An Investigation of the use of Linear Mixed Models under an Extreme Phenotype Sampling (EPS) design

Maryam Yetunde Onifade

Thesis submitted to the Faculty of Science in partial fulfillment of the requirements
for the degree of
Doctorate in Philosophy Mathematics and Statistics[1]

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

[1]The Ph.D. program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Abstract

Mixed models have been used in genome-wide association studies to correct for confounding by population stratification and other forms of hidden relatedness. This class of models includes linear mixed models (LMMs) and generalized linear mixed models (GLMMs). This thesis presents an investigation into the use and application of LMMs within the context of extreme phenotype sampling (EPS) designs where genetic covariates are missing for some participants since genotypes are only collected on samples having extreme response variable values.

We begin by exploring whether existing mixed model approaches correct for population stratification under an EPS design. These methods have been previously investigated with both continuous and case/control response variables. However, they have not been investigated in the context of EPS designs. We assess the performance of three mixed model approaches suitable for binary traits (GMMAT, LEAP and CARAT) and one linear mixed model approach (GEMMA) for continuous traits. Our investigation includes an overview of mixed model methodology applicable to binary response variables. We assess type 1 error rates and power using simulation studies with both common and rare variants scenarios. As a practical application of these mixed model techniques, we also compared methods when applied to a prostate cancer dataset collected as part of the PROtEUs study conducted in Québec, Canada

that is known to have population substructure. Our simulation results show that for a common candidate variant, both LEAP and GMMAT had type 1 error rate close to the nominal value and similar power. Similar type 1 error control was observed with the analysis on the PROtEUs dataset. However, for rare variants the false positive rate remains inflated even after correction with mixed model approaches.

Next, we present an Expectation Maximization (EM) algorithm for fitting linear mixed models with missing genetic covariates that was motivated by EPS designs. We used the method of weights adapted for linear mixed models to handle the missing genotypes. We derive two hypothesis tests for genetic association, a likelihood ratio test using importance sampling and a Monte-Carlo based Wald test.

The performance of our algorithm was then assessed. Simulation studies were used to estimate type 1 error and power. We observed type 1 error rates below the nominal values of 0.05, signifying a conservative test, and low power for all missing data scenarios considered. Moreover some point estimates appear biased. We applied our algorithm to analyze the PROtEUs dataset and although our algorithm was able to correctly estimate most of the model parameters, the genetic effect estimated using the EM approach was larger than values by other approaches. The false positive rate also seemed inflated based on the p-value distribution across 5000 genetic markers. More investigation is needed to ensure the EM-based procedure is a valid approach to handle missing genotype data, particularly from an EPS study.

Acknowledgements

I begin by expressing my deepest gratitude to Almighty Allah for guiding me through this long journey and granting me strength and perseverance.

I am deeply grateful to my supervisors, Dr. Kelly Burkett and Dr. David Sankoff. I especially thank Dr. Kelly Burkett for the opportunity to be her first PhD student. Her support, understanding, and kindness, especially during the most challenging times of my life, have been instrumental in my progress. I am particularly appreciative of her advice on managing stress and maintaining a balance between my health and the demands of the program. Thank you for closely collaborating with me at every stage of this journey. Your unique style of supervision has made a huge impact on me, and I have learned so much from your guidance. I also extend my sincere thanks to Dr. David Sankoff for his valuable contributions and support throughout my research. His insights and feedback have significantly enriched this thesis.

To my beloved Dad, Mum, and siblings: Your unwavering support and belief in me, even in moments of self-doubt, have been my pillars of strength. I am eternally grateful for your constant encouragement, prayers, and love.

I extend heartfelt thanks to my fellow students in the Department of Mathematics and Statistics—Diane, Clemonell, Yousouph, Sofiat, Jason, and Julie—for their friendship, companionship, and the cherished conversations over lunch. I am also

deeply grateful to Najla Aloraini, whose work was incredibly instrumental in my research. Thank you for sharing important research tips and advice; they provided the much-needed motivation that influenced my progress. A special thank you to Diane Demers, whose continuous support has been a source of encouragement since the beginning of my program. Her assistance and kindness have not gone unnoticed.

To my dear friend, sister, and unpaid therapist, Abisola, I am incredibly grateful for your readiness to babysit Ameer, even at the most inconvenient hours, and for listening when I just needed to be heard. Your support has been a great help and comfort to me.

Finally, to my dear son Ameer, thank you for being the sweetest little companion on this journey. You are the most understanding son I could ever wish for, and your patience and love, especially during the times when I was away or preoccupied, have meant the world to me. Your resilience, joy, and constant hugs have been my daily inspiration.

Contents

List of Figures

List of Tables

Notation used

- a scalar
- $\boldsymbol{a}$ vector
- $\boldsymbol{A}$ matrix
- $\boldsymbol{I}_n$ $n \times n$ identity matrix
- $\boldsymbol{a}'$ transpose of a vector $\boldsymbol{a}$
- $\boldsymbol{A}'$ transpose of a matrix $\boldsymbol{A}$
- log natural logarithm
- $|\boldsymbol{A}|$ determinant of a matrix $\boldsymbol{A}$
- tr trace
- E expectation
- V variance
- Cov covariance
- N Gaussian distribution

- L likelihood function
- ℓ log-likelihood function
- $\frac{\partial}{\partial \boldsymbol{\theta}}$ partial derivative with respect to the parameter vector $\boldsymbol{\theta}$.

Abbreviations

- EPS: Extreme Phenotype Sampling
- MAF: Minor Allele Frequency
- GWAS: Genome-Wide Association Study
- SNP: Single Nucleotide Polymorphism
- PC/PCA: Principal Components/Principal Component Analysis
- EM: Expectation-Maximization algorithm
- LMM: Linear Mixed Model
- GLMM: Generalized Linear Mixed Model
- LTM: Liability Threshold Model
- MAP: Maximum a posteriori estimate
- PML: Posterior mean of the Multivariate Liability
- MCMC: Markov chain Monte Carlo
- GRM: Genetic Relationship Matrix
- BMI: Body Mass Index

Chapter 1

Introduction and Background

In genetic studies involving human populations, researchers are interested in how genetic variation contributes to diseases. The Genome Wide Association Study (GWAS), which involves genotyping a large number of individuals at hundreds of thousands to millions of genetic markers has proved very useful in discovering the relationship between common variants (minor allele frequency $> 5\%$) and complex diseases [76].

GWAS have identified numerous genetic variants associated with many traits, for example cardiovascular diseases (coronary artery diseases and hypertension) and cancer (prostrate, breast, colorectal and lung cancer) [77]. Genetic variants associated with autoimmune diseases such as rheumatoid arthritis have also been identified by GWAS [76]. Though important discoveries have been made, variants identified through GWAS only explain a fraction of the variation that is thought to be due to genetic factors [88] (missing heritability). For this reason, researchers have explored other genetically inherited factors to explain the missing heritability. For example, rare variants - variants with a minor allele frequency of $< 1\%$ have been identified as

a possible explanation for the missing heritability.

Due to next generation sequencing or exome sequencing, more rare variants have been discovered and investigated in genetic association studies [78]. Next generation sequencing is expensive and after it was introduced, new study designs were explored to reduce costs while maintaining power even at lower sample sizes [30, 39]. One cost-saving approach is Extreme Phenotype Sampling (EPS), which is a design that is applicable when the continuous response variable (phenotype) has been measured on a cohort of individuals [36, 39]. Sequencing is only done on those in the extreme tails of the response distribution. This is motivated by the assumption that individuals who have phenotypic extremes in the population are more likely to have the genetic variants that influence the phenotype. This method was shown to be cost effective when considering the costs of genotyping a large cohort [35]. Recently, EPS has proved very important in rare variant studies. Emond et al. [17] have applied it to exome sequencing, Guey et al. [20] showed that sampling based on strict phenotypic extremes was likely to result in a more powerful test for rare variants and Barnett et al. [4] have also shown that EPS has better power to detect rare variants than random sampling.

Although the recent interest in EPS was motivated by the need to reduce sequencing costs, this is not a new study design. EPS was first proposed in quantitative genetics by Lander and Botstein [32] who referred to it as *selective genotyping* since it selectively samples individuals with "abnormal" phenotypes. Lander and Botstein used selective genotyping in the context of mapping quantitative trait loci (QTL).

In genetic study designs, population stratification describes differences in allele frequencies of genetic variants among members of a different strata within a larger population. Strata here is used to refer to members of certain genetically differentiated

subgroups (e.g ethnic or racial). Confounding from population stratification occurs when false associations are discovered between genotypes and phenotypes due to an association between ancestry (strata) and phenotype. In population based genetic association designs, we need to account for the confounding variable (ancestry) in the statistical association analyses to avoid false positives. A number of approaches have been developed to correct for confounding due to population stratification in population based study designs. These include Genomic Control [15], Structured association [59], Principal Component Analysis (PCA) [54] and the more recent Linear Mixed Effects models (LMMs) described in Yang et al. [81]. We gave a brief description of these approaches in chapter 2.

In this thesis, we explore the problem of correcting for population stratification in EPS studies using LMMs. In Chapter 1, we provide the genetic background and fundamental terminology essential for comprehending the subsequent chapters. In Chapter 2, we describe our investigation of existing mixed model approaches for correcting population stratification in the context of extreme phenotype sampling. We assessed their efficacy in terms of the type 1 error and power when the data comes from an EPS study design. In Chapter 3, we present a novel approach to fitting mixed models to data from genetic studies in order to correct for population stratification when genetic covariate data are missing. The approach uses the Expectation Maximization (EM) algorithm and is applicable to data that are missing at random, which includes EPS data. In Chapter 4, we evaluate the performance of the approach proposed in Chapter 3 in detecting and correcting the false positive rate through simulation studies and a real data analysis.

1.1 Genetic background

The *human genome* consists of the basic biological material that is transmitted from parents to offspring. The human genome is comprised of 23 pairs of chromosomes. A *genetic locus* refers to a particular location in the genome; if that locus contains more than one variant form it is called *polymorphic*. The alternate forms of a locus are called the *alleles*. An individual's *genotype* refers to the pair of alleles that were inherited from that individual's mother and father. If the two alleles are both the same type that locus is said to be homozygous; otherwise it is heterozygous. For example, a locus with two alleles labelled A and a will have three possible genotypes: AA, Aa and aa where AA and aa are homozygous and Aa is heterozygous. In modern genetic studies, we often only consider loci having two alleles (called biallelic loci) and we code the genotypes by the number of minor alleles. For example we would code AA, Aa and aa as $0, 1$ or 2 copies of the minor allele a.

Genetic markers or *markers* is a term used to describe polymorphic loci throughout the genome that have been selected for genotyping. *Single nucleotide polymorphism (SNP)* is a commonly used type of genetic marker that typically only has two alleles and the minor allele frequency greater than 1%.

Phenotypes are the observable physical traits exhibited by an individual influenced by genetic factors. Phenotypes can either be represented as categorical, binary or quantitative. For binary traits, the phenotype Y will take on values 0 or 1 representing the unaffected and affected categories, respectively. For many human traits, there is a complex relationship between the genotype and phenotype because the phenotype may depend on interactions between different genetic variants and on interactions between genetic factors and the environment [21].

Genetic Models

In probabilistic genetic models, the genotypes influence the probability of disease. A genetic model can be defined as the probability distribution for a trait conditional on the underlying genotype at a disease influencing locus.

The *penetrance function* is a set of conditional probabilities which determines the probabilistic effect of the individual's genotype on their phenotype. i.e $Pr(Y|G)$. If we consider the a allele to be the disease predisposing allele for a binary trait, then $Pr(Y = 1|G = 0)$ represents the background probability of disease for someone with no copies of the disease predisposing allele. For quantitative traits or continuous traits, the penetrance function is replaced by a set of conditional density functions $f(Y|G)$. The normal density, with mean that depends on the genotype while the variance does not, is a natural choice for modelling the distribution of a quantitative trait. Hence, we assume that $Y \sim N(\mu_G, \sigma^2)$.

The mode of inheritance refers to the way in which the mean of the phenotype (quantitative) or the probability of disease (binary) changes based on the number of pre-disposing alleles.

- The mode of inheritance is recessive if

$$Pr(Y = 1|G = 1) = Pr(Y = 1|G = 0) \text{ (Binary) or}$$

$$\mu_{G=1} = \mu_{G=0} \text{ (Quantitative)}$$

  and G is the genotype at the disease predisposing locus.

- The dominant mode is used when only one copy of the disease allele is required

to induce an effect on the phenotype; i.e

$$P(Y = 1|G = 2) = P(Y = 1|G = 1) \text{ (Binary) or}$$

$$\mu_{G=2} = \mu_{G=1} \text{ (Quantitative)}$$

- For the additive mode,

$$P(Y = 1|G = 2) = 2P(Y = 1|G = 1) \text{ (Binary) or}$$

$$\mu_{G=2} = 2\mu_{G=1} \text{ (Quantitative)}$$

- The codominant mode makes no assumptions about the relationship between the three penetrance functions only that they are different: i.e

$$P(Y = 1|G = 2) \neq P(Y = 1|G = 1) \text{ (Binary) or } \mu_{G=2} \neq \mu_{G=1} \text{ (Quantitative).}$$

1.2 Genetic Association

Studies to find genetic factors that influence human traits can be classified as linkage studies or association studies. Association studies can be further classified into family-based or population-based sampling. Figure 1.1 illustrates the framework of genetic association studies, showcasing the division between family-based and population-based designs, with the latter further categorized into case-control and cohort studies.This work focuses on the population based association study and therefore background about this type of approach is now provided.

An association study tests for a statistical association between a genotype at

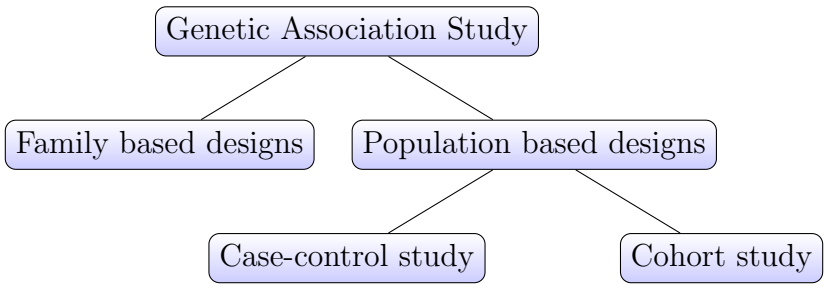


Figure 1.1: Flowchart of Genetic association designs

a locus in the genome and the phenotype. In population-based association studies, samples are independently drawn from a population using standard genetic epidemiological study designs. They can be classified into case-control studies and cohort studies. In case-control studies, samples are recruited based on their disease status, with cases being affected and controls unaffected. On the other hand, cohort studies recruit and follow a group of individuals sharing a common characteristic or experience over a defined period. They are often prospective in nature, for example a birth cohort initiated to investigate the development of diseases or outcomes of interest within a specific population.

Statistical methods are used to test for association between the genotype and phenotype. For binary traits, tests like the chi-square test and logistic regression can be used. ANOVA or linear regression approaches can be used when the phenotype is continuous. In practice, researchers test for association at many genetic markers. For example, in a GWAS, hundreds of thousands to millions of genetic markers might be genotyped and tested for association. To account for multiple testing, a genome-wide level of significance of 5×10^{-8} is used. Markers significantly associated with the trait are thought to be close to the true underlying but probably unobserved genetic variants that influence the disease. However, as will be discussed in the next section, population stratification can cause false positive associations in population

based association designs.

1.2.1 Population Stratification in Genetic-Association Studies

Population stratification occurs when individuals from genetically differentiated subpopulations, such as different ethnicities are sampled. Confounding from population stratification can occur when individuals from the different ancestries also have different phenotype distributions. Any locus with allele frequencies that also differ between subpopulations may be significantly associated with the phenotype. This leads to high false positive associations between genotype and phenotype [8]. Confounding from population stratification has been a factor in studies like Campbell et al. [8] and Knowler et al. [29]. Knowler et al. [29] demonstrated that a strong negative association between a GM gene variant and type-II diabetes was caused by confounding due to those with indigenous ancestry having both higher rates of type-II diabetes and a particular GM variant. Campbell et al [8] showed that a SNP with varying allele frequency across Europe was strongly associated with height; this apparent association was due largely to stratification as rematching individuals on the basis of ancestry greatly reduced the association.

As confounding due to population stratification is a known problem in genetic association studies, statistical approaches to correct for it have been proposed. An extensive review of the existing methods to correct for population stratification in the literature has been discussed by Chengqing et al. [79]. The first approach developed is called Genomic Control and it adjusts the association statistic by an inflation factor [15]. A second class of approaches use SNP data to infer ancestry in some way; structured association [60] and principal components analysis [57] are two examples.

Mixed effects model that incorporate the full covariance structure of the individuals have become popular recently [27, 84, 10]. More detail is now provided on each class of approach.

The Genomic Control (GC) approach makes use of a variance inflation factor, λ, which captures the extent to which test statistics are inflated [16]. The association statistic is computed on a set of neutral genetic markers. The set of statistics are used to compute tλ which is then used to scale the test statistic and adjust the p-value. A critique of GC is that it uses the same variance inflation factor as correction for all markers [57]; as some markers have larger differences in allele frequencies between populations than other markers, this constant correction may over-correct at some loci and under-correct at others. Hence, a constant inflation factor may be insufficient for markers which are highly differentiated across the population.

Structured association is implemented in a program called STRUCTURE [60] and it uses a Bayesian approach to assign individuals to subpopulations. An individual's posterior probability of being in a subpopulation is then included in the analysis to account for population stratification. It is an effective method for correcting for population stratification in datasets where population structure has been detected, but is computationally prohibitive with large datasets [58].

Principal Component Analysis (PCA) is a general statistical tool used in dimension reduction that has been applied to infer population structure in genetic data. To carry out PCA, an orthogonal transformation of the data is carried out: the eigenvalues and eigenvectors are computed and the largest eigenvector in the direction of the maximum variance is denoted as the first principal component. PCA was first explored in a population genetic context to detect and interpret underlying population structure in large datasets [45, 49]. PCA is also used to correct for population strati-

fication in association studies by including the top PCs in the model when computing asociation statistics [57].

Finally, mixed models are becoming popular for tackling the problem of population stratification in genetic studies because they are able to model population subdivision as well as hidden family structure and cryptic relatedness [86]. Mixed modeling uses a combination of fixed and random effects to model the phenotypes in the sample. A genetic similarity matrix is used to capture the pairwise relatedness among individuals and it is included as part of a random variance component in the mixed effects model. The fixed effects are usually the candidate SNPs and optional covariates. The mixed effects approach is a very appealing method but it is also computationally intensive. Computational improvements have been implemented that make the approach feasible on genome-wide data. These have been implemented in software such as EMMAX [27], FaSTLmm [37], BOLT-LMM [40] and more recently lme4qtl [87], lmekin [72] and Gaston [5]. More detailed information about mixed effects models are provided in Section 1.4.

1.3 Extreme Phenotype Sampling (EPS)

Due to the considerable costs associated with emerging genotyping technologies, genotyping a substantial number of samples might be economically unfeasible. An alternative that has been repeatedly proposed is "selective genotyping" of samples - a term initially introduced in quantitative genetics by Lebowitz et al. [33] and Lander and Botstein [32].

Selective genotyping refers to only genotyping subjects in a cohort whose phenotypes are in the extremes of the response distribution. Lander and Botstein proposed

the use of selective genotyping in the context of mapping quantitative trait loci using animal studies requiring multiple generations of animals to be reared and using linkage analysis. They used the term selective genotyping to mean raising a large population of animals but genotyping only those having phenotype that deviated substantially from the mean. Genotyping the extreme samples was motivated by the fact that some animals in the sample will provide more linkage information than others and, generally, those that contribute the most have genotypes that can be most clearly inferred from their phenotype. They showed that the strategy of selective genotyping will substantially increase efficiency whenever raising and phenotyping additional animals requires less effort than completely genotyping all available animals. They also showed that the method of selective genotyping had high power relative to completely genotyping the whole sample.

The work by Lander and Botstein formed the basis for the use of selective genotyping in genetic studies and since then several authors have used the method either directly or adapted it for uses beyond linkage analysis. For example, Dervasi et al. [12] showed that in the linkage analysis context, it is not important to sample more than the upper and lower 25% in a population. They also found that selective genotyping would greatly reduce the genotyping costs even in cases where costs of sampling and phenotyping are very great.

EPS was later proposed in the context of association studies. For example, EPS was used in Ball et al. [3] to test for the association between general cognitive ability as a behavioural trait and variation in candidate genes and in Vermissen et al [75] to identify genetic risk variants for coronary heart disease.

Gestel et al. [74] explored whether EPS is statistically more powerful than using the whole distribution when analyzing quantitative traits. In their study, a population

measured on a specific quantitative trait was modeled using a dominant/recessive model and an additive model. A bi-allelic genetic variant was assumed. Under the dominant/recessive model, the phenotypes for the two genotype classes were assumed to be normally distributed with mean that depended on the genotype, while for the additive model, three normally distributed subgroups with different means were assumed. Thus the marginal distribution of the phenotype was not itself normal but an admixture of normals. Subsamples consisting of the 2.5%, 5%, and 10% extremes were created and contingency tables of genotype by high/low phenotype category were analysed. Odds ratios were computed for scenarios having varying parameters such as the total sample size of the phenotyped group, allele frequency of the A alelle (0.1 to 0.9 in steps of 0.1), the mode of inheritance and the proportion of variance explained by the quantitative trait loci. As expected, the computed odd ratios increased as the extreme phenotype cutoff decreased; higher power is therefore expected.

For samples collected under the EPS design, the most straightforward approach for analysis is to consider the extremes as a binary trait; chi-square tests or logistic regression are often used to analyze such data. For example, Emond et al. [17] used logistic regression to test for association between the phenotype groups and variant scores collapsed by the gene. We will now review some additional analysis approaches that have been proposed specifically to handle EPS data.

1.3.1 EPS as a binary trait

Slatkin et al. [64] proposed a procedure for analyzing a slightly different design that consisted of samples from an extreme and a random sample from the same cohort. He developed two tests and combined the p-values from these tests to form an overall test. The first test compared the allele frequencies between the selected samples and

a random sample from the population and the second test compares the mean trait values among individuals with different genotypes using only those in the selected sample. Slatkin compared his test to a t-test and showed that his tests were more powerful than when the t-test was applied to a random sample with the same number of individuals. Chen et al. [11] improved Slatkin's design by replacing the random samples with a sample of individuals with unusually low trait values (i.e., an EPS design). Through simulation studies, they showed that their test was more efficient than Slatkin's.

Guey et al. [20] proposed an alternative approach to select extremes based on liability scores. Liability scores are computed by fitting a regression model to the data to correlate phenotype with observed non-genetic risk factors available on the full sample. Using this model, each individual is then assigned a predicted value, which is contrasted with the true phenotype to obtain a model residual called the liability score. If the phenotype is disease status, the liability scores are defined as the Pearson residuals [1]:

$$\frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}$$

where y_i is the observed disease status and $\hat{\pi}_i$ is the predicted probability of disease from the model for individual i. The individuals are then ranked in terms of their liability scores and the extremes are selected for sequencing. The selection of extreme of the liability scores is similar to the selection of extremes of the quantitative traits as carried out by Lander and Botstein [32] and Van et al. [74]. To estimate the power to detect a genetic association using a two tailed Fisher's exact test, they computed the proportion of simulations that gave rise to a rejection of the null hypothesis for a specified significance level. The simulation study showed results similar to [74]. Higher liability thresholds systematically increased the frequency of the genetic risk factor

in the affected individuals and decreased the frequency in unaffected individuals.

In the studies reviewed here, the extremes were treated as two groups and therefore the phenotype was binary. Treating the extremes as "cases" and "controls"" and using methods for binary response variables is a simple technique that is commonly used in this context. It is tempting to use the raw phenotype values measured on the extreme groups in subsequent analyses. However, Lin et al. [36] showed that using the actual trait values in an analysis without first accounting for the sampling could lead to increased type-1 error.

1.3.2 EPS analysed as a quantitative trait

Lin et al. [36] developed methods of analysis under EPS based on a primary trait observed on a cohort. Their approach also applied to studies where secondary traits were also measured on the same cohort. Genotypes were only measured on extremes based on the primary trait. Their approach employed an EM algorithm [13] to compute the maximum likelihood estimates of effects of genotype on primary and secondary traits. EM is usually employed when the maximum likelihood estimates cannot be easily obtained using the standard maximum likelihood techniques due to incomplete or missing data. Background on the use of an EM algorithm in handling missing data is provided in Section 1.5.3. The algorithm was used here since the genotypes of those not in the extremes are effectively missing. The genetic variable of interest, denoted G, was only available on the subset of size n_1. They also assumed a set of covariates, such as ancestry related variables, denoted Z, that are also available on the n_1 selected samples. The secondary trait, Y_2, can be potentially available on all the cohort subjects or on just the genotyped n_1. This implies that there will minimally be some data missing from Z and G. Y_1 and Y_2 can be formulated from the bivariate

linear model given as:

$$Y_1 = {\beta_1}^T G + {\gamma_1}^T Z + \epsilon_1 \quad \text{and} \quad Y_2 = {\beta_2}^T G + {\gamma_2}^T Z + \epsilon_2$$

and the observed data likelihood is:

$$\prod_{i=1}^{n_1} P(Y_{1i}|G_i, Z_i)P(G_i, Z_i) \prod_{i=n_1+1}^{n} \sum_{g,z} P(Y_{1i}|g, z)P(g, z) \prod_{i=1}^{n} P(Y_{2i}|Y_{1i}, G_i, Z_i)$$

where P is the conditional density function of the two trait variables. The summation in the middle is over possible values of g and z for the portion of the sample with phenotype values in the middle of the distribution (where g and z are missing). The parameters for the primary traits were obtained from the first two expressions of the conditional density function while the last expression gives us the parameters for the secondary observations.

1.3.3 Effects of population stratification in EPS

Spurious associations due to population stratification are a problem in both case control and quantitative trait analysis. However, it was unclear the extent to which the EPS design is affected by population stratification. To quantify false positive association rate due to EPS, Panarella and Burkett [52] compared the false associations in an EPS sample to an equivalently sized random sample from the same population. This work is now summarized as it is background for Chapter 2.

Assume a population of size N which consists of 2 subpopulations. Within subpopulation, assume that the phenotype is normally distributed. Let Y_{ij} represent

the phenotype of an individual j in population $i, i = 1, 2$; i.e

$$Y_{ij} \sim \mathrm{N}(\mu_i, \sigma^2) \;\; j = 1, \ldots \; j_i,$$

$\mu_1 = -\mu_2$ and assume that $\sigma^2 = 1$. Assume a biallelic genetic locus under Hardy-Weinberg equilibrium (HWE). HWE describes the distribution of the allele frequencies in a population. Under an independence assumption, for a diallelic locus with alleles A and a, let p_i represent the frequency of the A allele and $1 - p_i$ for the frequency of the a allele in each of population 1 and 2. Then assuming the conditions of HWE are met, the genotype frequencies in population i are given by:

$$\begin{aligned} P(G = AA) &= p_i^2 \\ P(G = Aa) &= 2p_i(1 - p_i) \\ P(G = aa) &= (1 - p_i)^2 \end{aligned}$$

Let ω_i represent the proportion of the full population from population i; therefore, $\omega_1 + \omega_2 = 1$. For a random sample of size n from the full population, the number of individuals sampled from population 1 is expected to be $n\omega_1$ and $n\omega_2$ individuals are expected to be sampled from population 2. The genotype counts within a subpopulation follows Hardy-Weinberg proportions and so the counts in the full population are given by:

$$\begin{aligned} n_{AA} &= n(\omega_1 p_1^2 + \omega_2 p_2^2) \\ n_{Aa} &= 2n(\omega_1 p_1(1 - p_1) + \omega_2 p_2(1 - p_2)) \\ n_{aa} &= n(\omega_1(1 - p_1)^2 + \omega_2(1 - p_2)^2) \end{aligned}$$

The distribution of the phenotypes is a mixture of two normal distributions which is given by:

$$F(y) = \omega_1 \Phi(\frac{y-\mu_1}{\sigma}) + \omega_2 \Phi(\frac{y-\mu_2}{\sigma}).$$

Due to population stratification, the phenotype distribution conditional on genotype category is a mixture distribution with normal components.

$$\begin{aligned}
F(y|G=AA) &= F(y|i=1, G=AA)\Pr(i=1|G=AA) \\
&\quad + F(y|i=2, G=AA)\Pr(i=2|G=AA) \\
&= \frac{w_1 p_1^2}{w_1 p_1^2 + w_2 p_2^2}\Phi(\frac{y-\mu_1}{\sigma}) + \frac{w_2 p_2^2}{w_1 p_1^2 + w_2 p_2^2}\Phi(\frac{y+\mu_1}{\sigma}) \\
&= p_{1|AA}\Phi(\frac{y-\mu_1}{\sigma}) + p_{2|AA}\Phi(\frac{y+\mu_1}{\sigma}) \qquad (1.3.1) \\
F(y|G=Aa) &= p_{1|Aa}\Phi(\frac{y-\mu_1}{\sigma}) + p_{2|Aa}\Phi(\frac{y+\mu_1}{\sigma}) \\
F(y|G=aa) &= p_{1|aa}\Phi(\frac{y-\mu_1}{\sigma}) + p_{2|aa}\Phi(\frac{y+\mu_1}{\sigma})
\end{aligned}$$

where $p_{i|g}$ denotes the probability of being in population i given genotype G. Given the phenotype distribution conditional on genotype, Panarella and Burkett [52] determined the expected counts in the phenotype extremes for the three different genotypes. With these counts, the probability of rejecting the null assuming a Pearson chi square test could be analytically computed since a closed form expression for the non-centrality parameter could be derived. However, a simulation study was carried out because it was not possible to analytically compute the false positive rate for the random sampling scheme. This is because the probability of rejecting the ANOVA statistic requires the distribution of the phenotypes conditional on genotypes, which is not Gaussian as seen in equation 1.3.1. Codominant, recessive and additive models were used for model evaluation. For each of these models considered, for the EPS

sample, the Cochran Armitage trend test, χ^2 test and a difference of proportions test was used for association testing. For the random sample, an ANOVA or linear regression was performed. The simulations were performed for combinations of mean phenotype values (μ_i), population mixing proportions (ω_i) and major allele frequencies (p_1, p_2). For example, for $\alpha = 0.05$, $\mu_1 = 0.1(\mu_2 = -\mu_2)$ and $p_1 = p_2 = 0.5$, the false positive rate equaled the nominal value of 0.05 which is expected since there was no form of population differentiation. For differences in the major allele frequencies, the false positive rates under an EPS study increased rapidly compared to the random sample. Specifically, when $p_1 = 0.5$ and $p_2 = 0.9$, the false positive rate for the EPS design is 0.5 while for the random sample it is 0.2. The tests were performed under all three genetic models considered and a similar pattern of results were obtained with the recessive model showing the lowest values. Therefore, when compared to random sampling, EPS sampling results in even higher false positive rates due to population stratification than with random sampling.

To assess PC-based correction for population stratification under the EPS design, Panarella and Burkett [52] simulated genotype data as in Price et al. [57]. They found that including the top principal components in a logistic regression model is sufficient for controlling the type 1 error rate and that the effects of confounding were not reduced even when the sample size was increased.

In Chapter 3, we extend this work to linear mixed models.

1.4 Linear Mixed Models in Genetic Association Studies

Linear mixed models (LMMs) have emerged as a powerful statistical tool for genetic association studies, offering enhanced capability to account for population stratification and cryptic relatedness and have largely replaced PCA as the tool of choice for population stratification correction [58]. This section will provide an overview of linear mixed models in the context of genetic association studies, highlighting their advantages, key components, and applications

Linear Mixed Models (LMMs) are statistical models that incorporate both fixed effects, which capture the effects of specific variables of interest, and random effects, which account for correlations within the data. An LMM can be generally expressed as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \epsilon \tag{1.4.1}$$

where $\boldsymbol{y}$ is a vector of observations, $\boldsymbol{X}$ is a matrix of known covariates, $\boldsymbol{\beta}$ is a vector of unknown regression coefficients, which are often called the fixed effects, $\boldsymbol{Z}$ is a known matrix, $\boldsymbol{b}$ is a vector of random effects and ϵ is a vector of residual errors. Both $\boldsymbol{b}$ and ϵ are unobservable. A basic assumption in the LMM is that $\boldsymbol{b}$ and ϵ have zero mean and finite variances.

Jiang [26] have classified LMMs as Gaussian or non-Gaussian. Gaussian LMMs assume that the random effects and residual errors are normally distributed. These forms of models are flexible in modelling compared to the non-normal counterparts. They can be further classified into ANOVA models, longitudinal LMMs and marginal LMMs.

Longitudinal linear mixed models are often used in the analysis of longitudinal data. One feature of these models is that the observations may be divided into independent groups with one random effect (or vector of random effects) corresponding to each group. In practice, these groups may correspond to different individuals involved in the longitudinal study. There maybe some serial correlations within each group which are in addition to the random effects. A general longitudinal model may be expressed as:

$$\boldsymbol{y_i} = \boldsymbol{X_i}\beta + \boldsymbol{Z_i}b_i + \epsilon_i, \quad i = 1, \ldots, m \tag{1.4.2}$$

where $\boldsymbol{y_i}$ represents the vector of observations for the ith individual; $\boldsymbol{X_i}$ and $\boldsymbol{Z_i}$ are known matrices associated with the fixed and random effects for a subject i; β is an unknown vector of regression (fixed effects) coefficients; $\boldsymbol{b_i}$ is a vector of random effects; and ϵ_i is a vector of errors. It is assumed that $\boldsymbol{b_i}$ and $\epsilon_i, i = 1, \ldots, m$ are independent with $\boldsymbol{b_i} \sim N(0, \boldsymbol{G_i}), \epsilon_i \sim N(0, \boldsymbol{R_i})$, where the covariance matrices $\boldsymbol{G_i}$ and $\boldsymbol{R_i}$ are known up to a vector $\boldsymbol{\theta}$ of variance. Longitudinal LMMs could also have time dependent covariates which may appear in $\boldsymbol{X}$, $\boldsymbol{Z}$ or in both matrices.

In genetic association studies, linear mixed model based approaches have been recently applied as an alternative method to principal component based approaches when adjusting for genetic relatedness in seemingly unrelated individuals [18]. The key idea behind using LMMs to correct for population stratification is that they allow for the use of a genetic relatedness matrix (GRM), which quantifies the degree of genetic similarity between pairs of individuals. The GRM is calculated based on genome-wide genetic data, typically using single nucleotide polymorphisms (SNPs). It captures the underlying genetic structure within the study population and serves to effectively model the correlation structure induced by hidden population substructure

or relatedness. We provide more details on the GRM in Section 1.4.1.

