	$\hat{\beta}_g^2 (\tilde{x_g}^T \hat{V}^{-1} \tilde{x_g}) \approx \hat{\beta}_g^2 \delta_g^2 \; tr(\hat{V}^{-1})$

Two study designs with a structured variance components, unrelated individuals with a non-genetic grouping factor and related individuals in families, are compared to the reference study design of unrelated individuals. The performance is evaluated to To make the study designs comparable, the sum of variance components in the V matrix is equal to one. (a) The effective size multiplier, estimated as $tr(V^{-1}\Sigma_g)/n$, depends on the variance explained. (b) When the variance explained fixed to 50% and the sample size varies, Notation: n, the sample size.

Ref 1.1

1.3.2 Analytical results for testing gene-environment interaction effect

Ref 1.2

1.3.3 Implication in association studies

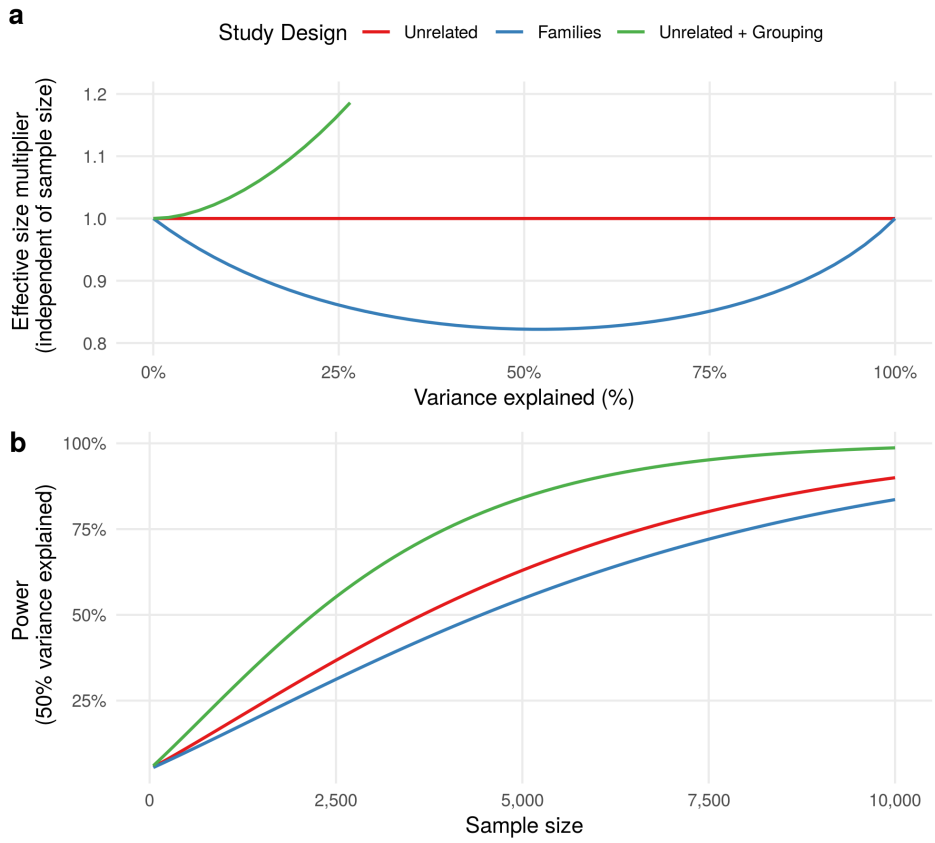


Figure 1.1: Three study designs are compared in terms of power to detect marginal genetic effect under the model $y \sim \mathcal{N}(\mu + \beta_g x_g, V)$ (see also Table 1.1). The reference study design "Unrelated" with $V = \sigma_r^2 I$ is; the study

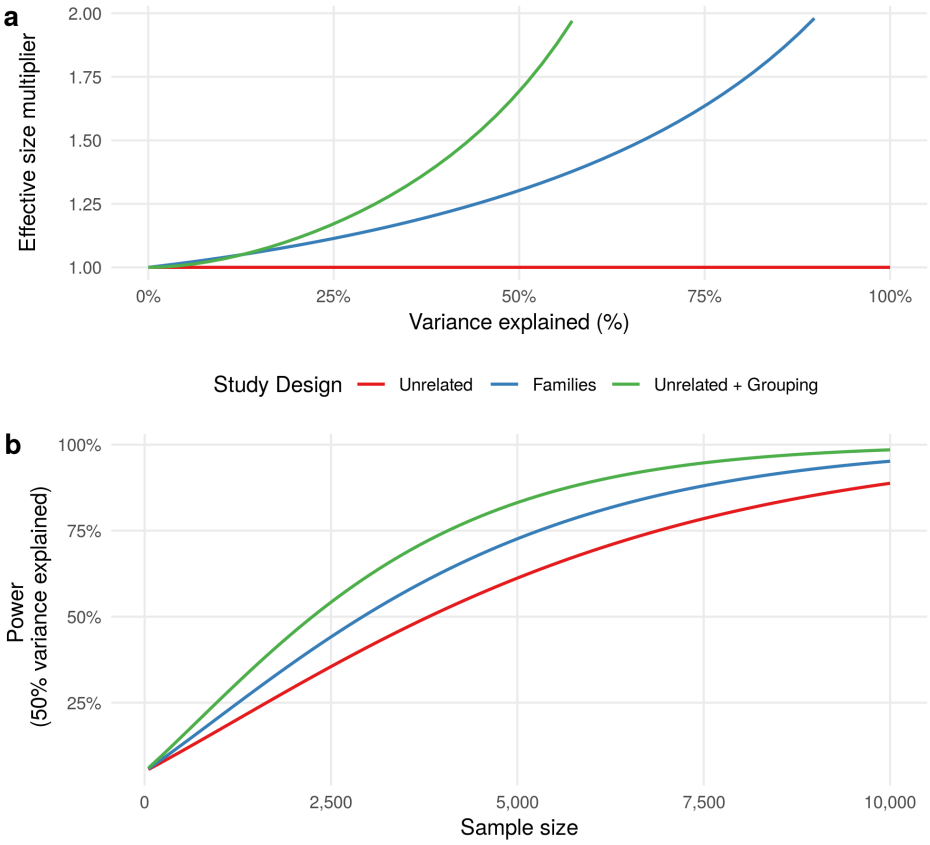


Figure 1.2: Comparison of study designs to detect interaction effect.

Chapter 2

Supplementary Material

2.1 Propositions

Quadratic form: If $\mathcal{X}$ is a vector of random variables with mean μ and (nonsingular) covariance matrix Σ, then the quadratic form $\mathcal{X}^T A\mathcal{X}$ is a scalar random variable:

$$E(\mathcal{X}^T A\mathcal{X}) = tr(A\Sigma) + \mu^T \Sigma\mu \tag{2.1}$$

$$Var(\mathcal{X}^T A\mathcal{X}) = 2tr(A\Sigma A\Sigma) + 4\mu A\Sigma A\mu \tag{2.2}$$

See [**?**, Appendix 3, pp. 843] for more details.

A linear transform of a random vector: If B is a constant matrix and $\mathcal{X}$ is a vector of random variables with mean μ and covariance matrix Σ, then $B\mathcal{X}$ is a vector of random variables:

$$E(B\mathcal{X}) = BE(\mathcal{X}) \tag{2.3}$$

$$Var(B\mathcal{X}) = BVar(\mathcal{X})B^T \tag{2.4}$$

The proof makes use of definitions of mean and variance.

