International Journal of
Environmental Research and Public Health

MDPI

Article

An Efficient Test for Gene-Environment Interaction in Generalized Linear Mixed Models with Family Data

Mauricio A. Mazo Lopera [1,2], Brandon J. Coombes [2] and Mariza de Andrade [2,*]

[1] School of Statistics, National University of Colombia, Medellín, Antioquia 050022, Colombia; mauromazo35@gmail.com
[2] Departament of Health Sciences Research, Mayo Clinic, Rochester, MN 55905, USA; coombes.brandon@mayo.edu
* Correspondence: mandrade@mayo.edu; Tel.: +1-507-284-1032

Received: 19 August 2017; Accepted: 25 September 2017; Published: 27 September 2017

Abstract: Gene-environment (GE) interaction has important implications in the etiology of complex diseases that are caused by a combination of genetic factors and environment variables. Several authors have developed GE analysis in the context of independent subjects or longitudinal data using a gene-set. In this paper, we propose to analyze GE interaction for discrete and continuous phenotypes in family studies by incorporating the relatedness among the relatives for each family into a generalized linear mixed model (GLMM) and by using a gene-based variance component test. In addition, we deal with collinearity problems arising from linkage disequilibrium among single nucleotide polymorphisms (SNPs) by considering their coefficients as random effects under the null model estimation. We show that the best linear unbiased predictor (BLUP) of such random effects in the GLMM is equivalent to the ridge regression estimator. This equivalence provides a simple method to estimate the ridge penalty parameter in comparison to other computationally-demanding estimation approaches based on cross-validation schemes. We evaluated the proposed test using simulation studies and applied it to real data from the Baependi Heart Study consisting of 76 families. Using our approach, we identified an interaction between BMI and the Peroxisome Proliferator Activated Receptor Gamma (*PPARG*) gene associated with diabetes.

Keywords: gene-environment interaction; generalized linear mixed model; variance component test; score test; ridge regression; best linear unbiased predictor; family data

1. Introduction

Linear mixed models (LMM) have been used to find associations between continuous phenotypes and genetic variants, genes, and gene-environment (GE) interactions in unrelated and related subjects in genome-wide association (GWA) analysis. For unrelated subjects, the analysis can be performed within the generalized linear model framework, however, for related subjects as in the case of family data, one has to include the kinship matrix to take into account the correlation among the relatives for each family. In this paper, we are interested in testing GE interaction for discrete phenotypes. Generalized linear mixed models (GLMM) proposed by Breslow and Clayton [1] is an ideal statistical approach to detect such an interaction with non-continuous phenotypes, because it can treat the familiar effect on the phenotype as a random effect.

Gene-based GE interaction tests have previously been proposed for independent subjects [2–4]. While each GE interaction can be tested individually using one single nucleotide polymorphism (SNP) at a time, it is known that single SNP association is not as powerful as the gene-based analysis [2] due to the linkage disequilibrium (LD) present among the SNPs in a gene. Lin et al. [2] proposed a variance component test (VCT) of the interactions by treating the interactions as a random effect.

This approach was extended to sequencing data with rare variants [3,5]. To overcome multicollinearity of the coefficients of the genetic markers, Lin et al. [2,3] applied ridge regression penalization of SNP coefficients and estimated the ridge penalty parameter with generalized cross-validation. However, this method is computationally demanding and their final test ignores the tuning of the ridge penalty parameter. Coombes [6] instead proposed treating the genetic coefficients as a random effect in a linear mixed model framework to perform the ridge penalization. This equivalence was initially proposed by Bishop and Tipping [7,8] for Bayesian ridge regression in linear models framework. While this approach was able to incorporate the ridge penalty into the test statistic, it was only developed for a quantitative phenotype [6].

Here, we propose a GLMM GE interaction framework for discrete and continuous phenotypes that treats the coefficients of genetic markers as random effects. Also, because the correlation among relatives cannot be ignored, this modeling framework incorporates the kinship matrix in the GLMM [9]. We test for GE interaction between a set of SNPs and an environment by treating interaction coefficients as random effects using a VCT. Our proposed model includes three random effects: the first for genetic variants, the second for gene-environment interaction, and the third for the inclusion of families. In the methods section, we develop the framework for this model. We present the VCT as proposed by Lin [10] in the presence of several random effects and adapt this VCT to accommodate the interactions [11]. We also prove that the corresponding best linear unbiased predictor (BLUP) in our GLMM model is equivalent to the ridge regression estimator, as proposed by Shen et al. [12]. In our simulations, we show that our model can be efficiently computed using the GMMAT package [13] in `R` and maintains appropriate type I error as well as sufficient power. Finally, we apply our methodology to test for genetic interactions with BMI associated with diabetes among the Baependi Heart Study [14], which consists of 76 families of varying pedigree size.

2. Methods

2.1. Generalized Linear Mixed Model

GLMMs have been widely applied in situations where the outcome is discrete and random components are involved in the linear predictor. GLMMs, as proposed by Breslow and Clayton [1], link a response variable y_i, for $i = 1, \ldots, n$, with vectors $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ of explanatory variables associated with the fixed and random effects. Given a r-dimensional vector $\boldsymbol{d}$ of random effects, the model is given by $g\left(\mu_i^d\right) = \boldsymbol{x}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{d}$, where $g(.)$ is known as the link function and $\mu_i^d = E(y_i|\boldsymbol{d})$ is the conditional mean. The conditional variance is given by $Var(y_i|\boldsymbol{d}) = \phi\omega_i\nu\left(\mu_i^d\right)$, with $\nu(.)$ is a known function, ϕ is a scale parameter and ω_i are known weights. Denoting the observation vector by $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ and the design matrices with rows $\boldsymbol{x}_i^T$ and $\boldsymbol{z}_i^T$ by $\boldsymbol{X}$ and $\boldsymbol{Z}$, the general formulation of GLMM is given by

$$g\left(\boldsymbol{\mu}^d\right) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{d} \tag{1}$$

with $\boldsymbol{\mu}^d = (\mu_1^d, \ldots, \mu_n^d)^T$, and where $\boldsymbol{d}$ is assumed multivariate normal distributed with mean $\boldsymbol{0}$ and covariance matrix $\boldsymbol{D} = \boldsymbol{D}(\boldsymbol{\pi})$ depending on an unknown vector $\boldsymbol{\pi}$ of variance components.

2.2. Generalized Linear Mixed Model with Gene-Environment

To set up the GLMM to model GE interaction in families, assume we have a random sample of N independent families from a study population, with n_i members in the ith family such the total number of individuals is $n = \sum_{i=1}^{N} n_i$. For the jth member in the ith family, let Y_{ij} be a discrete or continuous response variable for the phenotype of interest, $\boldsymbol{X}_{ij} = (X_{ij}^1, \ldots, X_{ij}^p)^T$, with p equal to the number of non-environmental covariates, $\boldsymbol{G}_{ij} = (G_{ij}^1, \ldots, G_{ij}^q)^T$ with q equal to number of observed

genotypes for SNPs in a gene, E_{ij} the environmental variable of interest, and $\boldsymbol{S}_{ij} = (E_{ij}G_{ij}^1, \ldots, E_{ij}G_{ij}^q)^T$ the GE interaction. The GE interaction GLMM for families may be written as

$$\begin{aligned} g\left[E(Y_{ij}|\alpha_{ij})\right] &= \boldsymbol{X}_{ij}^T\boldsymbol{\beta}_1 + E_{ij}\beta_2 + \boldsymbol{G}_{ij}^T\boldsymbol{\theta} + \boldsymbol{S}_{ij}^T\boldsymbol{\gamma} + \alpha_{ij} \\ Var(Y_{ij}|\alpha_{ij}) &= \phi\omega_{ij}^{-1}\nu\left[E(Y_{ij}|\alpha_{ij})\right] \\ \boldsymbol{\alpha}_i &= (\alpha_{i1}, \ldots, \alpha_{in_i})^T \sim N(\boldsymbol{0}, 2\sigma^2\boldsymbol{\Phi}_i) \end{aligned} \tag{2}$$

where $g(.)$ is a monotone known function, $Y_{ij}|\alpha_{ij}$ follows a distribution in the exponential family, $\nu(.)$ is a known function, ϕ is a scale parameter that may be known or may need to be estimated, ω_{ij} are known weights (commonly equal to 1), $\boldsymbol{\Phi}_i$ is the kinship matrix, and σ^2 is a parameter to be estimated.

