This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.
Genomic prediction models, selection tools and association studies for genotype by environment data

Daniel J. Tolhurst

The Roslin Institute and Royal (Dick) School of Veterinary Studies
College of Medicine and Veterinary Medicine
The University of Edinburgh

This thesis is submitted for the degree of
Doctor of Philosophy

November 2023
Declaration

I declare that this thesis and the data presented in it are original and my own work, unless otherwise specified. This work has not been submitted for any other degree or professional qualification. I have read and understood The University of Edinburgh guidelines on plagiarism and declare that this written dissertation is all my own work except where I indicate otherwise by proper use of quotes and references. This thesis is an account of work conducted by me whilst studying for the degree of Doctor of Philosophy at The University of Edinburgh.

Daniel J. Tolhurst
November 2023
Abstract

Plant breeding is complicated by the fact that genotypes respond differently to different environments, a phenomenon known as genotype by environment interaction (GEI). Despite its importance, however, many plant breeding programmes still use inefficient methods for handling GEI. This thesis develops a wide-array of methods that leverage GEI for efficient prediction, selection and discovery in plant breeding. The methods are demonstrated using a collaborating cotton breeding dataset from Bayer CropScience as well as publicly available and simulated datasets.

Chapter 1 presents a brief overview of plant breeding design and analysis, with a focus on genomic prediction models, selection tools and association studies for genotype by environment data.

Chapter 2 develops genomic prediction models that predict the response of different genotypes across different growing environments. The models are referred to as integrated factor analytic (IFA) models. The IFA models integrate known genotypic covariates derived from marker data and known environmental covariates derived from weather and soil data along with latent environmental covariates estimated directly from the phenotypic data. These models have great potential to improve predictive plant breeding in the presence of GEI.

Chapter 3 develops selection tools that provide breeders with information to select and deploy well-adapted genotypes to their target environments. The tools provide measures of overall performance and stability, which summarise average genotype performance across environments and the variability in performance. A new directional stability measure is also introduced that partitions genotype stability into components that reflect favourable and unfavourable adaptation. These tools are becoming increasingly important with the presence of rapidly changing environments amidst climate change.

Chapters 4 and 5 develop fast exact methods for conducting genome-wide association studies (GWAS). The methods produce all required test statistics from the fit of a single linear mixed model, instead of a very large number of models for all markers of interest. Fast methods are also introduced for GWAS using complex models for GEI. These methods have great potential to improve discovery in a wide-array of genetic studies, particularly with the advent of large-scale datasets and complex genotype by environment interactions.
Chapter 6 develops a general framework for simulating GEI using the class of multiplicative models. The framework can be used to simulate realistic multi-environment trial (MET) datasets and model breeding programmes that better reflect the complexity of real-world settings. This framework provides a general basis for plant breeders and researchers to evaluate different breeding methods in the presence of GEI.

Chapter 7 presents a discussion and concluding remarks, with a focus on placing the thesis in the wider agricultural community.
The world’s population is expected to increase to over 9 billion people by 2050. This poses a serious challenge to global agriculture, which must produce significantly more food while becoming more sustainable and more efficient amidst climate change. Plant breeding is at the heart of these efforts. It is responsible for developing new and improved individuals (genotypes) that are well-adapted to the different environmental conditions and biotic stresses experienced on farms all over the globe. Plant breeding continues to become more efficient with the advent of new technologies and improved breeding practices. However, the global demand for plant-based food is expected to increase by a further 70% by 2050, which poses an unprecedented challenge not only for global plant breeding programmes, but also for researchers, scientists and farmers working in the plant breeding industry.

Plant breeding involves multiple stages of testing, selection and advancement of candidate genotypes for eventual release to farmers. Many of these stages are complicated by the fact that the genotypes respond differently to different growing environments, a phenomenon known as genotype by environment interaction (GEI). This makes breeding for a single genotype that is well-adapted to many different environments very difficult. Historically, plant breeding programmes have handled GEI in one of three ways:

1. Ignore GEI by selecting genotypes with the highest average performance over all potential growing environments.
2. Reduce GEI by grouping similar environments together and then selecting high performing genotypes within each group of environments.
3. Leverage GEI by selecting high performing and stable genotypes for specific growing environments.

This thesis develops a wide-array of methods that leverage GEI in plant breeding. The methods are demonstrated using a collaborating cotton breeding dataset from Bayer CropScience as well as publicly available and simulated datasets. Chapter 1 presents a brief overview of plant breeding design and analysis, with a focus on genomic prediction models, selection tools and association studies for genotype by environment data. Chapter 2 develops genomic
prediction models that predict the response of different plant genotypes across different growing environments. These models can be used to predict the performance of a genotype prior to planting in a specific environment. Chapter 3 develops selection tools that provide measures of overall performance, responsiveness and stability for each genotype. These tools can be used to select and release well-adapted genotypes to specific growing environments. Chapters 4 and 5 develop fast exact methods for conducting genome-wide association studies (GWAS) in the presence of GEI. These studies can be used to identify genome regions responsible for producing genotypes with desirable adaptation and disease resistance across different environments. Chapter 6 develops a general framework for simulating genotype by environment data. This framework can be used to compare different plant breeding strategies and model plant breeding programmes that better reflect the complexity of real-world settings. Chapter 7 presents a discussion and concluding remarks, with a focus on placing the thesis in the wider agricultural community.

The methods and concepts developed in this thesis have great potential to improve the efficiency of plant breeding programmes, especially in the presence of complex genotype by environment interactions.
Acknowledgements

This PhD would not have been possible without the guidance and support of several people. I particularly thank my family and friends, both here in Edinburgh and back in Australia.

In their roles as supervisors, I sincerely thank Prof. Gregor Gorjanc and Dr. Chris Gaynor. Gregor, thank you for trusting in my science and allowing me to grow into an independent researcher. I have enjoyed our stimulating discussions and appreciate your guidance on more than just the science. Chris, thank you for sharing your extensive knowledge on plant breeding and your honest feedback. I look forward to continued collaboration with you both.

I greatly acknowledge Bayer CropScience for funding my PhD and Dr John Hickey for organising the collaboration. Thank you Dr. Brian Gardunia for taking on the role of unofficial supervisor and providing me with valuable insights into the technical side of plant breeding. Thanks also to the wider cotton breeding and pipeline design teams for their enthusiasm and willingness to test the methods developed in this thesis.

I thank the Highlander Lab for their ongoing encouragement, in particular Dr. Jon Bančič. Thank you for your fruitful collaborations and your honesty when my ideas have been far too abstract to be useful.

Lastly, I thank my partner Alexandra Sinclair. Thank you for coming on this journey with me and for supporting me like no other. This PhD is as much yours as it is mine.
Publications

Published manuscripts included in this thesis:


Additional manuscripts to be submitted from this thesis:


- **Tolhurst, DJ** and Gorjanc, G. (2024) Fast exact methods for genome-wide association studies (in prep)

- **Tolhurst, DJ**, Gaynor, RC, Gardunia, B and Gorjanc, G. (2024) Genome-wide association studies built on multiplicative models for genotype by environment interaction (in prep)

Additional manuscripts during candidature:


∗Joint first author
# Table of contents

## 1 Introduction

1.1 Background ............................................. 1
1.2 Genomic prediction models .............................. 3
1.3 Selection tools ........................................ 6
1.4 Association studies .................................... 8
1.5 Simulation studies ..................................... 9
1.6 Objectives ........................................... 10

## 2 Genomic prediction models

2.1 Prelude ................................................ 11
2.2 Concluding remarks ................................... 35

## 3 Selection tools

3.1 Prelude ................................................ 37
3.2 Linear mixed models .................................. 38
3.3 Models for the genotype by environment effects ........ 39
3.4 Selection tools ....................................... 41
   3.4.1 Overall performance ............................. 41
   3.4.2 A new rotation for disentangling non-crossover and crossover GEI 45
   3.4.3 Responsiveness ................................ 53
   3.4.4 Stability ...................................... 54
3.5 Measures of variance explained ........................ 57
3.6 Multiple target populations of environments ......... 58
3.7 Extension to the integrated factor analytic model .... 59
3.8 Application to example dataset ....................... 61
   3.8.1 Data description ................................ 61
   3.8.2 Model estimation ............................... 62
   3.8.3 Results ...................................... 62
Table of contents

3.9 Concluding remarks .................................................. 69

4 Association studies I ................................................. 71
  4.1 Prelude ................................................................. 71
  4.2 Introduction ......................................................... 72
  4.3 Linear mixed models .............................................. 73
    4.3.1 Baseline linear mixed model ............................... 74
    4.3.2 Marker-in linear mixed model ......................... 75
    4.3.3 Marker-out linear mixed model ..................... 76
  4.4 Fundamental expressions .......................................... 78
  4.5 Test statistics ................................................... 79
    4.5.1 A simple fix for current software ...................... 81
  4.6 Genotype to marker ............................................... 81
  4.7 Model estimation ................................................. 82
  4.8 Application to example dataset ............................... 84
    4.8.1 Data description ........................................... 84
    4.8.2 Models fitted ............................................... 84
    4.8.3 Results ...................................................... 86
    4.8.4 Discussion .................................................. 90
  4.9 Testing a set of markers at a time ........................... 92
    4.9.1 Baseline linear mixed model ............................... 92
    4.9.2 Set-in linear mixed model ................................. 93
    4.9.3 Set-out linear mixed model ............................... 95
    4.9.4 Test statistics .............................................. 96
  4.10 Testing a single marker while excluding a set ............... 97
    4.10.1 Chromosome-out linear mixed model ................. 98
    4.10.2 Baseline linear mixed model ............................. 99
  4.11 Modelling population structure ................................ 101
  4.12 Concluding remarks ............................................. 102

5 Association studies II .............................................. 103
  5.1 Prelude ............................................................. 103
  5.2 Linear mixed models ............................................ 104
  5.3 Preliminary baseline linear mixed model ...................... 104
  5.4 Compound symmetry model ....................................... 105
    5.4.1 Baseline linear mixed model ............................... 106
    5.4.2 Marker-out linear mixed model ....................... 107
5.5 Random regression model .......................................... 113  
5.5.1 Baseline linear mixed model ................................. 114  
5.5.2 Marker-out linear mixed model ......................... 115  
5.6 Factor analytic model ........................................... 122  
5.6.1 Baseline linear mixed model ................................. 123  
5.6.2 Rotating the loadings and scores ....................... 123  
5.6.3 Selection tools .................................................. 126  
5.7 Concluding remarks ............................................. 129  

6 Simulation studies .................................................. 131  
6.1 Prelude .............................................................. 131  
6.2 Introduction ....................................................... 132  
6.3 Simulation Framework .......................................... 133  
6.3.1 Simulating genotype by environment effects .............. 133  
6.3.2 Simulating main effects and interaction effects ........... 135  
6.3.3 Simulating a between-environment genetic variance matrix . 136  
6.3.4 Measures of variance explained .......................... 139  
6.3.5 Simulating phenotypes ....................................... 140  
6.3.6 Simulating one or more target population of environments . 141  
6.3.7 Measures of accuracy ....................................... 142  
6.4 Application .......................................................... 143  
6.4.1 MET dataset simulation ...................................... 143  
6.4.2 Breeding programme simulation ......................... 145  
6.5 Concluding remarks ............................................. 150  

7 General discussion ................................................. 151  
7.1 Genomic prediction models ..................................... 152  
7.2 Selection tools .................................................... 153  
7.3 Association studies ............................................... 154  
7.4 Simulation studies ............................................... 156  
7.5 Concluding remarks ............................................. 157  

References .......................................................... 159  

Appendix A Sparse formulation ...................................... 169  
A.1 Preliminaries ....................................................... 169  
A.2 Prediction of genotype scores and regression residuals ............ 171
A.3 Estimation of loadings and specific variances .......................... 172
A.4 Trimming the regression residuals ........................................ 174

Appendix B  Proof of invariance ................................................. 175
  B.1 Preliminaries ............................................................. 175
  B.2 Scale invariance .......................................................... 176
  B.3 Translational invariance .................................................. 177
  B.4 Rotational invariance ..................................................... 178

Appendix C  General extensions ................................................. 179
  C.1 Rotating the covariates and scores ..................................... 179
  C.2 Selection tools ............................................................ 183

Appendix D  Hypothetical examples ............................................ 185

Appendix E  Matrix results ...................................................... 195
Chapter 1

Introduction

1.1 Background

Plant breeding is complicated by the fact that genotypes respond differently to different environments, a phenomenon known as genotype by environment interaction (GEI). GEI can lead to significant increases in experimental error and reductions in genetic gain (Kang, 1990). Despite its importance, however, many plant breeding programmes still use inefficient methods for handling GEI. This thesis develops methods that leverage GEI for efficient prediction, selection and discovery in plant breeding. The methods are novel extensions built on the past 100+ years of experimental design and statistical analysis of genotype by environment data. The methods are demonstrated using a collaborating cotton breeding dataset from Bayer CropScience as well as publicly available and simulated datasets.

Plant breeding involves multiple stages of evaluation, selection and progression from the initial cross between parents to the commercial release as a variety or hybrid to growers. The early stages of breeding screen genotypes for agronomically important traits such as disease resistance and desired morphology (Bernardo, 2014). The later stages then evaluate genotypes for yield and other commercially important traits in formally designed comparative field experiments, known as field trials (Cochran and Cox, 1957). The field trials are conducted across a range of growing environments that reflect the intended region of commercial production (Comstock and Moll, 1963). Field evaluation is conducted in an incremental manner, with the last stage involving the most number of environments compared to the earlier stages but also the fewest number of genotypes (Allard, 1999). Figure 1.1 summarises the relative number of genotypes and environments in a hypothetical plant breeding programme with four stages of field evaluation. This figure highlights the incremental nature of plant breeding and the general structure of field evaluation that has been historically shaped by the presence of GEI.
Plant breeders select and progress superior genotypes during field evaluation by accumulating multi-environment trial (MET) datasets. A MET dataset is constructed with a sample of genotypes relevant to the current selection decisions and a sample of environments that generally span multiple years and locations (Smith et al., 2021a). An important consideration when constructing a MET dataset is the extent to which it represents the breeder’s target population of environments (TPE), i.e. the set of environments intended for commercial production (Comstock, 1977; Cooper et al., 1993, 2014). This alignment has ramifications on the realised rates of genetic gain observed by growers (Cooper et al., 2023).

Plant breeders also use MET datasets to gauge the extent and form of GEI observed in their TPE. Genotype by environment interaction can be broadly categorised as either non-crossover or crossover interaction, which reflect changes in the scale of genotype response between environments or changes in genotype rank (Allard and Bradshaw, 1964; Baker, 1988; Gail and Simon, 1985). Figure 1.2 demonstrates non-crossover and crossover GEI in terms of two hypothetical genotypes and two hypothetical environments. The scale of genotype response is higher in E2 compared to E1, while the rank of genotypes in E1 is different to E2. These are classical representations of non-crossover and crossover GEI, respectively. Often, crossover GEI is of particular importance to breeders because their selection decisions are more complicated by changes in genotype rank compared to changes in scale (Baker, 1990; Eisemann et al., 1990). The presence of crossover GEI has shaped the way plant breeding programmes operate.
1.2 Genomic prediction models

Genomic prediction is a form of marker-assisted prediction that can improve the genetic gain in animal and plant breeding (Bernardo and Yu, 2007; Meuwissen et al., 2001). Genomic prediction has been widely and successfully adopted in animal breeding, however, key points of difference were not fully considered in the transfer to plant breeding. The objective of Chapter 2 is to re-emphasise appropriate methods for analysing plant breeding MET datasets, with a particular focus on novel genomic prediction models for GEI which integrate known and latent environmental covariates.

Fig. 1.2 The hypothetical response of genotypes G1 and G2 in environments E1 and E2 for a single continuous trait. The figure demonstrates genotype response in terms of non-crossover and crossover GEI, which reflect changes in scale and rank between environments.

Historically, plant breeding programmes have handled GEI in one of three ways:

1. Ignore GEI by selecting genotypes with the highest average performance across all environments in the TPE.

2. Reduce GEI by grouping similar environments together and then selecting genotypes with the highest average performance in each group of environments.

3. Leverage GEI by selecting high performing and stable genotypes for specific environments in the TPE.

This thesis develops a wide-array of methods that leverage GEI for efficient prediction, selection and discovery in plant breeding. Each topic is introduced in the following.
Plant breeders use genomic prediction models to predict the future performance of genotypes in their TPE. The genotype by environment table can be considered in terms of a tested set of genotypes and environments with phenotypic data available and an untested set without phenotypic data (Malosetti et al., 2016). Predictions of untested genotypes are generally used to make selections prior to field evaluation (Gaynor et al., 2017), whereas predictions into untested environments are used to target and deploy well adapted genotypes for further evaluation or commercial release (Piepho et al., 1998). Figure 1.3 demonstrates the genotype by environment table for a hypothetical MET dataset with 20 genotypes and 20 environments. Half of the genotypes and environments are tested while the other half are untested. The two-way table is divided into four subsets; (i) tested genotypes and environments, (ii) tested genotypes but untested environments, (iii) untested genotypes but tested environments, and (iv) untested genotypes and environments. Subset (i) is incomplete because phenotypic data
is generally not available for all genotype by environment combinations in the MET dataset. Prediction into (i) is often regarded as the simplest operation because it utilises phenotypic data, whereas prediction into (iv) is the most difficult because it relies solely on known genotypic and environmental covariates (Meuwissen et al., 2001; Van Eeuwijk et al., 1996). Chapter 2 of this thesis focuses on integrating the known covariates within novel random regression approaches. A brief overview of current approaches is provided in the following.

Regressions on environmental covariates were first used in plant breeding by Yates and Cochran (1938). Their approach was later popularised by Finlay and Wilkinson (1963) and Eberhart and Russell (1966), and included a fixed coefficient regression on latent environmental mean yields (covariates). Hardwick and Wood (1972) extended the fixed regression model to include a more complex set of known environmental covariates derived from weather data (also see Freeman and Perkins, 1971; Fripp, 1972; Wood, 1976). These approaches have been widely adopted in plant breeding, however, they have distinct limitations when used to analyse MET datasets (Smith et al., 2005). An alternative approach is to use a linear mixed model with a random coefficient regression. This approach was popularised by Laird and Ware (1982), and involves fitting a variance matrix between the genotype intercepts and slopes. Recently, Heslot et al. (2014) extended the random regression model for genomic prediction using genotypic covariates derived from marker data. At a similar time, Jarquín et al. (2014) demonstrated a random regression model for a very large set of correlated environmental covariates. They found that the environmental covariates explained only 23% of the overall genetic variance. These examples highlight the current limitations of using known environmental covariates for genomic prediction. That is, they are often highly correlated and only explain a small proportion of GEI (Brancourt-Hulmel et al., 2000; Buntaran et al., 2021).

Regressions on latent environmental covariates were extended to linear mixed models for plant breeding by Gogel et al. (1995) (also see Oman, 1991; Patterson and Nabugoomu, 1992). Their approach included a fixed coefficient regression on random environmental effects (covariates), resulting in a factor analytic structure for the genotype dimension. Piepho (1997) extended the factor analytic approach to include more general forms of the latent environmental covariates. An alternative approach was then developed by Smith et al. (2001) which fits the factor analytic structure to the environment dimension, rather than genotype dimension (also see Piepho, 1998). Their approach also considered appropriate non-genetic effects, including within-trial spatial variation. The factor analytic linear mixed models above have been shown to provide an informative model for GEI in terms of a small number of common factors (Smith et al., 2005), and a good fit to MET datasets in general (Kelly et al., 2007). They also bear similarities to the conventional regression models with
Introduction

one important difference; the environmental covariates are estimated from the data as well as the genotype slopes. As a result, however, factor analytic linear mixed models cannot currently be used to predict the response of genotypes in untested environments.

The objective of Chapter 2 is to develop a genomic prediction models which combine known genotypic and environmental covariates with latent environmental covariates estimated directly from the phenotypic data. The models are hereafter referred to as the integrated factor analytic linear mixed models (IFA-LMMs).

1.3 Selection tools

Selection for broad and specific adaptation is an important breeding decision in the presence of GEI. Traditionally, these selection decisions have been aided by some measure of overall performance and stability for each genotype (Smith and Cullis, 2018). However, current approaches do not completely separate non-crossover and crossover GEI, and this may reduce the efficiency of selection. The objective of Chapter 3 is to develop selection tools for one or more TPE, with a particular focus on disentangling non-crossover and crossover GEI.

Plant breeders use the predictions from MET datasets to identify genotypes with desirable adaptation patterns within their TPE, i.e. those genotypes with broad or specific adaptation (Bernardo, 2014). Genotypes with broad adaptation often have a favourable response in terms of non-crossover GEI and a negligible response in terms of crossover GEI. These genotypes are of high interest to breeders because they are high performing and stable across

---

Fig. 1.4 The hypothetical response of genotypes G1, G2 and G3 in environments E1 and E2 for a single continuous trait. The figure demonstrates genotype response in terms of broad and specific adaptation.
all environments in the TPE (Piepho, 1996). In contrast, genotypes with specific adaptation are high performing in a subset of environments only. Figure 1.4 demonstrates the adaptation patterns of three hypothetical genotypes evaluated in two hypothetical environments. Genotype G1 is specifically adapted to E2 and G2 is specifically adapted to E1, whereas G3 is broadly adapted to both environments. As more genotypes and environments are sampled in the MET dataset, the number of potential combinations increases which makes selecting for desirable adaptation more difficult. This has lead to the development of various measures of overall genotype performance and stability. A brief summary is provided in the following.