Eigen-value decompostion (EVD): If K is the covariane matrix of size $n \times n$, that means K is symmetric and positive semi-definite. Furthemore, EVD of K is

$$K = QDQ^T = QDQ^{-1} \tag{2.5}$$

where Q is an $n \times n$ orthogonal matrix of eigen-vectors and D is a $n \times n$ diagonal matirx of eigen-values (λ_K^i with i from 1 to n).

EVD for the matrix inverse to K is

$$K^{-1} = QD^{-1}Q^T \tag{2.6}$$

EVD for the matrix such as $V = aK + bI$, where a and b are scalars, I is the $n \times n$ identity matrix, is

$$V = aK + bI = aQDQ^T + bI = aQDQ^T + bQIQ^T = Q(aK + bI)Q^T \tag{2.7}$$

Eigen-value decompostion (EVD) and the trace operator: For the covariance matrix K and the matrix $V = aK + bI$, we have the following serie of equation in relation to the trace operator.

$$
\begin{aligned}
tr(K) &= \sum_{i=1}^{n} \lambda_K^i \\
tr(K^{-1}) &= \sum_{i=1}^{n} (\lambda_K^i)^{-1} \\
tr(V) = tr(aK + bI) &= \sum_{i=1}^{n} (a\lambda_K^i + b) \\
tr(V^{-1}) = tr((aK + bI)^{-1}) &= \sum_{i=1}^{n} (a\lambda_K^i + b)^{-1} \\
tr(V^{-1}K) = tr((aK + bI)^{-1}K) = tr((aI + bK^{-1})^{-1}) &= \sum_{i=1}^{n} (a + b(\lambda_K^i)^{-1})^{-1}
\end{aligned}
\tag{2.8}
$$

In the last equation we used the following equality.

$$
\begin{aligned}
V^{-1}K = (aK + bI)^{-1}K &= (aK + bI)^{-1}(K^{-1})^{-1} \\
&= K^{-1}(aK + bI)^{-1} = (aI + bK^{-1})^{-1}
\end{aligned}
\tag{2.9}
$$

2.2 Analytical derivations

To study the impact of relatedness among individuals on modeling a continuous phenotype y, we consider the following linear mixed model:

$$
y = X\beta + \sum_{k=1}^{m} r_k + e \tag{2.10}
$$

where n is the number of individuals, p is the number of covariates or fixed effects, m is the number of structured random effects apart from the residuals errors, y is a phenotype vector of length n, X is a matrix of covariates of size $n \times p$, β is a vector of fixed effects of length p. The vectors of random effects r_k and e are mutually uncorrelated and multivariate normally distributed as $\mathcal{N}(0, \sigma_k^2 R_k)$ and $\mathcal{N}(0, \sigma_r^2 I)$. The variance-covariance matrices are parametrized with scalar parameters and constant matrices of size $n \times n$ that express relationships among n individuals. The first m random effects r_k are referred here as structured, whereas the last component e is simply the residual errors which are independent and identically distributed.

Thus, the phenotype follows a multivariate normal distribution (MVN) and Equation (2.10) can be rewritten:

$$
y \sim \mathcal{N}(X\beta, V) = \mathcal{N}(X\beta, \sum_{k=1}^{m} \sigma_k^2 R_k + \sigma_r^2 I) \tag{2.11}
$$

We further consider several parameterizations of the model in Equation (2.11) that depend on (i) whether marginal genetic or gene-environment interaction effect is under testing; (ii) whether structured random effects are included or only the residual errors. Consequently, the composition of fixed and random effects are updated accordingly via the matrices X and V, respectively.

2.2.1 Testing marginal genetic effect in unrelated individuals

We rewrite Equation (2.11) as following:

$$
y \sim \mathcal{N}(X\beta, V) = \mathcal{N}(\mu x_0 + \beta_g x_g, \sigma_r^2 I) \tag{2.12}
$$

where $x_0 = 1_n$ is a vector of n ones, μ is a mean of the phenotypic values, x_g is a vector of length of the genotypic values, β_g is the effect size of the genotype.

The ordinary least squares (OLS) solution for fixed effects is the following in the matrix form, $\hat{\beta} = (X^T X)^{-1} X y$. Further, the effect β_g can be estimated separately from the mean effect μ if vectors y and x_g are centered and, thus, the two vectors are uncorrelated. Hence, the estimated effect is expressed as $\hat{\beta_g} = (\tilde{x}_g^T \tilde{x}_g)^{-1} \tilde{x}_g \tilde{y}$, where $\tilde{x}_g$ and $\tilde{y}$ are centered vectors x_g and y, respectively. The variance of the estimate is $var(\hat{\beta_g}) = \sigma_r^2 / (\tilde{x}_g^T \tilde{x}_g)$ and the final expression is the following:

$$\hat{\beta}_g = (\tilde{x}_g^T \tilde{x}_g)^{-1} \tilde{x}_g^T \tilde{y} \sim \mathcal{N}(\beta_g, \sigma_r^2 / (\tilde{x}_g^T \tilde{x}_g)) \tag{2.13}$$

We next approximate the expression $\tilde{x}_g^T \tilde{x}_g$ by using the fact that that x_g is a realization of a vector of random variables $\mathcal{X}_\}$, which is a genotype in n unrelated individuals with a minor allele frequency p. Consequently, we denote $\mathcal{X}_\} \sim (\mu_g, \Sigma_g) = (p1_n, \delta_g^2 I) = (p1_n, 2p(1-p)I)$ and also $\tilde{\mathcal{X}}_\} \sim (0_n, \Sigma_g)$, where I is the identity matrix of size $n \times n$, 1_n is a vector of n ones and 0_n is a vector of n zeros. Applying the proposition for quadratic forms in Equation (2.2) for $\tilde{\mathcal{X}}_\}$, we obtain the approximation:

$$\tilde{x}_g^T \tilde{x}_g \approx E(\tilde{\mathcal{X}}_\}^T \tilde{\mathcal{X}}_\}) = tr(\delta_g^2 I) = \delta_g^2 n = 2p(1-p)n \tag{2.14}$$

The NCP parameter for testing the marginal genetic effect in unrelated individuals is approximated as following:

$$NCP_{unrel} = \hat{\beta}_g^2 / var(\hat{\beta}_g) \approx \hat{\beta}_g^2 \delta_g^2 n / \sigma_r^2 = \hat{\beta}_g^2 2p(1-p)n / \sigma_r^2 \tag{2.15}$$

If the the phenotype y is standardized, i.e. $var(y) = 1$ and the effect β_g is small, then we can further approximate $\sigma_r^2 \approx 1$ based on the following:

$$\begin{aligned} \sigma_r^2 &\approx \hat{\sigma}_r^2 = \hat{e}^T \hat{e} / (n-2) \\ &= (\tilde{y} - \hat{\beta}_g \tilde{x})^T (\tilde{y} - \hat{\beta}_g \tilde{x}) / (n-2) \approx \tilde{y}^T \tilde{y} / (n-2) \approx 1 \end{aligned} \tag{2.16}$$

Hence, we obtain the NCP estimation for the scaled phenotype:

$$NCP_{unrel} = \hat{\beta}_g^2 / var(\hat{\beta}_g) \approx \hat{\beta}_g^2 \delta_g^2 n = \hat{\beta}_g^2 2p(1-p)n \tag{2.17}$$

