Statistical models of gene-environment interactions

Matthew Kerin

Keble College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity Term, 2019

Acknowledgements

Firstly, I would like to thank Kevin Sharp and Jonathan Marchini for their guidance over the past four years. I am particularly grateful to Jonathan Marchini, who was always available to answer questions and provide insight into problems as they arose. Without him, this thesis would not exist. I thank David Steinsaltz for his supervision in my final year.

My time at the University of Oxford has been both a priviledge and a joy, and I owe a great debt of gratitude to the Wellcome Trust for their generous financial support and to the Genomic Medicine and Statistics programme for providing such a stimulating environment.

I thank my family and friends for their unwavering support in both sporting and academic life. Last but not least I thank my partner Sarah, for having the patience to put up with me.

Abstract

Despite longstanding interest in gene-by-environment (GxE) interactions, the contribution of GxE interactions to complex disease in humans remains poorly characterised. Studies of GxE interaction are statistically challenging due to the inherent high dimensionality of environmental exposures, the increased sample size required to robustly detect GxE effects and the historic lack of large datasets that combine genetic and environmental data. In this thesis I introduce a new approach for GxE interactions with multiple environmental variables on biobank scale datasets called LEMMA (Linear Environment Mixed Model Analysis). LEMMA uses a Bayesian whole genome approach to jointly model SNP marginal effects and SNP GxE interaction effects with an Environmental Score (ES). The ES is a linear combination of multiple environmental variables, and 'interaction weights' used to construct the ES are learnt in the same Bayesian framework. The ES can subsequently be interpreted directly, used in single SNP hypothesis or used for heritability estimation. Using variational inference to fir the model and a distributed computational architecture allows LEMMA to be computationally tractable on Biobank scale datasets.

LEMMA is compared to existing single-SNP approaches that jointly model the effects of multiple environmental variables at each locus. Results suggest that LEMMA has substantially greater power to detect GxE effects when the underlying model assumptions hold.

I have applied LEMMA to model GxE interactions in body mass index, systolic, diastolic and pulse pressure, using 42 environmental variables and approximately $280,000$ white British individuals from the UK Biobank. LEMMA uncovered 3 loci with GxE interactions at genome wide significance levels, and estimated that 9.3%, 3.9%, 1.6% and 12.5% of phenotypic variance is explained by GxE interactions, and that rare variants explain most of this variance.

In my final chapter I suggest a new model, referred to as RHE-NLS, that promises to achieve many of the same objectives as LEMMA with improved computational complexity.

Contents

List of Figures

List of Tables

1 Introduction

Contents

1.1 Background

1.1.1 Basic concepts

Deoxyribonucleic acid, or DNA, is the molecular storage unit used to pass hereditary information from one generation to the next by humans, or indeed by all animals. Most cells in the human body contain almost identical copies of the same DNA, which is transcribed to single-stranded messenger ribonucleic acid (mRNA) that is, in turn, translated into proteins. Proteins are sometimes called building blocks of the cell; patterns of protein expression govern cell type and behaviour.

The genetic information stored in DNA is encoded in sequences of four primary nucleobases; adenine (A), cytosine (C), guanine (G) and thymine (T). DNA typically exists in a double helix structure, with two strands of nucleotides that coil around each other and are joined by hydrogen bonds. Each nucleotide consists of a phosphate, a sugar molecule and a nucleobase. The nucleobases bind together according to base pairing rules due to hydrogen bonding, such that A bonds with T and C bonds with G and the strands in the double helix are complementary to one another. The human genome, the complete set of genetic material contained within a single cell, is estimated to contain approximately 3.3 billion base pairs. DNA is packaged into condensed structures called chromosomes. Humans possess 46 chromosomes organised into 23 pairs; 22 pairs of autosomes and one pair of sex-determining chromosomes (allosomes). Human cells are either haploid, with a single copy of each chromosome, such as egg or sperm cells, or diploid, meaning that they carry two homologous (identical in structure but small differences in DNA) copies of each autosome and two allosomes.

In humans, there are two key processes that introduce variation into the genome; mutation and recombination. Mutations are caused by errors in the cells' machinery during DNA replication or DNA repair. Mutations vary drastically in scale, from point mutations at a single nucleotide to insertion, deletion or repeated copying of regions of the genome. Genetic studies usually focus on point mutations, or single nucleotide polymorphisms (SNPs). Some mutations can be inherited; if a mutation occurs in a germline cell (progenitor of sperm or egg cell), it can be passed onto offspring in the next generation.

Recombination is the process during which homologous chromosomes physically exchange genetic material. Through complicated cellular machinery, matching sections are broken off from each chromosome, swapped and then reattached. Recombination occurs during meiosis; when a diploid cell divides into four haploid gametes each with a single copy of each chromosome. In sexually reproducing species, offspring inherit one combination from each parent and, due to recombination, each chromosome is a mosaic of the DNA inherited from the respective grandparents.

The International HapMap project[1] was a landmark study that published a catalogue of common SNPs present in the human genome and an estimate of the correlation structure between them. Because the breakpoints at which recombination occurs are highly conserved, the correlation structure of the human genome typically displays a block-diagonal-like structure. Blocks are known as linkage disequilibrium (LD) blocks and are separated by recombination hotspots, of which there are estimated to be approximately 20,000[2]. Although each LD block might cover many thousands of bases and thousands of commonly occurring SNPs, there are frequently only a small number of combinations of those SNPs (haplotypes) observed in a population. Thus researchers were able to identify a relatively small set of 500,000 'tag' SNPs that were representative of the majority of genetic variation from 10 million common SNPs [3, 4]. In association testing these 'tag' SNPs could be used to identify haplotypes that contained a causal SNP. Alternatively, comparing those same SNPs to reference panels of known haplotypes enabled researchers to impute the missing variants [5]. In subsequent genetic studies, much larger datasets could be assembled because individuals could be genotyped at a relatively small number of sites instead of using whole genome sequencing methods. In this thesis, the phrase 'genotyped SNPs' is indicative of SNPs that tag common variation in the genome.

1.1.2 Genome-wide association studies

Since 2007, the genome-wide association study (GWAS) has been one of the dominant statistical approaches to studying the genetic basis of complex disease[6, 7]. The GWAS is an agnostic statistical analysis that performs a univariate hypothesis test at each variant, from a set of variants that cover the genome, to assess if there is statistically significant correlation between carriers of the mutation and disease burden. To control for multiple hypothesis testing, a threshold of $p < 5 \times 10^{-8}$ is typically used to denote statistical significant[8, 9], which is roughly equivalent to a p-value threshold of 0.05 adjusted for the number of independent (common) variants in the genome. Crucial to the success of GWAS was the International HapMap project and the availability of cheap genotyping arrays. Genotyping arrays

offered a cheap way to detect the alleles present at up to 500,000 specific sites per individual, whilst researchers were able to choose sites that would capture most variation due to common SNPs first using the HapMap project[1] and subsequently the 1000 Genomes Project[10–12].

The first GWAS was a study of age-related macular degeneration using 96 cases and 50 controls, published in 2005[13]. By 2007, over 200 studies had been published including the Welcome Trust Case Control Consortium[7] which was the largest GWAS at the time (14,000) people and was awarded 'paper of the year' by the Lancet [14].

One of the main difficulties with GWASs can be controlling for confounding factors. Confounding occurs in a statistical analysis when a covariate C affects both the dependent variable X and the independent variable Y. Unless the effect of C is controlled for, the association between X and Y will be spuriously inflated. In GWASs, confounding can occur when related individuals have similar phenotype due to shared environmental exposures (family structure) or when two subpopulations with different genetic ancestry also have a mean difference in the phenotype (population structure) [15].

The two most common approaches to GWASs are linear regression and linear mixed models (LMMs). I define the dosage of a SNP as the (expected) number of alternative alleles present at a given site. Let y be a disease phenotype, x_{test} the dosage of the focal SNP to be tested, C a matrix of covariates and $\epsilon \sim N(0, \sigma_\epsilon^2 I)$ the residual variance. Both approaches test the hypothesis $\beta_{\text{test}} \neq 0$.

The basic model used in linear regression is given by

$$y = C\alpha + x_{\text{test}}\beta_{\text{test}} + \epsilon.$$

Population structure is usually controlled for by including the top genetic principal components (PCs) as covariates [16]. Related individuals are typically excluded from the analysis to control for family structure. However, this approach is imperfect and is still vulnerable to cryptic relatedness; a form of confounding due the sample structure between distantly related individuals with no known family structure[17].

LMMs have been used for quantitative genetics by animal breeders for decades [18]. The basic model used in LMMs is given by

$$y \sim \mathcal{N}(C\alpha + x_{\text{test}}\beta_{\text{test}}, \sigma_g^2 K + \sigma_e^2 I), \tag{1.1}$$

where the kinship matrix K models relatedness between samples. In animal breeding studies, the pedigree of all samples is known so K can simply be filled in with this information. In large human studies however, the pedigree is generally unknown. Thus the kinship matrix can be estimated from genotype data[19]. Specification of the kinship matrix K is still an area of active research [20–23] but the most simple approach is to let $K = XX^T/M$ where X is the $N \times M$ genotype matrix (where N is the number of samples and M is the number of SNPs) and columns of X are normalised to have mean zero and variance one.

LMMs are able to capture the effects of both family relatedness and population structure with the kinship matrix [24], so they have no need to exclude related samples. In addition they have the property that they have higher power than univariate linear regression in heritable traits, because they are able to implicitly condition on other causal SNPs in the genotype matrix X [25, 26].

1.1.3 Missing heritability

Heritability is defined as the proportion of trait variance in a population that is attributable to genetic effects [27]. Typically a phenotype (P) can be decomposed into contribution from genetic effects (G), from environmental effects (E) and interactions between the two (GxE).

$$P = G + E + G\text{x}E.$$

Assuming that $\text{Cov}(G, G\text{x}E) = 0$, $\text{Cov}(E, G\text{x}E) = 0$, and $\text{Cov}(G, E) = 0$, the phenotypic variance can then be written as

$$\text{Var}(P) = \text{Var}(G) + \text{Var}(E) + \text{Var}(G\text{x}E).$$

Historically the contribution of GxE effects to heritability has been ignored, partly because methods to estimate this component of variance have been lacking

[27]. 'Broad sense' heritability is defined as the proportion of trait variance explained by the total genetic variance ($H_g^2 = \frac{\text{Var}(G)}{\text{Var}(P)}$) whereas 'narrow sense' heritability is the proportion of trait variance explained by additive genetic effects ($h_g^2 = \frac{\text{Var}(\text{Additive}-G)}{\text{Var}(P)}$)).

Before GWASs started to allow geneticists to interrogate the genetic architecture of complex traits, it was expected that complex traits would be driven by a handful of moderate-effect loci[28]. By 2009, GWASs had been successful in identifying hundreds of robust associations between variants and complex human diseases[29]. However, it became clear that for many of these diseases, the sum of the independent effects estimated from GWASs accounted for a very small proportion of the total heritability estimated from family or twin studies. This became known as the problem of missing heritability [29, 30].

Multiple explanations for the missing heritability problem have been proposed[31]. Manolio et al. [29] argued that missing heritability might lie amongst the effects of rare or structural variants that are not well tagged by genotyping arrays, or amongst many small effects of common SNPs that GWASs at the time were not powered to discover. The latter was motivated by the observation that selection may act to keep effect sizes low, as variants of large effect are selected against[32]. Other explanations focussed on the heritability estimates of twin studies; as gene-gene iteractions[33], gene by environment interactions[34] or violations of the assumptions about shared environments[35] might caused heritability estimates from twin studies to be inflated.

A partial answer was provided in a seminal paper by Yang et al. [24]. The authors showed that the majority of missing heritability for height could be explained by genetic variation by common SNPs. Because LMMs are able to capture effects from population structure, family relatedness and causal SNPs in the kinship matrix[25], they observed that in a cohort of unrelated individuals with the same ancestry the kinship matrix will only capture effects from causal SNPs. Thus heritability could be estimated with the formula

$$\hat{h}_g^2 = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_e^2}$$

where $(\hat{\sigma}_g^2, \hat{\sigma}_e^2)$ are restricted maximum likelihood estimates from the LMM

$$y \sim \mathcal{N}(0, \sigma_g^2 K + \sigma_e^2 I), \tag{1.2}$$

where $K = XX^T/M$ and X is the normalised genotype matrix with column mean zero and column variance one.

Some of the implicit assumptions in this model encountered criticism[36]; in particular, the assumption that SNP effect sizes have a gaussian distribution, the implicit assumption that rare SNPs have larger effects or the assumption that SNP effect sizes are independent of linkage disequilibrium. However, the concept of heritability estimation in humans has been an active area of research ever since.

1.2 Gene by environment interactions

Despite longstanding interest in gene-by-environment (GxE) interactions[37, 38], the contribution of GxE interactions to complex disease in humans remains poorly characterised.

The term interaction has different meanings depending on the context[38]. A biological interaction can refer to two interacting factors in the body. A statistical GxE interaction is typically defined[39] as the joint effect of genetics and environment that causes a departure from the additive model

$$\mathbb{E}\left[Y|G, E\right] = a_1(G) + a_2(E).$$

In this thesis I focus on multiplicative interactions with a quantitative trait. That is, on interactions that take the form

$$Y = \beta_0 + G\beta_G + E\beta_E + G \times E\beta_{G\times E} + \epsilon.$$

Better characterisation of the contribution of GxE interactions to human genetics has many possible benefits. Identification of GxE interactions could offer greater insight into the biological pathways underlying complex traits[37, 38]. Alternatively, it could allow improved public health policy if subpopulations have significantly elevated risk of exposure to a given environmental hazard. From a genetic prediction

perspective, it could allow insight into the portability of genetic risk scores within and between ancestry groups [40]. Finally, if GxE effects do make a significant contribution to phenotypic variance then methods that better characterise GxE effects could inform the debate on missing heritability in complex traits [37, 41].

1.2.1 Challenges faced by GxE interaction studies

Studies of GxE interactions face several challenges over and above the usual challenges faced by GWASs of additive genetic effects. These challenges include problems of sample size, confounding and lack of exposure variation.

A commonly cited rule of thumb in epidemiology suggests that detection of interaction effects requires a sample size at least four times larger than that required to detect a main effect of comparable effect size [42]. This particular rule of thumb is based on a case control study and a single dichotomous environmental variable, however the general observation that GxE studies require larger sample sizes has been made many times[38, 43–45]. Higher sample sizes are required for two principal reasons; to offset the multiple testing burden imposed by the multiple plausible environmental exposures that could have interaction effects and to capture enough variation along each environmental exposure.

Confounding is another issue for which GxE studies must have additional care. Consider a phenotype y, variant x, environment E and confounder C. If C is confounded with G (ie it has causal effects on G and y) then unless the interaction between C and E is controlled for, the interaction between G and E may be biased [46]. In particular, this refers to population structure which is often controlled for using genetic PCs or LMMs in a conventional GWAS. In the case of adjustment by PCs, Keller [47] shows analytically that the $PC \times E$ should be controlled for in order to avoid confounding in the interaction effect. For LMMs, Sul et al. [48] observed empirically that the usual single component LMM failed to control for confounding in the GxE interaction test when the mean environmental exposure differed between two subpopulations for a dichotomous environmental variable.

In simulation, the authors further show that this form of confounding can be controlled for using the LMM given by

$$y \sim \mathcal{N}(C\alpha + x_{\text{test}}\beta_{\text{test}}, \sigma_g^2 K + \sigma_{GxE}^2 K^D + \sigma_e^2 I)$$

where $K_{ij}^D = K_{ij}$ if individuals i and j have the same level of exposure or $K_{ij}^D = 0$ otherwise. Whilst this model is simple to fit for a single categoric environment, a continuous environment would have to be binned according to some threshold(s) (as performed by Robinson et al. [49]). The MTG2 software package allows extensions to multiple environmental variables, and can accommodate either continuous or discrete variables [50, 51]. Finally, in the same way that genetic principal components (PCs) are analogous to controlling for population structure with a low dimensional representation of the kinship matrix, the PC×E terms are analogous to controlling for population structure with a low dimensional representation of K^D.

Lack of exposure variation is another important challenge that GxE studies can face [52, 53]. A compelling, and frequently used, example is provided by that of GxE interaction between physical activity (PA) and the FTO locus for body mass index. This GxE interaction was reported by several single cohort studies, one of which reported a statistically significant interaction (p=0.008) using a relatively small cohort ($N = 1,129$) of Indian ethnicity, but included individuals with a wide range of physical activity levels (from sedentary city dwellers to active farmworkers) [54]. On the other hand, a qualitatively similar interaction was replicated in a meta-analysis of Caucasian populations. However, due to lower variation in physical activity levels, a much larger sample size was required ($N > 200,000$) to achieve statistical significance ($\text{p} = 0.001$).

Jointly modelling multiple environmental variables is an alternative approach that may boost power, as a combination of multiple variables could act as a proxy for the true (unobserved) driver of interaction [55]. From the perspective of exposure variation, such a proxy would also have higher variability. Continuing with the previous example of Gene x PA in BMI; physical activity is typically measured using questionnaire data on the frequency and duration of exercise at different intensity

levels. However, other measures such as sedentary time, hours spent watching TV and diet can also act as proxies for an obesogenic environment [56]. In GxE study of BMI in the UK Biobank using $N = 89,552$ samples, Young, Wauthier, and Donnelly [57] reported GxE interactions at FTO with multiple environments including physical activity ($p = 3.1 \times 10^{-4}$), alcohol frequency ($p = 3.0 \times 10^{-4}$) and diet ($p = 5.0 \times 10^{-6}$), where each environmental was modelled separately. Using $N = 252,188$ samples, a similar set of environmental variables also from the UK Biobank and a novel method that jointly modelled GxE interaction from multiple environmental variables, Moore et al. [55] replicated GxE interaction at the FTO locus in BMI with much stronger statistical evidence ($p = 4.4 \times 10^{-16}$).

Using a conceptually similar framework, Kraft and Aschard [53] argue that testing for GxE interactions against a multi-SNP genetic risk score could have higher power to detect GxE interactions than tests against a single SNP because the multi-SNP approach captures a greater range of genetic variation. They acknowledge that this approach implicitly assumes that most SNPs in the genetic risk score interact with the same environmental exposure. This idea is present in several relatively recent studies that assess GxE interaction using a linear predictor of BMI-associated SNPs [58] or variance component tests [49, 55]. However, in each of these cases only a single environmental variable was used.

Finally, GxE studies are particularly vulnerable to scale effects. Consider for example a monogenic phenotype $Y = E\alpha + G\beta + \epsilon$, where Y, E, G, ϵ are all $N \times 1$ vectors denoting the phenotype, environment, genetic effect at a causal SNP and residual noise respectively and α and β are non-zero fixed effects. Then applying a non-linear transformation, say measuring Y^2, could induce spurious correlation between the phenotype Y and the environment E.

1.2.2 Current approaches

Tests for GxE effects have traditionally been performed using single-SNP, single-environment approaches. Testing all combinations of SNPs and plausible environments is both computationally and statistically burdensome, so a variety of

approaches have been employed to narrow down the search space. These approaches include hypothesis driven studies, 'two-step' approaches or, more recently, variance quantitative trait loci (QTL) methods. Separate from this, there has been growing interest in methods that model either SNPs or environmental variables jointly.

Many of the current GxE success stories have been identified through hypothesis driven (or candidate gene) studies, particularly those focusing on metabolic genes [59]. Examples include the interaction between variation in the *NAT2* gene and smoking for bladder cancer[60], or variation in the *ALDH2* gene and alcohol intake on esophageal cancer[61]. These approaches benefited from an extremely low multiple testing burden, but required strong prior knowledge regarding which genes and which environmental variables might be relevant. More agnostic methods will be required in order to more broadly characterise the total contribution of GxE effects to complex disease. Further, the fact that most GWAS hits for complex traits are located in non-coding regions[28] suggests that studies of GxE should also consider non-coding regions.

Several 'two-step' approaches have been proposed[62–64] that attempt to improve the efficiency of GWASs of GxE effects by reducing the number of GxE tests performed. These methods typically perform a normal GWAS scan for SNP main effects, and prioritise SNPs for testing GxE effects based on some p-value threshold from the main effects tests. The validity of this filtering approach depends on the independence of the main and GxE tests; if the two tests are dependent then this procedure will cause the GxE test statistics to be inflated. Consequently the number of tests performed, and thus the multiple testing correction, is reduced. This approach can be more powerful than performing an exhaustive scan, but is likely to miss any GxE effects with weak marginal effects[65].

A separate class of methods attempt to detect GxE effects by testing for variance quantitative trait loci (vQTLs). These methods leverage the insight GxE effects cause the phenotypic variance to change as a function of the genotype [66, 67]. This effect is known as conditional heteroskedasticity, and is shown in Figure 1.1. Several recent studies have applied these methods on biobank scale data[68, 69]. One

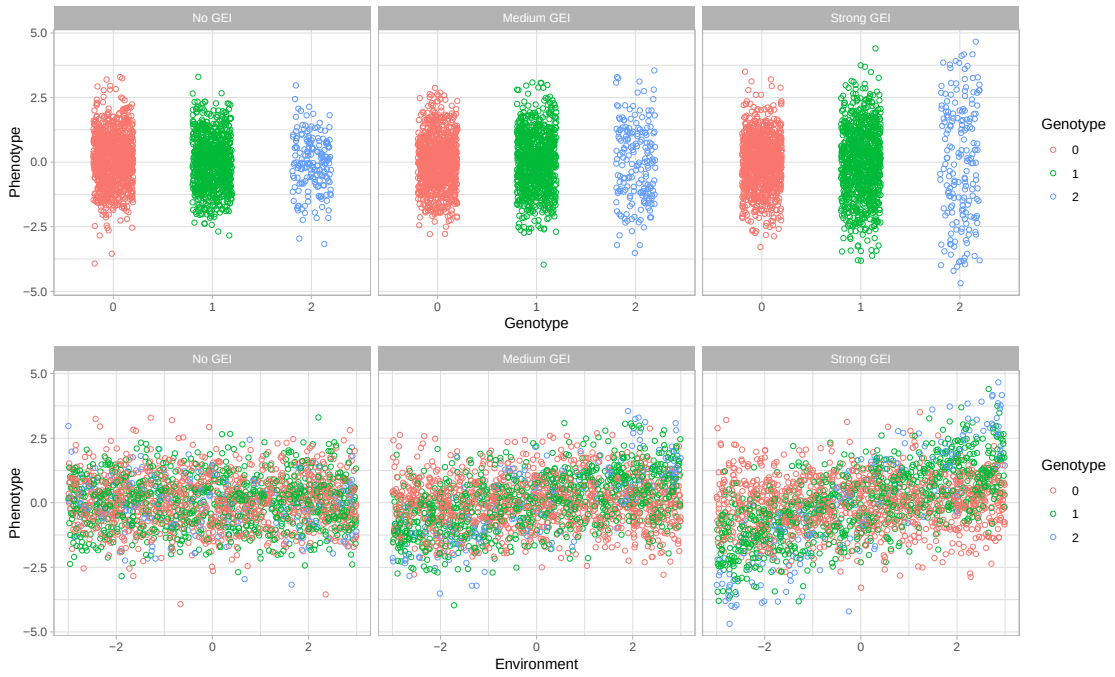

Figure 1.1: Conditional heteroskedasticity caused by a multiplicative GxE interaction. Differences in phenotypic variance by genotype group (top) and by strength of the environmental exposure (bottom). The phenotype was simulated using 2000 individuals on the basis of a multiplication interaction between a single genotype (minor allele frequency 0.3) and an environment (uniformly distributed over $[-2, 2]$). From left to right; the GxE interaction explained $0\%, 20\%, 40\%$ of trait variance.

advantage of these methods is that they do not require any data on environmental exposure, but are consequently less interpretable because they do not identify the interacting environment.

1.3 Methods

In this section, I provide a brief review of two methods commonly used in additive genetics that I build upon later on in my thesis.

1.3.1 Bayesian whole genome regression

The LMM in Equation (1.1) is mathematically equivalent to the Bayesian whole genome regression (WGR) model, given by

$$y = C\alpha + X\beta + \epsilon,$$
$$\beta \sim \mathcal{N}(0, \sigma_g^2).$$

Here β is a M-vector modelling the random effect of each SNP. This form corresponds to the so called 'infinitesimal model', where every SNP affects the phenotype, SNP effects are normally distributed, and SNP effects have identical variance. These are implicit assumptions encoded in the usual LMM[36]. Bayesian WGR models allow a generalisation of the model, using non-gaussian priors to model non-infinitesimal genetic architecture. This flexibility comes at some computational cost, as typically the effect sizes can no longer be analytically integrated out. Therefore Bayesian WGR methods must infer a posterior on β, using either variational inference or MCMC based methods.

This approach has been applied previously in human genetics [70–75] and by the 'Bayesian alphabet' of genomic prediction methods in the animal breeding literature [76–78].

The current state of the art WGR model, at least for the purposes of GWAS of quantitative traits, is BOLT-LMM[75, 79]. BOLT-LMM fits a whole genome regression model on genotyped SNPs using a mixture of gaussians prior on SNP effects. Inference of the latent variables is performed using variational inference, conditional on hyperparameters that are trained using cross validation. Loh et al. [75] show empirically that the algorithm runs in $\mathcal{O}(N^{1.5}M)$ time, for N samples and M genotyped SNPs. After convergence, the authors perform single SNP hypothesis testing on a much larger set of imputed SNPs against a residualised phenotype, which they show empirically to be well calibrated. They show that by using a mixture of gaussians prior instead of the implicit gaussian prior used by a standard LMM, BOLT-LMM has increased power to detect associated loci in many complex traits.

1.3.2 Haseman-Elston regression

The first of a growing family of approaches that measure heritability using genotype array data was 'genomic relationship matrix-REML' (GREML)[24], which assigned contribution of all SNPs to a single kinship matrix (as in Equation (1.2)). This is commonly referred to in the literature as the single component model. Subsequent research has shown that the single component model makes assumptions about

the relationship between MAF, LD and trait architecture that may not hold up in practice[36, 80]. There have been many follow up methods, including; generalisations that stratify variance into different components by MAF and LD [81], approaches that assign different weights for the kinship matrix [21, 36], methods that replace the gaussian prior with a spike and slab on SNP effect sizes [82] and methods that estimate heritability only from summary statistics and LD reference panels [20, 83].

An alternative method used to compute heritability is known as HE-regression. HE-regression is a method of moments estimator that optimises variance components (σ_g^2, σ_e^2) in order to minimise the squared difference between the observed and expected trait covariances. HE-regression methods are widely acknowledged to be more computationally efficient[84–86] and do not require any assumptions on the phenotype distribution beyond the covariance structure [84] (in contrast to maximum-likelihood estimators). However, HE-regression based estimates typically have higher variance [85], thus implying that they have less power.

Assuming a single component model

$$y \sim \mathcal{N}(0, \sigma_g^2 K + \sigma_e^2 I),$$

optimization of the variance components can be expressed in the regression framework

$$\text{vec}\left(yy^T\right) = \sigma_g^2 \text{vec}\left(K\right) + \sigma_e^2 \text{vec}\left(I\right) + \epsilon',$$

where $\text{vec}\left(A\right)$ is the vectorization operator that transforms an $N \times M$ matrix into an NM-vector. In matrix format, this corresponds to solving the following linear system

$$\begin{pmatrix} \text{tr}\left(KK\right) & \text{tr}\left(K\right) \\ \text{tr}\left(K\right) & N \end{pmatrix} \begin{pmatrix} \sigma_g^2 \\ \sigma_e^2 \end{pmatrix} = \begin{pmatrix} y^T K y \\ y^T y \end{pmatrix} \tag{1.3}$$

Recent method developments[86, 87] have shown that HE-regression can be computed efficiently on genetic datasets with hundreds of thousands of samples using a randomised approach called RHE. I introduce this in detail here as I have used this method in combination with my own work in subsequent sections.

Wu and Sankararaman [86] observed that Equation (1.3) can be solved efficiently without ever having to explicitly compute the kinship matrix K by using Hutchkinsons estimator [88], which states that $\mathrm{tr}\,(A) = \mathbb{E}\left[z^T A z\right]$ for any matrix where z is a random vector with mean zero and covariance given by the identity matrix. The proposed method involves approximating $\mathrm{tr}\,(K)$ and $\mathrm{tr}\,(K^2)$ using only matrix vector multiplications with the genotype matrix X with the following expressions

$$\mathrm{tr}\,(K) \approx \frac{1}{B}\frac{1}{M^2}\sum_b ||X^T z_b||_2^2,$$
$$\mathrm{tr}\left(K^2\right) \approx \frac{1}{B}\frac{1}{M^2}\sum_b ||XX^T z_b||_2^2.$$

Thus an approximate solution can be obtained in $\mathcal{O}(NMB)$ time, where B denotes a relatively small number of random samples.

Subsequent work by Pazokitoroudi et al. [87] showed that heritability estimates for the multiple component model

$$y \sim \mathcal{N}(0, \sum_k \sigma_k^2 K_k + I\sigma_e^2)$$

can also be efficiently obtained as solution to the linear system given by

$$\begin{pmatrix} T & b \\ b^T & N \end{pmatrix} \sigma^2 = \begin{pmatrix} \boldsymbol{c} \\ N \end{pmatrix} \tag{1.4}$$

where

$$T_{kl} = \mathrm{tr}\,(K_k K_l) \tag{1.5}$$
$$b_k = \mathrm{tr}\,(K_k) \tag{1.6}$$
$$c_k = y^T K_k y. \tag{1.7}$$

Using Hutchkinson's estimator as in the single component model allows the algorithm to scale in $\mathcal{O}(NMBK)$, where K is the number of components used. Pazokitoroudi et al. [87] also show how covariates can be efficiently projected out of the system, and how variance estimates can be computed with the block jacknife estimator without changing the computational complexity of the algorithm. These extensions are detailed in Appendix A.2.

1.4 Thesis outline

Chapter 2 describes the Linear Mixed Model Analysis (LEMMA) method that I have developed for identifying GxE interactions with a linear combination of multiple environmental variables. I give a description of the probabilistic model, followed by the inference procedure and details of the implementation required to enable LEMMA to scale to biobank scale datasets.

Chapter 3 contains two simulations studies; one that compares LEMMA to several existing single SNP methods when the phenotype is simulated from the LEMMA model, and one that compares the same set of methods in a scenario of model misspecification.

Chapter 4 contains results from applying LEMMA to body mass index (BMI) and three blood pressure traits using 42 environmental variables and approximately 280,000 white British individuals from the UK Biobank. In a GWAS of multiplicative GxE interactions, I find three loci with genome wide significant effects. For three of the four traits I estimate that multiplicative GxE interactions with the ES make non-negligible contributions to trait variance, and that this contribution to trait variance is concentrated amongst rare SNPs.

In Chapter 5 I present a novel approach that aims to estimate the same ES used by LEMMA, but from within the framework of randomised HE-regression. Such an approach promises to be more efficient.

Chapter 6 contains a discussion on the work presented in this thesis along with possible extensions of the methods.

2

The Linear Environment Mixed Model Analysis (LEMMA)

Contents

Models that consider environmental variables jointly can be advantageous, particularly if several environmental variables drive interactions at individual loci or if an unobserved environment that drives interactions is better reflected by a combination of observed environments. StructLMM[55] models the environmental

similarity between individuals (over multiple environments) as a random effect, and then tests each SNP independently for GxE interactions, but does not model the genome wide contribution of all the markers, which is often a major component of phenotypic variance.

In this chapter I present a new method called Linear Environment Mixed Model Analysis (LEMMA) which aims to combine the advantages of WGR and modelling GxE with multiple environments. Instead of assuming that the GxE effect over multiple environments is independent at each variant, as StructLMM does, LEMMA learns a single linear combination of environmental variables with a common role in interaction effects genome wide. I refer to this linear combination as an environmental score (ES). One advantage of this approach is that the ES provides a readily interpretable way to examine the combined effect of many environmental variables. The ES is estimated jointly with SNP effects within a Bayesian WGR model and variational inference is used to fit the model.

The LEMMA analysis has several distinct steps. First, the Bayesian WGR model is fitted using genotyped SNPs. Sections 2.1 to 2.3 give a formal specification of the WGR model, describe model inference and give details of the implementation. Section 2.4 shows how the estimated ES can then be used to estimate the proportion of phenotypic variability that is explained by GxE interactions (SNP GxE heritability), using randomised Haseman-Elston (RHE) regression on UK Biobank scale datasets [86, 87]. Finally Section 2.5 details how the ES can be used to perform hypothesis tests for multiplicative GxE interactions at each SNP in a larger set of imputed SNPs. I use “robust" standard errors when testing each variant for a GxE interaction, which helps to control for the conditional heteroskedasticity caused by GxE interactions.

2.1 LEMMA Model Structure

The usual Bayesian whole genome regression model is given by

$$y = C\alpha + X\beta + \epsilon,$$

where y is the centred and scaled $N \times 1$ vector of phenotypes, C is an $N \times L'$ matrix of covariates with fixed effects α, ϵ is an N-vector of residual effects, X is the $N \times M$ genotype matrix normalised to have column mean zero and column variance one, and β is a M-vector modelling the random effect of each SNP. I extend this setup to model multiplicative GxE interactions genome-wide with a linear combination of multiple environmental variables using

$$Y = C\alpha + X\beta + Z\gamma + \epsilon, \tag{2.1}$$

where

$$\begin{aligned} Z &= \operatorname{diag}(\eta)\, X, \\ \eta &= Ew, \\ w &\sim \mathcal{N}(0, I_L) \end{aligned}$$

where E is an $N \times L$ matrix of environmental variables that could potentially be involved in GxE interactions and η is a linear combination of those environments that is learnt in tandem with SNP effects. I use the phrase 'environmental score', or ES, to refer to η throughout this thesis. All environmental variables contained in E are also contained in C. The interaction weights w are modelled with a gaussian prior, but in theory one could consider sparser priors such as a spike and slab.

I place mixture of normals priors on both the main effects and the interaction effects, given by

$$\begin{aligned} \beta_j|\lambda_\beta, \sigma^2_{\beta,1}, \sigma^2_{\beta,2} &\sim \lambda_\beta \mathcal{N}(0, \sigma^2_{\beta,1}) + (1-\lambda_\beta)\mathcal{N}(0, \sigma^2_{\beta,2}), \\ \gamma_j|\lambda_\gamma, \sigma^2_{\gamma,1}, \sigma^2_{\gamma,2} &\sim \lambda_\gamma \mathcal{N}(0, \sigma^2_{\gamma,1}) + (1-\lambda_\gamma)\mathcal{N}(0, \sigma^2_{\gamma,2}) \end{aligned}$$

and standard normal priors on the covariate and error terms.

$$\begin{aligned} \epsilon|\sigma^2_e &\sim \mathcal{N}(0, \sigma^2_e), \\ \alpha|\sigma^2_e, \sigma^2_\alpha &\sim \mathcal{N}(0, \sigma^2_e \sigma^2_\alpha). \end{aligned}$$

2.2 Variational Inference

For notational convenience let $\mathcal{D} := \{X, E\}$ denote the genetic and environmental data, $\theta = \{\alpha, \beta, \gamma, w\}$ the set of variables and $\phi = \{\sigma_e^2, \{\sigma_{\beta,i}^2\}_{i=1}^2, \{\sigma_{\gamma,i}^2\}_{i=1}^2, \lambda_\beta, \lambda_\gamma, \sigma_\alpha^2\}$ the set of hyperparameters.

The posterior distribution $p(\theta|y, \mathcal{D}, \phi)$, given by

$$p(\theta|y, \mathcal{D}, \phi) \propto p(y|\theta, \mathcal{D}, \phi) \prod_c p(\alpha_c|\phi) \prod_l p(w_l) \prod_j p(\beta_j, u_j|\phi) \prod_j p(\gamma_j, v_j|\phi),$$

is intractable. The two principal tools available to statisticians to evaluate intractable distributions are Markov chain Monte Carlo (MCMC) and Variational Inference (VI) [89]. MCMC methods characterise the posterior distribution using an ergodic Markov chain that generates samples from that same distribution. MCMC methods are both asymptotically unbiased and quite flexible, and often all that is required is to be able to compute the density for a given sample up to a scalar factor. However, they can also be slow to converge on large datasets or complex models. Variational inference, on the other hand, generates a biased approximation of the posterior distribution in a manner that is usually faster and easier to scale to large datasets [90]. Typically this bias exists only in estimates of posterior variances, and estimates of the posterior mean are unbiased. As this method is intended for use on biobank scale datasets with hundreds of thousands of samples, I use variational inference for model inference.

Variational inference approximates the true posterior $p(\theta|y, \mathcal{D}, \phi)$ with a tractable alternative distribution $q(\theta; \nu)$ governed by (variational) parameters ν. Typically ν is high dimensional, allowing $q(\theta; \nu)$ to be highly flexible. The variational parameters ν are then optimized to make $q(\theta; \nu)$ as close as possible to the true posterior, where 'close' is defined in terms of the KL Divergence

$$KL(q||p) = \int \log \frac{p(\theta|y, \mathcal{D}, \phi)}{q(\theta; \nu)} q(\theta; \nu) d\theta = -\mathbb{E}_q \left[\log \frac{p(\theta|y, \mathcal{D}, \phi)}{q(\theta; \nu)} \right],$$

where $\mathbb{E}_q[.]$ denotes the expectation with respect to $q(\theta; \nu)$. Intuitively, the KL Divergence is a measure of the difference between two distributions. Viewed from

another perspective, minimising the KL Divergence is equivalent to maximising a lower bound on the marginal log-likelihood. One can see this by observing

$$\begin{aligned} KL(q||p) &= -\mathbb{E}_q\left[\log \frac{p(\theta|y,\mathcal{D},\phi)}{q(\theta;\nu)}\right], \\ &= -\mathbb{E}_q\left[\log \frac{p(\theta,y|\mathcal{D},\phi)}{q(\theta;\nu)}\right] + \mathbb{E}_q\left[\log p(y|\mathcal{D},\phi)\right], \\ &= -\mathbb{E}_q\left[\log \frac{p(\theta,y|\mathcal{D},\phi)}{q(\theta;\nu)}\right] + \log p(y|\mathcal{D},\phi), \end{aligned}$$

hence as $KL(q||p)$ is always greater than zero we can write

$$\mathcal{F}(\nu;\phi) := \mathbb{E}_q\left[\log \frac{p(\theta,y|\mathcal{D},\phi)}{q(\theta;\nu)}\right] \leq \log p(y|\mathcal{D},\phi)$$

As a consequence $\mathcal{F}(\nu;\phi)$ is known colloquially as the Evidence Lower Bound (ELBO)[90]. Alternatively, the ELBO can be decomposed into the expected log-likelihood conditional on the data and the KL Divergence between the variational distributions and their respective priors

$$\mathcal{F}(\nu;\phi) = \underbrace{\mathbb{E}_q\left[\log p(y|\theta,\mathcal{D},\phi)\right]}_{\text{conditional log-likelihood}} + \underbrace{KL(q(\theta)||p(\theta|\phi))}_{\substack{\text{KL-divergence between prior} \\ \text{and approximate posterior}}}$$

In this way the ELBO can be viewed as a balancing act between finding configurations of ν that explain patterns in the observed data, and forming variational densities that are close to the prior[90].