Equation (2) can be rewritten using the Cholesky decomposition of the kinship matrix for the ith family $2\boldsymbol{\Phi}_i = \boldsymbol{K}_i\boldsymbol{K}_i^T$ and assuming that $\boldsymbol{\alpha}_i = \boldsymbol{K}_i\boldsymbol{b}_i$ with $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{n_i})$, where $\boldsymbol{I}_{n_i}$ denotes a $(n_i \times n_i)$ identity matrix. Then, the family model is given by

$$g\left(\boldsymbol{\mu}_i^b\right) = \boldsymbol{X}_i\boldsymbol{\beta}_1 + \boldsymbol{E}_i\beta_2 + \boldsymbol{G}_i\boldsymbol{\theta} + \boldsymbol{S}_i\boldsymbol{\gamma} + \boldsymbol{K}_i\boldsymbol{b}_i \tag{3}$$

with $\boldsymbol{\mu}_i^b = E(\boldsymbol{Y}_i|\boldsymbol{b}_i)$, $g(\boldsymbol{\mu}_i^b) = (g(\mu_{i1}^b), \ldots, g(\mu_{in_i}^b))^T$, $\boldsymbol{X}_i = [\boldsymbol{X}_{i1} \ldots \boldsymbol{X}_{in_i}]^T$, $\boldsymbol{G}_i = [\boldsymbol{G}_{i1} \ldots \boldsymbol{G}_{in_i}]^T$, $\boldsymbol{E}_i = (E_{i1}, \ldots, E_{in_i})^T$, and $\boldsymbol{S}_i = [\boldsymbol{S}_{i1} \ldots \boldsymbol{S}_{in_i}]^T$.

To simplify the computational burden in the estimation process, we generalize Equation (3), by redefining its vectors and matrices as: $\boldsymbol{\mu}^b = E(\boldsymbol{Y}|\boldsymbol{b})$, $\boldsymbol{X} = [\boldsymbol{X}_1 \ldots \boldsymbol{X}_N]^T$, $\boldsymbol{G} = [\boldsymbol{G}_1 \ldots \boldsymbol{G}_N]^T$, $\boldsymbol{E} = (E_1, \ldots, E_N)^T$, $\boldsymbol{S} = [\boldsymbol{S}_1 \ldots \boldsymbol{S}_N]^T$, $\boldsymbol{K} = \text{diag}\{\boldsymbol{K}_1 \ldots \boldsymbol{K}_N\}$ and $\boldsymbol{b} = [\boldsymbol{b}_1 \ldots \boldsymbol{b}_N]^T$, and

$$g\left(\boldsymbol{\mu}^b\right) = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{G}\boldsymbol{\theta} + \boldsymbol{S}\boldsymbol{\gamma} + \boldsymbol{K}\boldsymbol{b} \tag{4}$$

with $\tilde{\boldsymbol{X}} = [\boldsymbol{X}\ \boldsymbol{E}]^T$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \beta_2)^T$. Our goal is to test the null hypothesis $H_0 : \boldsymbol{\gamma} = \boldsymbol{0}$ that the gene of interest has no GE interaction associated with response.

3. Proposed GE Interaction Test

As mentioned by Lin et al. [2], to treat $\boldsymbol{\gamma}$ as a fixed vector and proceed with a p degrees of freedom (DF) score test can result in loss of power to test for interaction. Another common strategy is to use a single SNP analysis of GE interaction, which assumes all SNPs are uncorrelated, however, this is usually not the case. In the majority of cases, the SNPs within a gene are highly correlated, thus, here we propose a test that accounts for correlation among SNPs and has uses less DF than the score test.

Using Equation (4), our proposed test assumes $\boldsymbol{\gamma}$ is a random vector following an arbitrary distribution with mean $\boldsymbol{0}$ and variance $\tau\boldsymbol{I}_q$. Thus, a test of the null hypothesis $H_0 : \tau = 0$ is equivalent to testing $H_0 : \boldsymbol{\gamma} = 0$. For simplicity, we assume $\boldsymbol{\gamma} \sim N(\boldsymbol{0}, \tau\boldsymbol{I}_q)$.

In order to account for LD among SNPs in a gene and avoid estimation issues related to multicollinearity, we use a ridge regression approach to impose a penalty on $\boldsymbol{\theta}$ in the PQL proposed for GLMM [1]. However, the selection of a penalty parameter can be computationally demanding [2]. Thus, to expedite and incorporate the selection of a penalty parameter used in our proposed test, we specify $\boldsymbol{d}_1 = \boldsymbol{G}\boldsymbol{\theta}$, as a random effect in Equation (4), with $\boldsymbol{\theta} \sim N(\boldsymbol{0}, \sigma_\theta^2\boldsymbol{I}_q)$. By using this approach, we demonstrate later in Section 3.1, under the null model, the best linear unbiased predictor (BLUP) of $\boldsymbol{d}_1$ is equivalent to the ridge regression estimator. Denoting $\boldsymbol{d}_2 = \boldsymbol{K}\boldsymbol{b}$ and $\boldsymbol{d}_3 = \boldsymbol{S}\boldsymbol{\gamma}$, and assuming $\boldsymbol{d}_1$, $\boldsymbol{d}_2$ and $\boldsymbol{d}_3$ to be independent, we can write Equation (4) in the GLMM form as

$$\begin{aligned} g\left(\boldsymbol{\mu}^{\boldsymbol{d}}\right) &= \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{d}_1 + \boldsymbol{d}_2 + \boldsymbol{d}_3 \\ &= \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{d} \end{aligned} \tag{5}$$

with $\boldsymbol{Z} = [\boldsymbol{I}_n\ \boldsymbol{I}_n\ \boldsymbol{I}_n]$, where $\boldsymbol{I}_n$ is the $(n \times n)$ identity matrix and $\boldsymbol{d} = \left(\boldsymbol{d}_1^T, \boldsymbol{d}_2^T, \boldsymbol{d}_3^T\right)^T \sim N(\boldsymbol{0}, \boldsymbol{D})$, where $\boldsymbol{D} = \text{diag}\left\{\sigma_\theta^2\left(\boldsymbol{G}\boldsymbol{G}^T\right), \sigma^2\left(\boldsymbol{K}\boldsymbol{K}^T\right), \tau\left(\boldsymbol{S}\boldsymbol{S}^T\right)\right\}$.

Based on the penalized quasi-likelihood (PQL) [1], Lin [10] developed a VCT for independent subjects in the framework of a GLMM. To test the null hypothesis $H_0 : \tau = 0$, the VCT uses the score statistic

$$\mathcal{U}_\tau\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\pi}}\right) = \frac{1}{2}\left\{\left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}\right)^T \boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{\Sigma}^{-1}\left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}\right) - \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{S}^T\right)\right\}\Big|_{\widehat{\boldsymbol{\beta}},\,\widehat{\boldsymbol{\pi}}} \quad (6)$$

where $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\pi}}$ are the maximum likelihood (ML) estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\pi} = (\sigma_\theta^2, \sigma^2, \phi)^T$, respectively, under the null model as described in Section 3.1. In addition, $\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{d}_1 + \boldsymbol{d}_2 + \boldsymbol{\varepsilon}$ is the corresponding working vector, where $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \phi\boldsymbol{W}^{-1})$ and $\boldsymbol{W} = \mathrm{diag}\left\{\omega_{ij}/\left[\nu(\mu_{ij}^d)g'(\mu_{ij}^d)^2\right]\right\}$ is calculated under the null model, where $g'(.)$ denotes the derivate of function $g(.)$. The covariance matrix for the working vector $\tilde{\boldsymbol{Y}}$ is given by $\boldsymbol{\Sigma} = \sigma_\theta^2\boldsymbol{G}\boldsymbol{G}^T + \sigma^2\boldsymbol{K}\boldsymbol{K}^T + \phi\boldsymbol{W}^{-1}$.