2.2.2 Testing marginal genetic effect in related individuals

We rewrite Equation (2.11) as following:

$$y \sim \mathcal{N}(X\beta, V) = \mathcal{N}(\mu x_0 + \beta_g x_g, \sum_{k=1}^{m} \sigma_k^2 R_k + \sigma_r^2 I) \tag{2.18}$$

The initial step in solving a linear mixed model is to estimate random effects parameters (σ_k^2 and σ_r^2) by maximum likelihood (ML), restricted maximum likelihood (REML) or other optimization technique [**?**]. Once the estimate of the variance-covariance matrix is found, $\hat{V} = \sum \hat{\sigma}_i^2 R_i + \hat{\sigma}_r^2 I$, the generalized least squares (GLS) solution for fixed effects is applied in the following matrix form, $\hat{\beta} = (X^T \hat{V}^{-1} X)^{-1} X \hat{V}^{-1} y$. This solution is obvious if both parts of Equation (2.18) are multiplied by $\hat{V}^{-0.5}$, thus removing the correlation structure in the random part.

$$\hat{V}^{-0.5} y \sim \mathcal{N}(\mu \hat{V}^{-0.5} x_0 + \beta_x \hat{V}^{-0.5} x, I) \tag{2.19}$$

The expression for the genetic effect estimate is obtained similarly to Equation (2.13) and working with centered vectors:

$$\hat{\beta}_g = (\tilde{x}_g^T \hat{V}^{-1} \tilde{x}_g)^{-1} \tilde{x}_g^T \hat{V}^{-1} \tilde{y} \sim \mathcal{N}(\beta_g, 1/(\tilde{x}_g^T \hat{V}^{-1} \tilde{x}_g)) \tag{2.20}$$

We further again consider the quadratic form $\tilde{x}_g^T \hat{V}^{-1} \tilde{x}_g$ and use its mean for approximation, as shown in Equation (2.2). A vector x_g of the genotypic values is a realization of a vector of random variables $\mathcal{X}_\} \sim (\mu_g, \Sigma_g) = (p1_n, \delta_g^2 K) = (p1_n, 2p(1-p)K)$, where p is a minor allele frequency of the genotype and K is the kinship matrix of size $n \times n$. We also introduce a centered vector of random variables $\mathcal{X}_\} \sim (0_n, \Sigma_g)$.

The matrix K expresses the genetic relatedness among n individuals and it is the identity matrix I for genetically unrelated individuals. The derivation presented here are appropriate for any form of the matrix K.

Treating the matrix $\hat{V}^{-1}$ as a (constant) transformation matrix A in Equation (2.2) for quadratic forms gives us the approximation:

$$\tilde{x}_g^T \hat{V}^{-1} \tilde{x}_g \approx E(\tilde{\mathcal{X}}_\}^T \hat{V}^{-1} \tilde{\mathcal{X}}_\}) = tr(\hat{V}^{-1} \Sigma_g) = \delta_g^2 tr(\hat{V}^{-1} K) = 2p(1-p) tr(\hat{V}^{-1} K) \tag{2.21}$$

The NCP parameter for testing the marginal genetic effect in related individuals is approximated as following:

$$\begin{aligned} NCP_{rel} &= \hat{\beta}_g^2 / var(\hat{\beta}_g) \approx \hat{\beta}_g^2 tr(\hat{V}^{-1} \Sigma_g) \\ &= \hat{\beta}_g^2 \delta_g^2 tr(\hat{V}^{-1} K) = \hat{\beta}_g^2 2p(1-p) tr(\hat{V}^{-1} K) \end{aligned} \tag{2.22}$$

2.2.3 Effective size multiplier for testing marginal genetic effect

We joint results from the previous two sections 2.2.1 and 2.2.2 to derive the formula for ratio between NCP_{rel} and NCP_{unrel}, as referred herein the effective size multiplier.

$$NCP_{rel} / NCP_{unrel} = tr(\hat{V}^{-1} K) / (n/\sigma_r^2) \tag{2.23}$$

If the variance of the phenotype y is standardized to 1 and the variance captured by the genotype is small, then we can approximate $\sigma_r^2 \approx 1$ in Equation (2.22) and further obtain:

$$NCP_{rel} / NCP_{unrel} = tr(\hat{V}^{-1} K) / n \tag{2.24}$$

The variance components in $\hat{V}$ are then considered as the proportions, since the variance of the phenotype y is standardized to 1.

2.2.4 Testing gene-environment interaction effect in unrelated individuals

We rewrite Equation (2.11) as following:

$$y \sim \mathcal{N}(X\beta, V) = \mathcal{N}(\mu x_0 + \beta_g x_g + \beta_e x_e + \beta_{ge} x_{ge}, \sigma_r^2 I) \tag{2.25}$$

where $x_0 = 1_n$ is a vector of n ones, μ is a mean of the phenotypic values, x_g is a genotype vector of length n, β_g is the effect size of the genotype, x_e is a environment exposure vector of length n, β_e is the effect size of the environment exposure, x_{ge} is a vector of length n of gene-environment interaction, β_{ge} is the interaction effect size.

The coding scheme of the genotypic and environmental variables to study gene-environment interaction under the standard linear model has been reviewed elsewhere [**?**]. Here, we work with centered variables $\tilde{x}_g$ and $\tilde{x}_e$, and define the interaction variable $\tilde{x}_{ge}$ by (i) element-wise multiplication of the two variables denoted as $\ddot{x}_{ge} = \tilde{x}_g \cdot \tilde{x}_e$, (ii) centering the resulted product $\ddot{x}_{ge}$. Hence, the effect size for each variable

(columns in X) can be estimated independently from the other variables under assumption that the two random variables of genotype and environmental exposure are independent [**?**, Appendix C]. Of a note, different coding schemes give different estimates of effect sizes, but the test statistic for gene-environment interaction is the same [**?**, Appendix B].

Therefore, the estimate of interest $\hat{\beta}_{ge}$ has the following distribution:

$$\hat{\beta}_{ge} = (\tilde{x}_{ge}^T \tilde{x}_{ge})^{-1} \tilde{x}_{ge}^T \tilde{y} \sim \mathcal{N}(\beta_{ge}, \sigma_r^2 / (\tilde{x}_{ge}^T \tilde{x}_{ge})) \tag{2.26}$$

To further approximate the quantity $\tilde{x}_{ge}^T \tilde{x}_{ge}$, we need to work with two random variables. The first one is a vector of random variables $\tilde{\mathcal{X}}_{\}} \sim (0_n, \Sigma_g) = (0_n, \delta_g^2 I) = (0_n, 2p(1-p)I)$, which we previously described. The second is a vector of random variables $\tilde{\mathcal{X}}_{ge} = \tilde{x}_e \cdot \tilde{\mathcal{X}}_g = E\tilde{\mathcal{X}}_g$, which is a transformed variable of $\tilde{\mathcal{X}}_{\}}$ with the transformation matrix $E = \text{diag}(\tilde{x}_e)$, defined as a diagonal matrix with values equal to those observed in the environmental exposure. The operator $\cdot$ denotes the element-wise multiplication (the Hadamard product).

We also consider a simple case for the environmental exposure when it is binary and the observed frequency of exposure is f. Then the values on diagonal of the matrix E are equal $-f$ and $1-f$, and we denote this matrix as E_b.