To make evaluating $KL(q||p)$ tractable, I use a popular assumption known as the Mean Field approximation, which assumes that the approximating distribution factorises. Consequently the approximate posterior $q(\theta;\nu)$ can be written as

$$q(\theta;\nu) = \prod_c q(\alpha_c) \prod_l q(w_l) \prod_j q(\beta_j, u_j) \prod_j q(\gamma_j, v_j).$$

2.2.1 Coordinate Ascent Variational Inference

Under the mean field assumption, the optimal distribution for the j'th parameter conditional on all other parameters is given by

$$q^*(\theta_j) \propto \exp \mathbb{E}_{-q(\theta_j)}\left[\log p(\theta_j|\theta_{-j},y,\mathcal{D},\phi)\right] \tag{2.2}$$

where $\mathbb{E}_{-q(\theta_j)}[.]$ denotes the expectation taken over all variables except θ_j (for clarity, the standard derivation[91] is provided in Appendix A.1).

Until this point nothing has been said about the densities used for the approximating distributions $q(\theta_j)$. In some instances it is popular to simply specify the densities, for example gaussian distributions are computationally and analytically advantageous, but this imposes an additional assumption on the inference strategy. However, when the priors and the conditional likelihood are conjugate the optimal approximating distribution (as well as optimal variational parameters for that form) can be simply obtained from Equation (2.2). For LEMMA, gaussian and mixture of gaussian priors are both conjugate priors. Thus the optimal posterior approximations for the set of latent variables are given by

$$\begin{aligned} q(w_l) &= \mathcal{N}(\mu_l^w, s_l^w), \\ q(\alpha_c) &= \mathcal{N}(\mu_c^\alpha, s_c^\alpha), \\ q(\beta_j, u_j) &= \psi_j^\beta \mathcal{N}(\mu_{j,1}^\beta, s_{j,1}^\beta) + (1-\psi_j^\beta)\mathcal{N}(\mu_{j,2}^\beta, s_{j,2}^\beta) \\ q(\gamma_j, v_j) &= \psi_j^\gamma \mathcal{N}(\mu_{j,1}^\gamma, s_{j,1}^\gamma) + (1-\psi_j^\gamma)\mathcal{N}(\mu_{j,2}^\gamma, s_{j,2}^\gamma), \end{aligned}$$

and the set of parameters of the variational approximation is given by

$$\nu = \{\boldsymbol{\mu}^w, \boldsymbol{s}^w, \boldsymbol{\mu}^\alpha, \boldsymbol{s}^\alpha, \boldsymbol{\psi}^\beta, \{\boldsymbol{\mu}_i^\beta\}_{i=1}^2, \{\boldsymbol{s}_i^\beta\}_{i=1}^2, \boldsymbol{\psi}^\gamma, \{\boldsymbol{\mu}_i^\gamma\}_{i=1}^2, \{\boldsymbol{s}_i^\gamma\}_{i=1}^2\}.$$

Each Coordinate Ascent Variational Inference (CAVI) update is convex, conditional on the current estimates for all of the other parameters. CAVI updates can then be applied in a cyclic coordinate ascent scheme. I refer to one 'iteration' as a full cycle of updates, where every parameter is updated once. Analytic updates for each of these variational parameters are given below.

VI parameter updates

In the following updates, I make frequent use of the following intermediate variables

$$\begin{aligned} y_m &= C\alpha + X\beta, \\ y_x &= X\gamma, \\ \eta &= Ew. \end{aligned}$$

These are not just for notational convenience; keeping copies of these variables updated is a frequently used trick in whole genome regression algorithms that reduces the cost of each CAVI iteration from $\mathcal{O}(NP)$ to $\mathcal{O}(N)$. I also use X_{-j} to denote the matrix X excluding the j'th column, and β_{-j} to denote β excluding the j'th element.

Updates for $q(\beta_j)$

I provide a derivation of the update for $q(\beta_j)$ as an example. Derivations of the other updates follow by similarity.

The CAVI update scheme (see Appendix A.1) states that

$$q^*(\theta_j) \propto \exp \mathbb{E}_{-q(\theta_j)} \left[\log p(y|\beta_j, \theta_{-j}, \mathcal{D}, \phi) + \log p(\beta_j)\right] \tag{2.3}$$

The conditional log likelihood can be written as

$$\log p(y|\beta_j, \theta_{-j}, \mathcal{D}, \phi) = -\frac{1}{2\sigma_e^2}\left(||X_j||_2^2\beta_j^2 - 2X_j^T y_{\text{resid, -j}}\beta_j\right) + Cnst, \tag{2.4}$$

where $Cnst$ is a constant independent of β_j and

$$y_{\text{resid},-\beta_j} = y - C\alpha - X_{-j}\beta_{-j} - \text{diag}\,(\eta)\, X\gamma.$$

Substituting eq. (2.4) into eq. (2.3) yields

$$q^*(\theta_j) \propto \exp\left(\mathbb{E}_{-q(\theta_j)}\left[\log p(y|\beta_j, \theta_{-j}, \mathcal{D}, \phi)\right]\right) \exp\left(\mathbb{E}_{-q(\theta_j)}\left[\log p(\beta_j)\right]\right), \tag{2.5}$$

$$= \exp\left(-\frac{||X_j||_2^2}{2\sigma_e^2}\beta_j^2 + \frac{1}{\sigma_e^2}X_j^T\mathbb{E}_{-q(\theta_j)}\left[y_{\text{resid, -j}}\right]\beta_j\right) \exp\left(\mathbb{E}_{-q(\theta_j)}\left[\log p(\beta_j)\right]\right), \tag{2.6}$$

$$= \exp\left(-\frac{||X_j||_2^2}{2\sigma_e^2}\beta_j^2 + \frac{1}{\sigma_e^2}X_j^T\mathbb{E}_{-q(\theta_j)}\left[y_{\text{resid, -j}}\right]\beta_j\right) \exp\left(\log p(\beta_j)\right), \tag{2.7}$$

$$= \exp\left(-\frac{||X_j||_2^2}{2\sigma_e^2}\beta_j^2 + \frac{1}{\sigma_e^2}X_j^T\mathbb{E}_{-q(\theta_j)}\left[y_{\text{resid, -j}}\right]\beta_j\right) p(\beta_j), \tag{2.8}$$

where eq. (2.7) follows because the prior on β_j is indepedent of θ_{-j}. Recall that the prior on β_j is given by

$$p(\beta_j) = \lambda_\beta \mathcal{N}(\beta_j; 0, \sigma_{\beta,1}^2) + (1-\lambda_\beta)\mathcal{N}(\beta_j; 0, \sigma_{\beta,2}^2),$$

and note that

$$\exp\left(-\frac{||X_j||_2^2}{2\sigma_e^2}\beta_j^2 + \frac{1}{\sigma_e^2}X_j^T\mathbb{E}_{-q(\theta_j)}\left[y_{\text{resid, -j}}\right]\beta_j\right)\mathcal{N}(\beta_j; 0, \sigma_{\beta,i}^2) = \exp\left(\frac{\mu_{j,i}^2}{2s_{j,i}^\beta}\right)\sqrt{\frac{s_{j,i}^\beta}{\sigma_{\beta,i}^2}}\mathcal{N}(\beta_j; \mu_{j,i}^\beta, s_{j,i}^\beta), \tag{2.9}$$

where

$$s_{j,i}^\beta = \frac{\sigma_e^2}{||X_j||_2^2 + \sigma_e^2/\sigma_{\beta,i}^2}, \qquad \text{for } i = 1, 2 \tag{2.10}$$

$$\mu_{j,i}^\beta = \frac{s_{j,i}^\beta}{\sigma_e^2}X_j^T\mathbb{E}_{-q(\beta_j)}\left[y_{\text{resid},-\beta_j}\right], \qquad \text{for } i = 1, 2. \tag{2.11}$$

Therefore we can rewrite Equation (2.8) as

$$q^*(\theta_j) \propto \lambda_\beta \exp\left(\frac{(\mu_{j,1}^\beta)^2}{2s_{j,1}^\beta}\right)\sqrt{\frac{s_{j,1}^\beta}{\sigma_{\beta,1}^2}}\mathcal{N}(\beta_j; \mu_{j,1}^\beta, s_{j,1}^\beta) + \tag{2.12}$$

$$(1-\lambda_\beta)\exp\left(\frac{(\mu_{j,2}^\beta)^2}{2s_{j,2}^\beta}\right)\sqrt{\frac{s_{j,2}^\beta}{\sigma_{\beta,2}^2}}\mathcal{N}(\beta_j; \mu_{j,2}^\beta, s_{j,2}^\beta), \tag{2.13}$$

$$= \psi\mathcal{N}(\beta_j; \mu_{j,1}^\beta, s_{j,1}^\beta) + (1-\psi)\mathcal{N}(\beta_j; \mu_{j,2}^\beta, s_{j,2}^\beta). \tag{2.14}$$

For $q^*(\theta_j)$ to be a valid distribution the mixture component ψ must sum to one. Therefore we obtain

$$\psi = \frac{\lambda_\beta \exp\left(\frac{(\mu_{j,1}^\beta)^2}{2s_{j,1}^\beta}\right)\sqrt{\frac{s_{j,1}^\beta}{\sigma_{\beta,1}^2}}}{\lambda_\beta \exp\left(\frac{(\mu_{j,1}^\beta)^2}{2s_{j,1}}\right)\sqrt{\frac{s_{j,1}^\beta}{\sigma_{\beta,1}^2}} + (1-\lambda_\beta)\exp\left(\frac{(\mu_{j,2}^\beta)^2}{2s_{j,2}^\beta}\right)\sqrt{\frac{s_{j,2}^\beta}{\sigma_{\beta,2}^2}}} \tag{2.15}$$

For convenience of notation, I rearrange Equation (2.15) to yield

$$\psi_j^\beta = \text{logistic}\left(\text{logit}\left(\lambda_\beta\right) - \frac{1}{2}\log\left(\frac{\sigma_{\beta,1}^2 s_{j,2}^\beta}{s_{j,1}^\beta\sigma_{\beta,2}^2}\right) + \frac{(\mu_{j,1}^\beta)^2}{2s_{j,1}^\beta} - \frac{(\mu_{j,2}^\beta)^2}{2s_{j,2}^\beta}\right).$$

Therefore, the update equations for $q(\beta_j)$ can be summarised as

$$\begin{aligned}
q(\beta_j) &= \psi_j^\beta\mathcal{N}(\beta_j|\mu_{j,1}^\beta, s_{j,1}^\beta) + (1-\psi_j^\beta)\mathcal{N}(\beta_j|\mu_{j,2}^\beta, s_{j,2}^\beta), \\
s_{j,i}^\beta &= \frac{\sigma_e^2}{||X_j||_2^2 + \sigma_e^2/\sigma_{\beta,i}^2}, && \text{for } i = 1, 2 \\
\mu_{j,i}^\beta &= \frac{s_{j,i}^\beta}{\sigma_e^2}X_j^T\mathbb{E}_{-q(\beta_j)}\left[y_{\text{resid},-\beta_j}\right], && \text{for } i = 1, 2 \\
\psi_j^\beta &= \text{logistic}\left(\text{logit}\left(\lambda_\beta\right) - \frac{1}{2}\log\left(\frac{\sigma_{\beta,1}^2 s_{j,2}^\beta}{s_{j,1}^\beta\sigma_{\beta,2}^2}\right) + \frac{(\mu_{j,1}^\beta)^2}{2s_{j,1}^\beta} - \frac{(\mu_{j,2}^\beta)^2}{2s_{j,2}^\beta}\right)
\end{aligned}$$

where

$$
\begin{aligned}
y_{\text{resid},-\beta_j} &= y - C\alpha - X_{-j}\beta_{-j} - \text{diag}\left(\eta\right) X\gamma, \\
\mathbb{E}_{-q(\beta_j)}\left[y_{\text{resid},-\beta_j}\right] &= y - C\mathbb{E}_q\left[\alpha\right] - X_{-j}\mathbb{E}_q\left[\beta_{-j}\right] - \text{diag}\left(\mathbb{E}_q\left[\eta\right]\right) X\mathbb{E}_q\left[\gamma\right]
\end{aligned}
$$

Updates for $q(\gamma_j)$

$$
\begin{aligned}
q(\gamma_j) &= \psi_j^\gamma \mathcal{N}(\gamma_j|\mu_{j,1}^\gamma, s_{j,1}^\gamma) + (1-\psi_j^\gamma)\mathcal{N}(\gamma_j|\mu_{j,2}^\gamma, s_{j,2}^\gamma), \\
s_{j,i}^\gamma &= \frac{\sigma_e^2}{\mathbb{E}_{-q(\gamma_j)}\left[||Z_j||_2^2\right] + \sigma_e^2/\sigma_{\gamma,i}^2}, && \text{for } i = 1,2 \\
\mu_{j,i}^\gamma &= \frac{s_{j,i}^\gamma}{\sigma_e^2} X_j^T \mathbb{E}_{-q(\gamma_j)}\left[\text{diag}\left(\eta\right) y_{\text{resid},-\gamma_j}\right], && \text{for } i = 1,2 \\
\psi_j^\gamma &= \text{logistic}\left(\text{logit}\left(\lambda_\gamma\right) - \frac{1}{2}\log\left(\frac{\sigma_{\gamma,1}^2 s_{j,2}^\gamma}{s_{j,1}^\gamma \sigma_{\gamma,2}^2}\right) + \frac{(\mu_{j,1}^\gamma)^2}{2s_{j,1}^\gamma} - \frac{(\mu_{j,2}^\gamma)^2}{2s_{j,2}^\gamma}\right),
\end{aligned}
$$

where

$$
\begin{aligned}
y_{\text{resid},-\gamma_j} &= y - C\alpha - X\beta - \text{diag}\left(\eta\right) X_{-j}\gamma_{-j}, \\
\mathbb{E}_{-q(\gamma_j)}\left[\text{diag}\left(\eta\right) y_{\text{resid},-\gamma_j}\right] &= \text{diag}\left(\hat{\eta}\right)\left(y - \hat{y}_m\right) + \text{diag}\left(\hat{\eta^2}\right) X_{-j}\mathbb{E}_q\left[\gamma_{-j}\right] \\
\mathbb{E}_{-q(\gamma_j)}\left[||Z_j||_2^2\right], &= \sum_i \mathbb{E}_q\left[(X_{ij}\eta_i)^2\right], \\
&= \sum_{l,l'} \mathbb{E}_q\left[w_l\right]\mathbb{E}_q\left[w_{l'}\right] \sum_i X_{ij}^2 E_{il} E_{il'} \\
&\quad + \sum_l \text{Var}_q\left(w_l\right) \sum_i X_{ij}^2 E_{il}^2.
\end{aligned}
$$

Naively recomputing the expected L2-norm of Z_j for every CAVI update would cost $\mathcal{O}(NL^2)$. Instead I precompute $\sum_i \sum_{l,l'} X_{ij}^2 E_{il} E_{il'}$ for every $1 \leq l < l' \leq L$ and $1 \leq j \leq M$, which is an $\mathcal{O}(NL^2)$ operation but need only be computed once. Therefore the cost of the expected L2-norm of Z_j is $\mathcal{O}(L^2)$ and the cost of the CAVI update is $\mathcal{O}(N)$.

Updates for $q(w_l)$

$$
\begin{aligned}
q(w_l) &= \mathcal{N}(w_l|\mu_l^w, s_l^w), \\
s_l^w &= \frac{\sigma_e^2}{\mathbb{E}_{-q(w_l)}\left[||B_l||_2^2\right] + \sigma_e^2}, \\
\mu_l^w &= \frac{s_l^w}{\sigma_e^2}\mathbb{E}_{-q(w_l)}\left[B_l^T y_{\text{resid},-w_l}\right],
\end{aligned}
$$

where

$$
\begin{aligned}
B &= \operatorname{diag}(X\gamma)\, E, \\
y_{\text{resid},-w_l} &= y - y_m - \operatorname{diag}(y_x)\, E_{-l} w_{-l}, \\
\mathbb{E}_{-q(w_l)}\left[B_l^T y_{\text{resid},-w_l}\right] &= E_l^T \operatorname{diag}(\hat{y}_x)\,(y - \hat{y}_M) - E_l^T \operatorname{diag}\left(\hat{y}_x^2\right)\hat{\eta}_{-l} \\
&\quad - \sum_j \operatorname{Var}_q(\gamma_j) \sum_{l' \neq l} \mathbb{E}_q[w_{l'}] \sum_i E_{il} E_{il'} X_{ij}^2, \\
\mathbb{E}_{-q(w_l)}\left[||B_l||_2^2\right] &= \sum_i \mathbb{E}_{-q(w_l)}\left[(E_{il} y_x)^2\right], \\
&= \sum_i E_{il}^2 \hat{y}_x^2 + \sum_j \operatorname{Var}_q(\gamma_j) \sum_i E_{il}^2 X_{ij}^2.
\end{aligned}
$$

Naive computation of $\mathbb{E}_{-q(w_l)}\left[B_l^T y_{\text{resid},-w_l}\right]$ and $\mathbb{E}_{-q(w_l)}\left[||B_l||_2^2\right]$ would cost $\mathcal{O}(NML)$ and $\mathcal{O}(NM)$ per CAVI update respectively. However given precomputation of $\sum_i E_{il}^2 X_{ij}^2$ for $1 \leq l < l' \leq M$ and $1 \leq j \leq M$ (similar to updates of γ_j), these calculations can be performed in $\mathcal{O}(ML)$ and $\mathcal{O}(M)$.

Updates for $q(\alpha_c)$

$$
\begin{aligned}
q(\alpha_c) &= \mathcal{N}(\alpha_c | \mu_c^\alpha, s_c^\alpha), \\
s_c^\alpha &= \frac{\sigma_e^2}{1/\sigma_\alpha^2 + (N-1)}, \\
\mu_c^\alpha &= \frac{s_c^\alpha}{\sigma_e^2} E_c^T \mathbb{E}_{-q(\alpha_c)}\left[y_{\text{resid},-\alpha_c}\right],
\end{aligned}
$$

where

$$
\begin{aligned}
y_{\text{resid},-\alpha_c} &= y - C_{-c}\alpha_{-c} - X\beta - \operatorname{diag}(\hat{\eta})\, X\gamma, \\
\mathbb{E}_{-q(\alpha_c)}\left[y_{\text{resid},-\alpha_c}\right] &= y - C_{-c}\mathbb{E}_q[\alpha_{-c}] - X\mathbb{E}_q[\beta] - \operatorname{diag}(\hat{\eta})\, \hat{y}_x.
\end{aligned}
$$

2.2.2 Evidence lower bound

Variational inference involves maximising the evidence lower bound (ELBO) $\mathcal{F}(\phi; \nu)$ on the model log-likelihood $\log p(y|\mathcal{D}, \phi)$. The ELBO can be separated into the expected conditional log-likelihood and the KL divergence between the variational

distribution and the respective priors. This is given by

$$\begin{aligned}\mathcal{F}(\phi;\nu) &= \mathbb{E}_q\left[\log p(y|\theta,\mathcal{D},\phi)\right] - \mathrm{KL}(q(\theta;\nu)||p(\theta|\phi)),\\ &= \mathbb{E}_q\left[\log p(y|\theta,\mathcal{D},\phi)\right] - \mathbb{E}_q\left[\log q(\theta;\nu) - \log p(\theta|\phi)\right],\\ &= \mathbb{E}_q\left[\log p(y|\theta,\mathcal{D},\phi)\right] - \sum_j \mathbb{E}_q\left[\log q(\theta_j;\nu_j) - \log p(\theta_j|\phi)\right],\\ &= \mathbb{E}_q\left[\log p(y|\theta,\mathcal{D},\phi)\right] - \sum_j \mathrm{KL}(q(\theta_j;\nu_j)||p(\theta_j|\phi)). \qquad (2.16)\end{aligned}$$

The expected conditional log-likelihood is given by

$$\begin{aligned}\mathbb{E}_q\left[\log p(y|\theta,\mathcal{D},\phi)\right] =& -\frac{N}{2}\log(2\pi\sigma_e^2)\\ &- \frac{1}{2\sigma_e^2}\left(||y - C\mathbb{E}_q[\alpha] - X\mathbb{E}_q[\beta] - \mathbb{E}_q[\eta]\odot X\mathbb{E}_q[\gamma]||_2^2\right)\\ &- \frac{1}{2\sigma_e^2}\left(\mathbb{E}_q[\gamma]^T X^T \mathrm{diag}\left(\mathbb{E}_q\left[\eta^2\right]\right) X\mathbb{E}_q[\gamma] - ||\mathbb{E}_q[\eta]\odot X\mathbb{E}_q[\gamma]||_2^2\right)\\ &- \frac{N-1}{2\sigma_e^2}\sum_l \mathrm{Var}_q(\alpha_l)\\ &- \frac{N-1}{2\sigma_e^2}\sum_k \mathrm{Var}_q(\beta_k).\end{aligned}$$

For two univariate gaussian distributions $u \sim \mathcal{N}(0,\sigma^2)$ and $v \sim \mathcal{N}(\mu,s)$, the KL Divergence is given by

$$KL(v\|u) = -\frac{1}{2} - \log(\sigma^2) + \log(s) - \frac{s+\mu^2}{2\sigma^2}.$$

The KL Divergence between two mixtures of gaussians is not analytically tractable. However the matched bound approximation [92] can be used to provide an upper bound when both have the same number of components. Thus for two mixtures of gaussians given by

$$\begin{aligned}u &\sim \lambda\mathcal{N}(0,\sigma_1^2) + (1-\lambda)\mathcal{N}(0,\sigma_2^2),\\ v &\sim \psi\mathcal{N}(\mu_1,s_1) + (1-\psi)\mathcal{N}(\mu_2,s_2),\end{aligned}$$

the matched bound on the KL divergence is given by

$$\begin{aligned}KL(v\|u) \leq& \psi\log\frac{\psi}{\lambda} + (1-\psi)\frac{1-\psi}{1-\lambda}\\ &+ -\frac{1}{2} + \psi\left(\log(s_1) - \log(\sigma_1^2) - \frac{s_1+\mu_1^2}{2\sigma_1^2}\right) + (1-\psi)\left(\log(s_2) - \log(\sigma_2^2) - \frac{s_2+\mu_2^2}{2\sigma_2^2}\right).\end{aligned}$$

Use of the matched bound approximation retains a valid variational algorithm, because it maintains the bound on the KL divergence[93]. Substituting these KL divergences into eq. (2.16) yields the full ELBO

$$
\begin{aligned}
\mathcal{F}(\phi;\nu) = &-\frac{N}{2}\log(2\pi\sigma_e^2) \\
&-\frac{1}{2\sigma_e^2}\left(||y - C\mathbb{E}_q\left[\alpha\right] - X\mathbb{E}_q\left[\beta\right] - \mathbb{E}_q\left[\eta\right]\odot X\mathbb{E}_q\left[\gamma\right]||_2^2\right) \\
&-\frac{1}{2\sigma_e^2}\left(\mathbb{E}_q\left[\gamma\right]^T X^T \mathrm{diag}\left(\mathbb{E}_q\left[\eta^2\right]\right) X\mathbb{E}_q\left[\gamma\right] - ||\mathbb{E}_q\left[\eta\right]\odot X\mathbb{E}_q\left[\gamma\right]||_2^2\right) \\
&-\frac{N-1}{2\sigma_e^2}\sum_l \mathrm{Var}_q\left(\alpha_l\right) \\
&-\frac{N-1}{2\sigma_e^2}\sum_k \mathrm{Var}_q\left(\beta_k\right) \\
&-\frac{1}{2\sigma_e^2}\sum_k \mathrm{Var}_q\left(\gamma_k\right)\left(X^T\mathrm{diag}\left(\mathbb{E}_q\left[\eta^2\right]\right)X\right)_{kk} \\
&+\frac{L'}{2}(1-\log(\sigma^2\sigma_\alpha^2)) + \frac{1}{2}\sum_m^{L'}\left(\log(s_m^\alpha) - \frac{s_m^\alpha + (\mu_m^\alpha)^2}{\sigma^2\sigma_\alpha^2}\right) \\
&-\sum_j\left[\psi_j^\beta\log\frac{\psi_j^\beta}{\lambda_\beta} + (1-\psi_j^\beta)\log\frac{1-\psi_j^\beta}{1-\lambda_\beta}\right] \\
&+\frac{1}{2}\sum_j\left[1 + \psi_j^\beta\left(\log\frac{s_{j,1}^\beta}{\sigma_{\beta,1}^2} - \frac{s_{j,1}^\beta + (\mu_{j,1}^\beta)^2}{\sigma_{\beta,1}^2}\right) + (1-\psi_j^\beta)\left(\log\frac{s_{j,2}^\beta}{\sigma_{\beta,2}^2} - \frac{s_{j,2}^\beta + (\mu_{j,2}^\beta)^2}{\sigma_{\beta,2}^2}\right)\right] \\
&-\sum_j\left[\psi_j^\gamma\log\frac{\psi_j^\gamma}{\lambda_\gamma} + (1-\psi_j^\gamma)\log\frac{1-\psi_j^\gamma}{1-\lambda_\gamma}\right] \\
&+\frac{1}{2}\sum_j\left[1 + \psi_j^\gamma\left(\log\frac{s_{j,1}^\gamma}{\sigma_{\gamma,1}^2} - \frac{s_{j,1}^\gamma + (\mu_{j,1}^\gamma)^2}{\sigma_{\gamma,1}^2}\right) + (1-\psi_j^\gamma)\left(\log\frac{s_{j,2}^\gamma}{\sigma_{\gamma,2}^2} - \frac{s_{j,2}^\gamma + (\mu_{j,2}^\gamma)^2}{\sigma_{\gamma,2}^2}\right)\right] \\
&+\frac{L}{2} + \frac{1}{2}\sum_l\left(\log(s_l^w) - (s_l^w + (\mu_l^w)^2)\right)
\end{aligned}
$$

2.2.3 Hyperparameter maximisation

For whole-genome regression methods using variational inference[71–73, 75], the core variational inference updates are typically very similar up to the choice of prior over SNP effects. There is however some diversity in the strategies used to optimise the hyperparameters, which for LEMMA are given by

$$\phi = \{\sigma_e^2, \{\sigma_{\beta,i}^2\}_{i=1}^2, \{\sigma_{\gamma,i}^2\}_{i=1}^2, \lambda_\beta, \lambda_\gamma, \sigma_\alpha^2\}$$

Some authors have essentially performed a grid search over possible hyperparameter combinations, using either predictive mean squared error [75] (ie cross-validation) or the approximate log-likelihood[73] to choose the optimal hyperparameter values. Others have optimised the hyperparameters within the variational inference framework, using either unconstrained maximisation of the ELBO[72] or penalised maximisation having placed priors on the hyperparameters[71] (sometimes known as hyperpriors).

For LEMMA, I set σ_α^2 to a large constant (1000) to create a flat prior on covariate effect sizes. This is similar to the unconstrained maximisation on covariate effect sizes used by Logsdon et al. [72]. A grid search over the remaining seven hyperparameters would be computationally demanding, both because the set of hyperparameters is larger and because LEMMA cannot efficiently perform multiple runs in parallel as done by Loh *et al.*[75]. Instead, I perform unconstrained maximisation of the ELBO with respect to the hyperparameters in a similar manner to that proposed previously[72]. In this manner, the LEMMA approach can be viewed as a variational expectation maximisation algorithm [94, 95].

Partial derivatives of the ELBO $F(\phi;\nu)$ with respect to the seven hyperpa-

rameters ϕ are given by

$$\frac{\partial F}{\partial \lambda_\beta} = \sum_j \left(\frac{\psi_j^\beta}{\lambda_\beta} - \frac{(1-\psi_j^\beta)}{1-\lambda_\beta} \right),$$
$$\frac{\partial F}{\partial \sigma_{\beta,1}^2} = \sum_j \frac{\psi_j}{2} \left(-\frac{1}{\sigma_{\beta,1}^2} + \frac{s_{j,1}^\beta + (\mu_{j,1}^\beta)^2}{(\sigma_{\beta,1}^2)^2} \right),$$
$$\frac{\partial F}{\partial \sigma_{\beta,2}^2} = \sum_j \frac{1-\psi_j}{2} \left(-\frac{1}{\sigma_{\beta,2}^2} + \frac{s_{j,2}^\beta + (\mu_{j,2}^\beta)^2}{(\sigma_{\beta,2}^2)^2} \right),$$

$$\frac{\partial F}{\partial \lambda_\gamma} = \sum_j \left(\frac{\psi_j^\gamma}{\lambda_\gamma} - \frac{(1-\psi_j^\gamma)}{1-\lambda_\gamma} \right),$$
$$\frac{\partial F}{\partial \sigma_{\gamma,1}^2} = \sum_j \frac{\psi_j}{2} \left(-\frac{1}{\sigma_{\gamma,1}^2} + \frac{s_{j,1}^\gamma + (\mu_{j,1}^\gamma)^2}{(\sigma_{\gamma,1}^2)^2} \right),$$
$$\frac{\partial F}{\partial \sigma_{\gamma,2}^2} = \sum_j \frac{1-\psi_j}{2} \left(-\frac{1}{\sigma_{\gamma,2}^2} + \frac{s_{j,2}^\gamma + (\mu_{j,2}^\gamma)^2}{(\sigma_{\gamma,2}^2)^2} \right),$$

$$\frac{\partial F}{\partial \sigma_e^2} = -\frac{N}{2\sigma_e^2} + \frac{1}{2(\sigma_e^2)^2} \mathbb{E}_q \left[||y - C\alpha - X\beta - \mathrm{diag}(\eta) X\gamma||_2^2 \right] - \frac{L'}{2\sigma_e^2} + \frac{1}{2(\sigma_e^2)^2 \sigma_\alpha^2} \sum_m \left(s_m^\alpha + (\mu_m^\alpha)^2 \right).$$

Setting the partial derivatives to zero yields the hyperparameter maximization steps

$$\hat{\lambda}_\beta = \frac{1}{M} \sum_j \psi_j^\beta,$$
$$\hat{\sigma}_{\beta,1}^2 = \frac{\sum_j \psi_j (s_{j,1}^\beta + (\mu_{j,1}^\beta)^2)}{\sum_j \psi_j^\beta},$$
$$\hat{\sigma}_{\beta,2}^2 = \frac{\sum_j (1-\psi_j)(s_{j,2}^\beta + (\mu_{j,2}^\beta)^2)}{\sum_j (1-\psi_j^\beta)},$$

$$\hat{\lambda}_\gamma = \frac{1}{M} \sum_j \psi_j^\gamma,$$
$$\hat{\sigma}_{\gamma,1}^2 = \frac{\sum_j \psi_j (s_{j,1}^\gamma + (\mu_{j,1}^\gamma)^2)}{\sum_j \psi_j^\gamma},$$
$$\hat{\sigma}_{\gamma,2}^2 = \frac{\sum_j (1-\psi_j)(s_{j,2}^\gamma + (\mu_{j,2}^\gamma)^2)}{\sum_j (1-\psi_j^\gamma)},$$

$$\hat{\sigma}_e^2 = \frac{\mathbb{E}_q \left[||y - C\alpha - X\beta - \mathrm{diag}\left(\eta\right) X\gamma||_2^2 \right] + \frac{1}{\sigma_\alpha^2} \sum_m \left(s_m^\alpha + (\mu_m^\alpha)^2 \right)}{N + L'}.$$

2.2.4 SQUAREM accelerator

Similar to the EM algorithm, the hyperparameter maximization step can lead to slow exploration of the hyperparameter space and thus to slow convergence of the LEMMA algorithm. I use an accelerator, SQUAREM [96], to speed up convergence. Given two estimates of the hyperparameters ϕ_{t-2} and ϕ_{t-1}, the maximised estimate ϕ_t can be adjusted as follows

$$\tilde{\phi}_t(v_t) = \phi_{t-2} - 2v_t\Delta\phi_{t-1} + v_t^2\Delta^2\phi_t,$$

where $\Delta\phi_{t-1} = \phi_{t-1} - \phi_{t-2}$ and $\Delta^2\phi_t = \phi_t - 2\phi_{t-1} + \phi_{t-2}$. Thus the new adjusted estimate $\tilde{\phi}_t(v_t)$ is a continuous function of the step size v_t, which yields the original estimate ϕ_t for $v_t = -1$. As recommended by Varadhan and Roland [96] I set $v_t = \min(-1, -||\Delta\phi_{t-1}||_2^2/||\Delta^2\phi_t||_2^2)$. Occasionally this leads to a parameter state with worse ELBO than the previous state, yields an estimate that is either outside of the domain of ϕ or causes numerical instability if the change in σ_e^2 is particularly large (ie from $\sigma_e^2 = 0.8$ to $\sigma_e^2 = 0.1$). To solve the later two issues I use a simple backtracking method of halving the distance between v_t and -1 until the proposed change in σ_e^2 is less than 0.05 and $\tilde{\phi}_t(v_t)$ is within the proper domain. To account for the former issue I simply judge model convergence when the absolute change in ELBO drops below a given threshold. I judge the algorithm to have converged when a full pass through all variables yields an absolute change of less than 0.01 in the aproximate log-likelihood (ELBO).

2.3 Implementation and computational efficiency

I implemented the LEMMA algorithm in C++, using a number of steps to improve computational and memory efficiency including parallel computing with OpenMPI, use of the well optimised Intel Math Kernel Library, compressed data formats, precomputation of some SNP-environment correlations (detailed below). I also take advantage of vectorisation using SIMD extensions in a manner similar to previous methods [75, 79], however, this is simply a matter of using the right compilation flags.

As no code was provided for the multi component model by Pazokitoroudi et al. [87] and the code provided by Wu and Sankararaman [86] does not contain extensions to multiplicative GxE I included my own implementation of the multi component RHE method within the LEMMA software package.

2.3.1 Computational efficiency

Using mean field variational inference, estimation of the posterior means of the variables β, γ, w can be reduced to an iterative algorithm that cycles through the variational parameters sequentially, updating each conditional on the values of the others. Taking the main effect of the j'th SNP as an example, the update scheme for β_j can be written heuristically as

$$\tilde{\beta}_j = X_j^T y_{\text{resid}}, \tag{2.17}$$

$$\hat{\beta}_j^t = \text{regularise}(\tilde{\beta}_j; \phi^t), \tag{2.18}$$

$$y_{\text{resid}}^t = y_{\text{resid}}^{t-1} - (\hat{\beta}_j^t - \hat{\beta}_j^{t-1}) X_j^T. \tag{2.19}$$

Equation (2.17) computes the correlation between the j SNP and the residual phenotype vector. Equation (2.18) computes the posterior mean of β_j which depends on the correlation with the residual phenotype, the prior on β_j and the current hyperparameters. Finally Equation (2.19) updates the residual phenotype vector.

CAVI is not an algorithm that lends itself naturally to parallelisation because it specifies that each posterior update should be computed in serial. For each update, the majority of computational time is spent on the dot product in Equation (2.17) and updating the residual phenotype in Equation (2.19). Both are $\mathcal{O}(N)$ BLAS Level 1 operations, which implies that memory access is often the principle bottleneck rather than the number of cores available. It is possible to step up to BLAS Level 2 by updating a block of SNPs in parallel (and performing adjustments with the covariance of that block of SNPs to obtain updates equivalent to treating each SNP one at a time) [75], however this is still a memory bound operation. Instead I use a parallel computing strategy suggested by [97] for use in whole genome regression, and subsequently used by [82], to compute the dot product and perform

the residual update in parallel using OpenMPI. Briefly, I partition the samples such that blocks of rows of the phenotype y, genotypes X and environmental variables E are assigned to each core. For a given update step, each core calculates the dot product for the locally held block of samples and then shares the local dot product with the rest of the network. From this, the dot product for the entire cohort can be reconstructed cheaply. After computing the posterior mean, each core then updates the residual phenotype for the block of samples stored locally.

2.3.2 Missing data

Samples with missing data in the phenotype, environmental variables or covariates are excluded. Missing genetic data is imputed to the mean dosage of that SNP.

Excluding samples with missing data could lead to a reduction in samples size if many environmental variables are included. When the amount of missing data is small this should not pose a problem. In cases when the amount of missing data is significant then data imputation methods could be applied as a preprocessing step. These methods can range in sophistication from mean imputing each variable independently, to PCA based methods that use covariance of the phenotypes, to methods that leverage covariance of both genetic and phenotypic data[98]. If LEMMA is applied in situations where the missing data structure is related to the phenotype of interest then simply excluding missing data could cause bias in the results.

2.3.3 Precomputed quantities

To improve computational efficiency of updates for w_l and γ_j, I precompute $\sum_i \sum_{l,l'} X_{ij}^2 E_{il} E_{il'}$ for every $1 \leq l < l' \leq M$ and $1 \leq j \leq M$. This has a one off cost of $\mathcal{O}(NML^2)$, but is embarrassingly parallel over SNPs so is feasible to compute on biobank scale data with tens of environmental variables on a cluster. I do this with a separate binary written in C++ which produces a (compressed) text file that LEMMA can read from during setup. For analyses of different traits using the same set of environmental variables this need only be performed once.

2.3.4 Parameter initialisation

I initialise the variational mean parameters of the SNP effects β and γ at zero and the mean parameters of the environment weights w at $\frac{1}{L}$ (ie a uniform weighting). Mean parameters of the covariate effects α are initialised at the least squares fit of the covariates on the phenotype.