Since $\boldsymbol{S}\boldsymbol{S}^T$ does not have a block diagonal structure, the score $\mathcal{U}_\tau\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\pi}}\right)$ cannot be written as a sum of N independent random variables, corresponding to the families. Therefore, the asymptotic distribution of $\mathcal{U}_\tau\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\pi}}\right)$ is not a normal distribution, in contrast to the VCT of Lin [10]. Instead, we follow the approach developed by Zhang and Lin [11], and propose, as score statistic, the first term in Equation (6), which corresponds to the quadratic form

$$U_\tau = U_\tau\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\pi}}\right) = \frac{1}{2}\left\{\left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}\right)^T \boldsymbol{\Sigma}^{-1}\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{\Sigma}^{-1}\left(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}\right)\right\}\Big|_{\widehat{\boldsymbol{\beta}}.\widehat{\boldsymbol{\pi}}}$$

To correct for bias, we use the restricted maximum likelihood (REML) estimators [1] in the GLMM framework to obtain $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\pi}}$ under the null hypothesis.

Zhang and Lin [11] showed that under $H_0 : \tau = 0$, U_τ follows approximately a mixture of one degree of freedom, independent chi-square distributions. However, for computational ease, we use the Satterthwaite method [15] to approximate the distribution of U_τ by a scaled chi-square distribution $\kappa\chi^2_\xi$, where the scale parameter κ and the degrees of freedom ξ can be calculated by equating the mean and variance of U_τ to those of $\kappa\chi^2_\xi$.

When REML estimates are used to calculate U_τ, Zhang and Lin [11] showed that the mean and variance of U_τ can be approximated, respectively, by

$$\mathrm{tr}\left(\boldsymbol{P}\boldsymbol{S}\boldsymbol{S}^T\right)\Big|_{\widehat{\boldsymbol{\pi}}} \quad \text{and} \quad \mathcal{I}_\tau = \frac{1}{2}\left\{\mathrm{tr}(\boldsymbol{P}\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{P}\boldsymbol{S}\boldsymbol{S}^T) - \boldsymbol{J}^T\boldsymbol{M}^{-1}\boldsymbol{J}\right\}\Big|_{\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\pi}}}, \text{ with}$$

$$\boldsymbol{J} = \begin{pmatrix} \mathrm{tr}[\boldsymbol{P}\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{P}\boldsymbol{G}\boldsymbol{G}^T] \\ \mathrm{tr}[\boldsymbol{P}\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{P}\boldsymbol{K}\boldsymbol{K}^T] \\ \hdashline \mathrm{tr}[\boldsymbol{P}\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{P}\boldsymbol{W}^{-1}] \end{pmatrix}, \quad \boldsymbol{M} = \left[\begin{array}{cc:c} \mathrm{tr}[\boldsymbol{P}\boldsymbol{G}\boldsymbol{G}^T\boldsymbol{P}\boldsymbol{G}\boldsymbol{G}^T] & \mathrm{tr}[\boldsymbol{P}\boldsymbol{G}\boldsymbol{G}^T\boldsymbol{P}\boldsymbol{K}\boldsymbol{K}^T] & \mathrm{tr}[\boldsymbol{P}\boldsymbol{G}\boldsymbol{G}^T\boldsymbol{P}\boldsymbol{W}^{-1}] \\ \mathrm{tr}[\boldsymbol{P}\boldsymbol{K}\boldsymbol{K}^T\boldsymbol{P}\boldsymbol{G}^T\boldsymbol{G}] & \mathrm{tr}[\boldsymbol{P}\boldsymbol{K}\boldsymbol{K}^T\boldsymbol{P}\boldsymbol{K}\boldsymbol{K}^T] & \mathrm{tr}[\boldsymbol{P}\boldsymbol{K}\boldsymbol{K}^T\boldsymbol{P}\boldsymbol{W}^{-1}] \\ \hdashline \mathrm{tr}[\boldsymbol{P}\boldsymbol{W}^{-1}\boldsymbol{P}\boldsymbol{G}^T\boldsymbol{G}] & \mathrm{tr}[\boldsymbol{P}\boldsymbol{W}^{-1}\boldsymbol{P}\boldsymbol{K}\boldsymbol{K}^T] & \mathrm{tr}[\boldsymbol{P}\boldsymbol{W}^{-1}\boldsymbol{P}\boldsymbol{W}^{-1}] \end{array}\right]$$

and $\boldsymbol{P} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{X}}\left(\tilde{\boldsymbol{X}}^T\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{\Sigma}^{-1}$.

Dashed lines in $\boldsymbol{J}$ and $\boldsymbol{M}$ represent the cases: (i) ϕ known, implying a (2×1) vector and (2×2) matrix, respectively and (ii) ϕ unknown, implying a (3×1) vector and (3×3) matrix, respectively.

As the mean and variance of $\kappa\chi^2_\xi$ are given by $\kappa\xi$ and $2\kappa^2\xi$, respectively, we obtain the equations $\mathrm{tr}(\widehat{\boldsymbol{P}}\boldsymbol{S}\boldsymbol{S}^T) = \kappa\xi$ and $\mathcal{I}_\tau = 2\kappa^2\xi$, where $\widehat{\boldsymbol{P}}$ denotes the matrix $\boldsymbol{P}$ evaluated in $\widehat{\boldsymbol{\pi}}$. By solving these equations, we demonstrate that $\kappa = \mathcal{I}_\tau/[2\,\mathrm{tr}(\widehat{\boldsymbol{P}}\boldsymbol{S}\boldsymbol{S}^T)]$ and $\xi = 2\left[\mathrm{tr}(\widehat{\boldsymbol{P}}\boldsymbol{S}\boldsymbol{S}^T)\right]^2/\mathcal{I}_\tau$.

Therefore, to test $H_0 : \tau = 0$, we propose the statistic

$$T_\tau = T_\tau\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\pi}}\right) = \frac{U_\tau\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\pi}}\right)}{\kappa} \quad (7)$$

which follows approximately a chi-square distribution with ξ degrees of freedom.

3.1. Null Model Estimation

Our proposed score test requires that we first fit the null model. Under the null hypothesis $H_0 : \tau = 0$, Equation (5) becomes

$$g\left(\boldsymbol{\mu}^{\boldsymbol{d}_0}\right) \quad = \quad \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{Z}_0\boldsymbol{d}_0 \tag{8}$$

where $\boldsymbol{d}_0 = (\boldsymbol{d}_1^T, \boldsymbol{d}_2^T)^T$, $\boldsymbol{Z}_0 = [\boldsymbol{I}_n\ \boldsymbol{I}_n]$ and $\boldsymbol{d}_0 \sim N(\boldsymbol{0}, \boldsymbol{D}_0)$ with $\boldsymbol{D}_0 = \text{diag}\{\sigma_\theta^2\left(\boldsymbol{G}\boldsymbol{G}^T\right), \sigma^2\left(\boldsymbol{K}\boldsymbol{K}^T\right)\}$.