We further give an example of the matrix E_b for 5 individuals under study with the first two unexposed and the last three exposed to the environment, i.e. $f = 0.6$.

$$E_b = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 0.6 \\ 0.6 \\ 0.6 \\ 0.6 \\ 0.6 \end{pmatrix} = \begin{pmatrix} -0.6 & 0 & 0 & 0 & 0 \\ 0 & -0.6 & 0 & 0 & 0 \\ 0 & 0 & 0.4 & 0 & 0 \\ 0 & 0 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & 0 & 0.4 \end{pmatrix}$$

We first need to derive the variance of the random variable $\tilde{\mathcal{X}}_{ge}$. We obtain from Equation (2.4):

$$\begin{aligned} \text{Var}(\tilde{\mathcal{X}}_{ge}) &= \text{Var}(E\tilde{\mathcal{X}}_g) = E\text{Var}(\tilde{\mathcal{X}}_g)E^T \\ &= E(\delta_g^2 I)E^T = \delta_g^2 EE^T = \delta_g^2 E^2 \end{aligned} \tag{2.27}$$

Applying the results for quadratic forms in Equation (2.2) gives us the approximation:

$$\tilde{x}_{ge}^T \tilde{x}_{ge} \approx E(\tilde{\mathcal{X}}_{ge}^T \tilde{\mathcal{X}}_{ge}) = tr(\delta_g^2 E^2) = \delta_g^2 tr(E^2) = 2p(1-p)tr(E^2) \tag{2.28}$$

When the exposure is binary, we can simplify this quantity using the following equality $tr(E_b) = f(1-f)n$:

$$\tilde{x}_{ge}^T \tilde{x}_{ge} \approx \delta_g^2 f(1-f)n = 2p(1-p)f(1-f)n \tag{2.29}$$

Next, the NCP parameter for testing the gene-environment interaction effect in unrelated individuals is approximated as following:

$$NCP_{unrel+int} = \hat{\beta}_{ge}^2 / var(\hat{\beta}_{ge}) \approx \hat{\beta}_{ge}^2 \delta_g^2 tr(E^2)/\sigma_r^2 = \hat{\beta}_{ge}^2 2p(1-p)tr(E^2)/\sigma_r^2 \tag{2.30}$$

When the exposure is binary:

$$NCP_{unrel+int} = \hat{\beta}_{ge}^2 / var(\hat{\beta}_{ge}) \approx \hat{\beta}_{ge}^2 \delta_g^2 tr(E_b^2)/\sigma_r^2 = \hat{\beta}_{ge}^2 2p(1-p)f(1-f)n/\sigma_r^2 \tag{2.31}$$

Additionally, we approximate $\sigma_r^2 \approx 1$ if the the phenotype y is standardized and the variance captured by all genetic, environmental and interaction effects is small.

2.2.5 Testing gene-environment interaction effect in related individuals

We rewrite Equation (2.11) as following:

$$y \sim \mathcal{N}(X\beta, V) = \mathcal{N}(\mu x_0 + \beta_g x_g + \beta_e x_e + \beta_{ge} x_{ge}, \sum_{k=1}^{m} \sigma_k^2 R_k + \sigma_r^2 I) \tag{2.32}$$

As in the previous derivation in Section 2.2.4, we apply the same coding scheme for genetic, environmental and gene-environmental interaction variables, $\tilde{x}_g$, $\tilde{x}_e$ and $\tilde{x}_{ge}$, respectively. As in the previous Section 2.2.2, we derive the distribution of $\hat{\beta}_{ge}$ conditionally on the estimate of the variance-covariance matrix $\hat{V} = \sum \hat{\sigma}_i^2 R_i + \hat{\sigma}_r^2 I$:

$$\hat{\beta}_{ge} = (\tilde{x}_{ge}^T \hat{V}^{-1} \tilde{x}_{ge})^{-1} \tilde{x}_{ge}^T \hat{V}^{-1} \tilde{y} \sim \mathcal{N}(\beta_{ge}, 1/(\tilde{x}_{ge}^T \hat{V}^{-1} \tilde{x}_{ge})) \tag{2.33}$$

Also as in the previous Section 2.2.2, we consider the two random vectors, $\tilde{\mathcal{X}}_{\}} \sim (0_n, \Sigma_g) = (0_n, \delta_g^2 K) = (0_n, 2p(1-p)K)$, and $\tilde{\mathcal{X}}_{ge} = \tilde{x}_e \cdot \tilde{\mathcal{X}}_g = E\tilde{\mathcal{X}}_g$. The later is a transformed variable of $\tilde{\mathcal{X}}_{\}}$ with the transformation matrix $E = \text{diag}(\tilde{x}_e)$,

In addition, we introduce a matrix D, which value at row i and column j is equal to the product of two diagonal entries i and j of E, i.e. $D_{i,j} = E_{i,i}E_{j,j}$. The use of this matrix D is explained below.

When the environmental exposure is binary with the exposure frequency f, we denote the matrix E as E_b and the matrix D as D_b. The values on diagonal of E_b are either f or $1-f$, while the values of D_b are either f^2, $(1-f)^2$ or $f(1-f)$.

We derive the variance of the random variable $\tilde{\mathcal{X}}_{ge}$ using proposition in Equation (2.4):

$$\begin{aligned} \text{Var}(\tilde{\mathcal{X}}_{ge}) &= \text{Var}(E\tilde{\mathcal{X}}_g) = E\text{Var}(\tilde{\mathcal{X}}_g)E^T = E\Sigma_g E^T \\ &= \delta_g^2 EKE^T = \delta_g^2 D \cdot K = \delta_g^2 K_D \end{aligned} \tag{2.34}$$

In the second part of derivation, we again used the fact that the matrix E is diagonal; that means the expression EAE^T for a given matrix A can be rewritten as $D \cdot A$, where the D was defined before and the operator $\cdot$ denotes the element-wise multiplication (the Hadamard product).

In Equation (2.34) we introduced a special kinship matrix K_D "masked" by the environmental exposure though the matrix D defined above.

$$K_D = D \cdot K \tag{2.35}$$

We note that the K_D matrix becomes the E^2 matrix in the previous section 2.2.4 when $K = I$, i.e. the case of unrelated individuals.

For an illustration example, we show how the matrices E_b, D_b, K and K_D look like for 5 individuals with the first two unexposed and the last three exposed to the environment, i.e. $f = 0.6$. The five individuals represent a nuclear family of two parents and three children.

$$E_b = \begin{pmatrix} -0.6 & 0 & 0 & 0 & 0 \\ 0 & -0.6 & 0 & 0 & 0 \\ 0 & 0 & 0.4 & 0 & 0 \\ 0 & 0 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & 0 & 0.4 \end{pmatrix}$$

$$D_b = \begin{pmatrix} -0.36 & -0.36 & -0.24 & -0.24 & -0.24 \\ -0.36 & -0.36 & -0.24 & -0.24 & -0.24 \\ -0.24 & -0.24 & 0.16 & 0.16 & 0.16 \\ -0.24 & -0.24 & 0.16 & 0.16 & 0.16 \\ -0.24 & -0.24 & 0.16 & 0.16 & 0.16 \end{pmatrix}$$

$$K = \begin{pmatrix} 1 & 0 & 0.5 & 0.5 & 0.5 \\ 0 & 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 1 \end{pmatrix}$$

$$K_{D_b} = \begin{pmatrix} -0.36 & 0 & -0.12 & -0.12 & -0.12 \\ 0 & -0.36 & -0.12 & -0.12 & -0.12 \\ -0.12 & -0.12 & 0.16 & 0.08 & 0.08 \\ -0.12 & -0.12 & 0.08 & 0.16 & 0.08 \\ -0.12 & -0.12 & 0.08 & 0.08 & 0.16 \end{pmatrix}$$