2.3.5 RAM usage

Naively storing the entire $N \times M$ centred and scaled genotype matrix in RAM would be hugely expensive for biobank scale datasets (approximately 730 gigabytes of RAM for $280,000$ participants and $650,000$ SNPs using single precision *float*s). To reduce RAM usage, LEMMA stores a compressed version of the genotype matrix using NM bytes. To do this LEMMA splits the interval $[0, 2]$ into 2^8 segments and stores the index of the segment that each dosage falls into, as well as the mean and variance for each SNP. Then when operating on a SNP, LEMMA reconstructs the centred and scaled dosages for that SNP. This approach is similar to that used by the BGEN data format [99] and results in a small loss of accuracy, but is more flexible that assuming dosages are hardcoded to $\{0, 1, 2\}$.

To describe this explicitly, let M be the compressed dosage matrix, $\mathbb{J}_{N\times M}$ an $N \times M$ matrix of ones and w be the width of the segment ($w = 2/2^8$). Then the reconstruction of the j'th column of X is given by

$$X_{.j} = \frac{w(M_{.j} + \frac{1}{2}\mathbb{J}_{N\times 1}) - \text{colMeans}(D)_j}{\text{colSds(D)}_j}.$$

2.3.6 Controlling for covariates

Unlike in BOLT-LMM [75], it is not possible to efficiently project covariates out of the model (y, X, Z), because the multiplicative interaction matrix Z changes after each full cycle of VI updates. Instead LEMMA models covariate effects within the variational framework with a gaussian prior. The variance of this prior is fixed at a large value $\sigma_\alpha^2 = 1000$ so the prior provides little penalisation of

covariate effect sizes. This is roughly equivalent to the unconstrained maximisation approach used by Logsdon et al. [72].

2.4 Partitioned heritability estimation

Originally I investigated heritability estimation using the output of the LEMMA VI method using the formula

$$h_g^2 = \frac{S_X^2 \mathrm{Var}\left(\beta\right)}{\mathrm{Var}\left(\beta\right) S_X^2 + \mathrm{Var}\left(\gamma\right) S_Z^2 + 1}, \tag{2.20}$$

$$\approx \frac{S_X^2 \mathrm{Var}_q\left(\beta\right)}{\mathrm{Var}_q\left(\beta\right) S_X^2 + \mathrm{Var}_q\left(\gamma\right) S_Z^2 + 1}, \tag{2.21}$$

where $\mathrm{Var}_q\left(\beta\right) = \hat{\lambda}_\beta \hat{\sigma}^2_{\beta,1} + (1 - \hat{\lambda}_\beta)\hat{\sigma}^2_{\beta,2}$, $\mathrm{Var}_q\left(\gamma\right) = \hat{\lambda}_\gamma \hat{\sigma}^2_{\gamma,1} + (1 - \hat{\lambda}_\gamma)\hat{\sigma}^2_{\gamma,2}$, $S_X^2 = M$ is the sum of column variances of X and S_Z^2 is the sum of column variances of Z.

There is some evidence in the literature that this should be a sensible approach. The first bayesian whole genome regression method to estimate heritability [73] (known as BVSR) implemented a model with a spike and slab prior on SNP effects, used VI for model inference and a similar formula for heritability. A follow up paper from the same group[74] compared BVSR with a new model BSLMM that used MCMC along with a mixture of gaussians prior on SNP effects. The authors were able to show that BVSR produced accurate heritability estimates in traits with a sparse genetic architecture but underestimated heritability as traits grew more polygenic, whereas BSLMM was accurate in both scenarios. The authors suggested that the poor performance of BVSR was due to the assumptions of sparsity inherent in its spike and slab prior.

However, I found that even with a mixture of gaussians prior, in simulation LEMMA still underestimated heritability in polygenic scenarios [results not shown]. I suggest that this tendency to underestimate heritability in polygenic scenarios is actually due to the mean field assumption. This can clearly be seen by substituting

the hyperparameter maximisation steps from Section 2.2.3 into $\widehat{\text{Var}\,(\beta)}$

$$\begin{aligned}\widehat{\text{Var}\,(\beta)} &= \hat{\lambda}_\beta \hat{\sigma}^2_{\beta,1} + (1-\hat{\lambda}_\beta)\hat{\sigma}^2_{\beta,2}, \\ &= \frac{1}{M}\sum_j \left(\psi^\beta_j (s^\beta_{j,1} + (\mu^\beta_{j,1})^2) + (1-\psi^\beta_j)(s^\beta_{j,2} + (\mu^\beta_{j,2})^2)\right), \\ &= \frac{1}{M}\sum_j \mathbb{E}_q\left[\beta_j^2\right].\end{aligned}$$

It is well documented that one drawback of the mean field assumption is that it causes variational inference algorithms to underestimate the variance of variables [90]. Thus $\text{Var}_q\,(\beta)$ is therefore likely to underestimate $\text{Var}\,(\beta)$ and hence this approach will underestimate h^2_g.

Lending strength to this argument, a more recent method[82] has subsequently shown that accurate heritability estimates can be obtained in a whole genome regression method with a spike and slab prior and MCMC inference. As far as I am aware, the fact that using VI causes whole genome regression methods to be inappropriate for heritability estimation has not been explicitly acknowledged in the literature. Finally I note that previous work has suggested that for estimation of SNP main effects, VI and MCMC frameworks have produced equivalent results [75].

Instead I use both single component and multi component implementations of the recent method called Randomised HE-regression (RHE)[86, 87] to esimate partitioned heritability attributable to additive genetic and multiplicative GxE effects. This has been introduced in Section 1.3.2. Briefly, the single component method estimates variance parameters $(\sigma^2_\beta, \sigma^2_\gamma, \sigma^2_e)$ from the model

$$y \sim \mathcal{N}\left(C\alpha, \sigma^2_\beta K_1 + \sigma^2_\gamma K_2(\hat{w}) + \sigma^2_e I\right), \tag{2.22}$$

where $K_1 = XX^T/M$, $K_2(\hat{w}) = \text{diag}\,(\hat{\eta})\, XX^T \text{diag}\,(\hat{\eta})\,/M$ and $\hat{\eta} = E\hat{w}$. Covariates C are projected out of the system using the residual maker matrix $W = I - C(C^TC)^{-1}C^T$. Estimates of $\sigma^2_\beta, \sigma^2_\gamma, \sigma^2_e$ are obtained by solving the following system of linear equations

$$\begin{pmatrix} \text{tr}\,(WK_1WK_1W) & \text{tr}\,(WK_1WK_2W) & \text{tr}\,(WK_1W) \\ \text{tr}\,(WK_1WK_2W) & \text{tr}\,(WK_2WK_2W) & \text{tr}\,(WK_1W) \\ \text{tr}\,(WK_1W) & \text{tr}\,(WK_2W) & \text{tr}\,(W) \end{pmatrix} \begin{pmatrix} \sigma^2_\beta \\ \sigma^2_\gamma \\ \sigma^2_e \end{pmatrix} = \begin{pmatrix} y^TWK_1Wy \\ y^TWK_2Wy \\ y^TWy \end{pmatrix}$$

When applying RHE to estimate heritability of multiplicative GxE effects with $\hat{\eta}$, I make one adjustment to the basic method (see Section 1.3.2). The usual equation for converting variance components to heritability, given by

$$h_i^2 = \frac{\sigma_k^2}{\sum_k \sigma_k^2},$$

relies on the assumption that $\text{sum}(K_k) = 0$ and $\text{tr}\,(K_k) = N$ [36], which is satisfied when X_k is standardised to have column mean zero and variance one. Both η and X are standardised to have column mean zero and column variance one, so the $K_2(\hat{w})$ will also satisfy $\text{sum}(K_2) = 0$ and $\text{tr}\,(K_2) = N$ in expectation if the covariance between $\hat{\eta}$ and each SNP is zero. To satisfy $\text{sum}(K_2) = 0$ more explicitly, I include an intercept amongst the covariates that are projected out of the system. Finally $\text{tr}\,(K_2)$ is already estimated as part of the inference process. Thus variance components can be converted to unbiased heritability estimates using the equations

$$\hat{h}_G^2 = \frac{\frac{\text{tr}(K_1)}{N}\hat{\sigma}_G^2}{\frac{\text{tr}(K_1)}{N}\hat{\sigma}_G^2 + \frac{\text{tr}(K_2)}{N}\hat{\sigma}_{GxE}^2 + \hat{\sigma}_e^2},$$

$$\hat{h}_{\text{GxE}}^2 = \frac{\frac{\text{tr}(K_2)}{N}\hat{\sigma}_{GxE}^2}{\frac{\text{tr}(K_1)}{N}\hat{\sigma}_G^2 + \frac{\text{tr}(K_2)}{N}\hat{\sigma}_{GxE}^2 + \hat{\sigma}_e^2}$$

2.5 Single SNP hypothesis testing

Posterior mean estimates of $\hat{\beta}, \hat{\gamma}$ and $\hat{\eta} = E\hat{w}$ are obtained after convergence of the LEMMA variational inference algorithm. From these I construct residualised phenotypes following a Leave One Chromosome Out (LOCO) scheme;

$$y_{\text{resid-LOCO}} = y - C\hat{\alpha} - X_{\text{LOCO}}\hat{\beta}_{\text{LOCO}} - \text{diag}(\hat{\eta})X_{\text{LOCO}}\hat{\gamma}_{\text{LOCO}}.$$

X_{LOCO} denotes the genotype matrix excluding SNPs on the same chromosome of the test SNP, and β_{LOCO} and γ_{LOCO} are constructed similarly. Whilst it is computationally feasible to leave out only the region of interest, I use a LOCO scheme both for simplicity and to match current convention in the genetics literature[25, 100]. Using a LOCO scheme has been shown to increase power in LMMs as the effect of the test SNP is conditioned on the effects on a large proportion of the rest of the genome [25, 100].

For each imputed SNP, I then perform hypothesis tests $\beta_{\text{test}} \neq 0$ and $\gamma_{\text{test}} \neq 0$ using the linear model

$$\begin{aligned} y_{\text{resid-LOCO}} &= x_{\text{test}}\beta_{\text{test}} + (\hat{\eta} \odot x_{\text{test}})\gamma_{\text{test}} + \epsilon, \\ &= H\tau + \epsilon, \end{aligned}$$

where H be formed from the column-wise concatenation of $[x_{\text{test}}, \hat{\eta} \odot x_{\text{test}}]$ and τ is the corresponding vector of fixed effects. Assuming that ϵ has mean zero and covariance matrix Ω, the standard OLS estimator can be used

$$\hat{\tau} = (H^T H)^{-1} H^T y,$$

which (under certain regularity conditions) is asympototically normally distributed with mean τ and variance $\text{Var}\,(\hat{\tau})$. By assuming the residual phenotype is homoskedastic, that is that $\Omega = \hat{\sigma}^2 I$, the usual variance estimator is given by

$$\text{Var}\,(\hat{\tau}) = \hat{\sigma}^2 (H^T H)^{-1}.$$

Almli et al. [101] have previously observed that GxE interaction tests are likely to suffer from conditional heteroskedasticity, and hence the homoskedastic variance estimator is likely to underestimate the true variance [102]. I give a detailed explaination of this phenomenon in the next section.

To overcome this I use robust standard errors, alternatively called Huber-White, sandwich or heteroskedastic-consistent errors [103, 104], that are standard tools in economics [105] and have previously been proposed for use in GxE interaction studies [101, 106, 107]. I further include a small adjustment that reduces bias in small samples [108]. This yields the variance estimator

$$\text{Var}\,(\hat{\tau}) = (H^T H)^{-1} H^T \hat{\Sigma} H (H^T H)^{-1},$$

where $\hat{\Sigma}$ is a diagonal matrix with $\hat{\Sigma}_{ii} = \frac{\hat{\epsilon}_i^2}{(1-h_{ii})^2}$, where $\hat{\epsilon} = y - H\hat{\tau}$ and $h = H(H^T H)^{-1} H^T$. Hence our GxE test statistic is given by

$$\frac{\hat{\gamma}_{\text{test}}^2}{\text{Var}\,(\hat{\gamma}_{\text{test}})}$$

and under the null hypothesis is asympototically distributed as χ_1^2. As main effects tests are not sensitive to assumptions of heteroskedacity in the same way that GxE tests are [101], I use a simple t-test to test the hypothesis $\beta_{\text{test}} \neq 0$.

2.5.1 Robust standard errors in GxE Studies

Figure 1.1 shows how a multiplicative GxE interaction effect on a quantitative trait can cause the conditional trait variance given an interacting SNP $\text{Var}(Y|g_0)$ to differ according to the interacting SNPs genotype. This is known as conditional heteroscedasticity and is the key insight behind several recent methods to detect SNPs with non-zero GxE effects in the UK Biobank [68, 69].

In the same figure, we can observe that the conditional trait variance $\text{Var}(y|E)$ also displays signs of conditional heteroskedasticity. Previous studies [101] have observed that methods that assume homoskedasticity can display substantial inflation when testing for GxE effects at SNPs where there is no true GxE effect. In simulations, I observed that inflation of GxE tests statistics from LEMMA-S and the F-test, both of which assume homoskedasticity, increased with SNP-GxE heritability. Below I give an explanation for this phenomenon.

Consider a polygenic quantitative trait Y that has multiplicative GxE interactions with the same environmental exposure E at multiple SNPs

$$y_i = \alpha e_i + \sum_{j=1}^{M} \beta_j g_{ij} + \sum_{j=1}^{M} \gamma_j e_i g_{ij} + \epsilon$$

where the coefficients represent true effects. For simplicity I assume that E and SNPs G_j are normalised to have mean zero and variance one, that the set of E with all causal SNPs $\{E\} \cup \{G_j : \beta_j \neq 0\}$ is pairwise independent and that the influence from population structure is negligible.

Suppose E has been identified as an environmental variable that may plausibly have GxE interactions with the phenotype and we then conduct a GWAS for GxE effects. Then at the k'th SNP we wish to test the hypothesis $\gamma \neq 0$ in the following linear model

$$\begin{aligned} y &= \alpha_k E + G_k \beta_k + E \cdot G_k \gamma_k + u, \\ &= X\tau + u, \end{aligned}$$

where in the second line $\tau = (\alpha_k, \beta_k, \gamma_k)^T$ is the vector of coefficients from the model fit of SNP k, and X is the corresponding design matrix encapsulating all fixed effects. Here u is a random effect that captures

$$u = \sum_{j \neq k} (G_j \beta_j + E G_j \gamma_j) + \epsilon, \tag{2.23}$$

all of which are unobserved. Further assume that $\mathbb{E}[u|X] = 0$. Then the usual least squares estimate of τ, $\hat{\tau} = (X^T X)^{-1} X^T y$, has asympotic distribution

$$\hat{\tau} \to \mathcal{N}(\tau, \mathrm{Var}(\hat{\tau})),$$

where

$$\begin{aligned} \mathrm{Var}(\hat{\tau}) &= \mathbb{E}_X [\mathrm{Var}(\hat{\tau}|X)] + \mathrm{Var}_X (\mathbb{E}[\hat{\tau}|X]), \\ &= \mathbb{E}_X [\mathrm{Var}(\hat{\tau}|X)] + \mathrm{Var}_X (\tau), \\ &= \mathbb{E}_X [\mathrm{Var}(\hat{\tau}|X)], \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}(\hat{\tau}|X) &= \mathrm{Var}\left(\tau + (X^T X)^{-1} X^T u | X\right), \\ &= (X^T X)^{-1} \mathrm{Var}\left(X^T u | X\right) (X^T X)^{-T}. \end{aligned}$$

The usual approach is to assume homoskedasticity, which yields the standard variance estimator $\mathrm{Var}(\hat{\tau}|X) = \sigma^2 (X^T X)^{-1}$. However, given the functional form of u from Equation (2.23), we can show that the conditional variance is given by

$$\begin{aligned} \mathrm{Var}(u|E = e, G_k = g_k) &= \mathrm{Var}\left(\sum_{j \neq k} (G_j \beta_j + e G_j \gamma_j) + \epsilon\right), \\ &= \sum_{j \neq k} \mathrm{Var}(\beta_j G_j) + \sum_{j \neq k} \mathrm{Var}(\gamma_j e G_j) + 2 \sum_{j \neq k} \mathrm{Cov}(\beta_j G_j, \gamma_j e G_j) + \sigma_e^2, \\ &\quad + \sum_{j \neq k, m \neq k} \mathrm{Cov}(\beta_j G_j, \beta_m G_m) + \sum_{j \neq k, m \neq k} \mathrm{Cov}(\gamma_j e G_j, \gamma_m e G_m) \\ &= \sum_{j \neq k} (\beta_j + e \gamma_j)^2 + \sigma_e^2, \end{aligned}$$

where the covariances in the second line are all zero due to pairwise independance of the set $\{E\} \cup \{G_j : \beta_j \neq 0\}$ and $\mathrm{Var}(\epsilon) = \sigma_e^2$. Thus the conditional trait variance

will vary depending on the strength of environmental exposure either if there are a few SNPs with GxE interactions of large effect or if there are many SNPs with small yet non-zero interaction effects, and in either case homoskedasticity is unlikely to be an appropriate assumption. As noted earlier, the Huber-White variance estimator[104] is a standard way to correct for this issue[105].

2.6 Discussion

In this chapter I have suggested a new method for characterising gene by environment interactions in biobank scale datasets with multiple environmental variables. The method estimates a linear combination of those environments (which I refer to as an environmental score; or ES) that interacts with a polygenic risk score. Both the ES and the polygenic risk score are modelled jointly. Interaction weights used to construct the ES can then be interpreted to discover which environmental variables are involved in GxE interactions, or the ES can then for single SNP hypothesis testing or in heritability estimation.

LEMMA is designed to be a method that runs on biobank scale datasets. The per-iteration complexity, $\mathcal{O}(NM)$, is independent of the number of environmental variables; however, there are no theoretical guarantees on the rate of convergence. This is tested empirically in simulation in the next chapter. Use of OpenMPI means that RAM (complexity $\mathcal{O}(NM)$) is limited only by the size of the cluster. OpenMPI should also help to make LEMMA computationally tractable for biobank scale datasets. Again, this is tested empirically in the next chapter.

LEMMA uses a mixture of gaussian priors to model SNP effects, a gaussian prior to model covariate main effects and a gaussian prior to model the interaction weights. A mixture of gaussians prior on SNP main effects has been shown to improve power to detect associated loci in GWAS[75], depending of the genetic architecture of the trait in question. Through maximisation of the hyperparameters, LEMMA has the flexibility to approximate a gaussian prior on SNP effects if the mixture of gaussians prior is not appropriate. The primary purpose of the gaussian prior on interaction weights is to make the model identifiable rather than to provide much

shrinkage (due to the fixed prior variance). Future improvements could explore using a sparse prior, such as a spike and slab, to provide more shrinkage.

At the core of this method is a hypothesis that if two loci have GxE interaction effects and we know that which environment (or combination of environments) drive GxE interactions at the first, then the same (combination of) environment(s) is likely to be relevant in GxE interactions at the second locus. LEMMA takes this assumption to the extreme by assuming that any SNPs with GxE effects will interact with the same ES. The simplest way to relax this assumption would be to partition SNPs into groups (perhaps based on annotations) and learn a single ES for each group. Learning more than one ES per group would be a more complex generalisation; as it is not clear if additional modelling assumptions (such as orthogonality of the different ESs) would need to be enforced.

3

Simulation Study

Contents

This chapter starts by comparing the performance of LEMMA against three existing methods using synthetic data where the phenotype is simulated according to the LEMMA model. In the second half of this chapter, I use simulated data to evaluate how all four methods perform in scenarios of model misspecification; where the phenotype depends non-linearly on a heritable environmental variable.

3.1 Method Comparisons

I compared LEMMA with three existing single-SNP methods that jointly model multiplicative GxE interactions with multiple environments. The first, StructLMM [55], is a recently published method that adopts a linear mixed model framework to test for and characterise loci that interact with multiple environmental variables. The second and third methods, which I refer to as the F-Test and the robust F-Test, similarly test for loci that interact with multiple environmental variables, but use standard techniques from linear modelling theory [105].

3.1.1 StructLMM

StructLMM [55] is a single SNP method that builds upon Linear Mixed Modelling theory used commonly in GWAS, using the random effects term u that is normally used to model genetic similarity to instead model environmental similarity between participants. As StructLMM does not account for genetic similarity with a kinship matrix, population structure is controlled for by including genetic PCs as covariates. This means that StructLMM does not benefit from the usual power boost that conventional LMMs applied to GWAS get from implicitly conditioning on other causal loci[25]. The StructLMM GxE interaction model is given by

$$y \sim \mathcal{N}(C'\alpha' + x_{\text{test}}\beta, \sigma^2_{\text{GxE}}\text{diag}\,(x_{\text{test}})\,\Sigma\text{diag}\,(x_{\text{test}}) + \sigma^2_e\Sigma + \sigma^2_n I), \tag{3.1}$$

where C' is the matrix of covariates with fixed effects α', x_{test} is an N-vector containing the dosage of the focal variant, $\Sigma = EE^T$ is the $N \times N$ environmental similarity matrix and E is the $N \times L$ matrix of environmental variables. For continuity with previously introduced notation, I use C' and α' to distinguish between the matrix of covariates including environmental main effects introduced as C in Section 2.1 and the covariate matrix that excludes environmental main effects used here.

StructLMM then tests the null hypothesis $\sigma^2_{\text{GxE}} = 0$, using the score-based variance component testing framework proposed by the Sequence Kernal Association Test (SKAT) method [109]. This approach is computationally advantageous as

it only requires the null model to be fitted. However even fitting the null model is prohibitively expensive in large samples due to the dense $N \times N$ covariance structure. Hence the authors additionally adopt methodological developments from the Fast-LMM method [100] which leverages that fact that Σ is low rank to perform a single spectral decomposition that allows hypothesis testing to scale linearly in the number of samples.

Finally I note that although StructLMM is also able to perform a joint test that jointly tests for non-zero main and interaction effects at each SNP, I only use the interaction test in the following section for compatibility with the other methods.

3.1.2 F-Test

The StructLMM model in Equation (3.1) is mathematically equivalently to the Bayesian linear regression given by

$$y = \underbrace{C'\alpha' + x_{\text{test}}\beta_{\text{test}}}_{\text{Fixed effects}} + \underbrace{E\alpha_E + x_{\text{test}} \odot Ew + \epsilon}_{\text{Random effects}}, \tag{3.2}$$

$$\alpha_E \sim \mathcal{N}(0, \frac{\sigma_E^2}{L}), \tag{3.3}$$

$$w \sim \mathcal{N}(0, \frac{\sigma_{\text{GxE}}^2}{L}), \tag{3.4}$$

$$\epsilon \sim \mathcal{N}(0, \sigma_e^2 I). \tag{3.5}$$

Dropping the priors on α_E and γ allows Equation (3.2) to be written as a linear regression model

$$y = C\alpha + x_{\text{test}}\beta_{\text{test}} + \text{diag}(x_{\text{test}})E\gamma + \epsilon, \tag{3.6}$$

where ϵ is the (unobserved) residual vector and $C = [C', E]$.

The alternate hypothesis $H_0 : \gamma \neq \vec{0}$ jointly tests for evidence that the combination of L environments have significant multiplicative GxE effects, and is analgous to the interaction test used by StructLMM. For notational convenience let H be formed from the column-wise concatenation of $[C, x_{\text{test}}, \text{diag}(x_{\text{test}})E]$ and τ is the corresponding vector of fixed effects. Assuming y is homoskedastic so

that ϵ is Gaussian with mean zero and covariance $\sigma_e^2 I$ means that the standard F-test can be used. The test statistic is given by

$$F_{\text{test}} = \frac{(R\hat{\tau})^T (R(H^T H)^{-1} R^T)^{-1} (R\hat{\tau})/L}{\hat{\sigma}^2},$$

where R is a binary matrix such that $R\tau = \gamma$ and $\hat{\tau}$ is the least squares estimator given by $(H^T H)^{-1} H^T y$. Under the null hypothesis $\gamma = \vec{0}$, F_{test} follows an $F_{d_1 - d_0, N - d_1}$ distribution, where d_1 is the column rank of H and d_0 is the column rank of H under the null hypothesis.

Moore et al. [55] do compare against the linear model given in Equation (3.6), but use different inference procedures. In their first approach, which the authors refer to as 'SingleEnv-Fenv', each environment is tested indepedently for GxE association with the alternate hypothesis $w_l \neq 0$. A joint p-value that corresponds to an alternative hypothesis of at least one environment having significant GxE effects is then constructed by taking the smallest p-value (from testing $w_l \neq 0$) and performing a Bonferroni correction. As one might expect, this incurred a substantial multiple testing burden and performed poorly against StructLMM[55], especially when a combination of several environments is required to represent the true environment driving the interation. For their second approach, refered to as 'MultiEnv-Fenv', the authors use both the score and the likelihood ratio tests (LRTs) to test the joint alternate hypothesis $w \neq \vec{0}$ using L degrees of freedom. The authors results show that 'MultiEnv-Fenv' was miscalibrated, but that calibration improved as sample size increased, in simulations with $1000, 2000$ and 5000 samples. I note that the score and LRTs depend on asymptotic results (ie the null distribution holds as sample size tends to infinity) whereas the F-Test is analytic. Further, 5000 samples is small compared to sample sizes available in modern genetic datasets such as the UK Biobank.

3.1.3 Robust F-test

The theoretical argument presented in Section 2.5.1 suggests that when performing hypothesis tests with single SNP GxE effects, the residual variance will become

increasingly heteroskedastic as GxE heritability increases. Thus assuming that $\text{Var}(\epsilon) = \sigma_e^2 I$ may be unwise. Instead, the robust Huber White[104] variance estimator can be used (in a similar manner to that of the LEMMA test statistic).

Let $\text{Var}(\epsilon) = \Omega$. Then under the null hypothesis, the least squares estimator $\hat{\gamma} = R(H^T H)^{-1} H^T y$ has asympototic distribution

$$\hat{\gamma} \sim \mathcal{N}(0, \text{Var}(\hat{\gamma})),$$

where $\text{Var}(\hat{\gamma}) = R(H^T H)^{-1} H^T \Omega H (H^T H)^{-1} R^T$. Using the Huber-White[103, 104] variance estimator results in the test statistic given by

$$F_{\text{robust}} = (R\hat{\tau})^T (R(H^T H)^{-1} H^T \hat{\Omega} H (H^T H)^{-1} R^T)^{-1} (R\hat{\tau}),$$

where $\hat{\Omega}$ is a diagonal matrix with $\hat{\Omega}_{ii} = \frac{\hat{\epsilon}_i^2}{(1-h_{ii})^2}$, $\hat{\epsilon} = y - H\hat{\tau}$ and $h = H(H^T H)^{-1} H$. Under the null hypothesis, F_{robust} is asympototically distributed as $\chi^2_{d_3}$ where d_3 is the rank of HR^T. Although this technically does not make use of the F-distribution, I refer to this as the robust F-test to make clear that it is analagous to the F-test but uses slightly different assumptions.

3.2 Baseline simulations

3.2.1 Data simulation

In all simulation studies I used genetic data subsampled from genotyped SNPs in the UK Biobank[110], drawing SNPs from all 22 chromosomes in proportion to chromosome length and using unrelated samples of mixed ancestry ($N = 25k$; 12500 white British, 7500 Irish and 5000 white European, $N = 50k$; 29567 white British, 7500 Irish and 12568white European, $N = 100k$; 79567 white British, 7500 Irish and 12568 white European; using self-reported ancestry in field *f.21000.0.0*). All samples were genotyped using the UKBB genotype chip and were included in the internal principal component analysis performed by the Uk Biobank. Environmental variables were simulated from a standard Gaussian distribution.

Phenotypes for the baseline simulations were all simulated according to the LEMMA model. More formally, given a set of parameters $\{h^2_\beta, h^2_{\text{GxE}}, N, M, M_{\text{causal, G}}, M_{\text{causal, GxE}} L, L_{\text{active}}\}$ each phenotype was constructed as

$$\begin{aligned} y|\alpha_{PC_1}, \beta_1, \beta_2, \gamma_1\gamma_2, \eta, \epsilon = PC_1\alpha_{PC_1} + X(\beta_1 + \beta_2) \\ + \text{diag}\,(\eta)\, X(\gamma_1 + \gamma_2) + \epsilon, \\ \eta|w = Ew, \\ \epsilon \sim \mathcal{N}(0, I), \end{aligned}$$

where X represents the $N \times M$ genotype matrix after columns have been standardised to have mean zero and variance one, PC_1 is the first principle component of the genotype matrix and E is the $N \times L$ matrix of environmental variables. In all simulations α_{PC_1} was set such that $PC_1\alpha_{PC_1}$ explained one percent of trait variance.

Non-zero elements of the interaction weight vector w were drawn from a decreasing sequence

$$w_i = \begin{cases} (-1)^i(1 - \frac{i}{2L_{\text{active}}}) & i \leq L_{\text{active}}, \\ 0 & o/w. \end{cases}$$

One advantage of modelling the interaction effects of multiple environmental variables jointly is that it should allow researchers to use many environmental variables that are plausibly relevant, with the intention that all environmental variables that truely have GxE interactions are included in the analysis. Varying the number of environmental variables with a non-zero interaction weight, which are referred from here as *active* environmental variables, thus allows the ability of LEMMA to distinguish *active* environmental variables from *inactive* ones to be tested.

Background genetic effects, denoted by β_1 and γ_1, were simulated from a spike and slab prior such that the number of non-zero elements was governed by $M_{\text{causal, G}}$ and $M_{\text{causal, GxE}}$ for main and interaction effects respectively. Non-zero elements were drawn from a standard Gaussian, and then subsequently rescaled to ensure that the heritability given by main and interaction effects was h^2_β and h^2_γ respectively. SNPs with non-zero main effects are referred to as causal through this simulation study.

In a similar fashion to Loh *et al.*[75], an additional 60 SNPs with non-zero effects were included in vectors β_2 and γ_2. Non-zero coefficients in these vectors standardised to each explain 0.00016% of trait variance, thus allowing direct power comparisons across different scenarios.

All non-zero effects were drawn from SNPs in the first half of each chromosome, leaving 'null' SNPs on the second half of each chromosome to quantify inflation in test statistics under the null hypothesis.

The default parameter choices were: $N = 25K; M = 100K; L = 30; L_{\text{active}} = 6, M_{\text{causal, G}} = 2500; M_{\text{causal, GxE}} = 1250; h^2_\beta = 20\%; h^2_\gamma = 5\%$.

The first principal component was provided as a covariate to all methods.

3.2.2 Simulation results

Inference of the Environmental Score

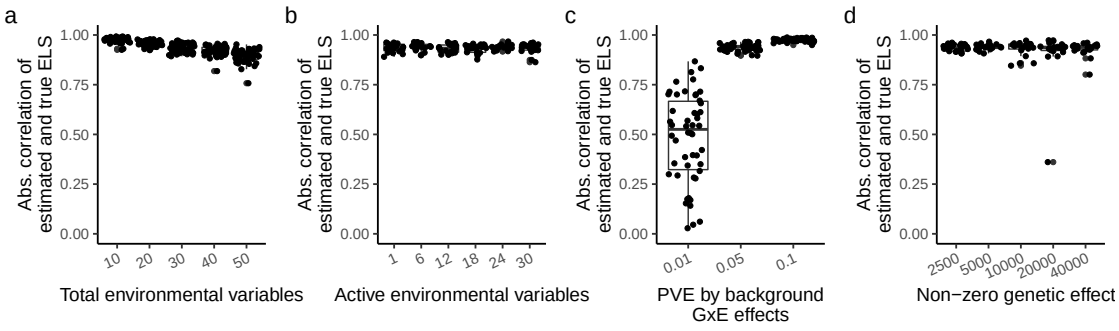

Figure 3.1: Estimation of the ES in simulation. Absolute correlation of the inferred ES with the true ES while varying (a) the total number of environmental variables, (b) active environmental variables, (c) proportion of variance explained by GxE effects, (d) the number of non-zero SNP effects. Simulations parameters used were the defaults given in Section 3.2.1. Abbreviations; PVE, proportion of variance explained.

One core aim of LEMMA is to collapse multiple environmental variables down to a single linear combination; referred to as the Environment Score (ES). The absolute correlation between the simulated ES and the inferred ES, which acts as a useful proxy for the accuracy of LEMMAs estimate of the ES, is shown in Figure 3.1. Simulation results showed that the correlation between the true ES and the inferred ES was largely invariant to the number of active environmental variables, and suffered only mild degradation as the number of environmental variables increased

from 30 to 50. Reducing the heritability of interaction effects will intuitively make inference harder, and using $N = 25K$ samples the mean absolute correlation with the true ES dropped from 0.938 at $h^2_{\text{GxE}} = 0.05$ to0.484 at $h^2_{\text{GxE}} = 0.01$.

Comparison of GxE interaction results

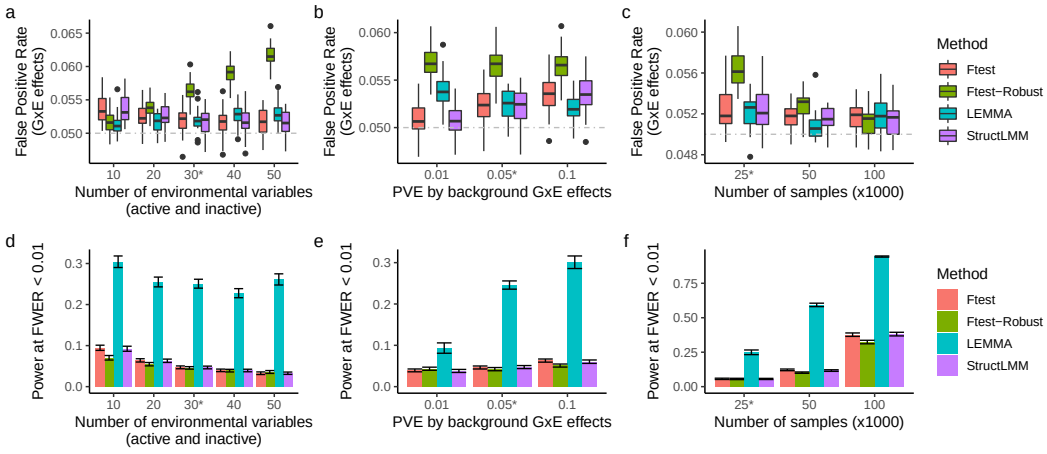

Figure 3.2: Type I error and Power of tests to detect GxE effects in simulation. (a-c) Comparison of false positive rate (FPR) as (a) the number of environments increases, (b) phenotype variance explained by GxE effects increases or (c) sample size increases. (e-f) Analagous comparison of the power to detect GxE interations. Default parameter values denoted by stars. FPR was assessed a the proportion of 'null' SNPs with p-value < 0.05. Power was assessed as the proportion of the 60 causal SNPs of standardised effect detected at a p-value threshold of 0.01. Results from 20 simulations shown.

Power and statistical calibration of the GxE interaction test is shown in Figure 3.2 for each of the four methods, while the number of environmental variables, background GxE heritability and sample size were varied. Power was assessed as the proportion of the 60 SNPs of standardised effect with p-value less than 0.01, and statistical calibration was assessed using False Positive Rate (FPR; the proportion of 'null' SNPs with a p-value less than 0.05).

In all scenarios LEMMA achieved a substantial power boost, suggesting that LEMMA gains a significant advantage from leveraging the assumption that SNPs with GxE interactions interact with the same linear combination of environments when the assumption holds. The LEMMA power increase occurs due to two sources. First, by collapsing multiple environmental variables down to a single linear combination, the LEMMA GxE interaction test statistic has one degree of

freedom rather than L. Second, by modelling the effects all SNPs jointly, LEMMA is able to reduce phenotypic variance (a feature common to all LMMs in genetics).

The standard F-test and StructLMM appeared to have equivalent power and false positive rate (FPR) in large sample sizes (although previous work suggests that StructLMM may outperform the F-test in small samples [55]). It is notable that the FPR of both of these methods seemed to increase with the background GxE heritability (Figure 3.2b). As both of these methods assume that the residual variance is homoskedastic, this result is consistent with the theory presented in Section 2.5.1 which suggests that heteroskedasticity in the phenotype variance will increase as a function of h^2_{GxE}.

Interestingly calibration of the robust F-test appeared to suffer as the number of environments increased (**Figure 3.2a**) and improve as the number of samples increased. This is consistent with the notion that the robust F-test is asymptotically correct, but not efficient, as described on page 193 of Greene [105].

In general, LEMMA controlled the FPR at least as well as other methods. The one scenario when LEMMA displayed slightly elevated FPR was when $h^2_{\text{GxE}} = 2\%; N = 25k$. This is the same scenario where LEMMA had poor ability to uncover the true ES; suggesting that there might be some possibility of overfitting with large L and a (relatively) small number of samples. When sample size was large however (N=100k; Figure 3.2c), all methods had reasonable control of FPR.

Comparison of main effects results

Figure 3.3 compares the power and statistical calibration of LEMMA, the F-test and BoltLMM to detect SNP main effects. Figure 3.3(a) suggests that LEMMA achieves similar control of FPR to BoltLMM; which also tests SNPs for association against a residualised phenotype. Both methods also achieve a slight power increase over the F-test, which is consistent with the notion that jointly modelling SNPs allows LEMMA and BoltLMM to reduce the phenotypic variance and thus achieve a power increase. One might expect LEMMA to have greater power as it has the opportunity to reduce phenotypic variance further by accounting for multiplicative GxE interactions,

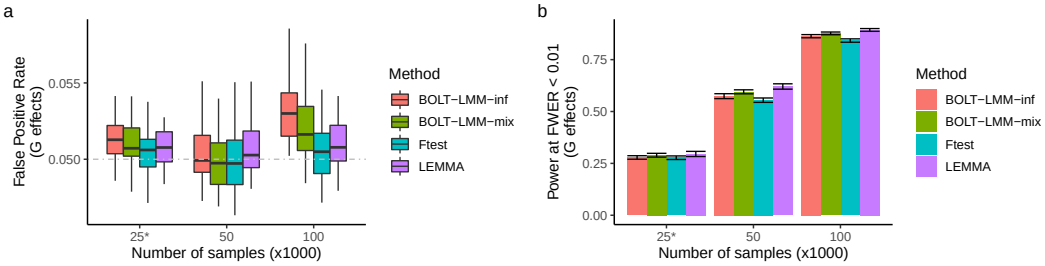

Figure 3.3: Type I error and Power of tests to detect main effects in simulation. Comparison of (a) false positive rate and (b) power as the sample size increases. FPR was assessed a the proportion of 'null' SNPs with p-value < 0.05. Power was assessed as the proportion of the 60 causal SNPs of standardised effect detected at a p-value threshold of 0.01. Results from 20 simulations shown.

however the heritable contribution of GxE effects in this simulation is relatively small so this effect would be subtle.