To estimate the parameters in Equation (8), Breslow and Clayton [1] proposed a Fisher scoring solution that may be expressed as the iterative solution to the system

$$\begin{bmatrix} \tilde{\boldsymbol{X}}^T\boldsymbol{W}\tilde{\boldsymbol{X}} & \tilde{\boldsymbol{X}}^T\boldsymbol{W}\boldsymbol{Z}_0 \\ \boldsymbol{Z}\boldsymbol{W}\tilde{\boldsymbol{X}} & \phi\boldsymbol{D}_0^{-1} + \boldsymbol{Z}_0^T\boldsymbol{W}\boldsymbol{Z}_0 \end{bmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{d}_0 \end{pmatrix} \quad = \quad \begin{pmatrix} \tilde{\boldsymbol{X}}^T\boldsymbol{W}\tilde{\boldsymbol{Y}} \\ \boldsymbol{Z}_0^T\boldsymbol{W}\tilde{\boldsymbol{Y}} \end{pmatrix} \tag{9}$$

This system is equivalent to the so called Henderson equations [16] for computing the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ and the best linear unbiased predictor (BLUP) of $\boldsymbol{d}_0$.

By re-expressing $\boldsymbol{d}_0$ to $(\boldsymbol{d}_1^T, \boldsymbol{d}_2^T)^T$ in Equation (9), we obtain the system

$$\begin{bmatrix} \tilde{\boldsymbol{X}}^T\boldsymbol{W}\tilde{\boldsymbol{X}} & \tilde{\boldsymbol{X}}^T\boldsymbol{W} & \tilde{\boldsymbol{X}}^T\boldsymbol{W} \\ \boldsymbol{W}\tilde{\boldsymbol{X}} & \frac{\phi}{\sigma_\theta^2}(\boldsymbol{G}\boldsymbol{G}^T)^{-1} + \boldsymbol{W} & \boldsymbol{W} \\ \boldsymbol{W}\tilde{\boldsymbol{X}} & \boldsymbol{W} & \frac{\phi}{\sigma^2}(\boldsymbol{K}\boldsymbol{K}^T)^{-1} + \boldsymbol{W} \end{bmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{d}_1 \\ \boldsymbol{d}_2 \end{pmatrix} = \begin{pmatrix} \tilde{\boldsymbol{X}}^T\boldsymbol{W}\tilde{\boldsymbol{Y}} \\ \boldsymbol{W}\tilde{\boldsymbol{Y}} \\ \boldsymbol{W}\tilde{\boldsymbol{Y}} \end{pmatrix} \tag{10}$$

This new system is equivalent to using a ridge regression penalization for parameter $\boldsymbol{\theta}$ in Equation (4) (see Appendix A for details). Assuming that $\boldsymbol{\pi} = (\sigma_\theta, \sigma, \phi)^T$ is known, it can be shown that the solution of Equation (10) is given by the following equations:

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\tilde{\boldsymbol{X}}^T\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{Y}} \\ \widehat{\boldsymbol{d}_0} &= \begin{pmatrix} \widehat{\boldsymbol{d}_1} \\ \widehat{\boldsymbol{d}_2} \end{pmatrix} = \boldsymbol{D}\boldsymbol{Z}^T\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}}) \\ &= \begin{pmatrix} \sigma_\theta^2\left(\boldsymbol{G}\boldsymbol{G}^T\right)\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}}) \\ \\ \sigma^2\left(\boldsymbol{K}\boldsymbol{K}^T\right)\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\widehat{\boldsymbol{\beta}}) \end{pmatrix} \end{aligned} \tag{11}$$

with $\boldsymbol{\Sigma} = \sigma_\theta^2\boldsymbol{G}\boldsymbol{G}^T + \sigma^2\boldsymbol{K}\boldsymbol{K}^T + \phi\boldsymbol{W}^{-1}$. Chen et al. [17] fitted their GLMM by defining, $\boldsymbol{d}_s = (\boldsymbol{d}_1 + \boldsymbol{d}_2)$, and assuming $\boldsymbol{d}_s \sim N\left(\boldsymbol{0}, \left[\sigma_\theta^2\boldsymbol{G}\boldsymbol{G}^T + \sigma^2\boldsymbol{K}\boldsymbol{K}^T\right]\right)$. However, the BLUP for $\boldsymbol{d}_s$ is identical to the sum of BLUPs in Equation (11). Therefore, we use their PQL-based estimation algorithm implemented in the R package GMMAT (see Chen et al. [17] and Chen and Conomos [13] for details) to estimate our null model. The iterative process uses REML to estimate the variance parameters vector $\boldsymbol{\pi}$ used in the score statistic Equation (7).

4. Simulations

For our simulations, we used SimPed [18] to generate 100 SNPs (50 independent and 50 in LD) for 1000 independent families with identical pedigree structures of size 10 Figure 1. For each simulation, we randomly sampled without replacement 100 families to obtain a sample of 1000 individuals.

The environment was simulated to be correlated within a family and depend on age and sex of the subject using the following model with parameters chosen to mimic our real data example:

$$E_{ij} = 2 + 0.01Age_{ij} + 0.1I(Female_{ij}) + \gamma_i + \varepsilon_{ij} \tag{12}$$

where $I(\cdot)$ is the indicator function, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{i10})^T \sim N(\mathbf{0}, 4\boldsymbol{I}_{10})$ where $\boldsymbol{I}_{10}$ is the (10×10) identity matrix, and $\gamma_i \sim N(0, 4)$.

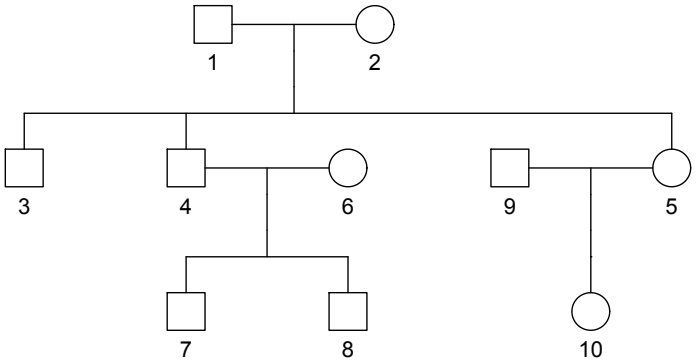

Figure 1. Pedigree for simulated data. Circle = Female, Square = Male.

In our simulations, we considered a SNP-set with 50 independent SNPs or in LD. The correlation structure for the SNPs in LD is shown in Figure 2. Using the `R` package `SimCorMultRes` [19], we simulated a correlated binary phenotype dependent on family using the following mean structure:

$$\begin{aligned}&\text{logit}\left[P\left(Y_{ij} = 1 | Age_{ij}, Female_{ij}, E_{ij}, G1_{ij}, G2_{ij}\right)\right] = 0.1 + 0.01Age_{ij} + 0.1I(Female_{ij}) \\ &+0.1E_{ij} + 0.3G1_{ij} + 0.3G2_{ij} + \gamma_1(G1_{ij} \times E_{ij}) + \gamma_2(G2_{ij} \times E_{ij}) + \alpha_{ij}\end{aligned}$$

with $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{i10})^T \sim N(\mathbf{0}, 2\sigma^2\boldsymbol{\Phi}_i)$, where $\boldsymbol{\Phi}_i$ is the kinship matrix corresponding to the family pedigree in Figure 1. Given `SimCorMultRes` only allows for specification of the correlation matrix, σ^2 is set equal to 1. $G1_{ij}$ and $G2_{ij}$ are either independent SNPs (MAF = 0.3 and 0.1) or the first and fifth SNPs with MAF = 0.3 and 0.17 respectively from Figure 2. Note that only the SNPs with a main effect interact with the environment in our model. We generated 10,000 and 1000 datasets to estimate type I error and empirical power, respectively, at an $\alpha = 0.05$ level. Using these datasets, we compared the performances of the score test, MinP test, and our proposed VCT. As previously mentioned, the score test treats γ as a fixed vector and results in a p DF test. The null model for this test was estimated as specified in Section 3.1. For the MinP test, which represents a single SNP analysis of GE interaction, we independently model the single SNP-by-environment interaction using Equation (4) where $\boldsymbol{G}$ and $\boldsymbol{S}$ are a vector, rather than a matrix, for a SNP and GE interaction, respectively. For each model, we calculated the p-value for the test of interaction and found the minimum p-value among all tests. We corrected for multiple testing by multiplying by the number of effective SNPs in the gene [20]. In the case of independence, the number of effective SNPs would be equivalent to the number of SNPs in the gene.