Mixed models were first applied to association mapping in maize, Arabidopsis, and potato panels, where a reduction in false positive associations compared to genomic control, structured association, and principal components methods was demonstrated [83, 85, 43]. However, the mixed model methods developed at the time had several limitations in the context of model organism association mapping. These limitations included potential inaccuracies in the computation of variance components and the high computational cost associated with the numerical optimization methods used [28]. These challenges led to the proposal of a new, more efficient mixed model method—EMMA (Efficient Mixed Model Association)—which was designed to correct for population structure and genetic relatedness in model organism mapping [28]. The model developed in EMMA laid the foundation for the use of LMMs in genetic association studies. The general form of LMM used in genetic association studies as given in Kang et al. [28] is of the form:

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \epsilon \tag{1.4.3}$$

where $\mathbf{y}$ is an $n \times 1$ vector of observed phenotypes, and $\boldsymbol{X}$ is an $n \times q$ matrix of fixed effects including the intercept, SNPs, and confounding variables. $\boldsymbol{\beta}$ is a $q \times 1$ vector representing coefficients of the fixed effects. $\boldsymbol{Z}$ is an $n \times t$ design matrix, $\mathbf{b}$ is the random effect of the mixed model with $\text{Var}(\mathbf{b}) = \sigma^2_{\text{b}}\boldsymbol{K}$, where $\boldsymbol{K}$ is the $n \times n$ GRM estimated from genotypes , and ϵ is an $n \times 1$ matrix of residual effect such that $\text{Var}(\mathbf{e}) = \sigma^2_{\epsilon} I_n$. Hence, the overall phenotypic variance-covariance matrix can be represented as $\boldsymbol{V} = \sigma^2_{\text{b}}\boldsymbol{Z}\boldsymbol{K}\boldsymbol{Z}' + \sigma^2_{\epsilon} I_n$.

In population based genetic association studies where the population consists

of large number of individuals from the same population background, applying an efficient implementation of a variance component method such as EMMA [28] could be computationally intractable for large datasets consisting of thousands of individuals due to the heavy computational limitation involved with the estimation of the variance parameters. Computational adjustments were made to the EMMA method which reduced the computational burden thus increasing the speed of computation and making human genetic association possible at the GWAS level. The improved expedited method from EMMA known as EMMAX (EMMA-eXpedited) leverages on the characteristics of complex traits in humans, namely the estimation of variance parameters only once for each dataset and then globally applying it for each marker. The models explored in EMMA and EMMAX provide approximate results with per-SNP computation time increasing quadratically with the number of individuals, it is hard to know how accurate these approximations are without an exact computation method. Hence, an exact efficient method that provides identical results to EMMA roughly n times faster was proposed. GEMMA - Genome wide efficient mixed model association makes exact computation for large GWAS feasible thereby eliminating the need for approximate methods [86]. GEMMA is based on the LMM model:

$$\mathbf{y} = \mathbf{W}\boldsymbol{\alpha} + \mathbf{x}\beta + \mathbf{Z}\mathbf{b} + \varepsilon \tag{1.4.4}$$

where:

$$\begin{aligned} \mathbf{b} &\sim \mathrm{MVN}_m\left(0, \lambda\tau^{-1}\mathbf{K}\right) \\ \boldsymbol{\varepsilon} &\sim \mathrm{MVN}_n\left(0, \tau^{-1}\mathbf{I}_n\right) \end{aligned} \tag{1.4.5}$$

where n is the number of individuals, m is the number of groups, strains or clusters, y is an $n1$ vector of quantitative traits, $\boldsymbol{W} = (w1, w2 \ldots wc)$ is an $n \times c$ matrix of covariates (fixed effects) including a column vector of 1, α is a $c \times 1$ vector of

corresponding coefficients including the intercept, $\boldsymbol{x}$ is an $n \times 1$ vector of marker genotypes (SNPs), β is the effect size of the marker, $\boldsymbol{Z}$ is an $n \times m$ loading matrix, b is an $m \times 1$ vector of random effects, ϵ is an $n \times 1$ vector of errors, τ^{-1} is the variance of the residual errors, λ is the ratio between the two variance components, $\boldsymbol{K}$ is a known $m \times m$ relatedness matrix, I_n is an $n \times n$ identity matrix and MVN denotes multivariate normal distribution.

In genetic studies involving human populations, where $m = n$, the matrix $\boldsymbol{Z}$ is typically an identity matrix I_n. This specification differs from the usual $m \times n$ $\boldsymbol{Z}$ matrix seen in typical longitudinal LMMs, where n corresponds to groups, clusters or time points. Despite this difference, the LMM used in GEMMA remains an identifiable model. The identifiability of the model depends on the proper specification of the variance components and the relatedness matrix $\boldsymbol{K}$ which ensures that the contributions of fixed and random effects can be distinctly and accurately estimated. The use of $\boldsymbol{Z} = \boldsymbol{I_n}$ simplifies the model by directly linking the random effects to the individuals, allowing the variance components to be uniquely estimated. Estimation of the fixed effect $(\boldsymbol{\alpha}, \beta)$ and random effect parameters b is achieved through standard maximum likelihood estimation (MLE) and restricted maximum likelihood estimation (REML) approaches, as described in [86]. These methods are well-suited to the LMM structure, ensuring that the model parameters can be correctly estimated even in the context of large-scale genetic association studies. Iterative methods like the EM algorithm have also been used to obtain the MLEs and in Chapter 3, we provide background on the EM approach.

Although LMMs have been widely used to correct for population structure in genetic association studies, they are not without limitations. LMMs can be computationally demanding, especially when applied to large-scale genetic association studies

involving thousands or millions of genetic variants and individuals [27]. Performing the necessary matrix operations for each variant and computing the GRM can be time-consuming and can require a significant amount of computational resources. In practice, this complexity and burden can limit the scalability and practicability of LMMs, especially for studies where computational infrastructure is a concern. To solve this problem, several implementations of the LMM in genetic association studies have emerged over the years. Examples of these methods includes EMMA [28], EMMAX [27], BOLT-LMM [40], GEMMA [86] and FASTLMM [37]. These methods have one feature in common: they all aim to improve the computational complexity associated with using LMMs in a GWAS. In Eu-Ahsunthornwattana et al. [18], a comparison of the performance of these methods was carried out and it was discovered that they all perform quite similarly in terms of the type 1 error and/or power, hence considerations such as speed and ease of use should be the factors to be considered when choosing a method to use.

1.4.1 Genetic Relationship Matrix for LMMs

In quantitative genetics, the Genetic Relationship Matrix (GRM) is a fundamental tool for quantifying the genetic relatedness among individuals in a population. It plays a crucial role in linear mixed models (LMMs) by accounting for the genetic covariance structure among individuals [65].

Let K denote the GRM, which is a symmetric $n \times n$ matrix, where n is the number of individuals in the population. Each element k_{ij} of K represents the estimated genetic covariance between individuals i and j. Methods for computing the GRM matrices are broadly divided into SNP-based and haplotype-based. SNP-based statistics, sometimes referred to as identity by state (IBS) methods, can be distin-

guished from haplotype-based methods (identity by descent (IBD) methods) using the idea of linkage. In this thesis, we consider only SNP-based measures for the GRM.

If $\boldsymbol{X}_s$ is a standardized $n \times p$ matrix of genotypes, a common method to compute a GRM is using:

$$\boldsymbol{K} = \frac{1}{p-1}\boldsymbol{X}_s\boldsymbol{X}_s'$$

where p is the number of SNPs. Some genetic software that incorporate the computation of GRMs into their program include GEMMA [86] and Gaston [5].

1.5 Concepts in Computational Statistics

The analysis of genetic data using LMMs are often computationally intensive for a number of reasons. LMMs usually involve working with large matrices, particularly the covariance matrices of the random effects and residuals. When dealing with large datasets, calculation on the large matrices can be computationally demanding. Also, the estimation of random effects in an LMM introduces an additional layer of complexity. The random effects, which are like nuisance variables, may require integration over their distribution, which may not be analytically feasible when the distribution is complex or multivariate. We now review some methods from computational statistics that are used to facilitate maximum likelihood estimation or hypothesis testing in this thesis.

1.5.1 Monte Carlo Integration

Monte Carlo integration is commonly used in statistics as a way of solving complex integrals. In statistics, integrals are often used to calculate probabilities or expected values of a random variable. However, in many cases, the integrals are too complex or have no closed-form solution, making them difficult or impossible to solve analytically. Monte Carlo integration provides a solution to this problem by using random sampling to estimate the integral. For more information about the Monte Carlo integration, see Tanner [69].

The basic idea is to generate a large number of random samples from the distribution of interest and then use these samples to approximate the integral. To illustrate this approach, consider the problem of calculating the expected value of a function $f(x)$ over a continuous distribution $p(x)$. The expected value is

$$\theta = E[f(x)] = \int f(x)p(x)dx \tag{1.5.1}$$

If $p(x)$ is a complex or high-dimensional distribution, it may be difficult to solve this integral analytically. In this case, Monte Carlo sampling can be used to estimate the expected value. With a random sample $x_1, x_2, \ldots x_n$ from the distribution $p(x)$, the Monte Carlo estimator of the expectation above is simply the sample mean:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} f(x_n) \tag{1.5.2}$$

where $\hat{\theta}$ is used to denote an estimate of $E[f(x)]$.

One factor affecting the accuracy of the Monte Carlo integration s the number of samples used in the simulation. Generally, the more samples that are generated,

the more accurate the estimate will be. However, even a relatively small number of samples can provide a reasonable estimate of the expected value, making Monte Carlo sampling a powerful tool for solving complex integrals in statistics.

On the other hand, estimation of some expected values, such as small probabilities, would require a large number of samples. Therefore, variations of Monte Carlo integration have been developed, such as importance sampling.

1.5.2 Importance Sampling

Importance sampling is an extension of Monte Carlo integration. It was initially proposed as a means to reduce the variance of the estimators compared to standard Monte Carlo methods [38]. We consider a situation where we want to calculate the expectation of a function $f(x)$ where x has a density function $p(x)$. But what if we can't sample easily from $p(x)$? With importance sampling, we are able to estimate the expectation (θ) based on some known and easily sampled distribution $q(x)$, by a simple transformation

$$\theta = E(f(x)) = \int f(x)p(x)dx = \int f(x)\frac{p(x)}{q(x)}q(x)dx.$$

Therefore, we can estimate θ with

$$\hat{\theta} \approx \frac{1}{m}\sum_{i=1}^{m} f(x_i)\frac{p(x_i)}{q(x_i)}$$

where x is now sampled from $q(x)$ and $q(x) \neq 0$. The quantity $\frac{p(x)}{q(x)}$ is known as the importance sampling weight to offset the sampling done from a different distribution,

and $q(x)$ is known as the "proposal distribution". When using importance sampling, it is important to carefully select the proposal distribution so that the variance of the estimator is not too large. The density $q(x)$ should be similar to $p(x)$ but with thicker tails so that the ratio $\frac{p(x)}{q(x)}$ doesn't get large [19].

1.5.3 The EM algorithm

The EM algorithm first appeared in the paper by Dempster et al. [13]. It was presented as a general approach to the iterative computation of ML estimates by viewing the observations as incomplete data [13]. Its popularity is partly due to its simplicity and also because a wide range of problems can be modified to fit into the incomplete data umbrella. The name of the method is obtained from the two iterative steps that make up the algorithm: an E-Step and an M-step.

A key idea in formulating the EM algorithm is the concept of the "complete data". These consist of the observed data denoted by $\mathbf{y}$ and the unobserved (missing) random variables or vector of random variables denoted as $\boldsymbol{\Omega}$. Let $\boldsymbol{w} = (\boldsymbol{y}, \boldsymbol{\Omega})$ denote the complete data with a probability density function $f(\boldsymbol{w}; \boldsymbol{\theta})$ which depends on a vector $\boldsymbol{\theta}$ of unknown parameters. The E-step of the algorithm computes the function Q, the conditional expectation of the complete data given the observed data and the current parameter estimates:

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta^{(t)}}) = \mathrm{E}\left\{\log f(\boldsymbol{w}; \boldsymbol{\theta}) \mid \mathbf{y}, \boldsymbol{\theta}^{(t)}\right\} \tag{1.5.3}$$

where $\boldsymbol{\theta^{(t)}}$ is the estimated $\boldsymbol{\theta}$ at the tth iteration. In the M-step, one maximizes $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$ to obtain the next step estimator, $\boldsymbol{\theta}^{(t+1)}$. The process is then iterated until convergence. It should be noted that the missing data must be

chosen appropriately so that the maximum likelihood estimation becomes trivial for the complete data [26].

1.6 Missing Data Mechanisms

Missing data is a common challenge encountered in genetic data analysis, as a result of factors such as genotyping errors, sample dropout, or the unavailability of certain measurements. Methods for studying missing data often depend on the pattern of missingness and the mechanism that generates the missing values. The missing data pattern describes which values are observed and which values are missing in the data matrix while the missing data mechanism concerns the relationship between missing data and the values of variables in the data matrix. Missing data mechanisms can be categorized as: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). We will give a brief discussion on these missing data mechanisms; for more information, see Little and Rubin [38] and Rubin [62].

Consider a dataset that consists of a vector of responses $\boldsymbol{y} = (y_1, \ldots y_n)$ that may contain missing values and an $n \times p$ matrix $\boldsymbol{X} = (x_1, \ldots x_n)$ of completely observed explanatory variables.

The missing values of y_i are MCAR if the probability of observing y_i is independent of X and the values of y_i that are observed or would have been observed. Under MCAR, the missingness pattern is entirely random and occurs by chance. Statistical analyses that assume MCAR typically do not introduce systematic bias into the results, as the missingness is not connected to any underlying data characteristics. Under MCAR, the missing data mechanism takes the form $f(r_i|X_i, \Phi)$ where Φ is a vector of unknown parameters and $R_i = (R_{i1} \ldots R_{im})'$ are the missing data

indicators; $R_{ij} = 1$ if Y_{ij} is observed and 0 otherwise. As an example of MCAR, in a large epidemiological study investigating the relationship between smoking and lung cancer, a batch of questionnaires was accidentally lost due to a clerical error or during data collection. If the lost questionnaires were from a random subset of participants, regardless of their smoking status or lung cancer diagnosis, this would be an example of MCAR. The missing data are completely unrelated to any participant characteristics.

Missing data is said to be MAR if given the observed data, the failure to observe a value does not depend on the values of y which are unobserved. While MAR acknowledges that the missingness is not completely random, it offers the opportunity to use the observed data to predict and potentially impute the missing values. Ignoring MAR and performing analyses based solely on complete cases can introduce bias if the missingness mechanism is not accounted for properly [38]. Under MAR, the missing data mechanism is represented as: $f(r_i|X_i, y_{obs,i}, \Phi)$ where $y_{obs,i}$ denotes the observed components of y_i. Consider a study on the effect of physical activity on obesity. Suppose that data on physical activity levels are more likely to be missing for older participants. However, within each age group, the likelihood of missing data does not depend on their obesity status or other health outcomes. In this case, the missingness is related to age (an observed variable) but not directly to the unobserved physical activity data. This is an example of MAR.

The missing data mechanism is said to be NMAR if the failure to observe a value depends on the value that would have been observed. Here, the probability of data being missing is related to the unobserved values themselves, even after accounting for observed data. The mechanism here is $f(r_i|X_i, y_{obs,i}, y_{mis,i}, \Phi)$. For example, in a cohort study on alcohol consumption and liver disease, individuals who consume

alcohol heavily might underreport or skip questions about their alcohol intake. This missing data is directly related to the level of alcohol consumption itself, making it MNAR

In the work presented in this thesis, we consider data missing MCAR and MAR only.

1.7 Overview of the thesis

This thesis is organized as follows:

In **Chapter 2**, we evaluate the performance of existing mixed model methodologies to handle confounding by population stratification within the context of EPS. While these methodologies have undergone extensive use in scenarios involving continuous variables and case-control studies, their applicability in situations where genetic covariates are exclusively available for samples exhibiting extreme values of the response variable remained unexplored. The mixed model methodologies employed in this investigation include the Linear Mixed Models (LMMs), Generalized Linear Mixed Models (GLMMs), and Liability Threshold Models (LTMs). The primary objective of this study was to conduct a comparative analysis of the efficacy of these methodologies in controlling Type 1 errors attributed to population stratification within the EPS data framework, and the response variable represented as a binary trait (high and low extreme). This study begins with an overview of mixed model methodology that is applicable to binary response variables. To assess Type 1 error rates and statistical power, simulation studies were employed. As a practical application of these mixed model techniques, we also examined their use in the context of a prostate cancer study. Specifically, we employed these methodologies to assess the

relationship between the BMI phenotype and genotype data in the ProTEUS study [73].

In **Chapter 3**, we present a novel approach based on the EM algorithm to estimate genetic effects in a LMM in the presence of missing genotype data. This work was initailly motivated as an extension of the Lin [36] linear model based method for EPS data to the linear mixed model context. Our algorithm is also applicable to sporadically missing genotype data. Our model draws upon the method of weights, as proposed by Ibrahim and Lipsitz [24] which involves assigning "weights" to the missing categorical covariates. We adapt this to the context of linear mixed models for genetic data analysis to effectively address the population structure inherent in population-based genetic association studies. In order to test the significance of the genetic effect, we devised a likelihood ratio test and Wald test statistic through a Monte Carlo sampling approach.

In **Chapter 4**, we evaluate the performance of our Expectation-Maximization (EM) approach, which we call EM-LMM, using data simulated under various scenarios and conditions. We simulated datasets with a population genetics model that reflects population structures commonly encountered in genetic association studies. We considered scenarios where genotype data were missing due to both EPS and sporadic missingness. We assessed the algorithm's performance in estimation of the model parameters, control of type 1 error rates and statistical power. We applied our approach in a real-world application to analyze a prostate cancer dataset [53] considered in Chapter 2. Our primary aim was to conduct a GWAS using the EM-algorithm, considering both an EPS design and a randomly missing scenario to mimic real world data complexities. By comparing the parameter estimates obtained from the EM algorithm to those derived from a traditional LMM model, we sought to

assess the algorithm's effectiveness.

In **Chapter 5** we summarize the work presented and we describe future research directions.

Chapter 2

Application of Mixed Model Methods in Correcting for Population Substructure

Population substructure is a broad term used in population genetics to refer to the phenomenon of systematic variation of the allele frequencies across individuals in the population. Population substructure includes population stratification, population admixture and cryptic relatedness [31] and if not properly accounted for, it can lead to spurious associations between genetic variants and phenotypes.

Recall that EPS designs were shown to have higher false positive rates compared to random sampling and that PC-based corrections adequately corrected the type 1 error except with rare variants Panarella and Burkett [52]. In this chapter, we present our work on the use of mixed model based approaches for correcting for population structure in extreme phenotype samples. The term "mixed-models" as used in this thesis refers to both linear mixed models (LMMs) and generalised linear

mixed models (GLMMs). The use of LMMs for correcting for population stratification has been reviewed in section 1.4. GLMMs are an extension of LMMs such that allow the response variable to come from different distributions besides Gaussian. As with GLMs, rather than modelling the responses directly, a link function is used to specify the relationship between the response and covariates.

Although LMMs offer a powerful approach to address population structure by explicitly modelling both the fixed effects (such as covariates and genetic variants) and the random effects (such as the genetic relatedness among individuals), they are computationally intensive. This has resulted in the development of a number of exact and approximate solutions designed to make LMM-based analyses feasible at the genome-wide level. All these methods assume continuous data.

Given the increasing popularity of mixed model based approaches, this chapter presents our investigation of the efficacy of mixed model based approaches to correct for confounding due to population stratification in the EPS context. Firstly we review mixed model based approaches for correcting for population stratification with a binary response variable. Secondly, we evaluate whether mixed model based correction is sufficient for controlling the inflated false positive rate due to confounding by population stratification under an EPS design. Using simulation studies, we estimate the type 1 error and power for both common and rare variants method. Third, we compare performance on all methods on a prostrate cancer dataset.

This work has been published in a BMC Genomics article titled "Comparison of mixed model based approaches for correcting for population substructure with application to extreme phenotype sampling" [50]. The published article has been copied in Section 2.2. The format of the journal requires that the methods be placed at the end of the article; therefore the methods are described on the 7th page of the

article.

2.1 Author Contributions

I am the first author and as such I contributed substantially to all aspects of this work. I designed the study, outlining the research objectives and identified relevant methodology. I programmed the simulations utilizing principles and techniques in statistical computing to ensure the accuracy of the simulations. I conducted the data analysis on the simulated and real data using R [71] and PLINK [61]. I summarized and interpreted the results. I also took the lead in drafting the manuscript.

2.2 The BMC Genomics article

Onifade *et al. BMC Genomics* (2022) 23:98
https://doi.org/10.1186/s12864-022-08297-y

BMC Genomics

RESEARCH ARTICLE Open Access

Comparison of mixed model based approaches for correcting for population substructure with application to extreme phenotype sampling

Maryam Onifade[1], Marie-Hélène Roy-Gagnon[2], Marie-Élise Parent[3] and Kelly M. Burkett[1*]

Abstract

Background: Mixed models are used to correct for confounding due to population stratification and hidden relatedness in genome-wide association studies. This class of models includes linear mixed models and generalized linear mixed models. Existing mixed model approaches to correct for population substructure have been previously investigated with both continuous and case-control response variables. However, they have not been investigated in the context of extreme phenotype sampling (EPS), where genetic covariates are only collected on samples having extreme response variable values. In this work, we compare the performance of existing binary trait mixed model approaches (GMMAT, LEAP and CARAT) on EPS data. Since linear mixed models are commonly used even with binary traits, we also evaluate the performance of a popular linear mixed model implementation (GEMMA).

Results: We used simulation studies to estimate the type I error rate and power of all approaches assuming a population with substructure. Our simulation results show that for a common candidate variant, both LEAP and GMMAT control the type I error rate while CARAT's rate remains inflated. We applied all methods to a real dataset from a Québec, Canada, case-control study that is known to have population substructure. We observe similar type I error control with the analysis on the Québec dataset. For rare variants, the false positive rate remains inflated even after correction with mixed model approaches. For methods that control the type I error rate, the estimated power is comparable.

Conclusions: The methods compared in this study differ in their type I error control. Therefore, when data are from an EPS study, care should be taken to ensure that the models underlying the methodology are suitable to the sampling strategy and to the minor allele frequency of the candidate SNPs.

Keywords: Population stratification, Extreme phenotype sampling, Generalized linear mixed models, Type 1 error, Genome-wide association study

*Correspondence: kburkett@uottawa.ca
[1]Department of Mathematics and Statistics, University of Ottawa, Ottawa, Canada
Full list of author information is available at the end of the article

Background

In genetic studies involving human populations, researchers are interested in determining how genetic variation contributes to disease. Genome-Wide Association Studies (GWAS), which involve genotyping a large number of individuals at hundreds of thousands of genetic markers, have been useful for discovering relationships between common variants and complex diseases. Recently, sequencing has been used to discover rare variants associated with human traits [1]. Although the cost of genetic association studies has decreased over the years, some technologies, including sequencing, remain relatively expensive [2]. Therefore study designs that reduce cost while maintaining power are desirable.

An example of a cost saving design is extreme phenotype sampling (EPS), a design where genetic data are collected only on individuals in the tails of the phenotype distribution. The use of this study design was proposed by Lander and Botstein [3] for linkage analysis. Extreme phenotype sampling was later used for candidate gene association studies. For example, the EPS design was used to investigate associations between genetic variants in the dopamine system genes and cognitive ability [4, 5]. This study design has also been used in GWAS, for example in Vermissen et al. [6] to identify genetic risk variants for coronary heart disease. Recently, EPS has been shown to be a powerful design to detect rare variants [2, 7–9].

As with all population-based genetic association designs, extreme phenotype sampling is prone to confounding by population structure or stratification. Differences in allele frequencies among members of a strata or subgroup in the population may lead to confounding if there are also differences in the phenotype distribution between the subgroups. Confounding is known to inflate the type I error rate, which can lead to spurious associations. Methods have been developed that can correct for the effects of population stratification using genomic data. The earliest approaches include Genomic Control [10] and STRUCTURE/STRAT [11]. Principal components (PC)-based corrections have also been shown to be sufficient for controlling the false positive rate [12, 13].

Mixed model methods have recently become popular due to their robustness in tackling other sources of confounding in the study, in particular cryptic relatedness [14]. Since mixed model based approaches are computationally intensive, a number of exact and approximate linear mixed model (LMM) methods have been developed for use in genome-wide association studies (for example, [15–17]). Each of these methods incorporate different strategies to make the LMM-based analyses feasible at the genome-wide level. Eu-ahsunthornwattana et al. [18] gives a comparison of these methods.

In human genetic studies, the phenotype of interest is often a binary trait, such as presence or absence of disease. To correct for population stratification, binary traits are sometimes analysed using LMMs [19–21] even though the response variable is not continuous. Pirinen et al. [22] gives a justification of this approach by deriving a mapping between the effect size estimates from the linear to the log-odds scale, which is the natural scale for binary traits. Although widely applied to binary traits, the LMM assumes a continuous phenotype with a constant residual variance. However, for binary traits in the presence of covariates, this assumption does not hold. Therefore, fitting a binary response with linear mixed models may fail to correct the type I error rate [23] or result in a loss of power [24].

Mixed model approaches that do not treat disease status as a continuous random variable have recently been developed. One such approach is based on the liability threshold model, which assumes that there is an unobserved normally distributed latent variable known as the 'liability' and that individuals having liability values above a threshold are classified as cases. Liability threshold-based methods have been implemented in the software LEAP [25] and LTMLM [26]. These methods estimate the latent liabilities and association is tested using these estimated latent response values. The generalized linear mixed model (GLMM) can also be used to model binary traits. For example, GMMAT [23] fits a logistic mixed model to the binary data, while CARAT [27] fits a retrospective model using a quasi-likelihood approach.

We have previously shown that the false positive rates due to population stratification are substantially inflated with EPS designs relative to random sampling [28]. Therefore, for EPS designs it is very important to include correction for population stratification. We have shown that including the top principal components in a logistic regression model adequately limits the type I error rate when the candidate variant was common; however, there was a slight inflation when the candidate variant was rare [28]. The mixed model-based approaches for correcting for population substructure were developed assuming binary traits from case-control type studies. In particular, the retrospective and liability threshold approaches model the underlying case-control ascertainment. However, the sampling scheme used in EPS designs is different from true case-control designs as both extremes of the phenotype distribution are included. Therefore, it is unclear whether these approaches will adequately control the false positive rate under the EPS ascertainment scheme when there is confounding due to population stratification. Given the increasing popularity of mixed model approaches, it is important to assess their performance in the EPS setting.

In this work, we aim to accomplish two goals. First, we present an overview of the mixed model-based approaches for correcting for population stratification with a binary response variable; we focus on the recently proposed algorithms LEAP, LTMLM, GMMAT and CARAT. Second, we compare the performance of these approaches and an LMM approach (GEMMA [17]) when the binary data comes from an EPS design. We use simulation to evaluate whether the type I error rate is adequately controlled when the candidate variant is both common and rare. We also examine the power of the methods shown to control the type I error rate. Finally, we compare these methods when applied to a real dataset collected as part of a case-control study conducted in Québec, Canada. The participants were collected from multiple ethnic groups and therefore we expect confounding by population stratification with this data.

Results

Evaluation of type I error - common variant

Table 1 shows the estimated type I error rates for the EPS samples of size 1000, 2000 and 4000, which correspond to full cohort sample sizes of 5000, 10,000 and 20,000 individuals. These results correspond to the simulations with the '1' allele frequency of $p_1 = 0.25$ and $p_2 = 0.85$ and the phenotypic means of $\mu_1 = 0.07$ and $\mu_2 = -0.07$ for subpopulations 1 and 2, respectively. LEAP and GMMAT show well controlled type I error rates, indicating adequate correction of the population structure in the data. For both approaches, the estimated type I error rate for all the sample sizes ranges between $0.041 - 0.052$. All but one of these estimates are slightly lower than the nominal level of 0.05; however, these small deviations from the true value can be explained by Monte Carlo sampling error. The type I error rate for the PCA approach is also close to the nominal value, though possibly slightly elevated; similar results for the PCA based correction were observed in our previous work [28]. CARAT shows higher type I error rates than the nominal level of 0.05. The false positive proportion ranges from 0.089 to 0.102, which is higher than can be explained by Monte Carlo simulation error alone. We therefore conclude that CARAT is not able to adequately correct for population stratification in the EPS setting.

We also evaluated the LMM approach GEMMA, where we coded the categorical phenotype as 0 and 1 for the two extreme groups and treated the 0/1 values as a continuous phenotype. Results in Table 1 show that the estimated type I error rates were around 0.05, which indicates that erroneously analysing as a continuous trait does not affect the correction for population substructure.

Figure 1 shows the results when the '1' allele frequency of the candidate SNP in subpopulation 2 was varied from 0.5 to 0.9, in increments of 0.1. When $p_1 = p_2$ there is no population stratification; as expected, under this case the type I error rate of the three methods are all close to the nominal value of 0.05. GMMAT and LEAP show no increase in the estimated type I error rates as p_2 increases; the estimated value remains around 0.05. However, for CARAT the type I error rates increases as the difference in the allele frequency between the two subpopulations increases, which again indicates inadequate correction for population stratification.

Evaluation of type I error - rare variants

Table 2 shows the estimated type I error rates of the Burden (SMMAT-B), SKAT (SMMAT-S), SKAT-O (SMMAT-O) and Hybrid efficient (SMMAT-E) statistics from SMMAT assuming a significance level of 0.05. The Burden test had an estimated type I error rate closest to the specified value (0.062 versus the expected 0.05). The three other rare variant statistics (SMMAT-S, SMMAT-O, SMMAT-E) have estimated type I error rates that range from 0.088 to 0.103. The inflation of the test statistics can also be seen in the QQ plot of the -log10 of the p-values (Fig. 2); the results with the Burden statistic appear closest to the identity line, which is what we would expect under no association, but there is still evidence of inflated test statistics. The deviations between the true and estimated type I error rates cannot be explained by simulation error alone; we conclude that under EPS, the type I error rate is not controlled using these rare variant tests.

Evaluation of power

Table 3 shows the estimated power to detect a common candidate variant when the phenotype also depends on subpopulation membership. We assessed power only for the methods with appropriate type I error control and we

Table 1 Estimated type I error rates for the three mixed model approaches for binary traits (LEAP, GMMAT and CARAT), the LMM method (GEMMA) and logistic regression with principal component based correction (PCA)

Cohort Sample Size (N)	Sub-sample Size (0.2N)	LEAP	GMMAT	CARAT	GEMMA	PCA
5000	1000	0.0405	0.04135	0.102.	0.061	0.0575
10000	2000	0.0417	0.0475	0.089	0.046	0.0605
20000	4000	0.0450	0.0515	0.0945*	0.052	0.0555

*Based on m=1999 simulations

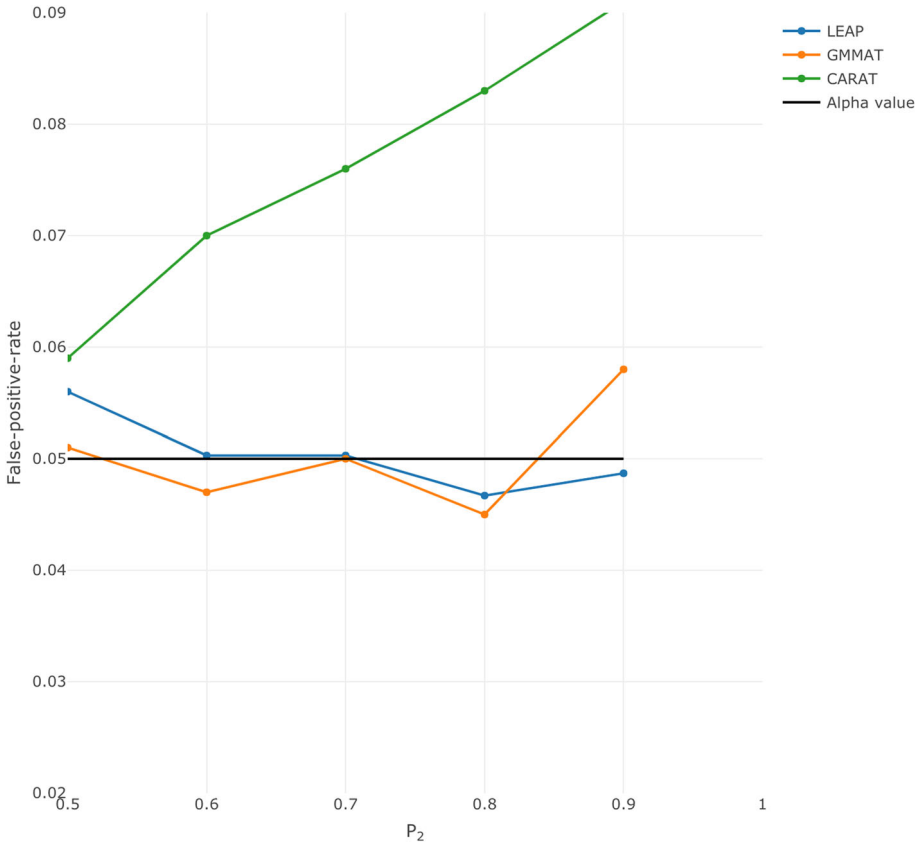


Fig. 1 Type 1 error rates for the three mixed model methods (LEAP, GMMAT and CARAT). The allele frequency in population 1, p_1, was fixed at 0.5. The allele frequency in population 2, p_2, ranged from 0.5 to 0.9. The x-axis corresponds to the p_2 value. The orange line represents GMMAT, the blue line represents LEAP, and the green line represents CARAT. The horizontal line indicates the alpha value of 0.05

evaluated two effect sizes. We note that overall power for all methods will depend on the effect size and sample size; therefore, we focus on comparing the estimates from each method to each other rather than on determining if power overall is high enough. At the larger effect size ($\beta = 0.25$), no method clearly outperforms the others. The estimated power for all four methods ranges from 0.48 to 0.52. LEAP has the lowest power and GEMMA the highest. The same pattern is seen with the lower effect size ($\beta = 0.15$); LEAP is lowest and GEMMA is highest. At the smaller effect size, GEMMA's estimated power is about 10% higher than the next lowest (PCA). If we perform a test of equality of proportions estimated for GEMMA and PCA, we would reject the hypothesis that they are equal.

Extreme BMI phenotype in the prostate cancer case-control study

Figure 3 shows the QQ plots of $-\log_{10}$ of the p-values from LEAP, GMMAT, GEMMA and the uncorrected logistic regression implemented in PLINK for the genome-wide association study using the extremes of the BMI phenotype from the prostate cancer case-control study. For reference purposes, Manhattan plots for each method are provided in Supplementary Figures 1-4 (Additional files 1, 2, 3 and 4), respectively.