These results serve as a useful common sense check, but are not regarded as primary results for LEMMA whose main purpose is to identify GxE interactions.

Heritability estimation

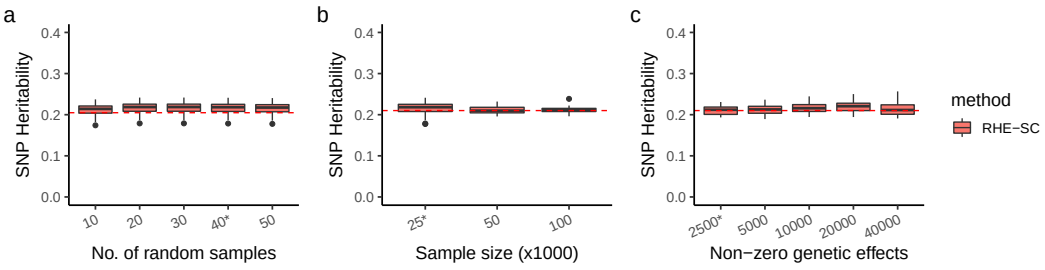

Figure 3.4: Heritability of additive genetic effects. Estimates of additive genetic heritability whilst varying the number of random draws (a), the number of samples (b) and trait polygenicity (c). The red dotted line denotes the true SNP-G heritability used whilst constructing the simulation. Stars denote default values of simulation parameters. Results for 20 replicates of each scenario shown.

Figure 3.4 shows estimates of the additive genetic heritability using my own implementation of randomised HE-regression, whilst varying the number of random draws used, sample size and trait polygenicity. As expected, higher sample size leads to more precise estimates and RHE regression is relatively robust to the number random draws[86]. Although HE-regression makes an implicit assumption that all SNPs contribute to the kinship matrix equally, RHE appeared quite robust

to reducing the polygenicity of the trait architecture (similar to observations made previously about GCTA-GREML [36]; see Figure 3.4(c)).

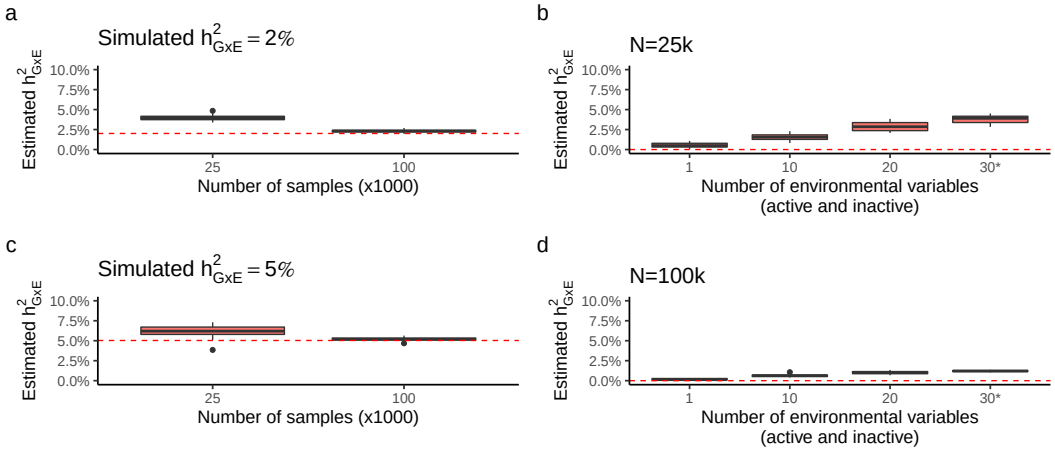

Figure 3.5: Estimation of GxE heritability in simulation. Estimates of SNP-GxE heritability whilst varying the number of environmental variables (b, d) and sample size (a, c). Simulations in the left column had $L = 30$ environmental variables, the red dotted line denotes the true SNP-GxE heritability used whilst constructing the simulation. We observed some upwards bias as the number of environmental variables increases (b, d), which is ameliorated with increased sample size (d). Phenotypes were constructed using $M = 100,000$ SNPs with $M_{\text{causal, G}} = 80,000$ causal main effects and $M_{\text{causal, GxE}} = 40,000$ causal interaction effects. Results for 20 replicates of each scenario shown.

When estimating the GxE heritability of the LEMMA ES using randomised HE regression with a single component (RHE-SC), I observed some upward bias (see **Figure 3.5**). The upwards bias increased with the number of environmental variables, was mitigated with increased sample size and was not present for heritability estimates of SNP main effects. This suggests that the upwards bias is a result of overfit of the LEMMA ES.. In twenty simulations with $L = 30$ environmental variables, $N = 100$k samples and true GxE heritability of 5% we observed mean GxE heritability of 5.2% (Figure 3.5(c)).

I then compared the performance of randomised HE-regression with a single component (RHE-SC) and with SNPs partitioned according to MAF and LD score (RHE-LDMS). As no code was provided by Pazokitoroudi et al. [87] for a multi-component analysis, I only show results from my own implementations. Simulation results using $N = 25k$ samples are shown in **Figure 3.6**. As expected, when causal

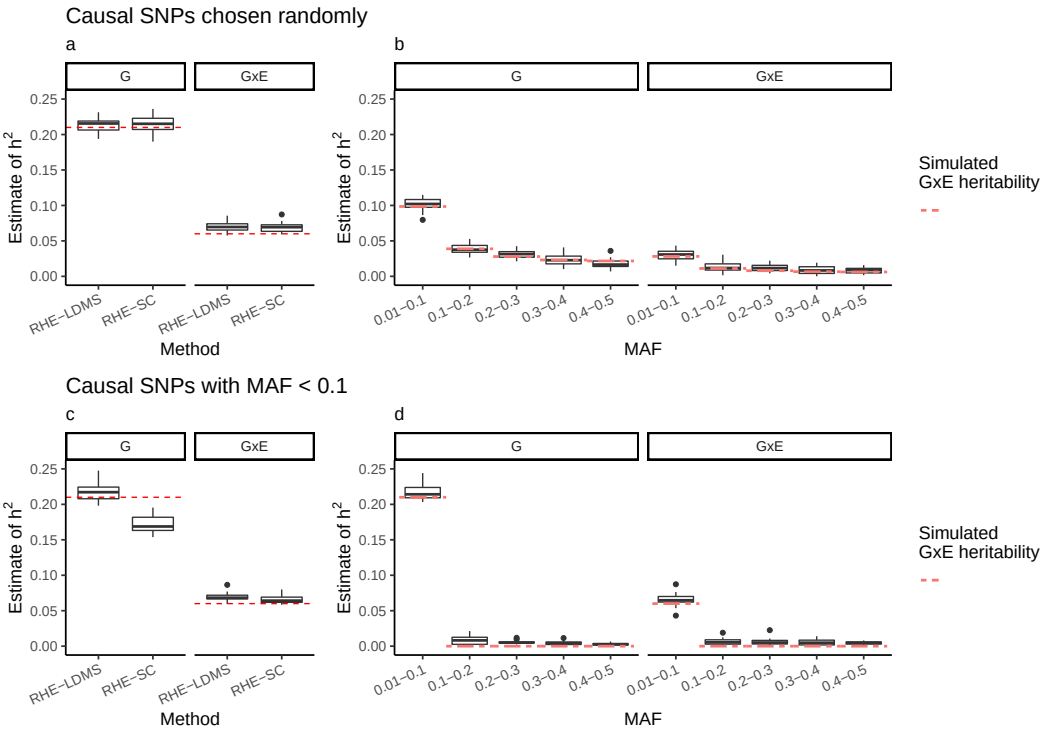

Figure 3.6: Heritability estimates stratified by LD and MAF in simulation. Comparison of heritability estimates using RHE-SC and RHE-LDMS when causal SNPs were drawn (a) at random or (c) only from rare (MAF < 0.1) SNPs. Heritability estimates (using RHE-LDMS) stratified by MAF when causal SNPs were drawn (b) at random or (d) only from rare (MAF < 0.1) SNPs. Simulations performed with $N = 25K$ samples, $M = 100K$ SNPs and the default baseline simulation parameters. Abbreviations; MAF, minor allele frequency; RHE-SC, randomised HE regression with a single SNP component[86]; RHE-LDMS, multi-component randomised HE-regression[87].

SNPs were chosen at random estimates RHEreg-SC were unbiased whereas when only rare SNPs were causal (MAF< 0.1) estimates from RHEreg-SC were biased downwards. In contrast RHEreg-LDMS was unbiased in both scenarios and was able to appropriately attribute the relative heritability for each SNP component.

Computational complexity

Figure 3.7(a, b) displays the strong scaling properties[1] of the LEMMA variational inference algorithm using OpenMP and OpenMPI architectures. Parallelising with OpenMP seems to give little improvement beyond two cores; intuitively this is because the LEMMA variational inference algorithm is dominated by BLAS level 1

[1]Strong scaling refers to how runtime varies with the number of cores whilst keeping data size constant (eg. the number of samples)

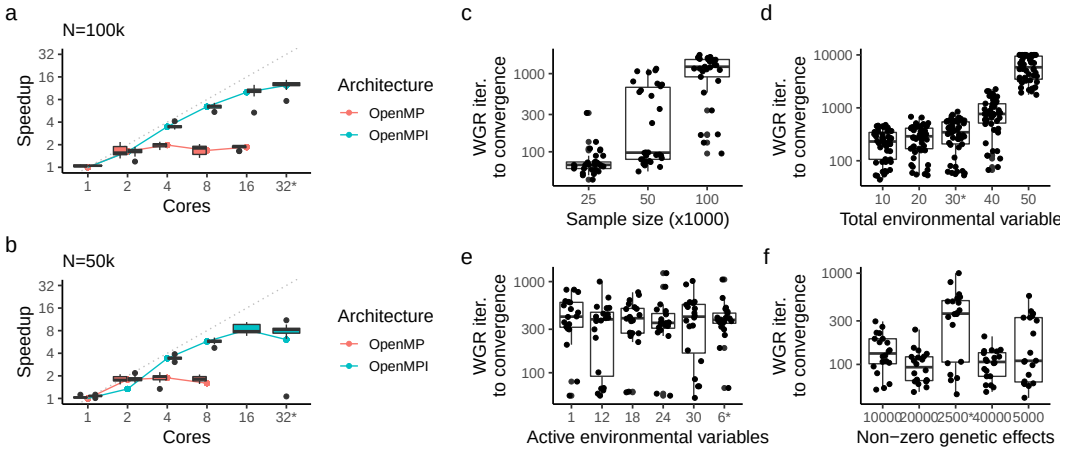

Figure 3.7: Computational complexity of LEMMA in simulation. Strong scaling of LEMMA WGR algorithm using MP and OpenMPI architectures with (a) $N = 50k$ samples and (b) $N = 100k$ samples. Comparison of iterations to convergence of the LEMMA WGR algorithm whilst varying the number of samples, the number of environments, the number of active environments and trait polygenicity. Default parameter values denoted by stars (default sample size was $N = 25k$). Results are shown for 20 replicates in each scenario. ŵhen using more than 16 cores the OpenMPI implementation was distributed across several nodes, otherwise it used only a single node.

and BLAS level 2 operations which are memory-bound rather than compute-bound. In contrast, partitioning samples with OpenMPI ensures that LEMMA makes best possible use of the available bandwidth from RAM to processors. Figure 3.7(a, b) suggest that parallelising with OpenMPI is highly efficient (ie doubling the number of cores almost doubles computational speed) up to the number of samples per core is a couple of thousand, after which adding extra cores yields diminishing returns. Using OpenMPI has the additional advantage of allowing users to utilise cores from across a cluster rather than being restricted to a single node.

Figure 3.7(c, d, e, f) show how the number of iterations required for convergence scales with sample size, the number of environments, trait polygenicity and the number of active environments. Loh *et al.*[75] have previously noted that for BOLT-LMM the number of iterations to convergence scales roughly as $N^{0.5}$, however, this seems too conservative for our LEMMA (see Figure 3.7(c)).

The number of iterations LEMMA required for convergence appeared to increase rapidly as a function of L (possibly polynomial scaling). This suggests that LEMMA is appropriate for running with tens of environmental variables, but

further methodological changes would be necessary to run LEMMA with hundreds of environments. Finally, iterations to convergence appeared to be independent of the number of active environmental variables and decreased slightly with increased polygenicity. The later possibly reflects the fact that a polygenetic trait with $10,000$ causal variants is closer to the mixture of Gaussians prior than a trait with 2500 causal variants.

3.2.3 Discussion

The primary result from this simulation study is that, if the assumptions behind the LEMMA model hold (namely that the same ES is driving GxE iterations with SNPs across the genome) then LEMMA can leverage these assumptions to gain a substantial power increase to detect SNPs with true GxE interactions over competing methods. I also show that, in general, LEMMA is at least as well calibrated as competing methods and is able to estimate the proportion of variance explained by GxE effects with a small amount of bias (that is ameliorated with increased sample size).

I also show that the F-test, which relies on well established linear modelling theory, behaves similarly to a recently published method StructLMM in large datasets. I confirm that both of these methods are likely to have spurious inflation in traits with non-negligible GxE heritability. Finally, I suggest a robust alternative to the F-Test, which also relies on standard results from linear models. In large datasets, this appears appropriate for use with tens of environmental variables but has poor calibration in smaller datasets.

3.3 Model misspecification simulations

The theory presented so far has assumed that the functional form of the relationship between phenotype and environmental variables is linear. In practice, there is no guarantee that relationships between measures of the environment in real datasets will be quite so convenient (for example age has a non-linear effect on complex traits such as blood pressure, and GWAS of blood pressure often adjust for both age and

age[2][111–113]). Previous work in GxE studies[106] has shown that misspecifying the functional form of the environment in a simple linear test of GxE interaction (ie the F-Test) can cause inflation. If the environment is independent of the genotypes, then this inflation is caused only be heteroskedasticity and can be solved with a robust variance estimator[106]. Correlation between the environment and genotypes, however, will cause bias in the least squares estimator and thus robust variance estimators are not sufficient [105].

Many environmental variables however have a genetic basis and thus will not be independent of genotypes. For example, drinking status can be used as an environmental variable and has been implicated in GxE interactions for esophageal squamous-cell carcinoma in Chinese populations[61, 114]. Drinking status is also known to be at least somewhat heritable as individuals who carry the ALDH2*2 allele are less likely to drink regularly or heavily[115] due to 'the alcohol flushing response'. In this section I perform simulations to characterise the bias in GxE tests from LEMMA, StructLMM, the F-test and the robust F-test caused when a phenotype depends on the non-linear (squared) effects of a heritable environmental variable[2]. I also suggest that relatively smooth non-linearities, such as squared effects, can be easily detected with standard regression methods in a preprocessing step and subsequently included as covariates.

3.3.1 Data simulation

The misspecified environmental variabled denoted by S was generated according to the model

$$S|\beta_s, \epsilon_s = X\beta_s + \epsilon_s,$$
$$\epsilon_s \sim \mathcal{N}(0, I),$$

where SNP effects β_s had a spike and slab prior

$$\beta_{s,j}|v_j \sim v_j\mathcal{N}(0, \sigma_s^2) + (1 - v_j)\delta_0(\beta_{s,j}),$$
$$v_j \sim \text{Ber}(\lambda_s).$$

[2]I thank Prof. Richard Durbin, who suggested this line of enquiry when this work was presented at the Bertinoro Computational Biology conference 2018 by Prof. Jonathan Marchini

Let y_{baseline} denote the baseline phenotype simulated using the procedure described in Section 3.2.1. Then the misspecified phenotype used in this simulation study was constructed as

$$y = \alpha_s S^2 + y_{\text{baseline}}.$$

SNP effects were sampled such that the causal SNPs for the misspecific environmental variable S were sampled from the first half of chromosomes $1-11$ only whilst causal SNPs for y were sampled from the first half of chromosomes $12-22$. This meant that it was possible to quantify the inflation in test statistics amongst SNPs with no effect on either trait and amongst SNPs that were causal for S but had no true effect on y separately.

By default S was simulated with 2500 causal SNPs and a heritability of 30%. The default parameter choices for the baseline phenotype were: $N = 25K; M = 20K; L = 5; L_{\text{active}} = 2, M_{\text{causal, G}} = 2500; M_{\text{causal, GxE}} = 1250; h^2_\beta = 30\%; h^2_\gamma = 5\%$, and a range of values for α_s was used to vary the strength of the non-linear relationship between y and S.

3.3.2 Simulation results

Initially, I tested the performance of LEMMA in simulation with no attempt to control for non-linear dependency between S and the phenotype (referred to as (-SQE) in the figures). Results shown in Figure 3.8 confirmed that this form of confounding does lead to inflated FPR at heritable sites of S (but not at null SNPs) and that the strength of the inflation varies as a function of the strength of the non-linear dependency (scale parameter α_s). However the power of standard regression techniques to identify squared effects should also increase as a function of the strength of the non-linear dependence. Therefore I investigated two simple processing strategies; one where significantly associated squared effects were first regressed out of the phenotype, and one where significantly associated squared effects were included as additional covariables. Finally, I have included results from running

LEMMA on the baseline phenotype using the same parameter settings as a control. These are referred to as (rmSQE), (+SQE) and (control) respectively in the figures.

Figure 3.8 shows the FPR of the GxE test statistics at null SNPs, the FPR of the GxE test statistics at heritable sites of S and power to detect true GxE interactions from all of the competing methods. Figure 3.8(b) shows that this form of model misspecification does cause LEMMA(-SQE) to have inflated GxE test statistics at heritable sites of S but not at null SNPs. This is consistent with the results of Tchetgen and Kraft [106], who showed that robust variance estimators can control for misspecification of the functional form of the environment but only if the environment is not correlated with the genotypes. Model misspecification of this form also causes LEMMA(-SQE) to place too much weight on S in the ES (Figure 3.9) which then results in a loss of power (Figure 3.8(c)).

Figure 3.8 also suggests that regressing detected squared environmental effects from the phenotype [LEMMA(-SQE)] mitigates the inflation in the FPR and increases power, but still has reduced power compared to the control. Including detected squared effects in the model achieves well-calibrated GxE test statistics and similar power to that of the control. As the cost of the (+SQE) procedure is small, I use the (+SQE) strategy for LEMMA by default for all analyses of real data.

Finally I compared the calibration and power of LEMMA, StructLMM, the F-Test and the robust F-test, with and without testing for significant squared effects. Simulation results shown in Figure 3.10 suggest that all methods are vulnerable to this form of model misspecification and that the (+SQE) strategy is effective at controlling for it. The difference in calibration between the F-test-(SQE) with and without robust variance estimators suggests that part, but not all, of the inflation in this scenario is due to heteroskedasticity; consistent with theory[105, 106].

3.3.3 Discussion

In this simulation study, I confirmed that if the phenotype depends non-linearly on a non-heritable environmental variable, then methods using robust variance estimators

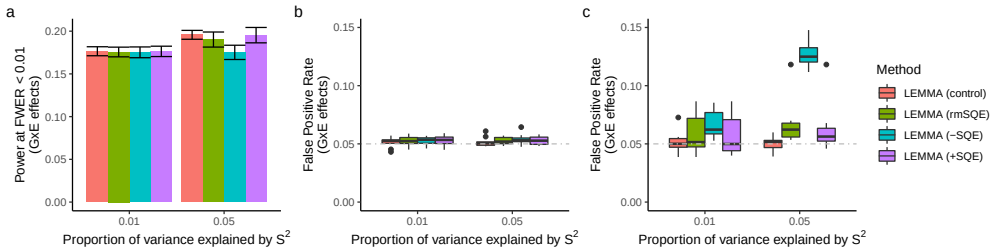

Figure 3.8: Comparison of three preprocessing strategies to control for a misspecified environmental variable. (a) Power to detect true GxE effects, and false positive rate of the GxE test statistic at (b) null SNPs and (c) heritable SNPs of S with no GxE effect. Results from 20 simulations shown. Abbreviations; (-SQE), no attempt to control for squared effects; (+SQE), squared effects with $p < 0.01$ (Bonferroni correction for multiple envs) included as covariates; (rmSQE), squared effects with $p < 0.01$ (Bonferroni correction for multiple envs) regressed from the phenotype; (control), $y = \alpha_s S + y_{\text{baseline}}$ used in phenotype construction.

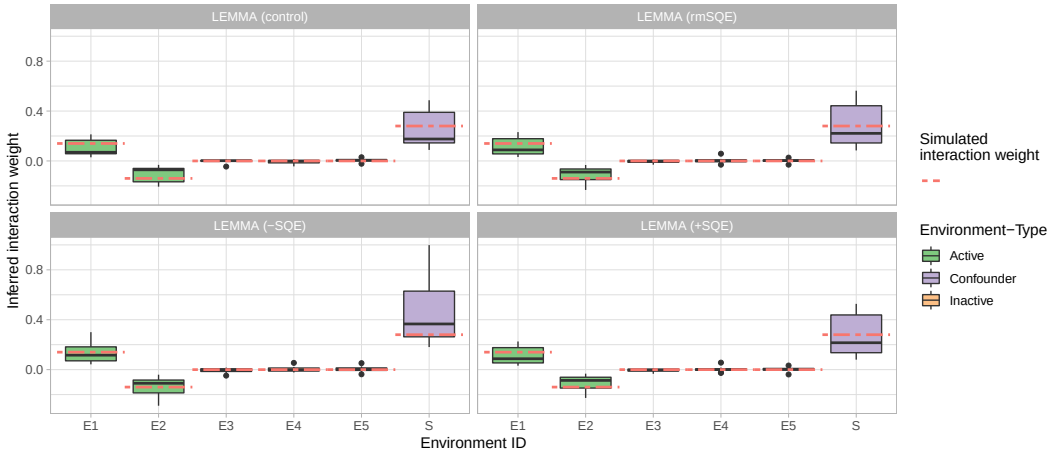

Figure 3.9: Average interaction weights inferred by LEMMA under different three preprocessing strategies. Results from 20 simulations shown. Abbreviations; (-SQE), no attempt to control for squared effects; (+SQE), squared effects with $p < 0.01$ (Bonferroni correction for multiple envs) included as covariates; (rmSQE), squared effects with $p < 0.01$ (Bonferroni correction for multiple envs) regressed from the phenotype; (control), $y = \alpha_s S + y_{\text{baseline}}$ used in phenotype construction.

will still be well calibrated. However when the environmental variable is heritable, all GxE methods will be miscalibrated at heritable sites of the environment.

To mitigate the problem, I suggest that when the non-linear effects are strong enough to be problematic, they are also strong enough to be detectable with standard regression methods. Thus non-linear dependencies (particularly if the relationship is squared) can be detected and controlled for with a simple preprocessing step. One caveat of this study is that I have not explored cases where the non-linearity

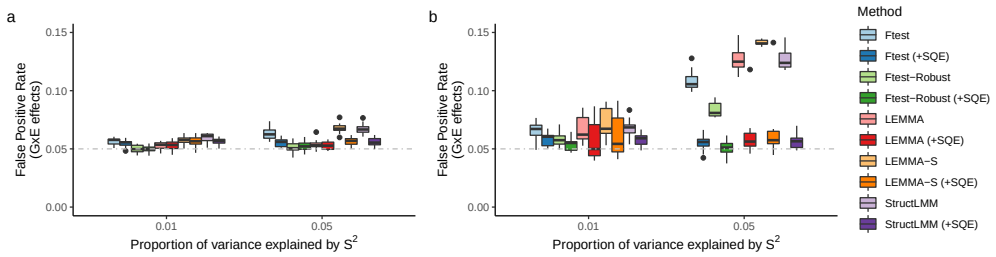

Figure 3.10: Comparison of false positive rate using the (-SQE) and (+SQE) preprocessing strategies for LEMMA, StructLMM, the F-Test and the robust F-Test. False positive rate of the GxE test statistic at (b) null SNPs and (c) heritable SNPs of S with no GxE effect. Results from 20 simulations shown. Abbreviations; (-SQE), no attempt to control for squared effects; (+SQE), squared effects with $p < 0.01$ (Bonferroni correction for multiple envs) included as covariates; (rmSQE).

has a functional form other than a squared effect. In cases where users suspect a different functional form of the non-linearity, this could be adopted very easily. Use of spline functions to model non-linear effects could be a more robust approach, however, this is reserved for future work.

Finally, I note that both simulation studies have assumed that all GxE interaction effects occur with a constant ES. It would be interesting to conduct a (perhaps more realistic) simulation study where the ES used in each GxE interaction is a random effect that occurs with some non-zero mean and variance. However this study is outside the scope of this thesis.

4

GxE analysis in the UK Biobank

Contents

4.1 Introduction

In the previous chapter, I showed simulation results characterising the ability of LEMMA to identify SNPs with GxE interactions and comparing LEMMA against several competing single SNP methods.

Until recently, searching for GxE interactions with multiple environmental variables has been difficult due to the lack of large datasets with high dimensional

phenotyping. The recently released UK Biobank [116] dataset, a large population cohort study with deep genotyping and sequencing and extensive phenotyping, offers a unique opportunity to explore GxE effects [49, 55, 58, 68, 117–119]. Therefore in this chapter, I used LEMMA to search for GxE interactions in body mass index (BMI), systolic blood pressure (SBP), diastolic blood pressure (DBP) and pulse pressure (PP) using $N = 280,000$ white British individuals from the full release of the UK Biobank. I also compare LEMMA against existing approach StructLMM and the F-Tests on logBMI.

The LEMMA analysis has several distinct steps. First, the WGR model was fitted using approximately 642,000 genotyped SNPs with minor allele frequency (MAF) greater than 0.01. The estimated environmental score (ES) was then used to estimate the proportion of phenotypic variability that is explained by GxE interactions (SNP GxE heritability), using recently developed randomised Haseman-Elston regression methods [86, 87]. Finally, a GWAS was performed on $10,295,038$ imputed SNPs, testing each SNP for multiplicative GxE interactions with the ES against a residualised phenotype. “Robust" standard errors were used when testing each variant for a GxE interaction, which helps to control for the conditional heteroskedasticity caused by GxE interactions.

4.2 Data

4.2.1 Genetic data

The UK Biobank is a large prospective cohort study of approximately $500,000$ middle-aged individuals living in the UK [116]. Participants were recruited between 2006 to 2010 from NHS patient registers, conditional on age (40-69 years) and proximity to one of 22 assessment centres spread across the UK in urban areas [120]. As well as extensive phenotypic and environmental data, the Biobank contains imputed genetic data of 92,693,895 SNPs, short indels and large structural variants from 487,442 individuals. Individuals were assayed using either the UK BiLeve Axiom™ Array (49,950 individuals) or the UK Biobank Axiom™ Array (438,427 individuals), and imputation was carried out on a combined reference panel from

the Haplotype Reference Consortium and the UK10K haplotype resource. The Biobank is both ethnically diverse, with over 22,000 participants from outside of Europe (self-reported), and to has elevated levels of relatedness (147731 individuals related to another participant in the Biobank, 3rd degree or closer) when compared to the general population [116].

To account for potential confounding effects of population structure, samples were first restricted to the set of 344,068 white British unrelated individuals used by Bycroft *et al.*[116] in a GWAS on human height. This represents unrelated individuals with white British ethnicity (self-reported) and whose genetic data projected onto principal components lies within the white British cluster. After subsetting down to individuals who had complete data across the phenotype, covariates and environmental factors (see below) I was left with approximately $280,000$ samples per trait (Table C.3).

Finally I filtered genetic data based on minor allele frequency (≥ 0.01) and IMPUTE info score (≥ 0.3), leaving approximately 642,000 genotyped variants and $10,295,038$ common imputed variants per trait (Table C.3). The phrase common imputed SNPs is used to refer to the set of imputed SNPs with MAF > 0.01 in the full UK Biobank cohort.

Multi-component HE-regression was performed after stratifying SNPs into 20 different components based on five MAF groups and four LD score quantiles. The MAF groups used were; MAF ≤ 0.1, $0.1 <$ MAF ≤ 0.2, $0.2 <$ MAF ≤ 0.3, $0.3 <$ MAF ≤ 0.4, $0.4 <$ MAF ≤ 0.5. LD score was defined as the sum of the r^2 values between the target variant and all variants within a 1 MB region centred on the variant. LD scores were computed using GCTA[121] on $10,000$ randomly chosen unrelated white British participants from the UK Biobank.

4.2.2 Phenotypes

BMI was computed as weight over height squared, using height and weight measurements made during the first assessment visit (instance '0' of field 21001). Participants whose BMI was more than six standard deviations from the population mean were

excluded. Finally, I applied a log transformation (for clarity I refer to the resulting variable as logBMI) as recommended by Young, Wauthier, and Donnelly [57], because BMI fits a log-normal distribution better than a normal distribution [57].

After calculating the mean SBP and DBP from two automated blood pressure readings made during the first assessment visit (fields 4080 and 4079), I adjusted for medication usage by adding 15mmHg and 10mmHg to SBP and DBP respectively [122]. Data from manual measurements (fields 93 and 94) was used in the rare instance that no automated reading was available. Participants whose SBP or DBP was more than four standard deviations from the population mean were excluded. PP was then calculated as SBP minus DBP.

4.2.3 Environments

From the data provided by the UK Biobank I included 7 continuous environmental variables ("Age when attended assessment centre", "Sleep duration", "Time spent watching television", "Number of days/week walked 10+ minutes", "Number of days/week of moderate physical activity 10+ minutes", "Number of days/week of vigorous physical activity 10+ minutes", "Townsend deprivation index at recruitment"), 1 ordinal environmental variable ("Alcohol intake frequency"), 9 dietary ordinal variables ("Salt added to food", "Oily fish intake", "Non-oily fish intake", "Processed meat intake", "Poultry intake", "Beef intake", "Lamb intake", "Pork intake", "Cheese intake") and 2 dietary continuous variables ("Tea intake", "Cooked vegetable intake"). We further derived one categoric variable ("Is Current Smoker" from the responses given in the UK Biobank field "Smoking status") and one continuous variable ("Sleep sd" as the number of standard deviations the participants estimated hours of sleep was from the population mean). For analyses of blood pressure, I additionally included one further continuous variable "Waist circumference". In all cases, where participants responded with "Prefer not to answer", "Do not know" or "None of the above" I set the value to missing. For three continuous variables ("Time spent watching television", "Tea intake", "Cooked vegetable intake") I removed the 99'th percentile and for "Sleep duration" I removed

both the 1st and 99th percentiles. This left 11 dietary variables and 10 non-dietary variables (11 for blood pressure traits). In addition, I included a binary indicator column for gender and for each non-dietary variable I included multiplicative interactions with age and gender. In total, this resulted in 42 environments in the BMI analysis and 45 environments for the blood pressure analyses. The set of environments used for the BMI analysis was similar to those used in previous GxE analyses of BMI in the UK Biobank [55, 57].

In all cases, where participants responded with "Prefer not to answer", "Do not know" or "None of the above" the value was set to missing. For three continuous variables ("Time spent watching television", "Tea intake", "Cooked vegetable intake") the 99'th percentile was removed and for "Sleep duration" both the 1st and 99th percentiles was removed.

Before running LEMMA each column was standardised using the formula

$$E_{ij} = \frac{E_{ij} - \text{mean}(E_{:,j})}{\text{sd}(E_{:,j})}.$$

Excluding participants with any missing data across phenotypes or environments left $281,149$ samples for the logBMI analysis and $280,749$ samples for the blood pressure analyses.

4.2.4 Covariates

In addition to controlling for the main effects of the environmental variables, I included age^3, age$^2\times$ gender, age$^3\times$ gender, a binary indicator for the genotype chip and the top 20 genetic principal components as covariates for each trait. For the blood pressure traits, I additionally controlled for BMI.

4.2.5 Post-processing

Recoding environmental variables

To aid interpretation, after running LEMMA I transformed the interaction weights to correspond to a recoded envirnonental data matrix E_1. Assuming that the

recoded matrix E_1 is a simple linear transformation of the original data matrix E, the recoded interactions weights w_1 can be obtained using the formula

$$\begin{aligned}\hat{w}_1 &= (E_1^T E_1)^{-1} E_1^T \hat{\eta}_{\text{LEMMA}}, \\ &= (E_1^T E_1)^{-1} E_1^T E \hat{w}\end{aligned}$$

Although multivariate linear regression is invariant to a rescaling of the design matrix, ridge regression is not due to the penalisation place on the magnitude of the learned parameters. However, as the magnitude of the weights from our UK Biobank analysis is typically small (less than 0.2) compared to the standard deviation of our gaussian prior (1) in this case the rescaling makes minimal difference.

Recoded data matrix E_1 was formed with one column for each of the 11 dietary variables (normalised to have mean zero and variance one), and three columns for each of the 10 (11 for blood pressure traits) non-dietary variables; the first augmented by a binary male indicator vector, the second by a binary female indicator vector and the third by a continuous vector of participant age. Columns augmented by male and female binary indicator vectors were normalised to have mean zero and variance one (not including zeros due to augmentation), apart from age which was scaled to represent the number of decades aged past 40. Columns augmented by age were normalised first and then multiplied by age on the per decade scale. Binary indicator columns for men and women were also included, which can be interpreted as gender-specific intercepts and is equivalent to including an intercept and a binary column for only one gender (men or women). This left 43 (46) columns where the extra column comes from including an intercept within the column space of E_1 and is necessary because some columns have mean not equal to zero. Thus the column space of E_1 is equivalent to E under the constraint that the ES has mean zero.

4.3 Results

4.3.1 Partitioned heritability estimates

I performed a partitioned heritability analysis, jointly estimating the proportion of trait variance explained by additive genetic effects (h^2_{G}) and by GxE interactions

with the ES (h^2_{GxE}). Previous studies have established that estimating heritability using a single component makes assumptions about the relationship between MAF, LD and trait architecture that may not hold up in practice[36, 80]. In comparison, stratifying SNPs into bins according to MAF and LD score (the LDMS approach) is relatively unbiased[80, 81, 123]. Therefore I performed three analyses using $M = 639,005$ genotyped SNPs with a single SNP component (SC), $M = 639,005$ genotyped SNPs stratified by LD score and MAF (LDMS), and $M = 10,270,052$ common imputed SNPs stratified by LD score and MAF (LDMS). LDMS analyses were performed using 5 MAF bins and 4 LD score quantiles.

Using imputed SNPs I estimated GxE heritability of 9.3%, 12.5%, 3.9% and 1.6% for logBMI, PP, SBP and DBP respectively (see Table 4.3). Table C.1 compares heritability estimates for each of the three runs and each trait. On genotyped SNPs the GxE heritability estimates were slightly lower for logBMI and PP ($h^2_{\text{GxE}} = 8.6\%$ and $h^2_{\text{GxE}} = 11.1\%$ respectively) and almost identical for SBP and DBP. For all traits the heritability of additive SNP effects were slightly higher on imputed data, consistent with previous results [81].

The heritability assigned to each SNP component is displayed in Figure 4.1 for each trait. Theoretical work on models of natural selection has shown that variance explained by additive genetic effects should be uniformly distributed as a function of MAF in a neutral evolutionary setting with constant population size[124]. For all four traits I found that additive genetic variance explained by low frequency SNPs was greater than would be expected under the neutral evolutionary model, consistent with observations of negative selection in these traits made previously[82]. Additionally, the distribution of additive genetic effects by MAF for logBMI was qualitatively similar to that found by GREML-LDMS is a previous study[81]. In contrast I found that variance explained by GxE effects was overwhelming attributed to low frequency SNPs (MAF < 0.1), especially those with low LD. However I am not aware of any theory linking the MAF distribution of GxE heritability to evolutionary models.

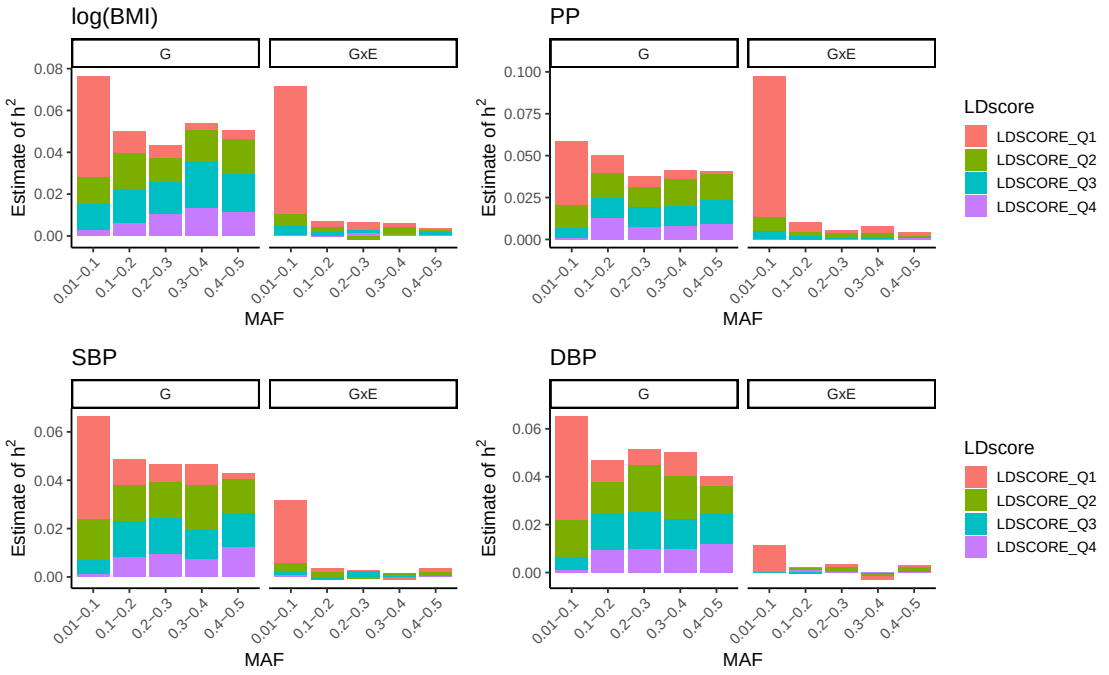

Figure 4.1: Partitioned heritability estimates for four quantitative traits in the UK Biobank. Heritability estimates partitoned into additive genetic and multiplicative GxE interaction effects for four quantitative traits in the UK Biobank using approximately 280,000 unrelated white British individuals (see Table C.3) and $M = 10,270,052$ common imputed SNPs (MAF > 0.01 in the full UK Biobank cohort). Multiplicative GxE iteractions were computed using the ES from each model fit. Heritability estimation was performed using a multi-component implementation of randomised HE-regression[86, 87] with 20 random samples and SNPs stratified into 20 components (5 MAF bins and 4 LD score quantiles).