Further applying the proposition for quadratic forms in Equation (2.2) gives us the approximation:

$$\begin{aligned} \tilde{x}_{ge}^T \hat{V}^{-1} \tilde{x}_{ge} &\approx E(\tilde{\mathcal{X}}_{ge}^T \hat{V}^{-1} \tilde{\mathcal{X}}_{ge}) = tr(\hat{V}^{-1} \delta_g^2 K_D) \\ &= \delta_g^2 tr(\hat{V}^{-1} K_D) = 2p(1-p) tr(\hat{V}^{-1} K_D) \end{aligned} \tag{2.36}$$

The NCP parameter for testing the gene-environment interaction effect in related individuals is approximated as following:

$$NCP_{rel+int} = \hat{\beta}_{ge}^2 / var(\hat{\beta}_{ge}) \approx \hat{\beta}_{ge}^2 \delta_g^2 tr(\hat{V}^{-1} K_D) = \hat{\beta}_{ge}^2 2p(1-p) tr(\hat{V}^{-1} K_D) \tag{2.37}$$

2.2.6 Effective size multiplier for testing marginal gene-environment interaction effect

We joint results from the previous two sections 2.2.4 and 2.2.5 and present the formula for ratio between $NCP_{rel+int}$ and $NCP_{unrel+int}$, as referred herein the effective size multiplier.

$$NCP_{rel+int} / NCP_{unrel+int} = tr(\hat{V}^{-1} K_D) / (tr(E^2)/\sigma_r^2) \tag{2.38}$$

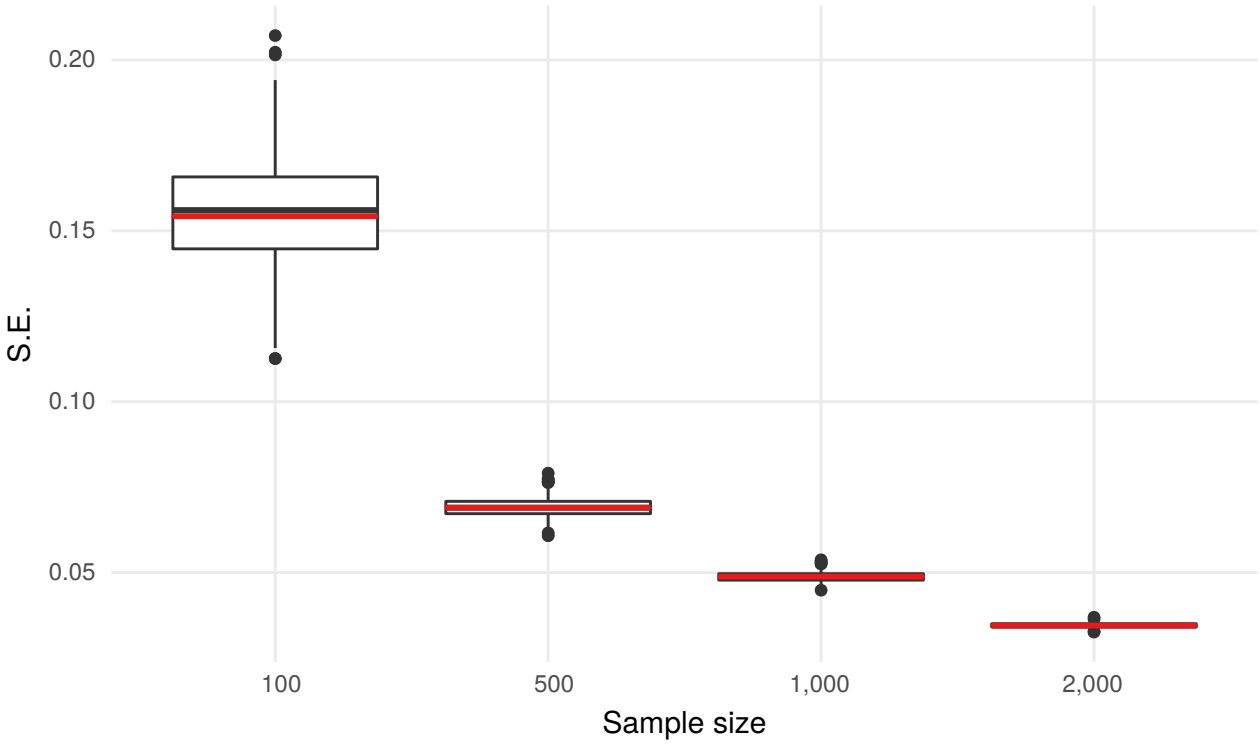
If the variance of the phenotype y is standardized to 1 and the variance captured by fixed effects is small, then we can approximate $\sigma_r^2 \approx 1$ in Equation (2.30) and further obtain:

$$NCP_{rel+int} / NCP_{unrel+int} = tr(\hat{V}^{-1} K_D) / tr(E^2) \tag{2.39}$$

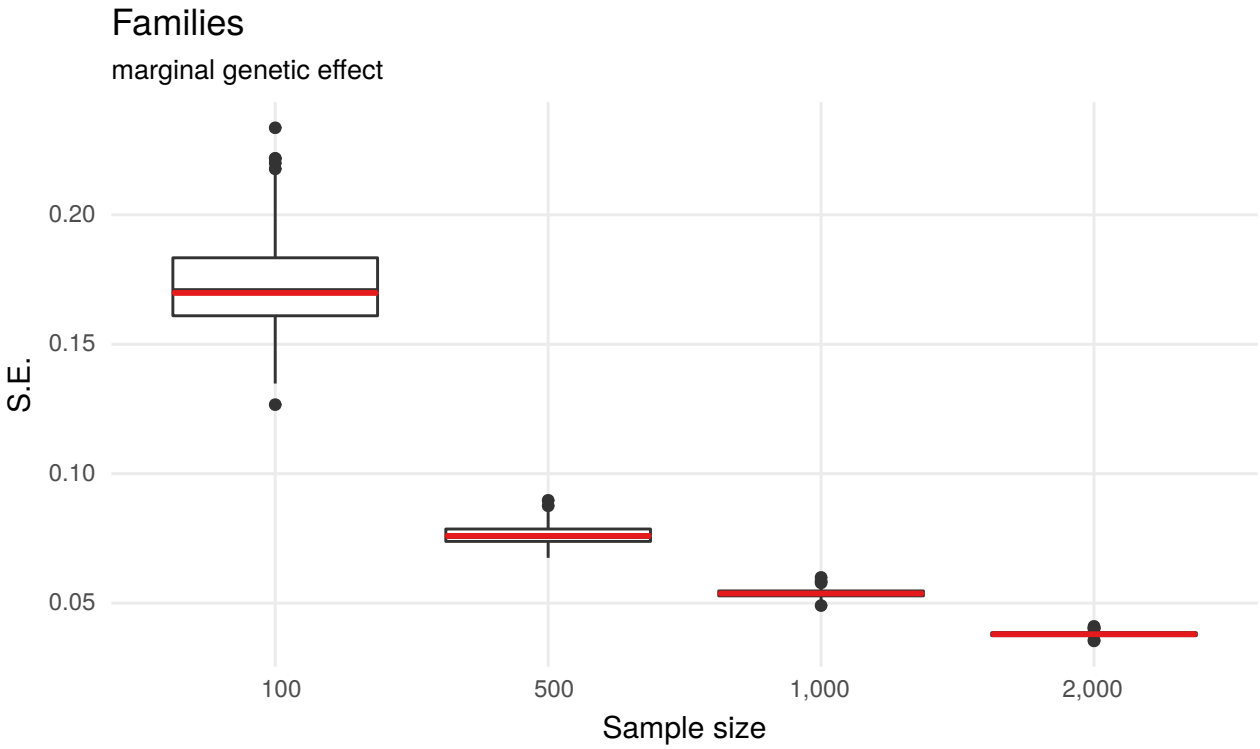
The variance components in $\hat{V}$ are then considered as the proportions, since the variance of the phenotype y is standardized to 1.