The results from LEAP, GEMMA, and GMMAT show well controlled type 1 error rates; in Fig. 3, the majority of p-values tend to fall close to the identity line although again GMMAT may slightly over-correct. The correction for relatedness does seem to alter the results; we can see that the points for the methods that offer correction (GMMAT, GEMMA and LEAP) are all below the points for the method which doesn't correct (PLINK).

Computational time and memory usage

We compared the computational time and memory requirements for LEAP, GMMAT, GEMMA and CARAT

Table 2 Estimated type I error rates for the rare variant mixed model methods implemented in SMMAT

SMMAT Method	Estimated Type I Error Rate
Burden (SMMAT-B)	0.0617
SKAT (SMMAT-S)	0.1039
SKAT-O (SMMAT-O)	0.1024
Efficient (SMMAT-E)	0.0883

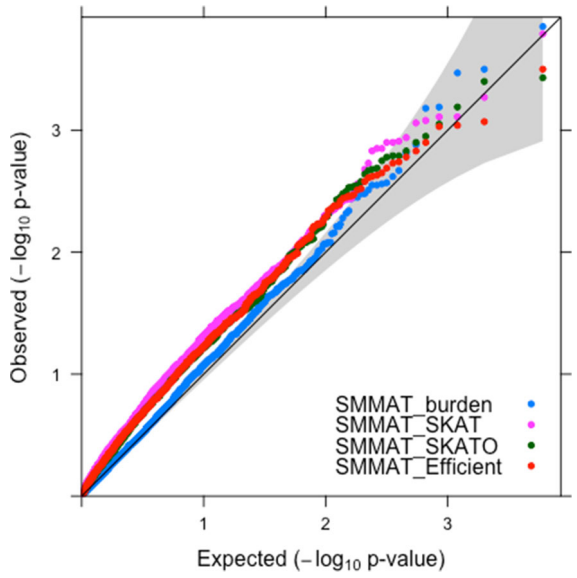


Fig. 2 Quantile-Quantile Plots of the $-\log_{10}$ of the *p*-values from the four SMMAT rare variant association tests (Burden, SKAT, SKAT-0 and Efficient)

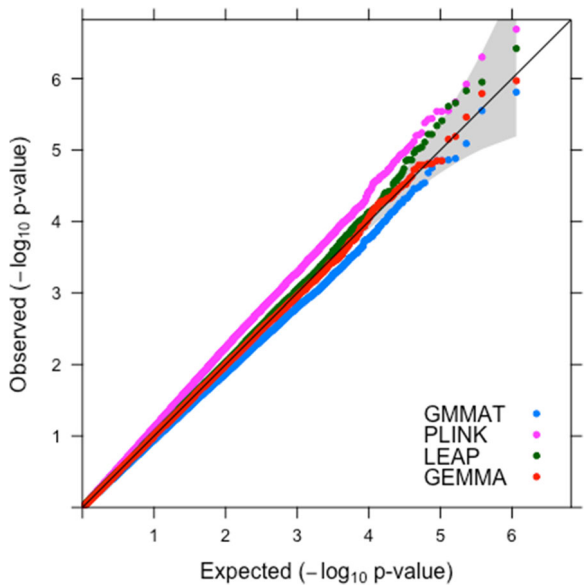


Fig. 3 Quantile-Quantile plot of population stratification adjusted GMMAT, LEAP, GEMMA and uncorrected PLINK in the GWAS analysis of the case-control dataset

on a dataset of $n = 1000$ individuals. With the exception of GMMAT, the methods included GRM calculation as part of the algorithm. For GMMAT, we used the GRM computed by GEMMA; this extra step increases the overall computational time. The average time (across simulations) to complete an association test on a single candidate SNP was approximately 22 seconds for GMMAT, 19 seconds for GEMMA, and 13 seconds for LEAP. Although the times are similar when analysing a single SNP, these differences between run times would be quite noticeable for a GWAS analysis. CARAT's run time was significantly longer at over 5 minutes per dataset; we therefore were unable to complete the same number of simulations for CARAT at the larger sample size (4000 samples from the phenotype extremes). Memory usage was comparable between methods, though LEAP's usage was higher (GMMAT: 3 GB, GEMMA: 2 GB; LEAP: 4.25 GB; CARAT 1.34 GB).

Table 3 Estimated power for detecting a causal variant of two different effect sizes ($\beta = 0.15$ and $\beta = 0.25$) in the presence of population stratification

Method	Estimated Power	
	β=0.15	β=0.25
LEAP	0.31	0.48
GMMAT	0.34	0.51
GEMMA	0.44	0.52
PCA	0.35	0.49

Discussion

In this work, we compared the performance under an EPS design of several mixed model-based association methods for binary phenotypes. We estimated the type I error rate for all methods under both a common variant and a rare variant scenario. We evaluated power for those methods with appropriate type I error control and we compared the computational requirements of the methods. We also applied the methods to a real dataset that was known to have population substructure.

For common variants, our simulations showed that methods based on the generalized linear mixed model (GMMAT), the LMM (GEMMA) and the liability threshold model in conjunction with an LMM (LEAP) all have a type 1 error rate that is close to, or at least not higher than, the specified value. Although, Chen et al. [23] note that the liability threshold mixed models may fail to control the type 1 error rates in the presence of moderate to strong population stratification, we did not observe such inflation in our simulations even when confounding due to population stratification would have been severe. On the other hand, we found that CARAT, which uses a retrospective model and a quasi-likelihood framework, did not adequately control the type 1 error rate. The CARAT method is based on a retrospective approach where the case-control ascertainment is modeled [24]. Though this is an appropriate approach for a true case-control design, extreme sampling represents a different type of ascertainment and therefore the retrospective model may not be appropriate in this case.

For rare variants, the type 1 error rate was inflated relative to the specified level for all statistics implemented in the GLMM-based approach SMMAT. The burden test had type 1 error closest to the nominal value of 0.05, which may be explained by the lower power of the burden test overall relative to the optimized variance component tests like SKAT-O [29] (SMMAT-O). Under population stratification there is a true difference between the genotype distributions in the two extreme groups, though this difference is not due to a causal association between the genetic variant and the phenotype. Therefore, methods that have higher power overall, like SKAT-O, will be more likely to detect this false association. Studies have shown that the inflation due to population stratification is higher with rare variants than with common variants (for example, [30]). Using a SKAT-type method incorporating a mixed model-based correction with common variants has been shown to adequately control the type 1 error under random sampling [31]. However, EPS inflates the effects of population stratification to a greater extent than random sampling [28]; therefore, conclusions drawn about corrections with rare variant approaches under random sampling scenarios may not apply to the EPS setting. In addition, the SNPs we simulated for the GRM calculations were all common. It is possible that in the context of EPS, a mixture of rare and common variants for capturing ancestry might have better type 1 error control since it was shown to be slightly conservative in the random sampling setting [31].

We also investigated the performance of LEAP, GMMAT and GEMMA for detecting genetic variants associated with the extremes of BMI in the prostate cancer case-control study. Although we do not know whether there are true associations in this dataset, we note that LEAP, GMMAT and GEMMA all have different genome-wide *p*-value distributions than the uncorrected results (logistic regression with PLINK), and that the corrected distributions appear to have less overall inflation of the test statistics. However, the results for GMMAT indicate a slight over-correction. In our common candidate variant simulations, we also observe some over-correction with both GMMAT and LEAP at the smaller sample sizes. Therefore, it is possible that the over-correction can be explained by the small sample size of the BMI EPS dataset.

The use of LMMs for binary traits has been discouraged due to the fact that this approach ignores the mean-variance relationship of the binary model and instead assumes a constant relationship [23]. Chen et al. [23] demonstrate both an increase and decrease in false positives with an LMM approach on a stratified asthma dataset by separating cases where the variance of the MAF was higher/lower in one ethnic group relative to remaining groups. In our simulations, GEMMA's LMM approach did not have an inflated false positive rate even under moderate to strong population stratification; though we note that our simulations were not designed to investigate this thoroughly. For example, in our simulations we did not vary the proportion of the full cohort from each subpopulation. In addition, in the real data analysis of the BMI phenotype, the results from GEMMA were actually closest to what would be expected if there were no true associations. Therefore, for both the simulated and real data, GEMMA had very good correction of the false positive rate when compared to the other methods.

A weakness of our simulation is the use of the Balding Nichol's model in simulating genotype data for GRM estimation. The Balding's Nichols's model allows the allele frequencies to differ between the subpopulations and guarantees a specific F_{st} value. However, for a given SNP, the actual allele frequency difference between the two subpopulations is small. In real data, some SNPs are highly differentiated between subpopulations [32]; these types of SNPs would not be simulated under this model.

In this study, we model the extreme phenotypes as binary and use methods suitable for analysing case-control or binary data. However, Barnett et al. [33] point out that analysing extremes as a binary phenotype rather than using the quantitative values might lead to a reduction in power to detect genotype-phenotype associations. However, if using the quantitative phenotype values, the extreme sampling mechanism must be modeled. For example, Lin et al. [34] showed that parameter estimates from the linear model are biased when the quantitative phenotypes are naively analysed without accounting for the sampling. Linear model-based methods that model the quantitative phenotype while accounting for the extreme sampling scheme have been developed [34, 35]. However, no such approach currently exists for the linear mixed model; this is therefore a topic for further research.

Conclusions

The mixed model-based methods for population stratification correction compared in this study do not all perform equally well when the data is taken from an extreme sampling design. For common variants, LEAP, GMMAT and GEMMA all had good type I error rates and power; however, CARAT did not adequately control the type I error rate. In addition, none of the available mixed model approaches for rare variants controlled the type 1 error rate. Therefore, when the data are from an EPS study, care should be taken to ensure that the underlying models used in the methods are suitable to the sampling strategy and to the minor allele frequency of the candidate SNPs. Our study highlights the need for the development of mixed model-based approaches for population stratification correction that model the underlying sampling

structure of the EPS design and are applicable to variants of all frequencies.

Methods

Overview of mixed model-based approaches for correcting for population stratification

In this section, we give a brief overview of mixed models and implementations that incorporate these models to correct for population structure. We focus on approaches that are suitable for binary response variables.

The linear mixed model and the generalized linear mixed model

A linear mixed model (LMM) to account for population substructure and/or hidden relatedness is given by:

$$Y = X\beta + Zb + \epsilon \tag{1}$$

where Y is the vector of phenotype values, X is the design matrix of genetic and non-genetic fixed-effect covariates including a column vector of 1, β is the vector of regression coefficients including the intercept, Z is a known design matrix corresponding to clustering that is the identity matrix in the simplest case and b is the vector of random effects. We assume the random effects, b, are $N(0, \sigma_a^2 G)$ distributed, where G is the known relationship matrix and σ_a^2 is the additive genetic variance, and $\epsilon \sim N(0, \sigma_e^2 I)$, where σ_e^2 is the error variance and I is the identity matrix. Therefore, the distribution of Y is:

$$Y \sim \mathcal{N}\left(X\beta + Zb, \sigma_a^2 G + \sigma_e^2 I\right) \tag{2}$$

We can infer from (2) that the matrix G imposes structure on the covariance matrix of Y; this forms the basis of using LMMs to correct for hidden relatedness in GWAS. With population-based samples, the relationship matrix G is estimated using genome-wide data.

Model (2) can be generalized to handle non-normal response variables. Given a vector of random effects b, the response variable Y is assumed to be from a distribution in the exponential family. That is, for the *ith* response,

$$f_i(y_i|b) = \exp\left\{\frac{y_i\varphi - b^*(\varphi)}{a_i(\phi)} + c_i(y_i, \phi)\right\}$$

where $b^*(.), a_i(.), c_i(.,.)$ are known functions that depend on the underlying distribution of Y, φ is a parameter that is associated with the conditional mean $\mu_i = E(Y_i|b)$, and ϕ is a dispersion parameter which may or may not be known. The linear predictor is $\eta_i = x_i\beta + z_i b$, where x_i and z_i are the covariates for the ith individual and β and b are as previously defined. The mean for individual i, μ_i, is related to the linear predictor via a link function:

$$g(\mu_i) = \eta_i.$$

In particular, the mixed logistic model for a binary response variable is given by

$$\text{logit}(p_i) = x_i\beta + z_i b, \tag{3}$$

where $p_i = \Pr(Y_i = 1|b)$ and x_i, z_i and b are as defined above.

Summary of mixed model implementations

Recently, several mixed model approaches for binary traits have been developed. In this section, we summarize the different approaches that have been implemented, which we classify as (i) approaches using the LMM, (ii) approaches using liability threshold models in conjunction with the LMM, and (iii) GLMM-based approaches. We provide more detail on the liability threshold (ii) and GLMM (iii) approaches since the LMM implementations (i) have been compared and summarized elsewhere [18, 36].

(i) Linear Mixed Model approaches

As previously mentioned, LMMs are used with binary traits even though the response variable is neither normal nor continuous. In order to fit LMMs in the GWAS context, large sample sizes are required to achieve sufficient statistical power. Unfortunately, the computational complexity associated with fitting LMMs increases cubicly with the number of individuals in the model [37]. This motivated the development of several variations of the LMM approach designed to increase computational speed and in turn make large scale GWAS feasible. Existing methods include EMMA [15], EMMAX [38], FASTLMM [16], BOLT-LMM [39, 40], GCTA [41], and GEMMA [17]. Some of these approaches have been designed to handle some specific forms of binary data. For example, BOLT-LMM is able to analyse balanced case-control data at large sample sizes [40].

(ii) Liability threshold models in conjunction with the LMM

In case-control studies, cases are over-sampled relative to the disease prevalence. The liability threshold model (LTM) assumes an underlying but unobserved latent trait that is normally distributed [42, 43]. Individuals with latent trait values beyond a threshold, t, are classified as cases ($Y = 1$) and all others are classified as controls ($Y = 0$). Hence the binary response variable for individual i, can be written as:

$$Y_i = \begin{cases} 1 & \text{if } z_i > t \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where Y_i is the observed binary trait and z_i is the unobserved liability score, which is assumed to be $N(0, 1)$. Since the liability scores are not observed, using the liability threshold model requires first estimating liability

scores for each individual. We now describe two implementations which differ in how the liability scores are estimated.

In the algorithm LEAP [25], the liability for individual i is assumed to be a sum of genetic and environmental components, $z_i = g_i + e_i$, where $g_i = X_i^t \beta_g$, X_i^t is the vector of genotype data and $e_i \sim N(0, \sigma_e^2)$. Estimation of z_i is achieved by first fitting a regularized probit model to estimate the parameters β_g. These are estimated with the maximum a posteriori estimate (MAP), also known as the posterior mode estimator. The liabilities are estimated by $\hat{z}_i = X_i^t \hat{\beta}_g$; these values are then used as the phenotype values for each individual. Tests for association are performed using a linear mixed model since the liabilities are assumed to be normally distributed.

LTMLM [26] is similar to LEAP in that it models the retrospective sampling and uses imputed liability scores; however, the liabilities are estimated using the posterior mean of the multivariate liability distribution (PMLs). A Gibbs sampler is used to sample from this distribution and the posterior mean is estimated by averaging over the Monte Carlo iterations. A score statistic is used to test for association between a candidate SNP and the imputed liabilities assuming a linear mixed model.

A comparison of the estimators used by LEAP and LTMLM showed that in the presence of population structure, the MAP yields more accurate liability estimates than the PML, often at a lower computational cost compared to the posterior mean estimator [25].

(iii) GLMM-based approaches

The logistic mixed model is a special case of the GLMM that can be used to analyse binary traits while accounting for population structure and hidden relatedness. However, this model has not been widely used for GWAS due to the computational complexity involved in fitting logistic mixed models for a large number of genetic variants. Chen et al. [23] developed GMMAT, a logistic mixed model that is computationally efficient enough to handle genome-wide data. GMMAT first fits a null logistic mixed model including fixed effects for any covariates and random effects for residual population stratification and relatedness. This fitted null model, which is the same for all genetic variants in the study, is then used to test for the association between a genetic variant and phenotype using a score test. The use of just one null model for testing all genetic variants greatly simplifies the model compared to fitting a full logistic mixed model for a large GWAS.

CARAT (Case control Retrospective Association Test) [27] is another mixed model approach for binary traits where the response variable is modeled using a mixed effects quasi-likelihood approach. In particular, only the conditional mean and covariance of the response variable given the genotypes and other covariates are specified. The conditional mean is selected to be the same as for the logistic model. The conditional covariance incorporates features of the logistic model and accounts for population substructure through the genetic relationship matrix. Like LTMLM, CARAT uses a retrospective model where the genotypes are treated as random and the association is performed conditional on the phenotypes and non-genetic covariates. However, unlike LTMLM, CARAT does not require the knowledge of disease prevalence. Like LTMLM, a score test is used to handle genome-wide data.

Simulations to evaluate type I error

In this section, we describe the simulation studies used to estimate the type I error rates of the mixed model software implementations that handle binary data. In particular, we focus on LEAP, GMMAT, CARAT and GEMMA as a representative LMM approach. We excluded LTMLM as we found that it took much longer to run than LEAP, which uses a similar liability threshold model.

Common candidate variant

We assumed a cohort consisting of two subpopulations of equal proportion. The total cohort size, N, was set to $5000, 10,000$ or 20,000. The F_{st} value between the two populations - a measure of genetic population differentiation - was set to 0.01; this value is higher than would be expected between typical European populations but it ensures substantial substructure [28].

Genetic data was simulated using the Balding-Nichols method [14, 44] as previously described [28]. For each individual, we simulated a total of $p = 5000$ bi-allelic SNPs. Though true genome-wide data would consist of much larger numbers of SNPs, our previous work with data simulated using this model has shown that this number of SNPs is sufficient to correct for population stratification [28]. We label the two alleles at each SNP as either '0' or '1'. For each SNP, the generating allele frequency, p, for the '1' allele was first sampled from a uniform $[0.1, 0.9]$ distribution. To mimic population differentiation, the '1' allele frequency within each of the two populations, p_1 and p_2, was then sampled from a Beta distribution with shape and scale parameters $\frac{p(1-F_{st})}{F_{st}}$ and $\frac{(1-p)(1-F_{st})}{F_{st}}$, respectively. This approach has been shown to generate genotype data having the desired F_{st} level [44]. Using the allele frequencies generated for each population, the genotype data was sampled assuming Hardy Weinberg equilibrium. The genotype data was coded as 0, 1 or 2 for genotypes 00, 01, and 11, respectively.

We simulated a candidate SNP separately. We first assumed that the '1' allele frequency for the candidate SNP was $p_1 = 0.25$ in the first subpopulation and $p_2 = 0.85$ in the second subpopulation. Although this allele frequency difference is probably not realistic in practice, it was chosen to reflect a 'worst case' scenario of a candidate

SNP showing extreme population differentiation. We also included a smaller simulation where we varied the '1' allele frequency difference between the populations; in particular, we set p_2 to range from $0.5 - 0.9$ while fixing p_1 at 0.5.

In order to obtain the EPS sample, we simulated phenotypes from a normal distribution with mean values $\mu_1 = 0.07$ and $\mu_2 = -0.07$ for subpopulation 1 and 2, respectively, and a common variance of $\sigma^2 = 1$. We note here that the genotypes and phenotypes have been simulated independently, which implies that the genotype at the candidate SNP is not causally associated with the phenotype. The EPS sample was then selected as the individuals in the upper and lower $10th$ percentile of the phenotype distribution. For the EPS sample, the binary response variable is membership in the upper or lower group; in practice, these are sometimes labelled as cases and controls, though it should be noted that there is no true control group in this design.

For methods requiring a GRM, the genetic data on the $p = 5000$ SNPs simulated under the Balding-Nichols method was used to compute the GRM; the candidate SNP for the association test was not included in the GRM calculation. LEAP and GEMMA compute the GRM as part of the algorithm. For GMMAT, the GRM must be computed externally and then passed to the program; we used the standardized GRM computed by GEMMA.

We simulated $m = 3000$ datasets under the scenario where the candidate variant '1' allele frequency was $p_1 = 0.25$ and $p_2 = 0.85$ in subpopulation 1 and 2, respectively. The computational time for CARAT is significantly longer than the other methods, particularly for the large sample sizes. Therefore, we were only able to complete CARAT analysis of $m = 1999$ simulated datasets for the simulation with full cohort size of $N = 10,000$. Due to limited computational time, we only performed $m = 1000$ simulations for each setting under the scenario where $p_1 = 0.5$ and p_2 varied. For these simulations, we chose to focus on the trend in the rate as p_2 varied for each method separately.

GMMAT is available as an R package [23]. LEAP, GEMMA and CARAT are stand-alone software packages that can be run at the command line on a Unix operating system. We used default settings for all packages. For comparison purposes, we also included a PC-based correction by including the top 5 principal components in a logistic regression model; this was also done in R. For each method, the type I error rate was estimated by the proportion of the simulated datasets where the null hypothesis was rejected at level $\alpha = 0.05$. Simulations were run in a cluster computing environment (CAC-FRONTENAC) and all analysis of the results was done in R [45].

Rare candidate variants

We also investigated the performance with a candidate region having rare variants. To simulate data for the candidate region, we simulated haplotype data in a 30kb region using the coalescent-based simulation program *ms* [46]. We simulated a total of 10,000 haplotypes assuming an effective population size of $N_e = 100,000$, a per-site mutation rate of $\mu = 10^{-8}$ and a per-adjacent site recombination rate of $\rho = 10^{-8}$. To incorporate population structure, we again assumed two subpopulations of equal size (i.e. 5000 haplotypes from each subpopulation) and a migration parameter $M = 10$, which is representative of the population differentiation parameter $F_{st} = 0.01$ in the case of a common variant [30]. To create genotypes in the candidate region for $N = 5,000$ individuals, the 10,000 haplotypes were randomly paired within subpopulation. The continuous phenotype values and genetic data at 5000 non-candidate SNPs (for GRM estimation) were generated as previously described for the common variant simulation study.

To test for association with rare variants while accounting for population structure, only the generalized linear model approach had software available. We used SMMAT (variant set mixed model association test) [47], which is a function available in the GMMAT package to perform several popular rare variant tests (burden test [48], SKAT [1], SKAT-O [49], and an efficient hybrid test that combines the burden and SKAT tests [47]) in the binary mixed model framework. We used the default values set in the software for all tests. As with the common variant scenarios, we estimated the type I error by the proportion of tests rejected at level $\alpha = 0.05$.

Simulation to investigate power

For methods that adequately controlled the type I error rate (LEAP, GMMAT, GEMMA, PCA), we conducted additional simulations in order to compare their performance with respect to power. We did not include a rare variant power simulation since none of the methods we tested adequately controlled the type I error rate.

For the power simulations, the genetic data for estimating ancestry was simulated using the same procedure as described for the type I error simulations. In particular, we continue to assume that there is hidden population subdivision. To simulate the candidate SNP, we assumed no differences in allele frequency between the two populations and an allele frequency of 0.2 for the causal allele. The genotypes were sampled assuming Hardy-Weinberg equilibrium. The phenotype was again simulated assuming a normal distribution, with variance 1 and mean $\mu_i + \beta G_{ij}$ where μ_i, the subpopulation means, are the same as for the type I error simulations, G_{ij} is the genotype of individual j in subpopulation i, and β is the effect size of the causal allele ($\beta = 0.15$ and $\beta = 0.25$). We simulated a full cohort size of $N = 10,000$ which gives us an EPS subsample of $n = 2000$ when we select the top and bottom 10%. The number of simulations for power estimation was set

to $m = 3000$ and power was estimated by the proportion of tests rejected at level $\alpha = 0.05$.

Analysis of BMI phenotype from a prostate cancer case-control study

We evaluated the mixed model methods for common variants on data collected from a population-based case-control study, conducted in Montréal, Canada. The study has been described elsewhere (e.g. [50]). Briefly, cases were men aged 76 and under who were newly diagnosed with prostate cancer between 2005-2009; age-matched controls (in 5 year age groups) were randomly recruited from the electoral list of men in the same districts as cases. Overall, 1933 cases and 1994 controls were recruited into the study. Genome-wide genotyping was done using the Illumina OmniExpress 12 platform. We performed quality control which included removing SNPs and individuals with a missingness level above 0.02, minor allele frequency (MAF) below 0.05 and those that deviated from the Hardy-Weinberg equilibrium at a p-value of 10^{-6}. We also checked that all the SNPs used were autosomal (i.e on chromosomes 1-22) and that all reported male individuals had an F value (based on the X chromosome inbreeding estimate) above 0.8. After quality control, genotype data was available on 574,885 SNPs and for 1295 cases and 1248 controls.

Data was collected on several continuous variables within this study. We found that body mass index (BMI) was not associated with prostate cancer status in this study (P-value=0.48); we therefore used this as our continuous phenotype and pooled the cases and controls. We selected those in the top and bottom 15% of BMI in our extreme sampling design. After data cleaning, we observed that 2520 of the men with complete genotype data also had BMI data. With these numbers, the sample size of the final EPS sample was about 756.

The study includes men from different ethnic backgrounds. About 77 of the men were Black, 28 were Asian, 1199 were European and 71 were of other nationalities. The ethnicity of 14 of the total sample collected could not be ascertained and therefore was marked as missing. As we are interested in methods for correcting for population stratification, we did not stratify our analysis by ethnicity.

We performed a GWAS comparing the methods GMMAT, LEAP and GEMMA. We excluded CARAT since we found that it had poor false positive rate correction in our simulations. Since this is a real dataset, we do not know whether there are true associations and whether population stratification is truly a problem. For this reason, we also used PLINK [51] to assess genome-wide association with no population stratification correction as a baseline comparison. For each method, we compute p-values of association for all available SNPs. We summarize the association results with Manhattan plots and we use QQ-plots of $-\log_{10}$ of the p-values to visually assess the inflation of test statistics. Both plots were created using the `qqman` R package [52].

Abbreviations

EPS: Extreme Phenotype Sampling; GWAS: Genome-Wide Assocation Study; SNP: Single Nucleotide Polymorphism; PC/PCA: Principal Components/Principal Component Analysis; LMM: Linear Mixed Model; GLMM: Generalized Linear Mixed Model; LTM: Liability Threshold Model; MAP: Maximum a posteriori estimate; PML: Posterior mean of the Multivariate Liability; MCMC: Markov chain Monte Carlo; GRM: Genetic Relationship Matrix; BMI: Body Mass Index; MAF: Minor Allele Frequency

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-022-08297-y.

Additional file 1: Title: Supplementary Figure 1. Description: Manhattan plot for results obtained from LEAP for the GWAS with the BMI phenotype. The y-axis shows -log10 of the p-values from the test for association between BMI extremes and genotype and the x-axis shows genomic position of the SNP. The blue line indicates the standard threshold for a suggestive association (p-value $< 1 \times 10^{-5}$). The red line indicates the standard threshold for a genome-wide significant association (p-value $< 5 \times 10^{-8}$).

Additional file 2: Title: Supplementary Figure 2. Description: Manhattan plot for results obtained from GMMAT for the GWAS with the BMI phenotype. The y-axis shows -log10 of the p-values from test for association between BMI extremes and genotype and the x-axis shows genomic position of the SNP. The blue line indicates the standard threshold for a suggestive association (p-value $< 1 \times 10^{-5}$). The red line indicates the standard threshold for a genome-wide significant association (p-value $< 5 \times 10^{-8}$).

Additional file 3: Title: Supplementary Figure 3. Description: Manhattan plot for results obtained from GEMMA for the GWAS with the BMI phenotype. The y-axis shows -log10 of the p-values from test for association between BMI extremes and genotype and the x-axis shows genomic position of the SNP. The blue line indicates the standard threshold for a suggestive association (p-value $< 1 \times 10^{-5}$). The red line indicates the standard threshold for a genome-wide significant association (p-value $< 5 \times 10^{-8}$).

Additional file 4: Title: Supplementary Figure 4. Description: Manhattan plot for results obtained from PLINK (uncorrected logistic regression) for the GWAS with the BMI phenotype. The y-axis shows -log10 of the p-values from test for association between BMI extremes and genotype and the x-axis shows genomic position of the SNP. The blue line indicates the standard threshold for a suggestive association (p-value $< 1 \times 10^{-5}$). The red line indicates the standard threshold for a genome-wide significant association (p-value $< 5 \times 10^{-8}$).

Acknowledgements

Computations were performed on resources and with support provided by the Centre for Advanced Computing (CAC) at Queen's University in Kingston, Ontario. The CAC is funded by: the Canada Foundation for Innovation, the Government of Ontario, and Queen's University. The abstract has been published on https://onlinelibrary.wiley.com/doi/10.1002/gepi.22431.

Authors' contributions

MO: designed the study, performed programming for the simulation study, performed data analysis, interpreted data, drafted and revised the manuscript. MEP: conception of the overall PROtEuS study, revised the manuscript; MHRG: guided the analysis, interpreted data, revised the manuscript; KMB: designed the study, guided the analysis, interpreted data, drafted and revised the manuscript. All authors reviewed and approved the final manuscript.

Funding
Student stipend and travel related to this work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) [reference no. RGPIN-2019-06051]. The case-control study was funded by the Canadian Cancer Society (grants 13149, 19500, 19864, and 19865), the Canadian Institutes of Health Research (grant 399507), the Cancer Research Society, the Fonds de Recherche du Québec–Santé (FRQS), the FRQS–Réseau de recherche en santé environnementale, and the Ministère du Développement Économique, de l'Innovation et de l'Exportation du Québec. Part of this work was undertaken while MO held a QEII scholarship. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript

Availability of data and materials
The computational scripts used for the simulation studies are available at https://github.com/statgen-uottawa/Onifade-EPS-PopStrat-GLMM. The prostate cancer case-control dataset that support the findings of this study are available from Dr. Parent but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

Declarations

Ethics approval and consent to participate
This study was approved by the Ethics Committees of the following institutions: Institut national de la recherche scientifique, Centre de Recherche du Centre Hospitalier de l'Université de Montréal, Hôpital Maisonneuve-Rosemont, Hôpital Jean-Talon, Hôpital Fleury, and Hôpital Charles-LeMoyne. All participants provided written informed consent. The prostate cancer case-control dataset was used under license from Dr. Parent.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Mathematics and Statistics, University of Ottawa, Ottawa, Canada. [2]School of Epidemiology and Public Health, University of Ottawa, Ottawa, Canada. [3]Centre Armand-Frappier Santé Biotechnologie, Institut national de la recherche scientifique, Université du Québec, Laval, Canada.

Received: 5 December 2021 Accepted: 6 January 2022
Published online: 04 February 2022

References

1. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011;89(1):82–93.
2. Guey LT, Kravic J, Melander O, Burtt NP, Laramie JM, Lyssenko V, Jonsson A, Lindholm E, Tuomi T, Isomaa B, et al. Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants. Genet Epidemiol. 2011;35(4):236–46.
3. Lander ES, Botstein D. Mapping mendelian factors underlying quantitative traits using rflp linkage maps. Genetics. 1989;121(1):185–99.
4. Petrill SA, Plomin R, McClearn GE, Smith DL, Vignetti S, Chorney MJ, Chorney K, Thompson LA, Detterman DK, Benbow C, et al. No association between general cognitive ability and the a1 allele of the d2 dopamine receptor gene. Behav Genet. 1997;27(1):29–31.
5. Ball D, Hill L, Eley TC, Chorney MJ, Chorney K, Thompson LA, Detterman DK, Benbow C, Lubinski D, Owen M, et al. Dopamine markers and general cognitive ability. Neuroreport. 1998;9(2):347–9.
6. Versmissen J, Oosterveer DM, Yazdanpanah M, Dehghan A, Hólm H, Erdman J, Aulchenko YS, Thorleifsson G, Schunkert H, Huijgen R, et al. Identifying genetic risk variants for coronary heart disease in familial hypercholesterolemia: an extreme genetics approach. Eur J Hum Genet. 2015;23(3):381.
7. Kang G, Lin D, Hakonarson H, Chen J. Two-stage extreme phenotype sequencing design for discovering and testing common and rare genetic variants: efficiency and power. Hum Hered. 2012;73(3):139–47.
8. Peloso GM, Rader DJ, Gabriel S, Kathiresan S, Daly MJ, Neale BM. Phenotypic extremes in rare variant study designs. Eur J Hum Genet. 2016;24(6):924–30.
9. Tong DMH, Hernandez RD. Population genetic simulation study of power in association testing across genetic architectures and study designs. Genet Epidemiol. 2019;0(0):. https://doi.org/10.1002/gepi.22264.
10. Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. Theor Popul Biol. 2001;60(3):155–66.
11. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155(2):945–59.
12. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904.
13. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2(12):190.
14. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 2010;11(7):459.
15. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. Efficient control of population structure in model organism association mapping. Genetics. 2008;178(3):1709–23.
16. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. Fast linear mixed models for genome-wide association studies. Nat Methods. 2011;8(10):833.
17. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012;44(7):821.
18. Eu-Ahsunthornwattana J, Miller EN, Fakiola M, Jeronimo SM, Blackwell JM, Cordell HJ, Wellcome Trust Case Control Consortium 2, et al. Comparison of methods to account for relatedness in genome-wide association studies with family-based data. PLoS Genet. 2014;10(7):1004445.
19. Fakiola M, Strange A, Cordell HJ, Miller EN, Pirinen M, Su Z, Mishra A, Mehrotra S, Monteiro GR, Band G, et al. Common variants in the hla-drb1–hla-dqa1 hla class ii region are associated with susceptibility to visceral leishmaniasis. Nat Genet. 2013;45(2):208.
20. Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, Moutsianas L, Dilthey A, Su Z, Freeman C, Hunt SE, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature. 2011;476(7359):214.
21. Tsoi LC, Spain SL, Knight J, Ellinghaus E, Stuart PE, Capon F, Ding J, Li Y, Tejasvi T, Gudjonsson JE, et al. Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. Nat Genet. 2012;44(12):1341.
22. Pirinen M, Donnelly P, Spencer CC, et al. Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. Ann Appl Stat. 2013;7(1):369–90.
23. Chen H, Wang C, Conomos MP, Stilp AM, Li Z, Sofer T, Szpiro AA, Chen W, Brehm JM, Celedón JC, 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(4):653–66.
24. Jiang D, Mbatchou J, McPeek MS. Retrospective association analysis of binary traits: overcoming some limitations of the additive polygenic model. Hum Hered. 2015;80(4):187–95.
25. Weissbrod O, Lippert C, Geiger D, Heckerman D. Accurate liability estimation improves power in ascertained case-control studies. Nat Methods. 2015;12(4):332.
26. Hayeck TJ, Zaitlen NA, Loh P-R, Vilhjalmsson B, Pollack S, Gusev A, Yang J, Chen G-B, Goddard ME, Visscher PM, Patterson N, Price AL. Mixed Model with Correction for Case-Control Ascertainment Increases Association Power. Am J Hum Genet. 2015;96(5):720–30. https://doi.org/10.1016/j.ajhg.2015.03.004.
27. Jiang D, Zhong S, McPeek MS. Retrospective binary-trait association test elucidates genetic architecture of crohn disease. Am J Hum Genet. 2016;98(2):243–55.
28. Panarella M, Burkett KM. A cautionary note on the effects of population stratification under an extreme phenotype sampling design. Front Genet. 2019;10:398.
29. Moutsianas L, Agarwala V, Fuchsberger C, Flannick J, Rivas MA, Gaulton KJ, Albers PK, Consortium G, McVean G, Boehnke M, Altshuler D, McCarthy MI. The Power of Gene-Based Rare Variant Methods to Detect Disease-Associated Variation and Test Hypotheses About Complex Disease. PLOS Genet. 2015;11(4):1005165. https://doi.org/10.1371/journal.pgen.1005165.

30. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. Nat Genet. 2012;44(3):243.
31. Luo Y, Maity A, Wu MC, Smith C, Duan Q, Li Y, Tzeng J-Y. On the substructure controls in rare variant analysis: Principal components or variance components?. Genet Epidemiol. 2018;42(3):276–87.
32. Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, Kittles R, Alarcon-Riquelme ME, Gregersen PK, Belmont JW, De La Vega FM, Seldin MF. Ancestry Informative Marker Sets for Determining Continental Origin and Admixture Proportions in Common Populations in America. Hum Mutat. 2009;30(1):69–78.
33. Barnett IJ, Lee S, Lin X. Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. Genet Epidemiol. 2013;37(2):142–51.
34. Lin D-Y, Zeng D, Tang Z-Z. Quantitative trait analysis in sequencing studies under trait-dependent sampling. Proc Natl Acad Sci. 2013;110(30): 12247–52.
35. Huang B, Lin DY. Efficient association mapping of quantitative trait loci with selective genotyping. Am J Hum Genet. 2007;80(3):567–76.
36. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. Advantages and pitfalls in the application of mixed-model association methods. Nat Genet. 2014;46(2):100.
37. Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, et al. Mixed linear model approach adapted for genome-wide association studies. Nat Genet. 2010;42(4):355.
38. Kang HM, Sul JH, Zaitlen NA, Kong S.-y., Freimer NB, Sabatti C, Eskin E, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010;42(4):348.
39. Loh P-R, Kichaev G, Gazal S, Schoech AP, Price AL. Mixed-model association for biobank-scale datasets. Nat Genet. 2018;50(7):906.
40. Loh P-R, Tucker G, Bulik-Sullivan BK, Vilhjalmsson BJ, Finucane HK, Salem RM, Chasman DI, Ridker PM, Neale BM, Berger B, et al. Efficient bayesian mixed-model analysis increases association power in large cohorts. Nat Genet. 2015;47(3):284.
41. Yang J, Lee SH, Goddard ME, Visscher PM. Gcta: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88(1):76–82.
42. Falconer DS. The inheritance of liability to certain diseases, estimated from the incidence among relatives. Ann Hum Genet. 1965;29(1):51–76.
43. Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet. 2011;88(3):294–305.
44. Balding DJ, Nichols RA. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica. 1995;96(1-2):3–12.
45. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2018. https://www.R-project.org.
46. Hudson RR. Generating samples under a wright–fisher neutral model of genetic variation. Bioinformatics. 2002;18(2):337–8.
47. Chen H, Huffman JE, Brody JA, Wang C, Lee S, Li Z, Gogarten SM, Sofer T, Bielak LF, Bis JC, et al. Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. Am J Hum Genet. 2019;104(2):260–74.
48. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009;5(2):1000384.
49. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Team ELP, Christiani DC, Wurfel MM, Lin X, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet. 2012;91(2):224–37.
50. Blanc-Lapierre A, Spence A, Karakiewicz PI, Aprikian A, Saad F, Parent M.-É.. Metabolic syndrome and prostate cancer risk in a population-based case–control study in montreal, canada. BMC Public Health. 2015;15(1): 913.
51. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
52. Turner SD. qqman: an r package for visualizing gwas results using q-q and manhattan plots. bioRxiv. 2014. https://doi.org/10.1101/005165.

Publisher's Note

2.2.1 Conclusion

In this work, we compared the type 1 error rate under confounding due to population stratification and the power of several mixed model-based association methods for binary traits under an EPS design. The results from this analysis showed that for common variants, methods like the LEAP, GMMAT and GEMMA all had type 1 error rates close to the nominal values and similar power. However, for rare variants, none of these methods adequately corrected false positive rates due to confounding. The results from this study highlights the need to develop mixed model based approaches that will correct population substructure in the EPS setting, in particular in the rare variant setting.

In the next chapter, we present our work developing a novel linear mixed model method for genetic association analyses that is applicable to the EPS setting.

Chapter 3

An EM algorithm for Linear Mixed Models with missing genotype data

3.1 Introduction

In the previous chapter, we investigated the performance of existing mixed model based approaches to correct for the confounding due to population stratification when the data comes from an EPS study. The results from that study highlighted the importance of developing a linear mixed model for correcting for population substructure when we have EPS data. In this chapter, we introduce a novel EM implementation designed to estimate genetic effects within a linear mixed model framework when there is missing genotype data. The motivation to implement this method was its application to genotype data that is missing due to EPS sampling; however, it is also applicable to sporadically missing genotype data. The EM algorithm [13] is a widely used tool for maximum likelihood estimation when there is missing data.

The chapter is structured as follows: Section 3.2 provides an overview of pa-

rameter estimation in a LMM. In Section 3.3 we derive an EM algorithm tailored to missing categorical genetic covariates. Section 3.4 describes how to compute appropriate Monte Carlo based likelihood ratio and Wald tests for assessing the hypothesis of no association.

3.2 The Linear Mixed Model

Recall from section 1.4, the linear mixed model to correct for population structure in genetic data analysis can be expressed as:

$$\mathbf{y} = \boldsymbol{G\beta_g} + \boldsymbol{X\beta_x} + \boldsymbol{Zb} + \boldsymbol{\epsilon}, \tag{3.2.1}$$

where $\mathbf{y}$ is a $n \times 1$ vector of observed phenotypes, $\boldsymbol{G}$ is a $n \times l$ known design matrix for the l genetic covariates, $\boldsymbol{\beta_g}$ is a $l \times 1$ vector representing the coefficients of the genetic effects, $\boldsymbol{X}$ is a $n \times q$ matrix of the non-genetic covariates with $\boldsymbol{\beta_x}$ the $q \times 1$ vector representing the coefficients of the fixed effects. $\boldsymbol{Z}$ is a known $n \times p$ matrix of covariates for the $p \times 1$ vector of random effect $\boldsymbol{b}$. In our modelling context, we assume that $\boldsymbol{Z}$ is the $n \times n$ identity matrix unless there is known clustering of individuals. In linear mixed models, the fixed effects are used for modelling the mean of $\mathbf{y}$ while the random effects govern the variance-covariance structure of $\mathbf{y}$. ϵ is a $n \times 1$ vector of model error. It is typically assumed that ϵ and $\boldsymbol{b}$ are independent and they jointly form the random components. They are distributed as: $\boldsymbol{b} \sim N_n(0, \sigma_b^2 \boldsymbol{K})$ and $\epsilon \sim N_n(0, \sigma_e^2 \boldsymbol{I_n})$. $N_n(0, \sigma_b^2 \boldsymbol{K})$ denotes the n-variate normal distribution with a $(n \times 1)$ mean vector of 0's and $n \times n$ covariance matrix $\boldsymbol{K}$, which is the genetic relationship matrix (GRM). The GRM models the relationship between individuals and is assumed to be a known symmetric and positive definite matrix. σ_b^2 is the unknown

variance component corresponding to population structure. The matrix $\boldsymbol{I}_n$ is the identity matrix and σ_e^2 is the model error. The overall variance-covariance matrix can be represented as $V_y = \sigma_b^2 \boldsymbol{ZKZ'} + \sigma_e^2 \boldsymbol{I_n}$.

The conditional distribution of $\mathbf{y}$ given the random effect $\boldsymbol{b}$ is given as:

$$\mathbf{y}|\theta, \boldsymbol{b} \sim N_n \left(\boldsymbol{G\beta_g} + \boldsymbol{X\beta_x} + \boldsymbol{Zb}, \sigma_e^2 \boldsymbol{I}_n \right) \tag{3.2.2}$$

where $\theta = (\boldsymbol{\beta_g}, \boldsymbol{\beta_x}, \sigma_g^2, \sigma_e^2)$, is the set of the model parameters. The marginal distribution of $\boldsymbol{y}$ is obtained by integrating out the random effect $\boldsymbol{b}$ from (3.2.1), which gives:

$$\mathbf{y}|\theta \sim N_n(\boldsymbol{G\beta_g} + \boldsymbol{X\beta_x}, \sigma_b^2 \boldsymbol{ZKZ'} + \sigma_e^2 \boldsymbol{I_n}) \tag{3.2.3}$$

In our setting, the variance components are regarded as nuisance parameters and the fixed effects are the parameters of interest. In the next section, we will discuss the methods of obtaining estimates for the $\boldsymbol{\beta}$'s , σ_b^2 and σ_e^2.

3.2.1 Parameter Estimation in a LMM (Complete Case)

In this section, we review methods for estimating the regression parameters and variance components in linear mixed models with no missing data. In the first approach, we present the direct approach to maximum likelihood estimation used for estimation of the fixed effects and in the second approach, we give an EM algorithm for finding the variance components that treat the random effects $\boldsymbol{b}$ as missing data.

3.2.2 Maximum Likelihood Estimation of Fixed Effects

The marginal density function of the data vector $\mathbf{y}|\theta \sim N(\boldsymbol{G\beta_g} + \boldsymbol{X\beta_x}, \sigma_b^2 \boldsymbol{ZKZ'} + \sigma_e^2 \boldsymbol{I_n})$ viewed as a function of the parameters $(\boldsymbol{\beta_g}, \boldsymbol{\beta_x}, \sigma_b^2, \sigma_e^2)$ is known as the likelihood function. This is given as:

$$L = L(\boldsymbol{\beta_g}, \boldsymbol{\beta_x}, \sigma_b^2, \sigma_e^2|\mathbf{y}) = \frac{e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{G\beta_g}-\boldsymbol{X\beta_x})'\mathbf{V_y}^{-1}(\boldsymbol{y}-\boldsymbol{G\beta_g}-\boldsymbol{X\beta_x})}}{(2\pi)^{\frac{1}{2}n}|\boldsymbol{V_y}|^{\frac{1}{2}}} \tag{3.2.4}$$

where $V_y = \sigma_b^2 \mathbf{ZKZ'} + \sigma_e^2 \boldsymbol{I_n}$. Maximum likelihood estimation maximizes (3.2.4) with respect to $\boldsymbol{\beta_g}$, $\boldsymbol{\beta_x}$, σ_b^2 and σ_e^2. By direct derivation, we will maximize the logarithm of (3.2.4) which we will denote as ℓ. For convenience in deriving estimates, let $\boldsymbol{X}^* = (\boldsymbol{G}, \boldsymbol{X})$ be the combined design matrix of fixed effects of dimension $(n \times (q + l))$ and $\boldsymbol{\beta}^* = (\boldsymbol{\beta_g}, \boldsymbol{\beta_x})$ the combined regression parameter vector. The corresponding log-likelihood is then:

$$\begin{aligned} \ell(\theta) &= \log(L) \\ &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{V_y}| - \frac{1}{2}\left(\mathbf{y} - \boldsymbol{X^*\beta^*}\right)' \boldsymbol{V_y}^{-1}(\mathbf{y} - \mathbf{X}^*\boldsymbol{\beta^*}) \end{aligned} \tag{3.2.5}$$

To maximize $\ell(\theta)$, we differentiate equation (3.2.5) first with respect to $\boldsymbol{\beta^*}$ using well known statistical properties of the quadratic form, which are stated below for reference. For a vector $\mathbf{x}$ and a symmetric matrix $\mathbf{A}$, the derivative of $\mathbf{x'Ax}$ with respect to $\mathbf{x}$ is given by:

$$\begin{aligned} \frac{\partial}{\partial \mathbf{x}}(\mathbf{x'Ax}) &= \frac{\partial \mathbf{x'}}{\partial \mathbf{x}}(\mathbf{Ax}) + \frac{\partial}{\partial \mathbf{x}}(\mathbf{Ax})'\mathbf{x} \\ &= \mathbf{Ax} + \mathbf{A'x} \end{aligned} \tag{3.2.6}$$

If $\mathbf{A}$ is symmetric, which is usually the case for statistical models, then

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{x}'\mathbf{A}\mathbf{x}) = 2\mathbf{A}\mathbf{x}.$$

Using equation (3.2.6), the derivative of equation (3.2.5) with respect to β^* is:

$$l_\beta = \frac{\partial \ell}{\partial \boldsymbol{\beta^*}} = (\mathbf{X}^*)'\boldsymbol{V_y}^{-1}\mathbf{y} - (\mathbf{X}^*)'\mathbf{V_y}^{-1}\mathbf{X}^*\boldsymbol{\beta^*} \tag{3.2.7}$$

For differentiating equation (3.2.5) with respect to σ_b^2 and σ_e^2, we use the following matrix relations governing the derivative of an inverse and determinants of a matrix $\mathbf{A}$:

- With a scalar t, define:
$$\frac{\partial \mathbf{A}}{\partial t} = \left\{ \frac{\partial a_{ij}}{\partial t} \right\}.$$
If $\mathbf{A}$ is non-singular then: $\mathbf{A}\mathbf{A}^{-1} = \boldsymbol{I}$ gives
$$\frac{\partial \mathbf{A}}{\partial t}\mathbf{A}^{-1} + \mathbf{A}\frac{\partial \mathbf{A}^{-1}}{\partial t} = 0$$
and so:
$$\frac{\partial \mathbf{A}^{-1}}{\partial t} = -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial t}\mathbf{A}^{-1}.$$

- This next identity was obtained from Searle and Khudri [63]. Suppose that elements of $\mathbf{A}$ are functions of the scalar y, then;
$$\frac{\partial}{\partial y}(\log|A|) = \frac{1}{|\mathbf{A}|}\frac{\partial |\mathbf{A}|}{\partial y} = \operatorname{tr}\left(\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial y}\right) \tag{3.2.8}$$

Using these properties, we differentiate equation (3.2.5) with respect to σ_b^2 and σ_e^2.

We obtain the derivative with respect to σ_b^2 as:

$$\frac{\partial \ell}{\partial \sigma_b^2} = \frac{1}{2}\left\{(\mathbf{y} - \mathbf{X}^*\boldsymbol{\beta}^*)'\boldsymbol{V_y}^{-1}\mathbf{ZKZ}'\mathbf{V_y}^{-1}(\mathbf{y} - \mathbf{X}^*\boldsymbol{\beta^*}) - \mathrm{tr}\left(\mathbf{V}_y^{-1}\mathbf{ZKZ}'\right)\right\}, \tag{3.2.9}$$

which was obtained by using:

$$\frac{\partial \boldsymbol{V_y}}{\partial \sigma_b^2} = \boldsymbol{ZKZ}' \tag{3.2.10}$$

The derivative with respect to σ_e^2 is given as:

$$\begin{aligned}\frac{\partial \ell}{\partial \sigma_e^2} &= \frac{1}{2}\left\{(\boldsymbol{y} - \boldsymbol{X^*\beta}^*)'\boldsymbol{V_y}^{-1}\boldsymbol{V_y}^{-1}(\boldsymbol{y} - \boldsymbol{X^*\beta}^*) - \mathrm{tr}\left(\boldsymbol{V_y}^{-1}\right)\right\} \\ &= \frac{1}{2}\left\{(\boldsymbol{y} - \boldsymbol{X^*\beta}^*)'(\boldsymbol{V_y}^{-1})^2(\boldsymbol{y} - \boldsymbol{X^*\beta}^*) - \mathrm{tr}\left(\boldsymbol{V_y}^{-1}\right)\right\}\end{aligned} \tag{3.2.11}$$

The procedure for finding the MLE is to solve the system of equations $\frac{\partial \ell}{\partial \boldsymbol{\beta^*}} = 0$, $\frac{\partial \ell}{\partial \sigma_b^2} = 0$ and $\frac{\partial \ell}{\partial \sigma_e^2} = 0$. Assuming that $\hat{\sigma}_b^2, \hat{\sigma}_e^2$ are the MLE's for $\hat{\sigma}_b^2$ and $\hat{\sigma}_e^2$ found by from setting equations (3.2.7), (3.2.9) and (3.2.11) to 0 and solving for $\boldsymbol{\beta}$, we obtain:

$$\hat{\boldsymbol{\beta}}^{\boldsymbol{*}} = \{(\boldsymbol{X}^*)'\hat{\mathbf{V}}_{\mathbf{y}}^{-\mathbf{1}}(\mathbf{X}^*)^{-1}\}(\boldsymbol{X}^*)'\hat{\mathbf{V}}_{\mathbf{y}}^{-\mathbf{1}}\mathbf{y} \tag{3.2.12}$$

where $\hat{\mathbf{V}}_{\mathbf{y}} = \mathbf{V}(\hat{\sigma}_b^2, \hat{\sigma}_e^2)$ is $\mathbf{V}_y$ with the variance components replaced by their MLE. Hence, once the MLE of the variance components are found, the MLE of the fixed effects are calculated using the closed form expression in equation (3.2.12).

To find the variance components, we discuss an EM algorithm for obtaining the maximum likelihood estimates of the linear mixed model with complete data.

3.2.3 Estimation using an EM approach

For the linear mixed model defined in equation (3.2.1), the complete data is $\boldsymbol{w} = (\boldsymbol{y}, \boldsymbol{b})$ with the assumption that the random effect $\boldsymbol{b}$ is the "missing" data and $\boldsymbol{y}$ is the observed data. The distribution of the complete data $\boldsymbol{w}$ can be written as a product of the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{b}$ and the marginal distribution of $\boldsymbol{b}$

$$f(\boldsymbol{w}) = f(\boldsymbol{y}, \boldsymbol{b}) = f(\boldsymbol{y} \mid \boldsymbol{b})\ f(\boldsymbol{b})$$

where $\boldsymbol{y} \mid (\boldsymbol{\beta_g}, \boldsymbol{\beta_x}, \sigma_e^2), \boldsymbol{b} \sim N_n(\boldsymbol{G\beta_g} + \boldsymbol{X\beta_x} + \boldsymbol{Zb}, \sigma_e^2 \boldsymbol{I_n})$ and $\boldsymbol{b} \sim N_n(0, \sigma_b^2 \boldsymbol{K})$.
The complete data log likelihood can now be written as:

$$\ell_c(\boldsymbol{\theta}) = \log\left(f(\boldsymbol{y}, \boldsymbol{b}|\boldsymbol{\theta})\right) \tag{3.2.13}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta_g}, \boldsymbol{\beta_x}, \sigma_b^2, \sigma_e^2)$

$$\begin{aligned}
\ell_c(\boldsymbol{\theta}) &= \log(f(\boldsymbol{y}, \boldsymbol{b}|\boldsymbol{\beta_g}, \boldsymbol{\beta_x}, \sigma_b^2, \sigma_e^2) \\
&= \log(f(\mathbf{y}|\boldsymbol{\beta_g}, \boldsymbol{\beta_x}, \sigma_e^2, \mathbf{b})) + \log f(\mathbf{b}|\sigma_b^2 \boldsymbol{K}) \\
&= \Big(-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma_e^2) \\
&- \frac{1}{2\sigma_e^2}(\boldsymbol{y} - \boldsymbol{G\beta_g} - \boldsymbol{X\beta_x} - \boldsymbol{Zb})'(\boldsymbol{y} - \boldsymbol{G\beta_g} - \boldsymbol{X\beta_x} - \boldsymbol{Zb})\Big) \\
&+ \left(-\frac{q}{2} log(2\pi) - \frac{1}{2}\log|\sigma_b^2 \boldsymbol{K}| - \frac{1}{2}\boldsymbol{b}'(\sigma_b^2 \boldsymbol{K})^{-1}\boldsymbol{b} \right)
\end{aligned} \tag{3.2.14}$$

In the E-step, for the expectation of $\ell_c(\boldsymbol{\theta})$ given the observed data and current parameter estimates $\boldsymbol{\theta^{(t)}}$ it has been shown that we need only the expected value of the

sufficient statistics [13]; the complete data sufficient statistics for σ_e^2 and σ_b^2 are:

$$\boldsymbol{e}'\boldsymbol{e} = (\boldsymbol{y} - \boldsymbol{G}\boldsymbol{\beta_g} - \boldsymbol{X}\boldsymbol{\beta_x} - \boldsymbol{Z}\boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{G}\boldsymbol{\beta_g} - \boldsymbol{X}\boldsymbol{\beta_x} - \boldsymbol{Z}\boldsymbol{b}) \text{ and}$$

$$\boldsymbol{b}'\,\boldsymbol{b}.$$

for σ_e^2 and σ_b^2, respectively. We therefore require $E(\boldsymbol{b}'\boldsymbol{b}|\boldsymbol{y}, \boldsymbol{\theta}^{(t)})$ and $E(\boldsymbol{e}'\boldsymbol{e}|\boldsymbol{y}, \boldsymbol{\theta}^{(t)})$. In the next section, we show that for the $(t+1)^{th}$ iteration, these can be written as:

$$\begin{aligned} E\Big(\boldsymbol{b}'\,\boldsymbol{b} \mid \boldsymbol{y}, \boldsymbol{\beta_g}^{(t)}, \boldsymbol{\beta_x}^{(t)}, (\sigma_b^2)^{(t)}, (\sigma_e^2)^{(t)}\Big) &= E\Big(\boldsymbol{b}|\boldsymbol{y}, \boldsymbol{\beta}_g^{(t)}, \boldsymbol{\beta}_x^{(t)}, (\sigma_b^2)^{(t)}, (\sigma_e^2)^{(t)}\Big) \\ &\quad E\Big(\boldsymbol{b}'|\boldsymbol{y}, \boldsymbol{\beta}_g^{(t)}, \boldsymbol{\beta}_x^{(t)}, (\sigma_b^2)^{(t)}, (\sigma_e^2)^{(t)}\Big) \\ &\quad + \text{Var}\Big(\boldsymbol{b}|\boldsymbol{y}, \boldsymbol{\beta}_g^{(t)}, \boldsymbol{\beta}_x^{(t)}, (\sigma_b^2)^{(t)}, (\sigma_e^2)^{(t)}\Big) \\ E\Big(\boldsymbol{e}'\boldsymbol{e}|\boldsymbol{y}, \boldsymbol{\beta}_g^{(t)}, \boldsymbol{\beta}_x^{(t)}, (\sigma_b^2)^{(t)}, (\sigma_e^2)^{(t)}\Big) &= \text{tr } E\Big(\boldsymbol{e}'\boldsymbol{e}|\boldsymbol{y}, \boldsymbol{\beta}_g^{(t)}, \boldsymbol{\beta}_x^{(t)}, (\sigma_b^2)^{(t)}, (\sigma_e^2)^{(t)}\Big). \end{aligned} \tag{3.2.15}$$

and we will provide expressions for each expectation and variance.

The M-step maximises the expected complete data log-likelihood to find the current parameter estimates. In the complete data context, the MLE's have been shown [13, 25] to be:

$$\begin{aligned} \hat{\sigma_e^2} &= \frac{\boldsymbol{e}'\boldsymbol{e}}{n} \qquad \text{and} \\ \hat{\sigma_b^2} &= \frac{\boldsymbol{b}'\boldsymbol{b}}{q} \end{aligned} \tag{3.2.16}$$

For the M step, we therefore replace $\boldsymbol{e}'\boldsymbol{e}$ and $\boldsymbol{b}'\boldsymbol{b}$ with their expected values, i.e $E(\boldsymbol{e}'\boldsymbol{e}|\boldsymbol{y}, \boldsymbol{\theta}^{(t)})$ and $E(\boldsymbol{b}'\boldsymbol{b}|\boldsymbol{y}, \boldsymbol{\theta}^{(\boldsymbol{t})})$ in equation 3.2.15 to estimate $(\sigma_b^2)^{(t+1)}$ and $(\sigma_e^2)^{(t+1)}$. After obtaining these parameter estimates, the variance $\boldsymbol{V_y}^{(t+1)}$ can now be calculated as:

$$\boldsymbol{V_y}^{(t+1)} = (\sigma_b^2)^{(t+1)}\boldsymbol{Z}\boldsymbol{K}\boldsymbol{Z}' + (\sigma_e^2)^{(t+1)}\boldsymbol{I_n}$$

and the fixed effects parameters are found using equation 3.2.12. The E and M steps are repeated until convergence has been reached.

To summarize, the EM algorithm for computing maximum likelihood estimates of the parameters of a linear mixed model with no missing data is as follows:

1. Let $t = 0$, choose initial values for $\beta^{*(0)}$, $(\sigma_b^2)^{(0)}$ and $(\sigma_e^2)^{(0)}$.

2. At the $(t+1)^{th}$ step:

   - E-Step: compute $E\left(\boldsymbol{b'b} \mid \boldsymbol{y}, \boldsymbol{\beta_g}, \boldsymbol{\beta_x}, \sigma_b^2, \sigma_e^2\right)$ and $E\left(\boldsymbol{e'e}|\boldsymbol{y}, \boldsymbol{\beta}_g, \boldsymbol{\beta}_x, \sigma_b^2, \sigma_e^2\right)$ using formulas in equation (3.2.28) and (3.2.30) derived in the next section.

   - M-Step: estimate the parameters:

$$
\begin{aligned}
(\sigma_e^2)^{(t+1)} &= \frac{E\left(\boldsymbol{e'e}|\boldsymbol{y}, \boldsymbol{\beta_g^{(t)}}, \boldsymbol{\beta_x^{(t)}}, (\sigma_b^2)^{(t)}, (\sigma_e^2)^{(t)}\right)}{n} \\
(\sigma_b^2)^{(t+1)} &= \frac{E\left(\boldsymbol{b'b} \mid \boldsymbol{y}, \boldsymbol{\beta_g^{(t)}}, \boldsymbol{\beta_x^{(t)}}, (\sigma_b^2)^{(t)}, (\sigma_e^2)^{(t)}\right)}{q} \\
\boldsymbol{V_y^{(t+1)}} &= (\sigma_b^2)^{(t+1)}\boldsymbol{ZKZ'} + (\sigma_e^2)^{(t+1)}\boldsymbol{I_n} \\
\boldsymbol{\beta^{*(t+1)}} &= \{(\boldsymbol{X}^*)'(\boldsymbol{V_y^{(t+1)}})^{-1}(X^*)^{-1}\}(\boldsymbol{X}^*)'(\boldsymbol{V_y^{(t+1)}})^{-1}\mathbf{y}
\end{aligned}
\tag{3.2.17}
$$

3. Repeat step 2 until convergence. The values obtained at convergence are the MLEs.

The mathematical details for obtaining the expected values of the sufficient statistics are now derived in the next section.

3.2.4 Expected values of the sufficient statistics given the observed data

In this section, we discuss the method for finding the mean and covariance matrix of the conditional distributions of $\boldsymbol{y}$ and $\boldsymbol{b}$ which are required to compute (3.2.15). Recall that the expected value $E(\boldsymbol{y})$ and variance of $\boldsymbol{y}, \text{Var}(\boldsymbol{y})$ for the marginal model $\boldsymbol{y}|\boldsymbol{\beta_g}, \boldsymbol{\beta_x}, \sigma_g^2, \sigma_e^2$ are:

$$\begin{aligned} E(\mathbf{y}) &= \mu_y = \boldsymbol{G\beta_g} + \boldsymbol{X\beta_x} \\ \text{Var}(\boldsymbol{y}) &= V_{\boldsymbol{y}} = \sigma_b^2 \boldsymbol{ZKZ'} + \sigma_e^2 \boldsymbol{I}_n \end{aligned} \tag{3.2.18}$$

and that the distribution of $\boldsymbol{y}$ is:

$$\mathbf{y}|\boldsymbol{\beta}_g, \boldsymbol{\beta}_x, \sigma_g^2, \sigma_e^2 \sim N_n(\mu_y, \mathbf{V_y})$$

The random effects $\boldsymbol{b}$ are regarded as missing with distribution:

$$\boldsymbol{b} \sim N_n(\mathbf{0}, \boldsymbol{V}_b)$$

where $\boldsymbol{V}_b = \sigma_b^2 \boldsymbol{K}$.

Hence the complete data vector $\boldsymbol{w}$ has a multivariate distribution with mean

$$\mu_w = (\mu_y, \mathbf{0})' = (\boldsymbol{G\beta_g} + \boldsymbol{X\beta_x}, \mathbf{0})'$$

and covariance matrix :

$$\boldsymbol{V}_w = \begin{bmatrix} \boldsymbol{V}_y & \boldsymbol{V}_{yb} \\ \boldsymbol{V}_{by} & \boldsymbol{V}_b \end{bmatrix} \tag{3.2.19}$$

where $\mathbf{V}_y$ is given in the second line of equation (3.2.18) and $\boldsymbol{V}_b = \sigma_b^2 \boldsymbol{K}$. The off diagonals of equation 3.2.19 are:

$$\begin{aligned} \boldsymbol{V}_{yb} =& \text{Cov}(\boldsymbol{G\beta_g} + \boldsymbol{X\beta_x} + \boldsymbol{Zb} + \boldsymbol{e}, \boldsymbol{b}) \\ =& \text{Cov}(\boldsymbol{Zb}, \boldsymbol{b}) \\ =& \boldsymbol{Z}E(\boldsymbol{bb'}) \\ =& \boldsymbol{ZV}_b = \boldsymbol{Z}(\sigma_b^2 \boldsymbol{K}) \\ =& \sigma_b^2 \boldsymbol{ZK} \end{aligned} \tag{3.2.20}$$

since $\boldsymbol{b}$ and $\boldsymbol{e}$ are independent, and $\boldsymbol{V}_{by} = \boldsymbol{V}'_{yb} = \sigma_b^2 \boldsymbol{Z'K}$.