4.3.2 Interpretation of the Environmental score

For logBMI LEMMA estimated an ES score that put high weight on alcohol intake frequency, Townsend Index and physical activity measures (**Figure 4.2c**). Almost all of the non-dietary environmental exposures had a higher effect in women than in men, with smoking status being the one exception. This is reflected in the facts that (a) the ES had much higher variance in women and (b) those with a negative ES were almost all female (97%) (see **Figure 4.2b**). The ES can alternatively be interrogated by examining how participants in the bottom or top 5% of the ES compare with the cohort as whole.

Summary statistics of participants on the tails of the ES against those of the whole cohort. When comparing the characteristics of those in the bottom 5% of the

ES to the whole cohort (using the mean for continuous variables and the mode for categoric), I found that those in the bottom 5% were predominantly female (100% vs 53%), younger (51 vs 56), had higher Townsend deprivation index (0.91 vs -1.74), drank less often ('Special occasions' vs 'Once or twice a week') and watched more TV (3.28 vs 2.69 daily hours of TV) (Table C.5). Positive values of the Townsend index indicate material deprivation, whereas negative values indicate relative affluence.

Table 4.1: Partitioned heritability estimates for four quantitative traits in the UK Biobank.

Trait	h^2_{G} (s.e)	h^2_{GxE} (s.e)
log BMI	0.274 (0.056)	0.093 (0.028)
PP	0.228 (0.051)	0.125 (0.028)
SBP	0.251 (0.05)	0.039 (0.023)
DBP	0.254 (0.05)	0.016 (0.02)

Heritability estimates obtained using common imputed SNPs (MAF$>$ 0.01 in the full UK Biobank cohort) with RHE-LDMS. GxE heritability estimates were were obtained using the ES from each model fit. All analyses controlled for the same covariates used in the WGR analysis (including the top 20 principal components). Abbreviations; s.e, standard error estimated using the block jacknife (see Appendix A.2); h^2_{G}, heritability due to additive genetic effects; h^2_{GxE}, heritability due to multiplicative GxE effects; RHE, randomised HE-regression[86, 87]; LDMS, SNPs stratified by minor allele frequency and LDscore (20 components).

Previous cross-sectional studies have reported GxE interactions between a linear predictor formed from BMI-associated SNPs and alcohol intake frequency [125], Townsend Index [125], physical activity measures [117, 125, 126] and time watching TV [125, 126], all of which are upweighted in the LEMMA lifestyle score. An alternative approach from Robinson *et al.*[49], called GCI-GREML, binned samples according to a single environmental exposure and tested for significant differences in SNP heritability using a likelihood ratio test. They reported strong interaction effects with age in a cohort of $43,407$ individuals whose ages spanned $18-80$, but only reported significant interactions with smoking in the UK Biobank interim release after testing eight environmental variables. This suggests that one might expect age to play a more dominant role in the logBMI ES in a cohort that included younger individuals. Finally, one category that is notably downweighted is the

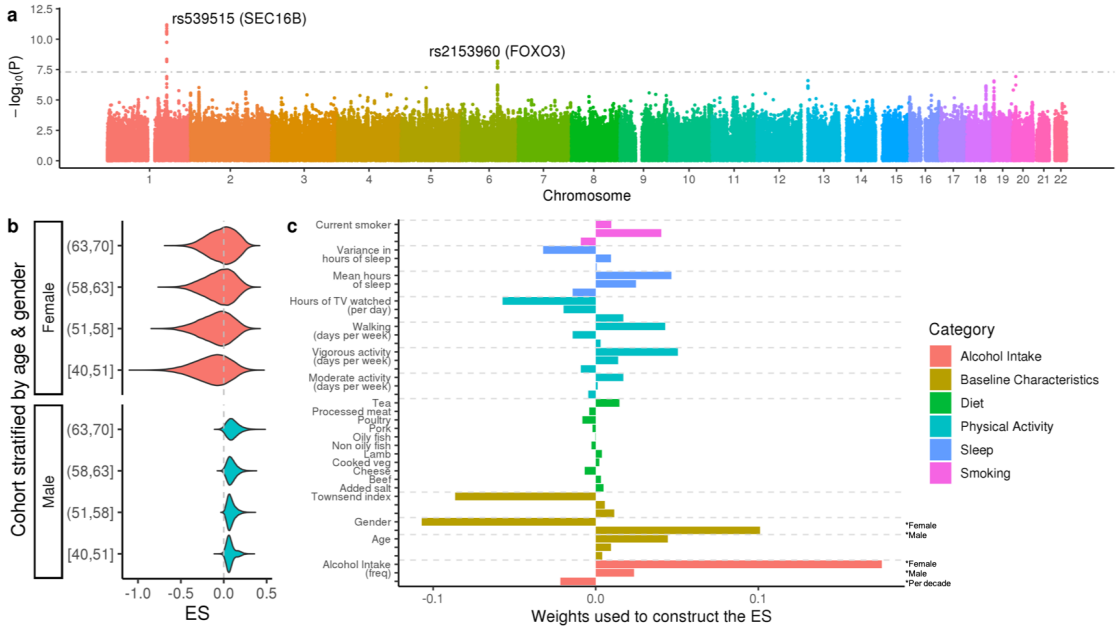

Figure 4.2: GxE analysis of log**BMI in the UK Biobank.** (a) LEMMA association statistics testing for multiplicative GxE interactions at each SNP. The horizontal grey line denotes ($p = 5 \times 10^{-8}$), p-values are shown on the $-\log_{10}$ scale. (b) Distribution of the environmental score (ES), stratified by gender and age quantile. (c) Weights used to construct the ES. Dietary variables have a single weight shown on the per standard deviation (s.d) scale. 'Gender' has two weights; a gender specific intercept for women (first) and men (second). Remaining non-dietary variables have three weights; (first) a per s.d effect for women only, (second) a per s.d effect for men only, (third) a per s.d per decade effect which is the same for both genders. s.d for the male and female specific weights is computed for each gender separately. Age is computed as the number of decades aged from 40. See Section 4.2.5 for details.

contribution from dietary variables. Although significant interactions with fried food consumption [127] and sugar-sweetened drinks [128] have previously been reported in a cohort of US health professionals, these dietary variables were not included in the diet questionnaire used by the UK Biobank.

The ES for PP was dominated by the effects of age and gender (age, age^2, age-x-gender, gender together explained 94.9% of variance in the ES). The magnitude of the ES was strongly associated with increased age whilst the sign of the ES was strongly associated with gender, implying that GxE effects were stronger in the elderly but acted in the opposite direction in men and women (Figure 4.3).

Similarly, the variance of the ES increased with age in both SBP and DBP. However age itself was not highly weighted in the ES; instead it appeared to

be interactions between age and other environmental. Specifically for SBP, age interactions with smoking, Townsend index and alcohol frequency explained 86% of variance in the ES (Figure 4.4). When compared to the cohort average, participants in the top 5% of the SBP ES were older (63 vs 58), had higher Townsend deprivation index (1.2 vs -1.74) and were more likely to smoke (59% vs 9%) whereas those in the bottom 5% were also older (65 vs 58), predominantly female (91.5%), rarely drank alcohol (43.9% drank "Never") and had low Townsend deprivation index (-2.9 vs -1.74) (Table C.7).

Finally, variance in the ES for DBP was notably higher amongst men, most of which appeared to be driven by high gender-specific weights for smoking status and alcohol frequency (Figure 4.5). Also, alcohol frequency and smoking status became increasingly influential with age. The total SNP-GxE heritability for this ES however was quite low.

4.3.3 Loci with significant multiplicative GxE interactions

For the final step in the LEMMA analysis I performed a GWAS on $10,295,038$ common imputed SNPs, testing each SNP for multiplicative interaction with the estimated ES against a residualised phenotype. This process was performed twice for each trait, with and without robust variance estimators. A comparison of the QQ plots from all eight GWASs makes it clear that robust variance estimators make a noticeable difference (Figure 4.6).

Using hypothesis tests with robust variance estimators, I identified two loci for logBMI (Figure 4.2a) and one locus for DBP (**Figure 4.5a**) with significant GxE interactions at the standard threshold for genome-wide significant ($p < 5 \times 10^{-8}$). The sentinel SNPs are shown in Table C.2). For logBMI, LEMMA identified GxE interactions at rs2153960 ($p = 6.5 \times 10^{-9}$; Figure 4.7) and at rs539515 ($p = 6.5 \times 10^{-12}$; Figure 4.8). rs2153960 is an intron in the *FOXO3* gene and has previously been associated with Insulin-like growth factor 1 (IGF-I) concentration in a cohort of $10k$ middle-aged Europeans [129]. IGF-I is known to be a central mediator of metabolic, endocrine and anabolic effects of growth hormone and is also

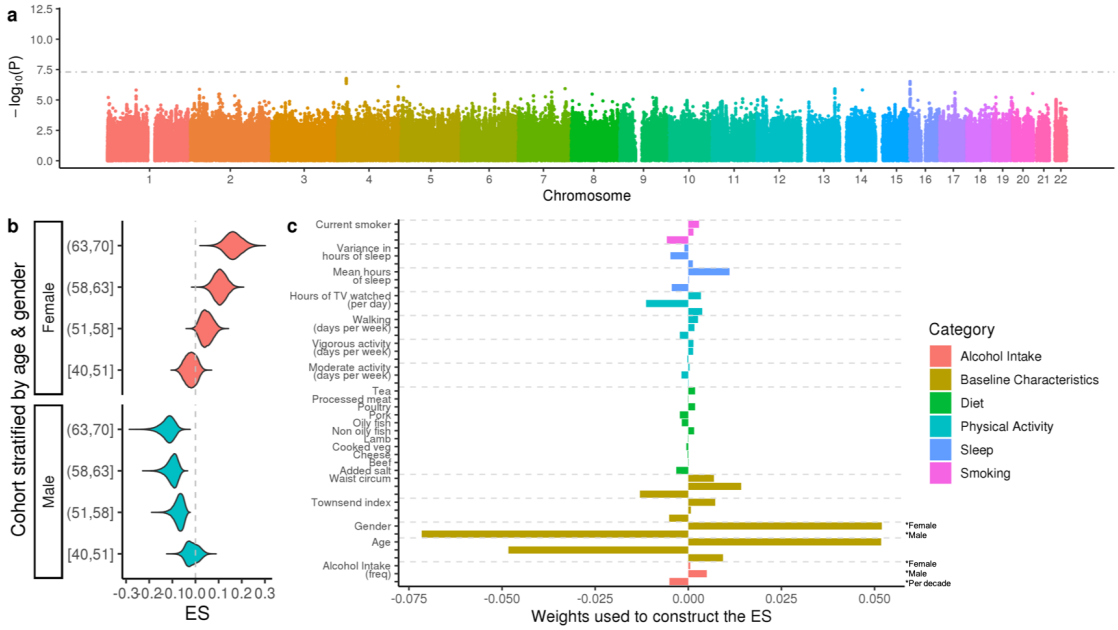

Figure 4.3: GxE analysis of PP in the UK Biobank. (a) LEMMA association statistics testing for multiplicative GxE interactions at each SNP. The horizontal grey line denotes ($p = 5 \times 10^{-8}$), p-values are shown on the $-\log_{10}$ scale. (b) Distribution of the environmental score (ES), stratified by gender and age quantile. (c) Weights used to construct the ES. Dietary variables have a single weight shown on the per standard deviation (s.d) scale. 'Gender' has two weights; a gender-specific intercept for women (first) and men (second). Remaining non-dietary variables have three weights; (first) a per s.d effect for women only, (second) a per s.d effect for men only, (third) a per s.d per decade effect which is the same for both genders. s.d for the male and female-specific weights is computed for each gender separately. Age was computed as the number of decades aged from 40. See Section 4.2.5 for details.

involved in carbohydrate homeostasis [129]. rs539515 is located 6kb downstream of *SEC16B*. Multiplicative GxE interactions have been reported at *Sec16B* with multiple environmental variables in a similar analysis in the UK Biobank[55], and with physical activity separately in Europeans [117] ($p = 0.025$) and in Hispanics [130] ($p = 8.1 \times 10^{-5}$). Highly significant variance effects ($p = 3.88 \times 10^{-17}$), which can be indicative of GxE, have also been reported at the *Sec16B* locus using $N = 456,422$ Europeans in the UK Biobank[131]. *SEC16B* transcribes one of the two mammalian homologues of the Sec16 protein, which has a key role in organising endoplasmic reticulum exit sites by interacting with COPII components [132]. Although several GWASs have identified associations between *SEC16B* and BMI[133, 134], the relevance of *SEC16B* to BMI is not well characterised [135]. Some

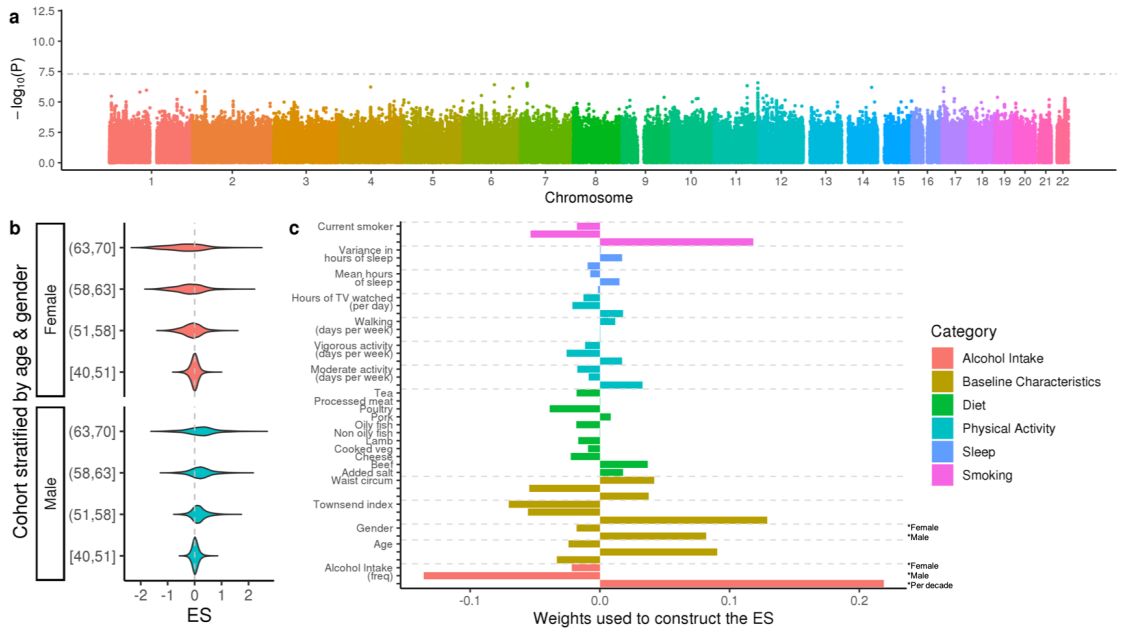

Figure 4.4: GxE analysis of SBP in the UK Biobank. (a) LEMMA association statistics testing for multiplicative GxE interactions at each SNP. The horizontal grey line denotes ($p = 5 \times 10^{-8}$), p-values are shown on the $-\log_{10}$ scale. (b) Distribution of the environmental score (ES), stratified by gender and age quantile. (c) Weights used to construct the ES. Dietary variables have a single weight shown on the per standard deviation (s.d) scale. 'Gender' has two weights; a gender-specific intercept for women (first) and men (second). Remaining non-dietary variables have three weights; (first) a per s.d effect for women only, (second) a per s.d effect for men only, (third) a per s.d per decade effect which is the same for both genders. s.d for the male and female-specific weights is computed for each gender separately. Age was computed as the number of decades aged from 40. See Section 4.2.5 for details.

evidence exists to suggest that *SEC16B* has a role in the transport of peroxisome biogenesis factors; peroxisomes being an organelle involved in the catabolism of long chain fatty acids found ubiquitously in eukaryotic cells. Previous authors have also speculated that the *SEC16B* might play a role in the transport of appetite regulatory peptides; however, I am not aware of any evidence for this theory.

For DBP one locus around rs8090962 reached genome-wide significance for multiplicative GxE with the ES. The interaction effect estimated for rs8090962 is qualitative[38] (see Figure 4.9); increased exposure is estimated to increase DBP in carriers and reduce it in non-carriers (and vice versa). Thus it is not surprising that LEMMA does not find a significant effect at rs8090962 or in the surrounding region (Figure 4.9). rs8090962 is located within an enhancer, approximately 100KB

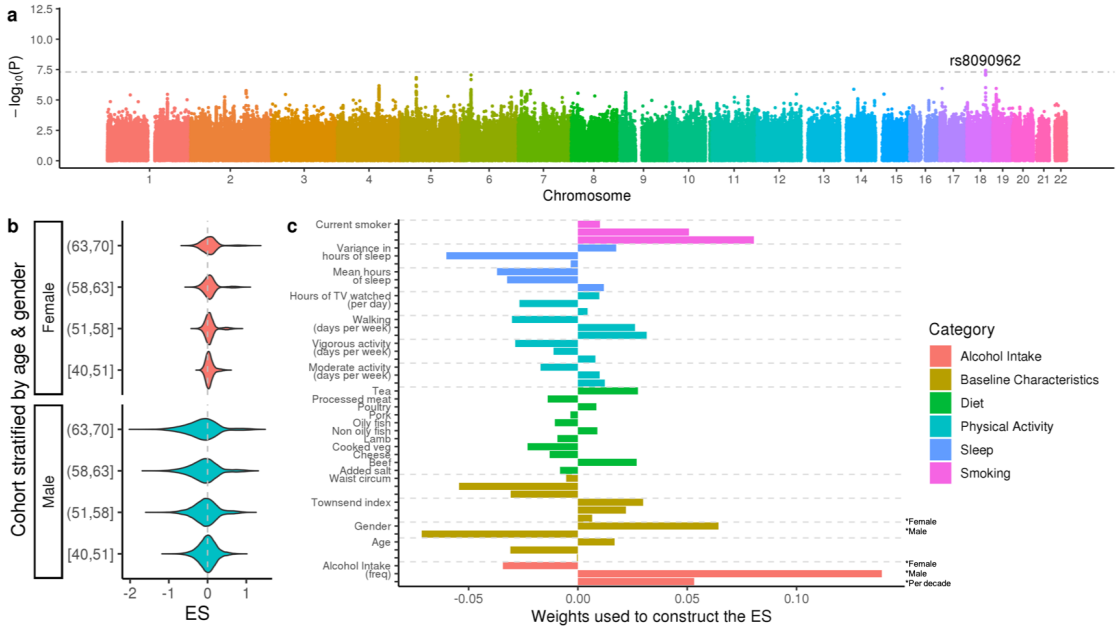

Figure 4.5: GxE analysis of DBP in the UK Biobank. (a) LEMMA association statistics testing for multiplicative GxE interactions at each SNP. The horizontal grey line denotes ($p = 5 \times 10^{-8}$), p-values are shown on the $-\log_{10}$ scale. (b) Distribution of the environmental score (ES), stratified by gender and age quantile. (c) Weights used to construct the ES. Dietary variables have a single weight shown on the per standard deviation (s.d) scale. 'Gender' has two weights; a gender-specific intercept for women (first) and men (second). Remaining non-dietary variables have three weights; (first) a per s.d effect for women only, (second) a per s.d effect for men only, (third) a per s.d per decade effect which is the same for both genders. s.d for the male and female-specific weights is computed for each gender separately. Age was computed as the number of decades aged from 40. See Section 4.2.5 for details.

downstream of the *SEC11C* gene and 50KB upstream of the *ZNF532* gene. However neither rs8090962 nor the two genes have previously been associated with blood pressure traits, the biological impact of rs8090962 is hard to interpret.

4.3.4 Comparison with other methods

Comparison with StructLMM, the F-Test and the robust F-Test

To compare LEMMA with existing single SNP methods I also ran StructLMM, the F-test and the robust F-test on logBMI using the same set of environmental variables as used by LEMMA (but not including the significant squared environments as covariates). Manhattan plots from each method are shown in Figure B.2 and the $-\log_{10}(p)$ values are compared directly in scatter plots in Figure 4.10. As expected,

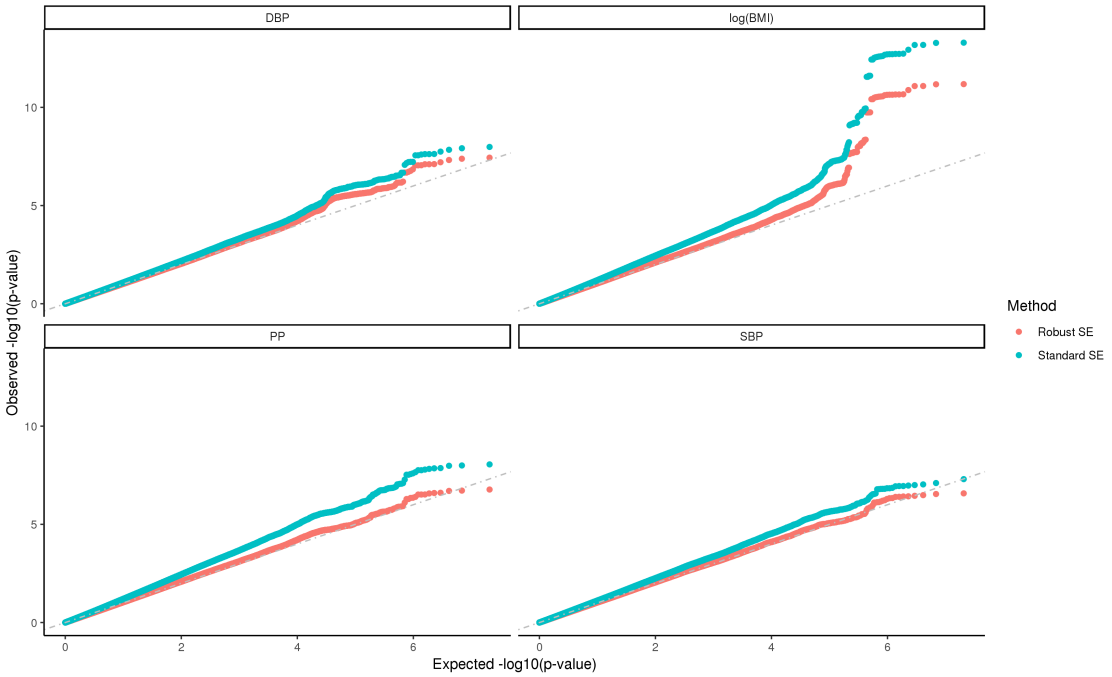

Figure 4.6: Effect of using robust standard errors for GxE interaction tests in the UK Biobank. QQ plots of the observed LEMMA $-\log_{10}(p)$ values for GxE interactions at imputed SNPs for four UK Biobank traits, with and without robust standard errors. The grey dotted line denotes expected $-\log_{10}(p)$-values under a null model. Association tests using 'Robust' standard errors are well calibrated in both homoskedastic and heteroskedastic regimes (see **Online Methods**) and are used in all follow up analysis. Genomic control statistics were $1.275, 1.271, 1.163, 1.111$ for logBMI, PP, SBP and DBP respectively using homoskedastic standard errors and $1.062, 1.047, 1.037, 1.027$ for logBMI, PP, SBP and DBP respectively using robust standard errors.

the F-test and StructLMM produced quite similar p-values, especially amongst SNPs with suggestive evidence of GxE Interaction results (Pearson $r^2 = 0.502$ amongst SNPs with $p < 1 \times 10^{-5}$; see Figure 4.10a). Furthermore, test statistics from both the F-test ($\lambda_{\text{GC}} = 1.37$) and StructLMM ($\lambda_{\text{GC}} = 1.235$) were substantially inflated when compared to the robust F-test and LEMMA ($\lambda_{\text{GC}} = 1.03$ and $\lambda_{\text{GC}} = 1.062$ respectively; see Table C.4), suggesting that StructLMM does not properly control for heteroskedasticity.

When comparing LEMMA and the robust F-test, I found low correlation between the respective p-values and that genome-wide significant loci identified by one method were not identified by the other and vice-versa (see Figure 4.10e). LEMMA relies on the assumption that all GxE interaction effects for a single trait share a com-

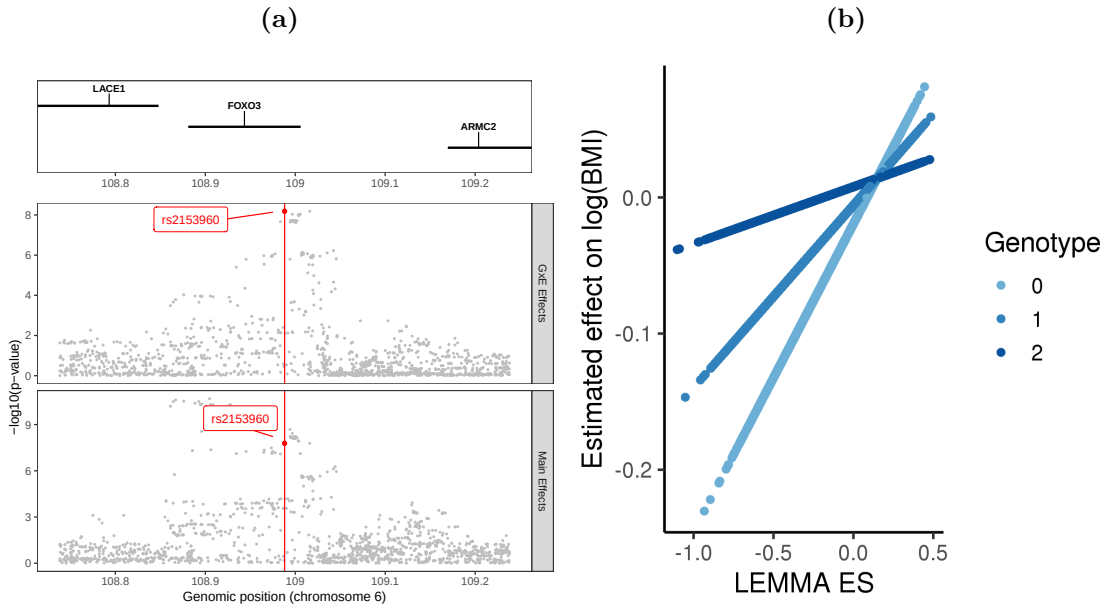

Figure 4.7: Estimated GxE effect rs2153960 on log**BMI.** (a) Regional plot of the main and interaction effects of SNPs within 250KB of rs2153960, (b) the estimated effect of rs2153960 on logBMI as a function of the environmental score (ES).

mon ES, and I have shown in simulation that when this assumption holds LEMMA achieves substantial increases in power. However I would expect LEMMA to have little power to detect SNPS which interact with a combination of environments that is not well correlated with the genome-wide ES estimated by LEMMA.

As an example the *FTO* is the strongest known susceptibility locus for BMI and was identified as genome-wide significant for GxE interactions by the robust F-test, the F-test and StructLMM in this analysis (Figure B.2) as well as in multiple previous studies with a variety of environmental variables [55, 57, 58, 68, 136]. After extracting an estimate of the SNP specific interaction profile at *FTO* using the robust F-test, I found that it's correlation with LEMMA's ES was low (Pearson $r^2 = 0.3$). In comparison, a similar analysis at *SEC16B* and *FOXO3* yielded much higher correlations (Pearson $r^2 = 0.725$ and $r^2 = 0.713$ respectively).

Comparison with MVRM

Finally I briefly compare LEMMA against the recently published Multivariate Reaction Normal Model (MRNM)[51]. The MRNM proposes to model GxE interaction with a single continuous environmental variable whilst carefully controlling

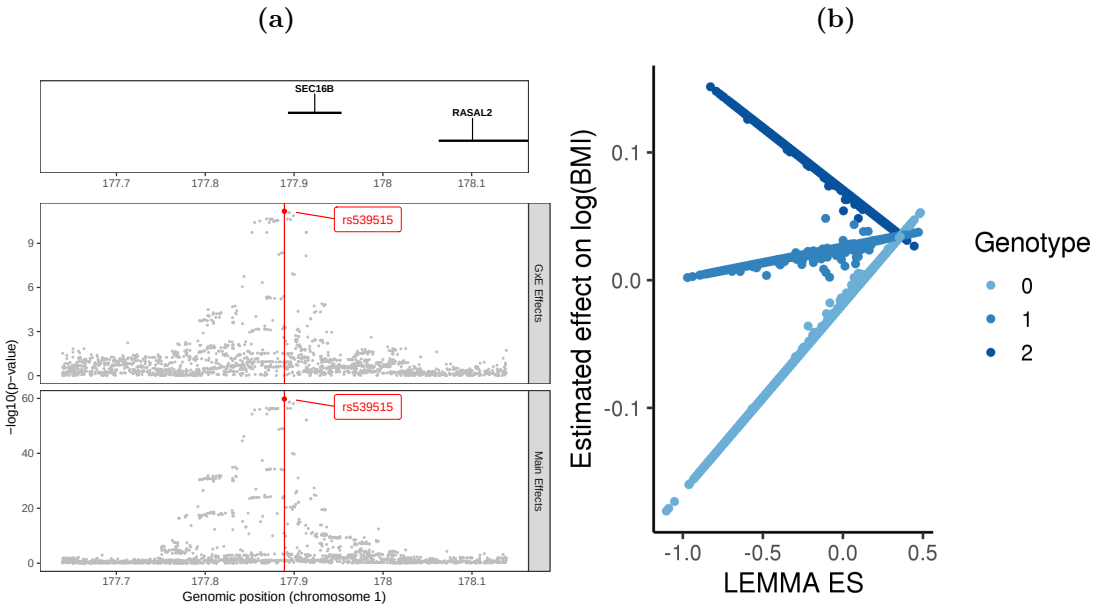

Figure 4.8: Estimated GxE effect rs539515 on log**BMI.** (a) Regional plot of the main and interaction effects of SNPs within 250KB of rs539515, (b) the estimaetd effect of rs539515 on logBMI as a function of the environmental score (ES).

for G-E correlation (eg shared genetic architecture), R-E correlation (eg shared environmental influences) and R-E interaction in a variance components framework.

The two methods have slightly different aims. LEMMA models GxE with an aggregate variable (the ES) made up of multiple environmental variables and then performs heritability estimation and single SNP testing. In contrast MRNM only performs a variance component analysis, but does so whilst explicitly account for confounding factors that LEMMA does not (G-E correlation, R-E correlation and R-E interaction). In particular, there is nothing in LEMMA's modelling framework to control for R-E interaction. However Ni et al. [51] note that in simulation "in the absence of G-E interactions but with R-C interactions, type I error rate was well controlled by all methods"; suggesting that this sort of effect is unlikely to bias estimates of GxE heritability.

In simulation Ni et al. [51] show that controlling for G-E correlation is important to obtain unbiased GxE estimates when comparing against the RNM model; a sub-model within the MRNM framework that makes no attempt to control for the main effect of E. Ni et al. [51] also acknowledge that simply regressing out the main effect

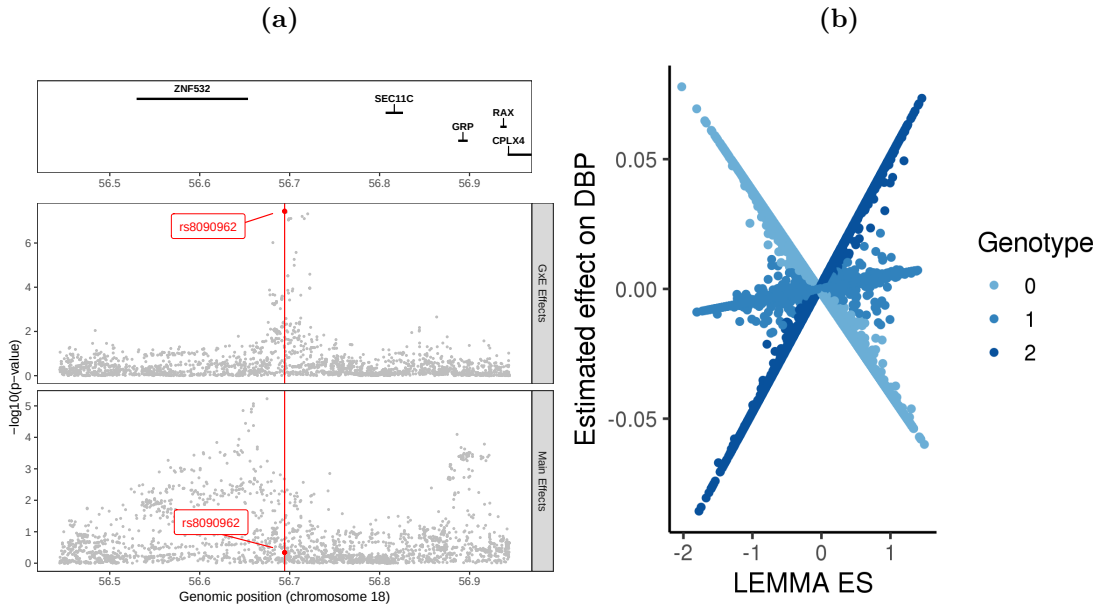

Figure 4.9: Estimated GxE effect rs8090962 on DBP. (a) Regional plot of the main and interaction effects of SNPs within 250KB of rs8090962, (b) the estimated effect of rs8090962 on DBP as a function of the environmental score (ES).

of E can provide a 'somewhat crude' method of controlling for G-E correlation. I find this somewhat unsurprising; standard linear modelling theory suggests that the best way to control for a variable (in this case E) is to model it jointly (as LEMMA does) or regress it out of all terms in the system of equations (as RHE does; see the Frisch–Waugh–Lovell theorem). I therefore suggest that LEMMA should be robust to G-E correlation - although this has not been empirically tested in simulation.

In terms of computational efficiency, LEMMA appears to be superior to MRNM. Ni et al. [51] state that the full MRNM model (ie including terms for R-E correlation and interaction) takes 300 hours on a dataset of $60K$ samples using 14 cores. As a consequence, Ni et al. [51] run their main data analysis using 66,281 samples from the UK Biobank first release and further split this cohort into two for a meta analysis approach. I finally note that Ni et al. [51] state that extending their framework to model GxE with multiple environments is theoretically simple, but chose not to do so due to computational limitations (as the number of parameters would increase exponentially). Thus this computational bottelneck has practical implications for their method.

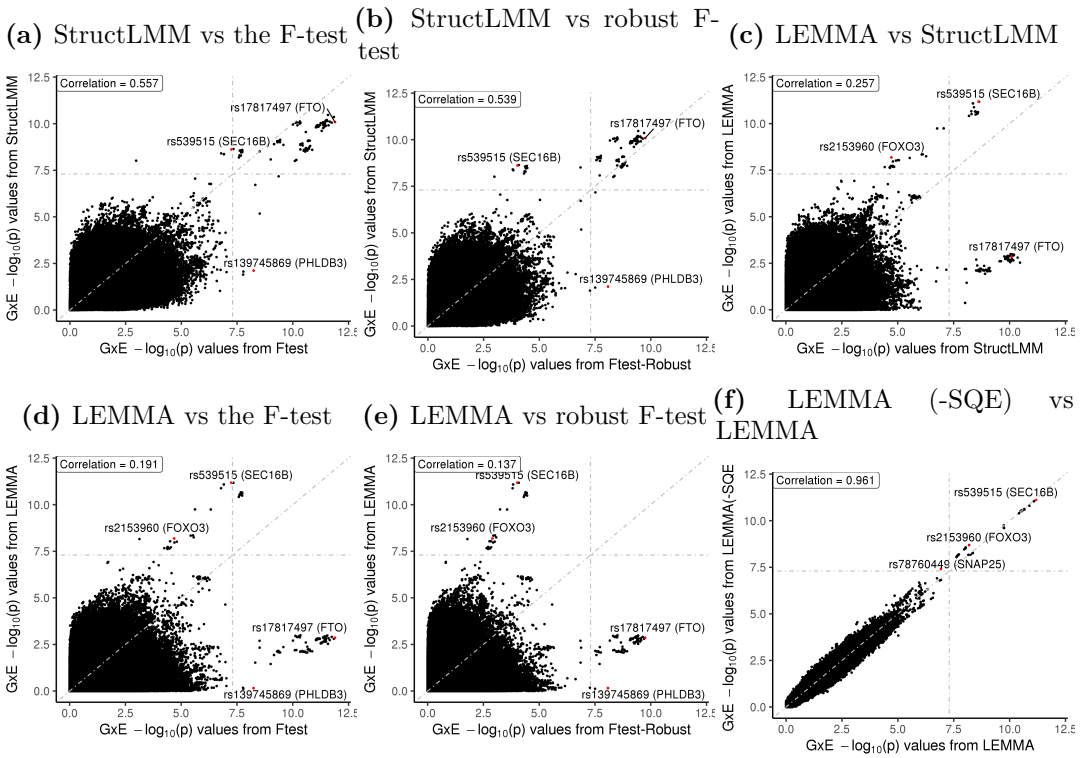

Figure 4.10: Comparison of GxE association statistics for log**BMI.** Comparison of negative $\log_{10}$ p values obtained from LEMMA, StructLMM, the F-test and the robust F-test in an analysis of logBMI in the UK Biobank. Grey lines denote ($p = 5 \times 10^{-8}$) and the $y = x$ axis. Pearson correlation is shown in a label at the top left of each plot. Red points denote the sentinel SNP for each locus.

Comparison with an ES estimated from marginal environmental effects

As a final comparison, I used the main effect of each environment (from a multivariate linear regression with all of the non-genetic covariates used by LEMMA) to construct an environmental score based on the marginal environmental effects. I refer to this as ES_{main}. If ES_{main} is similar to the LEMMA ES, then this would suggest that the entire LEMMA analysis could be simplified by using ES_{main} as a fixed effect in the Bayesian WGR instead of modelling the interactions weights as latent variables.