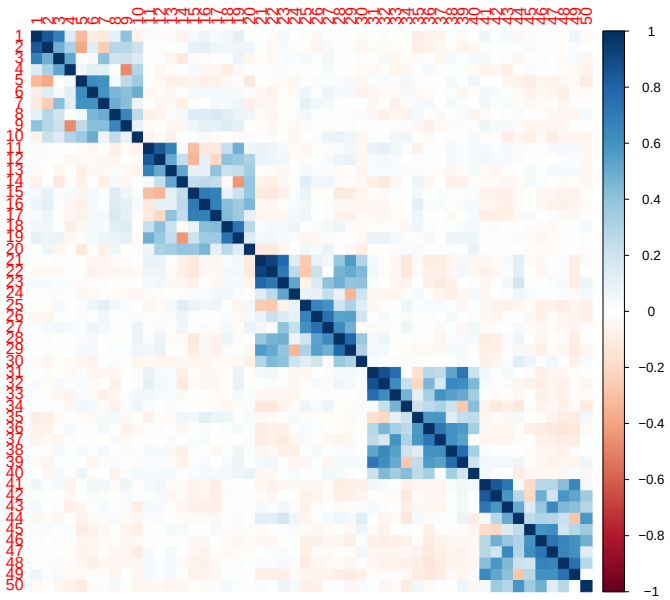

Figure 2. Correlation of the 50 simulated single nucleotide pollymorphisms (SNPs) in linkage disequilibrium (LD). The vertical color line to the right indicates the level of correlation between the SNPs. The dark blue (red) means high positive (high negative) correlation and the light blue (red) means low positive (low negative) correlation.

4.1. Type I Error

We first compared the empirical type I error of the different methods at 0.05 α-level. To evaluate type I error, we set $\gamma_1 = \gamma_2 = 0$ and varied the number of SNPs q in the gene. The SNPs were either independent or in LD. The empirical type I error rates are shown in Table 1 as well as the mean of the fitted $\hat{\sigma}^2$, $\hat{\sigma}^2_\theta$, and $\widehat{\lambda} = 1/\hat{\sigma}^2_G$ parameters from Equation (8) across all simulations. While the variance component for the random effect defined by the kinship matrix stays approximately constant, the penalization term $\widehat{\lambda}$ increases as the number of SNPs in the model increases. Thus, like in ridge regression, increasing the number of parameters results in an increased penalization of the model. All of the methods were conservative in our simulations, but as q increased, the score test became useless due to the large DF of the test.

Table 1. Type I error at 0.05 α-level for each method depending on the number of SNPs q and whether the SNPs are independent or in LD. Two of the SNPs in each scenario have a main effect.

SNPs Category	q	$\widehat{\sigma}^2$	$\widehat{\sigma}^2_\theta$	$\widehat{\lambda} = 1/\widehat{\sigma}^2_G$	Score	MinP	VCT
Independent	5	1.247	0.034	29.4	0.020	0.031	0.034
	10	1.240	0.017	58.8	0.023	0.026	0.024
	50	1.222	0.003	333	0.004	0.022	0.014
LD	5	1.243	0.021	47.6	0.025	0.026	0.031
	10	1.239	0.009	111	0.004	0.034	0.030
	50	1.222	0.002	500	0.000	0.028	0.022

4.2. Empirical Power for Independent SNPs

We next compared the empirical power of the different methods with either five, 10, or 50 independent SNPs. We varied the amount of interaction for the two selected SNPs by varying $\gamma_1 = \gamma_2$ from 0 to 0.1 by 0.01. Figure 3 shows that as the number of SNPs in the model increases, each of the methods lose power due to the increase in DF of each test. While the VCT performs best with five or

10 SNPs, the MinP test performs best for 50 SNPs because only two out of 50 SNPs have interaction. The MinP test will always perform best if very few of the SNPs in the set have interaction.

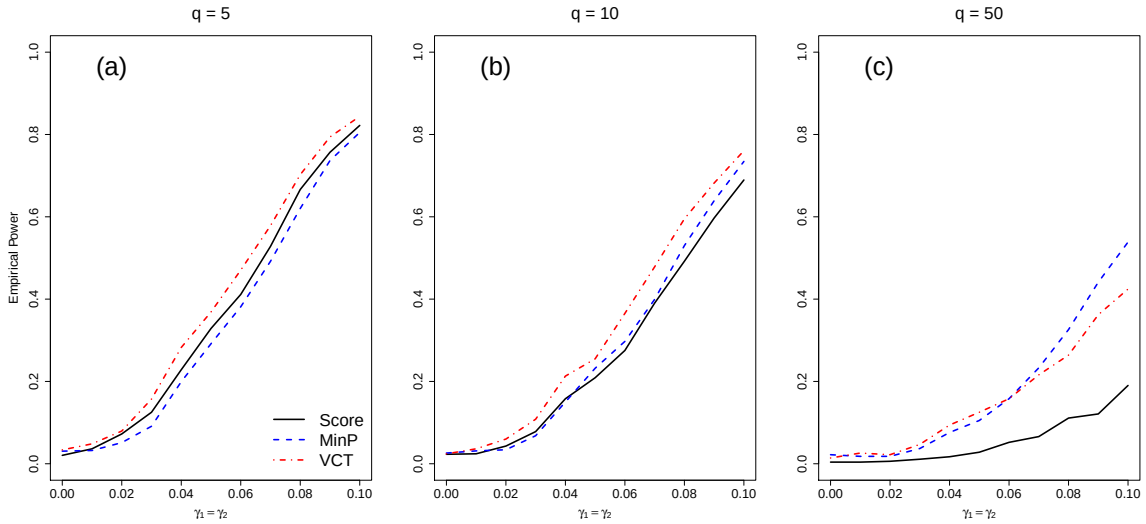

Figure 3. Empirical power at 0.05 α-level of the methods for q independent SNPs of which two of the SNPs have a main effect. The same two SNPs have an equal interaction with the environment. (**a**) $q = 5$ independent SNPs; (**b**) $q = 10$ independent SNPs; (**c**) $q = 50$ independent SNPs.

4.3. Empirical Power for SNPs in LD

Finally, we compared the empirical power of the different methods with either five, 10, or 50 SNPs in LD. We varied the amount of interaction for the two selected SNPs by varying $\gamma_1 = \gamma_2$ from 0 to 0.1 by 0.01. Figure 4 shows that as before, each of the methods lose power as q increases. The VCT outperforms the MinP test in all scenarios because the SNPs are correlated which the MinP test fails to account for.