2.3 Simulations

2.3.1 Unrelated: marginal genetic effect

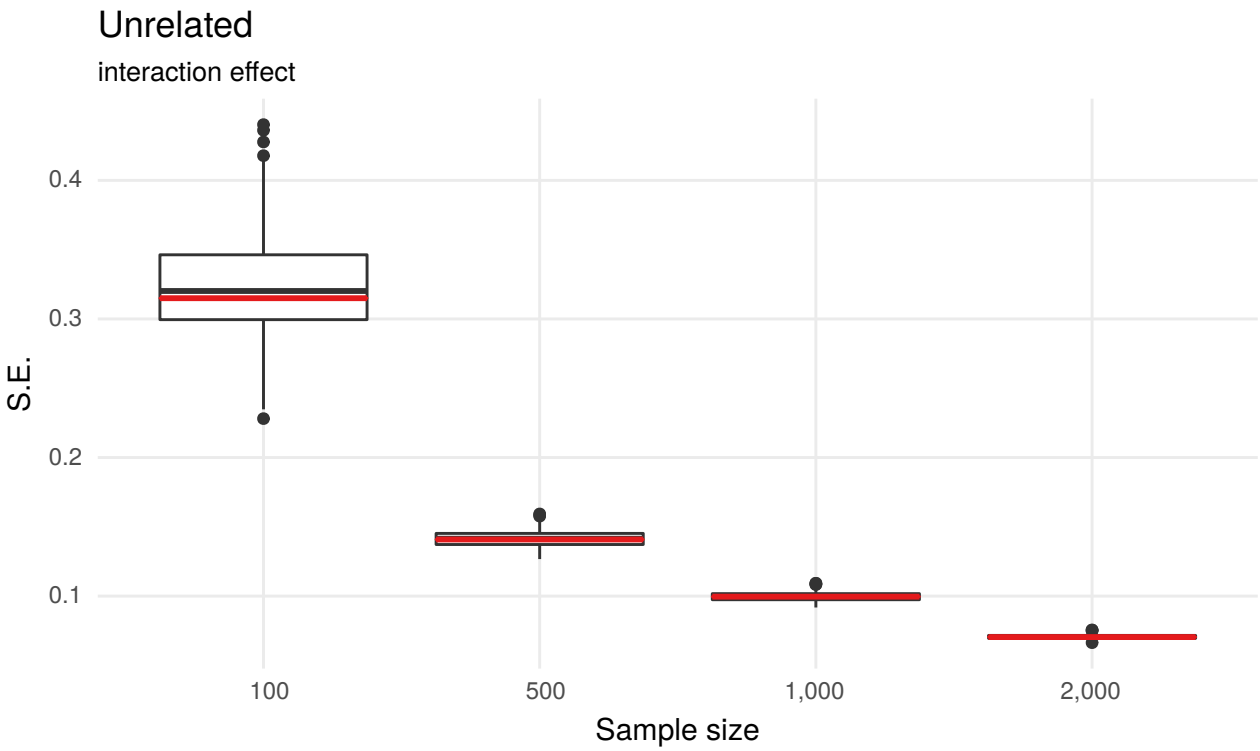


2.3.2 Families: marginal genetic effect

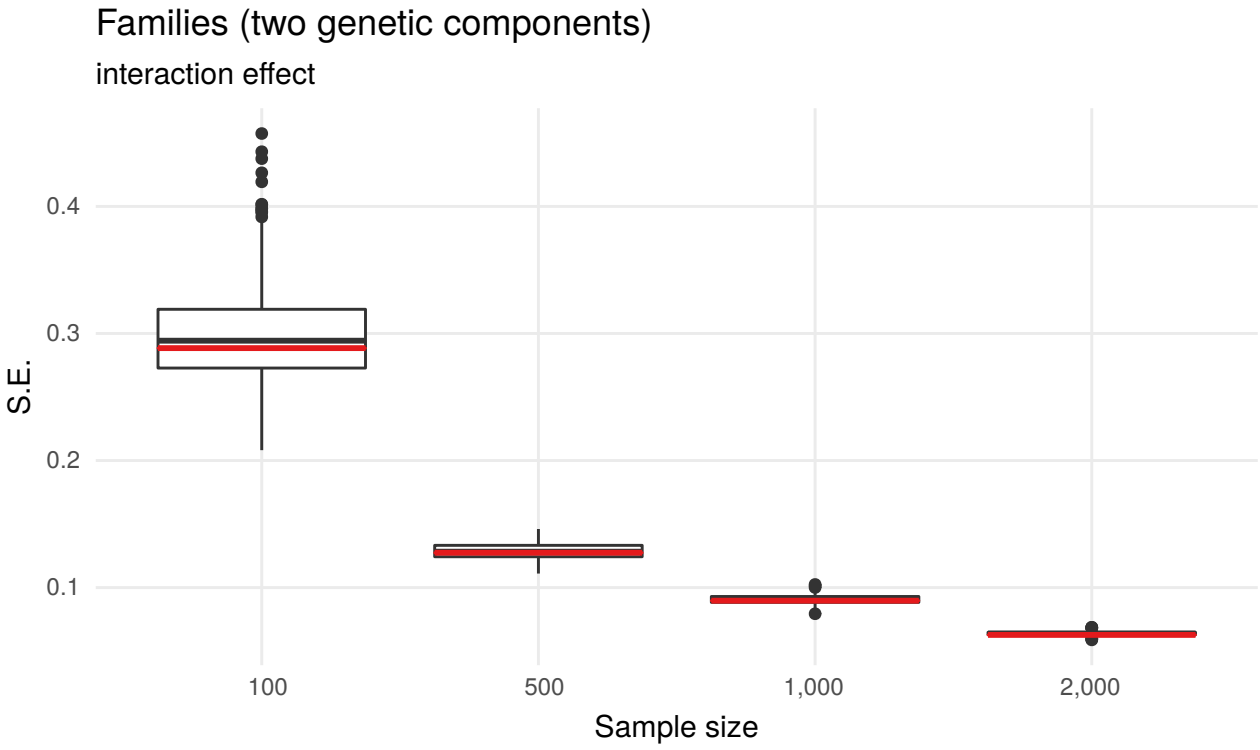


Sample Size	Trace Factor
100	0.8253
500	0.8253
1,000	0.8253
2,000	0.8253

2.3.3 Unrelated: interaction effect

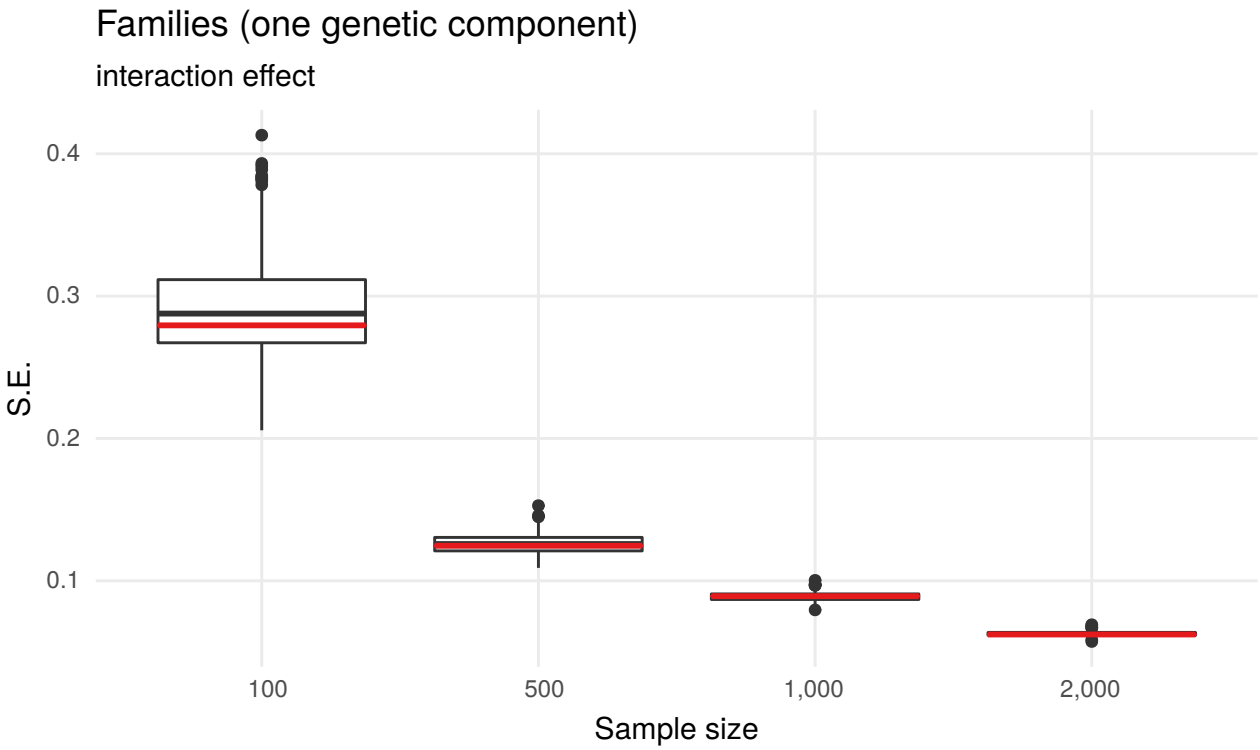


2.3.4 Familes (two genetic components): interaction effect