The correlation between the ESs estimated from the two approaches was $-0.062, -0.019, -0.297$ and -0.088 for logBMI, PP, SBP and DBP respectively. Low correlation here suggests that replacing the LEMMA ES with ES_{main} would be a poor heuristic, however confirming this is a possible direction of future research Figure B.3 shows a comparison of the interaction weights used to construct the

LEMMA ES and ES_{main} for each of the four traits. Visually the weights learnt through each approach look quite distinct. In particular, age, age^2 and age×gender appear to have stronger main effects in ES_{main} than in the LEMMA ES.

Correlation of beta and gamma

The predicted main and interaction effects learned by LEMMA, namely the N-vectors $X\beta$ and $X\gamma$ could be interpreted as polygenic risk scores. One simplifying assumption that could used when searching for GxE interactions is to replace the interaction effects PRS $X\gamma$ with the main effects PRS $X\beta$.

Table 4.2 displays the absolute correlation for $\text{cor}(\beta, \gamma)$ and $\text{cor}(X\beta, X\gamma)$ for each trait analysed by LEMMA. It is interesting to note that these correlations are all quite low, suggesting that LEMMA benefits from the additional flexibility allowed from modelling β and γ as separate effects.

Table 4.2: Correlation between main SNP effects and interaction SNP effects

Trait	abs(cor($X\beta$, $X\gamma$))	abs(cor(β, γ))
log(BMI)	0.13680	0.05810
PP	0.05306	0.03324
SBP	0.01741	0.00816
DBP	0.02464	0.00732

Correlation between main SNP effects and interaction SNP effects learnt during the LEMMA variational algorithm. Absolute correlation is used as γ is invariant to being multiplied by -1 (as LEMMA would apply the same transform to the ES).

4.3.5 Computational cost

The number of iterations and the time taken for the WGR model (the most computationally expensive part of the LEMMA algorithm) to converge is shown in Table C.3. Each run was performed using 196 GB of RAM and 32 cores distributed across a cluster with Xeon E5-2667 v4 3.2Ghz processors.

The number of iterations required for the WGR algorithm to converge ranged from 646 (on DBP) to 3821 (on PP). An earlier implementation of LEMMA using multithreading based on OpenMP took approximately 2000 seconds per iteration

Table 4.3: Partitioned heritability estimates for four quantitative traits in the UK Biobank.

Trait	h^2_{G} (s.e)	h^2_{GxE} (s.e)
log BMI	0.274 (0.056)	0.093 (0.028)
PP	0.228 (0.051)	0.125 (0.028)
SBP	0.251 (0.05)	0.039 (0.023)
DBP	0.254 (0.05)	0.016 (0.02)

Heritability estimates obtained using common imputed SNPs (MAF$>$ 0.01 in the full UK Biobank cohort) with RHE-LDMS. GxE heritability estimates were were obtained using the ES from each model fit. All analyses controlled for the same covariates used in the WGR analysis (including the top 20 principal components). Abbreviations; s.e, standard error estimated using the block jacknife (see Appendix A.2); h^2_{G}, heritability due to additive genetic effects; h^2_{GxE}, heritability due to multiplicative GxE effects; RHE, randomised HE-regression[86, 87]; LDMS, SNPs stratified by minor allele frequency and LDscore (20 components).

(using 16 cores and the same processor), meaning that the PP analysis would have taken over 88 days to converge. However using 32 cores and OpenMPI meant that the PP analysis took (only) 183 hours 26 minutes to converge. Thus use of OpenMPI was crucial to making LEMMA computationally tractable on biobank scale datasets.

Figure 4.12 shows the evolution of the interaction weights after each full cycle of VI updates, for all four traits. Many of the interaction weights display trajectories that are akin to slowly climbing a shallow plateau, which suggests that employing a momentum based inference strategy[137, 138] could lead to faster convergence.

Figure 4.11 compares the per-iteration improvement in the ELBO from running LEMMA with and without the SQUAREM algorithm on each of the four traits. The SQUAREM algorithm is an iterative scheme (implemented within LEMMA) that promises to accelerate the convergence of EM-like algorithms[96]. On average, running LEMMA with the SQUAREM algorithm reduced the number of iterations required for convergence by 543 and lowered the converged ELBO by 3.59. Variational inference is sensitive to initialisation[90]; thus, a small amount of variation around the converged ELBO is to be expected.

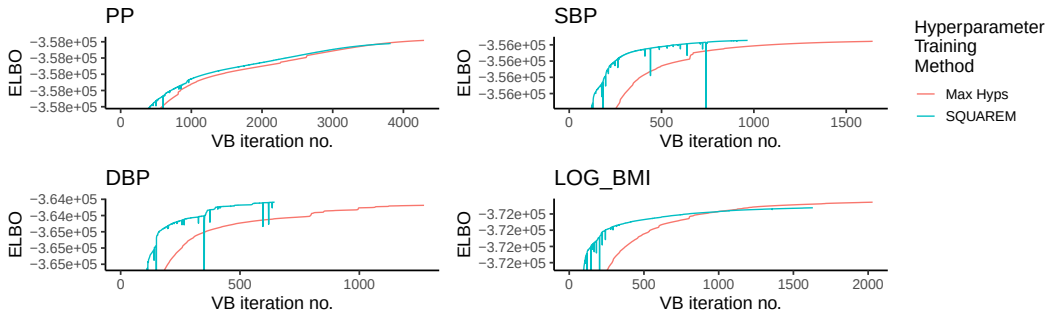

Figure 4.11: Comparison of hyperparameter maximization strategies on four quantitative traits in the UK Biobank.

4.4 Discussion

In this chapter, I have described results from applying LEMMA in a GxE analysis of four quantitative traits using data from the UK Biobank. First, a novel WGR framework was used on genotyped SNPs to jointly estimate SNP effects and the weights in a linear combination of environments that interacts with SNPs genome wide (referred to here as the ES). Second I estimated the proportion of trait variance attributable to multiplicative interactions with the ES and found that GxE effects among common imputed SNPs make a non-trivial contribution to trait variance for logBMI and PP (9.3% and 12.5% respectively), which is found primarily amongst low frequency SNPs (MAF < 0.1). Finally, I tested each SNP for non-zero interactions with the ES and identified three loci with statistically significant interactions. Two of these loci, rs539515 (*FOXO3*) and rs8090962, are novel and for the other, rs539515 (*SEC16B*), we show stronger evidence for than in the previous study[55].

The advantage of modelling mutliple environments within the LEMMA framework can be highlighted by comparing to Robinson *et al.*[49], who used GCI-GREML to quantify GxE heritability for BMI with a single environmental variable. GCI-GREML requires ordinal variables, but the authors convert continuous measures to ordinal using 4 quantiles. The authors performed GxE analyses with eight environmental variables (including measures of smoking, hours of TV watched and alcohol frequency) using data from the interim UK Biobank release and reported significant GxE heritability for smoking only ($h^2_{\text{GxE}} = 4\%$). In contrast, the ES

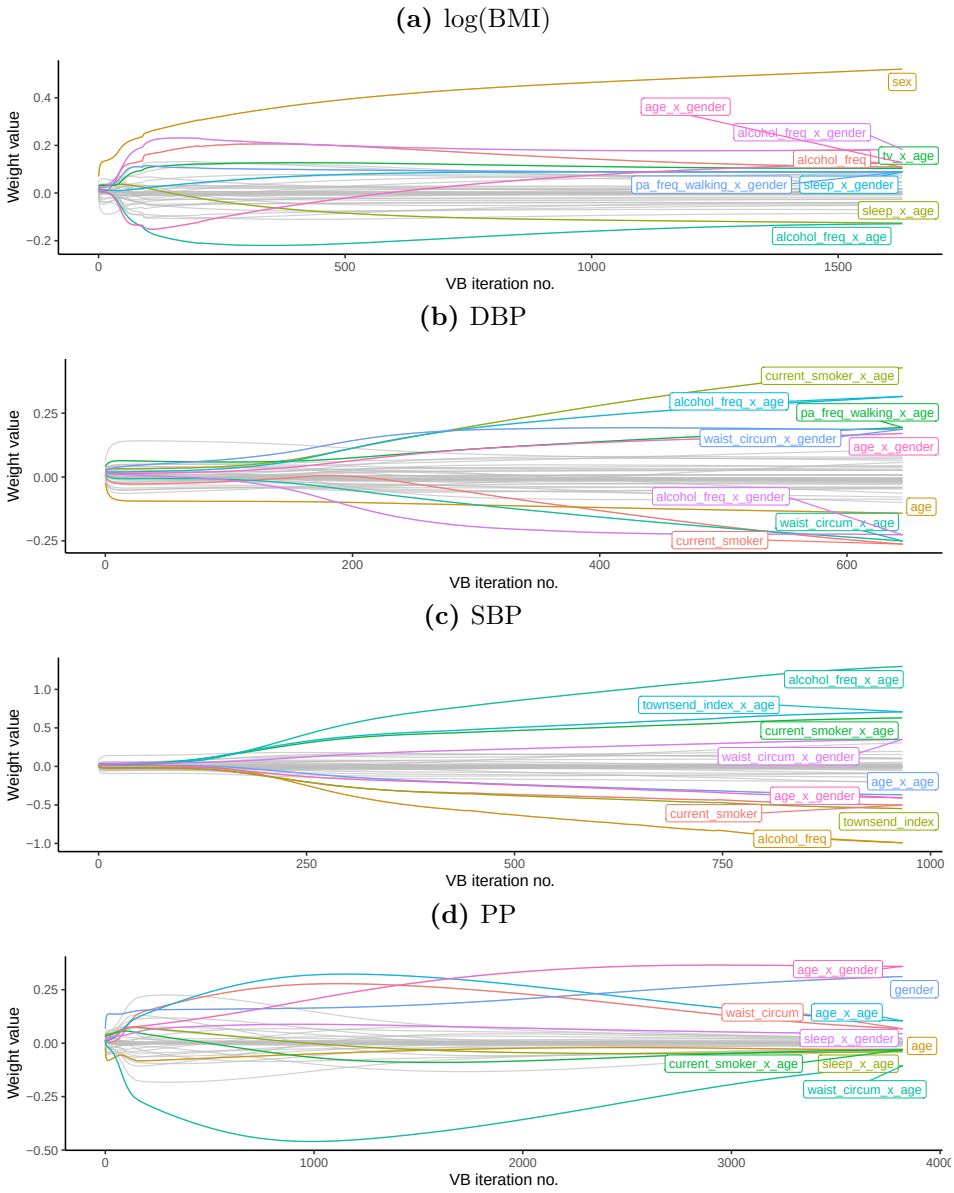

Figure 4.12: Inference of environmental score weights from GxE analyses of four quantitative traits in the UK Biobank. Evolution of the environmental score weights as LEMMA performs successive passes through the data.

estimated for logBMI in our analysis had upweighted contributions from many environmental variables, including hours of TV watched and smoking, suggesting that multiple environmental variables can influence on the genetic predisposition to BMI. Modelling these environmental variables jointly allowed LEMMA to capture a combination whose GxE interactions explained 9.3% of heritability.

In this chapter, I also compared LEMMA with three existing single SNP methods

(StructLMM, the F-test and a robust F-test), using logBMI as a model trait. The performance of StructLMM and the F-test appeared similar in this analysis; both identified similar sentinel SNPs and had inflated genomic control statistics (1.236 and 1.372 respectively). Use of robust variance estimators reduced genomic control for the F-Test to 1.034, suggesting that this inflation is due to heteroskedasticity and not to polygenicity of GxE effects. The two F-test methods further benefit from a wealth of existing theory[105] and, being theoretically simpler than StructLMM, could be easily implemented as an R-plugin with PLINK[139] (for example [101]). StructLMM is also arguable less interpretable than the F-Tests because identification of the relative importance of each environment is ascertained by comparing the log-marginal likelihood with and without that environment[55]. Therefore the robust F-test appears to be the most appropriate of the three single SNP methods to model GxE effects with tens of environments in biobank scale datasets. Finally, comparing LEMMA to the robust F-Test suggested that each method is powered to detect quite different GxE effects, consistent with the different assumptions made by each method.

5

GxE with multiple environmental variables using randomised HE-regression

Contents

This chapter introduces two extensions of the randomised HE-regression framework[86, 87] for modelling GxE interactions with multiple environmental variables. The first, which I refer to as RHE-multiE, is a simple extension of the randomised HE-regression framework that uses an additional L components to model multiplicative interactions, one for each of the L environmental variables.

The second, which I refer to as RHE-NLS, attempts to estimate the same environmental score used in LEMMA but using the framework of randomised HE-regression. Simulation results in this chapter suggest that such an approach could be a more computationally efficient way of achieving the same aims as LEMMA, as once

estimated the ES could be used to discover which environments are most relevant for GxE, used in single SNP hypothesis testing or to estimate the GxE heritability.

The work in this chapter was all done in the last month before submission, so the methods presented here are less mature than those presented earlier.

5.1 Methods

5.1.1 RHE-multiE

Here I extend the RHE framework defined in Section 1.3.2 to model GxE interactions with L environmental variables. Let E be an $N \times L$ matrix of environmental variables, where each column has been normalised to have mean zero and variance one. Then the RHE-multiE model is given by

$$y \sim \mathcal{N}(0, \sigma_\beta^2 K + \sum_l \sigma_{w_l}^2 E_l \odot K \odot E_l^T + \sigma_e^2 I). \tag{5.1}$$

For convenience of notation I define the matrices

$$K_k = \begin{cases} \frac{XX^T}{M} & \text{if } k = 1, \\ \frac{\text{diag}(E_{k-1}) X X^T \text{diag}(E_{k-1})}{M} & \text{if } 1 < k \leq L+1, \\ I & \text{if } k = L+2, \end{cases}$$

and let $\theta = \{\sigma_\beta^2, (\sigma_{w_l}^2)_{l=1}^L, \sigma_e^2\}$. Thus Equation (5.1) can be re-written as

$$y \sim \mathcal{N}\left(E\alpha, \sum_k \theta_k K_k\right).$$

As shown previously, fitting the variance components is done analytically by solving the following system of equations

$$T\theta = c \tag{5.2}$$

where

$$T_{kl} = \text{tr}\left(W K_k W K_l W\right) \tag{5.3}$$

$$c_k = y^T W K_k W y, \tag{5.4}$$

and $W = I - E(E^T E)^{-1} E^T$. As shown in the original RHE method[86, 87], Hutchinson's estimator can be used to efficiently estimate T_{kl}. Finally, the variance components are converted to heritability estimates using the following formula

$$\hat{h}_k^2 = \frac{\hat{\theta}_k \text{tr}\left(K_k\right)}{\sum_k \hat{\theta}_k \text{tr}\left(K_k\right)}$$

Efficient computation

To build the system of equantions in Equation (5.2), genotypes were streamed from file to compute $y^T W X X^T W y$ and the following N-vectors

$$u_b = X X^T W z_b,$$
$$v_{b,l} = X X^T E_l W z_b,$$

where z_b for $1 \leq b \leq B$ are random N-vectors drawn from a Gaussian distribution with mean zero and identity covariance.

Let ξ_b^k be defined as

$$\xi_b^k = \begin{cases} u_b & \text{if } k = 1, \\ v_{b,l} & \text{if } 1 < k \leq L + 1, \\ z_b & \text{if } k = L + 2. \end{cases}$$

Then Equation (5.2) can be filled using the following equations

$$T_{kl} = \frac{1}{M^2 B} \sum_b (\xi_b^k)^T \xi_b^k.$$

RHE-multiE costs $\mathcal{O}(NMLB)$ in compute and $\mathcal{O}(NLB)$ in RAM.

5.1.2 RHE-NLS

Here I describe a new method to collapse multiple environmental variables down to a single linear combination that is involved in multiplicative GxE interactions at SNPs genome-wide. As with LEMMA, this linear combination is referred to as an environmental score. The method aims to estimate the ES by minimising the squared loss between the expected and observed covariance matrices with respect to the variance parameters $\sigma_e^2, \sigma_G^2, \sigma_{GxE}^2$ and the interaction weights w.

Consider the Bayesian whole genome regression model given by

$$y = X\beta + \eta \odot X\gamma + \epsilon,$$
$$\eta = Ew,$$
$$\beta \sim \mathcal{N}(0, \sigma_\beta^2 I),$$
$$\gamma \sim \mathcal{N}(0, \sigma_\gamma^2 I).$$

This is the same model used by LEMMA in Equation (2.1), except the mixture of Gaussians priors on SNP effects have been replaced with Gaussian priors. Integrating out the SNP effects yields the model

$$y \sim \mathcal{N}(0, \sigma_\beta^2 K + \sigma_\gamma^2 K_2(w) + \sigma_e^2 I),$$

where

$$\begin{aligned} K &= \frac{XX^T}{M}, \\ K_2(w) &= \mathrm{diag}\left(Ew\right) K \mathrm{diag}\left(Ew\right), \\ &= \frac{1}{M}\mathrm{diag}\left(Ew\right) XX^T \mathrm{diag}\left(Ew\right), \\ &= \frac{1}{M}\sum_{l,m} w_l w_m \mathrm{diag}\left(E_l\right) XX^T \mathrm{diag}\left(E_m\right). \end{aligned}$$

Minisiming the squared loss between the expected and observed covariance is equivalent to the following regression problem

$$\begin{aligned} \mathrm{vec}\left(yy^T\right) &= \sigma_\beta^2 \mathrm{vec}\left(K\right) + \sigma_\gamma^2 \mathrm{vec}\left(K_2(w)\right) + \sigma_e^2 \mathrm{vec}\left(I\right) + \epsilon', \\ &= \sigma_\beta^2 \mathrm{vec}\left(K\right) + \sum_{l,m} \sigma_\gamma^2 w_l w_m \mathrm{vec}\left(\mathrm{diag}\left(E_l\right) K \mathrm{diag}\left(E_m\right)\right) + \sigma_e^2 \mathrm{vec}\left(I\right) + \epsilon'. \end{aligned} \tag{5.5}$$

In this format it is clear that optimising $\sigma_\beta^2, \sigma_\gamma^2, w, \sigma_e^2$ is a non-linear regression problem. Further, including a parameter for σ_γ^2 is no longer necessary. From here on I set $\tilde{w}_l = \sqrt{\sigma_\gamma^2} w_l$ and drop the $\tilde{.}$ parameterisation without loss of generality.

Levenburg-Marquardt algorithm

Unlike in linear regression, where ordinary least squares provides an analytic solution to the global minimum, there is no dominant algorithm for non-linear least squares. Linear approximations are effective once the current estimate is in the neighbourhood of a minimum, but can display erratic or divergent characteristics if initialised in the wrong place [105].

I first attempted to use the Nelder-Mead[140] algorithm to minimise the squared loss with respect to $\{\sigma_\beta^2, (\sigma_{w_l}^2)_{l=1}^L, \sigma_e^2\}$. Nelder-Mead is a numerical method used to find the minimum of a function by using a simplex to explore the function surface.

It uses a series of heuristics to expand, contract or reflect points in is the simplex as appropriate. Nelder-Mead is advantageous when the gradient is unknown or hard to compute, because it only requires the ability to evaluate the objective function at different points. However I found that it look a long time to converge, especially as the number of environmental variables increased (results not shown).

Instead I used the Levenburg-Marquardt[141] algorithm, which is commonly used for non-linear least squares problems. The algorithm effectively interpolates between the Gauss-Newton algorithm and the method of steepest gradient descent, by use of an adaptive damping parameter. In this manner, it is more robust than the straight forward Gauss Newtown algorithm but should have faster convergence than a gradient descent approach.

Without loss of generality, consider the following model non-linear regression problem

$$Y = f(\theta) + \epsilon. \tag{5.6}$$

Given a starting point θ_0, Levenburg-Marquardt proposes a new point $\theta_{\text{new}} = \theta_0 + \delta$ by solving the normal equations

$$\left(J(\theta_0)^T J(\theta_0) + \mu I\right) \delta = J(\theta_0)^T \epsilon(\theta_0), \tag{5.7}$$

where $J(\theta_0) = \frac{\delta f(\theta_0)}{\delta \theta_0}$ and $\epsilon(\theta_0) = Y - f(\theta_0)$ are respectively the Jacobian and the residual vector evaluated at θ_0.

If θ_{new} has lower squared error than θ_0, then the step is accepted and the adaptive damping parameter μ is reduced. Otherwise, μ is increased and a new step δ is proposed. For small values of μ Equation (5.7) approximates the quadratic step appropriate for a fully linear problem, whereas for large values of μ Equation (5.7) behaves more like steepest gradient descent. This allows the algorithm to defensively navigate regions of the parameter space where the model is highly non-linear. If $\theta + \delta$ reduces the squared error, then the step is accepted and μ is reduced, otherwise μ is increased and a new step δ is proposed. A complete description of the algorithm is given in Appendix A.3.

To apply Levenburg-Marquardt, all that is needed are efficient ways to compute $J(\theta)^T J(\theta)$, $J(\theta)^T \epsilon(\theta)$ and the squared error $S(\theta)$. Equation (5.6) is equivalent to Equation (5.5) if

$$\begin{aligned}
\theta &= \{\sigma_\beta^2, w, \sigma_e^2\}, \\
Y &= \text{vec}\left(yy^T\right), \\
f(\theta) &= \sigma_\beta^2 \text{vec}\left(K\right) + \sum_{l,m} w_l w_m \text{vec}\left(\text{diag}\left(E_l\right) K \text{diag}\left(E_m\right)\right) + \sigma_e^2 \text{vec}\left(I\right).
\end{aligned}$$

Preprocessing

In a preprocessing phase, the genotypes are streamed from file and the following N-vectors are stored in RAM

$$\begin{aligned}
u_b &= XX^T W z_b, \\
v_{b,l} &= XX^T E_l W z_b
\end{aligned}$$

where z_b for $1 \leq b \leq B$ are random N-vectors drawn from a Gaussian distribution with mean zero and identity covariance, W denotes the residual maker matrix $I - C^T(C^T C)^{-1} C$ for any covariates. If no covariates are present then this is just the identity.

In addition I also compute the following quantites

$$\begin{aligned}
&y^T W X X^T W y \\
&E_l^T \text{diag}\left(Wy\right) XX^T \text{diag}\left(Wy\right) E_m \text{for } 1 \leq l, m \leq L,
\end{aligned}$$

which can also be computed as genotypes are streamed from file.

Efficient computation

Define $v_b(w) = \sum_l w_l v_{b,l}$. In this manner the trace of $K_2(w)$ is efficiently be computed by

$$\text{tr}\left(K_2(w) K_2(w)\right) \approx \frac{1}{M^2 B} \sum_b ||v_b(w)||_2^2.$$

Let $(J^TJ)_{\theta_i,\theta_j}$ to denote the entry of the J^TJ that corresponds to $\frac{f(\theta)}{\partial\theta_i}^T\frac{f(\theta)}{\partial\theta_i}$ for $\theta_i,\theta_j \in \{w,\sigma^2_\beta,\sigma_e\}$. Therefore the $(L+2)\times(L+2)$ matrix $J(\theta)^TJ(\theta)$ is given by

$$\begin{aligned}
\left(J^TJ\right)_{w_l,w_m} &= \mathrm{tr}\left(\mathrm{diag}\left(\eta\right)K\mathrm{diag}\left(E_l\right)\mathrm{diag}\left(E_m\right)K\mathrm{diag}\left(\eta\right)\right),\\
&= \frac{1}{M^2B}\sum_b\left(v_b(w)^T\mathrm{diag}\left(E_l\right)\mathrm{diag}\left(E_m\right)v_b(w)\right),\\
\left(J^TJ\right)_{w_l,\sigma^2_\beta} &= \mathrm{tr}\left(\mathrm{diag}\left(\eta\right)K\mathrm{diag}\left(E_l\right)K\right),\\
&= \frac{1}{M^2B}\sum_b\left(v_b(w)^T\mathrm{diag}\left(E_l\right)u_b\right),\\
\left(J^TJ\right)_{\sigma^2_\beta,\sigma^2_\beta} &= \mathrm{tr}\left(KK\right),\\
&= \frac{1}{M^2B}\sum_b||u_b||^2_2,\\
\left(J^TJ\right)_{\sigma^2_\beta,\sigma^2_e} &= \mathrm{tr}\left(K\right),\\
&= \frac{1}{M^2B}\sum_b z_b^TWu_b,\\
\left(J^TJ\right)_{w_l,\sigma^2_e} &= \mathrm{tr}\left(\mathrm{diag}\left(\eta\right)K\mathrm{diag}\left(E_l\right)\right),\\
&= \frac{1}{M^2B}\sum_b z_b^TWv_b(w),\\
\left(J^TJ\right)_{\sigma^2_e,\sigma^2_e} &= \mathrm{tr}\left(W\right).
\end{aligned}$$

$J(\theta)^T\epsilon(\theta)$ is given by

$$\begin{aligned}
\left(J(\theta)^T\epsilon(\theta)\right)_{\sigma^2_\beta} &= \mathrm{tr}\left(y^TWKWy\right) - J(\theta)^TJ(\theta)\sigma^2_\beta,\\
\left(J(\theta)^T\epsilon(\theta)\right)_{w_l} &= \mathrm{tr}\left(y^TW\mathrm{diag}\left(E_l\right)K\mathrm{diag}\left(Ew\right)Wy\right) - J(\theta)^TJ(\theta)w_l,\\
\left(J(\theta)^T\epsilon(\theta)\right)_{\sigma^2_e} &= \mathrm{tr}\left(y^TWy\right) - J(\theta)^TJ(\theta)\sigma^2_e.
\end{aligned}$$

I note that $\mathrm{tr}\left(y^TWKWy\right)$ is computed in the preprocessing stage, and that

$$\mathrm{tr}\left(y^TW\mathrm{diag}\left(E_l\right)K\mathrm{diag}\left(Ew\right)Wy\right) = \sum_m E_l^T\mathrm{diag}\left(Wy\right)XX^T\mathrm{diag}\left(Wy\right)E_m.$$

Finally $S(\theta)$ is given by

$$\begin{aligned}
S(\sigma^2_\beta, w) &= ||(yy^T - \text{Cov}\,(y))||^2_F, \\
&= \text{tr}\left((yy^T - \text{Cov}\,(y))(yy^T - \text{Cov}\,(y))\right), \\
&= \text{tr}\left(yy^T yy^T\right) - 2\begin{pmatrix}\sigma^2_\beta \\ 1 \\ \sigma^2_e\end{pmatrix}^T \begin{pmatrix}\text{tr}\left(y^T K y\right) \\ \text{tr}\left(y^T K_2(w) y\right) \\ \text{tr}\left(y^T y\right)\end{pmatrix} \\
&\quad + \begin{pmatrix}\sigma^2_\beta \\ 1 \\ \sigma^2_e\end{pmatrix}^T \begin{pmatrix}\text{tr}\,(KK) & \text{tr}\,(KK_2(w)) & \text{tr}\,(K) \\ \text{tr}\,(KK_2(w)) & \text{tr}\,(K_2(w)K_2(w)) & \text{tr}\,(K_2(w)) \\ \text{tr}\,(K) & \text{tr}\,(K_2(w)) & N\end{pmatrix} \begin{pmatrix}\sigma^2_\beta \\ 1 \\ \sigma^2_e\end{pmatrix}
\end{aligned}$$

Complexity

The initial preprocessing step has costs $\mathcal{O}(NMLB + NML^2)$ in compute and $\mathcal{O}(NLB)$ in RAM.

The remaining algorithm does not require much RAM in addition to that required in the preprocessing step, so also costs $\mathcal{O}(NLB)$ in RAM. Construction of the summary variable $v_b(w) = \sum_l w_l v_{b,l}$ costs $\mathcal{O}(NLB)$ in compute. Each iteration of the Levenburg-Marquardt algorithm costs $\mathcal{O}(NL^2B)$.

It is possible to parallelise RHE-NLS using OpenMPI by partitioning samples across cores, in a similar manner to that used by LEMMA (see Section 2.3.1). Given that evaluating the objective function $S(\sigma^2_\beta, w)$ is characterised by BLAS level 1 array operations, a distributed algorithm using OpenMPI should have superior runtime versus an the same algorithm using OpenMP as well as providing RAM limited only by the size of a researchers compute cluster.

Initialisation

The Levenburg-Marquardt is only guaranteed to find a local minimum. Therefore I perform 10 repeats of the Levenburg-Marquardt algorithm with different initialisations, and keep results from the solution with lowest squared error $S(\hat{\theta})$.

Each run is initialised with a vector of interaction weights $\tilde{w}$, where each entry set to $\frac{1}{L}$ and a smal amount of Gaussian noise is added.

$$\tilde{w} = \frac{1}{L}\vec{1} + \mathcal{N}(0, \frac{2}{L^2} I_L).$$

To transform the initial weights vector $\tilde{w}$ to the initial parameters θ_0 I do the following. Let $(\hat{\sigma}^2_\beta, \hat{\sigma}^2_\gamma, \hat{\sigma}^2_e)$ be solutions to

$$(\hat{\sigma}^2_\beta, \hat{\sigma}^2_\gamma, \hat{\sigma}^2_e) = \text{argmin}_{\sigma^2_\beta, \sigma^2_\gamma, \sigma^2_e} ||yy^T - \left(\sigma^2_\beta K + \sigma^2_\gamma K_2(\tilde{w}) + \sigma^2_e I\right)||^2_F.$$

Then I initialise the RHE-NLS algorithm with $\theta_0 = (\hat{\sigma}^2_\beta, w, \hat{\sigma}^2_e)$ where $w = \sigma_\gamma \tilde{w}$.

5.1.3 Relationship between RHE-NLS and RHE-multiE

The RHE-NLS model is given by

$$y \sim \mathcal{N}(0, \sigma^2_\beta K + \sigma^2_\gamma K_2(w) + \sigma^2_e I), \tag{5.8}$$

which, when expanded into a Bayesian regression model, can be re-written as

$$\begin{aligned} y &= X\beta + (Ew) \odot X\gamma + \epsilon, \\ \beta &\sim \mathcal{N}(0, \sigma^2_\beta), \\ \gamma &\sim \mathcal{N}(0, \sigma^2_\gamma), \\ \epsilon &\sim \mathcal{N}(0, \sigma^2_e). \end{aligned} \tag{5.9}$$

The RHE-multiE model is given by

$$y \sim \mathcal{N}(0, \sigma^2_\beta K + \sum_l \sigma^2_{w_l} E_l \odot K \odot E_l^T + \sigma^2_e I), \tag{5.10}$$

which, when expanded into a Bayesian regression model, can be re-written as

$$\begin{aligned} y &= X\beta + \sum_l E_l \odot X\lambda_l + \epsilon, \\ \beta &\sim \mathcal{N}(0, \sigma^2_\beta), \\ \lambda_l &\sim \mathcal{N}(0, \sigma^2_{w_l}), \\ \epsilon &\sim \mathcal{N}(0, \sigma^2_e). \end{aligned} \tag{5.11}$$

Comparing Equation (5.9) with Equation (5.11), suggests that RHE-NLS is a contrained version of the RHE-multiE model (under the constraint that $\sum_l \lambda_{j,l} = \gamma_j \sum_l w_l$).

We can expect the two models to give similar heritability estimates, under the simplifying assumptions that GxE interactions do occur with a single linear

combination of the environments and that the set of random variables $\{g, (E_l \odot g)_{l=1}^L\}$ is mutually independent. Let $g \sim \mathcal{N}(0, K)$ and $\epsilon \sim \mathcal{N}(0, \sigma_e^2 I)$. Then connection between the two models is revealed by observing

$$\begin{aligned} Y &= \mathcal{N}(0, \sigma_\beta^2 K + \sigma_\gamma^2 K_2(w) + \sigma_e^2 I), \\ &= \sigma_\beta g + \left(\sigma_\gamma \sum_l w_l E_l\right) \odot g + \epsilon, \\ &= \sigma_\beta g + \sum_l \sigma_{w_l} E_l \odot g + \epsilon, \\ &= \mathcal{N}\left(0, \sigma_\beta^2 K + \sum_l \sigma_{w_l}^2 E_l \odot K \odot E_l^T + \sigma_e^2 I\right), \end{aligned}$$

where $\sigma_{w_l}^2 = \sigma_\gamma^2 w_l^2$. Thus we should expect both models to have the same estimate for the proportion of variance explained by GxE interaction effects.

Even in that case that RHE-multiE and RHE-NLS have the same expected heritability estimate, there are still some differences between the two. RHE-NLS is a constrained model, so the variance of its heritabiity estimates may be smaller. Further, although $\hat{\sigma}_{w_l}^2$ is proportional to the square of the weights used to construct the ES the sign of the interaction weight w_l has been lost. Thus it is not possible to reconstruct an ES for use in single SNP testing from RHE-multiE.

5.2 Simulation results

5.2.1 Data

As in Chapter 3, I used genetic data subsampled from genotyped SNPs in the UK Biobank[110], drawing SNPs from all 22 chromosomes in proportion to chromosome length and using unrelated samples of mixed ancestry ($N = 25k$; 12,500 white British, 7,500 Irish and 5,000 white European, $N = 50k$; 29,567 white British, 7,500 Irish and 12,568white European, $N = 100k$; 79,567 white British, 7,500 Irish and 12,568 white European; using self-reported ancestry in field *f.21000.0.0*). All samples were genotyped using the UKBB genotype chip and were included in the internal principal component analysis performed by the UK Biobank. Environmental variables were simulated from a standard Gaussian distribution.

I performed simulations where the phenotype was simulated from the LEMMA model and where the phenotype depended on the squared effect of a heritable environmental variable. These phenotype simulations strategies are described in detail in Sections 3.2.1 and 3.3.1 respectively.

The default parameter choices were: $N = 25K; M = 100K; L = 30; \text{active} - L = 6, M_{\text{causal, G}} = 2,500; M_{\text{causal, GxE}} = 1,250; h^2_\beta = 20\%; h^2_\gamma = 5\%$.

For avoidance of doubt, the model 'LEMMA' in the simulation results presented below refers to running the WGR model described in Chapter 2 and subsequently applying RHE to estimate the proportion of trait variance explained by multiplicative GxE effects as described in Section 2.4.

5.2.2 Results

Comparison on baseline simulations

I first compared LEMMA, RHE-NLS and RHE-multiE in simulations where the phenotype was simulated from the LEMMA model (referred to as the baseline phenotype).

The top row of Figure 5.1 compares estimates of the variance explained by GxE effects from all three methods. In general, all methods had upwards bias that decreased with sample size and increased with the number of environments. While heritability estimates from LEMMA and RHE-NLS appeared quite similar, estimates from RHE-multiE had much higher variance and also appeared to have higher upwards bias as the total number of environments increase (see 5.1f).

The bottom row of Figure 5.1 compares the absolute correlation between the simulated ES and the ES inferred by LEMMA and RHE-NLS. In general, the estimated ES from RHE-NLS had lower absolute correlation with the true ES than the estimated ES from LEMMA, implying that RHE-NLS has lower power. However, in large sample sizes ($N = 100k$), both methods achieve a correlation of over 0.98 with the simulated ES.

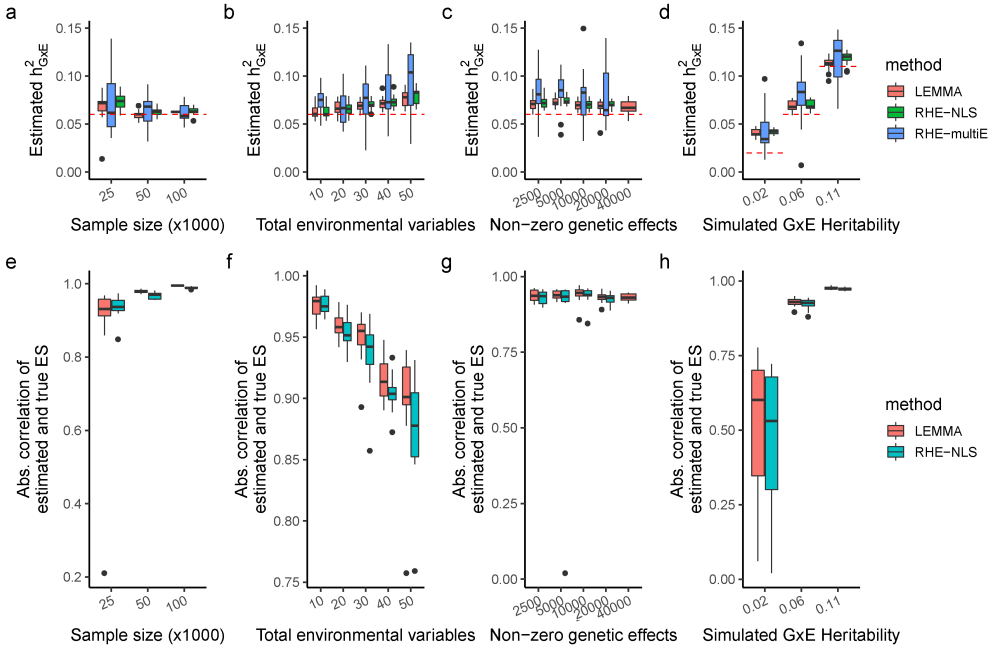

Figure 5.1: Comparison on baseline simulations. The top row shows estimates of the proportion of variance explained by GxE effects by LEMMA, RHE-NLS and RHE-multiE, whilst varying the number of environments, the number of active environments and the number of non-zero SNP effects and GxE heritability. The bottom row compares the absolute correlation between the true ES and the ES inferred by LEMMA and RHE-NLS. Results from 15 repeats shown.