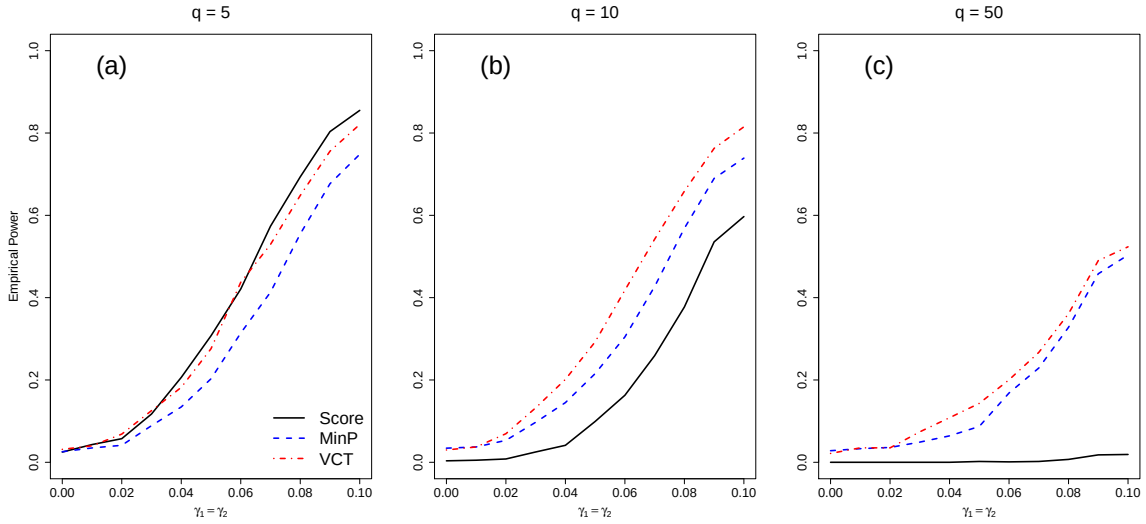

Figure 4. Empirical power at 0.05 α-level of the methods for q correlated SNPs of which two of the SNPs have a main effect. The same two SNPs have an equal interaction with the environment. (**a**) $q = 5$ correlated SNPs; (**b**) $q = 10$ correlated SNPs; (**c**) $q = 50$ correlated SNPs.

5. Application to Baependi Data

We use our proposed method to test for GE interactions in the Baependi Heart Study [14] between BMI and three different candidate genes that may be associated with type II diabetes (T2D). The first candidate gene we studied was the Peroxisome-Proliferator-Activated Receptors gamma (*PPARG*) gene, which is a key regulator of adipocyte differentiation and energy balance. Two of the mutations in the *PPARG* gene have been shown to be associated with obesity or diabetes-related phenotypes in different populations [21]. *PPARG2*, the predominantly isoform of *PPARG*, is expressed selectively and at a higher level in adipose tissue, where it modulates the expression of target genes implicated in adipocyte differentiation and glucose homeostasis [22]. Thus, the *PPARG2* gene is a major candidate gene for T2D or obesity, both being complex phenotypes determined by the combination of multiple genetic and environmental factors [23,24]. The second candidate gene studied was the Fat Mass and Obesity associated protein (*FTO*), which confers risk for obesity and BMI. Since obesity is known to be a predisposing factor for the development of T2D, it is not surprising that variants in FTO have been also found in T2D GWAS [25]. The final candidate gene studied was the cyclin-dependent kinase 5 regulatory subunit associated protein 1-like 1 (*CDKAL1*) gene which confers risk for obesity and T2D [26]. In our study, the *PPARG*, *FTO*, and *CDKAL1* genes had 16, 149, and 186 genetic variants genotyped, respectively. However these numbers of SNPs do not represent the number of effective SNPs discussed in Gao et al. [20], that is equivalent of the number of principal components to reach 99.5% of their total variation. Then, the effective number of SNPs associated with the *PPARG*, *FTO*, and *CDKAL1* genes are 10, 93, and 92, respectively. We used Equation (5) to specify a logistic GLMM to test for GE interaction of the aforementioned genes with BMI associated with T2D status (case/control). Our model included age, sex, and the first two principal components of the entire genotype data of Baependi data as covariates. Due to some individuals missing genotype information for some SNPs, the tests of each gene had different sample sizes. Table 2 describes the sample sizes (number of subjects and number of families) for cases and controls included in analysis of each gene.

Table 2. Summary of cases per subjects and families.

Gene	Subjects			Families		
	Control	Cases	Total	Control	Cases	Total
PPARG	845	83	928	43	42	85
FTO	712	71	783	47	38	85
CDKAL1	661	69	730	47	38	85

In Table 3, we report the p-values for each method. By comparing the p-values with respect to the corresponding α level, only the VCT identifies a significant GE interaction of BMI with *PPARG*. All other tests were non-significant for this gene as well as for other candidate genes.

Table 3. Sample size, GLMM parameters, p-values and execution times for the analysis of the Baependi dataset.

Gene	SNPs	Total Subjects	GLMM Parameters			Test	p-Value	α Level	Time (s)
			$\widehat{\sigma}^2$	$\widehat{\sigma}^2_\theta$	$\widehat{\lambda} = 1/\widehat{\sigma}^2_G$				
PPARG	16	928	0.4463	0.0029	344.8276	VCT	0.028	0.05	18.420
						MinP	0.019 *	0.005	100.261
						Score	0.595	0.05	9.025
FTO	149	783	0.3710	0.0033	303.0303	VCT	0.451	0.05	12.958
						MinP	0.031 *	0.0005	2675.907
						Score	0.992	0.05	6.197
CDKAL1	186	730	0.0918	0.0111	90.0901	VCT	0.635	0.05	9.907
						MinP	0.040 *	0.0005	1755.881
						Score	0.915	0.05	4.257

* Compare MinP test p-value with the corresponding corrected α, obtained by dividing 0.05 for the number of effective SNPs (which is equivalent to the number of principal components that reach 99.5% of the their total variation): 10 for *PPARG*, 93 for *FTO* and 92 for *CDKAL1*.

The variance estimates for families $\widehat{\sigma}^2$ and the ridge penalty $\widehat{\sigma}^2_\theta$ are reported in Table 3. In Section 4, we showed that the ridge penalty increases as the number of SNPs increases, however, our results for *PPARG*, *FTO* and *CDKAL1* suggest that λ also depends on the number of subjects. Finally, Table 3 also shows the execution times using the `R` version 3.3.1 and a processor Intel(R) Core(TM) i5-6500 CPU @ 3.20 GHz with a RAM memory 8.00 GB and operating system 64-bits. Computation times for the VCT and the score test were considerably lower than those for the MinP test. The time to compute each test increased with the increase in number of SNPs in a gene and number of subjects in the analysis.

6. Conclusions

We have proposed a variance component score test for testing for interactions between a set of SNPs in a gene and an environmental variable with family data. We specified the interaction coefficients as random variables with common variance and evaluated the null hypothesis that the variance is equal to zero. Given the LD among some SNPs in a gene, we fit the null model assuming the SNP coefficients as random effects and showed that the corresponding BLUP was equivalent to ridge regression estimator. This approach gives a direct estimation for the ridge penalization parameter in comparison with other computationally demanding procedures based on cross validation [2]. We compared, via simulations and a real data application, our approach with the so called MinP test and also with the traditional q degrees of freedom score test. The results showed that the proposed test is robust and performs well, with considerable power. Simulations and application presented in this paper were done assuming a binary phenotype and a continuous environmental variable, however, the GLMM admits phenotypes with distribution belonging to the exponential family. These other distributions are currently unexplored in this paper. In addition, it is possible to have multiple environmental factors as well as environmental factors that are discrete. The proposed model can easily incorporate these cases. It is important to note that the proposed model does not account for a possible correlation of the environment among family members. A possible extension of this work is to include in the GLMM a shared household factor by adding a random effect that follows a normal distribution with mean vector zero and with variance the matrix that characterizes household sharing. Finally, using GLMMs can be computationally intensive and may experience convergence issues. In the future, we plan to explore using generalized estimating equations as an alternative approach to testing for interactions in families.

Acknowledgments: The authors would like to thank the Brazilian National Counsel of Technological and Scientific Development (CNPq) for partially funding this work through the Genetics and Molecular Cardiology Laboratory at the Heart Institute, Medical School University of São Paulo and the Baependi Heart Study, the Doctoral Dissertation Fellowship from the University of Minnesota, Twin Cities, USA (Brandon J. Coombes),

and the COLCIENCIAS Institute Doctoral Fellowship, Bogotá, Colombia (Mauricio A. Mazo Lopera), and Mayo Foundation (Mariza de Andrade).