Measures of overall performance and stability were initially obtained from an analysis of variance (ANOVA) involving the two-way table of genotype by environment means. Overall performance was taken as the genotype main effect while stability was taken as some function of the residuals for each genotype (see, for example, Plaisted and Peterson, 1959; Shukla, 1972; Wricke, 1962). Measures were also obtained from the fixed coefficient regression by Finlay and Wilkinson (1963). Overall performance was taken as the genotype intercept from the regression while adaptability (stability) was taken as the genotype slope. Eberhart and Russell (1966) extended the concept of adaptability, or more specifically, stability, as the mean-square of the deviations around the regression (also see Rowe and Andrew, 1964). Other stability measures obtained from the fixed regression approach can be found in Perkins and Jinks (1968) and Tai (1971) (see Lin et al., 1986, for a complete review). The fixed regression approach has been widely adopted in plant breeding, however, it has distinct limitations when used to analyse MET datasets (Smith et al., 2005). The measures built on this approach also inherent the same limitations.

The selection tools of Smith and Cullis (2018) provide measures of overall performance and stability built on a factor analytic linear mixed model. The overall performance for a genotype is taken as the average value of the first factor regression while the stability is taken as the root mean square of the deviations around the regression. The conjecture here is that the first common factor predominately captures non-crossover GEI while the higher order factors predominately capture crossover GEI. However, this feature is a general artefact of plant breeding MET datasets, rather than a consequence of an approach that seeks to exclusively capture non-crossover GEI in the first factor and crossover GEI in the higher order factors.

The objective of Chapter 3 is to develop selection tools which provide measures of overall performance and stability for one or more TPE. The selection tools are built on a new approach that completely disentangles non-crossover and crossover GEI. A directional stability measure is also developed which partitions the conventional measures into components representing favourable and unfavourable adaptation.
1.4 Association studies

Genome-wide association studies (GWAS) are a powerful exploratory tool for biological discovery in a wide-array of genetic studies (Visscher et al., 2017; Zhu et al., 2008). Many of the current approaches are inherently inefficient, however, because they involve fitting a very large number of models to obtain the required test statistics for all markers of interest. The objective of Chapters 4 and 5 is to develop fast exact methods for GWAS that produce the required test statistics from the fit of a single linear mixed model.

Plant breeders use GWAS to identify and target genome regions of importance to their breeding objectives, e.g. those regions associated with agronomically important traits such as disease resistance and desired morphology. The most widely adopted approach involves a linear mixed model that sequentially fits and tests each marker of interest (Kang et al., 2008; Yu et al., 2006). The appealing feature of this approach is that population structure, kinship and other known covariance structures can be modelled simultaneously, which leads to a reduction in false positives (Li et al., 2014). Kinship is modelled by a random polygenic term that is parameterised by pedigree (Mrode, 2014) or marker data (Yu et al., 2006), with markers fitted directly or indirectly through a genomic relationship matrix (VanRaden, 2008). However, the use of marker data instead of pedigree has raised an important question in the literature; should the marker effect of interest also be included in the polygenic term? This has resulted in two alternative approaches for GWAS; (i) the marker-in approach which fits the marker effect as fixed and random in the polygenic term (Yu et al., 2006), and (ii) the marker-out approach which fits the marker effect as fixed but not random (Wang et al., 2014).

Despite their popularity, the marker-in and marker-out approaches are inherently inefficient because they involve fitting a separate model for each marker of interest. Zhou and Stephens (2012) addressed this issue for the marker-in approach by applying dimension reduction to the polygenic term (also see Lippert et al., 2011; Meyer and Tier, 2012). At a similar time, Kang et al. (2010) proposed an efficient approximation where the variance parameters are estimated using a baseline linear mixed model and then constrained for all subsequent models (also see Zhang et al., 2010). More recently, Gualdrón Duarte et al. (2014) proposed that all test statistics can be obtained from the fit of the same baseline linear mixed model, which elevates the need to fit any subsequent models (also see Bernal Rubio et al., 2016; Zhang et al., 2021). This approach was subsequently badged GWAS by GBLUP (Legarra et al., 2018). However, the marker-in approach fits the marker effect of interest as both fixed and random, which causes the two effects to become statistically aliased. GWAS by GBLUP emulates the marker-in approach, so that it also inherits this issue. The marker-out approach avoids the aliasing because it excludes the marker effect from the polygenic term, however, there is currently no equivalent of GWAS by GBLUP for the marker-out approach.
The objective of Chapter 4 is to develop fast exact methods for GWAS that produce the marker-in and marker-out test statistics from the fit of the same linear mixed model. The methods are generalised for any set of fixed and random effects and extended for testing a set of markers at a time. The methods are then extended for genotype by environment data in Chapter 5. The objective of this chapter is to develop fast exact methods for obtaining test statistics for the genomic prediction models in Chapter 2 and the selection tools in Chapter 3.

1.5 Simulation studies

Simulations are routinely used in plant breeding as a fast and cost-effective way to compare different breeding methodologies over time (Gaynor et al., 2021). Many of the current simulations, however, do not adequately capture the complexity of GEI observed in real-world plant breeding, which may result in spurious comparisons and overly optimistic projections of genetic gain. The objective of Chapter 6 is to develop a general framework for simulating GEI using the class of multiplicative models.

Numerous simulation packages have been developed to model plant breeding programmes, e.g. AlphaSimR (Gaynor et al., 2021), ADAM-Plant (Liu et al., 2019), ChromaX (Younis et al., 2023), GPOPSIM2 (Li et al., 2021), MOBPS (Pook et al., 2020) and QUGENE (Podlich and Cooper, 1998). These packages provide an efficient way to compare different breeding strategies, however, they generally over-simplify GEI which may produce spurious comparisons and overly optimistic projections of genetic gain. For example, AlphaSimR, ChromaX and GPOPSIM2 construct a single phenotype for each genotype comprising a genotype main effect, genotype by environment interaction effect and random error. The interaction effect is generally modelled through a single multiplicative term, where the environmental effect is randomly sampled each year of simulation. This produces a structure for GEI that is unrealistic and difficult to control. It is important to note that most of the current simulation packages do have the functionality to simulate multiple correlated traits (environments), so they do have the potential to include a more realistic framework for GEI.

Multiplicative models have gained popularity in plant breeding because they are effective at capturing non-crossover and crossover GEI. The most general model for GEI is the unstructured model, which fits a separate genetic variance for each environment and a separate genetic covariance between each pair of environments. The unstructured model captures the maximum amount of GEI in the data, however, it becomes computationally prohibitive and unnecessarily complicated as the number of environments increases. These issues can be overcome using multiplicative models (Mandel, 1971). The appealing feature of multiplicative models is that they capture a large proportion of GEI with a small number
of multiplicative terms. Some traditional examples include AMMI (Gauch, 1992; Kempton, 1984), GGE (Yan et al., 2000) and factor analytic models (Piepho, 1997; Smith et al., 2001). These models have been shown to provide an informative model for GEI and a good fit to MET datasets in general (Gauch et al., 2008; Kelly et al., 2007).

The objective of Chapter 6 is to develop a general framework for simulating GEI using multiplicative models. The framework can be used to simulate realistic MET datasets and model plant breeding programmes that better reflect the complexity of real-world settings.

### 1.6 Objectives

Genotype by environment interaction (GEI) was traditionally viewed as an impediment to plant breeding. Over the past 100+ years, however, methods have been developed that leverage GEI for efficient prediction, selection and discovery in plant breeding. This thesis will develop methods and concepts that build on the past 100+ years of experimental design and statistical analysis of genotype by environment data.

- The objective of Chapter 2 is to develop new genomic prediction models that combine known genotypic covariates derived from marker data and known environmental covariates derived from weather and soil data along with latent environmental covariates estimated directly from the phenotypic data. The models constitute the class of integrated factor analytic linear mixed models (IFA-LMMs).

- The objective of Chapter 3 is to develop selection tools that provide measures of overall performance, responsiveness and stability for one or more target population of environments (TPE). The selection tools are built on a new approach that completely disentangles non-crossover and crossover GEI.

- The objective of Chapters 4 and 5 is to develop fast exact methods for conducting genome-wide association studies (GWAS) using genotype by environment data. The methods produce all required test statistics from the fit of a single linear mixed model, rather than fitting a very large number of models for all marker effects of interest.

- The objective of Chapter 6 is to develop a general framework for simulating GEI using the class of multiplicative models. The framework can be used to simulate realistic multi-environment trial (MET) datasets and model plant breeding programmes that better reflect the complexity of real-world settings.

The methods will be demonstrated using a collaborating cotton breeding dataset from Bayer CropScience as well as publicly available and simulated datasets.
Chapter 2

Genomic prediction models

2.1 Prelude

This chapter contains the published manuscript *Genomic selection using random regressions on known and latent environmental covariates* by Tolhurst et al. (2022). The chapter presents an extensive review of conventional genomic prediction models for GEI, with a focus on random regression and factor analytic linear mixed models. It also introduces the integrated factor analytic linear mixed model (IFA-LMM), which combines known genotypic covariates derived from marker data and known environmental covariates derived from weather and soil data along with latent environmental covariates estimated directly from the phenotypic data.

There are three appealing features of the IFA-LMM:

1. The IFA-LMM captures GEI in terms of a small number of known and latent common factors, which simultaneously reduces the dimension of the known and latent environmental covariates.

2. Predictable GEI is modelled by the known genotypic and environmental covariates, which enables meaningful interpretation of GEI and genomic prediction of any tested or untested genotype into any tested or untested environment.

3. Observable GEI is modelled by the latent environmental covariates, which captures any remaining GEI not modelled by the known covariates, and thence ensures the IFA-LMM captures a large proportion of GEI overall.

The methods and concepts developed in this chapter have great potential to elucidate the biological drivers of GEI and improve predictive plant breeding, particularly with the emergence of rapidly changing environments and climate change.
Genomic selection using random regressions on known and latent environmental covariates

Daniel J. Tolhurst1 · R. Chris Gaynor2 · Brian Gardunia2 · John M. Hickey3 · Gregor Gorjanc1

Received: 5 October 2021 / Accepted: 28 June 2022 / Published online: 6 September 2022
© The Author(s) 2023

Abstract

Key message The integration of known and latent environmental covariates within a single-stage genomic selection approach provides breeders with an informative and practical framework to utilise genotype by environment interaction for prediction into current and future environments.

Abstract This paper develops a single-stage genomic selection approach which integrates known and latent environmental covariates within a special factor analytic framework. The factor analytic linear mixed model of Smith et al. (2001) is an effective method for analysing multi-environment trial (MET) datasets, but has limited practicality since the underlying factors are latent so the modelled genotype by environment interaction (GEI) is observable, rather than predictable. The advantage of using random regressions on known environmental covariates, such as soil moisture and daily temperature, is that the modelled GEI becomes predictable. The integrated factor analytic linear mixed model (IFA-LMM) developed in this paper includes a model for predictable and observable GEI in terms of a joint set of known and latent environmental covariates. The IFA-LMM is demonstrated on a late-stage cotton breeding MET dataset from Bayer CropScience. The results show that the known covariates predominately capture crossover GEI and explain 34.4% of the overall genetic variance. The most notable covariates are maximum downward solar radiation (10.1%), average cloud cover (4.5%) and maximum temperature (4.0%). The latent covariates predominately capture non-crossover GEI and explain 40.5% of the overall genetic variance. The results also show that the average prediction accuracy of the IFA-LMM is $0.02 - 0.10$ higher than conventional random regression models for current environments and $0.06 - 0.24$ higher for future environments. The IFA-LMM is therefore an effective method for analysing MET datasets which also utilises crossover and non-crossover GEI for genomic prediction into current and future environments. This is becoming increasingly important with the emergence of rapidly changing environments and climate change.

Introduction

This paper develops a single-stage genomic selection (GS) approach which integrates known and latent environmental covariates within a special factor analytic framework. The factor analytic linear mixed model of Smith et al. (2001) is an effective method for analysing multi-environment trial (MET) datasets, which includes a parsimonious model for genotype by environment interaction (GEI). The advantage of using random regressions on known environmental covariates, such as soil moisture and maximum temperature, is that the modelled GEI becomes predictable. The GS approach developed in this paper exploits the desirable features of both classes of model.

Genomic selection is a form of marker-assisted selection that can improve the genetic gain in animal and plant breeding programmes (Meuwissen et al. 2001). In plant breeding, however, GS is often restricted by the presence of GEI, that is the change in genotype response to a change in environment. There are two appealing features of using known environmental covariates for GS; (i) meaningful biological interpretation can be ascribed to GEI and (ii) predictions can be obtained for any tested or untested genotype into any current or future environment. These features...
represent two long-standing objectives of many plant breeding programmes.

Regressions on known environmental covariates were first used in plant breeding by Yates and Cochran (1938). Their approach was later popularised by Finlay and Wilkinson (1963), and includes a fixed coefficient regression on a set of environmental mean yields (covariates) with a separate intercept and slope for each genotype. Hardwick and Wood (1972) extended the fixed regression model to include a more complex set of environmental covariates, such as moisture and temperature (also see Wood 1976). These approaches have distinct limitations when used to analyse MET datasets, however (Smith et al. 2005). An alternative approach is to use a linear mixed model with a random coefficient regression. This approach was popularised by Laird and Ware (1982), and requires an appropriate variance model for the intercepts and slopes which ensures the regression is scale and translational invariant. Harsh et al. (2014) extended the random regression model for GS using a set of genotype covariates derived from marker data and a set of environmental covariates derived from weather data. They were unable to fit an appropriate variance model for the intercepts and slopes, however, so that the regression was not translational invariant. At a similar time, Jarquin et al. (2014) demonstrated an even simpler random regression model for a very large set of correlated environmental covariates. They found that the environmental covariates explained only 23% of the overall genetic variance. These examples highlight the current limitations of using known environmental covariates for GS. That is, they are often highly correlated and only explain a small proportion of GEI, and fitting an appropriate variance model is typically prohibitively expensive (Brancourt-Hulmel et al. 2000; Buntaran et al. 2021).

The factor analytic linear mixed model of Smith et al. (2001) includes a latent regression model for GEI in terms of a small number of common factors (also see Piepho 1997). This approach is a linear mixed model analogue to AMMI (Gauch 1992) and GGE (Yan et al. 2000), or more specifically factor analysis (Mardia et al. 1979), where the factors involve some combination of latent environmental covariates. It also bears similarities to the ordinary regression models with one important difference; the environmental covariates are estimated from the data as well as the genotype slopes. Several authors have discussed the addition of intercepts to the factor analytic model in an attempt to obtain a simple average (simple main effect) for each genotype, but note there are issues which limit their interpretability (Smith 1999).

The factor analytic linear mixed model has been widely adopted for the analysis of MET datasets (Ukrainetz et al. 2018). The two main variants involve pedigree or marker data (Oakley et al. 2007, 2016). Recently, Tolhurst et al. (2019) demonstrated a factor analytic linear mixed model for GS within a major Australian plant breeding programme. They demonstrated genomic selection tools to obtain a measure of overall performance (generalised main effect) and stability for each genotype (Smith and Cullis 2018). There is one limitation of this approach, however. The common factors are latent so the modelled GEI is observable, rather than predictable. This limitation has lead to ad hoc post processing of the latent factors with known covariates (Oliveira et al. 2020).

Until now, the analysis of MET datasets has involved only one set of known or latent environmental covariates. The aim of this paper is to extend the GS approach of Tolhurst et al. (2019) to integrate both known and latent environmental covariates. This new approach is hereafter referred to as the integrated factor analytic linear mixed model (IFA-LMM). There are three appealing features of the IFA-LMM:

1. The IFA-LMM includes a regression model for GEI in terms of a small number of known and latent common factors. This simultaneously reduces the dimension of the known and latent environmental covariates.
2. The regression model captures predictable GEI in terms of known covariates. This enables meaningful interpretation of GEI and genomic prediction into any current or future environment.
3. The regression model also captures observable GEI in terms of latent covariates, which are orthogonal to the known covariates. This enables the regression model to capture a large proportion of GEI overall, and thence enables the IFA-LMM to be an effective method for analysing MET datasets.

The IFA-LMM is demonstrated on a late-stage cotton breeding MET dataset from Bayer CropScience. The predictive ability of the IFA-LMM is compared to several popular random regression models.

**Materials and methods**

The Bayer CropScience Cotton Breeding Programme evaluates the commercial merit of test genotypes by annually conducting multi-environment field trials. There are two late-stages of field evaluation considered in this paper, referred to as preliminary commercial P1 and P2. The 2017 P1 MET dataset comprises the current set of environments and will be used to train all random regression models. The 2018 P2 MET dataset will be used to assess the predictive ability into future environments.
Data description

Experimental design and phenotypic data

Table 1 presents a summary of the 2017 P1 MET dataset for seed cotton yield. There were 72 field trials conducted in 24 environments across eight states in Southeast, Midsouth and Texas, USA (Fig. 1). A total of 208 genotypes were evaluated in all environments. Each environment consisted of three trials. Each trial was designed as a randomised complete block design with 144 plots comprising two replicate blocks of 68 test genotypes plus four checks. Yield data were recorded on most plots with 6.54% missing. The number of non-missing plots per test genotype ranged from 39 to 47, with mean of 45. The number of non-missing genotypes in common between environments ranged from 173 to 208, with mean of 204. The mean yield and generalised narrow-sense heritability (Oakey et al. 2006) varied substantially between environments and growing regions.

Supplementary Table 9 presents a summary of the 2018 P2 MET dataset for seed cotton yield. There were 20 field trials conducted in 20 environments across six states of USA (Fig. 1). Eleven trials were conducted in the same locations as the 2017 P1 trials and nine were conducted in new locations. A total of 55 genotypes were evaluated in all trials, with all genotypes previously evaluated in 2017 P1. Each trial was designed as a completely randomised design with a single replicate of all 55 genotypes. Note that only three environments were harvested in the Southeast due to severe weather.

Environmental data

Table 2 and Supplementary Table 10 present a summary of the known environmental covariates in the 2017 P1 and 2018 P2 MET datasets. There were 18 covariates available for all 44 environments, including latitude and longitude as well as 11 covariates derived from daily weather data.

Table 1 Summary of the 2017 P1 MET dataset for seed cotton yield

<table>
<thead>
<tr>
<th>State</th>
<th>Env</th>
<th>Trials</th>
<th>Genotypes*</th>
<th>Plots</th>
<th>Yield</th>
<th>NAs</th>
<th>Mean</th>
<th>$h^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>△ North Carolina</td>
<td>17NC1</td>
<td>3</td>
<td>208</td>
<td>15</td>
<td>189</td>
<td>4</td>
<td>432</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>17SC1</td>
<td>3</td>
<td>206</td>
<td>0</td>
<td>202</td>
<td>4</td>
<td>432</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>17SC2</td>
<td>3</td>
<td>183</td>
<td>52</td>
<td>127</td>
<td>4</td>
<td>432</td>
<td>107</td>
</tr>
<tr>
<td>△ South Carolina</td>
<td>17SC3</td>
<td>3</td>
<td>208</td>
<td>5</td>
<td>199</td>
<td>4</td>
<td>432</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>17GA1</td>
<td>3</td>
<td>208</td>
<td>2</td>
<td>202</td>
<td>4</td>
<td>432</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>17GA2</td>
<td>3</td>
<td>208</td>
<td>2</td>
<td>202</td>
<td>4</td>
<td>432</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>17GA3</td>
<td>3</td>
<td>208</td>
<td>2</td>
<td>202</td>
<td>4</td>
<td>432</td>
<td>2</td>
</tr>
<tr>
<td>△ Georgia</td>
<td>17GA4</td>
<td>3</td>
<td>208</td>
<td>2</td>
<td>202</td>
<td>4</td>
<td>432</td>
<td>2</td>
</tr>
<tr>
<td>° Missouri</td>
<td>17MO1</td>
<td>3</td>
<td>207</td>
<td>69</td>
<td>134</td>
<td>4</td>
<td>432</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>17AR1</td>
<td>3</td>
<td>207</td>
<td>18</td>
<td>185</td>
<td>4</td>
<td>432</td>
<td>20</td>
</tr>
<tr>
<td>° Arkansas</td>
<td>17AR2</td>
<td>3</td>
<td>205</td>
<td>2</td>
<td>199</td>
<td>4</td>
<td>432</td>
<td>9</td>
</tr>
<tr>
<td>° Mississippi</td>
<td>17MS1</td>
<td>3</td>
<td>204</td>
<td>9</td>
<td>191</td>
<td>4</td>
<td>432</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>17MS2</td>
<td>3</td>
<td>207</td>
<td>6</td>
<td>197</td>
<td>4</td>
<td>432</td>
<td>10</td>
</tr>
<tr>
<td>° Louisiana</td>
<td>17LA1</td>
<td>3</td>
<td>207</td>
<td>140</td>
<td>63</td>
<td>4</td>
<td>432</td>
<td>150</td>
</tr>
<tr>
<td>° Texas</td>
<td>17TX1</td>
<td>3</td>
<td>208</td>
<td>11</td>
<td>193</td>
<td>4</td>
<td>432</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>17TX2</td>
<td>3</td>
<td>208</td>
<td>1</td>
<td>203</td>
<td>4</td>
<td>432</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>17TX3</td>
<td>3</td>
<td>208</td>
<td>2</td>
<td>202</td>
<td>4</td>
<td>432</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>17TX4</td>
<td>3</td>
<td>208</td>
<td>4</td>
<td>199</td>
<td>4</td>
<td>432</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>17TX5</td>
<td>3</td>
<td>198</td>
<td>132</td>
<td>62</td>
<td>4</td>
<td>432</td>
<td>161</td>
</tr>
<tr>
<td></td>
<td>17TX6</td>
<td>3</td>
<td>206</td>
<td>29</td>
<td>173</td>
<td>4</td>
<td>432</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>17TX7</td>
<td>3</td>
<td>208</td>
<td>7</td>
<td>197</td>
<td>4</td>
<td>432</td>
<td>7</td>
</tr>
<tr>
<td>× Texas</td>
<td>17TX8</td>
<td>3</td>
<td>208</td>
<td>18</td>
<td>186</td>
<td>4</td>
<td>432</td>
<td>19</td>
</tr>
<tr>
<td>Overall</td>
<td>72</td>
<td>208</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>10,368</td>
<td>678</td>
</tr>
</tbody>
</table>

|                  |       |       |               |       |       |     |       |     |       |     |

Presented for each environment is the number of trials, genotypes (with one, two or more replicates) and plots (total and missing), as well as the mean yield (t/ha) and generalised narrow-sense heritability ($h^2$).