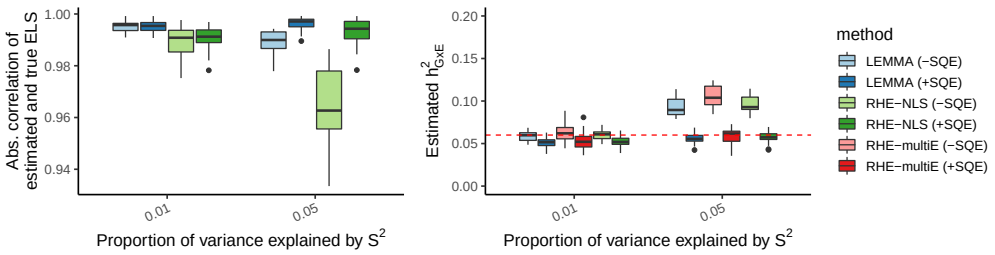

Figure 5.2: Comparison on simulations with a misspecified heritable environment Estimated proportion of trait variance explained by GxE effects is shown on the left, absolute correlation between the inferred ES and the true ES shown in the right. Results shown using LEMMA, RHE-NLS and RHE-multiE. Phenotypes simulated with a squared effect from a heritable confounder (see Section 3.3.1). Results from 20 repeats shown. Abbreviations; (-SQE), no attempt to control for squared effects; (+SQE), squared effects with $p < 0.01$ (Bonferroni correction for multiple envs) included as covariates

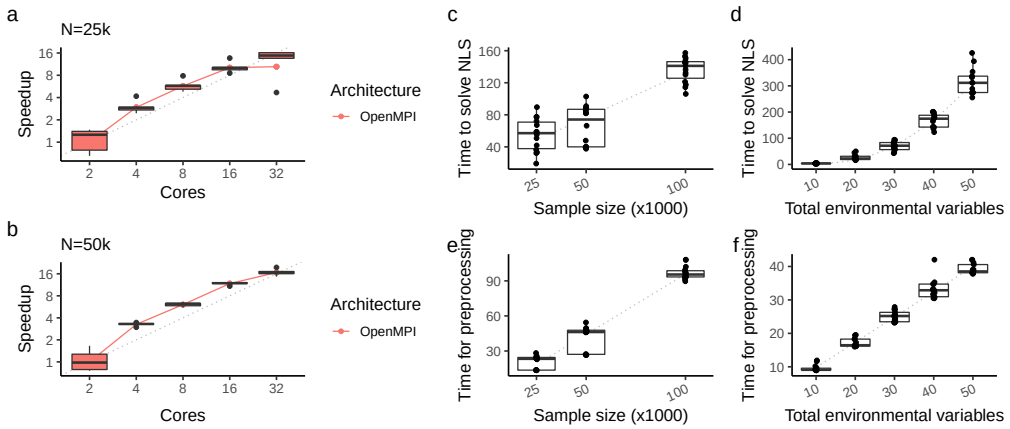

Figure 5.3: Computational complexity of RHE-NLS in simulation. Strong scaling of RHE-NLS using OpenMPI to parrallelise across cores with (a) $N = 25k$ samples and (b) $N = 50k$ samples. Comparison of the runtime of the Levenburg-Marquardt algorithm (c, d) and runtime of the preprocessing step (e, f). By default each run used; four cores, $N = 25k$ samples, $L = 30$ environments and 10 random starts of the Levenburg-Marquardt algorithm. Results from 15 repeats shown.

Comparison with a misspecified heritable environment

Subsequently I compared LEMMA, RHE-NLS and RHE-multiE in simulation where the functional form of a heritable environmental variable was misspecified (or more specifically; the phenotype depended on the squared effect of a heritable environment). Results are shown in Figure 5.2. Following the approach used in Section 3.3, I tested all methods without any attempt to control for model misspecification and using a preprocessing strategy where each environment was tested for squared effects and significant squared effects were included as covariates. These are referred to as *(-SQE)* and *(+SQE)* respectively in the figures. Using the *(-SQE)* strategy, all methods showed upwards bias in estimates of GxE heritability that increased with the strength of the squared effect on the phenotype. Model misspecification also caused bias in the ES of both RHE-NLS and LEMMA, however bias in the ES from RHE-NLS appeared to be much worse. Using the $(+SQE)$ strategy, all GxE heritability estimates were unbiased, consistent with earlier simulation results.

Computational complexity in simulation

Figure 5.3 displays simulation results on the computational complexity of RHE-NLS. Figures 5.3a and 5.3b show that RHE-NLS achieved perfect strong scaling[1] on the range of cores tested. This suggests that RHE-NLS has superior scalability to LEMMA, as for LEMMA the speedup due to increased cores began to decay after the number of samples per core dropped below $N = 3k$ (see Figure 3.7).

Time to compute the preprocessing step and solve the non-linear least squared problem are shown in Figures 5.3c to 5.3f, while the number of environments and sample size were varied. As expected, the preprocessing step appeared to be linear in both the number of environments and sample size. Time to solve the non-linear least squares problem appeared to be quadratic in the number of environments and approximately linear in sample size N. As a single Levenburg-Marquardt iteration should have complexity $\mathcal{O}(NL^2B)$, this suggests that the number of iterations required for convergence of the Levenburg-Marquardt algorithm was independent of sample size and the number of environments (at least for the range of values tested).

Finally, to give a direct comparison between LEMMA and RHE-NLS, I ran each method on simulated data with $N = 100k$ samples, $M = 100k$ SNPs and $L = 30$ environmental variables using 4 cores for each run. Over 20 repeats, LEMMA took an average of 648 minutes to run whereas RHE-NLS took an average of 233 minutes.

5.3 Discussion

In this chapter, I have presented two new approaches, referred to as RHE-multiE and RHE-NLS, for modelling GxE interactions with multiple environmental variables. The former only estimates GxE heritability, whereas the latter estimates both GxE heritability and an ES that interacts with SNPs genome-wide. This ES is very similar to the ES estimated by LEMMA; The difference being that the RHE-NLS model assumes explicitly that interaction weights are unconstrained and implicitly that a Gaussian is placed on SNP coefficients.

[1] A parallel algorithm has perfect strong scaling if the runtime on T processors is linear in $\frac{1}{T}$, including communication costs.

I compared all three of these methods in simulation. Estimates of GxE heritability from RHE-multiE had much higher variance than estimates from LEMMA and RHE-NLS, suggesting that the usefulness of RHE-multiE is limited. Although RHE-NLS had less power than LEMMA to estimate the true ES, both methods were reasonably accurate in estimating the ES in large samples. Both methods also produced similar estimates of GxE heritability. The primary advantage of RHE-NLS over LEMMA is in computational complexity, as the empirical complexity of RHE-NLS appeared to be linear in sample size whereas LEMMA was shown to be super-linear in Section 3.2.2. Therefore for the purposes of biobank scale datasets with hundreds of thousands of samples, the RHE-NLS is likely to be more computationally tractable than LEMMA.

A comparison with LEMMA is not entirely fair, because LEMMA is also able to perform single SNP hypothesis testing whereas RHE-NLS (currently) does not. Several approaches could use the ES estimated by RHE-NLS for hypothesis testing. The simplest approach would be to use standard linear regression, with genetic PCs to control for population structure, after running RHE-NLS. Explicitly, this could be achieved by fitting the model

$$y = C\alpha_C + \hat{\eta}\alpha_E + \hat{\eta} \odot C\alpha_{\text{CxE}} + x_{\text{test}}\beta_{\text{test}} + \hat{\eta} \odot x_{\text{test}}\gamma_{\text{test}} + \epsilon, \tag{5.12}$$

where $\hat{\eta}$ denotes the estimated ES and C is a matrix of covarites including genetic PCs. Alternatively, we could expect that an LMM with two kinship matrices (one for additive genetic similarity and one for multiplicative GxE similarity) might yield superior performance. This could be achieved with the model

$$y \sim \mathcal{N}(\hat{\eta}\alpha + x_{\text{test}}\beta_{\text{test}} + \hat{\eta} \odot x_{\text{test}}\gamma_{\text{test}}, \sigma_g^2 K + \sigma_{gxe}^2 \hat{\eta} \odot K \odot \hat{\eta} + \sigma_e^2 I), \tag{5.13}$$

where $K = \frac{XX^T}{M}$ and X is the genotype matrix normalised to have mean zero and variance one. In the same way that an LMM has better power and control of confounding that linear regression for a standard GWAS[25], the LMM in Equation (5.13) should have better power and control of confounding than linear regression in Equation (5.12). Naively fitting this model on large datasets with hundreds of

thousands of samples is computationally prohibitive. However recent advances[142] have suggested that, for a standard GWAS in biobank scale datasets, the kinship matrix K can be replaced with a sparse approximation. Preliminary evidence suggested that this approach yields vastly superior computational performance with minimal sacrifice in model performance[142]. Finally, the estimated ES could be used as the only environmental variable in the WGR regression algorithm implemented by LEMMA. However, exploring these approaches is reserved for future work.

6
Conclusion

This primary contribution of this thesis is a novel approach, LEMMA, for modelling multiplicative GxE interactions with multiple environmental variables in quantitative traits. The core of LEMMA is a Bayesian WGR algorithm that jointly models the main effect and GxE interaction effect of SNPs genome-wide. Instead of modelling a separate GxE interaction effect for each SNP and each environment, LEMMA assumes that each SNP has a GxE interaction with a weighted average of the environmental variables (referred to as the environmental score or ES). This assumption corresponds to a hypothesis that the same environmental variable (or combination of variables) might be involved in GxE interactions at multiple locations across the genome.

In Chapter 3, I compared LEMMA to existing single SNP methods that model multiplicative GxE interactions with multiple environmental variables. In simulations where the hypothesis underlying LEMMA was true, I showed that LEMMA has much higher power to identify GxE interactions than existing methods. I also showed that the estimated ES could be combined with randomised HE-regression to estimate the proportion of variance explained by multiplicative GxE interaction effects. This approach is biased upwards, but with bias that reduces with sample size.

Chapter 4 contained an analysis of GxE interactions in BMI, PP, SBP and DBP, using approximately 280,000 white British participants and 42 environmental variables from the UK Biobank. I estimated that 9.3%, 12.5%, 3.9% and 1.6%, respectively, of phenotypic variance, is explained by GxE interactions with the ES, and that rare variants explain most of this variance. LEMMA also identified three loci that interact with the estimated environmental scores at genome-wide significance levels ($-\log_{10} p > 7$). In a comparison of LEMMA against StructLMM, the F-Test and the robust F-test on BMI with the same set of environmental variables, I showed that LEMMA is powered to detect loci that are distinct from those detected by the single SNP methods. This suggests that the underlying hypothesis of LEMMA has potential (at least for BMI). I also showed that state-of-the-art method StructLMM[55] shows substantial inflation when applied to BMI, which I suggest is due to conditional heteroskedasticity.

There are several ways in which LEMMA could be improved. The central hypothesis (that all GxE interactions occur with a single environmental score) is a strong one. I have suggested that this hypothesis could be relaxed by partitioning SNPs into groups (perhaps based on annotations) and learning a single ES for each group. Alternatively, I have suggested that a more complex generalisation could be achieved by learning more than one ES per group. Both of these could be pursued in future work.

At the moment LEMMA has only been tested in traits with a continuous phenotype. BOLT-LMM[75, 79], a WGR algorithm that models the SNP additive effects in a manner similar to LEMMA, has been shown to be statistically well-calibrated for balanced case-control studies and to have inflated type I error in unbalanced case-control studies[143]. This suggests that LEMMA may also be able to run on balanced case-control studies; however, this has not yet been tested. Extensions to case-control studies are a topic for future work.

Finally, LEMMA is a computationally intensive algorithm. Even though I have implemented several approaches to make LEMMA scalable, such as use of OpenMPI for parallelisation or the SQUAREM algorithm, LEMMA still required

over a week to converge using 32 cores in the analysis of PP presented in Chater 4. In chapter 5, I explored a different approach, RHE-NLS, that estimates a similar ES to that estimated by LEMMA and its contribution to GxE heritability. RHE-NLS appeared to have less power than LEMMA but superior computational efficiency, suggesting that RHE-NLS might be more appropriate for GxE analyses with hundreds of thousands of samples. In simulation, RHE-NLS had empirical computational complexity that was linear in sample size and quadratic in the number of environments. I have also suggested ways in which RHE-NLS could be extended to perform single SNP hypothesis testing in a manner that is computationally efficient and is likely to have similar power and calibration as LEMMA does. However this approach is, again, the subject of future work.

Appendices

A
Appendix

Contents

A.1 Derivation of mean field Coordindate Ascent Variational Inference

We use Coordinate Ascent Variational Inference to optimize the ELBO $\mathcal{F}(\nu;\phi)$ [91].

Recall that the evidence lower bound (ELBO) is defined by

$$\begin{aligned}\mathcal{F}(\nu;\phi) &:= \mathbb{E}_q\left[\log \frac{p(\theta, y|\mathcal{D},\phi)}{q(\theta;\nu)}\right], \\ &:= \mathbb{E}_q\left[\log p(y|\mathcal{D},\phi) + \log p(\theta|y,\mathcal{D},\phi) - \log q(\theta;\nu)\right],\end{aligned}$$

Using the mean field assumption, that the variational distributions factorise, we

can rewrite the ELBO as

$$\mathcal{F}(\nu;\phi) = \log(p(y|\phi) + \mathbb{E}_q\left[\log p(\theta|y,\phi)\right] - \sum_j \mathbb{E}_q\left[\log q(\theta_j)\right].$$

For notational convenience, dependance on data $\mathcal{D}$ and variational parameters ν have been suppressed. Using C to denote a constant that is not necessarily the same between lines, we can consider $\mathcal{F}$ as a function of $q(\theta_j)$

$$\begin{aligned}
\mathcal{F}_j &= \mathbb{E}_q\left[\log p(\theta|y,\phi)\right] - \mathbb{E}_q\left[\log q(\theta_j)\right] + C, \\
&= \int \prod_i q(\theta_i)\left\{\log p(\theta|y,\phi)\right\} d\theta - \int q(\theta_j)\log q(\theta_j) d\theta_j + C, \\
&= \int q(\theta_j)\left\{\int \prod_{i\neq j} q(\theta_i)\log p(\theta|y,\phi) d\theta_{-j}\right\} d\theta_j - \int q(\theta_j)\log q(\theta_j) d\theta_j + C, \\
&= \int q(\theta_j)\left\{\mathbb{E}_{-q(\theta_j)}\left[\log p(\theta|y,\phi)\right]\right\} d\theta_j - \int q(\theta_j)\log q(\theta_j) d\theta_j + C, \\
&= -KL\left(q(\theta_j)|\mathbb{E}_{-q(\theta_j)}\left[\log p(\theta|y,\phi)\right]\right) + C,
\end{aligned}$$

where $\mathbb{E}_{-\theta_j}$ is the expectation taken over all variables except θ_j. From the last line it is clear that to maximise $\mathcal{F}_j$ we minimise the KL divergence between $\log q(\theta_j)$ and $\mathbb{E}_{-q(\theta_j)}\left[\log p(\theta|y,\phi)\right]$. This minimization occurs when

$$\log q^*(\theta_j) = \mathbb{E}_{-q(\theta_j)}\left[\log p(\theta|y,\mathcal{D},\phi)\right] + C.$$

Equivalently, the CAVI identity could be stated as one of the following

$$q^*(\theta_j) \propto \exp \mathbb{E}_{-q(\theta_j)}\left[\log p(\theta|y,\mathcal{D},\phi)\right], \tag{A.1}$$

$$q^*(\theta_j) \propto \exp \mathbb{E}_{-q(\theta_j)}\left[\log p(\theta_j|\theta_{-j},y,\mathcal{D},\phi)\right] \tag{A.2}$$

A.2 Details of Randomised HE-regression

HE-regression is a method of moments estimator that minimizes the squared difference between the observed and expected covariance from the folowing model

$$y \sim \mathcal{N}(0, \sigma_g^2 K + \sigma_e^2 I), .$$

More explicitly, HE-regression solves the following optimization problem

$$(\hat{\sigma}_g^2, \hat{\sigma}_e^2) = \text{argmin}_{\sigma_g^2,\sigma_e^2} ||yy^T - \left(\sigma_g^2 K + \sigma_e^2 I\right)||_F^2.$$

Recent work by Wu and Sankararaman [86] showed that HE-regression can be computed efficiently using Hutchkinsons estimator [88]. This is termed randomised HE-regression. Subsequently, Pazokitoroudi et al. [87] showed that heritability estimates for the multiple component model

$$y \sim \mathcal{N}(0, \sum_k \sigma_k^2 K_k + I\sigma_e^2)$$

can also be efficiently obtained as solution to the linear system given by

$$\begin{pmatrix} T & b \\ b^T & N \end{pmatrix} \sigma^2 = \begin{pmatrix} \boldsymbol{c} \\ N \end{pmatrix} \tag{A.3}$$

where

$$T_{kl} = \operatorname{tr}(K_k K_l) \tag{A.4}$$

$$b_k = \operatorname{tr}(K_k) \tag{A.5}$$

$$c_k = y^T K_k y. \tag{A.6}$$

Using Hutchkinsons estimator as in the single trait model allows the algorithm to scale in $\mathcal{O}(NMB)$.

A.2.1 Controlling for covariates

To control for covariates, the matrix of covariates C can be efficiently regressed out of all terms in the model without changing how the algorithm scales. Let $W = I - C^T(C^TC)^{-1}C$. Then by re-writing Appendix A.2 as a bayesian regression model

$$\begin{aligned} y|\beta, \epsilon &= X\beta + \epsilon, \\ \beta &\sim \mathcal{N}(0, \sigma_\beta^2), \\ \epsilon &\sim \mathcal{N}(0, \sigma_e^2), \end{aligned}$$

it is clear that applying W to all terms in the bayesian regression has the effect of projecting covariates out of the system

$$Wy \sim \mathcal{N}(0, \sigma_\beta^2 WKW + W\sigma_e^2).$$

Therefore Equation (1.4) with covariates projected out is given by

$$\begin{aligned} T_{kj} &= \operatorname{tr}\left(WK_kWK_jW\right), \\ b_k &= \frac{\operatorname{tr}\left(WX_kX_k^T\right)}{M_k} \\ c_k &= y^TWK_kWy, \end{aligned}$$

using the property that W is idempotent.

A.2.2 Block jacknife variance estimator

Statistical resampling methods are a useful tool in computational statistics for parametric estimation, hypothesis testing and model validation. Commonly used resampling methods include permutation tests, the bootstrap and the jacknife.

The jacknife has previously been used to estimate the variance of HE-regression heritability estimates, with one individual is left out at a time [85]. Such an approach is not efficient using the randomised HE regssion framework. Instead Pazokitoroudi *at al.*[87] use the block jacknife, leaving out blocks of SNPs.

Let

$$A = \begin{Bmatrix} T & b \\ b^T & N \end{Bmatrix}$$

so that the system we want to solve is given by

$$A\boldsymbol{\sigma^2} = \begin{Bmatrix} c \\ y^Ty \end{Bmatrix}$$

Partition the columns of genotype matrix X into J disjoint blocks $X^1, ..., X^J$ such that $X = [X^1, ..., X^J]$, where X^j is an $N \times M_j$ matrix.

To compute the j'th jacknife replicate of h_k^2, the heritability estimate for the k'th component, we first solve

$$A_{[j]}\boldsymbol{\sigma}^{\mathbf{2}}_{[j]} = \begin{Bmatrix} c_{[j]} \\ y^Ty \end{Bmatrix}$$

where $A_{[j]}$ and $\boldsymbol{c}_{[j]}$ refer to A and $\boldsymbol{c}$ computed excluding SNPs from the j'th block. Then we compute $h^2_{k,[j]}$ as

$$h^2_{k,[j]} = \frac{\sigma^2_{k,[j]}}{\sum_k \sigma^2_{k,[j]}}.$$

Note that *a priori* we expect $h^2_{k,[j]}$ to be an ascending function of the number of SNPs used. Thus we would expect the jacknife estimates of h^2_k to systematically be smaller than the full data estimate. That is that $h^2_{k,[j]} < h^2_k$ on average. Therefore we define the per-SNP heritability

$$\tilde{h}^2_k = \frac{h^2_k}{M}$$
$$\tilde{h}^2_{k,[j]} = \frac{h^2_{k,[j]}}{M - M_j}$$

The per-SNP jacknife variance is then

$$\mathrm{Var}_{\mathrm{Jack}}\left(\tilde{h}^2_k\right) = \frac{J-1}{J}\sum_j \left(\tilde{h}^2_{k,[j]} - \tilde{h}^2_{k,[:]}\right)^2,$$

where

$$\tilde{h}^2_{k,[:]} = \frac{1}{J}\sum_j \tilde{h}^2_{k,[j]}$$

is the mean per-SNP h^2 estimate for the k'th component over all jacknife samples. Finally the jacknife variance over all SNPs is

$$\mathrm{Var}_{\mathrm{Jack}}\left(h^2_k\right) = M^2\mathrm{Var}_{\mathrm{Jack}}\left(\tilde{h}^2_k\right).$$

A.2.3 Efficient computation

Let

$$u^k_b = X_k X_k^T W z_b$$

where k refers to the k'th SNP component, b is the b'th random draw and W is the identity if there are no covariates. Then Equation (A.3) can be solved efficiently by substituting in the following expressions

$$\begin{aligned}
T_{kj} &= \mathrm{tr}\left(WK_kWK_jW\right), \\
&= \frac{1}{BM_kM_j}\sum_b (u^k_b)^T W u^k_b, \\
b_k &= \frac{\mathrm{tr}\left(WX_kX_k^T\right)}{M_k} \\
&= \frac{1}{BM_k}\sum_b (u^k_b)^T W z_b, \\
c_k &= y^T W K_k W y, \\
&= \frac{1}{M_k}||X_k^T W y||_2^2,
\end{aligned}$$

To efficiently compute block jacknife estimates, further define

$$u_{b,j}^k = X_k^{(j)}(X_k^{(j)})^T z_b.$$

Thus

$$y^T K_k^{-j} y = y^T K_k y - y^T K_k^j y, \tag{A.7}$$

$$\mathrm{tr}\left(K_k^{-j}\right) = \frac{1}{B}\frac{1}{M_{k,j}}\sum_b z_b^T K_k z_b - z_b^T K_k^j z_b, \tag{A.8}$$

$$= \frac{1}{B}\frac{1}{M_{k,j}}\sum_b z_b^T u_b^k - z_b^T u_{b,j}^k, \tag{A.9}$$

$$\mathrm{tr}\left(K_k^{-j} K_l^{-j}\right) = \frac{1}{B}\frac{1}{M_{k,j}M_{l,j}}\sum_b z_b^T (K_k - K_k^j)(K_l - K_l^j) z_b, \tag{A.10}$$

$$= \frac{1}{B}\frac{1}{M_{k,j}M_{l,j}}\sum_b (u_b^k - u_{b,j}^k)^T (u_b^l - u_{b,j}^l). \tag{A.11}$$

So we just need to store N-vectors $u_{b,j}^k$ for $1 \le b \le B$, $1 \le j \le J$, $1 \le k \le K$ which costs $8NBJK$ bytes of RAM.

A.3 Levenburg-Marquardt algorithm

The Levenburg Marquardt algorithm described here is almost identical to that described by Madsen, Nielsen, and Tingleff [144], however it is included here for completeness.

Levenburg Marquardt[141] is a commonly used algorithm for non-linear least squares problems, which interpolates between the Gauss-Newton algorithm and the method of steepest gradient descent by using an adaptive damping parameter. Consider the problem

$$y = f(x;\theta) + \epsilon.$$

If $f(x;\theta)$ is linear in the parameters then the normal least squares solution can be applied. If not then there is no analytic expression for the global maximum. Instead Levenburg-Marquardt applies the iterative solution

$$(J^T J + \mu I)\delta = J^T(y - f(x;\theta)),$$
$$= J^T \epsilon(\theta))$$

Levenburg Marquardt is an iterative algorithm. At time t, the algorithm proposes a new point $\theta_{\text{new}} = \theta_t + \delta$ by solving the normal equations

$$\left(J(\theta_t)^T J(\theta_t) + \mu I\right) \delta = J(\theta_t)^T \epsilon(\theta_t), \tag{A.12}$$

where μ is an adaptive damping parameter, and $J(\theta_t) = \frac{\delta f(\theta_t)}{\delta \theta_t}$ and $\epsilon(\theta_t) = Y - f(\theta_t)$ are respectively the jacobian and the residual vector evaluated at θ_t.

The proposed step θ_{new} is accepted if it results in a positive value of the *gain ratio*. The gain ratio ρ is defined by

$$\rho = \frac{S(\theta_t) + S(\theta_{\text{new}})}{\frac{1}{2}\delta^T(\mu\delta + J(\theta_t)^T \epsilon(\theta_t)}.$$

I use two stopping criteria. At a local minimum, one would expect the gradient $\frac{\delta}{\delta\theta} S(\theta_t) = J(\theta_t)^T \epsilon(\theta_t) = \vec{0}$. Therefore the first criterion should reflect if the algorithm is at a local minimum, and is given by

$$||J(\theta_t)^T \epsilon(\theta_t)||_\infty \leq \epsilon_1,$$

where ϵ_1 is some small constant. The second stopping criterion is given by

$$||\delta||_2 \leq \epsilon(||\theta_t||_2 + \epsilon),$$

where ϵ_2 is some small constant. The second criterion terminates the algorithm if the proposed step size is too small.

The adaptive damping scheme is given by

$$\begin{aligned} \rho &= \frac{S(\theta) - S(\theta + \delta)}{L(0) - L(\delta)}, \\ &= \frac{S(\theta) - S(\theta + \delta)}{0.5 * (\mu\delta^T\delta + \delta^T J^T \epsilon(\theta))} \end{aligned}$$

The full algorithm is described in Algorithm 1, and a function for the gain ratio is given in Algorithm 2. Thus the Levenburg marquardt algorithm can be applied given values for $S(\theta), J(\theta)^T J(\theta)$ and $J(\theta)^T \epsilon(\theta)$.

Input: θ_0
Output: $\hat{\theta}$
begin
  `// Constants`
  $v = 2; \quad \tau := 10^{-3}; \quad \epsilon_1 = 10^{-15}; \quad \epsilon_2 = 10^{-15}$
  `// Initialise parameters`
  stop = false
  $t = 0$
  $A := J(\theta_t)^T J(\theta_t); \quad g := J(\theta_t)^T \epsilon(\theta_t)$
  $\mu := \tau \times \max\{\text{diag}\,(A)\}$
  `// Start loop`
  while $t < T_{max}$ *and !stop* **do**
    $A := J(\theta_t)^T J(\theta_t); \quad g := J(\theta_t)^T \epsilon(\theta_t)$
    $\delta = (A + \mu I)^{-1} g$
    if $||\delta||_2 \leq \epsilon_2(||\theta_t||_2 + \epsilon_2)$ **then**
      stop = true
    else
      $\theta_{\text{new}} = \theta_t + \delta$
      $\rho :=$ `GainRatio(`θ_t, δ, μ`)`
      if $\rho > 0$ **then**
        `// Accept step`
        $\theta_{t+1} = \theta_{\text{new}}$
        $t = t + 1$
        `// Reduce damping parameter`
        $\mu = \mu \times \max\{1/3, 1 - (2\rho - 1)^3\}$
        $v = 2$
        `// Check termination`
        $g := J(\theta_t)^T \epsilon(\theta_t)$
        stop = $(||g||_\infty \leq \epsilon_1)$
      else
        `// Increase damping parameter`
        $\mu = \mu \times v$
        $v = 2v$
      end
    end
  end
  $\hat{\theta} = \theta_t$
end

Algorithm 1: Levenburg-Marquardt algorithm

Function `GainRatio`(θ, δ, μ)**:**

`/* Compute gain ratio between current and proposed parameter values */`

$$\rho := \frac{S(\theta) - S(\theta + \delta)}{0.5 \times \delta^T \left(\mu\delta + J(\theta)^T \epsilon(\theta)\right)}$$

return ρ

end

Algorithm 2: Gain ratio function

B

Supplementary figures

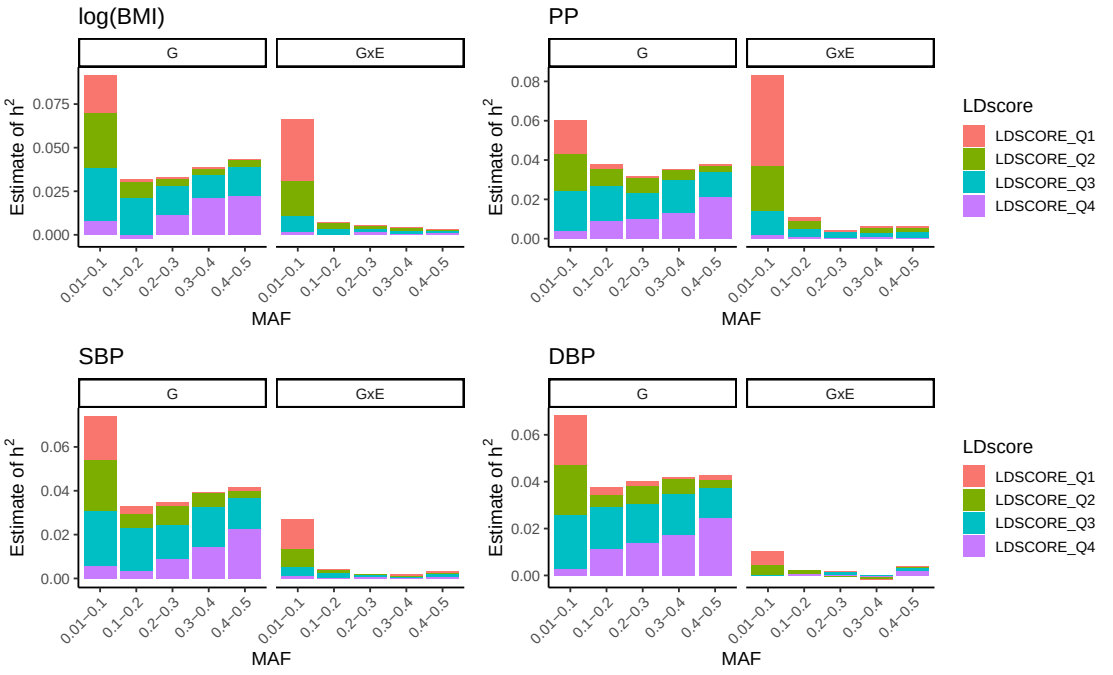

Figure B.1: Partitioned heritability estimates using genotyped SNPs for four quantitative traits in the UK Biobank. Heritability estimates partitoned into additive genetic and multiplicative GxE interaction effects for four quantitative traits in the UK Biobank using approximately 280,000 unrelated white British individuals (see Table C.3) and $M = 639,005$ genotyped SNPs. Multiplicative GxE iteractions were computed using the ELS from each model fit. Heritability estimation was performed using a multi-component implementation of randomised HE-regression[86, 87] with SNPs stratified into 20 components (5 MAF bins and 4 LD score quantiles).

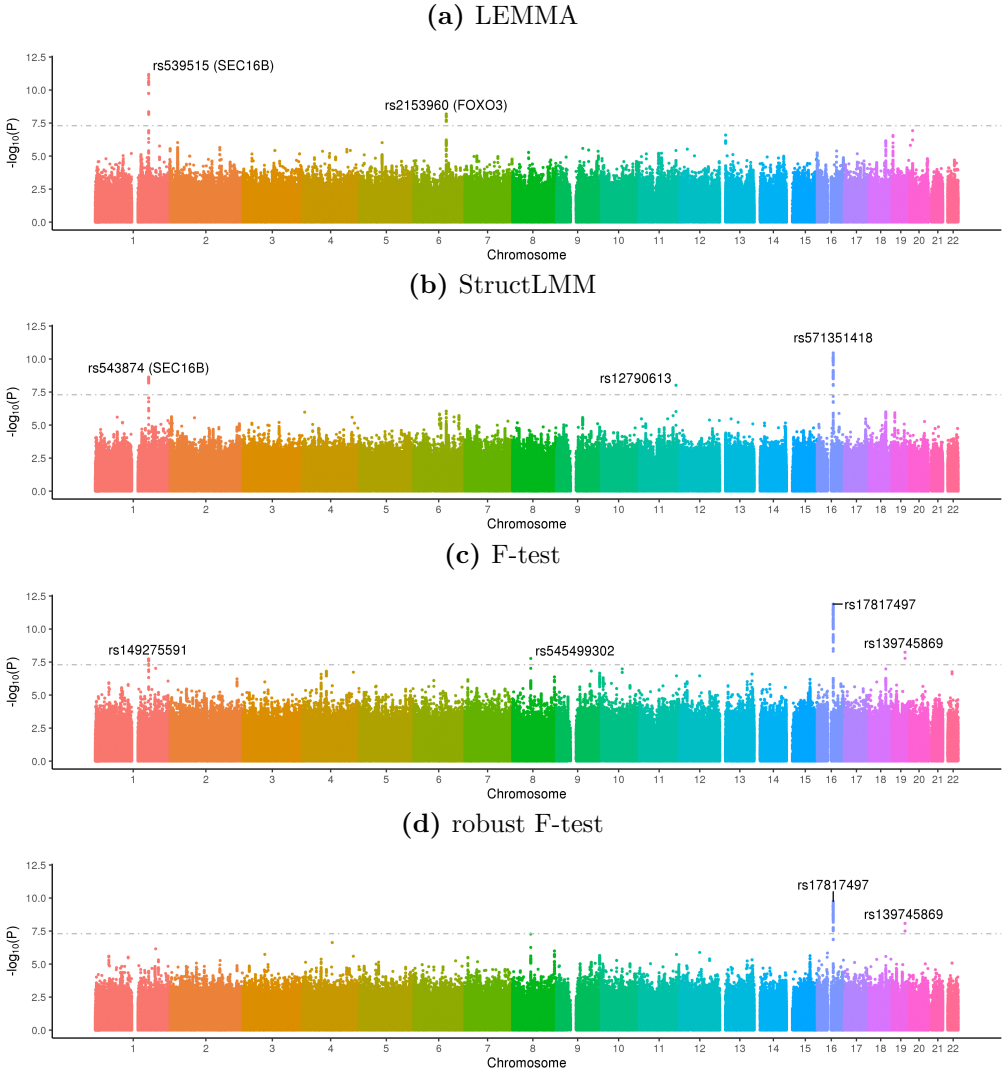

Figure B.2: GxE association statistics for log**BMI.** Manhattan plots displaying the negative $\log_10$ p values from GxE interaction tests at $10,295,038$ imputed SNPs applied to logBMI in the UK Biobank. GxE interaction tests were computed using (a) LEMMA, (b) StructLMM, (c) the F-test and (d) the robust F-test. The horizontal grey line denotes ($p = 5 \times 10^{-8}$).

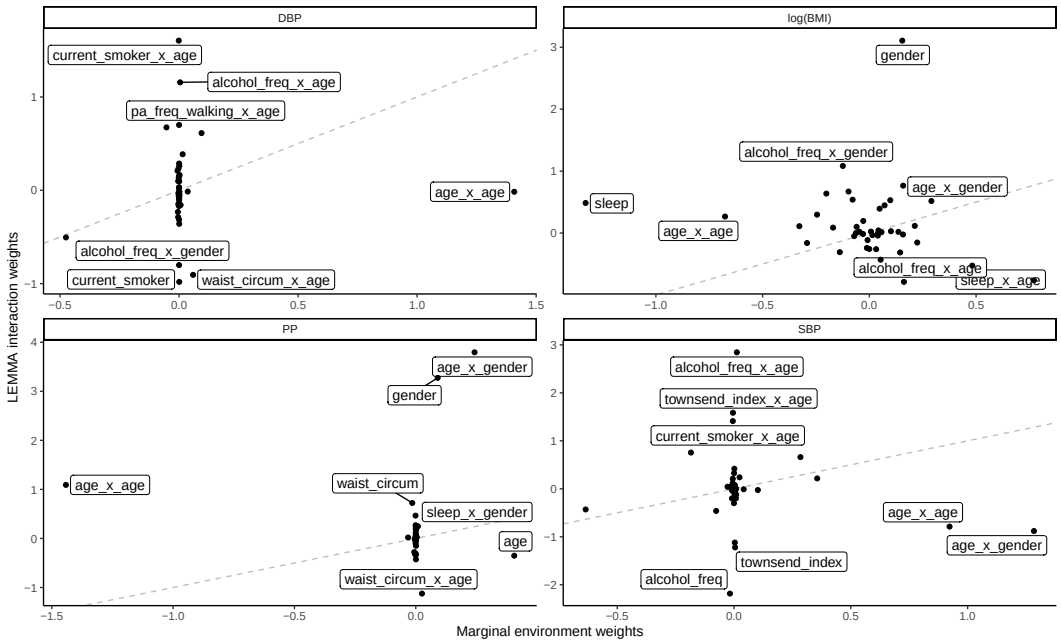

Figure B.3: Comparison of the LEMMA vs marginal environmental score. Interaction weights of the marginal environmental score were estimated from a multivariate linear regression, using all the non-genetic covariates used by LEMMA. Interactions weights were all rescaled so that the corresponding ES had variance one.

C
Supplementary tables

Table C.1: Comparison of partitioned heritability estimates using genotyped and imputed SNPs for four quantitative traits in the UK Biobank

Trait	Genotyped (SC)		Genotyped (LDMS)		CommonImputed (LDMS)	
	h^2_{G} (s.e)	h^2_{GxE} (s.e)	h^2_{G} (s.e)	h^2_{GxE} (s.e)	h^2_{G} (s.e)	h^2_{GxE} (s.e)
log BMI	0.259 (0.069)	0.071 (0.009)	0.237 (0.126)	0.086 (0.024)	0.274 (0.056)	0.093 (0.028)
PP	0.233 (0.039)	0.075 (0.018)	0.203 (0.084)	0.111 (0.021)	0.228 (0.051)	0.125 (0.028)
SBP	0.24 (0.053)	0.033 (0.003)	0.223 (0.095)	0.038 (0.017)	0.251 (0.05)	0.039 (0.023)
DBP	0.277 (0.034)	0.014 (0.001)	0.231 (0.079)	0.016 (0.017)	0.254 (0.05)	0.016 (0.02)

Comparison of the heritability estimates obtained using genotyped SNPs with RHE-SC, genotyped SNPs with RHE-LDMS, and common imputed SNPs (MAF$>$ 0.01 in the full UK Biobank cohort) with RHE-LDMS. GxE heritability estimates were were obtained using the ELS from each model fit. All analyses controlled for the same covariates used in the WGR analysis (including the top 20 principal components). Abbreviations; s.e, standard error estimated using the block jacknife (see Appendix A.2); h^2_{G}, heritability due to additive genetic effects; h^2_{GxE}, heritability due to multiplicative GxE effects; RHE, randomised HE-regression[86, 87]; SC, single SNP component; LDMS, SNPs stratified by minor allele frequency and LDscore (20 components).