Hence $\boldsymbol{V}_w$ is:

$$\boldsymbol{V}_w = \begin{bmatrix} \sigma_b^2 \boldsymbol{ZKZ'} + \sigma_e^2 \boldsymbol{I} & \sigma_b^2 \boldsymbol{ZK} \\ \sigma_b^2 \boldsymbol{KZ'} & \sigma_b^2 \boldsymbol{K} \end{bmatrix} = \begin{bmatrix} \sigma_b^2 (\boldsymbol{ZKZ'} + \frac{\sigma_e^2}{\sigma_b^2} \boldsymbol{I}) & \sigma_b^2 \boldsymbol{KZ} \\ \sigma_b^2 \boldsymbol{KZ'} & \sigma_b^2 \boldsymbol{K} \end{bmatrix} \tag{3.2.21}$$

The complete-data log likelihood is given by:

$$\begin{aligned} \ell_c(\boldsymbol{\theta}) &= \log f(\boldsymbol{w}|(\boldsymbol{\mu}_w, \boldsymbol{V}_w)) \\ &= -\frac{1}{2} \log |\boldsymbol{V}_w| - \frac{1}{2} (\boldsymbol{w} - \boldsymbol{\mu}_w)' \boldsymbol{V}_w^{-1} (\boldsymbol{w} - \boldsymbol{\mu}_w) \end{aligned} \tag{3.2.22}$$

The next step would be to compute the conditional distribution of $\boldsymbol{b}$ given the observed data $\boldsymbol{y}$. We will use the following notation:

$$\begin{aligned} \boldsymbol{E}(\boldsymbol{b}|\boldsymbol{y}) &\equiv \mu_{\boldsymbol{b}|\boldsymbol{y}} \text{ and} \\ \boldsymbol{V}(\boldsymbol{b}|\boldsymbol{y}) &\equiv \boldsymbol{V}_{\boldsymbol{b}|\boldsymbol{y}} \end{aligned} \tag{3.2.23}$$

Using facts from multivariate normal theory, we are able to find the mean and variance of the conditional distribution $\boldsymbol{b} \mid \boldsymbol{y}$. Generally, for two normal random variables $\boldsymbol{x}$ and $\boldsymbol{y}$ such that:

$$\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} \sim \boldsymbol{N} \left(\begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{bmatrix}, \begin{bmatrix} \boldsymbol{V}_x & \boldsymbol{V}_{xy} \\ \boldsymbol{V}_{yx} & \boldsymbol{V}_y \end{bmatrix} \right),$$

then

$$\boldsymbol{x} \mid \boldsymbol{y} \sim N\left(\boldsymbol{\mu}_{x|\boldsymbol{y}}, \boldsymbol{V}_{x|y}\right),$$

where

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y}} &= \boldsymbol{\mu}_{\boldsymbol{x}} + \boldsymbol{V}'_{\boldsymbol{yx}} \boldsymbol{V}_{\boldsymbol{y}}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{y}}), \\ \boldsymbol{V}_{\boldsymbol{x}|\boldsymbol{y}} &= \boldsymbol{V}_{\boldsymbol{x}} - \boldsymbol{V}'_{\boldsymbol{yx}} \boldsymbol{V}_{\boldsymbol{y}}^{-1} \boldsymbol{V}_{\boldsymbol{yx}}, \end{aligned} \tag{3.2.24}$$

$\boldsymbol{\mu}_{\boldsymbol{x}}$ is the mean vector of $\boldsymbol{x}$, $\boldsymbol{V}_{\boldsymbol{x}}$ and $\boldsymbol{V}_{\boldsymbol{y}}$ are the covariance matrices of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. $\boldsymbol{V}_{\boldsymbol{yx}}$ is the covariance matrix of $\boldsymbol{y}$ and $\boldsymbol{x}$. Recall the multivariate quadratic form:

$$\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y} = tr(\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y}) = tr(\boldsymbol{A}\boldsymbol{y}\boldsymbol{y}') \tag{3.2.25}$$

The first equality can be proven by noting that $\boldsymbol{y}'\boldsymbol{A}\boldsymbol{y}$ is a quadratic form, which is a scalar quantity. The second equality is due to the cyclic property of the trace operator. Now using the above facts, we can obtain the conditional distribution of $\boldsymbol{b} \mid \boldsymbol{y}$ as multivariate normal with mean equal to:

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{b}|\boldsymbol{y}} &= \boldsymbol{\mu}_{\boldsymbol{b}} + \boldsymbol{V}'_{by} \boldsymbol{V}_y^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{y}}) \\ &= 0 + \sigma_b^2 \boldsymbol{K}\boldsymbol{Z}' (\sigma_b^2)^{-1} (\sigma_b^2 \boldsymbol{Z}\boldsymbol{K}\boldsymbol{Z}' + \sigma_e^2 \boldsymbol{I})^{-1} (\boldsymbol{y} - \boldsymbol{X}^*\boldsymbol{\beta}^*) \\ &= \boldsymbol{K}\boldsymbol{Z}' \left(\boldsymbol{Z}\boldsymbol{K}\boldsymbol{Z}' + \frac{\sigma_e^2}{\sigma_b^2} \boldsymbol{I} \right)^{-1} (\boldsymbol{y} - \boldsymbol{X}^*\boldsymbol{\beta}^*) \end{aligned} \tag{3.2.26}$$

and the variance is obtained as:

$$\begin{aligned}
\boldsymbol{V}_{b|y} &= \boldsymbol{V}_b - \boldsymbol{V}'_{yb}\boldsymbol{V}_y^{-1}\boldsymbol{V}_{by} \\
&= (\sigma_b^2\boldsymbol{K}) - \sigma_b^2\boldsymbol{K}\boldsymbol{Z}'(\sigma_b^2\boldsymbol{Z}\boldsymbol{K}\boldsymbol{Z}' + \sigma_e^2\boldsymbol{I})^{-1}(\sigma_b^2)\boldsymbol{K}\boldsymbol{Z}' \\
&= (\sigma_b^2\boldsymbol{K}) - \sigma_b^2\boldsymbol{K}\boldsymbol{Z}'(\sigma_b^2)^{-1}\left(\boldsymbol{Z}\boldsymbol{K}\boldsymbol{Z}' + \frac{\sigma_e^2}{\sigma_b^2}\boldsymbol{I}\right)^{-1}(\sigma_b^2\boldsymbol{K}\boldsymbol{Z}') \\
&= (\sigma_b^2\boldsymbol{K}) - \sigma_b^2\boldsymbol{K}\boldsymbol{Z}'\boldsymbol{V}_{\boldsymbol{y}}^{-1}\boldsymbol{K}\boldsymbol{Z}'
\end{aligned} \tag{3.2.27}$$

We have now obtained the values of the conditional expectations necessary to obtain the expected values of the sufficient statistic, $\boldsymbol{b}'\boldsymbol{b}$. Hence:

$$\begin{aligned}
E\left(\boldsymbol{b}'\boldsymbol{b} \mid \boldsymbol{y}, \boldsymbol{\beta_g}, \boldsymbol{\beta_x}, \sigma_b^2, \sigma_e^2\right) &= E\left(\boldsymbol{b}|\boldsymbol{y}, \boldsymbol{\beta_g}, \boldsymbol{\beta_x}, \sigma_b^2, \sigma_e^2\right) E\left(\boldsymbol{b}'|\boldsymbol{y}, \boldsymbol{\beta_g}, \boldsymbol{\beta_x}, \sigma_b^2, \sigma_e^2\right) \\
&\quad + \mathrm{Var}\left(\boldsymbol{b}|\boldsymbol{y}, \boldsymbol{\beta_g}, \boldsymbol{\beta_x}, \sigma_b^2, \sigma_e^2\right) \\
&= \boldsymbol{K}\boldsymbol{Z}'\left(\boldsymbol{Z}\boldsymbol{K}\boldsymbol{Z}' + \frac{\sigma_e^2}{\sigma_b^2}\boldsymbol{I}\right)^{-1}(\boldsymbol{y} - \boldsymbol{X}^*\boldsymbol{\beta}^*) \\
&\quad + \left\{\boldsymbol{K}\boldsymbol{Z}'\left(\boldsymbol{Z}\boldsymbol{K}\boldsymbol{Z}' + \frac{\sigma_e^2}{\sigma_b^2}\boldsymbol{I}\right)^{-1}\right\} \\
&\quad \left(\boldsymbol{y} - \boldsymbol{X}^*\boldsymbol{\beta}^*\right)' + (\sigma_b^2\boldsymbol{K}) - \sigma_b^2\boldsymbol{K}\boldsymbol{Z}'\boldsymbol{V}_{\boldsymbol{y}}^{-1}\boldsymbol{K}\boldsymbol{Z}'.
\end{aligned} \tag{3.2.28}$$

To obtain $E\left(\boldsymbol{e}'\boldsymbol{e}|\boldsymbol{y}, \boldsymbol{\beta_g}, \boldsymbol{\beta_x}, \sigma_b^2, \sigma_e^2\right)$, we again use the formulas for the expectation of a quadratic form and the conditional variance as follows:

$$\begin{aligned}
E\left(\boldsymbol{e}'\boldsymbol{e}|\boldsymbol{y}, \boldsymbol{\beta_g}, \boldsymbol{\beta_x}, \sigma_b^2, \sigma_e^2\right) &= \mathrm{tr}E\left(\boldsymbol{e}\boldsymbol{e}'|\boldsymbol{y}, \boldsymbol{\beta_g}, \boldsymbol{\beta_x}, \sigma_b^2, \sigma_e^2\right) \\
&= tr\left[E\left(\boldsymbol{e}|\boldsymbol{y}, \boldsymbol{\beta_g}, \boldsymbol{\beta_x}, \sigma_b^2, \sigma_e^2\right) E\left(\boldsymbol{e}'|\boldsymbol{y}, \boldsymbol{\beta_g}, \boldsymbol{\beta_x}, \sigma_b^2, \sigma_e^2\right)\right. \\
&\quad \left.+\mathrm{Var}\left(\boldsymbol{e}|\boldsymbol{y}, \boldsymbol{\beta_g}, \boldsymbol{\beta_x}, \sigma_b^2, \sigma_e^2\right)\right].
\end{aligned} \tag{3.2.29}$$

Using similar reasoning as the computations above involving the conditional distri-

bution of the random effects $\boldsymbol{b}$, we would have $\boldsymbol{e}|\boldsymbol{y}$ as:

$$\boldsymbol{e}|\boldsymbol{y} \sim N(\boldsymbol{\mu}_{e|y}, \boldsymbol{V}_{e|y}).$$

where $\boldsymbol{\mu}_{e|y}$ and $\boldsymbol{V}_{e|y}$ are obtained using formulas for the conditional mean and variance given in equation (3.2.24). Using this, we obtain

$$\boldsymbol{\mu}_{e|y} = \sigma_e^2 \boldsymbol{I_n} \boldsymbol{V}_y^{-1}(\boldsymbol{y} - \boldsymbol{X}^*\boldsymbol{\beta}^*)$$

and

$$\boldsymbol{V}_{e|y} = \sigma_e^2 \boldsymbol{I_n} - \sigma_e^4 \boldsymbol{V}_y^{-1}.$$

Substituting these into equation (3.2.29), we obtain:

$$\begin{aligned} E\left(\boldsymbol{e}'\boldsymbol{e}|\boldsymbol{y}, \boldsymbol{\beta}_g, \boldsymbol{\beta}_x, \sigma_b^2, \sigma_e^2\right) &= tr\left[E\left(\boldsymbol{e}|\boldsymbol{y}, \boldsymbol{\beta_g}, \boldsymbol{\beta_x}, \sigma_b^2, \sigma_e^2\right) E\left(\boldsymbol{e}'|\boldsymbol{y}, \boldsymbol{\beta_g}, \boldsymbol{\beta}_x, \sigma_b^2, \sigma_e^2\right)\right. \\ &\quad \left. + \text{Var}\left(\boldsymbol{e}|\boldsymbol{y}, \boldsymbol{\beta}_g, \boldsymbol{\beta}_x, \sigma_b^2, \sigma_e^2\right)\right] \\ &= tr\left[\sigma_e^4 \boldsymbol{V}_y^{-1}(\boldsymbol{y} - \boldsymbol{X}^*\boldsymbol{\beta}^*)(\boldsymbol{y} - \boldsymbol{X}^*\boldsymbol{\beta}^*)'\boldsymbol{V}_y^{-1} + \sigma_e^2 \boldsymbol{I} - \sigma_e^4 \boldsymbol{V}_y^{-1}\right] \end{aligned} \tag{3.2.30}$$

In this section, we have discussed direct optimisation and an EM approach for obtaining the MLEs of the linear mixed model when the data is complete. In the next section, we derive an EM algorithm for the linear mixed model with missing genotypes that is based on the method of weights [23]. The method of weights was adapted for LMMs by Ibrahim et al. [24, 66], which is the approach our algorithm is based on.

3.3 Derivation of the EM algorithm when genotype data are missing

Recall the linear mixed model introduced in section 3.2:

$$\boldsymbol{y} = \boldsymbol{G}\boldsymbol{\beta}_g + \boldsymbol{X}\boldsymbol{\beta}_x + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}. \tag{3.3.1}$$

We now assume that we are considering a single candidate locus so that $l = 1$ and the genotypes at that locus are stored in a vector labelled $\boldsymbol{g}$. As we are considering a single biallelic locus, an individual's genotype can be denoted by 0, 1 or 2 depending on the number of minor alleles they have. $\boldsymbol{Z}_{n\times p}$ is a known design matrix and in our model it is the identity matrix. With these assumptions, the model (3.3.1) can now be written as:

$$\boldsymbol{y} = \boldsymbol{g}\beta_g + \boldsymbol{X}\boldsymbol{\beta}_x + \boldsymbol{I}_n\boldsymbol{b} + \boldsymbol{\epsilon} \tag{3.3.2}$$

where $\boldsymbol{g}$ is the $n\times 1$ vector of genetic fixed effects and the regression parameters for the fixed effects are denoted by β_g and $\boldsymbol{\beta}_x$ for the genetic and non-genetic fixed effects respectively. $\boldsymbol{b}_{n\times 1}$ is the vector of random effects assumed to be $\boldsymbol{b} \sim N_n(0, \sigma_b^2\boldsymbol{K})$ where $\boldsymbol{K}$ is the known genetic relationship matrix (GRM). In our context, the design matrix for the random effects is the identity matrix and so we substitute $\boldsymbol{Z}$ with $\boldsymbol{I}_n$. The model errors $\boldsymbol{\epsilon} \sim N_n(0, \sigma_e^2\boldsymbol{I}_n)$. We assume that all $\boldsymbol{b}$ and $\boldsymbol{\epsilon}$ are mutually independent.

We are primarily interested in estimation and testing of the genetic effect (β_g) in the presence of other covariates and population substructure; the variance components are regarded as nuisance parameters. We assume that the genetic data $\boldsymbol{g}$ are not available on all study participants; for example genotypes in an EPS design are only

available on the top and bottom of the phenotype distribution. We can consider these observations missing at random since the missing data mechanism depends only on the observed data $\boldsymbol{y}$. To fit the model, we therefore need to handle the missing covariate data. We will do this using an EM algorithm called the method of weights. To simplify our approach, we will assume that the only missing data are the genetic covariates. We will also assume that the genetic covariate is uncorrelated with the non-genetic covariate and the random effect. We will also need to assume a model for the missing genotypes. Since we are considering only a single locus at a time, the genotypes can be modeled with a multinomial distribution with parameter $\boldsymbol{\gamma} = (\boldsymbol{\gamma_0}, \boldsymbol{\gamma_1}, \boldsymbol{\gamma_2})$ for the probabilities corresponding to each genotype.

Since $\boldsymbol{X}$ is fully observed and considered fixed and known, the density function is then given as:

$$f_{\boldsymbol{Y},B,G}(\boldsymbol{y}, \boldsymbol{b}, \boldsymbol{g}; \boldsymbol{X}, \boldsymbol{\theta}) = f_{Y,B|g}(\boldsymbol{y}, \boldsymbol{b}|\boldsymbol{g}, \boldsymbol{X}; \boldsymbol{\theta}) f_G(\boldsymbol{g}; \boldsymbol{\gamma}). \tag{3.3.3}$$

It is often helpful for parameter estimation purposes to write the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{b}$ as a product of the conditional distribution of $\boldsymbol{y}|\boldsymbol{b}$, i.e conditioning on the random effects and marginal distribution of $\boldsymbol{b}$. We can rewrite equation (3.3.3) with the subscripts suppressed as:

$$f(\boldsymbol{y}, \boldsymbol{g}, \boldsymbol{b}|\boldsymbol{X}; \theta) = f\left(\boldsymbol{y} \mid \boldsymbol{b}, \boldsymbol{g}, \boldsymbol{X}; \beta_g, \boldsymbol{\beta_x}, \sigma_e^2\right) f\left(\boldsymbol{b} \mid \sigma_b^2\right) f(\boldsymbol{g} \mid \boldsymbol{\gamma}) \tag{3.3.4}$$

The complete data log-likelihood for the LMM is:

$$\ell_c(\boldsymbol{\theta}; \boldsymbol{Y}, \boldsymbol{G}, \boldsymbol{X}) = \log f(\boldsymbol{y}|\boldsymbol{g}, \boldsymbol{b}; \boldsymbol{\beta_x}, \beta_g, \sigma_e^2) + \log f(\boldsymbol{b}|\sigma_b^2) + \log f(\boldsymbol{g}|\boldsymbol{\gamma}) \tag{3.3.5}$$

If all genotypes were observed, the log-likelihood would simply be the first two terms of equation (3.3.5); maximization of this log-likelihood has already been discussed in Section 3.2.1. Here, we also have to take into account the missing data and we will use the method of weights, which is an EM algorithm for handling missing categorical covariates. In the next section, we present the method of weights.

3.3.1 Method of weights for LMMs

The EM algorithm used for the generalised linear models with missing categorical covariate has been named the method of weights [23]. The E-step is a problem of computing a weight of one or less for each possible value of the categorical covariate. In our context, an individual with unknown genotype can have one of three genotypes. The weight is the probability of the genotype conditional on the observed data. When the underlying model is a LMM or GLMM, then the original method of weights [23] would not be suitable, as we have to deal with the random effects $\boldsymbol{b}$. Ibrahim et al. [24] and Stubbendick and Ibrahim [66] proposed an extension to the method of weights that handles both mixed models and data that are non-ignorably missing. However, since our data is missing at random (MAR) we do not need to model the missing data mechanism.

Let $g_i = 0, 1$ or 2 be individual i's genotype, which may be observed or missing. If the genotype is observed, we may also label the genotype as $g_{obs,i}$; if the genotype is not observed, we label it as $g_{mis,i}$. The set of observed genotypes on the full dataset is denoted $\boldsymbol{g}_{obs}$ and the set of missing genotypes is denoted $\boldsymbol{g}_{mis}$. The complete data genotype vector is $\boldsymbol{g} = (\boldsymbol{g}_{obs}, \boldsymbol{g}_{mis})$.

Let $\boldsymbol{\theta}^{(t)} = \left(\boldsymbol{\beta}_{\boldsymbol{g}}^{(t)}, \boldsymbol{\beta}_x^{(t)}, (\sigma_b^2)^{(t)}, (\sigma_e^2)^{(t)}, \boldsymbol{\gamma}^{(t)}\right)$ be the parameter estimates at the tth iteration. The additional parameter $\boldsymbol{\gamma}$ parameterizes the genetic covariate distribu-

tion. Normally, when maximizing the log-likelihood for a linear mixed model, this term would not need to be included as we can consider the genetic covariates fixed and known.

The Expectation step

The Expectation step involves computing $Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\right)$, the conditional expected log-likelihood of the complete-data $\ell_c(\boldsymbol{\theta}; \boldsymbol{g}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{b})$ given the observed data $(\boldsymbol{g}_{\text{obs}}, \boldsymbol{X}, \boldsymbol{y})$ and the current parameter estimates $\boldsymbol{\theta}^{(t)}$:

$$Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\right) = E\left[\ell_c(\boldsymbol{\theta}) \mid \boldsymbol{g}_{obs}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}^{(t)}\right] \tag{3.3.6}$$

Recall the complete data log likelihood $\ell_c(\boldsymbol{\theta} \mid \boldsymbol{X}, \boldsymbol{g}, \boldsymbol{b}, \boldsymbol{y})$ is expressed in equation (3.3.5) with:

$$\boldsymbol{b} \sim N_n(\mathbf{0}, \boldsymbol{V}_b) \text{ where } \boldsymbol{V}_b = \sigma_b^2 \boldsymbol{K},$$

$$f(\boldsymbol{g}; \boldsymbol{\gamma}) = \prod_{i=1}^{n} \Pr(g_i|\gamma)$$

where $g_i \sim Multinomial(2, \gamma)$ and

$$\boldsymbol{y}|\boldsymbol{g}, \boldsymbol{b} \sim N_n(\boldsymbol{g}\beta_g + \boldsymbol{X}\boldsymbol{\beta}_{\boldsymbol{x}} + \boldsymbol{b}, \sigma_e^2 \boldsymbol{I}).$$

As the covariance matrix is diagonal, the y_i are all conditionally independent therefore

$$f(\boldsymbol{y}|\boldsymbol{g}, \boldsymbol{b}) = \prod_{i=1}^{n} f(y_i|g_i, b_i)$$

where

$$y_i|g_i, b_i \sim N(\beta_g g_i + X_i\beta_x + b_i, \sigma_e^2).$$

Hence in this case, the complete data log-likelihood can be further simplifed to:

$$\ell_c(\boldsymbol{\theta} \mid \mathbf{x}, \mathbf{g}, \mathbf{b}, \mathbf{y}) = \sum_{i=1}^{n} \log f(y_i|g_i, b_i) + \log f(\boldsymbol{b}) + \sum_{i=1}^{n} \log \Pr(g_i|\gamma)$$

However, for simplicity in computations, we will often express the distributions for $\boldsymbol{y}$ and $\boldsymbol{g}$ in their multivariate forms.

The approach of Ibrahim et al. [24] deals with computational problems encountered with the presence of random effects $\boldsymbol{b}$ in the linear mixed model. The E-step of this model deals with these problems by integrating out the random effect $\boldsymbol{b}$. This procedure makes the E-step more efficient for the normal random effects model and greatly reduces the amount of iterations needed for convergence. The E-step is obtained as:

$$\begin{aligned} Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\right) &= \sum_{\boldsymbol{g}_{mis} \in \mathcal{G}} \int_b \Big(\log f(\boldsymbol{y}|\boldsymbol{g}, \beta_g, \beta_x, \boldsymbol{X}, \sigma_e^2, \boldsymbol{b}) + \log f(\boldsymbol{b}|\boldsymbol{D}) + \log p(\boldsymbol{g}|\boldsymbol{\gamma}) \Big) \\ &\quad \times p(\boldsymbol{g_{mis}}, \boldsymbol{b} \mid \boldsymbol{g_{obs}}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}^{(t)}) d\boldsymbol{b} \\ &\equiv \sum_{\boldsymbol{g} \in \boldsymbol{G}} (I_1 + I_2 + I_3) \end{aligned} \tag{3.3.7}$$

where $\boldsymbol{G}$ is the set of possible genotypes for those with missing genotype data, given the observed genotype data $\boldsymbol{g}_{obs}$, $p(\boldsymbol{g_{mis}}, \boldsymbol{b} \mid \boldsymbol{g_{obs}}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}^{(t)})$ is the distribution of the missing genotypes and random effects conditional on the observed data, and

$$\begin{aligned} I_1 &= \int_b \Big(p(\boldsymbol{g_{mis}}, \boldsymbol{b} \mid \boldsymbol{g_{obs}}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}^{(t)}) \log f(\boldsymbol{y}|\boldsymbol{g}, \beta_g, \beta_x, \boldsymbol{X}, \sigma_e^2, \boldsymbol{b}) \Big) d\boldsymbol{b} \\ I_2 &= \int_b \Big(p(\boldsymbol{g_{mis}}, \boldsymbol{b} \mid \boldsymbol{g_{obs}}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}^{(t)}) \log f(\boldsymbol{b}|\boldsymbol{D}) \Big) d\boldsymbol{b} \\ I_3 &= \int_b \Big(p(\boldsymbol{g_{mis}}, \boldsymbol{b} \mid \boldsymbol{g_{obs}}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}^{(t)}) \log p(\boldsymbol{g}|\boldsymbol{\gamma}) \Big) d\boldsymbol{b} \end{aligned} \tag{3.3.8}$$

To estimate the parameters of interest in this model, we must maximize each of I_1, I_2 and I_3. Following Ibrahim et al. [24], we first integrate out the random effects $\boldsymbol{b}$ from the conditional density using the expressions previously obtained for the conditional mean and variance for the complete case LMM in section 3.2.1.

Consider I_1:

$$I_1 = \int_{\mathbb{R}^n} \log f(\boldsymbol{y}|\boldsymbol{g}, \beta_g, \boldsymbol{\beta}_x, \boldsymbol{X}, \sigma_e^2, \boldsymbol{b}) \times p\left(\mathbf{g}_{mis}, \boldsymbol{b} \mid \boldsymbol{g}_{obs}, \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}^{(t)}\right) d\boldsymbol{b} \tag{3.3.9}$$

$$= \left[\int_{\mathbb{R}^n} \log f(\boldsymbol{y}|\boldsymbol{g}, \beta_g, \boldsymbol{\beta}_x, \boldsymbol{X}, \sigma_e^2, \boldsymbol{b}) p(\boldsymbol{b} \mid \mathbf{y}, \boldsymbol{X}, \boldsymbol{g}, \boldsymbol{\theta}^{(t)}) d\boldsymbol{b}\right] p(\boldsymbol{g}_{mis}|\boldsymbol{g}_{obs}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}^{(t)}) \tag{3.3.10}$$

Let I_1^* denote the integral in I_1 :

$$I_1^* = \int_{\mathbb{R}^n} \log p(\boldsymbol{y}|\boldsymbol{g}, \beta_g, \boldsymbol{\beta_x}, \boldsymbol{X}, \sigma_e^2, \boldsymbol{b}) \times p(\boldsymbol{b} \mid \mathbf{y}, \boldsymbol{X}, \boldsymbol{g}, \boldsymbol{\theta}^{(t)})\, d\boldsymbol{b} \tag{3.3.11}$$

$$= \int_{\mathbb{R}^n} \left\{ -\frac{1}{2} \log |\sigma_e^2 I_n| - \frac{1}{2\sigma_e^2} (\boldsymbol{y} - \boldsymbol{g}\beta_g - \boldsymbol{X}\boldsymbol{\beta_x} - \boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{g}\beta_g - \boldsymbol{X}\boldsymbol{\beta_x} - \boldsymbol{b}) \right\}$$

$$p(\boldsymbol{b} \mid \mathbf{y}, \boldsymbol{X}, \boldsymbol{g}, \boldsymbol{\theta}^{(t)}) d\boldsymbol{b}$$

$$= -\frac{1}{2} \log |\sigma_e^2 I_n| - \int_{\mathbb{R}^n} \left\{ \frac{1}{2\sigma_e^2} (\boldsymbol{y} - \boldsymbol{g}\beta_g - \boldsymbol{X}\boldsymbol{\beta_x} - \boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{g}\beta_g - \boldsymbol{X}\boldsymbol{\beta_x} - \boldsymbol{b}) \right\}$$

$$p(\boldsymbol{b} \mid \mathbf{y}, \boldsymbol{X}, \boldsymbol{g}, \boldsymbol{\theta}^{(t)}) d\boldsymbol{b}$$

The first term in the last equality is obtained by noting that it does not depend on $\boldsymbol{b}$. Thus $\frac{1}{2} \log |\sigma_e^2 I_n|$ can be factored out of the integral and the integral that remains evaluates to 1.

To integrate out the random effects from the second term, we expand the quadratic

form and evaluate any term involving $\boldsymbol{b}$. Letting $\boldsymbol{b}^{(t)} = E_{\boldsymbol{\theta}^{(t)}}[\boldsymbol{b}|\boldsymbol{y}]$, we get

$$\begin{aligned} E[\boldsymbol{y}'\boldsymbol{b}|\boldsymbol{y}] &= \boldsymbol{y}'\boldsymbol{b}^{(t)} \\ E[(\boldsymbol{g}\beta_g)'\boldsymbol{b}|\boldsymbol{y}] &= (\boldsymbol{g}\beta_g)'\boldsymbol{b}^{(t)} \\ E[(\boldsymbol{X}\beta_X)'\boldsymbol{b}|\boldsymbol{y}] &= (\boldsymbol{X}\beta_X)'\boldsymbol{b}^{(t)} \\ E[\boldsymbol{b}'\boldsymbol{y}|\boldsymbol{y}] &= (\boldsymbol{b}^{(t)})'\boldsymbol{y} \\ E[\boldsymbol{b}'(\boldsymbol{g}\beta_g)|\boldsymbol{y}] &= (\boldsymbol{b}^{(t)})'\boldsymbol{g}\beta_g \\ E[\boldsymbol{b}'(\boldsymbol{X}\beta_X)|\boldsymbol{y}] &= (\boldsymbol{b}^{(t)})'(\boldsymbol{X}\beta_X) \end{aligned} \tag{3.3.12}$$

To compute $E[\boldsymbol{b}'\boldsymbol{b}|\boldsymbol{y}]$, we first note the following matrix algebra property: for $\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}$ where $\boldsymbol{A}$ is a symmetric $n \times n$ matrix, the expectation of the quantity $\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}$ can be given as:

$$E(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}) = \mu'\boldsymbol{A}\mu + tr(\boldsymbol{A}\Sigma)$$

where μ and Σ are the mean and variance of $\boldsymbol{X}$ respectively. Applying this result to $E[\boldsymbol{b}'\boldsymbol{b}|\boldsymbol{y}]$ yields

$$E[\boldsymbol{b}'\boldsymbol{b}|\boldsymbol{y}] = (\boldsymbol{b}^{(t)})'\boldsymbol{b}^{(t)} + tr(\boldsymbol{V}^{(t)}) \tag{3.3.13}$$

where $\boldsymbol{V}^{(t)} = \text{var}_{\boldsymbol{\theta}^{(t)}}[\boldsymbol{b}|\boldsymbol{y}]$. Expressions for $\boldsymbol{b}^{(t)}$ and $\boldsymbol{V}^{(t)}$ were derived in [24] and in our model are given as:

$$\boldsymbol{b}^{(t)} = \left((\sigma_e^2)^{(t)}\right)^{-1}\boldsymbol{V}^{(t)}(\boldsymbol{y} - \boldsymbol{g}\beta_g^{(t)} - \boldsymbol{X}\boldsymbol{\beta}_{\boldsymbol{X}}^{(t)}) \tag{3.3.14}$$

$$\boldsymbol{V}^{(t)} = \left((\sigma_e^2)^{(t)} + (\boldsymbol{D}^{(t)})^{-1}\right)^{-1}. \tag{3.3.15}$$

By substituting equations (3.3.12) to (3.3.13) into I_1 and re-writing as a quadratic

form, we get:

$$I_1 = -\frac{1}{2}\log(\sigma_e^2) - \frac{1}{2\sigma_e^2}tr(\boldsymbol{V}^{(t)}) - \frac{1}{2\sigma_e^2}(\boldsymbol{y} - \boldsymbol{g}\beta_g - \boldsymbol{X}\boldsymbol{\beta_x} - \boldsymbol{b}^{(t)})' \tag{3.3.16}$$
$$(\boldsymbol{y} - \boldsymbol{g}\beta_g - \boldsymbol{X}\boldsymbol{\beta_x} - \boldsymbol{b}^{(t)})p(\boldsymbol{g}_{mis}|\boldsymbol{g}_{obs}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}^{(t)})$$

We observe that in the quadratic form, the random effects have been replaced with their conditional expectation and therefore they are no longer random variables.

We now consider the sum of I_1 over the missing genotypes in equation 3.3.7. Since the first two terms are not functions of the genotypes they can be factored out of the sum. The term remaining is the sum of a probability distribution over its support and so it sums to 1. We therefore get

$$\sum_{\boldsymbol{g}_{mis}\in\boldsymbol{G}} I_1 = -\frac{1}{2}\log(\sigma_e^2) - \frac{1}{2\sigma_e^2}tr(\boldsymbol{V}^{(t)}) - \frac{1}{2\sigma_e^2}\sum_{\boldsymbol{g}_{mis}\in\boldsymbol{G}}(\boldsymbol{y} - \boldsymbol{g}\beta_g - \boldsymbol{X}\boldsymbol{\beta_x} - \boldsymbol{b}^{(t)})' \tag{3.3.17}$$
$$(\boldsymbol{y} - \boldsymbol{g}\beta_g - \boldsymbol{X}\boldsymbol{\beta_x} - \boldsymbol{b}^{(t)})p(\boldsymbol{g}_{mis}|\boldsymbol{g}_{obs}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}^{(t)})$$
$$= -\frac{1}{2}\log(\sigma_e^2) - \frac{1}{2\sigma_e^2}tr(\boldsymbol{V}^{(t)}) - \frac{1}{2\sigma_e^2}\sum_{\boldsymbol{g}_{mis}\in\boldsymbol{G}}\Big(\sum_{i=1}^{n}(y_i - g_i\beta_g - \boldsymbol{x}_i\boldsymbol{\beta}_x - b_i^{(t)})^2\Big)$$
$$p(\boldsymbol{g}_{mis}|\boldsymbol{g}_{obs}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}^{(t)})$$

The distribution of the missing genotype vector conditional on the observed data can be written as a product of the probabilities for each individual genotype. However, because each $b_i^{(t)}$ depends on the entire genotype vector $\boldsymbol{g}$, we need to compute $\boldsymbol{b}$ for each $\boldsymbol{g}_{mis} \in \boldsymbol{G}$. For this reason, computing the sum over the genotype vector requires enumeration of all genotype vectors compatible with the observed data. As this sum is too large to compute, we follow Ibrahim et al. [24] and use a Monte Carlo

approximation. We first sample missing genotypes from the conditional distribution

$$\Pr\left(\boldsymbol{g}_{mis} \mid \boldsymbol{g}_{obs}, \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}^{(t)}\right).$$

The genotypes are independent, so we can achieve this by sampling the missing genotype for individual i from $\Pr(g_i|\boldsymbol{x}_i, y_i, \theta^{(t)})$; these probabilities are called the weights and the procedure for sampling from them is described in Section 3.3.1. By doing this for all the individuals with missing genotypes, we obtain a draw of size M from the conditional genotype distribution. The sum over $\boldsymbol{g}_{mis} \in \boldsymbol{G}$ is then approximated by:

$$\frac{1}{M}\sum_{l=1}^{M}(\boldsymbol{y}-\boldsymbol{g_l}\beta_g-\boldsymbol{X}\boldsymbol{\beta_x}-\boldsymbol{b}_l^{(t)})'(\boldsymbol{y}-\boldsymbol{g_l}\beta_g-\boldsymbol{X}\boldsymbol{\beta_x}-\boldsymbol{b}_l^{(t)}) = \frac{1}{M}\sum_{l=1}^{M}\sum_{i=1}^{n}(y_i-g_{i,l}\beta_g-\boldsymbol{x}_i\boldsymbol{\beta}_x-b_{i,l}^{(t)})^2$$

where $\boldsymbol{g}_l = (\boldsymbol{g}_{obs}, \boldsymbol{g}_{mis,l})$ is the lth sampled genotype vector, $\boldsymbol{b}_l^{(t)}$ is the conditional expectation of the random effects for the lth sampled genotype vector and $b_{i,l}^{(t)}$ is the ith element of $\boldsymbol{b}_l^{(t)}$. The expression for I_1 now becomes:

$$\begin{aligned} I_1 = &-\frac{1}{2}\log(\sigma_e^2) - \frac{1}{2\sigma_e^2}tr(\boldsymbol{V}^{(t)})- \\ &\frac{1}{2\sigma_e^2}\frac{1}{M}\sum_{l=1}^{M}(\boldsymbol{y}-\boldsymbol{g}_l\beta_g-\boldsymbol{X}\boldsymbol{\beta_x}-\boldsymbol{b}_l^{(t)})'(\boldsymbol{y}-\boldsymbol{g}_l\beta_g-\boldsymbol{X}\boldsymbol{\beta_x}-\boldsymbol{b}_l^{(t)}) \\ = &-\frac{1}{2}\log(\sigma_e^2) - \frac{1}{2\sigma_e^2}tr(\boldsymbol{V}^{(t)}) - \frac{1}{2\sigma_e^2}\frac{1}{M}\sum_{l=1}^{M}\sum_{i=1}^{n}(y_i-g_{i,l}\beta_g-\boldsymbol{x}_i\beta_x-b_{i,l}^{(t)})^2 \end{aligned} \tag{3.3.18}$$

For I_2, we have

$$\begin{aligned} I_2 &= \int_{\mathbb{R}^n} \log p(\boldsymbol{b}|\boldsymbol{D}) p(\boldsymbol{g_{mis}}, \boldsymbol{b} \mid \boldsymbol{g_{obs}}, \boldsymbol{X}, \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{\theta}^{(t)}) d\boldsymbol{b} \\ &= \left[\int_{\mathbb{R}^n} \log p(\boldsymbol{b}|\boldsymbol{D}) p(\boldsymbol{b} \mid \boldsymbol{g}, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\theta}^{(t)}) d\boldsymbol{b} \right] p(\boldsymbol{g_{mis}} \mid \boldsymbol{g_{obs}}, \boldsymbol{X}, \boldsymbol{z}, \boldsymbol{y}, \boldsymbol{\theta}^{(t)}) \end{aligned} \tag{3.3.19}$$

To integrate out the random effect $\boldsymbol{b}$, consider the integral in equation 3.3.19 denoted as I_2^*:

$$\begin{aligned} I_2^* &= \int_{\mathbb{R}^n} \log p(\boldsymbol{b}|\boldsymbol{D}) \times p(\boldsymbol{b}|\boldsymbol{y}, \boldsymbol{g}, \boldsymbol{X}, \boldsymbol{\theta}^{(t)}) \, d\boldsymbol{b} \\ &= -\frac{1}{2} \log |\boldsymbol{D}| - \frac{1}{2} \int_{\mathbb{R}^n} \boldsymbol{b}' \boldsymbol{D}^{-1} \boldsymbol{b} \times p(\boldsymbol{b}|\boldsymbol{y}, \boldsymbol{g}, \boldsymbol{X}, \boldsymbol{\theta}^{(t)}) d\boldsymbol{b} \end{aligned} \tag{3.3.20}$$

We again use the trace formula for the expectation of the quadratic form for a random vector: for $\boldsymbol{x} \sim N(\mu, \boldsymbol{\Sigma})$ and a symmetric positive definite matrix $\boldsymbol{A}$, the expectation of the quadratic form

$$\mathrm{E}\left[\boldsymbol{x}' \boldsymbol{A} \boldsymbol{x}\right] = \mu' \boldsymbol{A} \mu + \mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}). \tag{3.3.21}$$

Hence

$$E\left[\boldsymbol{b}' \boldsymbol{D}^{-1} \boldsymbol{b}\right] = (\boldsymbol{b}^{(t)})' \boldsymbol{D}^{-1} \boldsymbol{b}^{(t)} + tr(\boldsymbol{D}^{-1} \boldsymbol{V}^{(t)}) \tag{3.3.22}$$

where $\boldsymbol{b}^{(t)}$ and $\boldsymbol{V}^{(t)}$ are the conditional expectations of $p(\boldsymbol{b}|\boldsymbol{y}, \boldsymbol{g}, \boldsymbol{X}, \boldsymbol{\theta}^{(t)})$ with formulas given in equations 3.3.14 and 3.3.15. Therefore

$$I_2^* = -\frac{1}{2} \log |\mathbf{D}| - \frac{1}{2} tr\left(\boldsymbol{D}^{-1} \boldsymbol{V}^{(t)}\right) - \frac{1}{2} \left(\mathbf{b}^{(t)\prime} \mathbf{D}^{-1} (\mathbf{b})^{(t)}\right) \tag{3.3.23}$$

and the expression for I_2 now becomes:

$$I_2 = -\frac{1}{2}\log|\mathbf{D}| - \frac{1}{2}tr\left(\boldsymbol{D}^{-1}\boldsymbol{V}^{(t)}\right) - \frac{1}{2}\sum_{\boldsymbol{g}_{mis}\in\boldsymbol{G}}\left(\boldsymbol{b}^{(t)'}\mathbf{D}^{-1}(\mathbf{b})^{(t)}\right)\times Pr\left(\mathbf{g_{mis}} \mid \mathbf{g_{obs}}, \mathbf{y}, \mathbf{X}, \boldsymbol{\theta}^{(t)}\right) \tag{3.3.24}$$

where we note that $\boldsymbol{b}^{(t)}$ is a function of the genotype vector $\boldsymbol{g}$.