Note: Symbols distinguish the △ Southeast, ° Midsouth and × Texas growing regions

*Total number after missing plots removed
Fig. 1 Map of the cotton growing environments in the 2017 P1 and 2018 P2 MET datasets. Note: States and years are distinguished by colour and growing regions are distinguished by shape.

Table 2 Summary of the known environmental covariates in the 2017 P1 MET dataset

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Description (units)</th>
<th>△ Southeast</th>
<th>◦ Midsouth</th>
<th>× Texas</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAT</td>
<td>latitude (°)</td>
<td>31.0</td>
<td>33.0</td>
<td>35.4</td>
</tr>
<tr>
<td>LONG</td>
<td>longitude (°)</td>
<td>−84.7</td>
<td>−81.7</td>
<td>−78.0</td>
</tr>
<tr>
<td>avgCCR</td>
<td>average cloud cover (%)</td>
<td>53.4</td>
<td>56.0</td>
<td>59.1</td>
</tr>
<tr>
<td>minHUM</td>
<td>min humidity (%)</td>
<td>43.7</td>
<td>47.7</td>
<td>53.7</td>
</tr>
<tr>
<td>maxDSR</td>
<td>max downward solar radiation (W/m²)</td>
<td>0.74</td>
<td>0.76</td>
<td>0.77</td>
</tr>
<tr>
<td>maxNSR</td>
<td>max net solar radiation (W/m²)</td>
<td>0.62</td>
<td>0.64</td>
<td>0.66</td>
</tr>
<tr>
<td>maxPRP</td>
<td>max precipitation (mm/hr)</td>
<td>2.4</td>
<td>2.9</td>
<td>3.4</td>
</tr>
<tr>
<td>totPRP</td>
<td>total precipitation (mm/day)</td>
<td>3.2</td>
<td>3.5</td>
<td>4.2</td>
</tr>
<tr>
<td>maxDPT</td>
<td>max dew point temperature (°C)</td>
<td>20.5</td>
<td>21.1</td>
<td>22.1</td>
</tr>
<tr>
<td>maxTMP</td>
<td>max temperature (°C)</td>
<td>28.5</td>
<td>30.3</td>
<td>31.5</td>
</tr>
<tr>
<td>minTMP</td>
<td>min temperature (°C)</td>
<td>19.0</td>
<td>20.1</td>
<td>21.0</td>
</tr>
<tr>
<td>minWSP</td>
<td>min wind speed (km/hr)</td>
<td>4.9</td>
<td>5.2</td>
<td>5.7</td>
</tr>
<tr>
<td>avgWDR</td>
<td>average wind direction (azimuth degrees)</td>
<td>166.7</td>
<td>175.8</td>
<td>181.5</td>
</tr>
<tr>
<td>maxST1</td>
<td>max soil temperature 1 (°C)</td>
<td>27.6</td>
<td>29.9</td>
<td>31.3</td>
</tr>
<tr>
<td>minST1</td>
<td>min soil temperature 1 (°C)</td>
<td>19.8</td>
<td>21.8</td>
<td>23.2</td>
</tr>
<tr>
<td>avgSM3</td>
<td>soil moisture 3 (%)</td>
<td>7.0</td>
<td>23.8</td>
<td>42.3</td>
</tr>
<tr>
<td>avgSM4</td>
<td>soil moisture 4 (%)</td>
<td>10.0</td>
<td>29.5</td>
<td>44.6</td>
</tr>
<tr>
<td>minST4</td>
<td>min soil temperature 4 (°C)</td>
<td>20.0</td>
<td>22.4</td>
<td>24.2</td>
</tr>
</tbody>
</table>

Note: Values presented are prior to centring and scaling
Presented for each covariate is the minimum, mean and maximum for the △ Southeast, ◦ Midsouth and × Texas growing regions.
and 5 covariates derived from daily soil data. These tables show that the known covariates vary substantially within and between growing regions, as well as between years. Each covariate was then centred and scaled to unit length for all subsequent analyses. The practical implication of this will be discussed in “Regressions on latent covariates”.

Marker data

Marker data were available for 204 (of the 208) genotypes in 2017 P1, which included all 55 genotypes in 2018 P2. The markers correspond to a high confidence set of 36,009 single-nucleotide polymorphisms. Genotypes were coded as either −1, 0 or 1 for the homozygous minor, heterozygous or homozygous major alleles at each marker. The frequency of heterozygous markers was low given the level of selfing accumulated up to the P1 stage. Monomorphic markers were then removed and missing markers were imputed using the $k$-nearest neighbour approach of Troyanskaya et al. (2001), with $k = 10$. Note that the four genotypes without marker data are of no practical interest (see Tolhurst et al. 2019, for further details).

The genomic relationship matrix was constructed using the pedicure package (Butler 2019) in R (R Core Team 2021). The default settings in pedicure were used as filters, with minor allele frequency $> 0.002\%$ and missing marker frequency $< 0.998\%$. A total of 24,265 markers were retained using this criteria. The diagonal elements of the relationship matrix ranged from 0.004 to 2.022, with mean of 1.234. The off-diagonals ranged from −0.388 to 1.322, with mean of −0.006.

Statistical models

Preliminaries

Assume the MET dataset comprises $v = 204$ genotypes evaluated in $t = 72$ field trials conducted across $p = 24$ environments, where $t = \sum_{j=1}^{q} t_j$ and $t_j = 3$ is the number of trials in environment $j$. Let the $n$-vector of phenotypic data be given by $y = (y_{1j}, y_{2j}, ..., y_{pj})$, where $y_{1j} = (y_{11j}, y_{12j}, ..., y_{13j})$ is the $n_{1j}$-vector for environment $j$ and $y_{1j}$ is the $n_{1j}$-vector for trial $k$ in environment $j$. The length of $y$ is therefore given by:

$$ n = \sum_{j=1}^{q} n_{1j} = \sum_{j=1}^{p} n_{1j}. $$

Lastly, assume all $p = 24$ environments have $q = 18$ known covariates available, that is assume $p > q$. Let the $p \times q$ matrix of covariates be given by $S = [s_1, s_2, ..., s_q]$, with columns given by the centred and scaled environment scores for each covariate, such that $s_i s_i = 1$.

Linear mixed model

The linear mixed model for $y$ can be written as:

$$ y = X \tau + Z u_p + e, $$

where $\tau$ is a vector of fixed effects with design matrix $X$, $u_p$ is a $p \times q$ matrix of known covariates, $Z$ is a $v \times q$ matrix of random genotypes by environment (GE) effects with $n \times q$ design matrix $Z$, $u_p$ is a vector of random non-genetic peripheral effects with design matrix $Z_p$ and $e$ is the $n$-vector of residuals.

The vector of fixed effects, $\tau$, includes the mean parameter for each environment. This vector is fitted as fixed following a classical quantitative genetics approach where the GE effects in different environments are regarded as different correlated traits (Falconer and Mackay 1996). This vector can be extended to a regression on known environmental covariates, with:

$$ \tau = \mu + S \tau_s + \omega, $$

where $\mu$ is the overall mean parameter (intercept), $S$ is the $p \times q$ matrix of known covariates, $\tau_s$ is a $q$-vector with elements given by the mean response of genotypes to each covariate and $\omega$ is a $p$-vector of residual environmental effects, with $\omega \sim N(0, \sigma^2 I_p)$.

The vector of random non-genetic effects, $u_p$, accommodates the plot structures of trials and environments (Bailey 2008). This vector is fitted as random to enable recovery of information across incomplete blocks and trials (Patterson and Thompson 1971). Other effects in $u_p$ may accommodate extraneous variations across field columns and rows (Gilmour et al. 1997).

It is assumed that:

$$ \begin{bmatrix} u \\ u_p \\ e \end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0_p \\ 0 \\ 0 \\ 0_p \\ 0 \end{bmatrix}, \begin{bmatrix} G & 0 & 0 \\ 0 & G_p & 0 \\ 0 & 0 & R \end{bmatrix} \right). $$

Following Tolhurst et al. (2019), $G_p = \bigotimes_{j=1}^p G_{j}$ is diagonal with a separate variance component model for the $j$th environment and $R = \bigotimes_{j=1}^p R_{j}$ is block diagonal with a two-dimensional spatial model for the $j$th environment. The form of $G$ is presented below, but note that all variance matrices in Eq. 3 are fitted at the environment level, not trial level. This completely aligns the non-genetic and residual variance models with the genetic variance model.
Variance model for the GE effects

The GE effects are modelled using $r = 24,265$ markers, and therefore referred to as the additive GE effects. This model is an extension of the univariate GBLUP model (Stranden and Garrick 2009), with:

$$u = (I_p \otimes M)u_m \quad \text{and} \quad G = G_e \otimes \Sigma^{-1}/m = G_e \otimes G_g. \quad (4)$$

where $M = [\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_v]$ is a $v \times r$ design matrix with columns given by the centred genotype scores for each marker, $u_m$ is a $r \times 1$-vector of additive marker by environment effects, $G_e$ is a $p \times p$ additive genetic variance matrix between environments and $G_g = MM'/m$ is the $v \times v$ genomic relationship matrix between genotypes (VanRaden 2008).

The random regression models for $u$ considered in this paper include:

1. Latent covariates; models with simple or generalised main effects.
2. Known covariates; models with or without translational invariance.
3. Known and latent covariates; models with generalised main effects and translational invariance.

All regression models are summarised in Table 3, with full details provided below.

### Table 3: Summary of the variance models for the additive GE effects considered in this paper

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
<th>$G_e$</th>
<th>Parameters</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>Identity</td>
<td>$\sigma^2_{ge} I_p$</td>
<td>1</td>
<td>Patterson et al. (1977)</td>
</tr>
<tr>
<td>diag</td>
<td>Diagonal</td>
<td>$\Sigma_{ge}$</td>
<td>$\sigma^2_{ge} \otimes \Sigma_{ge}$</td>
<td>$p$</td>
</tr>
<tr>
<td>comp</td>
<td>Compound symmetry</td>
<td>$\sigma^2_{ge} I_p + \sigma^2_{p} I_p$</td>
<td>$\Sigma_{ge}$</td>
<td>$\sigma^2_{ge} \otimes \Sigma_{ge}$</td>
</tr>
<tr>
<td>mdiag</td>
<td>Main effects plus diagonal</td>
<td>$\Sigma_{ge}$</td>
<td>$\sigma^2_{ge} I_p + \Sigma_{ge}^2$</td>
<td>$p + 1$</td>
</tr>
<tr>
<td>FAMk</td>
<td>Factor analytic plus main effects</td>
<td>$\sigma^2_{ge} I_p + \Theta A^T + \Psi$</td>
<td>$D = \Theta \otimes \Psi$</td>
<td>$p(k + 1) - k(k - 1)/2 + 1$</td>
</tr>
<tr>
<td>FAk</td>
<td>Factor analytic</td>
<td>$\Theta A^T + \Psi$</td>
<td>$\Psi = \Theta \otimes \Psi$</td>
<td>$p(k + 1) - k(k - 1)/2$</td>
</tr>
<tr>
<td>rreg1</td>
<td>Random regression</td>
<td>$\sigma^2_{ge} I_p + \sigma^2_{SS} I_p$</td>
<td>$\Sigma_{ge}$</td>
<td>$\sigma^2_{ge} \otimes \Sigma_{ge}^2$</td>
</tr>
<tr>
<td>rreg2</td>
<td>Random regression</td>
<td>$\sigma^2_{ge} I_p + \Theta S S + \Psi$</td>
<td>$\Sigma_{ge}$</td>
<td>$\Theta \otimes \Sigma_{ge}^2$</td>
</tr>
<tr>
<td>FARK</td>
<td>Factor analytic regression</td>
<td>$[\mathbf{I}_p^* \otimes \mathbf{S}][\mathbf{A} \otimes (\mathbf{I}_p^* \mathbf{S})] + \Psi$</td>
<td>$\Lambda_k = \left[ \begin{array}{c} \Lambda_k^1 \ \Lambda_k^2 \end{array} \right]$</td>
<td>$p(k + 1) - k(k - 1)/2$</td>
</tr>
<tr>
<td>IFARK</td>
<td>Integrated factor analytic</td>
<td>$[\mathbf{S} \mathbf{\Gamma} \otimes \mathbf{A} \otimes (\mathbf{S} \mathbf{\Gamma})^\top + \Psi$</td>
<td>$\Lambda_k = \left[ \begin{array}{c} \Lambda_k^1 \ \Lambda_k^2 \end{array} \right]$</td>
<td>$p(k + 1) - k(k - 1)/2$</td>
</tr>
</tbody>
</table>

Presented for each model is the structure of the additive genetic variance matrix between environments ($G_e$), number of estimated variance parameters and the reference

**Note**: The $vp$-vector of additive GE effects is given by $u$ with $\text{var}(u) = G_e \otimes G_g$, where $G_e^{vp}$ is the variance matrix between environments and $G_g^{vp}$ is the genomic relationship matrix between genotypes. Also note that $\mathbf{I}_p^* = I_p/\sqrt{p}$, $A^{vp}$ is a matrix of latent covariates with $p$ environments and $k$ factors, $S^{qvp}$ is a matrix of known covariates with $q$ covariates and $\mathbf{F}^{vp(q-v)q}$ is an orthogonal projection matrix, with $\mathbf{S} \mathbf{\Gamma} = \mathbf{0}$.
Following Smith et al. (2021), the loadings are assumed to have orthonormal columns, with $\Lambda\Lambda = I_k$, and the scores are assumed to be independent across factors, with non-unit variance. It therefore follows that:

$$
\begin{bmatrix}
Y_1 \\
\delta
\end{bmatrix}
\sim N
\begin{bmatrix}
0 \\
0
\end{bmatrix}
\begin{bmatrix}
p\sigma^2_l & 0 & 0 \\
0 & D & 0 \\
0 & 0 & \Psi
\end{bmatrix}
\otimes G_k.
$$

where $\sigma^2_l$ is the intercept variance, $D = \otimes_{i=1}^p d_i$ is a diagonal matrix in which $d_i$ is the score variance for the $i$th latent factor ordered as $d_1 > d_2 > \ldots > d_k$ and $\Psi = \otimes_{j=1}^k \psi_j$ is a diagonal matrix in which $\psi_j$ is the specific variance for the $j$th environment. The variance matrix for $u$ is then given by:

$$
G = \left( \begin{bmatrix}
1_p \Lambda^* & \begin{bmatrix}
p\sigma^2_l & 0 \\
0 & D
\end{bmatrix}
\end{bmatrix}
\begin{bmatrix}
\Lambda^* \\
\Psi
\end{bmatrix}
\right)
\otimes G_k.
$$

where $G_k \equiv \sigma^2_l^2 I_p + AD\Lambda^* + \Psi$ and $J_p = 1_p 1_p^\top$. This variance matrix highlights the analogy to a random regression without translational invariance, that is where the intercepts and slopes are independent (see Eq. 14).

Note that the intercepts in $Y_1$ reflect the fitted value of each genotype at zero values of the environmental loadings. In order for the intercepts to reflect true main effects, however, the average values of the loadings must also be zero. The analogy to ordinary regression models is when the known covariates are column centred, so that the intercepts will reflect main effects taken at average (zero) values of the covariates.

Smith (1999) use a Gram-Schmidt process to column centre the environmental loadings (see “Appendix”). The variance matrix in Eq. 6 can therefore be written as:

$$
G = \left( \begin{bmatrix}
1_p \Lambda^* \\
D_{12}^\top \\
D_{22}^\top
\end{bmatrix}
\right)
\begin{bmatrix}
1_p \Lambda^* \\
\Psi
\end{bmatrix}
\otimes G_k.
$$

where $G_k \equiv \sigma^2_l^2 I_p + AD\Lambda^* + \Lambda^* D_{12}^\top + \Lambda^* D_{22}^\top + \Lambda^* D_{12}^\top + \Lambda^* D_{22}^\top + \psi$. with $\Lambda^* 1_p = 0$. This variance matrix highlights the analogy to a random regression with translational variance, that is where the main effects and slopes are dependent (see Eq. 19). This variance matrix also highlights the analogy to a special FAmk model, where the first factor loadings are constrained to be equal and the higher order loadings sum to zero.

The simple main effects are now equivalent to simple averages across environments, with:

$$
y_g = y_1 + \sqrt{p} \sum_i \hat{d}_i \hat{t}_i \quad \text{and} \quad y_g \sim N \left( 0, p\sigma^2_l G_k \right),
$$

where $\sigma^2_l = \sigma^2_l + \sum_i d_i^2 \hat{d}_i^2$ is the simple main effect variance and $\hat{d}_i = 1_p^\top \hat{d}_i \hat{t}_i / p$ is the mean loading for the $i$th latent factor. The distinguishing feature compared to the intercepts in Eq. 5 is that the simple main effects now reflect the fitted value of each genotype at average (zero) values of the loadings.

The percentage of additive genetic variance explained by the simple main effects is given by:

$$
v_g = 100 p\sigma^2_l / \text{tr}(G_k),
$$

where $G_k$ is defined in Eq. 7.

Models with generalised main effects

The conventional factor analytic (FAk) model is a simplification of the FAmk model in Eq. 5, with:

$$
u = (A \otimes I_1) y + \delta \quad \text{and} \quad G = (A\Lambda^* + \Psi) \otimes G_k
$$

where $G_k = A\Lambda^* + \Psi$. The distinguishing feature of this model is that intercepts are not explicitly fitted for each genotype (see “Appendix”).

Smith and Cullis (2018) discuss the ability of factor analytic models to capture heterogeneity of scale variance, that is non-crossover GEI, within the first factor. They proposed a set of generalised main effects based on this factor, with:

$$
y_g^* = \hat{d}_i \hat{t}_i \quad \text{and} \quad y_g^* \sim N \left( 0, d_i^2 \hat{d}_i^2 G_k \right),
$$

where $\hat{d}_i = 1_p^\top \hat{d}_i \hat{t}_i / p$ and $\hat{d}_i$ is the $p$-vector of first factor loadings which are assumed to be positive. The generalised main effects can therefore be viewed as weighted averages across environments. This highlights an important difference to the simple main effects in the FAmk model, which are simple averages across environments.

The percentage of additive genetic variance explained by the generalised main effects is equivalent to the variance explained by the first factor, which is given by:

$$
v_1 = 100 d_i^2 / \text{tr}(G_k),
$$

where $G_k$ is defined in Eq. 10. This measure will be compared to the variance explained by the simple main effects in “Results”.

Regressions on known covariates

The ordinary random regression model is given by:

$$
u = (1_p^* \otimes I_1) y + (s_i \otimes I_1) y_s + \ldots + (s_q \otimes I_1) y_s + \delta = (1_p^* \otimes I_1) y + (S \otimes I_1) r + \delta.
$$

where $y_g = (y_{g,1}, y_{g,2}, \ldots, y_{g,p}^q)$ is the $v$-vector of simple main effects, $S = [s_1, s_2, \ldots, s_q]$ is the $p \times q$ matrix of centred and scaled known environmental covariates, $r = [r_s^1, r_s^2, \ldots, r_s^q]^\top$.
is a $vq$-vector of genotype slopes in which $\mathbf{y}_k$ is the $v$-vector for the $k^{th}$ known covariate and $\mathbf{\delta} = (\delta_1, \delta_2, \ldots, \delta_p)$ is the $vp$-vector of regression residuals. This specification highlights the analogy to the FAM$k$ model in Eq. 5. Note, however, that the known covariates are already column centred so that the intercepts already reflect simple main effects.

**Models without translational invariance**

The random regression model in Heslot et al. (2014) assumes independent main effects and slopes with:

$$
\begin{bmatrix}
\mathbf{y}_k \\
\mathbf{r}_k
\end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ \rho \sigma^2 + \mathbf{0} \end{bmatrix}, \mathbf{\Sigma}_k \otimes \mathbf{G}_k \right).
$$

where $\sigma^2_k$ is the simple main effect variance and $\Sigma_k = \otimes_{i=1}^{p} \sigma_i^2$ is a diagonal matrix in which $\sigma_i^2$ is the slope variance for the $i^{th}$ known covariate. The distributional assumption for $\mathbf{y}_k$ may restrict interpretation, however, when the mean response to specific covariates is expected to be nonzero. The regression form of $\mathbf{r}$ in Eq. 2 overcomes this issue, with $\mathbf{r}_k \sim N(\mathbf{r}_k \otimes I, \mathbf{\Sigma}_k \otimes \mathbf{G}_k)$. The variance matrix for $\mathbf{u}$ is then given by:

$$
\mathbf{G} = \left(\begin{bmatrix} \mathbf{I}_p^* \mathbf{S} \\ \mathbf{0} \end{bmatrix} \mathbf{\Sigma}_k \otimes \mathbf{G}_k \right),
$$

where $\mathbf{G}_k \equiv \sigma^2_k \mathbf{I}_p + \mathbf{S} \mathbf{\Sigma}_k \mathbf{S}^\top + \Psi$.