Sample Size	Trace Factor
100	1.1921
500	1.2219
1,000	1.2333
2,000	1.2598

2.3.5 Familes (one genetic component): interaction effect



Sample Size	Trace Factor
100	1.2702
500	1.2773
1,000	1.2463
2,000	1.2775

2.4 Supplementary Figures

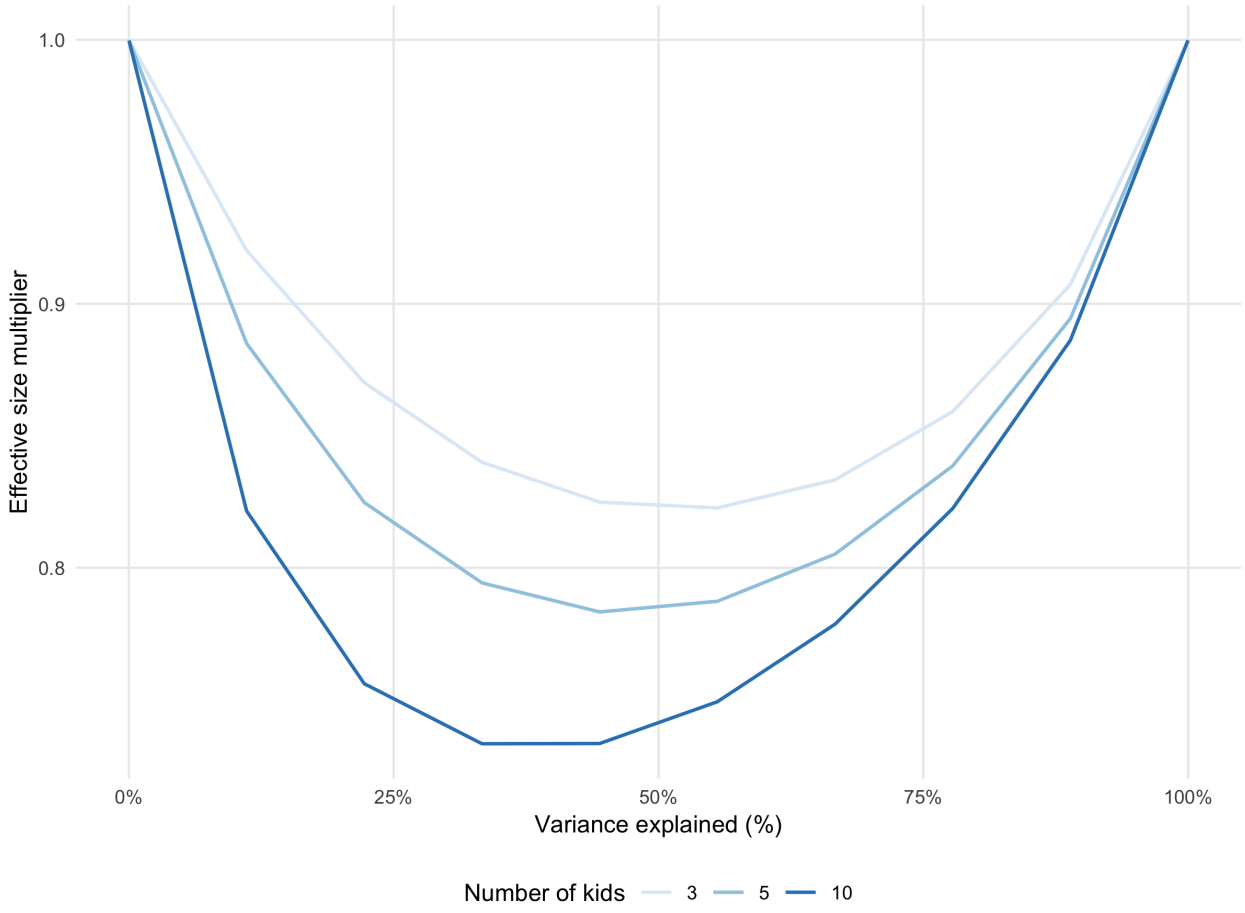


Figure 2.1: Influence of family structure to detect marginal genetic effect.