As with the calculation for I_1, the number of terms to sum over is too high. We instead use the Monte Carlo approximation by drawing M genotype vectors with individual i's genotype sampled from $\Pr(g_i|\boldsymbol{x_i}, y_i, \theta^{(t)})$. We approximate I_2 with

$$I_2 = -\frac{1}{2}\log|\mathbf{D}| - \frac{1}{2}tr\left(\boldsymbol{D}^{-1}\boldsymbol{V}^{(t)}\right) - \frac{1}{2M}\sum_{l=1}^{M}\left(\boldsymbol{b}_l^{(t)'}\boldsymbol{D}^{-1}\boldsymbol{b}_l^{(t)}\right) \tag{3.3.25}$$

where $\boldsymbol{b}_l^{(t)}$ is defined as previously.

For I_3, $\log(p(\boldsymbol{g}|\gamma))$ does not depend on $\boldsymbol{b}$ and the genotypes are considered independent. Therefore, I_3 can be written as

$$\begin{aligned} I_3 &= \sum_{\boldsymbol{g}\in\mathcal{G}}\log p(\boldsymbol{g}|\gamma)\times p(\boldsymbol{g}_{mis} \mid \boldsymbol{g}_{obs}, \mathbf{y}, \mathbf{X}, \boldsymbol{\theta}^{(t)}) \\ &= \sum_{i=1}^{n}\sum_{j=0}^{2}\log p(g_i = j|\gamma)\times p(g_i = j \mid y_i, \boldsymbol{x_i}, \boldsymbol{\theta}^{(t)}) \end{aligned} \tag{3.3.26}$$

We note that the final versions of I_1, I_2 and I_3 do not involve random variables $\boldsymbol{b}$. The full objective function from the E-step is:

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = &-\frac{1}{2}\log(\sigma_e^2) - \frac{1}{2\sigma_e^2}tr(\boldsymbol{V}^{(t)}) \\ &- \frac{1}{2M\sigma_e^2}\sum_{l=1}^{M}(\boldsymbol{y} - \boldsymbol{g}_l\beta_g - \boldsymbol{X}\boldsymbol{\beta}_{\boldsymbol{x}} - \boldsymbol{b}_l^{(t)})'(\boldsymbol{y} - \boldsymbol{g}_l\beta_g - \boldsymbol{X}\boldsymbol{\beta}_{\boldsymbol{x}} - \boldsymbol{b}_l^{(t)}) \end{aligned} \tag{3.3.27}$$

$$-\frac{1}{2}\log|\mathbf{D}| - \frac{1}{2}tr\left(\boldsymbol{D}^{-1}\boldsymbol{V}^{(t)}\right) - \frac{1}{2M}\sum_{l=1}^{M}\left(\boldsymbol{b}_l^{(t)\prime}\boldsymbol{D}^{-1}\boldsymbol{b}_l^{(t)}\right)$$
$$+\sum_{i=1}^{n}\sum_{j=0}^{2}\log p(g_i=j|\gamma) \quad \Pr(\boldsymbol{g}_{mis} \mid \boldsymbol{g}_{obs}, \mathbf{y}, \mathbf{X}, \boldsymbol{\theta}^{(t)})$$

In the M step, we maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ to obtain the new parameter estimates. However, we first describe how to model the distribution of the missing genotypes conditional on the observed data. These probabilities are called the "weights".

Computing weights for sampling genotypes

To compute I_1 to I_3, we require the distribution for missing genotypes conditional on the observed data: $\Pr(\boldsymbol{g}_{mis} \mid \boldsymbol{g}_{obs}, \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}^{(t)})$. Assuming that genotypes are conditionally independent of each other and the non-genetic covariates $\boldsymbol{X}$

$$\Pr(\boldsymbol{g}_{mis} \mid \boldsymbol{g}_{obs}, \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}^{(t)}) = \prod_{i=1}^{n}\Pr(g_i = j \mid \mathbf{x}_i, \mathbf{y}_i, \boldsymbol{\theta}^{(t)}) \tag{3.3.28}$$

The term $\Pr(g_i = j \mid \mathbf{x}_i, \mathbf{y}_i, \boldsymbol{\theta}^{(t)})$ is called the weight $w_{ij}^{(t)}$. Note that if an individuals genotype is observed, the weight is 1 for their actual genotype and 0 for the other two genotypes. In the last line, the term $\Pr(g_i = j \mid \mathbf{x}_i, \boldsymbol{\theta}^{(t)})$ simplifies to $\Pr(g_i = j \mid \boldsymbol{\gamma}^{(t)})$ since genotypes are independent of non-genetic covariates and depend only on the genotype frequency parameters $\boldsymbol{\gamma}^{(t)}$. In a saturated genetic model where an individual's genotype has a multinomial distribution, the probability of the covariate vector taking on value j at the t^{th} iteration is $\boldsymbol{\gamma}_j^{(t)}$, the probability of that genotype. These parameters are estimated as part of the EM algorithm; however, an alternative option would be to fix these parameters at values estimated from population-based data.

The term $P\left(y_{\mathrm{i}} \mid g_i = j, \mathbf{x}_i, \boldsymbol{\theta}^{(\mathrm{t})}\right)$ is the density of the marginal normal distribution for y_i evaluated when $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)} = (\boldsymbol{\beta}_g, \boldsymbol{\beta}_x, \sigma_b^2, \sigma_e^2)^{(\mathrm{t})}$ where $\boldsymbol{\beta}_g, \boldsymbol{\beta}_x, \sigma_b^2, \sigma_e^2$ are all parameters of the linear model.

To summarize, the E-step involves computing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, the conditional expected value of the complete-data log-likelihood given the observed data. At the $t+1$th iteration, computing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ first requires determining the weights $w_{ij}^{(t)}$ for $i = 1, \ldots, n$ and $j = 0, 1, 2$. To estimate the sums in equation (3.3.27) using Monte Carlo, a genotype with level 0, 1 or 2 is sampled for each individual with missing genotype based on the weights of each genotype for that individual. This is repeated M times to get M genotype vectors. The weights are also required to calculate the last term of equation (3.3.27).

The Maximization step

The conditional expected values of the complete-data log-likelihood given the observed data and parameter estimates are maximized in the M step to find the new parameter estimates. The expected complete-data log-likelihood $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is the sum of three portions, a term that depends on the regression model parameters, a term that depends on the random effects model parameters and the term that depends on the genotype parameter. Since $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ breaks into the three portions, the maximization to update $\boldsymbol{\beta}$, the variance components (σ_b^2, σ_e^2) and $\boldsymbol{\gamma}$ estimates can all be derived separately.

Estimation of $\boldsymbol{\gamma}$ is the simplest as it is similar to the maximization of a multinomial likelihood. If no assumptions are made about the covariate distribution for the

genetic model, the objective function to maximize is:

$$\sum_{i=1}^{n}\sum_{j=0}^{2} w_{ij}^{(t)} \log P(g_i = j; \boldsymbol{\gamma}) = \sum_{i=1}^{n}\sum_{j=0}^{2} w_{ij}^{(t)} \log \gamma_j$$

We take derivatives with respect to γ_0 and γ_1; $\gamma_2 = 1 - \gamma_0 - \gamma_1$. Since the genotype weights for an individual sum to 1, we have $w_{i2}^{(t)} = 1 - w_{i0}^{(t)} - w_{i1}^{(t)}$. This gives:

$$\begin{aligned}
\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \gamma_k} &= \sum_{i=1}^{n}\left(\frac{w_{ik}^{(t)}}{\gamma_k} - \frac{w_{i2}^{(t)}}{1-\gamma_0-\gamma_1}\right) \quad k = 0, 1 \\
&= \frac{\sum_{i=1}^{n} w_{ik}^{(t)}}{\gamma_k} - \frac{n - \sum_{i=1}^{n} w_{i0}^{(t)} - \sum_{i=1}^{n} w_{i1}^{(t)}}{1-\gamma_0-\gamma_1} \quad k = 0, 1
\end{aligned}$$

Since $\sum_{i=1}^{n}\sum_{j=0}^{2} w_{ij}^{(t)} = n$, the maximization is similar to that of a multinomial distribution, which has solution

$$\hat{\gamma}_k = \frac{W_k^{(t)}}{n}$$

where $W_k^{(t)} = \sum_{i=1}^{n} w_{ik}^{(t)}$.

We now consider the regression parameter $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_x, \beta_g)$. The portion of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ that depends on $\boldsymbol{\beta}$ is:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = -\frac{1}{2M\sigma_e^2}\sum_{l=1}^{M}(\boldsymbol{y} - \boldsymbol{g}_l\beta_g - \boldsymbol{X}\boldsymbol{\beta}_{\boldsymbol{x}} - \boldsymbol{b}_l^{(t)})'(\boldsymbol{y} - \boldsymbol{g}_l\beta_g - \boldsymbol{X}\boldsymbol{\beta}_{\boldsymbol{x}} - \boldsymbol{b}_l^{(t)}) \quad (3.3.29)$$

To simplify notation, we will let $\boldsymbol{X}_l^* = (\boldsymbol{X}, \boldsymbol{g}_l)$ be the combined design matrix. Hence

in the first step, we will obtain the likelihood equation for $\boldsymbol{\beta}^*$ as follows:

$$\begin{aligned} 0 = \frac{\partial Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}\right)}{\partial \boldsymbol{\beta^*}} &= -\frac{1}{2M\sigma_e^2}\sum_{l=1}^{M}\frac{\partial}{\partial \boldsymbol{\beta^*}}\left\{(\boldsymbol{y}-\boldsymbol{X}_l^*\boldsymbol{\beta}^*-\boldsymbol{b}_l^{(t)})'(\boldsymbol{y}-\boldsymbol{X}_l^*\boldsymbol{\beta}^*-\boldsymbol{b}_l^{(t)})\right\} \\ &= -\frac{1}{2M\sigma_e^2}\sum_{l=1}^{M}-2(\boldsymbol{y}-\boldsymbol{X}_l^*\boldsymbol{\beta}^*-\boldsymbol{b}_l^{(t)})'\boldsymbol{X}_l^* \\ &= \frac{1}{M\sigma_e^2}\sum_{l=1}^{M}\left((\boldsymbol{y}-\boldsymbol{b}_l^{(t)})'\boldsymbol{X}_l^*-\boldsymbol{\beta^{*\prime}}\boldsymbol{X}_l^{*'}\boldsymbol{X}_l^*\right) \end{aligned} \tag{3.3.30}$$

We solve the likelihood equation for $\boldsymbol{\beta^*}$ by noting that $\sigma_e^2 > 0$ to obtain the updating equation for $\boldsymbol{\beta}$ at the $(t+1)$th iteration:

$$(\boldsymbol{\beta^*})^{(t+1)} = \Big((\sum_{l=1}^{M}\boldsymbol{X}_l^{*'}\boldsymbol{X}_l^*)^{-1}\Big)'\Big(\sum_{l=1}^{M}(\boldsymbol{y}-\boldsymbol{b}_l^{(t)})'\boldsymbol{X}_l^*\Big)' \tag{3.3.31}$$

Note that $\boldsymbol{X}_l^{*'}\boldsymbol{X}_l^*$ is a symmetric $n \times n$ matrix. The inverse of a symmetric matrix is also symmetric, the sum of symmetric matrices is symmetric, and the transpose of a symmetric matrix is itself. Therefore

$$(\boldsymbol{\beta^*})^{(t+1)} = \Big((\sum_{l=1}^{M}\boldsymbol{X}_l^{*'}\boldsymbol{X}_l^*)^{-1}\Big)\Big(\sum_{l=1}^{M}(\boldsymbol{y}-\boldsymbol{b}_l^{(t)})'\boldsymbol{X}_l^*\Big)' \tag{3.3.32}$$

By differentiating equation 3.3.27 with respect to σ_e^2, the updating equation for σ_e^2 is obtained as:

$$(\sigma_e^2)^{(t+1)} = tr(\boldsymbol{V}^{(t)}) + \frac{1}{M}\sum_{l=1}^{M}(\boldsymbol{y}-\boldsymbol{g}_l\beta_g-\boldsymbol{X}\boldsymbol{\beta_x}-\boldsymbol{b}_l^{(t)})'(\boldsymbol{y}-\boldsymbol{g}_l\beta_g-\boldsymbol{X}\boldsymbol{\beta_x}-\boldsymbol{b}_l^{(t)}) \tag{3.3.33}$$

For the estimate and updating equation of the variance component σ_b^2, we replace $\boldsymbol{D} = \sigma_b^2 K$ and find the derivative with respect to σ_b^2.

$$\begin{aligned}\frac{\partial Q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(l)}\right)}{\partial \sigma_b^2} &= \frac{\partial}{\partial \sigma_b^2}\left\{-\frac{1}{2}\log|\sigma_b^2\boldsymbol{K}| - \frac{1}{2}tr\left((\sigma_b^2\boldsymbol{K})^{-1}\boldsymbol{V}^{(t)}\right) - \frac{1}{2M}\sum_{l=1}^{M}\boldsymbol{b}_l^{(t)\prime}(\sigma_b^2\boldsymbol{K})^{-1}\boldsymbol{b}_l^{(t)}\right\} \\ &= -\frac{n}{2\sigma_b^2} + \frac{1}{2(\sigma_b^2)^2}tr(\boldsymbol{K^{-1}V^{(t)}}) + \frac{1}{2M(\sigma_b^2)^2}\sum_{l=1}^{M}\boldsymbol{b}_l^{(t)\prime}\boldsymbol{K}^{-1}\boldsymbol{b}_l^{(t)}\end{aligned} \tag{3.3.34}$$

Equating to 0 and solving for σ_b^2, we obtain the updating equation as:

$$(\sigma_b^2)^{(t+1)} = \frac{1}{n}\left(tr(\boldsymbol{K}^{-1}\boldsymbol{V}^{(t)}) + \frac{1}{M}\sum_{l=1}^{M}\boldsymbol{b}_l^{(t)\prime}\boldsymbol{K}^{-1}\boldsymbol{b}_l^{(t)}\right) \tag{3.3.35}$$

To summarize, the method of weights employed in the E-step has enabled us to obtain approximate maximum likelihood estimates for the fixed effects and variance component parameters. In particular, we note that our estimates at each iteration do not require knowledge of the random effects $\boldsymbol{b}$.

3.3.2 Summary of our EM algorithm

We now provide a summary of the EM algorithm used to handle missing genotypes in a linear mixed model setting.

- **Step 1:** We obtain starting values for the parameters, i.e $\boldsymbol{\beta}^{*(0)}$, $\sigma_b^{2(0)}$, and $\sigma_e^{2(0)}$, by fitting a LMM to the observed data only using the R packages `Gaston` [5] or `lme4qtl` [87]. When we used Gaston, the GRM $\boldsymbol{K}$ is obtained directly from the program. In contrast, lme4qtl does not provide an estimate of the GRM. Therefore, we utilized GEMMA [83], which does estimate the GRM, and then fed this GRM into lme4qtl for the analysis.

- **Step 2:** For the $(t+1)th$ iteration:

– For individual i with missing genotype data, compute the weights for the jth genotype, $j = 0, 1, 2$ using;

$$w_{ij}^{(t+1)} = \frac{P\left(y_{\mathrm{i}}, \mid g_i = j, \mathbf{x}_i, (\beta_g, \boldsymbol{\beta}_x, \sigma_b^2, \sigma_e^2)^{(\mathrm{t})}\right) \Pr(g_i = j \mid \boldsymbol{\gamma})}{\sum_{k=0}^{2} P\left(y_{\mathrm{i}}, \mid g_i = k, \mathbf{x}_i, (\beta_g, \boldsymbol{\beta}_x, \sigma_b^2, \sigma_e^2)^{(\mathrm{t})}\right) \Pr(g_i = k \mid \boldsymbol{\gamma}^{(t)})}.$$

Sample each individual's genotype M times from this distribution to obtain genotype vectors $\boldsymbol{g}_l = (\boldsymbol{g}_{obs}, \boldsymbol{g}_{mis,l})$, $l = 1, \ldots, M$.

– Compute $\boldsymbol{b}_l^{(t+1)}$ and $V_l^{(t+1)}$ from expressions

$$\boldsymbol{b}_l^{(t+1)} = \left((\sigma_e^2)^{(t)}\right)^{-1} \boldsymbol{V}^{(t)}(\boldsymbol{y} - \boldsymbol{g}_l \beta_g^{(t)} - \boldsymbol{X}\boldsymbol{\beta}_{\boldsymbol{X}}^{(t)}) \tag{3.3.36}$$

$$\boldsymbol{V}_l^{(t+1)} = \left((\sigma_e^2)^{(t)} + \left((\sigma_b^2)^{(t)} \boldsymbol{K}\right)^{-1}\right)^{-1}. \tag{3.3.37}$$

– Compute $\boldsymbol{\beta}^{*(t+1)}$ from the expression:

$$(\boldsymbol{\beta^*})^{(t+1)} = \left(\left(\sum_{l=1}^{M} \boldsymbol{X}_l^{*\prime} \boldsymbol{X}_l^*\right)^{-1}\right)' \left(\sum_{l=1}^{M} (\boldsymbol{y} - \boldsymbol{b}_l^{(t+1)})' \boldsymbol{X}_l^*\right)'$$

– Compute $\sigma_e^{2(t+1)}$ using the equation:

$$\begin{aligned}(\sigma_e^2)^{(t+1)} = tr(\boldsymbol{V}^{(t+1)}) + \frac{1}{M} \sum_{l=1}^{M} (\boldsymbol{y} - \boldsymbol{X}_l^* (\boldsymbol{\beta}^*)^{(t+1)} - \boldsymbol{b}_l^{(t+1)})' \\ (\boldsymbol{y} - \boldsymbol{X}_l^* (\boldsymbol{\beta}^*)^{(t+1)} - \boldsymbol{b}_l^{(t+1)})\end{aligned} \tag{3.3.38}$$

– Compute $\sigma_b^{2(t+1)}$ using equation:

$$(\sigma_b^2)^{(t+1)} = \frac{1}{n} \left(tr(\boldsymbol{K}^{-1} \boldsymbol{V}^{(t+1)}) + \frac{1}{M} \sum_{l=1}^{M} \boldsymbol{b}_l^{(t+1)\prime} \boldsymbol{K}^{-1} \boldsymbol{b}_l^{(t+1)} \right) \tag{3.3.39}$$

– Compute $\boldsymbol{\gamma}^{(t+1)}$ using:

$$\gamma_k^{(t+1)} = \frac{W_k^{(t+1)}}{n} \; k = 0, 1, 2$$

where $W_k^{(t)} = \sum_{i=1}^{n} w_{ik}^{(t)}$.

– Compute convergence criteria and check if the values are smaller than the chosen tolerance values. If yes, go to Step 3. Otherwise set $t = t + 1$ and repeat Step 2. The convergence criteria used are:

1. The number of iterations must be lower than the maximum number of iterations. The maximum will be set to 50.
2. The maximum of the absolute difference from the tth to the $(t+1)$th iteration between any of the regression parameter estimates is below the tolerance. The tolerance will be set to 10^{-3}.
3. The difference in complete data likelihood from iteration t to $t+1$ is below a specified tolerance. The tolerance will be set to 10^{-3}.

- **Step 3:** Return the final estimates as the desired MLEs.

3.4 Hypothesis Testing

Once the MLE estimates have been obtained using the EM algorithm, we need to conduct a hypothesis test to determine whether the genetic factor under investigation is significantly associated with the outcome. For our linear mixed model, we are primarily interested in testing the hypothesis of no association against the alternative; i.e $H_0 : \beta_g = 0$ vs $H_1 : \beta_g \neq 0$.

There are several commonly used likelihood-based hypothesis testing frameworks

that can be used depending on the nature of the data and the research question at hand. In section 3.4.1, we briefly review the likelihood ratio test and describe an approximate procedure for computing the likelihood ratio statistic in our context. In section 3.4.3, we describe a Wald test.

3.4.1 Likelihood Ratio Test

Likelihood ratio tests (LRTs) are a class of general tests that are almost always applicable in likelihood based inference and in most cases an optimal choice [9]. In the context of linear mixed models, LRTs can be used to compare the fit of different fixed and random effects models [26].

Let $L(\theta|x)$ be the likelihood function of a random variable $X_1, \ldots X_n$ with a pdf $f(x|\theta)$. We can define the LRT statistic for testing $H_0 : \theta \in \Theta_0$ versus the alternative $H_1 : \theta \in \Theta_0^c$ as

$$\Lambda = -2 \log W$$

where W is expressed as:

$$W(\boldsymbol{x}) = \frac{\sup_{\Theta_0} L(\theta|\boldsymbol{x})}{\sup_{\Theta} L(\theta|\boldsymbol{x})} \tag{3.4.1}$$

where Θ is used to denote the unrestricted parameter space and $\boldsymbol{\theta_0}$ is the null parameter space. The numerator of (3.4.1) is the maximum probability of the observed sample computed over the parameters in the null hypothesis while the denominator is the maximum probability of the observed sample over all the possible parameters. If $W(\boldsymbol{x})$ is small, then there are parameter points in H_1 for which the observed sample is much more likely than for any parameter point in the null parameter space and we will reject H_0 in this situation.

To compute $W(\boldsymbol{x})$, let $\hat{\theta}$ be the MLE of θ obtained by an unrestricted maximiza-

tion of $L(\theta|\boldsymbol{x})$. Consider also the MLE of θ, $\hat{\theta}_0$, obtained by a restricted maximization over Θ_0 as the parameter space. It is clear that $\hat{\theta}_0 = \hat{\theta}_0(\boldsymbol{x})$ is the value $\theta \in \Theta_0$ that maximizes the likelihood $L(\theta|\mathbf{x})$. Equation (3.4.1) can be written as:

$$W(\boldsymbol{x}) = \frac{L(\hat{\theta}_0|\boldsymbol{x})}{L(\hat{\theta}|\boldsymbol{x})} \tag{3.4.2}$$

From (3.4.2), we see that the likelihood ratio test depends on the likelihood (or log-likelihood) functions of the model under the null and alternative. In the context of missing data, this likelihood must be computed based on the observed data likelihood, which may not be possible to compute. In particular, let $\boldsymbol{x}$ be the vector of missing (continuous) data and $\boldsymbol{y}$ the vector of observed data. The observed data likelihood is the marginal density of $\boldsymbol{y}$, which is an integral expression as shown below:

$$L(\boldsymbol{\theta}) = f_Y(\boldsymbol{y};\boldsymbol{\theta}) = \int f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\theta})d\boldsymbol{x} \tag{3.4.3}$$

As with parameter estimation, this integral may not be available in closed form. Therefore likelihood ratio tests in the context of missing data are also not as straightforward as the non-missing data context. In fact, the literature on how to construct and conduct a likelihood ratio test when there is missing data is quite sparse, as noted by Yang and Kim [82] and Sung [68].

We have used the Monte Carlo EM for obtaining MLE estimates in section 3.3.1. For approximating the observed data likelihood, we used a Monte Carlo approximation method to obtain the observed likelihood function as in Yang and Kim [82].

3.4.2 Likelihood Ratio Test for the EM-LMM algorithm

In this subsection, we describe our approach for carrying out an LRT for testing the hypothesis of no association in our model. The approach is based on the method of Yang and Kim [82] which uses the concept of importance sampling to approximate the observed likelihood.

Recall the Monte Carlo samples generated in section 4.3.1. We are not able to use these samples to impute the data in order to compute the complete-data log-likelihood because these samples are generated from a distribution with fixed parameter values [82]. We use the importance sampling technique for approximating the observed log-likelihood function. The observed density function is given as:

$$f_{obs}(\boldsymbol{y}, \boldsymbol{g}, \boldsymbol{b}; \boldsymbol{X}, \boldsymbol{\theta}) = \sum_{\boldsymbol{g_{mis}}} f(\boldsymbol{y}|\boldsymbol{g}, \boldsymbol{b}, \boldsymbol{X}; \boldsymbol{\theta})\ f(\boldsymbol{g}; \gamma) \tag{3.4.4}$$

where the summation is taken over the missing genotype vector.

Using the marginal model we rewrite the observed density as:

$$f_{obs}(\boldsymbol{y}, \boldsymbol{g}; \boldsymbol{\theta}) = \sum_{\boldsymbol{g_{mis}}} \left[f(\boldsymbol{y}|\boldsymbol{g}; \beta_g, \beta_x, \sigma_b^2, \sigma_e^2)\ f(\boldsymbol{g}; \gamma)\ \frac{Pr(\boldsymbol{g_{mis}} \mid \boldsymbol{g_{obs}}, \boldsymbol{X}, \boldsymbol{y}, \hat{\boldsymbol{\theta}})}{Pr(\boldsymbol{g_{mis}} \mid \boldsymbol{g_{obs}}, \boldsymbol{X}, \boldsymbol{y}, \hat{\boldsymbol{\theta}})} \right] \tag{3.4.5}$$

where the distribution $Pr(\boldsymbol{g}_{mis}|\boldsymbol{g}_{obs}, \boldsymbol{X}, \boldsymbol{y}, \hat{\boldsymbol{\theta}})$ now acts as the proposal distribution, denoted $h(\boldsymbol{g_{mis}})$. Our choice of this distribution stems from the fact that we are able to sample missing genotypes quite easily from it, as done in section 3.3.1.

This sum can be written as an expectation over the distribution:

$$Pr(\boldsymbol{g_{mis}} \mid \boldsymbol{g_{obs}}, \boldsymbol{X}, \boldsymbol{y}, \hat{\boldsymbol{\theta}})$$

as:

$$E\left[\frac{f(\boldsymbol{y} \mid \boldsymbol{g}; \beta_g, \beta_x, \sigma_b^2, \sigma_e^2)\ f(\boldsymbol{g}; \gamma)}{Pr(\boldsymbol{g_{mis}} \mid \boldsymbol{g_{obs}}, \boldsymbol{X}, \boldsymbol{y}, \hat{\boldsymbol{\theta}})}\right] \tag{3.4.6}$$

The expectation can then be approximated by a simple Monte Carlo to obtain the observed log-likelihood as:

$$\ell_{obs}(\boldsymbol{\theta}) = \log(f_{obs}(\boldsymbol{y}, \boldsymbol{g}; \boldsymbol{\theta})) \approx \log\left(\frac{1}{M}\sum_{l=1}^{M}\left[\frac{f(\boldsymbol{y} \mid \boldsymbol{g}_l; \beta_g, \beta_x, \sigma_b^2, \sigma_e^2)\ f(\boldsymbol{g}_l; \gamma)}{h(\boldsymbol{g_{mis}})}\right]\right) \tag{3.4.7}$$

The LRT statistic is now of the form:

$$\Lambda = -2\{\ell_{obs}(\boldsymbol{\theta}_0) - \ell_{obs}(\hat{\boldsymbol{\theta}})\}. \tag{3.4.8}$$

Under the null H_0 and under certain regularity conditions,

$$\Lambda \xrightarrow{d} \chi^2_{(p)}$$

where p is the difference in the number of parameters in H_0 and H_1. We reject H_0 if

$$\Lambda > \chi^2_{(p),1-\alpha}$$

where α is the significance level.

The procedure for constructing the LRT for our model can be summarized as:

1. Under the null, we need to obtain the log-likelihood function for the marginal linear mixed model with no genetic covariates ($\beta_g = 0$). The density function corresponding to this can be expressed as the density of the marginal normal

distribution for y and the multinomial density of g. i.e

$$f(\boldsymbol{y}|\boldsymbol{X}; \hat{\beta}_{x(0)}, \hat{\sigma}^2_{b(0)}, \hat{\sigma}^2_{e(0)})\; p(\boldsymbol{g_l} \mid \hat{\boldsymbol{\gamma}}) \tag{3.4.9}$$

where

$$\boldsymbol{y} \mid \beta_x, \sigma^2_b, \sigma^2_e \sim N_n\left(\boldsymbol{X}\beta_x, V_y\right)$$

and

$$p(\boldsymbol{g}; \boldsymbol{\gamma}) = \prod_{i=1}^{n} Pr(g_i \mid \gamma)$$

where $g_i \sim \text{Multinomial}(2, \hat{\gamma})$.

2. For the observed likelihood under H_1, we compute the approximate log-likelihood $\ell_{obs}(\hat{\theta})$ by:

   - Sample missing genotypes from the proposal distribution $h(\boldsymbol{g_{mis}})$ given by

   $$Pr(\boldsymbol{g_{mis}} \mid \boldsymbol{g_{obs}}, \boldsymbol{X}, \boldsymbol{y}, \hat{\boldsymbol{\theta}}).$$

   - This distribution has been named "weights" in section 3.3.1 but here, we sample from this distribution by letting $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. i.e θ at the MLE. For each missing genotype sampled, we would need to keep track of the sample and its associated weight.
   - By multiplying these weights, we will obtain the density of the proposal distribution which is the denominator in the equation 3.4.7 of the observed log likelihood ℓ_{obs}.

- For the numerator, we compute two densities: the marginal density evaluated at the MLE i.e

$$f(\boldsymbol{y} \mid \boldsymbol{g}_l; \hat{\beta}_g, \hat{\beta}_x, \hat{\sigma_b^2}, \hat{\sigma_e^2})$$

and

$$f(\boldsymbol{g}_l; \hat{\boldsymbol{\gamma}}) = \prod_{i=1}^{n} Pr(g_i \mid \hat{\gamma}).$$

- The observed log-likelihood evaluated at the MLE is now obtained using equation (3.4.7) above.

3. The likelihood ratio statistic is then computed based on Wilk's theorem as:

$$\Lambda = -2\left\{l_{obs}\left(\hat{\boldsymbol{\theta}}_0\right) - l_{\text{obs}}\left(\hat{\boldsymbol{\theta}}\right)\right\} \sim \chi_p^2 \tag{3.4.10}$$

3.4.3 The Wald Test

Another statistical test that can be used for inference on β_g is an approximate Wald test. The standard error of the $\boldsymbol{\beta_g}$ component in $\boldsymbol{\beta^*}$ is given as:

$$\text{SE}(\hat{\boldsymbol{\beta}}_g) = \sqrt{V(\hat{\boldsymbol{\beta}}_g)}$$

where $\hat{\beta}_g$ is the MLE estimate of β_g. Given the standard error estimate, the null and alternative hypotheses $\text{H}_0 : \beta_g = 0$ versus $\text{H}_\text{A} : \beta_g \neq 0$ can be tested by using the following Wald statistic:

$$Z = \frac{\hat{\beta}_g}{\text{SE}(\hat{\boldsymbol{\beta}}_g)}$$

where $\text{SE}(\hat{\boldsymbol{\beta}}_g)$ is the standard error of $\hat{\boldsymbol{\beta}}_g$.

The asymptotic variance of the estimators is approximated with the inverse of

the Fisher information matrix (FIM) $V(\hat{\boldsymbol{\theta}}) \approx [\mathcal{I}(\hat{\boldsymbol{\theta}})]^{-1}$. The FIM, $\mathcal{I}(\boldsymbol{\theta})$ is usually approximated by the observed information matrix, $I(\boldsymbol{\theta})$ in a missing data context. The standard error can be calculated using Louis's method [41] given as:

$$I(\boldsymbol{\theta}) = E_{\theta}\left[\boldsymbol{I}_c(\boldsymbol{\theta}) \mid \boldsymbol{g}_{obs}, \boldsymbol{X}, \boldsymbol{y}\right] - \operatorname{var}\left[S_c(\boldsymbol{\theta}) \mid g_{obs}, \boldsymbol{X}, \boldsymbol{y}\right] \tag{3.4.11}$$

where $I(\boldsymbol{\theta})$ is the negative hessian of the observed data log-likelihood, $I_c(\boldsymbol{\theta})$ is the negative Hessian of the complete-data log-likelihood function and $S_c(\boldsymbol{\theta})$ is the score of the complete-data log-likelihood function. When the expectations in equation 3.4.11 are not easily computed, a Monte Carlo approach by Tanner [69] can be used to estimate the observed matrix. The estimated observed information matrix for $\boldsymbol{\theta}$ at $\hat{\boldsymbol{\theta}}$ is given by:

$$I(\widehat{\boldsymbol{\theta}}) = \frac{1}{M}\sum_{i=1}^{M} \frac{\partial^2 \log p\left(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{z_i}\right)}{\partial \boldsymbol{\theta}^2}\bigg|_{\hat{\boldsymbol{\theta}}} + \frac{1}{M}\sum_{i=1}^{M}\left(\frac{\partial \log p\left(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{z_i}\right)}{\partial \boldsymbol{\theta}}\bigg|_{\hat{\boldsymbol{\theta}}}\right)^2 \tag{3.4.12}$$

Using the values sampled from the Monte Carlo, we have:

$$\begin{aligned} I(\hat{\boldsymbol{\theta}}) &= E_{\theta}\left[\boldsymbol{I}_c(\boldsymbol{\theta}) \mid \boldsymbol{g}_l, \boldsymbol{X}, \boldsymbol{y}\right] - \operatorname{var}\left[S_c(\boldsymbol{\theta}) \mid \boldsymbol{g}_l, \boldsymbol{X}, \boldsymbol{y}\right] \\ &= \frac{\partial^2}{\partial \boldsymbol{\theta}\boldsymbol{\theta}'}\boldsymbol{Q}\left(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}\right)|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} - \frac{1}{M}\sum_{l=1}^{M} S\left(\widehat{\boldsymbol{\theta}}; \boldsymbol{X}_l, \boldsymbol{y}, \hat{\boldsymbol{b}}_l\right) S\left(\widehat{\boldsymbol{\theta}}; \boldsymbol{X}_l, \boldsymbol{y}, \hat{\boldsymbol{b}}_l\right)' \end{aligned} \tag{3.4.13}$$

Here, the missing genotypes have been replaced by their sampled values from the conditional distribution at convergence and we know that $\boldsymbol{X}_l = [\boldsymbol{g}_l, \boldsymbol{X}]$. We are interested in the observed information for the fixed effect parameter β^* and we now outline the procedure for obtaining this quantity using the Monte Carlo approach of Tanner.