The random regression model in Jarquín et al. (2014) uses an even simpler variance matrix for the slopes, with $\text{var}(\mathbf{r}_k) = \sigma^2_k \mathbf{I}_p \otimes \mathbf{G}_k$, where $\sigma^2_k$ is the slope variance across all known covariates. The variance matrix for $\mathbf{u}$ is then given by:

$$
\mathbf{G} = \left(\begin{bmatrix} \mathbf{I}_p^* \mathbf{S} \\ \mathbf{0} \end{bmatrix} \mathbf{\Sigma}_k \otimes \mathbf{G}_k \right),
$$

where $\mathbf{G}_k \equiv \sigma^2_k \mathbf{I}_p + \mathbf{S} \mathbf{\Sigma}_k \mathbf{S}^\top + \Psi$. Note that this random regression is neither scale nor translational invariant.

**Models with translational invariance**

Jennrich and Schlueter (1986) proposed an extension of the random regression model which includes a factor analytic model for the known environmental covariates. This extension will be referred to as the FAR$k$ model, where $k$ denotes the number of known factors. The FAR$k$ model for the simple main effects and slopes in Eq. 13 is given by:

$$
\mathbf{y}_k = (\Lambda_k \otimes I) \mathbf{f} + \mathbf{\delta}_k \quad \text{and} \quad \mathbf{r}_k = (\Lambda_k \otimes I) \mathbf{f} + \mathbf{\delta}_k.
$$

where $\mathbf{f} = (\mathbf{f}_1', \mathbf{f}_2', \ldots, \mathbf{f}_p')$ is the $vk$-vector of genotype scores which correspond to the $k$ known factors. The FAR$k$ model constructs a joint regression across the main effects and slopes, with loadings given by:

$$
\Lambda_k = \begin{bmatrix}
\lambda_{k1} & \lambda_{k2} & \ldots & \lambda_{kp}
\end{bmatrix} \quad \text{and} \quad \mathbf{A}_k = \begin{bmatrix}
\lambda_{k1} & \lambda_{k2} & \ldots & \lambda_{kp}
\end{bmatrix},
$$

where $\Lambda_k$ is a $k$-vector and $\mathbf{A}_k$ is a $q \times k$ matrix. The deviations in Eq. 16 are given by:

$$
\mathbf{\delta}_k = \begin{bmatrix}
\delta_{k1} \\
\delta_{k2} \\
\vdots \\
\delta_{kp}
\end{bmatrix} \quad \text{and} \quad \mathbf{\delta}_k = \begin{bmatrix}
\delta_{k1} \\
\delta_{k2} \\
\vdots \\
\delta_{kp}
\end{bmatrix}.
$$

The inclusion of the deviations in Eq. 16 may be unnecessary, however, particularly for higher order FAR$k$ models in which the percentage of variance explained by these effects is small. This leads to a reduced rank factor analytic model for the simple main effects and slopes (Kirkpatrick and Meyer 2004), with:

$$
\mathbf{y}_k = (\Lambda_k \otimes I) \mathbf{f} \quad \text{and} \quad \mathbf{r}_k = (\mathbf{A}_k \otimes I) \mathbf{f}.
$$

The main effects and slopes are assumed to be dependent, with:

$$
\mathbf{y}_k \sim N\left(\begin{bmatrix} 0 \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{A}_k \mathbf{DA}_k^\top & \mathbf{A}_k \mathbf{DA}_k^\top
\end{bmatrix} \otimes \mathbf{G}_k \right).
$$

where $\mathbf{D} = \otimes_{i=1}^{k} \mathbf{d}_i$ is the score variance matrix with diagonal elements ordered as $d_1 > d_2 > \ldots > d_k$. The FAR$k$ model is then obtained by substituting the vectors in Eq. 17 into Eq. 13, which gives:

$$
\mathbf{y}_k = \begin{bmatrix} \mathbf{I}_p^* \mathbf{A}_k \mathbf{S} + \mathbf{S} \mathbf{A}_k \mathbf{S} \mathbf{A}_k^\top + \Psi
\end{bmatrix} \mathbf{f} + \mathbf{\delta}.
$$

The variance matrix for $\mathbf{u}$ is then given by:

$$
\mathbf{G} = \left(\begin{bmatrix} \mathbf{A}_k \mathbf{DA}_k^\top & \mathbf{A}_k \mathbf{DA}_k^\top
\end{bmatrix} \mathbf{A}_k \mathbf{DA}_k^\top + \Psi \right) \otimes \mathbf{G}_k.
$$

where $\mathbf{G}_k \equiv \mathbf{A}_k \mathbf{DA}_k^\top + \Psi$. Note that this variance matrix is equivalent to the conventional FA$k$ variance matrix in Eq. 10 when $\mathbf{A}$ is square and has full rank.

**Regressions on known and latent covariates**

The integrated factor analytic (IFA$k$) model is an extension of the FAR$k$ model to include generalised main effects based on latent environmental covariates, instead of simple main effects. The IFA$k$ model can also be viewed as a special FA$k$ model with loadings constrained to be linear combinations
of two orthogonal sources of GEI, that is known and latent environmental covariates. The loadings matrix in Eq. 5 can therefore be written as:

$$\Lambda = S\Lambda_k + \Gamma \psi$$

where \( \Lambda = [\Lambda_k \Gamma \Lambda_r] \) is a \( p \times k \) matrix of basis functions which is assumed to have full rank. \( S = [s_1, s_2, \ldots, s_q] \) is the \( p \times q \) matrix of known environmental covariates and \( \Gamma = [r_1, r_2, \ldots, r_q] \) is a \( p \times (p - q) \) matrix of latent environmental covariates. The two loadings matrices in Eq. 20 correspond to the dependent and independent formulations of the IFAk model. The dependent formulation is translational invariant, and hence the focus of this paper. No further reference will be made to the independent formulation, but full details are provided in the Supplementary Material.

The dependent formulation constructs a joint regression across the known and latent environmental covariates. The \( p \times k \) matrix of joint factor loadings is given by:

$$[\Lambda_k \Lambda_r] = [\lambda_{k1} \lambda_{k2} \ldots \lambda_{kq} \lambda_{r1} \lambda_{r2} \ldots \lambda_{rk}] \tag{21}$$

where \( \Lambda_k \) is a \( q \times k \) matrix corresponding to the known covariates and \( \Lambda_r \) is a \( (p - q) \times k \) matrix corresponding to the latent covariates. The common factors underlying \( \Lambda_k \) and \( \Lambda_r \) are therefore referred to as the known and latent factors, and collectively as the joint factors.

The projection matrix in Eq. 20 is chosen to ensure that \( B \) has full rank and that the known and latent factors are orthogonal. This is achieved by projecting \( \Lambda_k \) onto the orthogonal complement to the space spanned by \( S \). A convenient choice for \( \Gamma \) is the first \( (p - q) \) columns in:

$$[I_p - SS^T S] \tag{22}$$

assuming that \( p > q \). This choice ensures that the same number of variance parameters are estimated as the conventional FAk model in Eq. 10. When \( p \gg q \), however, it may be desirable to take fewer than \( (p - q) \) columns in Eq. 22, and hence estimate fewer variance parameters. This enables the IFAk model to be scalable to a very large number of environments.

The IFAk model is obtained by substituting the first loadings matrix in Eq. 20 into Eq. 10, which gives:

$$u = ([S\Lambda_k + \Gamma \psi] \otimes I_k) f + \delta \tag{23}$$

where \( f = [f_1, f_2, \ldots, f_k] \) is the \( k \)-vector of genotype scores which correspond to the \( k \) joint factors.

The main difference to the FARk model in Eq. 18 is that there are now two vectors of slopes, with:

$$r_s = (A_s \otimes I_k) f \quad \text{and} \quad r_r = (A_r \otimes I_k) f \tag{24}$$

where \( r_s \) is a \( vq \)-vector corresponding to the known covariates and \( r_r \) is a \( v(p-q) \)-vector corresponding to the latent covariates. Another important difference is the addition of generalised main effects in \( r_r \), with:

$$y = N(0, \hat{\Lambda}^2_i G_k) \tag{25}$$

where \( \hat{\lambda}_i \) is the \( v \)-vector of genotype scores corresponding to the known and latent environmental covariates. The common factors underlying \( \Lambda_k \) and \( \Lambda_r \) are therefore referred to as the known and latent factors, and collectively as the joint factors.

### Model estimation

All variance models for the additive GE effects were implemented within the linear mixed model in Eq. 1. The two factor analytic linear mixed models with simple and generalised main effects are referred to as the FAM-LMM and FA-LMM, respectively. The other two linear mixed models developed in this paper are derived below.

The factor analytic regression linear mixed model (FAR-LMM) is obtained by substituting Eq. 18 into Eq. 1, which gives:

$$y = X r + Z_{\Lambda_k} f + Z\delta + Z_p u_p + e \tag{27}$$

where \( Z_{\Lambda_k} = Z(A\Lambda_k \otimes I_k) \). In this model, the covariances between the simple main effects and slopes are based on a reduced rank factor analytic model.

The integrated factor analytic linear mixed model (IFA-LMM) is obtained by substituting Eq. 23 into Eq. 1, which gives:

$$y = X r + Z_{\Lambda_k} f + Z\delta + Z_p u_p + e \tag{28}$$
where $Z_{A_h} = Z(BA_h \otimes I_1)$ and $A_h = \begin{bmatrix} A_s \\ A_g \end{bmatrix}$. In this model, the covariances between the known and latent environmental covariates are based on a reduced rank factor analytic model. The IFA-LMM will now be used to demonstrate all remaining methods. Similar results can be obtained for the other three linear mixed models where required.

**Rotation of loadings and scores**

Constraints are required in the IFA-LMM during estimation to ensure unique solutions for $\begin{bmatrix} A_s \\ A_g \end{bmatrix}$ and $D$. Following Smith et al. (2021), the upper right elements of $\begin{bmatrix} A_s \\ A_g \end{bmatrix}$ are set to zero when $k > 1$ and $D$ is set to $I_k$. Let the loadings and scores with these constraints be denoted by $\begin{bmatrix} A_s^* \\ A_g^* \end{bmatrix}$ and $f^*$, with $f^* \sim N(0, I_k \otimes G_k)$. The loadings and scores can be rotated back to their original form in Eq. 23 for interpretation. This rotation is given by:

$$
\begin{bmatrix} A_s \\ A_g \end{bmatrix} = \begin{bmatrix} A_s^* \\ A_g^* \end{bmatrix} V D^{-1/2} \quad \text{and} \quad f = (D^{1/2}V^T \otimes I_k) f^*, \quad (29)
$$

where $V$ is a $k \times k$ orthonormal matrix of right singular vectors and $D^{1/2}$ is a $k \times k$ diagonal matrix of singular values sorted in decreasing order, with $f \sim N(0, D \otimes G_k)$. These matrices can be obtained from the singular value decomposition given by:

$$
B \begin{bmatrix} A_s^T \\ A_g^T \end{bmatrix} = UD^{1/2}/V^T \quad \text{or} \quad A^* = UD^{1/2}/V^T, \quad (30)
$$

where $U$ is a $p \times k$ orthonormal matrix of left singular vectors, with $\begin{bmatrix} A_s^T \\ A_g^T \end{bmatrix} \equiv B^{-1}U$ and $\begin{bmatrix} A_s^T \\ A_g^T \end{bmatrix} \equiv B^{-1}A^*$, where $A^*$ is the loadings matrix in Eq. 10 with upper right elements set to zero (see “Appendix”). This demonstrates how the factor loadings in the IFA-LMM can be obtained directly from the fit of the conventional FA-LMM.

**Computation**

The IFA-LMM was coded in R (R Core Team 2021) using open source libraries. The computational approach for fitting the IFA-LMM is provided in the Supplementary Material. This approach obtains REML estimates of the variance parameters using an extension of the sparse formulation of the average information algorithm (Thompson et al. 2003). Let the REML estimates of the key variance parameters be denoted by $\tilde{A}$ and $\tilde{\Psi}$, with EBLUPs of the key random effects denoted by $\tilde{f}$ and $\tilde{\delta}$. All linear mixed models were also fitted in ASReml-R (Butler 2020), with known environmental covariates included using the mbf argument. An example R script is provided in the Supplementary Material.

**Model selection**

Order selection in the IFA-LMM was achieved using a combination of formal and informal criteria. Formal selection was achieved using the Akaike Information Criterion (AIC) and informal selection was achieved using two measures of variance explained. These measures are an extension of Smith et al. (2021) to include known environmental covariates, and are similar to the $R^2$ goodness-of-fit statistic in multiple regression. These measures are derived in the Supplementary Material.

The percentage of additive genetic variance explained by the known covariates and overall by the known and latent covariates is given by:

$$
\bar{v}_s = 100 \frac{\text{tr}(S \bar{A} \bar{A}^T S^T)}{\text{tr}(G_e)} \quad \text{and} \quad \bar{v} = 100 \frac{\text{tr}(D) }{\text{tr}(G_e)}, \quad (31)
$$

where $G_e$ is defined in Eq. 26. Similar measures are also obtained for the $j^{th}$ environment, that is $v_s$ and $v_j$. The final model order is typically chosen such that $\bar{v}_s$ and $\bar{v}$ are sufficiently high and the number of environments with low values of $v_s$ and $v_j$ is small. Note that this may require a different number of known and latent factors, that is $k_s$ and $k_j$.  

**Model assessment**

Model assessment of the IFA-LMM was achieved using the prediction accuracy for current and future environments. Prediction into current environments was assessed using leave-one-out cross-validation, where yield data for a single environment were excluded and then predicted. The additive GE effects for environment $j$ were predicted as:

$$
\tilde{u}_j = \left( [S_j \bar{A}_{s_j} + \bar{A}_{r_j}] \otimes I_q \right) \tilde{f}_j, \quad (32)
$$

where $S_j$ is a $q$-vector of known covariates, $\tilde{f}_j$ is a $q \times k$-vector of predicted scores for the $j^{th}$ genotype in the $j^{th}$ current environment and $\bar{A}_{r_j} = \mathbf{1}_{p-q \times k} \bar{A}_{r_j} / (p-1)$ ensures the scores are appropriately scaled by the latent covariates. Note that the factor loadings, $\bar{A}_{s_j}$ and $\bar{A}_{r_j}$, are estimated using data on the $(p-1)$ environments excluding the $j^{th}$ environment. The prediction accuracy for environment $j$ was then calculated as:
where \( \mathbf{\bar{T}}_j \) is a \( v_j \)-vector of genotype mean yields for the \( j^{th} \) current environment.

Prediction into future environments was assessed using a similar measure, but note that yield data for the entire year were excluded at once. The additive GE effects for environment \( j \) were then predicted as:

\[
\mathbf{\hat{u}}_j = \left( \mathbf{S}_j^T \mathbf{\hat{A}}_s + \mathbf{\hat{A}}_r \right) \otimes \mathbf{I}_j \mathbf{\hat{f}}_j^*.
\]

where \( \mathbf{S}_j \) is a \( q \)-vector, \( \mathbf{\hat{f}}_j^* \) is a \( v_j \times k \)-vector for the \( v_j \) genotypes in the \( j^{th} \) future environment and \( \mathbf{\hat{A}}_s = \mathbf{I}_{p-q} \mathbf{\hat{A}}_s / p \). In this case, the factor loadings, \( \mathbf{\hat{A}}_s \) and \( \mathbf{\hat{A}}_r \), are estimated using data on the \( p \) current environments only.

**Model summaries and interpretation**

The main limitation of the conventional FA-LMM is that the common factors are latent so they cannot be used for interpretation or prediction. The IFA-LMM overcomes this limitation since it integrates known environmental covariates into the common factors. Interpretation is then achieved using a series of regression plots and four measures of variance explained. The regression plots are an extension of Cullis et al. (2014) and the measures of variance explained are an extension of Eq. 31.

The percentage of additive genetic variance explained by known covariate \( i \) is given by:

\[
v_i = 100 \frac{\mathbf{S}_i^T \mathbf{\hat{S}}_i \mathbf{\hat{S}}_i^T}{\mathbf{\hat{S}}_i^T \mathbf{\hat{S}}_i \mathbf{\hat{S}}_i^T \text{tr}(\mathbf{G}_s)}. \tag{35}
\]

where \( \mathbf{G}_s \) is defined in Eq. 26. Note that \( \mathbf{\hat{V}}_i \neq \sum_{i=1}^q \mathbf{V}_i \) since the known covariates are not orthogonal. This issue is addressed in the Supplementary Material.

The percentage of additive genetic variance explained by known factor \( l \) and by joint factor \( l \) is given by:

\[
v_l = 100 \frac{\mathbf{A}_l^T \mathbf{S} \mathbf{\hat{S}}_l}{\text{tr}(\mathbf{G}_s)} \quad \text{and} \quad v_{jl} = 100 \frac{\mathbf{A}_l^T \mathbf{S} \mathbf{\hat{S}}_l}{\text{tr}(\mathbf{G}_s)}. \tag{36}
\]

Note that \( \mathbf{\hat{V}}_i = \sum_{i=1}^q \mathbf{V}_i \) and \( \mathbf{\hat{v}} = \sum_{i=1}^q \mathbf{v}_i \) since the known and joint factors are orthogonal.

Lastly, the percentage of additive genetic variance in joint factor \( l \) explained by known covariate \( i \) is given by:

\[
v_{li} = 100 \left( \mathbf{s}_i \mathbf{S}_i \mathbf{\hat{A}}_s \right)^2. \tag{37}
\]

The percentage of variance explained by all covariates is then given by \( v_i = 100 \left( \mathbf{S}_i \mathbf{S}_i \mathbf{\hat{A}}_s \right)^2 \), which is equivalent to \( v_i/v_l \) in Eq. 36.

**Results**

This section presents the results of model fitting using the 2017 P1 MET dataset and model assessment using the 2018 P2 MET dataset. The P1 dataset is summarised in Tables 1 and 2, and comprises \( v = 204 \) genotypes evaluated in \( p = 24 \) current environments with \( q = 18 \) known covariates. The P2 dataset is summarised in Supplementary Tables 9 and 10, and comprises \( v^* = 55 \) (of the 204) genotypes evaluated in \( p^* = 20 \) future environments with the same known covariates. The results are presented according to model selection, assessment and interpretation.

### Table 4: Linear mixed models with random regressions on latent environmental covariates

<table>
<thead>
<tr>
<th>Model</th>
<th>Pars</th>
<th>Loglik</th>
<th>AIC</th>
<th>( v )</th>
<th>( \mathbf{\bar{V}} )</th>
<th>Model</th>
<th>Pars</th>
<th>Loglik</th>
<th>AIC</th>
<th>( v )</th>
<th>( \mathbf{\bar{V}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>comp</td>
<td>2</td>
<td>10,504.2</td>
<td>20,478.4</td>
<td>36.2</td>
<td>36.2</td>
<td>id</td>
<td>1</td>
<td>10,156.9</td>
<td>20,055.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mdia</td>
<td>25</td>
<td>10,563.6</td>
<td>20,821.1</td>
<td>33.6</td>
<td>33.6</td>
<td>diag</td>
<td>24</td>
<td>10,249.3</td>
<td>20,194.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAM1</td>
<td>49</td>
<td>10,765.4</td>
<td>21,176.8</td>
<td>36.8</td>
<td>54.4</td>
<td>FA1</td>
<td>48</td>
<td>10,667.1</td>
<td>20,982.2</td>
<td>43.2</td>
<td>43.2</td>
</tr>
<tr>
<td>FAM2</td>
<td>72</td>
<td>10,893.8</td>
<td>21,387.6</td>
<td>37.2</td>
<td>67.5</td>
<td>FA2</td>
<td>71</td>
<td>10,827.4</td>
<td>21,256.8</td>
<td>44.1</td>
<td>60.4</td>
</tr>
<tr>
<td>FAM3</td>
<td>94</td>
<td>10,942.9</td>
<td>21,441.8</td>
<td>38.2</td>
<td>72.0</td>
<td>FA3</td>
<td>93</td>
<td>10,940.3</td>
<td>21,438.5</td>
<td>43.8</td>
<td>70.7</td>
</tr>
<tr>
<td>FAM4</td>
<td>115</td>
<td>10,981.7</td>
<td>21,477.5</td>
<td>38.1</td>
<td>76.9</td>
<td>FA4</td>
<td>114</td>
<td>10,978.3</td>
<td>21,472.5</td>
<td>43.8</td>
<td>75.2</td>
</tr>
<tr>
<td>FAM5</td>
<td>135</td>
<td>11,011.2</td>
<td>21,496.5</td>
<td>38.7</td>
<td>80.0</td>
<td>FA5</td>
<td>134</td>
<td>11,010.1</td>
<td>21,496.1</td>
<td>44.3</td>
<td>79.0</td>
</tr>
</tbody>
</table>

Presented for each model is the number of estimated genetic variance parameters, residual log-likelihood, AIC and percentage of variance explained by the simple \( \mathbf{v}_i \) or generalised \( \mathbf{v}_i \) main effects and overall \( \mathbf{\bar{V}} \).

Note: 128 non-genetic and residual variance parameters estimated in all models. The selected FAM4 and FA4 models are distinguished with **bold font**.

†Models where intercepts are not explicitly fitted.
Model selection

Tables 4 and 5 present the model selection criteria previously described in “Model selection”. The important results from each model fit are detailed below.

Baseline linear mixed models

The analyses commenced by fitting a linear mixed model with a diagonal model for the additive GE effects (diag; Table 4b). This approach reflects the initial single-site analyses routinely performed on MET datasets, where the additive GE effects in different environments are assumed to be independent. The single-site analyses are typically used to inspect the experimental design, address spatial variations and identify potential outliers.

The analyses continued by fitting a linear mixed model with a compound symmetry model for the additive GE effects (comp; Table 4a). This approach reflects many current applications of GS in plant breeding, where the additive GE effects in different environments are assumed to be correlated. The compound symmetry model is very restrictive, however, since it comprises a single variance component for the simple genotype main effects and genotype by environment interaction effects. This model can be extended to include heterogeneous interaction variances across environments, that is the main effects plus diagonal model (mdiag; Table 4a). The AIC for this model is much lower, and thence much better, than the standard compound symmetry model. There are negligible differences between the overall additive genetic variance explained, however, with $\bar{v} \approx 35\%$ for both models.