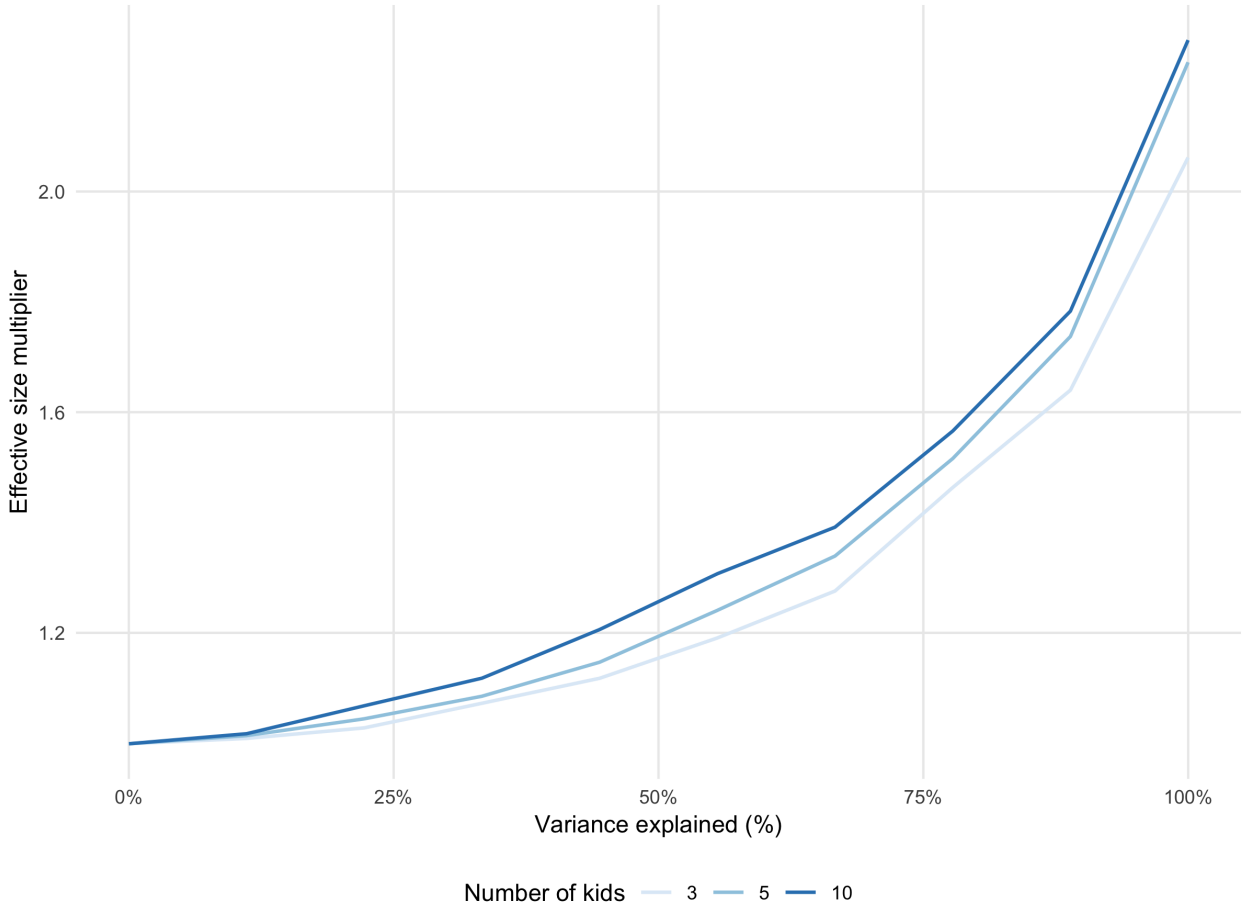


Figure 2.2: Influence of family structure to detect interaction genetic effect.

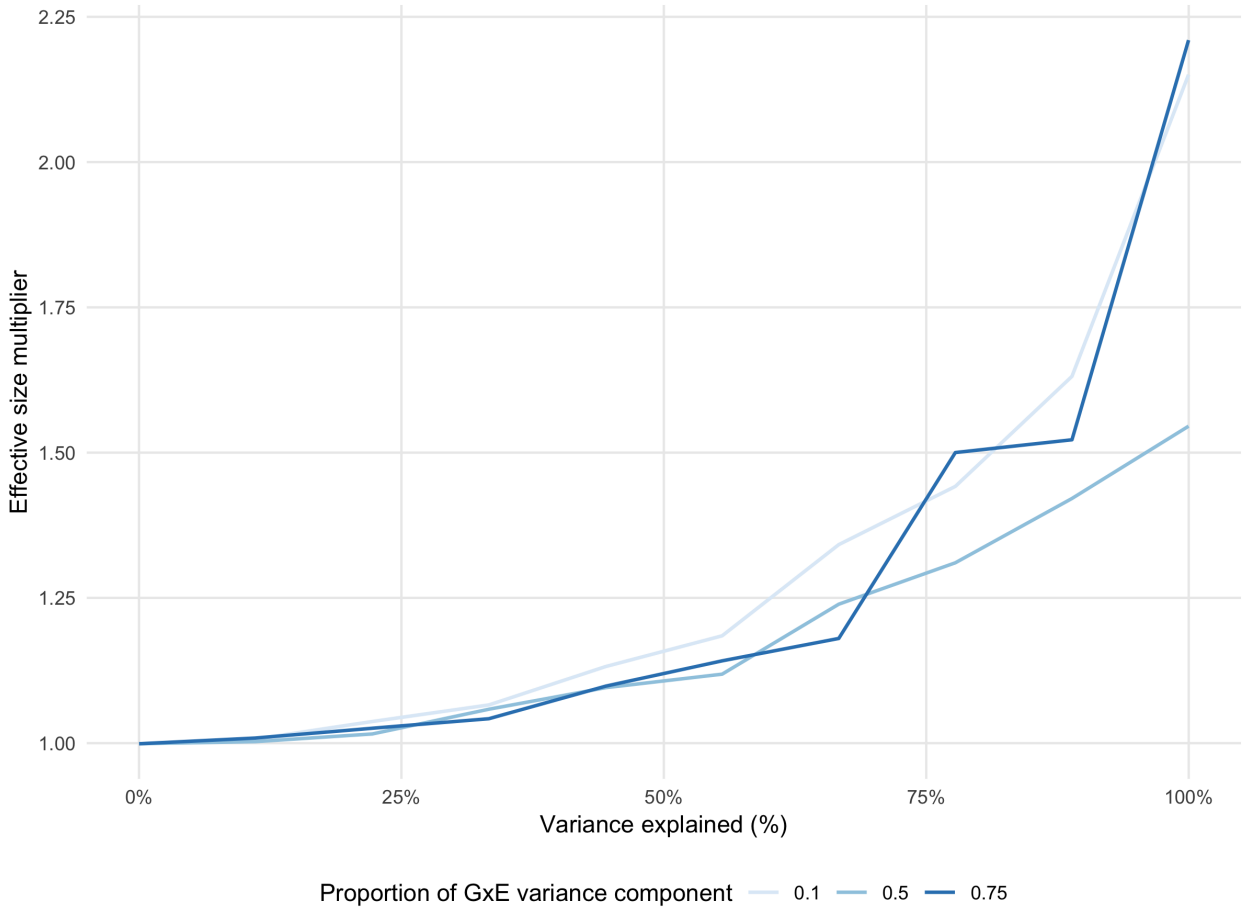


Figure 2.3: Influence of GxE variance component to detect interaction effect.