- The observed information matrix for $\boldsymbol{\beta}^*$, $I(\boldsymbol{\beta}^*)$ can be expressed as:

$$I(\hat{\boldsymbol{\beta}}^*) = -\frac{\partial^2}{\partial \boldsymbol{\beta}^* \boldsymbol{\beta}^{*\prime}} \boldsymbol{Q}\left(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}\right) |_{\boldsymbol{\beta}^*=\hat{\boldsymbol{\beta}}^*} - \frac{1}{M}\sum_{l=1}^{M} S\left(\hat{\boldsymbol{\beta}}^*; \boldsymbol{X}_l, \boldsymbol{y}, \hat{\boldsymbol{b}}_l\right) S\left(\hat{\boldsymbol{\beta}}^*; \boldsymbol{X}_l, \boldsymbol{y}, \hat{\boldsymbol{b}}_l\right)' \tag{3.4.14}$$

- We can approximate each of these quantities by sampling from $Pr(\boldsymbol{g_{mis}} \mid \boldsymbol{g_{obs}}, \boldsymbol{X}, \boldsymbol{y}, \hat{\boldsymbol{\theta}})$ which are our weights from equation **??**, evaluated at the MLE. The first term is approximated by:

$$\frac{1}{M}\sum_{l=1}^{M} \frac{\partial^2}{\partial \boldsymbol{\beta}^* \boldsymbol{\beta}^{*\prime}} \boldsymbol{Q}\left(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}\right)$$

where

$$\frac{\partial Q\left(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}\right)}{\partial \boldsymbol{\beta}^*} = \frac{1}{M\hat{\sigma}_e^2}\sum_{l=1}^{M}\left(\left(\boldsymbol{y} - \hat{\boldsymbol{b}}_l\right)' \boldsymbol{X}_l^* - \hat{\boldsymbol{\beta}}^{*\prime} \boldsymbol{X}_l^{*\prime} \boldsymbol{X}_l^*\right).$$

and

$$\frac{\partial^2 Q\left(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}\right)}{\partial \boldsymbol{\beta}^* \boldsymbol{\beta}^{*\prime}} = \frac{1}{M\hat{\sigma}_e^2}\sum_{l=1}^{M} \boldsymbol{X}_l^{*\prime} \boldsymbol{X}_l^*$$

- The score function is approximated by:

$$S\left(\hat{\boldsymbol{\beta}}^*; \boldsymbol{X}_l, \boldsymbol{y}, \boldsymbol{b}_l\right) = \frac{\partial \ell_c(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}^*} = \frac{1}{\widehat{\sigma}_e^2}\sum_{l=1}^{M}\left((\boldsymbol{y} - \hat{\boldsymbol{b}}_l)' \boldsymbol{X}_l^* - \hat{\boldsymbol{\beta}}^{*\prime} \boldsymbol{X}_l^{*\prime} \boldsymbol{X}_l^*\right) \tag{3.4.15}$$

- The observed information matrix for $I(\hat{\boldsymbol{\beta}}^*)$ under a Monte Carlo approximation

is:

$$I(\hat{\boldsymbol{\beta}}^{*}) = \frac{1}{M\hat{\sigma_e^2}} \sum_{l=1}^{M} \boldsymbol{X}_l^{*\prime} \boldsymbol{X}_l^{*} + \frac{1}{M\hat{\sigma_e^2}} \sum_{l=1}^{M} \left((\boldsymbol{y} - \hat{\boldsymbol{b}}_l)' \boldsymbol{X}_l^{*} - \hat{\boldsymbol{\beta}}^{*\prime} \boldsymbol{X}_l^{*^{\prime}} \boldsymbol{X}_l^{*} \right) \left((\boldsymbol{y} - \hat{\boldsymbol{b}}_l)' \boldsymbol{X}_l^{*} - \hat{\boldsymbol{\beta}}^{*\prime} \boldsymbol{X}_l^{*^{\prime}} \boldsymbol{X}_l^{*} \right)' \tag{3.4.16}$$

- The asymptotic variance $V(\hat{\boldsymbol{\beta}}^{*})$ based on Tanner [70] is obtained as the inverse of the fisher information for $\boldsymbol{\beta}^{*}$ i.e

$$V(\hat{\boldsymbol{\beta}}^{*}) = [I(\hat{\boldsymbol{\beta}}^{*})]^{-1}.$$

3.5 Conclusion

In this chapter, we have developed an Expectation-Maximization (EM) algorithm tailored to address missing genotype data in the context of linear mixed models (LMMs). We began by introducing the concept of linear mixed models and how to estimate the fixed and random effects parameters when we have complete data. We then described the EM algorithm in detail, outlining the two main steps: the expectation (E) step, where missing values are imputed based on the current parameter estimates, and the maximization (M) step, where model parameters are updated using the complete data. Next, we outlined our LMM model and derived the EM algorithm for modelling the type of data we were interested in. We derived the EM updates for the parameters of interest, including fixed effects coefficients and variance components. Finally we derived two hypothesis tests and procedure. In the next chapter, we apply our method to both simulated and real data to evaluate the type 1 error, bias and power.

Chapter 4

Evaluating the Performance of the EM-LMM algorithm

4.1 Introduction

In this chapter, we evaluate the performance of the EM-LMM algorithm on real and simulated datasets. We first present a simulation study with missing data and population stratification to assess the type 1 error and power under various conditions of missingness. Type 1 error, also known as false positive rate, refers to the probability of erroneously rejecting a true null hypothesis. We varied the proportion of missing data, the missing data mechanism (EPS and randomly missing in the statistical sense i.e MCAR), and the sample size to assess their effects on the algorithm's accuracy and robustness. We also conducted simulation experiments to assess the power of our methods in these scenarios. By manipulating these factors in a controlled manner, we aim to identify conditions under which the EM algorithm may exhibit inflated or deflated type 1 error rates, thereby providing guidance for its practical application

in real-world research settings. Quantifying the probability of this error is crucial in determining the reliability and credibility of this approach. Power is the probability of correctly rejecting the null. By comparing power of the EM-LMM approach to analysis with the full dataset, we aim to determine the effect of missing data.

We also present an application of our EM algorithm on a real population-based dataset collected to investigate prostate cancer risk factors that is known to have population substructure. We selected extremes of the BMI response variable in order to evaluate the estimation of model parameters.

This chapter is organized as follows: Section 4.2 first describes the data simulation strategy for simulating the genotypes and phenotypes for assessing the type 1 error and power. This Section also describes the PROtEuS study, the variables in the dataset and a comprehensive quality control carried out on the genetic data. The analysis methods are then summarized for both the simulation study and real-data analysis. In Section 4.3 we display and summarize the results from our analysis in terms of the type 1 error, power and the model parameters and in Section 4.4, we provide concluding remarks.

4.2 Data and Methods

4.2.1 Data Simulation Strategy

In this section, we describe the approach used to simulate genetic covariates and phenotypes in our simulation studies. Assuming common variants, our approach to simulating the genetic covariates and candidate SNP follows that described in Onifade et al. [50].

We assumed a cohort consisting of two subpopulations of equal proportion. The

total cohort size N, was set to $5,000$ and $10,000$. The F_{st} value, a measure of population substructure between the two populations was set to 0.01. This is higher than would be expected between typical European populations but it ensures substantial population substructure as noted previously in [52]. We first simulated the genetic data using the Balding-Nichols method [2, 58] and as previously described in [52]. For each individual, we simulated a total of $p = 5000$ SNPs. Though true genome-wide data would consist of much larger number of SNPs, our previous work with data simulated using this model has shown that this number of SNPs is sufficient to correct for population stratification [52]. For each SNP, the generating allele frequency, p, was sampled from a uniform $[0.1, 0.9]$ distribution. To mimic population differentiation, the allele frequency within each of the two populations, p_1 and p_2, was sampled from a Beta distribution with shape and scale parameters $\frac{p(1-F_{st})}{F_{st}}$ and $\frac{(1-p)(1-F_{st})}{F_{st}}$, respectively. This approach has been shown to generate genotype data having the desired F_{st} level [2].

Using the allele frequencies p_1 and p_2 the genotype data was sampled assuming Hardy Weinberg equilibrium within each subpopulation. The genotype data was coded as 0, 1 or 2 for the number of minor alleles. To test the hypothesis of no association, we simulate a single locus to act as our candidate SNP. To simulate the candidate SNP, we assumed that the '1' alelle frequency was $p_1 = 0.25$ in the first subpopulation and $p_2 = 0.85$ in the second subpopulation. Although this allele frequency difference is probably not realistic in practice, it was chosen to reflect a 'worst case' scenario of a SNP that showed extreme population differentiation. Since our mixed model-based method required the use of a GRM matrix to correct for population stratification, the genetic data based on the $p = 5000$ SNPs simulated under the Balding Nichols model was used to compute the GRM , K, using Gaston;

a software for fitting genetic mixed models in R [5]. The candidate SNP for the association test was not included in the computation of the GRM.

Simulations to assess type 1 error

After simulating genotypes, phenotypes were simulated using a linear mixed model which required specifying the fixed effects of the genotypes and covariates, the random effect components, accounting for the population structure in the study and observational noise [46]. There are several existing software packages that are capable of generating realistic phenotypes by specifying these components; these include PhenotypeSimulator [46], MultiPhen [51], MultitraitGWAS [55] and Gaston [5]. We used Gaston mainly due to the ease of use. We selected a total genetic variance of 0.4 while leaving about 0.6 to be explained by the noise terms. Since we are simulating data to estimate the type 1 error, the genetic fixed effect is set to 0. We did not simulate any non-genetic covariates. To mimic EPS sampling, we ordered the phenotype variable and retained genotype data for only the top and bottom 20% of the phenotype distribution. For comparison purposes, we also generated datasets with sporadically missing genotype data. We randomly removed 5%, 10% and 20% for our randomly missing scenarios.

For each scenario, we simulated 1000 datasets. For each simulated dataset, we analyzed the data using the methods described in section 4.2.3 which was programmed in R. Due to inconsistent results with the LRT, hypothesis testing used the Wald test. We stored whether we reject or fail to reject the null hypothesis of $\beta_g = 0$ at a significance level of 0.05. The type I error rate for each scenario considered is estimated by the proportion of the simulated datasets where the null hypothesis was rejected at level $\alpha = 0.05$. If the method has good type 1 error control, we expect

Table 4.1: Parameters for type 1 error simulations

Description	Values
Population differentiation (F_{st})	0.01
Number of individuals N	5000
Number of subpopulations	2
Number of SNPs p	5000
Allele frequency in population 1 p_1	0.25
Allele frequency in population 2 p_2	0.85
Total number of simulations n	1000
Phenotype mean μ	0
σ_b^2, σ_e^2	$0.4, 0.6$

this proportion to be approximately 0.05. Simulations were run in a cluster computing environment known as the Digital Research Alliance of Canada (CAC-FRONTENAC, GRAHAM and CEDAR) and all analysis of the results was done in R [71]. Table 4.2 gives a summary of the values used in simulating both the genotypes and phenotypes for the type 1 error estimation.

Simulations to investigate Power

To evaluate the ability of our model to detect true associations between genetic variants and the phenotype, we carried out a simulation study aimed at assessing the power when there is no confounding due to population stratification. We simulated genetic data for estimating ancestry using the same procedure as for the type 1 error simulations. To simulate the candidate SNP, we assumed no differences in allele frequency between the two populations and an allele frequency of 0.2 for the causal allele. The genotypes at the candidate SNP were sampled assuming Hardy-Weinberg equilibrium. The phenotype was again simulated under a linear mixed model where the genetic covariate effects β_g, are set to $\beta_g = 0.10$ or $\beta_g = 0.20$. The EPS and

Table 4.2: Parameters for power simulations

Description	Values
Population differentiation (F_{st})	0.01
Number of individuals N	5000
Number of subpopulations	2
Number of SNPs p	5000
Causal Allele frequency	0.2
Effect Sizes β_g	0.10, 0.20

randomly missing genotypes are obtained using the same approach as for the type 1 error. For each scenario, we simulated 2000 datasets. The power is estimated by the proportion of hypothesis tests rejected at $\alpha = 0.05$. Simulations were again run in a cluster computing environment.

4.2.2 The PROtEuS Study

prostate cancer is the most frequent cancer diagnosed in men in the western world and the known risk factors leading to its development include age, family history of prostate cancer and ancestry [67]. The data for this study was obtained from the prostate Cancer and Environment Study (PROtEus), a case-control study undertaken in Montreal to evaluate the role of environmental and lifestyle factors in prostate cancer development [53]. In this section, we provide a brief summary of the PROtEuS study; a detailed description has been described elsewhere [6, 7, 53].

The PROtEuS study contains phenotype, exposure variable and SNP (genotype) data. The phenotypic data was available for 3969 men with about 1933 cases and 1994 controls. The eligible subjects for the study were men who are Canadian citizens, under 76 years of age at the time of diagnosis or selection, are resident in the greater Montreal area and registered on Quebec's permanent electoral list. Data col-

lection included face-to-face interviews with participants providing information about socio-demographic characteristics, a wide range of lifestyle-related factors, a prostate cancer screening history and detailed occupational histories. Ethics committees at all participating institutions approved the protocol and subjects provided informed consent. Cases were ascertained through pathology departments across all 11 French hospitals, of the total of 14 hospitals that diagnose prostate cancer in the Montreal area. All patients diagnosed with primary, histologically confirmed prostate cancer (International Classification of Diseases, 10th revision, code C61) between September 2005 and June 2008 were included. Control subjects were selected concurrently from the population-based provincial electoral French-speaking list, and frequency-matched to cases by 5-year age groups. The electoral list is thought to represent a nearly complete listing of Canadian citizens residing in the province of Quebec. Controls were drawn randomly from an area comprising 39 electoral districts corresponding to those of the case series. Subjects were still eligible if they had a history of cancer other than prostate cancer.

Data was collected on several continuous variables in this study. These include socio-demographic characteristics, anthropometric, life style and environmental factors. Due to individuals with diverse self reported ancestry in the data, we expect the data to have some level of population substructure. Body mass index (BMI) was measured on participants and is one of the indicators of physical health that indicates whether or not a respondent is considered overweight. The BMI values for adults are normally categorized as follows: < 18.5 is considered underweight, an individual between $[18.5, 24.9]$ is considered healthy while > 30 is obese.

To apply our method, we require a continuous phenotype; we chose BMI from the available variables. Since the cases were oversampled relative to the general

population we carried out an ANOVA to test for the association between the BMI and prostate cancer status. BMI was not associated with the prostate cancer status (P-value=0.48). We also plotted a distribution of the BMI for four groups of cancer severity: control, low severity of prostate cancer (Pca), high severity of Pca and those with unknown status. As observed from Figure 4.1, the distribution of BMI is similar across the control and case groups. Since there weren't any significant differences in the mean BMI among the groups, cases and controls were combined into a single sample for our association analysis.

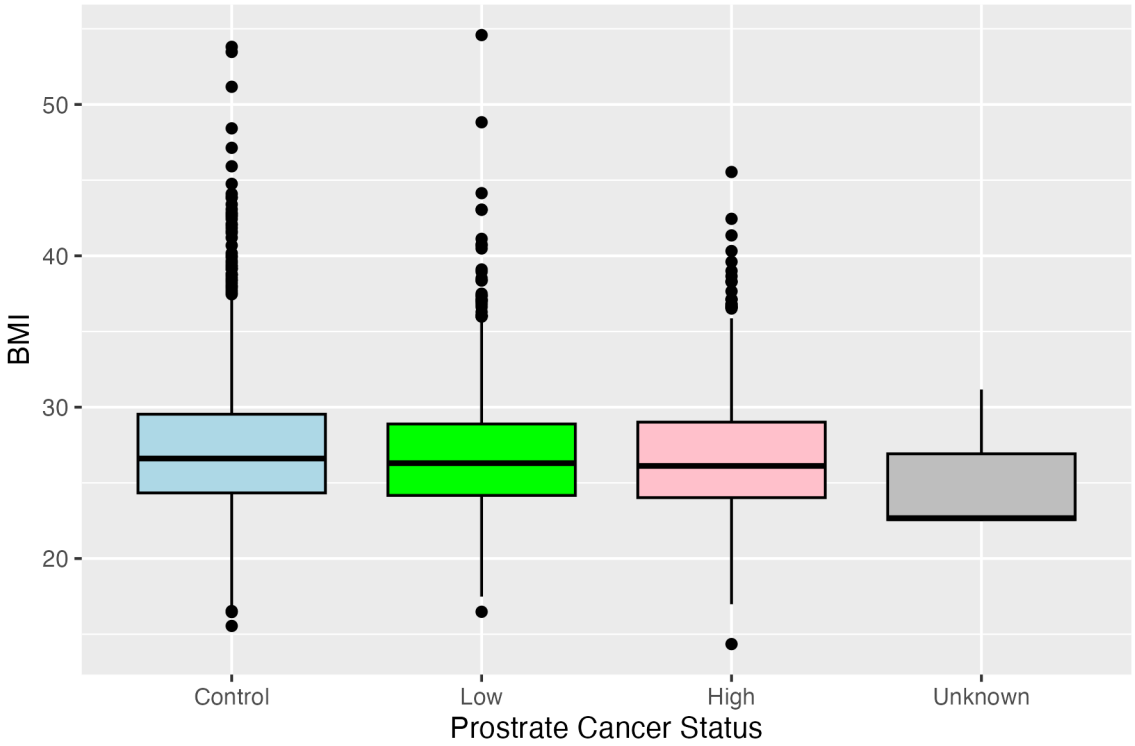


Figure 4.1: Distribution of BMI by prostate cancer status. These categories are based on the Gleason score: Low refers to a less severe form of cancer with Gleason $<= 7$ and high refers to a Gleason score > 7.

Table 4.3 provides a summary of the study population, including demographic information, baseline characteristics, exposure and outcome variables and other rel-

evant information. The median age of cases and controls was around 65 years and cases were slightly younger than controls. Cases are also more likely than controls to have a family history of prostate cancer (Pca). Compared to men of European ancestry, African men had higher risk while Asian men had lower risks of developing Pca. Cases and controls had similar values in terms of education.

Table 4.3: Selected summary statistics among the cases and controls in the PROtEuS study population

Variables	**Cases** ($n = 1921$)	**Controls** ($n = 1982$)
Age, median (IQR)	$64(59 - 69)$	$65(61 - 70)$
BMI, kg/m^2; median (IQR)	$26.8(24.1 - 29.1)$	$26.6(24.3 - 29.5)$
Ancestry, n **(%)**		
European	1680 (87.5%)	1687 (84.8%)
African	90 (4.7%)	130 (6.7%)
Asian	73 (3.8%)	24 (1.2%)
Other	31 (1.6%)	29 (1.5%)
Do not know	14 (0.7%)	12 (0.6%)
Family history of Pca, n **(%)**		
No	1728 (89.9%)	1411 (71.2%)
Yes	198 (10.3%)	448 (22.6%)
Do not know	56 (2.9%)	62 (3.1%)
PCa grade, n **(%)**		
Low-grade PCa	1728 (89.9%)	807 (40.7%)
High-grade PCa	198 (10.3%)	1114 (56.2%)
Education		
Primary	429 (21.5%)	449 (23.2%)
Secondary/College	953 (47.8%)	891 (46.1%)
University	611 (30.7%)	592 (30.6%)

Quality Control (QC) for genetic data from PROtEuS Study

A vital step before genetic data analysis is quality control (QC). The aim of QC is to remove genotype data more likely to have genotyping errors. Genotyping errors can be due to poor quality of individual samples, cross-contamination between samples either during or after collection and processing or storage that introduced foreign genetic material. Genotyping errors can lead to erroneous conclusions if used in the association analysis. Addressing these sources of errors involves implementing quality control measures and validation techniques. We followed the QC steps from protocol [44].

1. Remove SNPs and individuals with high levels of missingness. We removed (filtered out) any individuals who have $> 2\%$ missing genotype data and removed SNPs that have $> 2\%$ missing. It is important to perform SNP filtering before individual filtering. After this step, $11,784$ SNPs and 15 individuals were removed.

2. Sex discrepancy: In this step, we checked for discrepancies between sex of the individuals recorded in the dataset and their sex based on X chromosome heterozygosity/homozygosity rates from the genotype data. We checked that for all the individuals reported, the X chromosome homozygosity estimate (F-value) is > 0.8. We obtained an individual with an unknown sex and removed them from the dataset.

3. Autosomal SNPs: We filtered out non-autosomal SNPs. These are the SNPs that are not between chromosomes 1-22. We deleted $17,799$ non autosomal SNPs.

4. Minor allele frequency (MAF): In this study, we are interested in using common variants or SNPs with a MAF > 0.05. With the study design, we have little to no power for detecting rare SNP-phenotype associations [78]. Hence, we removed $100,283$ SNPs that didn't meet this condition. (Figure 4.2)

5. Hardy-Weinberg equilibrium (HWE): We also excluded SNPs / markers which deviated from Hardy-Weinberg equilibrium. In general, for quantitative traits, a HWE p-value of $< 1e-6$ is recommended for GWAS [44]. 1279 SNPs were removed.

Lastly, we removed the individuals with missing phenotypes. All QC and preprocessing was done with PLINK (version 1.90b), a genome association analysis toolset that is designed to conduct analyses in a computationally efficient manner [61]. After the QC steps and preprocessing, we obtained a genetic sample of 2527 individuals and $574,561$ SNPs. A summary of these steps and the number of individuals and SNPs removed at each step is shown in Figure 4.3.

4.2.3 Models and software implementation

Model for Simulation study

For the data simulated in Section 4.2.1, we are interested in carrying out a candidate SNP study using our EM-LMM algorithm. Recall the LMM model for genetic data analysis discussed in Chapter 3, section 3.3. We consider the model with only genetic effects and this is given as:

$$\boldsymbol{y} = \boldsymbol{X}^*\boldsymbol{\beta}^* + \boldsymbol{I}\boldsymbol{b} + \epsilon \tag{4.2.1}$$

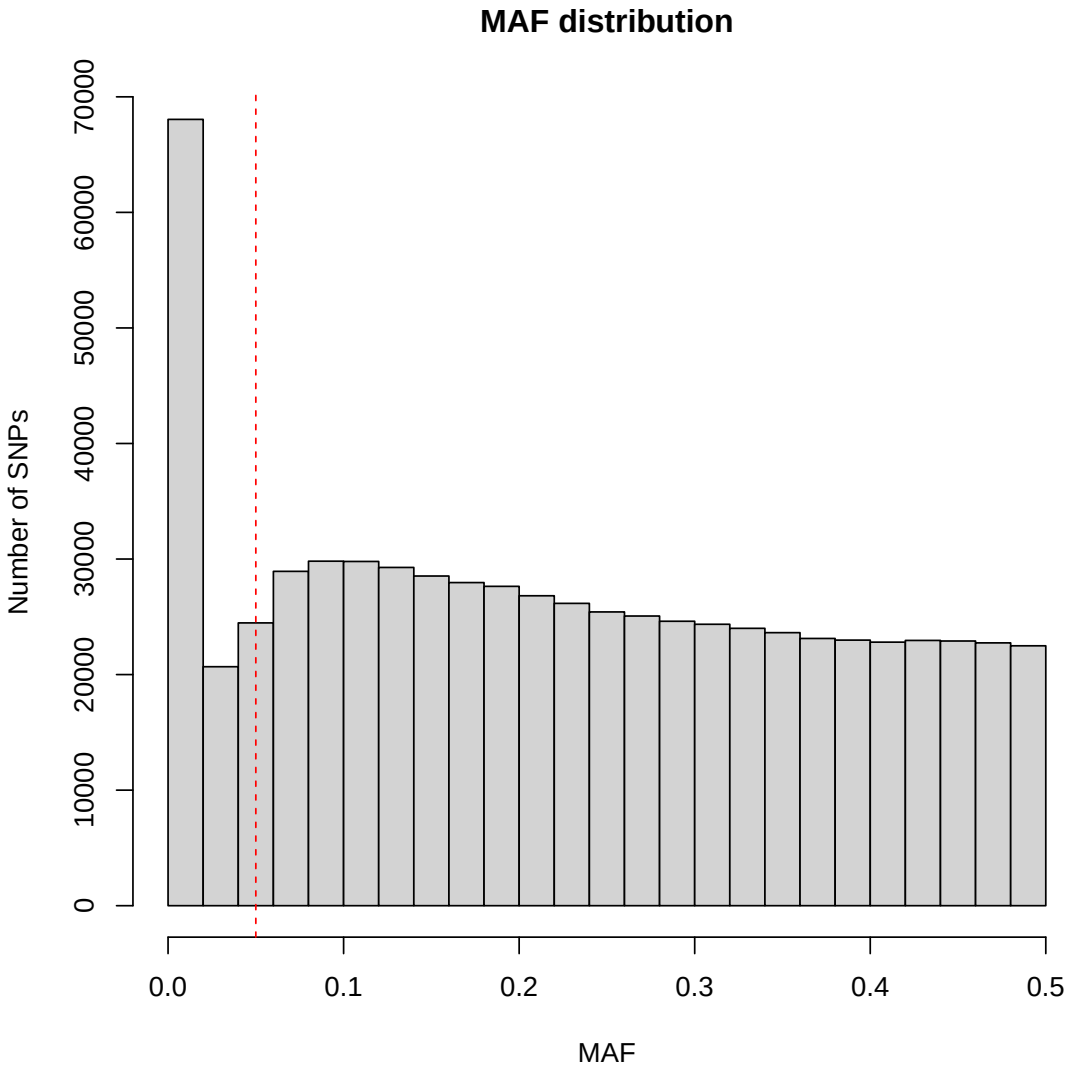
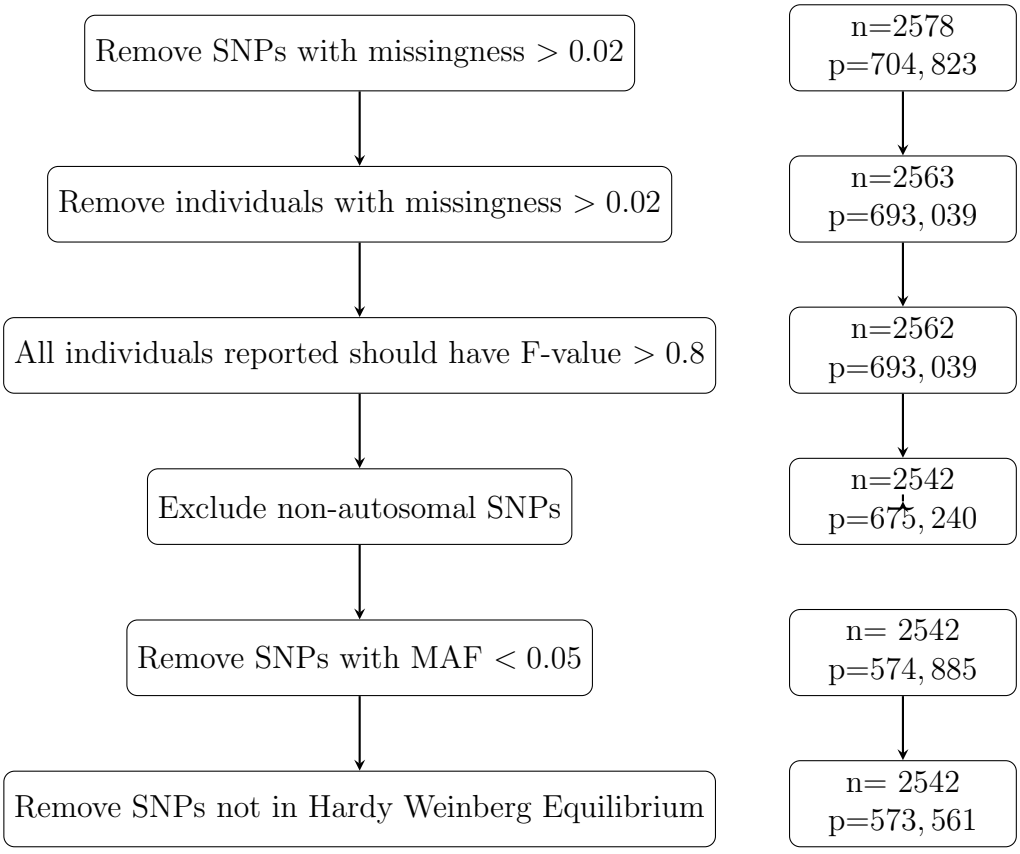


Figure 4.2: The histogram showing MAF distribution of the SNPs in the PROtEuS study. The red vertical line indicates the SNPs that will be removed at MAF < 0.05 These are SNPs to the left of the vertical line.

Figure 4.3: Summary of the quality control steps carried out: n is the number of individuals while p is the number of SNPs.



where $\boldsymbol{y}$ is the $n \times 1$ simulated phenotype values, $\boldsymbol{X^*} = [\mathbf{1}, \boldsymbol{g}]$ is the design matrix corresponding to the intercept and candidate SNP, $\boldsymbol{\beta^*} = [\beta_0, \beta_g]'$ is the vector of regression coefficients with β_0 the intercept and β_g the genetic effect. $\boldsymbol{b_{n\times 1}}$ is the vector of random effects assumed to be $\boldsymbol{b} \sim N_n(0, \sigma_b^2 \boldsymbol{K})$ where $\boldsymbol{K}$ is the $n \times n$ genetic relationship matrix (GRM) estimated from the $p = 5000$ SNPs simulated under the Balding-Nichols model. The model errors are $\boldsymbol{\epsilon} \sim N_n(0, \sigma_e^2 \boldsymbol{I}_n)$. We assume that all $\boldsymbol{b}$ and $\boldsymbol{\epsilon}$ are mutually independent.

Fitting the EM-LMM model involves estimation of β_0 and β_g and the variance components σ_b^2 and σ_e^2. The model was programmed in R and the initial values (fixed effects and variance components) for starting the EM algorithm was obtained from running a linear mixed model. For the initial genotype parameters, we used the calculated genotype frequencies from the available SNP data. We ran the algorithm on both the EPS and sporadically missing datasets. We used the Wald test described in section 3.4.3 for testing the hypothesis of no association in the EM-LMM model. Gaston and GEMMA methods.

Model for the PROtEuS study

Since the PROtEuS study consists of men from different ethnic backgrounds, we expect that there might be some level of population subdivision in the dataset and therefore it is necessary to correct for population stratification. Linear mixed modelling is an appropriate analysis strategy for this type of data. However since this is real data, we are not sure if there are true associations or population structure.

A LMM model based on this dataset can be written as follows:

$$\boldsymbol{y} = \boldsymbol{g}\beta_g + \boldsymbol{I}\boldsymbol{b} + \epsilon \tag{4.2.2}$$

where $\boldsymbol{y}$ is the $n \times 1$ vector of BMI for $n = 2539$ individuals. Performing a GWAS using an LMM involves fitting the model to each SNP in the genome hence the genetic factor $\boldsymbol{g}$ is a $n \times 1$ vector of genotypes for a particular SNP. β_g is the genetic effect of each SNP and $\boldsymbol{b}$ is the random effect assumed to be distributed as $\boldsymbol{b} \sim N(0, \sigma_b^2 \boldsymbol{K})$ where $\boldsymbol{K}$, the GRM is estimated from the genomewide SNPs using GEMMA. The model error is $\boldsymbol{e} \sim N_n(0, \sigma_e^2 \boldsymbol{I}_n)$. We first analyzed a single SNP (rs903924) from the ProtEus study as the genetic covariate and the BMI as our response variable using Gaston [5] and lme4qtl [87] . The estimates obtained from these two methods were similar so we used the values from Gaston as the initial values for the EM-LMM algorithm.

4.3 Results

4.3.1 Simulation Study

Type 1 error estimation

We evaluated the type 1 error rate for a total cohort sample size of 5000 and across different levels of missingness. Table 4.4 summarizes the findings regarding the type 1 error rate for various missing scenarios. Notably, the results indicate consistently low values for the type 1 error rate across all simulated conditions. At a sample size of 5000, the type 1 error rates of EM-LMM at $\alpha = 0.05$ for the EPS scenario was 0.017. Therefore the method is conservative. We also estimated the type 1 error rate for randomly missing genotype scenarios (called RM). The type 1 error rate remained consistently low across different levels of missingness for the RM scenarios (Table 4.4, row 1.) The type 1 error decreased as missing genotypes increased. When we

compared these values to the type 1 error rate obtained from using Gaston on the full dataset, we obtained a type 1 error rate of 0.021, which is close to what we obtained with the EM-LMM analyses. For example, when 5% was missing, the type 1 error with EM-LMM was 0.026 while it was 0.011 when 20% was missing. For GEMMA, we observe values close to the nominal value of 0.05 for both the full and randomly missing datasets indicating adequate control of the type 1 error. These results suggests that the LMM approaches we investigated are conservative and that EM-LMM may become more conservative when there is additional uncertainly due to missing data.

Table 4.4: Estimated type 1 error rates for different software under various missing data scenarios

Number of Samples	**Method**	**Type 1 error**				
		Full Dataset	**EPS**	**RM 5%**	**RM 10%**	**RM 20%**
5000	EM-LMM	-	0.0169	0.0264	0.0252	0.0108
	GASTON	0.021	-	0.021	0.022	0.025
	GEMMA	0.053	-	0.053	0.051	0.052

[a] EPS: Extreme Phenotype Sample – A subset of the data selected based on 20% extreme values of the phenotype.
[b] RM: Randomly Missing 5%, 10%, and 20% indicate cases where 5%, 10%, and 20% of the data are randomly missing, respectively.
[c] EM-LMM: our EM Algorithm for Linear Mixed Models

In Table 4.5, we present a comparison of the average estimates and biases obtained from two different estimation methods: EM-LMM and GASTON. We excluded GEMMA since it doesn't provide variance component estimates by default. These estimates provide insights into the performance of each method in accurately estimating the parameters of interest under a simulated scenario. For β_0, the EM-LMM method provides an average estimate that is very close to the true value of 1, with

negligible bias, for both the EPS and random missing scenarios. This indicates accurate estimation of β_0. Gaston method is also close to the true value. For β_g, both EM-LMM provide estimates close to the true value of 0 for β_g. For both EPS and RM scenarios. GASTON applied to the full dataset also shows little bias. For the variance component σ_b^2, only GASTON yields an estimate close to the true value of 0.4 with negligible biases. With our EM algorithm, we had a highly biased estimate of σ_b^2 indicating that our method is overestimating this parameter. For the EPS sample, the average estimate is 4.5720, which also leads to a high bias of 4. Similar bias is observed for the RM scenario. In contrast, Gaston provides a more accurate estimates, with a bias that is close to 0. The biases for σ_e^2 for EPS and RM are also observed with σ_e^2. In contrast, Gaston provides a more accurate estimate, with a bias close to 0. The genetic parameters γ_0, γ_1, and γ_2 are only available for our EM method. Compared to the true values, both the EPS and randomly missing (RM) forms of our algorithm provide estimates for these parameters with negligible biases indicating accurate estimation.