Regressions on latent covariates

A series of factor analytic linear mixed models were then fitted with either (a) simple or (b) generalised main effects (Table 4). The most notable differences between the FAM-LMMs and FA-LMMs are observed in the lower orders, where the overall additive genetic variance explained by the latent common factors is low. At the higher orders, where the overall variance explained is sufficiently high, the differences are negligible. Both models required $k = 4$ latent factors to reach a sufficient percentage of additive genetic variance explained for individual environments and overall, with $v_j > 40\%$ and $\bar{v} > 75\%$. Lastly, note that the generalised main effects in (b) explain 5.7% more variance than the simple main effects in (a), despite very similar overall variance explained. This feature is now discussed.

The simple and generalised main effects are demonstrated in Fig. 2. This figure presents a series of regression plots for checks C1 and C2 in terms of the (a) FAM4 and (b/c) FA4 models. Recall that the FAM4 model can be viewed as a special FA5 model where the first factor loadings are

![Fig. 2](image-url)

Fig. 2 Regression plots for checks C1 and C2 in terms of the first two factors obtained from the (a) FAM4 and (b/c) FA4 models. Note: The simple main effects in (a) and the generalised main effects in (b) are denoted with closed circles and the growing regions are distinguished by shape. The percentage of additive genetic variance explained by each factor is labelled. The additive GE effects in (c) have been adjusted for those in (b).
equal and correspond to the simple main effects, whereas the higher order loadings sum to zero and correspond to the interaction effects. The first two factors are plotted for the FAM4 model in Fig. 2a where the simple main effects are denoted by the fitted values of the second factor regressions at the mean loading of zero, that is 0.06 and −0.09 t/ha for C1 and C2. In contrast, the generalised main effects for the FA4 model in Fig. 2b are denoted by the fitted values of the first factor regressions at the mean loading of 0.19, that is 0.05 and −0.06 t/ha. There are two important differences between these approaches:

1. The generalised main effects capture heterogeneity of scale variance, that is non-crossover GEI, whereas the simple main effects do not capture GEI. This is demonstrated in Fig. 2b where the regression lines diverge across environments so the genotype rankings never crossover, whereas the first factor regression lines in the FAM4 model are always parallel (not shown).

2. The higher order factors in the FA4 model predominately capture crossover GEI only, whereas those in the FAM4 model capture some mixture of non-crossover and crossover GEI. This is demonstrated in Fig. 2c where the regression lines intersect so the genotype rankings crossover, whereas the regression lines in Fig. 2a diverge as well as crossover.

Regressions on known covariates

The next two linear mixed models fitted include random regressions without translational invariance. The random regression in Jarquín et al. (2014) reflects a popular application of GS in plant breeding (rreg; Table 5a). Like the compound symmetry model, however, this model is very restrictive since it only comprises two variance components. The only difference is that the interaction effects are now parametrised by known environmental covariates. This model can be extended to include heterogeneous interaction variances across covariates (rreg; Table 5a). The AIC for the random regression in Heslot et al. (2014) is much better than the simpler random regression. There are negligible differences between the additive genetic variance explained, however, with $\tilde{\nu}_j \approx 23\%$ and $\tilde{\nu} \approx 58\%$ for both models. Interestingly, the former measure matches that reported in Jarquin et al. (2014).

A series of FAR-LMMs with translational invariance were then fitted (Table 5a). This approach required $k = 4$ known factors to reach a sufficient percentage of additive genetic variance explained for individual environments and overall, with $\tilde{\nu}_j > 40\%$ and $\tilde{\nu} = 70.7\%$. The AIC for the FAR4 model is substantially better than the random regressions in Jarquín et al. (2014) and Heslot et al. (2014). The FAR4 model also explains more additive genetic variance in the known covariates, with $\tilde{\nu}_j = 33.2\%$ compared to only 20.8 and 23.2 %. This demonstrates the importance of appropriately modelling the variance structure between known covariates.

Regressions on known and latent covariates

The analyses concluded by fitting a series of IFA-LMMs with generalised main effects and translational invariance (Table 5b). This approach required $k = 4$ known and $k_s = 3$ latent factors to reach a sufficient percentage of additive genetic variance explained for individual environments and overall, with $\tilde{\nu}_j > 45\%$ and $\tilde{\nu} = 74.9\%$. The AIC for the IFA4-3 model is substantially better than the FAR4 model.

---

**Table 5** Linear mixed models with random regressions on known and latent environmental covariates

<table>
<thead>
<tr>
<th>Model</th>
<th>Pars</th>
<th>Loglik</th>
<th>AIC</th>
<th>$\tilde{\nu}_j$</th>
<th>$\tilde{\nu}$</th>
<th>Model</th>
<th>Pars</th>
<th>Loglik</th>
<th>AIC</th>
<th>$\tilde{\nu}_j$</th>
<th>$\tilde{\nu}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{rreg}_1$</td>
<td>26</td>
<td>10,721.2</td>
<td>−21,134.3</td>
<td>20.8</td>
<td>57.1</td>
<td>id</td>
<td>1</td>
<td>10,156.9</td>
<td>−20,055.9</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>$\text{rreg}_2$</td>
<td>43</td>
<td>10,750.7</td>
<td>−21,159.3</td>
<td>23.2</td>
<td>58.5</td>
<td>diag</td>
<td>24</td>
<td>10,249.3</td>
<td>−20,194.7</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>FAR1</td>
<td>43</td>
<td>10,636.7</td>
<td>−20,931.4</td>
<td>6.2</td>
<td>40.0</td>
<td>IFA1</td>
<td>48</td>
<td>10,667.1</td>
<td>−20,982.2</td>
<td>7.0</td>
<td>43.2</td>
</tr>
<tr>
<td>FAR2</td>
<td>61</td>
<td>10,791.4</td>
<td>−21,204.8</td>
<td>19.2</td>
<td>57.0</td>
<td>IFA2</td>
<td>71</td>
<td>10,827.4</td>
<td>−21,256.8</td>
<td>20.1</td>
<td>60.4</td>
</tr>
<tr>
<td>FAR3</td>
<td>78</td>
<td>10,887.0</td>
<td>−21,361.9</td>
<td>29.2</td>
<td>66.7</td>
<td>IFA3</td>
<td>93</td>
<td>10,940.3</td>
<td>−21,438.5</td>
<td>30.1</td>
<td>70.7</td>
</tr>
<tr>
<td>FAR4</td>
<td>94</td>
<td>10,911.7</td>
<td>−21,373.4</td>
<td>33.2</td>
<td>70.7</td>
<td>IFA4-3</td>
<td>108</td>
<td>10,971.9</td>
<td>−21,471.9</td>
<td>34.4</td>
<td>74.9</td>
</tr>
<tr>
<td>FAR5</td>
<td>109</td>
<td>10,931.3</td>
<td>−21,388.7</td>
<td>36.2</td>
<td>73.8</td>
<td>IFA5-3</td>
<td>122</td>
<td>10,996.4</td>
<td>−21,492.8</td>
<td>36.2</td>
<td>78.0</td>
</tr>
</tbody>
</table>

Presented for each model is the number of estimated genetic variance parameters, residual log-likelihood, AIC and percentage of variance explained by the known covariates ($\tilde{\nu}_j$) and overall ($\tilde{\nu}$).  

Note: 128 non-genetic and residual variance parameters estimated in all models. The models $\text{rreg}_1$ and $\text{rreg}_2$ correspond to the random regressions in Jarquin et al. (2014) and Heslot et al. (2014). The selected FAR4 and IFA4-3 models are distinguished with **bold font**.

*Models where intercepts are not explicitly fitted
The IFA4-3 model also explains more overall variance, that is $\bar{\nu} = 74.9\%$ compared to 70.7\%, despite similar variance explained by the known covariates, with $\bar{\nu}_1 \approx 35\%$ for both models. This demonstrates the advantage of including generalised main effects based on latent environmental covariates, instead of simple main effects.

**Model comparison**

The IFA4-3 model provides a good fit to the MET dataset and captures a large proportion of additive genetic variance (Table 5). The FAM4 and FA4 models also provide a good fit and capture a large proportion of variance, but they cannot be used for prediction into future environments (Table 4). The random regression models in Jarquín et al. (2014) and Heslot et al. (2014) can be used for prediction, but they provide a poor fit, capture the lowest variance of all models and are not translational invariant. The FAR4 model provides a better fit and captures more variance than the simpler random regression models, and is translational invariant. The IFA4-3 model provides an even better fit, captures more variance than the FAR4 model and is also translational invariant; making it the preferred method of analysis in this paper.

**Model assessment**

The mean prediction accuracy of the IFA4-3 model is considerably higher than all other random regression models (Table 6). The prediction accuracy was calculated in terms of 24 current environments in 2017 P1 and 20 future environments in 2018 P2. The most notable differences between models are observed for the 2018 environments in Texas, where the accuracy of the IFA4-3 model is at least 0.22 higher. In the Southeast and Midsouth, the accuracies are at least 0.06 and 0.10 higher, respectively. The differences in Texas are negligible for the 2017 environments, where the accuracies are generally higher for all models. In the Southeast and Midsouth, however, the accuracies of the IFA4-3 model are still at least 0.09 higher.

**Model summaries and interpretation**

Tables 7, 8 and Figs. 3, 4 present the model summaries previously described in “Model summaries and interpretation”. These summaries are presented for the IFA4-3 model in terms of environments, covariates and genotypes.

**Summary of environments and covariates**

Table 7 presents a summary of the growing environments in the 2017 P1 MET dataset. The additive genetic variance of individual environments range from 0.01 to 0.06, with mean of 0.03. These variances are obtained from the diagonal elements of the denominator in Eq. 31. The overall variance explained by the known and latent covariates is much higher than the variance explained by the known covariates alone, that is $v_1 = 44.3 - 100.0 \%$ with $\bar{\nu} = 74.9\%$ compared to $v_1 = 12.5 - 85.4 \%$ with $\bar{\nu}_1 = 34.4\%$. Most variance is explained overall in the Midsouth (84.9\% compared to only 66.6 and 69.3\%), whereas most variance is explained by the known covariates in Texas (41.1\% compared to only 28.4 and 33.4\%). Table 7 also presents REML estimates of the joint factor loadings. The first factor comprises positive loadings only, and explains $v_1 = 43.7\%$ of the additive genetic variance. The higher order factors comprise both positive and negative loadings, and explain $v_1 = 4.0 - 16.2 \%$, with 31.2\% in total. The sign of the loadings indicate that the first factor captures non-crossover GEI only, whereas the

Table 6  Summary of the prediction accuracies for the 2017 current and 2018 future environments

<table>
<thead>
<tr>
<th>Year</th>
<th>Model</th>
<th>△ Southeast</th>
<th>△ Midsouth</th>
<th>△ Texas</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min</td>
<td>Mean</td>
<td>Max</td>
<td>Min</td>
<td>Mean</td>
</tr>
<tr>
<td>2017</td>
<td>rreg1</td>
<td>0.27</td>
<td>0.51</td>
<td>0.68</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>rreg2</td>
<td>0.27</td>
<td>0.52</td>
<td>0.69</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>FAR4</td>
<td>0.25</td>
<td>0.50</td>
<td>0.66</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>IFA4-3</td>
<td>0.33</td>
<td>0.60</td>
<td>0.76</td>
<td>0.45</td>
</tr>
<tr>
<td>2018</td>
<td>rreg1</td>
<td>0.58</td>
<td>0.60</td>
<td>0.64</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>rreg2</td>
<td>0.58</td>
<td>0.61</td>
<td>0.64</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>FAR4</td>
<td>0.58</td>
<td>0.61</td>
<td>0.67</td>
<td>0.26</td>
</tr>
</tbody>
</table>

Presented for each model is the minimum, mean and maximum prediction accuracy for the △ Southeast, △ Midsouth and △ Texas, as well as overall across all regions.

Note: The models rreg1 and rreg2 correspond to the random regressions in Jarquín et al. (2014) and Heslot et al. (2014). The highest accuracy is distinguished with bold font.
higher order factors predominately capture crossover GEI only (Smith and Cullis 2018).

Table 8 presents a similar summary for the known environmental covariates in the MET dataset. The additive genetic covariance of individual covariates range from −0.33 to 0.25, with mean of 0.01. These covariances are obtained from the square-root of the elements in Eq. 37. The variance explained by individual covariates is $\tilde{v}_s = 0.10.1\%$, with $\bar{v}_s = 34.4\%$. The most notable covariates are maxDSR (10.1%), avgCCR (4.5%) and maxTMP (4.0%). Table 8 also presents REML estimates of the known factor loadings. The interpretation of these loadings is similar to above, but note that the higher order factors explain more additive genetic variance than the first factor, with 29.0% in total compared to only 5.4%. This will be discussed further below.

**Correlations between environments**

Figure 3 presents heatmaps of the additive genetic correlation matrices between environments in terms of the (a) known covariates and (b) known and latent covariates. These matrices are ordered based on the dendrogram constructed using the *agnes* function in the *cluster* package (Maechler et al. 2019). This dendrogram generally places environments closer together that have more similar GEI patterns than those further apart. Figure 3 suggests there is structure to the GEI underlying the heatmaps. There are three notable features:

1. The overall correlations based on the known and latent covariates are considerably higher than the correlations based on the known covariates alone.
2. The highest overall correlations generally occur between environments in the same growing region. Environments in the Southeast and Midsouth are also well correlated.
3. The overall correlations between environments in the same growing region are less than one. This indicates that crossover GEI is present within regions.
Table 8  The selected IFA4-3 model, Part 2: Summary of known environmental covariates

Presented are the REML estimates of additive genetic covariance, percentage of variance explained by individual known covariates ($v_s$) and estimates of the known factor loadings ($\hat{\lambda}_u$).

Note: The percentage of variance explained by all known covariates ($\bar{v}_s$) and by individual factors ($v_s$) is presented in the final row.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Covar</th>
<th>$v_s$</th>
<th>$\hat{\lambda}_u$</th>
<th>$\hat{\lambda}_k$</th>
<th>$\hat{\lambda}_l$</th>
<th>$\hat{\lambda}_h$</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAT</td>
<td>0.02</td>
<td>0.4</td>
<td>0.02</td>
<td>0.10</td>
<td>−0.21</td>
<td>−0.20</td>
</tr>
<tr>
<td>LONG</td>
<td>0.05</td>
<td>0.5</td>
<td>−0.18</td>
<td>0.04</td>
<td>0.56</td>
<td>0.33</td>
</tr>
<tr>
<td>avgCCR</td>
<td>−0.18</td>
<td>4.5</td>
<td>−0.37</td>
<td>0.31</td>
<td>−0.02</td>
<td>0.29</td>
</tr>
<tr>
<td>maxDPT</td>
<td>0.25</td>
<td>3.7</td>
<td>0.47</td>
<td>−0.46</td>
<td>−0.68</td>
<td>−0.22</td>
</tr>
<tr>
<td>maxDSR</td>
<td>0.25</td>
<td>10.1</td>
<td>−0.30</td>
<td>0.41</td>
<td>−0.10</td>
<td>0.17</td>
</tr>
<tr>
<td>minHUM</td>
<td>−0.33</td>
<td>3.5</td>
<td>−0.62</td>
<td>0.24</td>
<td>1.03</td>
<td>1.10</td>
</tr>
<tr>
<td>maxNSR</td>
<td>0.04</td>
<td>1.9</td>
<td>0.05</td>
<td>0.11</td>
<td>−0.19</td>
<td>−0.29</td>
</tr>
<tr>
<td>maxPRP</td>
<td>−0.01</td>
<td>0.1</td>
<td>0.04</td>
<td>0.05</td>
<td>−0.18</td>
<td>−0.55</td>
</tr>
<tr>
<td>totPRP</td>
<td>0.03</td>
<td>1.6</td>
<td>0.11</td>
<td>−0.01</td>
<td>0.05</td>
<td>−0.15</td>
</tr>
<tr>
<td>maxTMP</td>
<td>0.18</td>
<td>4.0</td>
<td>−0.31</td>
<td>0.09</td>
<td>0.58</td>
<td>0.32</td>
</tr>
<tr>
<td>minTMP</td>
<td>−0.18</td>
<td>3.1</td>
<td>−0.05</td>
<td>0.44</td>
<td>−0.67</td>
<td>−1.00</td>
</tr>
<tr>
<td>minWSP</td>
<td>0.01</td>
<td>0.1</td>
<td>−0.13</td>
<td>−0.09</td>
<td>0.31</td>
<td>0.16</td>
</tr>
<tr>
<td>avgWDR</td>
<td>−0.03</td>
<td>1.5</td>
<td>0.03</td>
<td>0.14</td>
<td>−0.01</td>
<td>−0.33</td>
</tr>
<tr>
<td>maxST1</td>
<td>−0.04</td>
<td>1.0</td>
<td>0.09</td>
<td>0.06</td>
<td>−0.27</td>
<td>−0.25</td>
</tr>
<tr>
<td>minST1</td>
<td>0.04</td>
<td>0.1</td>
<td>0.37</td>
<td>−0.48</td>
<td>0.15</td>
<td>0.96</td>
</tr>
<tr>
<td>avgSM3</td>
<td>−0.02</td>
<td>0.4</td>
<td>0.10</td>
<td>0.12</td>
<td>0.10</td>
<td>0.19</td>
</tr>
<tr>
<td>avgSM4</td>
<td>0.05</td>
<td>1.2</td>
<td>−0.10</td>
<td>0.15</td>
<td>−0.25</td>
<td>−0.41</td>
</tr>
<tr>
<td>minST4</td>
<td>0.09</td>
<td>1.4</td>
<td>−0.30</td>
<td>0.32</td>
<td>0.10</td>
<td>−0.48</td>
</tr>
<tr>
<td>Overall</td>
<td>0.01</td>
<td>34.4</td>
<td>5.4</td>
<td>15.3</td>
<td>9.8</td>
<td>3.9</td>
</tr>
</tbody>
</table>

Fig. 3  Heatmaps of the additive genetic correlation matrices between environments in terms of the (a) known covariates and (b) known and latent covariates. Note: Both matrices are ordered using the dendrogram applied to (b). Black lines distinguish the △ Southeast, ° Midsouth and × Texas cotton growing regions. The colourkey ranges from 1 (agreement in rankings) through zero (dissimilarity in rankings) to −1 (reversal of rankings).
Figure 4a presents a series of regression plots for checks C1 and C2 in terms of the $k = 4$ joint factors in the IFA4-3 model. These plots are used to assess genotype performance and stability in response to the known and latent environmental covariates. These plots show that check C1 is generally higher performing than C2 since it has a higher predicted slope for the first factor regression, that is 0.26 compared to $-0.32$. Both checks are considerably unstable, however, since they have large slopes for the higher order factors and therefore have large deviations about the first factor regression. Figure 4a also suggests that the second factor is correlated with longitude (Pearson’s $r = 0.80$), where the loadings on the left correspond to the Southeast and Midsouth while the loadings on the right correspond to Texas. This highlights an important limitation of the conventional FA-LMM, where interpretation is often limited to post-processing of the latent factors. This will be discussed further below.
Figure 4b presents direct interpretation of the factors in terms of the variance explained by the known environmental covariates. This figure suggests there is structure to the GEI underlying the regression plots. There are three notable features:

1. The known covariates predominately model crossover GEI, with $v_1 \approx 89.2 - 97.9\%$ of the additive genetic variance explained in the higher order factors compared to only $v_1 \approx 12.4\%$ explained in the first factor. These measures are obtained from Eq. 36, and are equivalent to $v_a/v_l$ in Tables 7 and 8.

2. The second factor is well explained by multiple known covariates. This demonstrates the biological drivers of crossover GEI in this factor, that is the drivers of crossover GEI due to changes in LONG.

3. The third and fourth factors are not well explained by individual covariates. This indicates that crossover GEI in these factors is driven by a combination of known covariates as well as their interaction.

### Discussion

This paper developed a single-stage GS approach which integrates known and latent environmental covariates within a special factor analytic framework. The FA-LMM of Smith et al. (2001) is an effective method for analysing MET datasets, but has limited practicality since the underlying factors are latent so the modelled GEI is observable, rather than predictable. The advantage of using random regressions on known environmental covariates is that the modelled GEI becomes predictable. The IFA-LMM developed in this paper includes a model for predictable and observable GEI in terms of a joint set of known and latent environmental covariates.

Regressions on known environmental covariates were first used in plant breeding by Yates and Cochran (1938). Their work was later popularised by Finlay and Wilkinson (1963), and includes a fixed coefficient regression on a set of environmental mean yields (covariates). Despite its popularity, however, there is a fundamental problem with using mean yields as covariates (Knight 1970; Freeman and Perkins 1971). This problem can be overcome by implementing environmental covariates which are independent of the genotypes under study, such as soil moisture and daily temperature (Hardwick and Wood 1972; Fripp 1972). Several authors have also used fixed regressions on genotype covariates, such as disease resistance and maturity, in addition to the environmental covariates. This approach is often referred to as fixed factorial regression (Denis 1980, 1988).

An alternative approach is to use a linear mixed model with a random coefficient regression. This approach was popularised by Laird and Ware (1982), and requires an appropriate variance model for the intercepts and slopes which ensures the regression is scale and translational invariant. An appropriate choice is the fully unstructured variance model, however, this model becomes computationally prohibitive as the number of covariates increases. Recently, Heslot et al. (2014) extended the random regression model for GS, but they were unable to fit an appropriate variance model (also see Jarquin et al. 2014). The FAR-LMM developed in this paper includes a reduced rank factor analytic variance model for the intercepts and slopes. This ensures the regression is computationally efficient as well as both scale and translational invariant, regardless of the number of covariates. The selected FAR-LMM also provides a substantially better fit and captures more additive genetic variance than the simpler random regression models.