Table C.2: Loci with genome-wide significant GxE interaction effects with the ELS

Trait	SNP	Chr	BP	A0	A1	AF	GxE Effect Size	GxE Effect SE	p-values		Nearest gene
									Main effect	GxE	
log(BMI)	rs539515	1	177,889,025	A	C	0.208	-0.01166	0.00170	1.6×10^{-60}	6.5×10^{-12}	*SEC16B*
log(BMI)	rs2153960	6	108,988,184	G	A	0.713	-0.00984	0.00170	1.6×10^{-8}	6.5×10^{-9}	*FOXO3*
DBP	rs8090962	18	56,694,404	A	G	0.443	0.00867	0.00157	4.5×10^{-1}	3.6×10^{-8}	*OACYLP / SEC11C*

Independent loci with genome-wide significant ($P < 5 \times 10^{-8}$) GxE interaction effects with the environmental latent score (ELS). Loci atleast 0.5cM apart were judged to be independent. SNP locations follow the GrCh37 human genome assembly. Abbrevations are as follows; Chr, chromosome; A0, reference allele; A1, alternative allele; AF, reference allele frequency; cM, centimorgan. All loci had IMPUTE INFO score > 0.99

Table C.3: Quality control and time to convergence of the WGR analyses

Trait	No. samples	No. SNPs	No. envs	No. envs-sq	No. covars total	Iterations for WGR converge	Time for WGR to converge*
log(BMI)	281149	642095	42	30	96	1631	78 hours 18 mins
PP	280749	642102	45	13	83	3821	183 hours 26 mins
SBP	280749	642102	45	15	85	967	46 hours 25 mins
DBP	280749	642102	45	15	85	646	31 hours 00 mins

Time for the whole genome regression analysis to converge is reported for each analysis of UK Biobank traits, as well as the number of SNPs and samples passing quality control and the number of covariates controlled for. 'Other covariates' consisted of the top 20 genetic principal components as reported by the UK Biobank, age^3, age$^2\times$ gender, age$^3\times$ gender, a binary indicator for the genotype chip and (for blood pressure traits only) BMI. The environmental variables used (including lower orders of age and gender) is described in (Section 4.2.2). To control for potential bias due to non-linear dependance between the phenotype and heritable environmental variables, I tested each environmental variable and included any significant squared effects as additional covariates (as described in Section 3.3). *based on the average per-iteration cost of 243 seconds, using 32 cores distributed across a cluster with Xeon E5-2667 v4 3.2Ghz processors.

Table C.4: Comparison of the number of genome-wide significant GxE associations and genomics control statistics from a GxE analysis of log**BMI in the UK Biobank**

Method	No. signals	Genomic control (χ^2)	Genomic control (p-values)*
LEMMA	2	1.062	1.038
LEMMA-S	5	1.275	1.164
StructLMM (-SQE)	3	NA	1.236
F-test (-SQE)	4	NA	1.372
robust F-test (-SQE)	2	NA	1.034
LEMMA (-SQE)	3	1.065	1.04
LEMMA-S (-SQE)	6	1.288	1.171

The number of independent loci (atleast 0.5cM apart) with genome-wide significant GxE interaction effects and genomic control statistics for seven different methods applied to logBMI in the UK Biobank. Genomic control is computed from GxE interaction tests statistics from $10,295,038$ imputed SNPs. Abbreviations; LEMMA-S, LEMMA with a homoskedastic test statistic (see **Online Methods**); (-SQE), significant squared environmental variables (Bonferroni correction) not included as additional covariates.
*The test statistics from StructLMM, F-test and the robust F-test are not χ^2_1 distributed. Hence for these methods we use $\lambda_{\text{GC}} = \log_{10}(m)/\log_1 0(0.5)$, where m is the median p-value, to denote the genomic control statistic as suggested by Moore *et al.*[55].

Table C.5: Summary statistics of the environment for participants in the top and bottom 5% of the environmental latent score for log**BMI**. The mean and standard deviation (in brackets) are shown for continuous variables, counts and proportions (in brackets) of each factor are shown for ordinal variables.

	Bottom 5%	Top 5%	Whole cohort
Mean hours of sleep			
Mean (SD)	7.18 (1.08)	7.41 (0.95)	7.28 (0.89)
Median	7	7	7
Variance of hours of sleep			
Mean (SD)	1.02 (0.65)	0.86 (0.62)	0.81 (0.56)
Median	0.7968124	0.7968124	0.7968124
TV watched (hours/day)			
Mean (SD)	3.28 (1.68)	3.03 (1.92)	2.69 (1.49)
Median	3	3	3
Current smoker			
Non-smoker	11,970 (85.15%)	10,006 (71.18%)	255,381 (90.83%)
Current-smoker	2,088 (14.85%)	4,052 (28.82%)	25,768 (9.17%)
Walking (days/week)			
Mean (SD)	4.32 (2.29)	5.81 (1.93)	5.38 (1.94)
Median	5	7	6
Moderate activity (days/week)			
Mean (SD)	2.56 (2.28)	4.11 (2.48)	3.59 (2.31)
Median	2	4	3
Vigorous activity (days/week)			
Mean (SD)	0.81 (1.36)	2.40 (2.28)	1.84 (1.92)
Median	0	2	1
Alcohol Intake (freq)			
Daily or almost daily	15 (0.11%)	6,801 (48.38%)	63,359 (22.54%)
Three or four times a week	135 (0.96%)	2,200 (15.65%)	71,457 (25.42%)
Once or twice a week	1,111 (7.90%)	1,750 (12.45%)	74,047 (26.34%)
One to three times a month	2,497 (17.76%)	868 (6.17%)	30,468 (10.84%)
Special occasions only	5,541 (39.42%)	1,258 (8.95%)	26,289 (9.35%)
Never	4,759 (33.85%)	1,181 (8.40%)	15,529 (5.52%)
Townsend index			
Mean (SD)	0.91 (3.31)	-0.83 (3.87)	-1.74 (2.82)
Median	0.818634	-2.1192	-2.47572
Age			
Mean (SD)	51.68 (7.40)	59.58 (8.05)	56.60 (8.02)
Median	51	62	58
Gender			
Female	14,058 (100.00%)	6,961 (49.52%)	149,093 (53.03%)
Male	0 (0.00%)	7,097 (50.48%)	132,056 (46.97%)

Table C.6: Summary statistics of the environment for participants in the top and bottom 5% of the environmental latent score for PP. The mean and standard deviation (in brackets) are shown for continuous variables, counts and proportions (in brackets) of each factor are shown for ordinal variables.

	Bottom 5%	Top 5%	Whole cohort
Mean hours of sleep			
Mean (SD)	7.96 (1.03)	7.36 (0.97)	7.28 (0.89)
Median	8	7	7
Variance of hours of sleep			
Mean (SD)	1.13 (0.77)	0.89 (0.60)	0.81 (0.56)
Median	0.7968124	0.7968124	0.7968124
TV watched (hours/day)			
Mean (SD)	3.53 (1.70)	3.57 (1.56)	2.69 (1.49)
Median	3	4	3
Current smoker			
Non-smoker	9,363 (66.70%)	13,849 (98.65%)	255,018 (90.83%)
Current-smoker	4,674 (33.30%)	189 (1.35%)	25,731 (9.17%)
Walking (days/week)			
Mean (SD)	5.71 (1.82)	5.59 (1.79)	5.38 (1.94)
Median	7	6	6
Moderate activity (days/week)			
Mean (SD)	3.82 (2.40)	4.02 (2.31)	3.59 (2.31)
Median	4	4	3
Vigorous activity (days/week)			
Mean (SD)	1.71 (2.04)	1.69 (1.92)	1.84 (1.92)
Median	1	1	1
Waist circum			
Mean (SD)	103.81 (11.78)	80.33 (9.28)	89.87 (13.20)
Median	103	80	89
Alcohol Intake (freq)			
Daily or almost daily	5,804 (41.35%)	1,473 (10.49%)	63,271 (22.54%)
Three or four times a week	3,399 (24.21%)	1,941 (13.83%)	71,378 (25.42%)
Once or twice a week	2,883 (20.54%)	3,229 (23.00%)	73,959 (26.34%)
One to three times a month	827 (5.89%)	1,976 (14.08%)	30,409 (10.83%)
Special occasions only	700 (4.99%)	3,082 (21.95%)	26,235 (9.34%)
Never	424 (3.02%)	2,337 (16.65%)	15,497 (5.52%)
Townsend index			
Mean (SD)	0.06 (3.35)	-2.26 (2.48)	-1.74 (2.82)
Median	-0.477834	-2.848015	-2.47572
Age			
Mean (SD)	65.26 (2.99)	67.12 (1.74)	56.59 (8.02)
Median	66	67	58
Gender			
Female	0 (0.00%)	14,038 (100.00%)	148,857 (53.02%)
Male	14,037 (100.00%)	0 (0.00%)	131,892 (46.98%)

Table C.7: Summary statistics of the environment for participants in the top and bottom 5% of the environmental latent score for SBP. The mean and standard deviation (in brackets) are shown for continuous variables, counts and proportions (in brackets) of each factor are shown for ordinal variables.

	Bottom 5%	Top 5%	Whole cohort
Mean hours of sleep			
Mean (SD)	7.41 (0.98)	7.37 (0.95)	7.28 (0.89)
Median	7	7	7
Variance of hours of sleep			
Mean (SD)	0.91 (0.61)	0.87 (0.61)	0.81 (0.56)
Median	0.7968124	0.7968124	0.7968124
TV watched (hours/day)			
Mean (SD)	2.94 (1.54)	3.20 (1.71)	2.69 (1.49)
Median	3	3	3
Current smoker			
Non-smoker	14,022 (99.89%)	5,737 (40.87%)	255,018 (90.83%)
Current-smoker	16 (0.11%)	8,301 (59.13%)	25,731 (9.17%)
Walking (days/week)			
Mean (SD)	5.43 (1.92)	5.69 (1.84)	5.38 (1.94)
Median	6	7	6
Moderate activity (days/week)			
Mean (SD)	3.50 (2.38)	4.17 (2.36)	3.59 (2.31)
Median	3	5	3
Vigorous activity (days/week)			
Mean (SD)	1.43 (1.84)	1.91 (2.13)	1.84 (1.92)
Median	1	1	1
Waist circum			
Mean (SD)	84.52 (12.03)	95.21 (12.25)	89.87 (13.20)
Median	83	95	89
Alcohol Intake (freq)			
Daily or almost daily	0 (0.00%)	8,868 (63.17%)	63,271 (22.54%)
Three or four times a week	0 (0.00%)	3,049 (21.72%)	71,378 (25.42%)
Once or twice a week	232 (1.65%)	1,643 (11.70%)	73,959 (26.34%)
One to three times a month	1,704 (12.14%)	318 (2.27%)	30,409 (10.83%)
Special occasions only	5,939 (42.31%)	129 (0.92%)	26,235 (9.34%)
Never	6,163 (43.90%)	31 (0.22%)	15,497 (5.52%)
Townsend index			
Mean (SD)	-2.90 (1.92)	1.20 (3.39)	-1.74 (2.82)
Median	-3.22619	1.26566	-2.47572
Age			
Mean (SD)	64.43 (3.49)	62.98 (4.14)	56.59 (8.02)
Median	65	63	58
Gender			
Female	12,841 (91.47%)	3,722 (26.51%)	148,857 (53.02%)
Male	1,197 (8.53%)	10,316 (73.49%)	131,892 (46.98%)

Table C.8: Summary statistics of the environment for participants in the top and bottom 5% of the environmental latent score for DBP. The mean and standard deviation (in brackets) are shown for continuous variables, counts and proportions (in brackets) of each factor are shown for ordinal variables.

	Bottom 5%	Top 5%	Whole cohort
Mean hours of sleep			
Mean (SD)	7.49 (1.15)	7.24 (0.89)	7.28 (0.89)
Median	7	7	7
Variance of hours of sleep			
Mean (SD)	1.09 (0.71)	0.81 (0.56)	0.81 (0.56)
Median	0.7968124	0.7968124	0.7968124
TV watched (hours/day)			
Mean (SD)	3.30 (1.69)	3.02 (1.64)	2.69 (1.49)
Median	3	3	3
Current smoker			
Non-smoker	14,013 (99.82%)	496 (3.53%)	255,018 (90.83%)
Current-smoker	25 (0.18%)	13,542 (96.47%)	25,731 (9.17%)
Walking (days/week)			
Mean (SD)	4.23 (2.32)	5.75 (1.76)	5.38 (1.94)
Median	5	7	6
Moderate activity (days/week)			
Mean (SD)	2.82 (2.32)	3.93 (2.40)	3.59 (2.31)
Median	2	4	3
Vigorous activity (days/week)			
Mean (SD)	1.37 (1.85)	1.76 (2.07)	1.84 (1.92)
Median	0	1	1
Waist circum			
Mean (SD)	105.08 (13.33)	88.64 (11.47)	89.87 (13.20)
Median	104	89	89
Alcohol Intake (freq)			
Daily or almost daily	155 (1.10%)	6,300 (44.88%)	63,271 (22.54%)
Three or four times a week	607 (4.32%)	3,006 (21.41%)	71,378 (25.42%)
Once or twice a week	2,819 (20.08%)	2,465 (17.56%)	73,959 (26.34%)
One to three times a month	2,657 (18.93%)	944 (6.72%)	30,409 (10.83%)
Special occasions only	3,838 (27.34%)	902 (6.43%)	26,235 (9.34%)
Never	3,962 (28.22%)	421 (3.00%)	15,497 (5.52%)
Townsend index			
Mean (SD)	-1.81 (2.81)	-0.41 (3.37)	-1.74 (2.82)
Median	-2.55106	-1.12539	-2.47572
Age			
Mean (SD)	61.73 (5.84)	59.47 (6.04)	56.59 (8.02)
Median	63	60	58
Gender			
Female	78 (0.56%)	6,211 (44.24%)	148,857 (53.02%)
Male	13,960 (99.44%)	7,827 (55.76%)	131,892 (46.98%)

References

[1] John W. Belmont et al. “A haplotype map of the human genome”. In: *Nature* 437.7063 (Oct. 2005), pp. 1299–1320.

[2] Simon Myers et al. “Genetics: A fine-scale map of recombination rates and hotspots across the human genome”. In: *Science* 310.5746 (2005), pp. 321–324.

[3] Itsik Pe’er et al. “Evaluating and improving power in whole-genome association studies using fixed marker sets”. In: *Nature Genetics* 38.6 (2006), pp. 663–667.

[4] Paul I.W. De Bakker et al. “Efficiency and power in genetic association studies”. In: *Nature Genetics* 37.11 (2005), pp. 1217–1223.

[5] Jonathan Marchini and Bryan Howie. “Genotype imputation for genome-wide association studies”. In: *Nature Reviews Genetics* 11.7 (June 2010), pp. 499–511.

[6] Elizabeth Pennisi. “Human genetic variation”. In: *American Association for the Advancement of Science* 3.1 (2007), pp. 9–11.

[7] Paul R. Burton et al. “Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls”. In: *Nature* 447.7145 (2007), pp. 661–678.

[8] John D Storey and Robert Tibshirani. “Statistical significance for genomewide studies”. In: *Proceedings of the National Academy of Sciences* 100.16 (Aug. 2003), pp. 9440–9445.

[9] Itsik Pe’er et al. “Estimation of the multiple testing burden for genomewide association studies of nearly all common variants”. In: *Genetic Epidemiology* 32.4 (May 2008), pp. 381–385.

[10] Richard M. Durbin et al. “A map of human genome variation from population-scale sequencing”. In: *Nature* 467.7319 (2010), pp. 1061–1073.

[11] David M. Altshuler et al. “An integrated map of genetic variation from 1,092 human genomes”. In: *Nature* 491.7422 (2012), pp. 56–65.

[12] Adam Auton et al. “A global reference for human genetic variation”. In: *Nature* 526.7571 (2015), pp. 68–74.

[13] Jonathan L Haines et al. “Complement Factor H Variant Increases the Risk of Age-Related Macular Degeneration”. In: *Science* 308.5720 (Apr. 2005), 419 LP –421.

[14] William Summerskill. “Paper of the year 2007”. In: *The Lancet* 371.9610 (Feb. 2008), pp. 370–371.

[15] William Astle and David J. Balding. “Population Structure and Cryptic Relatedness in Genetic Association Studies”. In: *Statistical Science* 24.4 (2010), pp. 451–471.

[16] Alkes L. Price et al. “Principal components analysis corrects for stratification in genome-wide association studies”. In: *Nature Genetics* 38.8 (2006), pp. 904–909.

[17] Alkes L. Price et al. “New approaches to population stratification in genome-wide association studies”. In: *Nature Reviews Genetics* 11.7 (2010), pp. 459–463.

[18] C R Henderson et al. “The Estimation of Environmental and Genetic Trends from Records Subject to Culling”. In: *Biometrics* 15.2 (1959), pp. 192–218.

[19] Peter M. Visscher et al. “Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings”. In: *PLoS Genetics* 2.3 (2006), pp. 0316–0325.

[20] Hilary K Finucane et al. “Partitioning heritability by functional annotation using genome-wide association summary statistics”. In: *Nature Genetics* 47.11 (2015), pp. 1228–1235.

[21] Doug Speed et al. “Reevaluation of SNP heritability in complex human traits”. In: *Nature Genetics* 49.7 (2017), pp. 986–992.

[22] Steven Gazal et al. “Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection”. In: *Nature Genetics* 49.10 (2017), pp. 1421–1427.

[23] Doug Speed and David J. Balding. “SumHer better estimates the SNP heritability of complex traits from summary statistics”. In: *Nature Genetics* 51.2 (2019), pp. 277–284.

[24] Jian Yang et al. “Common SNPs explain a large proportion of the heritability for human height”. In: *Nature Genetics* 42.7 (July 2010), pp. 565–569.

[25] Jian Yang et al. *Advantages and pitfalls in the application of mixed-model association methods*. Feb. 2014.

[26] Jennifer Listgarten et al. “Improved linear mixed models for genome-wide association studies”. In: *Nature Methods* 9.6 (2012), pp. 525–526.

[27] Peter M. Visscher, William G. Hill, and Naomi R. Wray. “Heritability in the genomics era - Concepts and misconceptions”. In: *Nature Reviews Genetics* 9.4 (2008), pp. 255–266.

[28] Evan A Boyle, Yang I Li, and Jonathan K Pritchard. “Leading Edge Perspective An Expanded View of Complex Traits: From Polygenic to Omnigenic”. In: *Cell* 169.7 (2017), pp. 1177–1186.

[29] Teri A. Manolio et al. *Finding the missing heritability of complex diseases*. Oct. 2009.

[30] Brendan Maher. “Personal genomes: The case of the missing heritability”. In: *Nature* 456.7218 (Nov. 2008), pp. 18–21.

[31] Evan E. Eichler et al. “Missing heritability and strategies for finding the underlying causes of complex disease”. In: *Nature Reviews Genetics* 11.6 (2010), pp. 446–450.

[32] Jonathan K Pritchard and Nancy J Cox. “The allelic architecture of human disease genes: common disease-common variant or not”. In: *Human Molecular Genetics* 11.20 (Oct. 2002), pp. 2417–2423.

[33] O. Zuk et al. "The mystery of missing heritability: Genetic interactions create phantom heritability". In: *Proceedings of the National Academy of Sciences* 109.4 (2012), pp. 1193–1198.

[34] Shaun Purcell. "Variance Components Models for Gene–Environment Interaction in Twin Analysis". In: *Twin Research* 5.6 (2002), pp. 554–571.

[35] Jacob Felson. "What can we learn from twin studies? A comprehensive evaluation of the equal environments assumption". In: *Social Science Research* 43 (2014), pp. 184–199.

[36] Doug Speed et al. "Improved heritability estimation from genome-wide SNPs". In: *American Journal of Human Genetics* (2012).

[37] David J. Hunter. "Gene-environment interactions in human diseases". In: *Nature Reviews Genetics* 6.4 (Apr. 2005), pp. 287–298.

[38] D Thomas. "Gene-environment-wide association studies: emerging approaches". In: *Nat Rev Genet* 11.4 (Mar. 2010), pp. 259–272.

[39] Amy Berrington de Gonzalez and D R Cox. "Interpretation of interaction: A review". en. In: *Ann. Appl. Stat.* 1.2 (2007), pp. 371–385.

[40] Hakhamanesh Mostafavi et al. "Variable prediction accuracy of polygenic scores within an ancestry group". In: *bioRxiv* (Jan. 2019), p. 629949.

[41] Duncan Thomas. "Methods for Investigating Gene-Environment Interactions in Candidate Pathway and Genome-Wide Association Studies". In: *Annual Review of Public Health* 31.1 (2010), pp. 21–36.

[42] NE Smith, PG and Day. "The design of case-control studies: The influence of confounding and interaction effects". In: *International Journal of Epidemiology* 13.3 (1984), pp. 356–365.

[43] Cassandra E Murcray et al. "Sample size requirements to detect gene-environment interactions in genome-wide association studies". eng. In: *Genetic epidemiology* 35.3 (Apr. 2011), pp. 201–210.

[44] Hugues Aschard. "A perspective on interaction effects in genetic association studies". In: *Genetic Epidemiology* 40.8 (2016), pp. 678–688.

[45] Hugues Aschard et al. "Challenges and opportunities in genome-wide environmental interaction (GWEI) studies". In: *Human Genetics* 131.10 (2012), pp. 1591–1613.

[46] Vincent Y. Yzerbyt, Dominique Muller, and Charles M. Judd. "Adjusting researchers' approach to adjustment: On the use of covariates when testing interactions". In: *Journal of Experimental Social Psychology* 40.3 (2004), pp. 424–431.

[47] Matthew C. Keller. *Gene × environment interaction studies have not properly controlled for potential confounders: The problem and the (simple) solution.* Jan. 2014.

[48] Jae Hoon Sul et al. "Accounting for Population Structure in Gene-by-Environment Interactions in Genome-Wide Association Studies Using Mixed Models". In: *PLOS Genetics* 12.3 (Mar. 2016). Ed. by Nicholas J. Schork, e1005849.

[49] Matthew R Robinson et al. “Genotype-covariate interaction effects and the heritability of adult body mass index”. In: *Nature Genetics* 49.8 (2017), pp. 1174–1181.

[50] S. H. Lee and J. H.J. Van Der Werf. “MTG2: An efficient algorithm for multivariate linear mixed model analysis based on genomic information”. In: *Bioinformatics* 32.9 (2016), pp. 1420–1422.

[51] Guiyan Ni et al. “Genotype-covariate correlation and interaction disentangled by a whole-genome multivariate reaction norm model”. In: *bioRxiv* (2018), p. 377796.

[52] Stephanie L. Stenzel et al. “The impact of exposure-biased sampling designs on detection of gene–environment interactions in case–control studies with potential exposure misclassification”. In: *European Journal of Epidemiology* 30.5 (2015), pp. 413–423.

[53] Peter Kraft and Hugues Aschard. “Finding the missing gene–environment interactions”. In: *European Journal of Epidemiology* 30.5 (2015), pp. 353–355.

[54] Steven C. Moore et al. “Common genetic variants and central adiposity among Asian-Indians”. In: *Obesity* 20.9 (2012), pp. 1902–1908.

[55] Rachel Moore et al. “A linear mixed model approach to study multivariate gene-environment interactions”. In: *Nat Genet* (2018), p. 270611.

[56] Jessica Tyrrell et al. “Gene–obesogenic environment interactions in the UK Biobank study”. In: *International Journal of Epidemiology* 46.2 (Jan. 2017), pp. 559–575.

[57] Alexander I Young, Fabian Wauthier, and Peter Donnelly. “Multiple novel gene-by-environment interactions modify the effect of FTO variants on body mass index”. In: *Nature Communications* 7 (2016), p. 12724.

[58] Tuomas O. Kilpeläinen et al. “Physical activity attenuates the influence of FTO variants on obesity risk: A meta-analysis of 218,166 adults and 19,268 children”. In: *PLoS Medicine* 8.11 (2011).

[59] Beate R. Ritz et al. “Lessons Learned from Past Gene-Environment Interaction Successes”. In: *American Journal of Epidemiology* 186.7 (2017), pp. 778–786.

[60] Montserrat García-Closas et al. “NAT2 slow acetylation, GSTM1 null genotype, and risk of bladder cancer: results from the Spanish Bladder Cancer Study and meta-analyses”. In: *The Lancet* 366.9486 (Aug. 2005), pp. 649–659.

[61] Chen Wu et al. “Genome-wide association analyses of esophageal squamous cell carcinoma in Chinese identify multiple susceptibility loci and gene-environment interactions”. In: *Nature Genetics* 44.10 (2012), pp. 1090–1097.

[62] Charles Kooperberg and Michael Leblanc. “Increasing the power of identifying gene x gene interactions in genome-wide association studies”. eng. In: *Genetic epidemiology* 32.3 (Apr. 2008), pp. 255–263.

[63] James Y Dai et al. “Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction”. In: *Biometrika* 99.4 (Sept. 2012), pp. 929–944.

[64] Pingye Zhang et al. “Detecting Gene-Environment Interactions for a Quantitative Trait in a Genome-Wide Association Study”. eng. In: *Genetic epidemiology* 40.5 (July 2016), pp. 394–403.

[65] David M. Evans et al. “Two-stage two-locus models in genome-wide association”. In: *PLoS Genetics* 2.9 (2006), pp. 1424–1432.

[66] Guillaume Paré et al. “On the use of variance per genotype as a tool to identify quantitative trait interaction effects: A report from the women’s genome health study”. In: *PLoS Genetics* 6.6 (2010), pp. 1–10.

[67] Maksim V. Struchalin et al. “Variance heterogeneity analysis for detection of potentially interacting genetic loci: Method and its limitations”. In: *BMC Genetics* 11 (2010).

[68] Alexander I Young, Fabian L Wauthier, and Peter Donnelly. “Identifying loci affecting trait variability and detecting interactions in genome-wide association studies”. In: *Nature Genetics (in press)* 50.November (2018).

[69] Huanwei Wang et al. “Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank”. In: *bioRxiv* (2019), pp. 1–24.

[70] Yongtao Guan and Matthew Stephens. “Bayesian variable selection regression for genome-wide association studies and other large-scale problems”. In: *Annals of Applied Statistics* 5.3 (2011), pp. 1780–1815.

[71] Benjamin A. Logsdon, Gabriel E. Hoffman, and Jason G. Mezey. “A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis”. In: *BMC Bioinformatics* 11 (2010).

[72] Benjamin A. Logsdon et al. “A novel variational Bayes multiple locus Z-statistic for genome-wide association studies with Bayesian model averaging”. In: *Bioinformatics* 28.13 (2012), pp. 1738–1744.

[73] Peter Carbonetto and Matthew Stephens. “Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies”. In: *Bayesian Analysis* 7.1 (2012), pp. 73–108.

[74] Xiang Zhou et al. “Polygenic Modeling with Bayesian Sparse Linear Mixed Models”. In: *PLoS Genetics* 9.2 (Feb. 2013). Ed. by Peter M. Visscher, e1003264.

[75] Po Ru Loh et al. “Efficient Bayesian mixed-model analysis increases association power in large cohorts”. In: *Nature Genetics* 47.3 (2015), pp. 284–290.

[76] BJ Hayes, ME Goddard, et al. “Prediction of total genetic value using genome-wide dense marker maps”. In: *Genetics* 157.4 (2001), pp. 1819–1829.

[77] Gustavo de los Campos et al. “Whole-genome regression and prediction methods applied to plant and animal breeding”. In: *Genetics* 193.2 (2013), pp. 327–345.

[78] M Erbe et al. “Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels”. In: *Journal of dairy science* 95.7 (2012), pp. 4114–4129.

[79] Po Ru Loh et al. “Mixed-model association for biobank-scale datasets”. In: *Nature Genetics* 50.7 (2018), pp. 906–908.

[80] "Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits". In: *Nature Genetics* 50.5 (2018), pp. 737–745.

[81] Jian Yang et al. "Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index". In: *Nature Genetics* 47.10 (2015), pp. 1114–1120.

[82] Joseph E. Powell et al. "Signatures of negative selection in the genetic architecture of human complex traits". In: *Nature Genetics* 50.5 (2018), pp. 746–753.

[83] Brendan K Bulik-Sullivan et al. "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies". In: *Nature Genetics* 47.3 (Feb. 2015), pp. 291–295.

[84] David Golan, Eric S. Lander, and Saharon Rosset. "Measuring missing heritability: Inferring the contribution of common variants". In: *Proceedings of the National Academy of Sciences* 111.49 (2014), E5272–E5281.

[85] Jian Yang et al. "Concepts, estimation and interpretation of SNP-based heritability". In: *Nature Genetics* 49.9 (2017), pp. 1304–1310.

[86] Yue Wu and Sriram Sankararaman. "A scalable estimator of SNP heritability for biobank-scale data". In: *Bioinformatics* 34.13 (2018), pp. i187–i194.

[87] Ali Pazokitoroudi et al. "Scalable multi-component linear mixed models with application to SNP heritability estimation". In: *bioRxiv* (2019), p. 522003.

[88] M F Hutchinson. "A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines". In: *Communications in Statistics - Simulation and Computation* 19.2 (Jan. 1990), pp. 433–450.

[89] Martin J Wainwright, Michael I Jordan, et al. "Graphical models, exponential families, and variational inference". In: *Foundations and Trends® in Machine Learning* 1.1–2 (2008), pp. 1–305.

[90] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877.

[91] Christopher M Bishop. *Pattern Recognition and Machine Learning*. 1st ed. Springer-Verlag New York, 2006, p. 738.

[92] J R Hershey and P A Olsen. "Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models". In: *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*. Vol. 4. 2007, pp. IV–317–IV–320.

[93] Chong Wang and David M. Blei. "Variational inference in nonconjugate models". In: *Journal of Machine Learning Research* 14.1 (2013), pp. 1005–1031. arXiv: 1209.4360.

[94] Kevin P Murphy and Francis Bach. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012, p. 1096.

[95] Matt Hoffman et al. "Stochastic Variational Inference". In: 14 (June 2012), p. 390984.

[96] Ravi Varadhan and Christophe Roland. "Simple and globally convergent methods for accelerating the convergence of any em algorithm". In: *Scandinavian Journal of Statistics* 35.2 (2008), pp. 335–353.

[97] Rohan L. Fernando, Jack C.M. Dekkers, and Dorian J. Garrick. "A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses". In: *Genetics Selection Evolution* 46.1 (2014), pp. 1–13.

[98] Andrew Dahl et al. "A multiple-phenotype imputation method for genetic studies". In: *Nature Genetics* 48.4 (2016), pp. 466–472.

[99] Gavin Band and Jonathan Marchini. "BGEN: a binary file format for imputed genotype and haplotype data". In: *BioRxiv* (2018), p. 308296.

[100] Christoph Lippert et al. "FaST linear mixed models for genome-wide association studies". In: *Nature Methods* 8.10 (2011), pp. 833–835.

[101] Lynn M. Almli et al. "Correcting systematic inflation in genetic association tests that consider interaction effects application to a genome-wide association study of posttraumatic stress disorder". In: *JAMA Psychiatry* 71.12 (2014), pp. 1392–1399.

[102] Rand R Wilcox. *Introduction to robust estimation and hypothesis testing*. Academic press, 2011.

[103] Peter J Huber et al. "The behavior of maximum likelihood estimates under nonstandard conditions". In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 1. University of California Press. 1967, pp. 221–233.

[104] Halbert White et al. "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity". In: *econometrica* 48.4 (1980), pp. 817–838.

[105] William H Greene. *Econometric Analysis 5th edition*. Pearson Education India, 2003.

[106] Eric J Tchetgen Tchetgen and Peter Kraft. "On the robustness of tests of genetic associations incorporating gene-environment interaction when the environmental exposure is misspecified". In: *Epidemiology* 22.2 (2011), pp. 257–261.

[107] Arend Voorman et al. "Behavior of QQ-plots and Genomic Control in studies of gene-environment interaction". In: *PLoS ONE* 6.5 (2011).

[108] J Scott Long and Laurie H Ervin. "Using heteroscedasticity consistent standard errors in the linear regression model". In: *The American Statistician* 54.3 (2000), pp. 217–224.

[109] Michael C. Wu et al. "Rare-variant association testing for sequencing data with the sequence kernel association test". In: *American Journal of Human Genetics* 89.1 (2011), pp. 82–93.

[110] Clare Bycroft et al. "The UK Biobank resource with deep phenotyping and genomic data". In: *Nature* 562.7726 (2018), pp. 203–209.

[111] Thomas J Hoffmann et al. "Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation". In: *Nature Genetics* 49.1 (Nov. 2016), pp. 54–64.

[112] Helen R Warren et al. “Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk”. In: *Nature Genetics* 49.3 (Jan. 2017), pp. 403–415.

[113] Evangelos Evangelou et al. “Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits”. In: *Nature Genetics* 50.10 (2018), pp. 1412–1425.

[114] Chen Wu et al. “Genome-wide association study identifies three new susceptibility loci for esophageal squamous-cell carcinoma in Chinese populations”. In: *Nature Genetics* 43.7 (2011), pp. 679–684.

[115] Philip J. Brooks et al. “The alcohol flushing response: An unrecognized risk factor for esophageal cancer from alcohol consumption”. In: *PLoS Medicine* 6.3 (2009), pp. 0258–0263.

[116] Clare Bycroft et al. “Genome-wide genetic data on ˜ 500 , 000 UK Biobank participants”. In: *bioRxiv* (2017).

[117] Shafqat Ahmad et al. “Gene × Physical Activity Interactions in Obesity: Combined Analysis of 111,421 Individuals of European Ancestry”. In: *PLoS Genetics* 9.7 (2013), pp. 1–9.

[118] Christiaan A. de Leeuw et al. “Conditional and interaction gene-set analysis reveals novel functional pathways for blood pressure”. In: *Nature Communications* 9.1 (2018).

[119] Thomas Burgoine et al. “Examining the interaction of fast-food outlet exposure and income on diet and obesity: Evidence from 51,361 UK Biobank participants”. In: *International Journal of Behavioral Nutrition and Physical Activity* 15.1 (2018), pp. 1–12.

[120] UK Biobank Coordinating Centre. “UK Biobank: Protocol for a large-scale prospective epidemiological resource UK Biobank Coordinating Centre Stockport”. In: *Design* 06.March (2007), pp. 1–112.

[121] Jian Yang et al. “GCTA: a tool for genome-wide complex trait analysis”. In: *The American Journal of Human Genetics* 88.1 (2011), pp. 76–82.

[122] Martin D. Tobin et al. “Adjusting for treatment effects in studies of quantitative traits: Antihypertensive therapy and systolic blood pressure”. In: *Statistics in Medicine* 24.19 (Oct. 2005), pp. 2911–2935.

[123] Kangcheng Hou et al. “Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture”. In: *Nature Genetics* 51.8 (2019), pp. 1244–1251.

[124] P. M. Visscher et al. “Evidence-based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses”. In: *Molecular Psychiatry* 17.5 (2012), pp. 474–485.

[125] Mathias Rask-Andersen et al. “Gene-environment interaction study for BMI reveals interactions between genetic factors and physical activity, alcohol consumption and socioeconomic status”. In: *PLOS Genetics* 13.9 (Sept. 2017). Ed. by Ulrike Peters, e1006977.

[126] Qibin Qi et al. “Television watching, leisure time physical activity, and the genetic predisposition in relation to body mass index in women and men”. In: *Circulation* 126.15 (2012), pp. 1821–1827.

[127] Qibin Qi et al. “Fried food consumption, genetic risk, and body mass index: Gene-diet interaction analysis in three US cohort studies”. In: *BMJ (Online)* 348.March (2014), pp. 1–12.

[128] Qibin Qi et al. “Sugar-Sweetened Beverages and Genetic Risk of Obesity From the Departments of Nutrition (Q”. In: *NEJM.org. N Engl J Med* 15.11 (2012), pp. 1387–96.

[129] Robert C. Kaplan et al. “A genome-wide association study identifies novel loci associated with circulating IGF-I and IGFBP-3”. In: *Human Molecular Genetics* 20.6 (2011), pp. 1241–1251.

[130] A. S. Richardson et al. “Moderate to vigorous physical activity interactions with genetic variants and body mass index in a large US ethnically diverse cohort”. In: *Pediatric Obesity* 9.2 (2014), e35–e46.

[131] Huanwei Wang et al. “Genotype-by-environment interactions inferred from genetic effects on phenotypic variability in the UK Biobank”. In: *Science Advances* 5.8 (Aug. 2019), eaaw3538.

[132] Dibyendu Bhattacharyya and Benjamin S. Glick. “Two Mammalian Sec16 Homologues Have Nonredundant Functions in Endoplasmic Reticulum (ER) Export and Transitional ER Organization”. In: *Molecular biology of the cell* 18.December (2007), pp. 986–994.

[133] Gudmar Thorleifsson et al. “Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity”. In: 41.1 (2009), pp. 18–24.

[134] Kikuko Hotta et al. “Association between obesity and polymorphisms in SEC16B, TMEM18, GNPDA2, BDNF, FAIM2 and MC4R in a Japanese population”. In: *Journal of Human Genetics* 54.12 (2009), pp. 727–731.

[135] Peter M Schmid et al. “Expression of fourteen novel obesity-related genes in zucker diabetic fatty rats”. In: *Cardiovascular Diabetology* 11 (2012), pp. 1–11.

[136] Dolores Corella et al. “Statistical and Biological Gene-Lifestyle Interactions of MC4R and FTO with Diet and Physical Activity on Obesity: New Effects on Alcohol Consumption”. In: *PLoS ONE* 7.12 (2012).

[137] Yin Tat Lee and Aaron Sidford. “Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems”. In: *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS* (2013), pp. 147–156.

[138] Yurii Nesterov and Sebastian U. Stich. “Efficiency of the Accelerated Coordinate Descent Method on Structured Optimization Problems”. In: *SIAM Journal on Optimization* 27.1 (2017), pp. 110–123.

[139] Christopher C Chang et al. “Second-generation PLINK: rising to the challenge of larger and richer datasets”. In: *GigaScience* 4.1 (Dec. 2015), p. 7.

[140] J A Nelder and R Mead. “A Simplex Method for Function Minimization”. In: *The Computer Journal* 7.4 (Jan. 1965), pp. 308–313.

[141] AA Zolfaghari et al. "An algorithm for the least-squares estimation of nonlinear parameters." In: *International Journal of Soil Science* 3.2 (2005), pp. 270–277.

[142] Longda Jiang et al. "A resource-efficient tool for mixed model association analysis of large-scale data". In: *bioRxiv* (2019), p. 598110.

[143] Wei Zhou et al. "Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies". In: *Nature Genetics* 50.9 (2018), pp. 1335–1341.

[144] Kaj Madsen, Hans Bruun Nielsen, and Ole Tingleff. "Methods for non-linear least squares problems". In: (1999).