Author Contributions: Mariza de Andrade conceived and designed the research project; Mauricio A. Mazo Lopera wrote the R programs and Brandon J. Coombes and Mauricio A. Mazo Lopera analyzed the data; Mariza de Andrade. contributed materials and analysis tools; Mauricio A. Mazo Lopera, Brandon J. Coombes, and Mariza de Andrade wrote the paper.

Conflicts of Interest: None declared. The authors do not have conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LMM	Linear Mixed Model
GLMM	Generalized Linear Mixed Model
GE	Gene-Environment
SNP	Single Nucleotide Pollymorphism
BLUP	Best Linear Unbiased Predictor
PPARG	Peroxisome Proliferator Activated Receptor Gamma
GWA	Genome Wide Association
LD	Linkage Disequilibrium
VCT	Variance Components Test
ML	Maximum Likelihood
REML	Restricted Maximum Likelihood
PQL	Penalized Quasi-Likelihood

Appendix A. Ridge Regression and BLUP Equivalence in GLMM

Using the same notation of Section 2.2, where $\boldsymbol{d}_2 = \boldsymbol{Kb}$, the Equation (4) under the null hypothesis $H_0 : \gamma = \mathbf{0}$ becomes to

$$g(\boldsymbol{\mu}^{d_2}) = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{G}\boldsymbol{\theta} + \boldsymbol{d}_2$$

with $\boldsymbol{\mu}^{d_2} = E(\boldsymbol{Y}|\boldsymbol{d}_2)$. Following Breslow and Clayton [1], the penalized quasi-likelihood (PQL) for this model is given by

$$\begin{aligned} ql(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi_R, \sigma_R) &= -\frac{1}{2}\log\left|\frac{\sigma_R^2}{\phi_R}(\boldsymbol{K}\boldsymbol{K}^T)\boldsymbol{W}_R + \boldsymbol{I}_n\right| \\ &\quad + \sum_{i=1}^{N}\sum_{j=1}^{n_i} ql_{ij}(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi_R; \tilde{\boldsymbol{d}}_2) - \frac{1}{2}\tilde{\boldsymbol{d}}_2^T\left(\sigma_R^2\boldsymbol{K}\boldsymbol{K}^T\right)^{-1}\tilde{\boldsymbol{d}}_2 \end{aligned}$$

with $\tilde{\boldsymbol{d}}_2$ is choosen to maximize the sum of the last two terms, $\boldsymbol{W}_R = \text{diag}\left\{\omega_{ij}/\left[\nu(\mu_{ij}^{d_2})g'(\mu_{ij}^{d_2})^2\right]\right\}$ and

$$ql_{ij}(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi_R; \boldsymbol{d}_2) = \int_{Y_{ij}}^{\mu_{ij}^{d_2}} \frac{\omega_{ij}(Y_{ij} - \mu)}{\phi_R \nu(\mu)} d\mu$$

Subindex R denotes that the parameters are being estimated under Ridge regression method. We maximize the PQL with respect to $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and $\boldsymbol{d}_2 = \tilde{\boldsymbol{d}}_2$ jointly, obtaining the partial derivates

$$\begin{aligned} \frac{\partial ql(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi_R, \sigma_R)}{\partial \boldsymbol{\beta}} &= \frac{1}{\phi_R}\tilde{\boldsymbol{X}}^T\boldsymbol{W}_R\boldsymbol{\Delta}_R(\boldsymbol{Y} - \boldsymbol{\mu}^{d_2}) \\ \frac{\partial ql(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi_R, \sigma_R)}{\partial \boldsymbol{\theta}} &= \frac{1}{\phi_R}\boldsymbol{G}^T\boldsymbol{W}_R\boldsymbol{\Delta}_R(\boldsymbol{Y} - \boldsymbol{\mu}^{d_2}) \\ \frac{\partial ql(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi_R, \sigma_R)}{\partial \boldsymbol{d}_2} &= \frac{1}{\phi_R}\boldsymbol{W}_R\boldsymbol{\Delta}_R(\boldsymbol{Y} - \boldsymbol{\mu}^{d_2}) - \left(\sigma_R^2\boldsymbol{K}\boldsymbol{K}^T\right)^{-1}\boldsymbol{d}_2 \end{aligned} \quad \text{(A1)}$$

with $\boldsymbol{\Delta}_R = \text{diag}\left\{g'\left(\boldsymbol{\mu}^{d_2}\right)\right\}$ (see Chen et al. [17] for details). Ridge regression estimator of $\boldsymbol{\theta}$ is obtained by minimizing the function

$$[ql(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi_R, \sigma_R) \times \phi_R] - \frac{1}{2}\lambda\boldsymbol{\theta}^T\boldsymbol{\theta}$$

where λ is a penalizing factor and function $ql(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi_R, \sigma_R)$ is multiplied by ϕ_R to be consistent with the definition of ridge regression estimator under normality [27]. Therefore, the second line of Equation (A1) becomes

$$\frac{\partial ql(\boldsymbol{\beta}, \boldsymbol{\theta}, \phi_R, \sigma_R)}{\partial \boldsymbol{\theta}} = \boldsymbol{G}^T\boldsymbol{W}_R\boldsymbol{\Delta}(\boldsymbol{Y} - \boldsymbol{\mu}^{d_2}) - \lambda\boldsymbol{\theta}$$

and the system of equations is therefore

$$\begin{aligned} \tilde{\boldsymbol{X}}^T\boldsymbol{W}_R\boldsymbol{\Delta}_R(\boldsymbol{Y} - \boldsymbol{\mu}^{d_2}) &= \boldsymbol{0} \\ \boldsymbol{G}^T\boldsymbol{W}_R\boldsymbol{\Delta}_R(\boldsymbol{Y} - \boldsymbol{\mu}^{d_2}) - \lambda\boldsymbol{\theta} &= \boldsymbol{0} \\ \boldsymbol{W}_R\boldsymbol{\Delta}_R(\boldsymbol{Y} - \boldsymbol{\mu}^{d_2}) - \frac{\phi_R}{\sigma_R^2}\left(\boldsymbol{K}\boldsymbol{K}^T\right)^{-1}\boldsymbol{d}_2 &= \boldsymbol{0} \end{aligned} \tag{A2}$$

and left-multiplying by $(\boldsymbol{G}\boldsymbol{G}^T)^{-1}\boldsymbol{G}$ the second line of Equation (A2), we have

$$\begin{aligned} \tilde{\boldsymbol{X}}^T\boldsymbol{W}_R\boldsymbol{\Delta}_R(\boldsymbol{Y} - \boldsymbol{\mu}^{d_2}) &= \boldsymbol{0} \\ \boldsymbol{W}_R\boldsymbol{\Delta}_R(\boldsymbol{Y} - \boldsymbol{\mu}^{d_2}) - \lambda(\boldsymbol{G}\boldsymbol{G}^T)^{-1}\boldsymbol{G}\boldsymbol{\theta} &= \boldsymbol{0} \\ \boldsymbol{W}_R\boldsymbol{\Delta}_R(\boldsymbol{Y} - \boldsymbol{\mu}^{d_2}) - \frac{\phi_R}{\sigma_R^2}\left(\boldsymbol{K}\boldsymbol{K}^T\right)^{-1}\boldsymbol{d}_2 &= \boldsymbol{0} \end{aligned} \tag{A3}$$