The FAR-LMM includes a set of simple main effects which reflect simple averages across environments. Smith and Cullis (2018) discuss the limitations of simple main effects, and demonstrate how generalised main effects can be obtained from FA-LMMs. They also discuss how the generalised main effects capture heterogeneity of scale variance, that is non-crossover GEI, whereas the simple main effects do not. The generalised main effects can therefore be viewed as weighted averages across environments which are based on differences in scale variance. This highlights an important difference to the simple main effects, which are more restrictive and based on a single genetic variance across environments. This feature is demonstrated in Fig. 2 for the FA-LMM and the FAM-LMM, where the generalised main effects capture $\sim 6\%$ more additive genetic variance than the simple main effects.

The IFA-LMM is an effective method for analysing MET datasets which also utilises crossover and non-crossover GEI for genomic prediction into current and future environments. The IFA-LMM is effective since it exploits the desirable features of the FAR-LMM and the FA-LMM. That is, it exploits the ability of random regression models to capture crossover GEI for prediction using known covariates and the ability of factor analytic models to capture non-crossover GEI using latent covariates. The IFA-LMM can therefore be viewed as a random factorial regression, with known genotype covariates derived from marker data, known environmental covariates derived from weather and soil data as well as latent environmental covariates estimated from the phenotypic data itself. The IFA-LMM can also be viewed as a linear mixed model analogue to redundancy analysis (Van Den Wollenberg 1977), where the factors are constrained to be linear combinations of known and latent environmental covariates. The selected IFA-LMM provides a substantially...
better fit and captures more additive genetic variance than the selected FAR-LMM and the simpler random regression models. There are three appealing features of the IFA-LMM which address several long-standing objectives of many plant breeding programmes:

1. The IFA-LMM includes a regression model for GEI in terms of a small number of known and latent factors. This simultaneously reduces the dimension of the known and latent environmental covariates.
2. The regression model captures predictable GEI in terms of known environmental covariates. This is predominately in the form of crossover GEI, and enables meaningful interpretation and prediction into any current or future environment.
3. The regression model also captures observable GEI in terms of latent environmental covariates, which are orthogonal to the known covariates. This is predominately in the form of non-crossover GEI, and enables a large proportion of GEI to be captured by the regression model overall.

The IFA-LMM was demonstrated on a late-stage cotton breeding MET dataset. This dataset is an example of a small in situ training population which comprises a subset of current test genotypes and growing environments in 2017. A larger MET dataset across multiple years and locations is required, however, to capture the extent of transient and static GEI in the cotton growing regions of USA. This will ensure the scope of the known and latent covariates are relevant for prediction into future environments. Computational challenges are anticipated for these larger MET datasets and finding efficient ways to scale the IFA-LMM is the topic of current research.

There are four important points from “Results”:

1. The IFA4-3 model has fewer genetic variance parameters compared to the FA4 and FAM4 models, despite very similar model selection criteria (Tables 4 and 5). This highlights an important advantage of implementing known environmental information into the common factors. The IFA4-3 model also has better selection criteria than the FAR4 model. This also highlights the advantage of implementing generalised main effects based on latent environmental covariates, instead of simple main effects.
2. The known environmental covariates explain $\bar{v}_k = 34.4\%$ of the overall additive genetic variance, which represents 93.0% of the crossover GEI captured by the regression model. This is at least 11% more variance compared to the random regression models in Jarquín et al. (2014) and Heslot et al. (2014).
3. The latent environmental covariates explain 40.5% of the overall additive genetic variance, which represents 87.6% of the non-crossover GEI. This feature can be visualised in Fig. 3 where the overall correlations based on the known and latent covariates are much higher than those based on the known covariates alone.
4. The mean prediction accuracy of the IFA4-3 model is 0.02–0.10 higher than all other random regression models for current environments and 0.06–0.24 higher for future environments (Table 6). This highlights another important advantage of implementing known environmental information into the common factors.

Point 4 is now discussed further. The mean prediction accuracy of the IFA4-3 model was considerably higher than all other random regression models, especially for future environments in Texas. The prediction accuracy was calculated in terms of 24 current environments in 2017 P1 and 20 future environments in 2018 P2 (Table 6). The accuracy of all models were generally low for Texas in 2018, with mean of 0.20–0.44 for all models. This suggests that GEI is more complex in Texas and that there is substantial transient GEI present across years in addition to static GEI across locations (Cullis et al. 2000). It also suggests that the crossover GEI captured by the known covariates may not be repeatable across years and that the generalised main effects based on the latent covariates may not accurately capture the true non-crossover GEI across years. That is, the current scope of the known and latent covariates is less relevant for Texas compared to the Southeast and Midsouth. The application of a larger multi-year MET dataset should overcome these issues.

Another key feature of the IFA-LMM is the ability to identify the biological drivers of GEI, such as maximum downward solar radiation and average cloud cover. Interpretation within the IFA-LMM was demonstrated using a series of regression plots (Fig. 4). These plots are used to assess genotype performance and stability in response to the known and latent environmental covariates. Previously, interpretation within factor analytic linear mixed models was limited to post-processing of model terms, for example by correlating known covariates with latent factors (Oliveira et al. 2020) or by examining the response of reference genotypes in different environments (Mathews et al. 2011). The distinguishing feature of the IFA-LMM is the ability to ascribe direct biological interpretation to the modelled GEI. This feature has three important practical implications:

1. The first factor captures non-crossover GEI only, and is predominately explained by the latent environmental covariates. The higher order factors capture crossover GEI, and are predominately explained by the known environmental covariates. This enables the drivers of GEI across a set of target environments to be identified.
2. The importance of known covariates as drivers of GEI can be quantified. This provides information on which covariates should be measured with high accuracy, say, and which covariates may be less important or don’t need to be measured at all. This is particularly appealing with the advent of high-throughput environmental data.

3. Genomic selection tools can be applied to obtain measures of overall performance and stability for each genotype. This will enable the drivers of genotype performance and stability across a set of target environments to be identified. This is the topic of a subsequent paper.

The IFA-LMM is an effective method for analysing MET datasets which also utilises crossover and non-crossover GEI for genomic prediction into current and future environments. This is becoming increasingly important with the emergence of rapidly changing environments and climate change.

Appendix: Orthogonal matrix rotations

This appendix demonstrates how simple or generalised main effects can be obtained from factor analytic models regardless of whether intercepts are explicitly fitted. The simple main effects require rotation of the loadings and scores using a Gram-Schmidt process, whereas the generalised main effects require rotation to a principal component solution. The two rotations are detailed below.

Gram-Schmidt process

Smith (1999) discuss the need to column centre the environmental loadings in the FAMk model so they are orthogonal to the simple main effects. This is achieved using a Gram-Schmidt process, with:

\[
\begin{bmatrix} 1^*_p \cdot \Lambda^* \end{bmatrix} = \begin{bmatrix} 1^*_p \cdot \Lambda \end{bmatrix} U^{-1} \quad \text{and} \quad \begin{bmatrix} \gamma_k \cdot f^* \end{bmatrix} = \left( U \otimes \mathbf{1}_p \right) \begin{bmatrix} \gamma_k \cdot f \end{bmatrix}.
\]

with \(\Lambda^* \cdot \mathbf{1}_p = \mathbf{0}\), where \(\mathbf{1}_p = \mathbf{1}_p / \sqrt{p}\) and \(U = Q^* [1^*_p \cdot \Lambda]\) is a \((k + 1) \times (k + 1)\) upper triangular matrix in which the columns given by:

\[
\mathbf{q}_1 = 1^*_p \quad \text{and} \quad \mathbf{q}_{i+1} = \left( \lambda_i - \sum_{k=1}^{i} \mathbf{q}_k \cdot \lambda_k \right) / c_{i+1}, \quad (38)
\]

where \(l = 1, 2, \ldots, k\) and \(c_{i+1}\) is a constant chosen to ensure \(\mathbf{q}_{i+1}\) has unit length.

It is assumed that:

\[
\begin{bmatrix} 1^*_p \mbox{ and } \mathbf{0} \end{bmatrix}^\top \begin{bmatrix} \rho \sigma^2 \mbox{ and } \sigma^2 \end{bmatrix} \begin{bmatrix} D^*_{12} \mbox{ and } D^*_{22} \end{bmatrix} \otimes \mathbf{G}_k \sim N \left( \mathbf{0} \otimes \mathbf{I}_p, \mathbf{D} \otimes \mathbf{G}_k \right). \quad (39)
\]

where \(D^* = U \begin{bmatrix} \rho \sigma^2 \mathbf{0} \mbox{ and } \mathbf{0} \end{bmatrix} U\), \(\sigma^2 = \sigma^2_i + \sum_{l=1}^{k} d_l \lambda_l^2\) and \(\lambda_l = 1^*_p \cdot \lambda_l / p\). The FAMk variance matrix in Eq. 6 is now given by:

\[
\begin{bmatrix} 1^*_p \cdot \Lambda^* \end{bmatrix}^\top \begin{bmatrix} \rho \sigma^2 \mbox{ and } \sigma^2 \end{bmatrix} \begin{bmatrix} D^*_{12} \mbox{ and } D^*_{22} \end{bmatrix} \left[ 1^*_p \cdot \Lambda^* \right]^\top + \Psi \otimes \mathbf{G}_k. \quad (40)
\]

The conventional FAk model can be viewed as a special FAMk model where the intercept variance, \(\sigma^2\), is constrained to zero. The variance matrix in Eq. 10 can therefore be written as:

\[
\begin{bmatrix} 1^*_p \cdot \Lambda \end{bmatrix}^\top \begin{bmatrix} \rho \sigma^2 \mbox{ and } \sigma^2 \end{bmatrix} \begin{bmatrix} D^*_{12} \mbox{ and } D^*_{22} \end{bmatrix} \left[ 1^*_p \cdot \Lambda \right]^\top + \Psi \otimes \mathbf{G}_k. \quad (41)
\]

Simple main effects can be obtained from this model using a similar Gram-Schmidt process as above. The FAk variance matrix in Eq. 41 is now given by:

\[
\begin{bmatrix} 1^*_p \cdot \Lambda^* \end{bmatrix}^\top \begin{bmatrix} \rho \sigma^2 \mbox{ and } \sigma^2 \end{bmatrix} \begin{bmatrix} D^*_{12} \mbox{ and } D^*_{22} \end{bmatrix} \left[ 1^*_p \cdot \Lambda^* \right]^\top + \Psi \otimes \mathbf{G}_k. \quad (42)
\]

where \(\mathbf{G}_k \equiv \hat{\lambda}^2 \mathbf{I}_p + 1^*_p \cdot \mathbf{D}_{12} \cdot \mathbf{A}^\top + \Lambda^* \cdot \mathbf{D}_{12} \cdot 1^*_p \cdot \mathbf{A}^\top + \Lambda^* \cdot \mathbf{D}_{22} \cdot \mathbf{A}^\top + \Psi\) and \(\hat{\lambda}^2 = \sum_{l=1}^{k} d_l \lambda_l^2\) is the simple main effect variance, which is equal to \(\sigma^2_{s\cdot1}\) in Eq. 40 when \(\sigma^2_{s\cdot1} = 0\).

The FAk model in Eq. 10 can therefore be written as:

\[
\mathbf{u} = (1^*_p \otimes \mathbf{1}_p) \gamma_k + (\Lambda^* \otimes \mathbf{1}_p) \mathbf{f}^* + \delta, \quad (43)
\]

where \(\gamma_k\) is a \(v\)-vector of simple main effects, with:

\[
\gamma_k \sim N \left( \mathbf{0}, \rho \hat{\lambda}^2 \mathbf{G}_k \right). \quad (44)
\]

Principal component rotation

Constraints are required in the FAM-LMM and FA-LMM during estimation to ensure unique solutions for \(\Lambda\) and \(\mathbf{D}\). Following Smith et al. (2021), the upper right elements of \(\mathbf{A}\) are set to zero when \(k > 1\) and \(\mathbf{D}\) is set to \(\mathbf{I}_p\). Let the loadings and scores with these constraints be denoted by \(\mathbf{A}^*\) and \(\mathbf{f}^*\), with \(\mathbf{f}^* \sim N \left( \mathbf{0}, \mathbf{I}_p \otimes \mathbf{G}_k \right)\). The loadings and scores can be rotated back to their original form for interpretation. This rotation is given by:

\[
\mathbf{A} = \mathbf{A}^* \mathbf{V} D^{-1/2} \quad \text{and} \quad \mathbf{f} = (\mathbf{D}^{1/2} \mathbf{V}^\top \otimes \mathbf{1}_p) \mathbf{f}^*, \quad (45)
\]
where $V$ is a $k \times k$ orthonormal matrix of right singular vectors and $D^{1/2} = \oplus_{i=1}^{k} \sqrt{d_i}$ is a diagonal matrix of singular values sorted in decreasing order, with $f \sim N(0, D \otimes G_k)$.

These matrices are obtained from the singular value decomposition $\mathbf{A}' = U D^{1/2} V'$, where $U$ is a $p \times k$ orthonormal matrix of left singular vectors and $\Lambda \equiv U$ in Eq. 45.

The loadings and scores can then be rotated using the Gram-Schmidt process in the previous section to obtain simple main effects for either model. Alternatively, generalised main effects can be obtained for the FA-LMM using Eq. 11. In terms of the FAM-LMM, however, an alternative rotation is required which consumes the intercept variance, $\sigma^2_1$ into the factors. This rotation is given by:

$$\mathbf{A}' = [1_p \times \Lambda] V' \mathbf{D}^{-1/2}$$

and

$$\mathbf{f}' = (\mathbf{D}^{1/2} V' \otimes 1) \left( \begin{array}{c} f_1 \\ f \end{array} \right).$$

where $V'$ is a $(k+1) \times (k+1)$ orthonormal matrix and $\mathbf{D}^{1/2} = \oplus_{i=1}^{k} \sqrt{d_i}$ is a diagonal matrix, with $f' \sim N(0, D' \otimes G_k)$. These matrices are obtained from the singular value decomposition $[1_p \times \Lambda] = U' \mathbf{D}^{-1/2} V'$, where $U'$ is a $p \times (k+1)$ orthonormal matrix and $\Lambda' \equiv U'$.

The FAMk model in Eq. 5 can therefore be written as:

$$\mathbf{u} = (\mathbf{A}' \otimes 1) \mathbf{f}' + \delta,$$

where $\mathbf{A}'$ is a $p \times (k+1)$ matrix and $\mathbf{f}'$ is a $\nu (k+1)$-vector. The generalised main effects are based on the first factor, with:

$$y_k^* = \tilde{x}_k^* \mathbf{f}'_1$$

and

$$y_k^* \sim N(0, d_1^2 \tilde{x}_k^2 G_k),$$

where $\tilde{x}_k^* = 1_p \times \tilde{x}_k^2 / \nu$.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1017/CBO9780511611483.

**Acknowledgements** The authors thank Bayer CropScience for funding and use of their data. We also thank Kolbyn Joy, Nicholas Ames and Nilesh Dighe for their stimulating discussions and insights into the cotton breeding programme. Lastly, we sincerely thank the referees whose comments have led to an improved manuscript.

**Author contribution statement** DT conceived and developed the methodology, curated the data, conducted the analyses and wrote the manuscript. CG and BG provided input on plant breeding perspectives. JH organised the research project and secured funding. GG provided input on quantitative genetics perspectives. All authors have read and approved the final manuscript.

**Funding** This research was funded by Bayer CropScience through collaboration with The Roslin Institute.

**Data availability** The data that support the findings of this study are available from Bayer CropScience. Restrictions apply to the availability of these data, which were used under license for this study.

**Code availability** The R scripts to fit all linear mixed models in Table 3 using ASReml-R are provided in the Supplementary Material.

**Declarations**

**Conflict of Interest** The authors declare that they have no conflict of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

**References**


Butler DG (2020) asreml: Fits the Linear Mixed Model. http://vsni.co.uk/software/asreml-r, R package version 4.1.0


Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
2.2 Concluding remarks

This chapter introduced a new genomic prediction approach that integrates known and latent environmental covariates within a special factor analytic framework. The IFA-LMM was developed as an extension of an already effective approach for analysing MET datasets, i.e. the factor analytic linear mixed model. The objective here was to supplement that approach with a way to (i) ascribe direct biological interpretation to the underlying factors, and (ii) obtain forward predictions that were competitive with the conventional random regression approaches. The objectives were traditionally addressed using multiple processes and multiple methods of analysis. The IFA-LMM developed in this chapter addressed both objectives within a single, unified approach.

The IFA-LMM was demonstrated using an example cotton breeding dataset from Bayer CropScience. The results show that the known environmental covariates predominately captured crossover GEI and the latent environmental covariates predominately captured non-crossover GEI. The most notable known covariates were maximum downward solar radiation (10.1% of overall GEI), average cloud cover (4.5%) and maximum temperature (4.0%). The known covariates therefore provided a good basis to capture changes in genotype rank between environments for the example dataset but not changes in scale, which involved more complex (latent) interactions. This is an important result that can provide key insights into the nature of non-crossover and crossover GEI in plant breeding.

The results also show that the predictive ability of the IFA-LMM was higher than the conventional random regression approaches for tested and untested environments, particularly for the more variable environments in the example dataset. This demonstrates the value of integrating both known and latent environmental covariates to capture both non-crossover and crossover GEI. It also provides the first example of using the class of factor analytic linear mixed models for forward prediction, which addresses a longstanding objective of many breeders and researchers.

The methods and concepts developed in this chapter have great potential to elucidate the biological drivers of GEI and improve predictive plant breeding. At the very least, they demonstrate how known and latent environmental covariates can be integrated into a single, unified approach. Several formulations of the IFA-LMM were also developed, which provide plant breeding programmes with a general basis on which to tailor their analyses.

Areas of active research include covariate selection, non-linear smoothing between the known and latent covariates and scaling the IFA-LMM to large plant breeding datasets using novel extensions of the sparse formulation and data augmentation.
Chapter 3

Selection tools

3.1 Prelude

This chapter contains the unpublished manuscript *Disentangling non-crossover and crossover genotype by environment interaction for selection* by Tolhurst et al. (2024a). The chapter extends a recent set of selections tools for the IFA-LMM to provide measures of overall performance, responsiveness and stability that are tailored to one or more TPE. It also introduces a directional stability measure, which partitions the conventional stability measure into components representing favourable and unfavourable adaptation.

There are four measures of interest:

1. Overall performance, representing the average genotype response across environments.
2. Responsiveness, representing the specific genotype response to environmental factors.
4. Directional stability, representing the difference in variability above and below the average genotype response.

The selection tools are built on a new rotation that places all non-crossover variation into the first common factor and all crossover variation into the higher order factors. Previously, a principal component rotation was used that does not guarantee this feature. The selection tools are also generalised for any random regression type model to provide measures of responsiveness to different environmental covariates.

The methods and concepts developed in this chapter are becoming increasingly important when selecting in the presence of GEI, particularly with high levels of crossover interaction. The generalisations now enable a wide-array of genomic prediction models to be considered.
3.2 Linear mixed models

Assume a MET dataset comprises $v$ genotypes evaluated in $p$ environments, with $n$ plots in total. Let the $n$-vector of phenotypic data be given by $y = (y_1^\top, \ldots, y_p^\top)^\top$, where $y_j$ is the $n_j$-vector for environment $j$. Also assume that all $v$ genotypes have $m$ markers available and all $p$ environments have $q$ covariates available, such that $v \leq m$ and $p \geq q$.

The linear mixed model for $y$ can be written as:

$$ y = X\tau + Zu + e, \quad (3.1) $$

where $\tau$ is a vector of fixed effects with design matrix $X$, $u$ is a $vp$-vector of random genotype by environment (GE) effects with design matrix $Z$ and $e$ is the $n$-vector of residuals. The vector of fixed effects accommodates the mean parameter for each environment and genotypes without marker data (Tolhurst et al., 2019). This vector is initially fitted as fixed following a classical quantitative genetics approach where the GE effects in different environments are regarded as different correlated traits (Falconer and Mackay, 1996). The vector of residuals accommodates the experimental design (Bailey, 2008), spatial modelling (Gilmour et al., 1997) and residual variance heterogeneity (Smith et al., 2001). The methods developed in this chapter are generalised for any set of fixed and random effects.

The vector of fixed effects can be extended to include a regression on the known environmental covariates, which is given by:

$$ \tau = \mu 1_p + S\tau_s + \epsilon, \quad (3.2) $$

where $\mu$ is the overall mean parameter (intercept), $S = [s_1 \ldots s_q]$ is a $p \times q$ matrix of column centred and scaled known environmental covariates, $\tau_s$ is a $q$-vector with elements given by the mean response of genotypes to each covariate and $\epsilon$ is a $p$-vector of residual environmental effects, with $\epsilon \sim N(0, \sigma^2_\epsilon I_q)$. Other (random) regressions can be used where required.

It is assumed that:

$$ \begin{bmatrix} u \\ e \end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} G_a \otimes G & 0 \\ 0 & R \end{bmatrix} \right), \quad (3.3) $$

where $G_a \otimes G$ is a $vp \times vp$ additive genetic variance matrix with $p \times p$ between-environment variance matrix $G_a$ and $v \times v$ between-genotype relationship matrix $G$ (VanRaden, 2008) and $R$ is a $n \times n$ residual variance matrix, which is assumed to be completely general. The form of $G_a$ is developed below, but note that all variance matrices in Eq. 3.3 are assumed to be positive (semi)-definite.
3.3 Models for the genotype by environment effects

Following Smith et al. (2001), a factor analytic model is initially proposed for the GE effects. The factor analytic model is effective for modelling the covariances between GE effects in terms of a small number of latent common factors (Kelly et al., 2007). This model is referred to as the FA$_k$ model, where $k$ denotes the number of latent common factors.