Power

Table 4.6 presents the estimated power for the two effect sizes ($\boldsymbol{\beta_g}$) across the missing data scenarios. We observed that the EM-LMM demonstrates a low power of 0.30 and 0.32 for effect sizes of $\beta = 0.10$ and $\beta = 0.20$, respectively. The maximum power would be what is obtained from the full dataset which is roughly 0.55. Power is slightly higher but still low (0.34) under random missing scenarios. Since the approach was shown to be conservative with missing data, the power may also be affected since overall the procedure results in fewer rejections than expected.

Table 4.5: Comparison of the average estimates and biases for all the parameters estimated from the EM-LMM model and the Gaston method based on 1000 simulated datasets.

Parameter	True Value	EPS		GASTON		RM (5%)	
		$\bar{\boldsymbol{\theta}}_{EPS}$	Bias	$\bar{\boldsymbol{\theta}}_{\boldsymbol{G}}$	Bias	$\bar{\boldsymbol{\theta}}_{RM}$	Bias
β_0	1	0.9996	-0.0003	0.9980	-0.0001	1.0008	0.0008
β_g	0	0.0007	0.0007	-0.0003	-0.0003	0.0007	-0.0007
σ_b^2	0.4	4.9720	4.5720	0.3999	-0.0000	4.9720	4.5720
σ_ε^2	0.6	0.0177	-0.5822	0.6001	0.0001	0.0177	-0.5822
γ_0	0.3918	0.3923	0.0005	-	-	0.3922	0.0004
γ_1	0.3108	0.3150	0.0042	-	-	0.3149	0.0041
γ_2	0.2914	0.2927	0.0012	-	-	0.2927	0.0014

[a] The EM-LMM method was applied to both the EPS (20% in the extremes) and randomly missing(RM) 5% scenarios while the Gaston software was applied to the full dataset.

4.3.2 Results from the PROtEuS Study

Table 4.7 shows the parameter estimates from our proposed EM-LMM method on EPS with missing genotypes and the estimates from the full dataset estimated using Gaston and Lme4qtl. To obtain these estimates, we performed a single variant association test with our EM-LMM model for all the cases of missingness considered. To better understand performance, we used the "rs903924" SNP with a minor allele frequency 0.40 randomly selected from the list of SNPS after QC.

We observe that the estimates of fixed effects and variance components from our method with the EPS design are generally similar to what would be obtained by analyzing the full dataset using lme4qtl. The estimates from Gaston are also similar except for the estimate for the model error σ_e^2. This difference could be attributed to the different forms of GRM matrix used in fitting both models. Our EM approach, GEMMA and lme4qtl use the same approach for calculating the GRM while Gaston

Table 4.6: Estimation of power for all forms of missingness considered

Method	**Scenario**	**Estimated Power**	
		$\boldsymbol{\beta_g = 0.10}$	$\boldsymbol{\beta_g = 0.20}$
EM-LMM	EPS	0.30	0.32
	5% Missing	0.34	0.37
Gaston	Full dataset	0.43	0.56
	5% Missing	0.48	0.50
GEMMA	Full dataset	0.54	0.58
	5% Missing	0.48	0.49

[a] We compared power obtained from the EPS and 5% missing with the maximum power obtainable from the full dataset determined by GASTON and GEMMA.

uses a standardised GRM computed directly in the program. We also note that the regression parameter for the fixed effect, β_g is higher using the EM-LMM approach on the EPS data when compared to the estimates obtained by lme4qtl and Gaston on the full dataset.

Table 4.7: Fixed effect and variance effect estimates for EM-LMM analysis on the EPS dataset and Lme4qtl and Gaston on the full dataset for SNP rs903924

Parameter	EM-LMM	lme4qtl	GASTON
β_0	26.6675	27.6220	26.9490
β_g	0.4687	0.1092	0.151
σ_b^2	7.1236	8.356	7.0309
σ_e^2	7.6301	8.4838	10.7414

To assess whether the inflated β_g value is related to the EPS sampling design, we also considered randomly missing genotypes. Table 4.8 displays the estimated

parameters assuming an EPS study design and randomly missing scenarios. We observe that the estimates are generally close to each other except for the fixed effect parameter β_g. The β_g estimate for the randomly missing scenarios are close to those obtained by Gaston. It is therefore possible that the inflated β_g estimate in Table 4.7 is due to the EPS sampling rather than the EM-LMM algorithm.

Table 4.8: Parameter estimates from EM-LMM analysis for various missing data scenarios.

	Parameter Estimates			
Parameter	EPS	5%	10%	20%
β_0	26.6675	26.9176	27.0797	26.9537
β_g	0.4687	0.1873	0.1678	0.1465
σ_b^2	7.1236	7.1231	7.1229	7.1229
σ_e^2	7.6301	7.6142	7.6039	7.6172
γ_0	0.3238	0.3313	0.3373	0.3417
γ_1	0.4705	0.4777	0.4612	0.4559
γ_2	0.2057	0.2010	0.201	0.2022

In order to fully assess the performance of our model for each scenario of missing data, we tested for association of $50,000$ randomly chosen SNPs from the genomewide data. We plotted the observed p-values using a Q-Q plot (quantile-quantile plot). A Q-Q plot is a graphical tool used to assess whether the distribution of observed p-values in the association test deviates from the expected distribution under the null

hypothesis of no association.

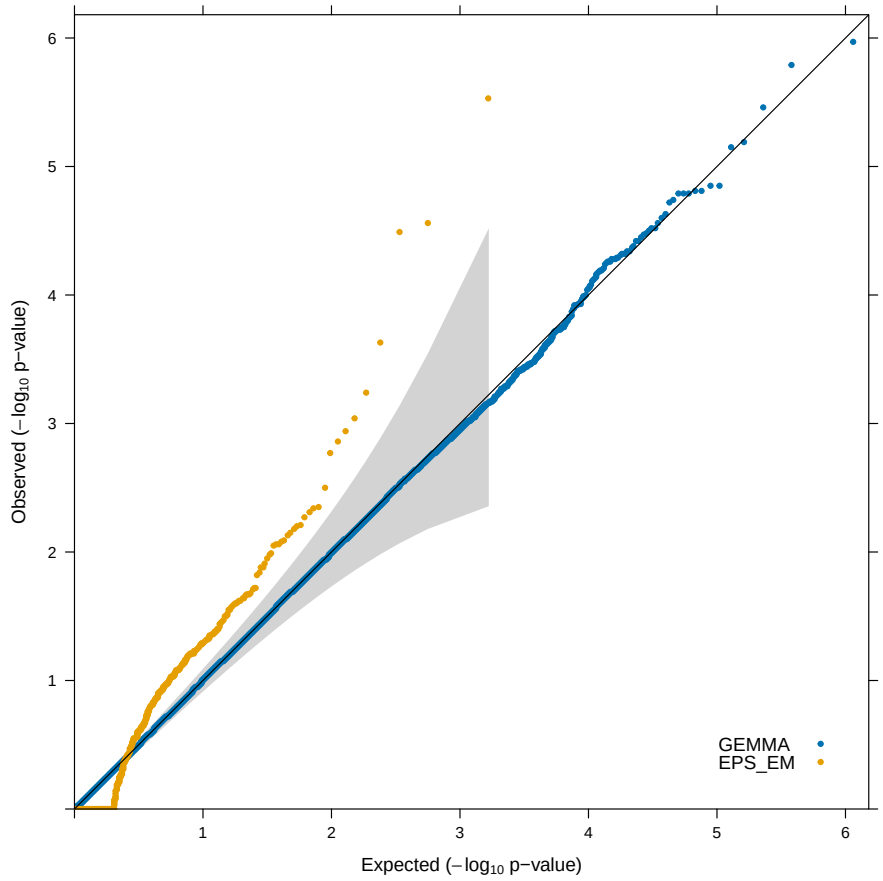


Figure 4.4: Q-Q Plot of p-values obtained from EM-LMM and GEMMA

Figure 4.4, shows the Q-Q plot plotted for the EPS scenario using our EM-algorithm. We also plotted the observed p-values from GEMMA. The EPS samples were obtained from retaining 20% of the top and bottom genotypes. However, given that we expect few SNPs to be significant, if the test statistics are not inflated, we would expect all points to fall on the line $x = y$. We obtained points that have larger

than normal values compared to the expected values (yellow points) which shows that the test statistics are massively inflated. The distribution of p-values for GEMMA is close to expected under the null and therefore we do not have to worry about an inflated false positive rate. We repeated the QQ plot with a 5% missing scenario and the Q-Q plot was similar.

4.4 Discussion

In this chapter, we applied the Expectation-Maximization (EM -LMM) algorithm to analyze simulated datasets and a real Prostate cancer dataset and compared results to existing LMM approaches.

Our results from the analysis of the PROtEuS dataset were unexpected given the simulation study. When looking at a single SNP, the genetic effect estimate was higher under the EPS design than the estimate obtained on the dataset without missing genotypes, but the variance components were similar. On the other hand, the estimates under randomly missing scenarios were all similar to those obtained with the full dataset. However, when looking at the distribution of P-values from the randomly missing scenario across 50,000 SNPs from the PROtEus study, we see massive inflation of the number of significant hypothesis tests. This result would indicate that the EM-LMM approach is not a valid analysis strategy. On the other hand, the simulation study showed that the EM-LMM approach is conservative, with estimated type 1 error rates lower than the nominal value. The approach becomes more conservative as the proportion of missing genotypes increases. However, we have shown that the LMM approach overall is conservative; our type 1 error estimates are close to those estimated using the LMM approach of Gaston on the full dataset. In

looking at power, we note that power under EPS is lower than randomly missing. This can be explained by the conservative nature of this procedure as well as the fact that under EPS, there are more missing data. Power was comparable to Gaston. Finally, the EM-LMM approach estimates the fixed effects with high accuracy. However the variance components are not estimated accurately. Further work is needed to determine the reason for the biases of the variance components.

The most likely explanation for the discrepancy of results between the simulated and real data analysed is an error in the real data analysis itself. We are currently investigating this hypothesis. However, if no errors are found, then our conclusion is that the EM-LMM approach to missing genotype data is not an appropriate analysis strategy. For EPS designs, this means that the data should be analysed as binary, as described in Chapter 2. For sporadically missing data, an imputation approach is preferred.

Chapter 5

Conclusion and Future Work

In this thesis, we have conducted an investigation into the use of Linear Mixed Models (LMMs) for handling the effects of population stratification in extreme phenotype samples (EPS) and we have proposed a novel LMM algorithm for the analysis of such data. Population stratification, arising from genetic ancestry differences among study participants, poses a significant challenge in genetic association studies, particularly when analyzing samples from an EPS design. Through our research, we have explored LMMs as a statistical tool for accounting for population structure in both extreme phenotype samples and when genotype data are sporadically missing. Our work highlights the importance of considering population stratification in genetic analyses involving EPS samples and demonstrates the use of LMMs as a tool to effectively controlling for these confounding factors. As we conclude this thesis, we summarize the main findings from our investigation, their implications and discuss potential directions for future research in this field.

In Chapter 2, we compared the type 1 error and power of several mixed model based approaches in correcting for population substructure when we have extreme

phenotype samples. We modelled the extreme phenotypes as binary and used methods suitable for analysing case-control or binary data. We simulated genetic data using the Balding Nichols model and normally distributed phenotypes for each subpopulation. We considered an extreme form of population stratification for this study by setting $p_1 = 0.25$ and $p_2 = 0.85$ in each subpopulation respectively. We also considered a real dataset. Our findings revealed adequate control of the type 1 error rate for most of the mixed-model based approaches when the genetic variant was common. We also considered rare variants using the available mixed-model based approaches. We observed inadequate correction for the type 1 error under this scenario.

Drawing upon the results in Chapter 2, in Chapter 3 we developed a novel EM algorithm in a linear mixed model framework that is applicable for missing genotype data. Our method is an extension of the method of weights by Ibrahim et al. [23] and we assume that only the genetic data at the candidate SNP are missing. We call this method EM-LMM. We also derived a likelihood ratio statistic and a Wald statistic for testing the hypothesis of no association; however, we focus on the Wald test in subsequent work. Our Wald test was based on the Monte Carlo approximation method of Tanner [69] as it was difficult to obtain analytical expressions for the standard errors using the Louis' method.

In Chapter 4, we evaluated our EM-LMM algorithm using simulation studies and a real data analysis from the PROtEUS study. Our investigation aimed to assess the algorithm's efficacy in terms of controlling the type 1 error, power and how well the fixed effects and variance components are estimated by the model. These are all crucial for accurately drawing inference about the model we have developed and its use in genetic association studies. Notably, our simulation studies demonstrated that the algorithm exhibited a conservative type 1 error rate and low power for all

forms of missing scenarios considered. Furthermore, the algorithm produced unbiased estimates for the fixed effects, suggesting its ability in capturing the true effect sizes of genetic variants on the phenotype of interest. However, the simulation study also uncovered a concerning aspect of the algorithm's performance. Specifically, we observed highly biased estimates for variance components. This indicates a systematic deviation from the true underlying variability within the dataset. This finding raises important considerations regarding the algorithm's suitability for applications where accurate estimation of variance components is critical.

For the real data analysis when we examined a single randomly chosen SNP, we didn't observe bias for the variance component estimates; they were all within an acceptable range when compared with other LMM methods applied to the full dataset. However, the fixed genetic effect estimate from EM-LMM was larger than the estimates from other methods. We also analysed a large subset of the genome wide data using the EM-LMM model when data is EPS-missing and randomly missing. Despite the rigorous testing of our algorithm, the results did not align closely with the expected values, as evidenced by the Quantile-Quantile (Q-Q) plot. While it is not clear if this discrepancy is from our Wald statistic or the number of SNPs that were analysed, the results suggests potential challenges or limitations associated with the use of EM-LMM when we have real genetic data. Identifying the underlying causes of this discrepancy is an area for further investigation.

Overall, the evaluation of our EM-LMM algorithm points out the importance of thoroughly assessing a proposed methodology in genetic association studies. While the simulation studies indicated that the algorithm demonstrates promise in controlling the type 1 error rate and estimating fixed effects accurately, the analysis of the ProtEus data pointed to a high false positive rate. At this time, to analyze EPS data,

we recommend using the methods for binary data evaluated in Chapter 2. Current and future research efforts should focus on addressing the algorithm's limitations to enhance its applicability in controlling for population stratification in extreme phenotype samples.

5.1 Limitations of the Study

While this study offers insights into the use of linear mixed models applied to missing genotype data, it is important to acknowledge its limitations.

First, our approach like that of Lin et al. [36] requires genotype data for ancestry estimation/GRM to be available on all samples. This would apply for example if a GWAS was first conducted on the full cohort, but targeted sequencing of selected regions was later only done on a subsample. However it is possible that in real studies genomic data is not available. One of the approaches we tried was the conditional likelihood approach used by a number of authors like Huang et al. [22], Li et al. [35] and Barnett et al. [4]. After expressing the likelihood as a truncated normal distribution taking into account the extreme sampling, we proceeded to use the method of Derkach et al. [14] to estimate an appropriate score test for the hypotheses of no association. Unfortunately, this could not be solved analytically and we therefore decided to consider the problem as a missing data problem so that the EM algorithm could be used.

In Chapter 3, we also developed a likelihood ratio test to evaluate the hypothesis of no association. However the literature on the likelihood ratio test with missing data is sparse because the observed likelihood, the marginal density of the observed part of the data, is an integral expression [82]. Drawing from Yang and Kim [82], who

proposed an approximation for the observed log likelihood, we employed importance sampling to develop an approximate likelihood ratio statistic for handling missing data at random. Despite implementing the likelihood ratio test, we encountered computational errors that we hypothesize are due to extremely small numbers during summation and division. Consequently, we explored alternative methods and our work in Chapter 4 is based on Wald tests. However, given the scarcity of literature on hypothesis testing on missing data, there are concerns surrounding its application for missing genetic data.

Another limitation of our approach is the computational time. First we do not have an inbuilt function for computing the GRM matrix. We had to rely on external approaches to compute the GRM. Since working with extremely high dimensional matrices requires implementation of numerical techniques, we had to rely on other software for GRM calculation which increases the computational time. Second, our solutions to the missing data rely on sampling many observations from the distributions defined by the weights. This greatly increases the time needed to analyse a single SNP. Since LMM-based analyses are already computationally demanding for genome level analyses, this limits our algorithm's utility to settings where cluster computing is available.

5.2 Ongoing and Future Work

We have focused on using common variants in this study with hopes that we would extend the algorithm to include rare variants, as our work in Chapter 2 showed a lack of suitable mixed model approaches for rare variants when data are from an EPS study. Typically, a rare variant is a genetic variant with minor allele frequency (MAF)

$< 1\%$. Due to their low frequency in the population, the ability to detect associations are often poorly captured by traditional genetic association approaches. Methods for tackling the problem of association testing for rare variants include Burden tests such as the Combined Multivariate and Collapsing method (CMC) [34], Cohort Allelic Sums test [47] and the Weighted Sum test [42]. These tests combine information from all rare variants within a gene and collapse them into a single genetic variable which is then tested for association with a phenotype of interest. The Sequence Kernel Association test [80] aggregates evidence of individual variant effects across the region using kernel function and uses a computationally efficient mixed model variance component test to test for association. In the context of EPS, very limited statistical methods have been developed for studying rare variants effects [4].

To extend our EM-LMM algorithm to accommodate rare variants, we proposed using the gene-based collapsing method of Povysil et al. [56]. In this approach, variants that satisfy a specific criteria (qualifying variants) are binned together as equivalent and each individual is categorized as either having the "qualifying variant" or not. This approach is easily accommodated in our algorithm as we would only need to recode the expression for computing the "weights" for the missing genotypes as allowing only 2 categories (0 = does not have "qualifying" variant, 1 = has a "qualifying" variant) rather than 3. However, after laying down the background work to extend the EM algorithm to rare variants, the outcomes observed upon applying the EM-LMM algorithm to common variants from the ProtEus study suggests a cautious approach when considering its application to an even more complex scenario like rare variant association. Hence, we would consider this extension as future work.

The present investigation into the use of LMMs under an EPS design has provided valuable insights into applying an EM algorithm to handle missing genotype data.

We will now outline some alternative approaches and potential directions for further inquiry and development.

A common approach used to handle missing genotypes in genetic association analyses has been to perform imputation to "fill-in" the missing genotypes. Single imputation of ungenotyped SNPs using resources like 1000 Genomes is already very common [48]. Multiple imputation, as opposed to single imputation, is a general statistical technique for handling missing data. The key idea here is to replace each missing value with a set of plausible values thereby creating multiple completed datasets. Each of these resulting completed datasets can then be analyzed using standard complete-data methods. Imputation has been used in methods for carrying out GWAS (e.g GEMMA [86]) as it offers a way to tackle the missing genotypes typically found in large genetic datasets even after quality control has been carried out. An area of future work would be to compare the multiple imputation approach to our EM-LMM algorithm especially for the real data and observe the method that offers a better correction of the type 1 error rate due to population stratification.

In conclusion, the application of the EM algorithm in the framework of linear mixed models represents a valuable approach for analyzing complex genetic data. While challenges were encountered in this study, the insights gained contribute to advancing our understanding of statistical methodologies for genetic association analysis and inform future research directions in this field.

Appendix A

Appendix A

A.1 Code for the EM-LMM algorithm

- The R code for implementing the EM-LMM algorithm are all available on the github page (https://github.com/monif064/EM-LMM-Analysis for the research work.

Bibliography

[1] Alan Agresti and Maria Kateri. Categorical data analysis. In *International encyclopedia of statistical science*, pages 206–208. Springer, 2011.

[2] David J Balding and Richard A Nichols. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica*, 96(1-2):3–12, 1995.

[3] David Ball, Linzy Hill, Thalia C Eley, Michael J Chorney, Karen Chorney, Lee A Thompson, Douglas K Detterman, Camilla Benbow, David Lubinski, Michael Owen, et al. Dopamine markers and general cognitive ability. *Neuroreport*, 9(2):347–349, 1998.

[4] Ian J Barnett, Seunggeun Lee, and Xihong Lin. Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. *Genetic epidemiology*, 37(2):142–151, 2013.

[5] Marcelo Bertalan. *gaston: Genetic Data Handling (QC, GRM, LD, PCA) & Linear Mixed Models*, 2022. R package version 1.5.6.

[6] Audrey Blanc-Lapierre, Andrea Spence, Pierre I Karakiewicz, Armen Aprikian, Fred Saad, and Marie-Élise Parent. Metabolic syndrome and prostate cancer

risk in a population-based case–control study in montreal, canada. *BMC public health*, 15:1–11, 2015.

[7] Audrey Blanc-Lapierre, Deborah Weiss, and Marie-Élise Parent. Use of oral anticoagulants and risk of prostate cancer: a population-based case–control study in montreal, canada. *Cancer Causes & Control*, 25:1159–1166, 2014.

[8] Catarina D Campbell, Elizabeth L Ogburn, Kathryn L Lunetta, Helen N Lyon, Matthew L Freedman, Leif C Groop, David Altshuler, Kristin G Ardlie, and Joel N Hirschhorn. Demonstrating stratification in a european american population. *Nature genetics*, 37(8):868, 2005.

[9] George Casella and Roger L Berger. *Statistical inference*. Cengage Learning, 2021.

[10] Han Chen, Chaolong Wang, Matthew P Conomos, Adrienne M Stilp, Zilin Li, Tamar Sofer, Adam A Szpiro, Wei Chen, John M Brehm, Juan C Celedón, et al. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics*, 98(4):653–666, 2016.

[11] Zehua Chen, Gang Zheng, Kaushik Ghosh, and Zhaohai Li. Linkage disequilibrium mapping of quantitative-trait loci by selective genotyping. *The American Journal of Human Genetics*, 77(4):661–669, 2005.

[12] A Darvasi and M Soller. Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. *Theoretical and applied Genetics*, 85(2-3):353–359, 1992.

[13] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[14] Andriy Derkach, Jerald F Lawless, and Lei Sun. Score tests for association under response-dependent sampling designs for expensive covariates. *Biometrika*, 102(4):988–994, 2015.

[15] B Devlin, Kathryn Roeder, and Larry Wasserman. Genomic control, a new approach to genetic-based association studies. *Theoretical population biology*, 60(3):155–166, 2001.

[16] Bernie Devlin and Kathryn Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, 1999.

[17] Mary J Emond, Tin Louie, Julia Emerson, Wei Zhao, Rasika A Mathias, Michael R Knowles, Fred A Wright, Mark J Rieder, Holly K Tabor, Deborah A Nickerson, et al. Exome sequencing of extreme phenotypes identifies dctn4 as a modifier of chronic pseudomonas aeruginosa infection in cystic fibrosis. *Nature genetics*, 44(8):886, 2012.

[18] Jakris Eu-Ahsunthornwattana, E Nancy Miller, Michaela Fakiola, Wellcome Trust Case Control Consortium 2, Selma MB Jeronimo, Jenefer M Blackwell, and Heather J Cordell. Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLoS genetics*, 10(7):e1004445, 2014.

[19] James E Gentle. *Computational statistics*. Springer, 2010.

[20] Lin T Guey, Jasmina Kravic, Olle Melander, Noël P Burtt, Jason M Laramie, Valeriya Lyssenko, Anna Jonsson, Eero Lindholm, Tiinamaija Tuomi, Bo Isomaa, et al. Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants. *Genetic epidemiology*, 35(4):236–246, 2011.

[21] Daniel L Hartl, Andrew G Clark, and Andrew G Clark. *Principles of population genetics*, volume 116. Sinauer associates Sunderland, 1997.

[22] BE Huang and Danyu Y Lin. Efficient association mapping of quantitative trait loci with selective genotyping. *The American Journal of Human Genetics*, 80(3):567–576, 2007.

[23] Joseph G Ibrahim. Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 85(411):765–769, 1990.

[24] Joseph G Ibrahim, Ming-Hui Chen, and Stuart R Lipsitz. Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika*, 88(2):551–564, 2001.

[25] Joseph G Ibrahim and Geert Molenberghs. Missing data methods in longitudinal studies: a review. *Test*, 18(1):1–43, 2009.

[26] Jiming Jiang and Thuan Nguyen. *Linear and generalized linear mixed models and their applications*, volume 1. Springer, 2007.

[27] Hyun Min Kang, Jae Hoon Sul, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, Eleazar Eskin, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4):348, 2010.

[28] Hyun Min Kang, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, 2008.

[29] William C Knowler, RC Williams, DJ Pettitt, and A Gm Steinberg. Gm3; 5, 13, 14 and type 2 diabetes mellitus: an association in american indians with genetic admixture. *American journal of human genetics*, 43(4):520, 1988.

[30] Gregory V Kryukov, Alexander Shpunt, John A Stamatoyannopoulos, and Shamil R Sunyaev. Power of deep, all-exon resequencing for discovery of human trait genes. *Proceedings of the National Academy of Sciences*, 106(10):3871–3876, 2009.

[31] Nan M Laird and Christoph Lange. *The fundamentals of modern statistical genetics*. Springer, 2011.

[32] Eric S Lander and David Botstein. Mapping mendelian factors underlying quantitative traits using rflp linkage maps. *Genetics*, 121(1):185–199, 1989.

[33] RJ Lebowitz, M Soller, and JS Beckmann. Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. *Theoretical and Applied Genetics*, 73:556–562, 1987.

[34] Bingshan Li and Suzanne M Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3):311–321, 2008.

[35] Dalin Li, Juan Pablo Lewinger, William J Gauderman, Cassandra Elizabeth Murcray, and David Conti. Using extreme phenotype sampling to identify the

rare causal variants of quantitative traits in association studies. *Genetic epidemiology*, 35(8):790–799, 2011.

[36] Dan-Yu Lin, Donglin Zeng, and Zheng-Zheng Tang. Quantitative trait analysis in sequencing studies under trait-dependent sampling. *Proceedings of the National Academy of Sciences*, 110(30):12247–12252, 2013.

[37] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. Fast linear mixed models for genome-wide association studies. *Nature methods*, 8(10):833–835, 2011.

[38] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[39] Dajiang J Liu and Suzanne M Leal. A flexible likelihood framework for detecting associations with secondary phenotypes in genetic studies using selected samples: application to sequence data. *European journal of human genetics*, 20(4):449–456, 2012.

[40] Po-Ru Loh, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjálmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, Bonnie Berger, et al. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature genetics*, 47(3):284–290, 2015.

[41] Thomas A Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 44(2):226–233, 1982.

[42] Bo Eskerod Madsen and Sharon R Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS genetics*, 5(2):e1000384, 2009.

[43] M Malosetti, CG Van der Linden, Bl Vosman, and FA Van Eeuwijk. A mixed-model approach to association mapping using pedigree information with an illustration of resistance to phytophthora infestans in potato. *Genetics*, 175(2):879–889, 2007.

[44] Andries T Marees, Hilde de Kluiver, Sven Stringer, Florence Vorspan, Emmanuel Curis, Cynthia Marie-Claire, and Eske M Derks. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International journal of methods in psychiatric research*, 27(2):e1608, 2018.

[45] Paolo Menozzi, Alberto Piazza, and L Cavalli-Sforza. Synthetic maps of human gene frequencies in europeans. *Science*, 201(4358):786–792, 1978.

[46] Hannah Verena Meyer and Ewan Birney. Phenotypesimulator: A comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships. *Bioinformatics*, 34(17):2951–2956, 2018.

[47] Stephan Morgenthaler and William G Thilly. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (cast). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 615(1-2):28–56, 2007.

[48] Adam C Naj. Genotype imputation in genome-wide association studies. *Current Protocols in Human Genetics*, 102(1):e84, 2019.

[49] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, et al. Genes mirror geography within europe. *Nature*, 456(7218):98, 2008.

[50] Maryam Onifade, Marie-Hélène Roy-Gagnon, Marie-Élise Parent, and Kelly M Burkett. Comparison of mixed model based approaches for correcting for population substructure with application to extreme phenotype sampling. *BMC genomics*, 23(1):98, 2022.

[51] Paul F O'Reilly, Clive J Hoggart, Yotsawat Pomyen, Federico CF Calboli, Paul Elliott, Marjo-Riitta Jarvelin, and Lachlan JM Coin. Multiphen: joint model of multiple phenotypes can increase discovery in gwas. *PloS one*, 7(5):e34861, 2012.

[52] Michela Panarella and Kelly M Burkett. A cautionary note on the effects of population stratification under an extreme phenotype sampling design. *Frontiers in genetics*, 10:398, 2019.

[53] Marie-Élise Parent, Mark S. Goldberg, Dan L. Crouse, Nancy A. Ross, Hong Chen, Marie-France Valois, and Alexandre Liautaud. Traffic-related air pollution and prostate cancer risk: a case-control study in Montreal, Canada. *Occupational and Environmental Medicine*, 70(7):511–518, July 2013.

[54] Nick Patterson, Alkes L Price, and David Reich. Population structure and eigenanalysis. *PLoS genetics*, 2(12):e190, 2006.

[55] Heather F Porter and Paul F O'Reilly. Multivariate simulation framework reveals performance of multi-trait gwas methods. *Scientific reports*, 7(1):38837, 2017.

[56] Gundula Povysil, Slavé Petrovski, Joseph Hostyk, Vimla Aggarwal, Andrew S Allen, and David B Goldstein. Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nature Reviews Genetics*, 20(12):747–759, 2019.

[57] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904, 2006.

[58] Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459, 2010.

[59] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[60] Jonathan K Pritchard, Matthew Stephens, Noah A Rosenberg, and Peter Donnelly. Association mapping in structured populations. *The American Journal of Human Genetics*, 67(1):170–181, 2000.

[61] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3):559–575, 2007.

[62] Donald B Rubin. Wiley series in probability and mathematical statistics. applied probability and statistics, 1987.

[63] Shayle R Searle and Andre I Khuri. *Matrix algebra useful for statistics*. John Wiley & Sons, 2017.

[64] Montgomery Slatkin. Disequilibrium mapping of a quantitative-trait locus in an expanding population. *The American Journal of Human Genetics*, 64(6):1765–1773, 1999.

[65] Doug Speed and David J Balding. Relatedness in the post-genomic era: is it still useful? *Nature Reviews Genetics*, 16(1):33–44, 2015.

[66] Amy L Stubbendick and Joseph G Ibrahim. Maximum likelihood methods for nonignorable missing responses and covariates in random effects models. *Biometrics*, 59(4):1140–1150, 2003.

[67] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[68] Yun Ju Sung and Charles J Geyer. Monte carlo likelihood inference for missing data models. 2007.

[69] Martin A Tanner. *Tools for statistical inference*, volume 3. Springer, 1993.

[70] Martin A Tanner. *Tools for statistical inference: observed data and data augmentation methods*, volume 67. Springer Science & Business Media, 2012.

[71] R Core Team et al. R: A language and environment for statistical computing. 2013.

[72] Terry Therneau. The lmekin function. *Rochester, MN: Mayo Clinic*, 2012.

[73] Karine Trudeau, Marie-Claude Rousseau, and Marie-Élise Parent. Extent of food processing and risk of prostate cancer: The proteus study in montreal, canada. *Nutrients*, 12(3):637, 2020.

[74] Sofie Van Gestel, Jeanine J Houwing-Duistermaat, Rolf Adolfsson, Cornelia M van Duijn, and Christine Van Broeckhoven. Power of selective genotyping in genetic association analyses of quantitative traits. *Behavior genetics*, 30(2):141–146, 2000.

[75] Jorie Versmissen, Daniëlla M Oosterveer, Mojgan Yazdanpanah, Abbas Dehghan, Hilma Hólm, Jeanette Erdman, Yurii S Aulchenko, Gudmar Thorleifsson, Heribert Schunkert, Roeland Huijgen, et al. Identifying genetic risk variants for coronary heart disease in familial hypercholesterolemia: an extreme genetics approach. *European Journal of Human Genetics*, 23(3):381–387, 2015.

[76] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

[77] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

[78] J Dylan Weissenkampen, Yu Jiang, Scott Eckert, Bibo Jiang, Bingshan Li, and Dajiang J Liu. Methods for the analysis and interpretation for rare variants associated with complex traits. *Current protocols in human genetics*, 101(1):e83, 2019.

[79] Chengqing Wu, Andrew DeWan, Josephine Hoh, and Zuoheng Wang. A comparison of association methods correcting for population stratification in case–control studies. *Annals of human genetics*, 75(3):418–427, 2011.

[80] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011.

[81] Jian Yang, Noah A Zaitlen, Michael E Goddard, Peter M Visscher, and Alkes L Price. Advantages and pitfalls in the application of mixed-model association methods. *Nature genetics*, 46(2):100, 2014.

[82] Shu Yang and Jae Kwang Kim. Likelihood-based inference with missing data under missing-at-random. *Scandinavian Journal of Statistics*, 43(2):436–454, 2016.

[83] Jianming Yu, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, Brandon S Gaut, Dahlia M Nielsen, James B Holland, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*, 38(2):203–208, 2006.

[84] Zhiwu Zhang, Elhan Ersoz, Chao-Qiang Lai, Rory J Todhunter, Hemant K Tiwari, Michael A Gore, Peter J Bradbury, Jianming Yu, Donna K Arnett, Jose M Ordovas, et al. Mixed linear model approach adapted for genome-wide association studies. *Nature genetics*, 42(4):355, 2010.

[85] Keyan Zhao, María José Aranzana, Sung Kim, Clare Lister, Chikako Shindo,

Chunlao Tang, Christopher Toomajian, Honggang Zheng, Caroline Dean, Paul Marjoram, et al. An arabidopsis example of association mapping in structured samples. *PLoS genetics*, 3(1):e4, 2007.

[86] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821–824, 2012.

[87] Andrey Ziyatdinov, Miquel Vázquez-Santiago, Helena Brunel, Angel Martinez-Perez, Hugues Aschard, and Jose Manuel Soria. lme4qtl: linear mixed models with flexible covariance structure for genetic studies of related individuals. *BMC bioinformatics*, 19(1):1–5, 2018.

[88] Or Zuk, Eliana Hechter, Shamil R Sunyaev, and Eric S Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, 109(4):1193–1198, 2012.