But $\boldsymbol{\Delta}_R(\boldsymbol{Y} - \boldsymbol{\mu}^{d_2}) = \tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta} - \boldsymbol{G}\boldsymbol{\theta} - \boldsymbol{d}_2$, so the Equation (A3) becomes

$$\begin{aligned} \tilde{\boldsymbol{X}}^T\boldsymbol{W}_R\tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{X}}^T\boldsymbol{W}_R\boldsymbol{G}\boldsymbol{\theta} + \tilde{\boldsymbol{X}}^T\boldsymbol{W}_R\boldsymbol{d}_2 &= \tilde{\boldsymbol{X}}^T\boldsymbol{W}_R\tilde{\boldsymbol{Y}} \\ \boldsymbol{W}_R\tilde{\boldsymbol{X}}\boldsymbol{\beta} + [\boldsymbol{W}_R + \lambda(\boldsymbol{G}\boldsymbol{G}^T)^{-1}]\boldsymbol{G}\boldsymbol{\theta} + \boldsymbol{W}_R\boldsymbol{d}_2 &= \boldsymbol{W}_R\tilde{\boldsymbol{Y}} \\ \boldsymbol{W}_R\tilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{W}_R\boldsymbol{G}\boldsymbol{\theta} + \left[\boldsymbol{W}_R + \frac{\phi_R}{\sigma_R^2}\left(\boldsymbol{K}\boldsymbol{K}^T\right)^{-1}\right]\boldsymbol{d}_2 &= \boldsymbol{W}_R\tilde{\boldsymbol{Y}} \end{aligned} \tag{A4}$$

and denoting $\boldsymbol{d}_1 = \boldsymbol{G}\boldsymbol{\theta}$, we have the matricial representation of Equation (A4)

$$\begin{bmatrix} \tilde{\boldsymbol{X}}^T\boldsymbol{W}_R\tilde{\boldsymbol{X}} & \tilde{\boldsymbol{X}}^T\boldsymbol{W}_R & \tilde{\boldsymbol{X}}^T\boldsymbol{W}_R \\ \boldsymbol{W}_R\tilde{\boldsymbol{X}} & \boldsymbol{W}_R + \lambda(\boldsymbol{G}\boldsymbol{G}^T)^{-1} & \boldsymbol{W}_R \\ \boldsymbol{W}_R\tilde{\boldsymbol{X}} & \boldsymbol{W}_R & \boldsymbol{W}_R + \frac{\phi_R}{\sigma_R^2}(\boldsymbol{K}\boldsymbol{K}^T)^{-1} \end{bmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{d}_1 \\ \boldsymbol{d}_2 \end{pmatrix} = \begin{pmatrix} \tilde{\boldsymbol{X}}^T\boldsymbol{W}_R\tilde{\boldsymbol{Y}} \\ \boldsymbol{W}_R\tilde{\boldsymbol{Y}} \\ \boldsymbol{W}_R\tilde{\boldsymbol{Y}} \end{pmatrix}$$

which is equivalent to the Equation (10), where $\boldsymbol{d}_1$ is assumed as a normal random effect with mean $\boldsymbol{0}$ and variance $\sigma_\theta^2\boldsymbol{G}\boldsymbol{G}^T$ and, also, the penalization factor λ is identical to ϕ/σ_θ^2.

References

1. Breslow, N.; Clayton, D. Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.* **1993**, *88*, 9–25.
2. Lin, X.; Lee, S.; Chistiani, D.; Lin, X. Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics* **2013**, *14*, 667–681.
3. Lin, X.; Lee, S.; Wu, M.; Wang, C.; Chen, H.; Li, Z.; Lin, X. Test for rare variants by environment interactions in sequencing association studies. *Biometrics* **2016**, *72*, 156–164.
4. Coombes, B.; Basu, S.; Mcgue, M. A combination test for detection of gene-environment interaction in cohort studies. *Genet. Epidemiol.* **2017**, *41*, 396–412.

5. Wu, M.; Lee, S.; Cai, T.; Li, Y.; Boehnke, M.; Lin, X. Rare variant association testing for sequencing data using the Sequence Kernel Association Test (SKAT). *Am. J. Hum. Genet.* **2011**, *89*, 82–93.
6. Coombes, B. Tests for Detection of Rare Variants and Gene-Environment Interaction in Cohort and Twin Family Studies. Ph.D. Thesis, University of Minnesota, Minneapolis, MN, USA, 2016.
7. Bishop, C.; Tipping, M. Variational Relevance Vector Machines. In Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, Stanford, CA, USA, 30 June–3 July 2000; pp. 46–53.
8. Bishop, C.; Tipping, M. Bayesian regression and classification. *Nato Sci. Ser. Sub Ser. III Comput. Syst. Sci.* **2003**, *190*, 267–288.
9. Wang, T.; He, P.; Ahn, K.; Wang, X.; Ghosh, S.; Laud, P. A re-formulation of generalized linear mixed models to fit family data in genetic association studies. *Front. Genet.* **2015**, *6*, 120.
10. Lin, X. Variance component testing in generalised linear models with random effects. *Biometrika* **1997**, *84*, 309–326.
11. Zhang, D.; Lin, X. Hyphotesis testing in semiparametric additive mixed models. *Biostatistics* **2003**, *4*, 57–74.
12. Shen, X.; Alam, M.; Fikse, F.; Ronnegard, L. A novel generalized ridge regression method for quantitative genetics. *Genetics* **2013**, *193*, 1255–1268.
13. Chen, H.; Conomos, M. *GMMAT: Generalized Linear Mixed Model Association Tests*; R Package Version 0.7; 2016. Available online: https://content.sph.harvard.edu/xlin/dat/GMMAT_user_manual_v0.7.pdf (accessed on 16 January 2017).
14. Oliveira, C.; Pereira, A.; de Andrade, M.; Soler, J.; Krieger, J. Heritability of cardiovascular risk factors in a brazilian population: Baependi heart study. *BMC Med. Genet.* **2008**, *9*, 32.
15. Satterthwaite, F. Synthesis of variance. *Psychometrika* **1941**, *6*, 309–316.
16. Henderson, C. Best linear unbiased estimation and prediction under a selection model. *Biometrics* **1975**, *31*, 423–447.
17. Chen, H.; Wang, C.; Conomos, M.; Stilp, A.; Li, Z.; Sofer, T.; Szpiro, A.; Chen, W.; Brehm, J.; Celedon, J.; et al. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *Am. J. Hum. Genet.* **2016**, *98*, 653–666.
18. Leal, S.; Yan, K.; Muller-Myhsokb, B. SimPed: A Simulation Program to Generate Haplotype and Genotype Data for Pedigree Structures. *Hum. Hered.* **2005**, *60*, 119–122.
19. Touloumis, A. Simulating Correlated Binary and Multinomial Responses under Marginal Model Specification: The SimCorMultRes Package. *R J.* **2016**, *8*, 79–91.
20. Gao, X.; Starmer, J.; Martin, E. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet. Epidemiol.* **2008**, *32*, 361–369.
21. Spiegelman, B. PPAR-gamma: Adipogenic regulator and thiazolidinedione receptor. *Diabetes* **1998**, *47*, 507–514.
22. Latruffe, N.; Vamecq, J. Peroxisome proliferator activated receptor (PPARs) as regulators of lipid metabolism. *Biochimie* **1997**, *79*, 81–94.
23. Scott, L.; Mohlke, K.; Bonnycastle, L.; Willer, C.; Li, Y.; Duren, W.; Erdos, M.; Stringham, H.; Chines, P.; Jackson, A.; et al. A genome wide association study of Type 2 Diabetes in Finns detects multiple susceptibility variants. *Science* **2007**, *316*, 1341–1345.
24. Kevin, J.; Mathew, E.; Qianghua, X.; Struan, F. Genetic Susceptibility to Type 2 Diabetes and Obesity: Follow-up of Findings from Genome-Wide Association Studies. *Int. J. Endocrinol.* **2014**, *2014*, 769671.
25. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. *Nature* **2007**, *447*, 661–678.
26. Wood, A.; Tyrrell, J.; Beaumont, R.; Jones, S.; Tuke, M.; Ruth, K.; Yaghootkar, H.; Freathy, R.; Murray, A.; Frayling, T.; et al. Variants in the FTO and CDKAL1 loci have resessive effects on risk of obesity and type 2 diabetes, respectively. *Diabetologia* **2016**, *59*, 1214–1221.
27. Vlaming, R.; Groenen, P. The current and future use of ridge regression for prediction in quantitative genetics. *BioMed Res. Int.* **2015**, *2015*, 143712.