The FA$_k$ model is given by:

$$u = (\lambda_1 \otimes I_v) f_1 + \ldots + (\lambda_k \otimes I_v) f_k + \delta$$

$$= (\Lambda \otimes I_v) f + \delta,$$  \hspace{1cm} (3.4)

where $\Lambda = [\lambda_1 \ldots \lambda_k]$ is a $p \times k$ orthogonal matrix of latent environmental loadings (covariates), $f = (f_{1\top}, \ldots, f_{k\top})\top$ is a $vk$-vector of genotype scores (slopes) and $\delta = (\delta_{1\top}, \ldots, \delta_{p\top})\top$ is a $vp$-vector of regression residuals (deviations), which are specific to individual environments. This specification highlights the analogy to an ordinary regression, with the key difference that the environmental covariates are estimated from the data as well as the genotype slopes.

Following Smith and Cullis (2018), the vector of GE effects can be rewritten as:

$$u = (\Lambda \otimes I_v) f + \delta$$

$$= \beta + \delta,$$ \hspace{1cm} (3.5)

where $\beta = (\Lambda \otimes I_v) f$ is a $vp$-vector of common GE effects and $\delta$ is a $vp$-vector of specific GE effects. The common GE effects will be used to construct the selection tools in Section 3.4, but note that the specific GE effects can also be included where appropriate.

Chapter 2 extended the FA$_k$ model to include a regression on known and latent environmental covariates. This extension is referred to as the IFA$_k$ model, where $k$ denotes the number of known and latent common factors. The IFA$_k$ model can be viewed as a special FA$_k$ model with loadings constrained to be linear combinations of known and latent environmental covariates, i.e. $\Lambda = S\Lambda_s + \Lambda_r$.

The IFA$_k$ model is given by:

$$u = ([S\Lambda_s + \Lambda_r] \otimes I_v) f + \delta,$$  \hspace{1cm} (3.6)

where $\Lambda_s = [\lambda_{s1} \ldots \lambda_{sk}]$ is a $q \times k$ matrix of loadings corresponding to the known environmental covariates and $\Lambda_r = [\lambda_{r1} \ldots \lambda_{rk}]$ is a $p \times k$ matrix of loadings corresponding to the latent environmental covariates. The common factors underlying $\Lambda_s$ and $\Lambda_r$ are referred to as the known and latent factors.
The IFAk model in Eq. 3.6 is a variant of the dependent formulation in Chapter 2 without the projection matrix. The exclusion of the projection matrix ensures that the regression is appropriately stabilised, which is particularly important when the known covariates are poorly scaled or when the number of environments is not substantially larger than the number of covariates, i.e. $p \gg q$. The practical implication is that when the known environmental covariates contribute no information to the regression, i.e. $\Lambda_s = 0$, the IFAk model will revert back to the conventional FAk model in Eq. 3.4. Note, however, the variant in Eq. 3.6 does not guarantee a unique solution for $\Lambda_s$ and $\Lambda_r$ (see Section 3.8.2).

It is assumed that:

$$
\begin{bmatrix}
 f \\
 \delta
\end{bmatrix} \sim N \left( \begin{bmatrix}
 0 \\
 L \ 0 \\
 0 \ 0 \Psi
\end{bmatrix}, \left[ S \ I_p \right] \otimes G \right),
$$

(3.7)

where $L = \oplus_{r=1}^{k} I_r$ is a $k \times k$ diagonal variance matrix comprising a separate variance for each common factor which are sorted in decreasing order and $\Psi = \oplus_{j=1}^{p} \psi_j$ is a $p \times p$ diagonal variance matrix comprising a separate variance for each environment.

It then follows that:

$$
\text{var}(u) = \left( \left[ S \ I_p \right] \left[ \begin{array}{cc}
 \Lambda_s \Lambda_s^T & \Lambda_s \Lambda_r^T \\
 \Lambda_r \Lambda_s^T & \Lambda_r \Lambda_r^T
\end{array} \right] \left[ S \ I_p \right]^T + \Psi \right) \otimes G,
$$

(3.8)

where $G_q = \Lambda \Lambda^T + \Psi$ and $\Lambda = S \Lambda_s + \Lambda_r$. Note that the IFAk variance matrix is equivalent to the conventional FAk variance matrix because the $p \times (p+q)$ matrix of basis functions $[S \ I_p]$ has full rank for all values of $q$. The fit of the IFAk model in Eq. 3.6 is therefore exactly the same as the FAk model in Eq. 3.4.

Chapter 2 also extended the ordinary random regression model to include a factor analytic variance matrix for the genotype main effects (intercepts) and slopes. This extension is referred to as the FAR$k$ model, where $k$ denotes the number of common factors. The FAR$k$ model can be viewed as a special IFAk model with latent environmental covariates constrained to be proportional to the unit vector, i.e. $\Lambda \approx 1_p \lambda_1 + S \Lambda_s$.

The FAR$k$ model is given by:

$$
u = \left( [1_p \lambda_1 + S \Lambda_s] \otimes I_d \right) f + \delta,
$$

(3.9)

where $\lambda_1 = \begin{bmatrix} \lambda_{11} & \ldots & \lambda_{1k} \end{bmatrix}$ is a $k$ row-vector of loadings corresponding to the genotype main effects (intercepts) and $\Lambda_s = \begin{bmatrix} \lambda_{s1} & \ldots & \lambda_{sk} \end{bmatrix}$ is a $q \times k$ matrix of loadings corresponding to the genotype slopes.
The FAR\(k\) model can be rewritten in terms of the main effects and slopes as:

\[
\mathbf{u} = (\mathbf{1}_p \otimes \mathbf{u}_1) + (\mathbf{S} \otimes \mathbf{I}_v) \mathbf{u}_s + \mathbf{\delta},
\]

(3.10)

where \(\mathbf{u}_1 = (\mathbf{\lambda}_1 \otimes \mathbf{I}_v) \mathbf{f}\) is a \(v\)-vector of genotype main effects (intercepts) and \(\mathbf{u}_s = (\mathbf{\Lambda}_s \otimes \mathbf{I}_v) \mathbf{f}\) is a \(vq\)-vector of genotype slopes. This highlights the analogy to an ordinary random regression, with the difference that the genotype main effects and slopes are linear combinations of the genotype scores in \(\mathbf{f}\).

It then follows that

\[
\text{var}(\mathbf{u}) = \left(\left[\mathbf{1}_p \; \mathbf{S}\right] \begin{bmatrix} \mathbf{\Lambda}_1 \mathbf{\Lambda}_1^\top & \mathbf{\Lambda}_1 \mathbf{\Lambda}_s^\top \\ \mathbf{\Lambda}_s \mathbf{\Lambda}_1^\top & \mathbf{\Lambda}_s \mathbf{\Lambda}_s^\top \end{bmatrix} \left[\mathbf{1}_p \; \mathbf{S}\right]^\top + \mathbf{\Psi}\right) \otimes \mathbf{G},
\]

(3.11)

where \(\mathbf{G}_a \simeq \mathbf{\Lambda} \mathbf{\Lambda}^\top + \mathbf{\Psi}\) and \(\mathbf{\Lambda} \simeq \mathbf{1}_p \mathbf{\lambda}_1 + \mathbf{S} \mathbf{\Lambda}_s\). Note that the FAR\(k\) variance matrix will be equivalent to the conventional FA\(k\) variance matrix when the \(p \times (q + 1)\) matrix of basis functions \([\mathbf{1}_p \; \mathbf{S}]\) has full rank. The fit of the FAR\(k\) model in Eq. 3.9 is therefore different to the FA\(k\) model in Eq. 3.4 and the IFA\(k\) model in Eq. 3.6 when \([\mathbf{1}_p \; \mathbf{S}]\) does not have full rank.

### 3.4 Selection tools

The following sections extend the selection tools of Smith and Cullis (2018) to produce measures of overall performance, responsiveness and stability that are tailored to one or more TPE. The selection tools are built on a new rotation that places all non-crossover variation into the first common factor and all crossover variation into the higher order factors. The selection tools are then extended for the IFA\(k\) model in Section 3.7 and generalised for the FAR\(k\) model in Appendix C.

#### 3.4.1 Overall performance

Smith and Cullis (2018) discussed the ability of factor analytic models to capture non-crossover GEI within the first common factor. They proposed a measure of overall performance based on this factor given by:

\[
\text{OP} = \bar{\mathbf{\lambda}}_1 \mathbf{f}_1,
\]

(3.12)

with \(\bar{\mathbf{\lambda}}_1 = \mathbf{1}_p^\top \mathbf{\lambda}_1 / p\), where \(\mathbf{\lambda}_1\) is the \(p\)-vector of environmental loadings for the first common factor and \(\mathbf{f}_1\) is the \(v\)-vector of genotype scores. The OP for a genotype is therefore equal to the fitted value of the first factor regression at the mean loading of \(\bar{\mathbf{\lambda}}_1\) (Figure 3.1).
Fig. 3.1 Latent regression plot showing the overall performance measures for two genotypes, which are denoted by open circles. The deviations reflect the common GE effects for the higher order factors, with genotypes and environments distinguished by shape.

When all environmental loadings for the first factor are positive, the regression lines for any pair of genotypes may diverge but never intersect within the range of the data, so they capture heterogeneity of scale associated with perfect (positive) genetic correlation between environments. This is a classical representation of non-crossover GEI, and means that the first factor captures the discriminating ability of the environments with regards to the trait under study (Figure 3.1). It is this feature that Smith and Cullis (2018) exploit in their OP measure. However, when the first factor contains both positive and negative loadings, the regression lines intersect and the first factor captures both non-crossover and crossover GEI, i.e. the regression also captures genotype contrasts between environments (Figure 3.2). There is currently ambiguity in the literature as to whether the OP measure in Eq. 3.12 is appropriate for such cases.

Chapter 2 demonstrated how genotype main effects can be obtained from factor analytic models, regardless of whether intercepts are explicitly fitted. The genotype main effects are given by:

$$u_j = \sum_{r=1}^{k} \tilde{\lambda}_r f_r,$$  \hspace{1cm} (3.13)

with $\tilde{\lambda}_r = \mathbf{1}_p^\top \lambda_r / p$, where $\lambda_r$ is the $p$-vector of environmental loadings for common factor $r$. 
3.4 Selection tools

Fig. 3.2 Latent regression plots showing non-crossover and crossover GEI in the first common factor. The deviations reflect the common GE effects for the higher order factors, with genotypes and environments distinguished by shape.

and \( f_r \) is the \( v \)-vector of genotype scores. The main effect for a genotype is therefore equal to the average performance across all environments based on all factors. This highlights an important difference to the OP measure in Eq. 3.12, which can be viewed as a weighted average across environments based on the first factor only. Importantly, the conventional OP measure can be adjusted so that it becomes equivalent to \( u_1 \) in Eq. 3.13.

The adjusted OP measure is given by:

\[
\text{OP}^* = \text{OP} + \sum_{r=2}^{k} \tilde{\lambda}_r f_r,
\]  

(3.14)

where \( \sum_{r=2}^{k} \tilde{\lambda}_r f_r \) represents the contribution from the higher order factors. The adjusted OP for a genotype is also equal to the fitted value of the first factor regression at the mean loading of \( \bar{\lambda}_1 \), but this regression now includes an intercept for each genotype based on the higher order factors given by \( \sum_{r=2}^{k} \tilde{\lambda}_r f_r \) (Figure 3.3). When there is no contribution from the higher order factors, i.e. \( \tilde{\lambda}_r = 0 \) for \( r \geq 2 \), the intercepts reduce to zero and the OP measures become equivalent. This is an important result that will be utilised below.

The choice of OP measure raises an important question as to which is more appropriate. The answer lies in the assumptions made on the environmental loadings and genotype scores, i.e. the columns of \( \Lambda \) are orthogonal and the variances in \( L \) are sorted in decreasing order.
Fig. 3.3 Latent regression plot showing the adjusted overall performance measures for two genotypes, which are denoted by open circles. The deviations reflect the common GE effects for the higher order factors, with genotypes and environments distinguished by shape.

This is generally achieved using a principal component rotation after estimation, rather than a constraint imposed during estimation (see Smith et al., 2021b, for a complete discussion). Although the rotation typically provides desirable features for modelling GEI in practice, e.g. the first common factor predominately captures non-crossover GEI and the higher order factors predominately capture crossover GEI, it does not guarantee them. These features are a consequence of the data structure, rather than a consequence of an approach that seeks to find a first factor that exclusively captures non-crossover GEI and higher order factors that exclusively capture crossover GEI. A new rotation is developed in the next section which addresses these objectives, and places all non-crossover variation into the first common factor and all crossover variation into the higher order factors. As a result, the fitted value of the first factor regression for a genotype will equal it’s average performance (main effect), meaning that there is no contribution from the higher order factors, i.e. $\hat{\lambda}_r = 0$ for $r \geq 2$. This also suggests that the adjusted OP measure in Eq. 3.14 is more appropriate than the conventional measure in Eq. 3.12, provided the objective is to obtain a measure of average performance across all environments in the MET dataset.
### 3.4.2 A new rotation for disentangling non-crossover and crossover GEI

Constraints are required in factor analytic models to ensure unique solutions for $\Lambda$ and $L$ during estimation. Following Smith et al. (2021b), the upper right elements of $\Lambda$ are set to zero when $k > 1$ and $L$ is set to $I_k$. Let the environmental loadings and genotype scores with these constraints be denoted by $\Lambda^*$ and $f^*$, such that $f^* \sim N(0, I_k \otimes G)$. The loadings and scores are generally rotated after estimation to enable interpretation.

Two rotations are considered below:

1. A principal component rotation which ensures the common factors are orthogonal and sorted in decreasing order.

2. A new rotation which ensures the first common factor exclusively captures non-crossover GEI and the higher order factors are orthogonal, sorted in decreasing order and exclusively capture crossover GEI.

The principal component rotation was initially proposed for the FA$k$ model by Cullis et al. (2010), and has since been widely adopted for the analysis of MET datasets (see, for example, Smith et al., 2015; Ukrainetz et al., 2018). However, it does not guarantee the first common factor exclusively captures non-crossover GEI, nor does it guarantee the OP for a genotype equals it’s average performance. The new rotation overcomes both of these issues. It will be shown how to apply the new rotation and how it can be used to obtain more appropriate measures of responsiveness and stability based on crossover GEI only.

**Principal component rotation**

The objective of the principal component rotation is to improve the interpretability of the common factors. This is achieved by rotating the environmental loadings to a solution where the first common factor captures the largest amount of variation in the GE effects, the second common factor captures the next largest amount and is orthogonal to the first, and so on.

Following Smith et al. (2021b), the principal component rotation is given by:

$$\Lambda = \Lambda^* V L^{-1/2} \quad \text{and} \quad f = (L^{1/2} V^\top \otimes I_v) f^*, \quad (3.15)$$

which is based on the singular value decomposition given by:

$$\Lambda^* = U L^{1/2} V^\top,$$

where $U = [u_1 \ldots u_k]$ is a $p \times k$ orthogonal matrix of left singular vectors, $V = [v_1 \ldots v_k]$ is a
\(k \times k\) orthogonal matrix of right singular vectors and \(L = \oplus_{r=1}^{k} l_r\) is a \(k \times k\) diagonal matrix of singular values sorted in decreasing order, i.e. \(l_1 > \ldots > l_k\).

It then follows that:

\[
\Lambda \equiv U \quad \text{and} \quad f \sim N(0, L \otimes G), \tag{3.16}
\]

such that the common factors are orthogonal, i.e. \(\Lambda^\top \Lambda = I_k\). The conventional OP measure in Eq. 3.12 is then constructed using the rotated loadings and scores in Eq. 3.15.

**The new rotation**

The objective of the new rotation is to construct a first common factor which exclusively captures non-crossover GEI and higher order factors which exclusively capture crossover GEI. This is achieved by transferring all variation associated with the genotype main effects to the first common factor and all other variation to the higher order factors.

The new rotation can be summarised by three key steps:

1. Obtain genotype main effects and common GE effects based on the latent regression.

2. Construct generalised main effects by transferring all non-crossover variation to the genotype main effects, adjust and update the common GE effects.

3. Construct the first factor based on the generalised main effects and higher order factors based on the common GE effects, then rotate to their principal component solution.

Each step is detailed in the following.

1. The genotype main effects and common GE effects are given by:

\[
u_1 = (\Lambda^* \otimes I_v) f^* \quad \text{and} \quad \beta = (\Lambda^* \otimes I_v) f^*,
\]

where \(\Lambda^* = I_p \Lambda^*/p\) is a \(k\) row-vector with elements given by the mean loading for each common factor. Note that the expression for \(u_1\) is an equivalent form of Eq. 3.13.

It then follows that:

\[
\begin{bmatrix}
u_1 \\
\beta
\end{bmatrix} \sim N\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix}
\Lambda^* \Lambda^*^\top & \Lambda^* \Lambda^*^\top \\
\Lambda^* \Lambda^*^\top & \Lambda^* \Lambda^*^\top
\end{bmatrix} \otimes G\right),
\]

such that a covariance is present between the genotype main effects and common GE effects.
The variance matrix can be generalised as (ignoring the relationship matrix):

\[
\begin{bmatrix}
\sigma_1^2 & \Sigma_{12} \\
\Sigma_{21} & \Sigma_{22}
\end{bmatrix} = \begin{bmatrix}
\bar{\Lambda} \bar{\Lambda}^T & \bar{\Lambda} \Lambda^T \\
\Lambda^T \Lambda^T & \Lambda^T \Lambda^T
\end{bmatrix},
\] (3.17)

where \(\sigma_1^2\) is the genotype main effect variance, \(\Sigma_{21} = \Sigma_{12}\) is a \(p\)-vector with elements given by the covariance between the genotype main effects and common GE effects for each environment and \(\Sigma_{22}\) is a \(p \times p\) variance matrix for the common GE effects. It is important to note that the covariances in \(\Sigma_{21}\) (and \(\Sigma_{12}\)) capture variation associated with the genotype main effects, so they capture non-crossover GEI (see Falconer, 1990; Waters et al., 2023, for a similar argument). Any remaining variation in \(\Sigma_{22}\), after adjusting for \(\Sigma_{21}\), will then capture crossover GEI. This result will be utilised in the following step to disentangle non-crossover and crossover GEI in the common GE effects.

2. The genotype main effects and common GE effects can be reconstructed as:

\[
u_1^* = \left(\frac{\Sigma_{21}}{\sigma_1^2} \otimes u_1\right) \quad \text{and} \quad \beta^* = \beta - u_1^*,
\] (3.18)

where \(u_1^*\) is a \(vp\)-vector of generalised main effects and \(\beta^*\) is a \(vp\)-vector of common GE effects adjusted for \(u_1^*\), which now sum to zero for each genotype. This form of the FAk model will be demonstrated using hypothetical examples in Appendix D.

It then follows that:

\[
\begin{bmatrix}
u_1^* \\
\beta^*
\end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix}
\Sigma_{21}\Sigma_{12}/\sigma_1^2 & 0 \\
0 & \Sigma_{22} - \Sigma_{21}\Sigma_{12}/\sigma_1^2
\end{bmatrix} \otimes G\right),
\] (3.19)

such that the generalised main effects and common GE effects are now independent because the covariances in \(\Sigma_{21}\) have been transferred entirely to \(u_1^*\). When all covariances in \(\Sigma_{21}\) are positive, all environments will have some variation explained by the main effects, meaning they have some scaled main effect variance. In this case, the generalised main effects will exclusively capture non-crossover GEI, i.e. \(\text{cor}(u_1^*, u_{1j}^*) = G\) for all pairs of environments (Figure 3.4). When this is not the case, and at least one covariance in \(\Sigma_{21}\) is negative, the generalised main effects will capture some crossover GEI, i.e. \(\text{cor}(u_1^*, u_{1j}^*) = -G\) for at least one pair of environments. Simple constraints can be imposed so that only variation in \(\Sigma_{21}\) due to non-crossover GEI is transferred to the generalised main effects. The constraints set the environment with the minimum negative covariance as a baseline. This environment has no main effect variance so that all variation is attributed to crossover GEI. All other environments are then set relative to the baseline, so they have at least some scaled main
Fig. 3.4 Latent regression plots showing non-crossover and crossover GEI in the generalised main effects. The genotype main effects are denoted by *crosses* and the generalised main effects are denoted by *stars*. The deviations reflect the common GE effects adjusted for the generalised main effects, with genotypes and environments distinguished by *shape*.

Effect variance and therefore at least some variation due to non-crossover GEI (Figure 3.5).

The generalised main effects and common GE effects can then be updated as:

\[ \mathbf{u}_1^* = \left( \mathbf{\Sigma}_{21}^*/\sigma^2_1 \otimes \mathbf{u}_1 \right) \quad \text{and} \quad \mathbf{\beta}^* = \mathbf{\beta} - \mathbf{u}_1^*, \quad (3.20) \]

where \( \mathbf{\Sigma}_{21}^* = \mathbf{\Sigma}_{21} - \sigma_{21}^* \mathbf{1}_p \) is an adjusted covariance term, \( \sigma^2_1 = \sigma_1^2 - \sigma^2_{21} \) is an adjusted main effect variance and \( \sigma^2_{21} \) is the minimum negative covariance in \( \mathbf{\Sigma}_{21} \), i.e. \( \sigma^2_{21} = \min(\mathbf{\Sigma}_{21}, 0) \). Note that setting \( \sigma^2_{21} \) in this manner enables the methods to be generalised for cases where there are no negative covariances, i.e. where \( \sigma^2_{21} = 0 \). Also note that in practice the adjusted baseline covariance in \( \mathbf{\Sigma}_{21}^* \) is set to a sufficiently small value to avoid computational issues.

The generalised main effects can be rewritten as:

\[ \mathbf{u}_1^* = \left( \mathbf{1}_p \otimes \mathbf{u}_1 \right) + \mathbf{\eta}_1, \quad (3.21) \]

where \( \mathbf{u}_1 \) is the \( v \)-vector of genotype main effects and \( \mathbf{\eta}_1 = (\mathbf{\Sigma}_{21} - \sigma^2_1 \mathbf{1}_p)/\sigma_1^2 \otimes \mathbf{u}_1 \) is a \( vp \)-vector of deviations around \( \mathbf{u}_1 \). It can be shown that \( \text{cor}(\mathbf{u}_1, \mathbf{\eta}_{1j}) = \pm \mathbf{G} \) for all environments, meaning that the new rotation constructs intercepts for the \( \mathbf{FAk} \) model which are perfectly correlated with the first factor slopes (Figure 3.6). The sign of the correlation will
Fig. 3.5 Latent regression plots showing the generalised main effects before and after imposing constraints. The genotype main effects are denoted by crosses and the generalised main effects are denoted by stars. The deviations reflect the common GE effects adjusted for the generalised main effects, with genotypes and environments distinguished by shape.

be indicative of how much main effect variance is present, e.g. a positive sign means that environment $j$ has more variation explained by the genotype main effects than $\sigma_1^2$ while a negative sign means that environment $j$ has less variation explained than $\sigma_1^2$. In the extreme, $\eta_{1j} = -u_1$ for the baseline environment because it has no main effect variance.

It then follows that:

$$
\begin{bmatrix}
  u_1^* \\
  \beta_1^*
\end{bmatrix}
\sim N\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix}
  \sigma_1^2 \Sigma_{21} \Sigma_{12}^+ & \Omega_{12} - \sigma_1^2 \Sigma_{21} \Sigma_{12}^+ \\
  \Omega_{21} - \sigma_1^2 \Sigma_{21} \Sigma_{12}^+ & \Omega_{12} - \sigma_1^2 \Sigma_{21} \Sigma_{12}^+ + \sigma_2^2 \Sigma_{21} \Sigma_{12}^+
\end{bmatrix} \right) / (\sigma_1^2)^2 \otimes G,
$$

where $\Sigma_{22}^+ = (\sigma_1^2)^2 \Sigma_{22}$, $\Omega_{21} = \sigma_1^2 \Sigma_{21} \Sigma_{12}^+$, $\Omega_{12} = \Omega_{21}^\top$ and $\Sigma_{12}^+ = \Sigma_{21}^\top$. The generalised main effects and common GE effects are now dependent because the part of the covariances in $\Sigma_{21}$ attributed to crossover GEI remains in the common GE effects. It can be shown that $u_{1j}^* = 0$ for the baseline environment, since there is no main effect variance for this environment. It then follows that $\beta_{1j}^* = \beta_{1j}$ for the baseline environment, since the common GE effects for this environment are entirely attributed to crossover GEI. Note, however, this is not the case when all covariances in $\Sigma_{21}$ are positive, and the variance matrix above reverts back to the form in Eq. 3.19, where the generalised main effects and common GE effects are completely independent and all environments have some scaled main effect variance.
The first common factor is based on the generalised main effects, and is constructed as:

$$\lambda_1 = \Sigma_{21}^* \sqrt{\Sigma_{12}^* \Sigma_{21}}$$

and

$$f_1 = \sqrt{\Sigma_{12}^* \Sigma_{21}} u_1 / \sigma_{12}^2,$$  \hspace{1cm} (3.22)

such that $\lambda_1^\top \lambda_1 = 1$ and $f_1 \sim N(0, l_1 G)$, where $l_1 = \Sigma_{12}^* \Sigma_{21}^* \sigma_{12}^2 / (\sigma_{12}^2)^2$. The first factor can be written in terms of the genotype main effects (intercepts) in a similar manner to Eq. 3.21, by centring and scaling $\lambda_1$ to unit length.

The higher order factors are based on the common GE effects adjusted for the generalised main effects, and are constructed as:

$$\Lambda_{22} = [\Lambda^* - \Sigma_{21}^* \bar{\Lambda}^* / \sigma_{12}^2] V_{22} L_{22}^{-1/2}$$

and

$$f_{22} = (L_{22}^{1/2} V_{22}^\top \otimes I_r) f^*,$$  \hspace{1cm} (3.23)

which is based on the singular value decomposition given by:

$$[\Lambda^* - \Sigma_{21}^* \bar{\Lambda}^* / \sigma_{12}^2] = U_{22} L_{22}^{1/2} V_{22},$$

where $U_{22} = [u_2 \ldots u_{k+1}]$ is a $p \times k$ orthogonal matrix of left singular vectors, $V_{22} = [v_2 \ldots v_{k+1}]$ is a $k \times k$ orthogonal matrix of right singular vectors and $L_{22} = \otimes_{r=2}^{k+1} l_r$ is
a $k \times k$ diagonal matrix of singular values sorted in decreasing order, i.e. $l_2 > \ldots > l_{k+1}$.

It then follows that:

$$\Lambda_{22} \equiv U_{22} \quad \text{and} \quad f_{22} \sim N(0, L_{22} \otimes G),$$

such that the higher order factors are orthogonal and sum to zero, i.e. $\Lambda_{22}^T \Lambda_{22} = I_k$ and $1_p^T \Lambda_{22} = 0$. This produces desirable features for modelling and interpreting GEI in practice.

The common GE effects are then constructed by multiplying the rotated environmental loadings and genotype scores in Eqs. 3.22 and 3.23, which gives:

$$\beta = (\Lambda \otimes I_v) f,$$

where $\Lambda = [\lambda_1 \Lambda_{22}]$ is a $p \times (k+1)$ matrix and $f = (f_1^T, f_{22}^T)^T$ is a $(k+1)$-vector. The new rotation therefore constructs $k+1$ factors from the original $k$ factors in order to completely disentangle non-crossover and crossover GEI. Note, however, this is not the case when all covariances in $\Sigma_{21}$ are positive and all variation in $\Lambda_{22}$ will be explained by the first $k-1$ factors, i.e. $l_{k+1} = 0$. For simplicity, let $k$ denote the appropriate number of factors hereafter.

It also follows that:

$$f \sim N(0, L \otimes G) \quad \text{and} \quad L = \begin{bmatrix} l_1 & L_{12} \\ L_{21} & L_{22} \end{bmatrix},$$

where $L_{21} = \sqrt{\Sigma_{22}^\times \Sigma_{21}^\times} V_{22}^\top \tilde{\Lambda}^\top / \sigma_1^2$ and $L_{12} = L_{21}^\top$. It can be shown that $L_{21} = 0$ when all covariances in $\Sigma_{21}$ are positive, so that $L$ will revert to a diagonal matrix.

Lastly, the OP measure based on the new rotation is obtained as:

$$\text{OP} = \tilde{\lambda}_1 f_1,$$

which has the same form as the conventional measure in Eq. 3.12, but is equal to the genotype main effects in Eq. 3.13 because there is no contribution from the higher order factors, i.e. $\tilde{\lambda}_r = 0$ for $r \geq 2$. The new rotation is particularly favourable over the principal component rotation in the presence of high levels of crossover GEI, i.e. when the first principal component contains both positive and negative loadings (Figure 3.7).

Measures of responsiveness and stability will also be built on the new rotation in the following sections. As a result, the measures will be appropriately adjusted for non-crossover GEI, i.e. they will only measure the response of genotypes with regards to crossover GEI. The new rotation is generalised for any random regression type model in Appendix C and further demonstrated using hypothetical examples in Appendix D.
Fig. 3.7 Latent regression plots showing the principal component rotation and the new rotation developed in this chapter. The overall performance measures are denoted by open circles, which are equivalent to the y-intercepts in the top right plot. The deviations in the top and middle plots reflect the common GE effects adjusted for the proceeding factor/s, with genotypes and environments distinguished by shape.
3.4 Selection tools

3.4.3 Responsiveness

Smith and Cullis (2018) proposed a measure of responsiveness for each factor given by:

\[
\text{RESP}_r = (\bar{\lambda}_r^+ - \bar{\lambda}_r^-) f_r, \tag{3.26}
\]

where \(\bar{\lambda}_r^+\) is the mean of the positive environmental loadings for common factor \(r\) and \(\bar{\lambda}_r^-\) is the mean of the negative loadings. The responsiveness of a genotype is therefore equal to the fitted value of the regression at the mean contrast of \(\bar{\lambda}_r^+ - \bar{\lambda}_r^-\), i.e. the fitted value at the mean positive loading minus the fitted value at the mean negative loading (Figure 3.8). Note that the responsiveness for the first factor reduces to the OP measure when all loadings are positive, which is always the case for the new rotation developed in the previous section.

When the environmental loadings for a particular (higher order) factor are a mixture of positive and negative values, the regression lines for any pair of genotypes will intersect so they will capture variation due to lack of perfect (positive) genetic correlation between environments. This is a classical representation of crossover GEI, and means that the factor captures genotype contrasts between environments. It is this feature that Smith and Cullis (2018) exploit in their measure of responsiveness (and stability). When the mean of the positive and negative loadings are very different, \(\bar{\lambda}_r^+ - \bar{\lambda}_r^- \gg 0\), the regression lines will

![Latent regression plots showing the responsiveness measures for factors two and three, which are given by the vertical distance between the upward and downward triangles. The deviations in the left plot reflect the common GE effects for the third common factor, with genotypes and environments distinguished by shape.](image)
mostly capture variation due to heterogeneity of scale rather than lack of genetic correlation
(Figure 3.8, also see Waters et al., 2023, for further details). When the mean of the positive
and negative loadings are equal, $\bar{\lambda}_+^+ - \bar{\lambda}_-^-$ = 0, the regression lines will mostly capture
variation due to lack of genetic correlation rather than heterogeneity of scale. In this case,
the responsiveness measure in Eq. 3.26 for all genotypes will equal zero, i.e. $\text{RESP}_r = 0$.
Developing an appropriate responsiveness measure is the topic of active research.

### 3.4.4 Stability

Smith and Cullis (2018) proposed a measure of stability based on the higher order factors
given by:

$$\text{RMSD} = \sqrt{\text{diag}(EE^\top)} / p,$$

where $E = FA^\top - f_1 \lambda^\top_1$ is a $v \times p$ matrix of common GE effects adjusted for the first factor
and $F = [f_1 \ldots f_k]$ is a $v \times k$ matrix of genotype scores. The RMSD for a genotype is therefore
equal to the root mean square of the deviations around the first factor regression (Figure 3.9).
Note that the RMSD measure can be written succinctly as
$$\text{RMSD} = \sqrt{\sum_{i=2}^k f_i^2} / p,$$
since the higher order factors are orthogonal for the principal component rotation and the new rotation.

![Fig. 3.9 Latent regression plot showing the stability measures for two genotypes, which are
given by the root mean square of the deviations around the regression lines. The deviations
reflect the common GE effects for the higher order factors, with different genotypes and
environments distinguished by shape.](image-url)
3.4 Selection tools

The RMSD measure can be adjusted so that it complements the OP* measure in Eq. 3.14. The adjusted RMSD measure is given by:

\[ \text{RMSD}^* = \sqrt{\text{diag}(E^*E^{*\top})} / p, \] (3.28)

where \( E^* = E - \sum_{r=2}^{k} \lambda_r f_r I_p \) is a \( v \times p \) matrix of common GE effects adjusted for the mean contribution of the higher order factors. The adjusted RMSD for a genotype is also equal to the root mean square of the deviations around the first factor regression, but this regression now includes an intercept for each genotype based on the higher order factors given by \( \sum_{r=2}^{k} \lambda_r f_r \) (Figure 3.3). When there is no contribution from the higher order factors, i.e. \( \lambda_r = 0 \) for \( r \geq 2 \), the intercepts reduce to zero and the RMSD measures become equivalent. Note that this is always the case for the new rotation because the common GE effects for the higher order factors sum to zero for each genotype (see Eq. 3.18).

The OP and RMSD measures can be combined into a selection index, which is often achieved by penalising OP by RMSD. However, this approach is flawed because RMSD is based on the deviations above and below the first factor regression, so that the selection index penalises favourable and unfavourable adaptation. A new stability measure is developed below which separates these two components, so that favourable adaptation is not penalised and an appropriate selection index can be constructed.

**Directional stability**

The conventional RMSD measure can be partitioned according to the deviations above and below the first factor regression. The positive and negative stability measures are given by:

\[ \text{RMSD}^+ = \sqrt{\text{diag}(E^+E^{+\top})} / p \quad \text{and} \quad \text{RMSD}^- = \sqrt{\text{diag}(E^-E^{-\top})} / p, \] (3.29)

where \( E^+ \) is a \( v \times p \) matrix given by the positive elements in \( E \) and \( E^- \) is a \( v \times p \) matrix given by the negative elements. All other elements in \( E^+ \) and \( E^- \) are set to zero, such that \( E^+ \geq 0 \) and \( E^- \leq 0 \). The positive and negative RMSD measures are therefore equal to the root mean square of the deviations above or below the first factor regression, respectively (Figure 3.10).

The measure of directional stability is then given by:

\[ d\text{RMSD} = \text{RMSD}^+ - \text{RMSD}^-, \] (3.30)

which is equal to the root mean square of the deviations above the first factor regression minus those below the regression (Figure 3.10). When the common GE effects for the higher order
Fig. 3.10 Latent regression plot showing the directional stability measures for two genotypes, which are given by the difference between the root mean square of the deviations above and below the regression lines. The deviations reflect the common GE effects for the higher order factors, with different genotypes and environments distinguished by shape.

Factors sum to zero for each genotype, the deviations around the first factor regression will sum to zero but not necessarily the root mean square of the deviations. The dRMSD measure therefore summarises the variation above and below the first factor regression, with positive values indicating the potential for specific adaptation. The conventional measure does not have this feature. It summarises all variation around the first factor regression, regardless of its form. In this case, a genotype will appear stable if it has little or no deviations around the regression or unstable (variable) if it has large deviations. This approach is particularly inefficient when combining OP and RMSD within a selection index because it will equally penalise favourable and unfavourable adaptation. It is also inefficient during the early stages of a breeding programme because those genotypes with the potential for specific adaptation will be removed in favour of those which have little variability around their average. Ideally, a mixture of genotypes should be selected, i.e. those which are high performing and stable for their potential for broad adaptation as well as those which are high performing and variable, or more specifically responsive, for their potential for specific adaptation. The two measures are therefore complementary, and should be tailored depending on the breeding objectives.
3.5 Measures of variance explained

An important supplement to the selection tools are measures of variance explained. There are two measures of interest for the factor analytic models:

1. The proportion of variance explained by each common factor and overall.

2. The proportion of non-crossover and crossover variance.

The measures will be derived for the new rotation, but links to the principal component rotation will be highlighted in text.

1. The proportion of variance explained by common factor \( r \) is given by:

\[
v_r = \frac{l_r}{\text{tr}(G_a)},
\]

where \( l_r \) is diagonal element \( r \) of \( L \) and \( \lambda_r^\top \lambda_r = 1 \). This measure has the same form for the principal component rotation and the new rotation. However, when at least one covariance in \( \Sigma_{21} \) from Eq. 3.17 is negative, a covariance is induced in the new rotation between the first factor and the higher order factors (see Eq. 3.24).

The proportion of covariance between factor 1 and factor \( r \) is then given by:

\[
v_{1r} = \frac{2l_{1r} \lambda_1^\top \lambda_r}{\text{tr}(G_a)},
\]

where \( l_{1r} \) is element \( r \) of the first row of \( L \) and \( \lambda_1^\top \lambda_r \neq 1 \) for \( r \neq 1 \). When all covariances in \( \Sigma_{21} \) are positive, the first common factor will be independent of the higher order factors and the covariance \( l_{1r} \) will reduce to zero, i.e. \( v_{1r} = 0 \) (also see Eq. 3.24).

The overall proportion of variance explained is therefore given by:

\[
\bar{v} = \sum_{r=1}^{k} v_r + v_{1r},
\]

which is equivalent to the variance explained by all common factors. A similar measure can be obtained for the principal component rotation, but note that \( v_{1r} = 0 \) for \( r \geq 2 \).

2. The proportion of non-crossover variance is equal to the variance explained by the first common factor, which is given by:

\[
v_n = \frac{l_1}{\text{tr}(G_a)},
\]

where \( l_1 = \Sigma_{12} \Sigma_{21} \sigma_1^2 / (\sigma_1^2)^2 \) from Eq. 3.22. It can be shown that \( \Sigma_{12} \Sigma_{21} / (\sigma_1^2)^2 \geq 1 \), which
means that the proportion of non-crossover variance is greater than or equal to the proportion of main effect variance. This is an expected result because the non-crossover variance captures all heterogeneity of scale associated with the generalised main effects, i.e. all variation associated with perfect (positive) genetic correlation between environments. The crossover variance will then capture all remaining variation including heterogeneity of scale that is not perfectly (positively) correlated with the main effects. This includes all variation in the higher order factors, as well as any variation remaining in the covariance term between the first and higher order factors.

The proportion of crossover variance is therefore given by:

\[
v_c = \sum_{r=2}^{k} \frac{l_r + 2l_{1r}^\top \lambda_r}{\text{tr}(G_a)},
\]

such that \(v_c = 1 - v_n\). An equivalent measure for the principal component rotation is not readily available because it does not exclusively capture non-crossover GEI in the first factor.

3.6 Multiple target populations of environments

Smith et al. (2021b) discuss the appropriateness of selecting for broad adaptation in the presence of high levels of crossover GEI, particularly when the first common factor captures both non-crossover and crossover GEI. The new rotation overcomes this issue because it places all non-crossover variation into the first factor and all crossover variation into the higher order factors. Selection for broad adaptation can then be achieved using the OP and RMSD measures presented in the previous sections, provided there is sufficient main effect variance to make genetic gain. This approach is particularly useful in the early stages of a breeding programme where the material is not yet targeted toward specific environments.

Selection for broad adaptation becomes less efficient later in the breeding programme, particularly when breeder’s target multiple TPE. The selection tools can be extended to provide tailored measures of overall performance, responsiveness and stability for multiple TPE. The extension requires a separate rotation for each TPE, which can be achieved using a principal component rotation or the new rotation developed in Section 3.4.2. Let the rotated environmental loadings and genotype scores for TPE \(i\) be given by \(\Lambda_i\) and \(f_i\), respectively.

The measure of overall performance for TPE \(i\) is given by:

\[
\text{OP}_i = \bar{\lambda}_{i1} f_{i1},
\]

where \(\bar{\lambda}_{i1}\) is the mean environmental loading for the first common factor and \(f_{i1}\) is the \(v\)-vector
of genotype scores. The adjusted OP measure in Eq. 3.14 is a straightforward extension.

The measure of responsiveness for TPE $i$ is given by:

$$\text{RESP}_{ir} = \left( \bar{\lambda}_{ir}^+ - \bar{\lambda}_{ir}^- \right) f_{ir},$$

where $\bar{\lambda}_{ir}^+$ is the mean of the positive environmental loadings for factor $r$, $\bar{\lambda}_{ir}^-$ is the mean of the negative environmental loadings and $f_{ir}$ is the $v$-vector of genotype scores.

The measure of stability for TPE $i$ is given by:

$$\text{RMSD}_i = \sqrt{\text{diag}(E_i E_i^\top)/p_i},$$

where $E_i = F_i \Lambda_i^\top - f_{i1} \lambda_1^\top$ is a $v \times p_i$ matrix of common GE effects adjusted for the first factor and $F_i = [f_{i1} \ldots f_{ik}]$ is a $v \times k$ matrix of genotype scores. The adjusted RMSD measure in Eq. 3.28 is also a straightforward extension, but note that by design the adjusted measures are exactly the same as the conventional measures for the new rotation.

Lastly, the measure of directional stability for TPE $i$ is given by:

$$\text{dRMSD}_i = \text{RMSD}^+_i - \text{RMSD}^-_i,$$

where $\text{RMSD}^+_i$ is a $vp_i$-vector with non-zero elements given by the root mean square of the positive deviations above the first factor regression and $\text{RMSD}^-_i$ is a $vp_i$-vector with non-zero elements given by the root mean square of the negative deviations.

### 3.7 Extension to the integrated factor analytic model

The main limitation of the conventional FA$k$ model is that the common factors are latent, so they do not have direct biological interpretation. The IFA$k$ model in Eq. 3.6 overcomes this limitation because it integrates known environmental covariates into the common factors, i.e. $\Lambda = S \Lambda_s + \Lambda_r$. This section extends the selection tools for the IFA$k$ model to provide predictable and interpretable measures of overall performance, responsiveness and stability. Forward predictions can be obtained by updating the environmental covariates in $S$ for new or untested environments, which results in updated measures for all environments together or each TPE separately. Interpretation is then achieved using measures of variance explained in terms of the known environmental covariates. The measures summarise the non-crossover and crossover variation in the GE effects with regards to potential biological drivers of GEI.
There are three measures of interest for the IFAk model:

1. The proportion of variance in the first common factor explained by each known environmental covariate.
2. The proportion of variance in the higher order factors explained by each known environmental covariate.
3. The overall proportion of variance in the higher order factors explained by each known environmental covariate.

The measures will be derived for the new rotation developed in Section 3.4.2, but links to the principal component rotation will be highlighted in text. Note that the measures include a covariance term between the known and latent factors because they are not orthogonal in the variant of the IFAk model considered in this chapter, i.e. $S^\top \Lambda = 0$. Also note that the known covariates are assumed to be centred and scaled to unit length, i.e. $s_h^\top s_h = 1$.

1. The proportion of variance in the first common factor explained by known covariate $h$ is:

$$ v_{1h} = (s_h^\top [S\lambda_{s1} + \lambda_{r1}])^2, \quad (3.35) $$

such that the variance explained by all known covariates is $v_1 = \lambda_{s1}^\top S^\top [S\lambda_{s1} + \lambda_{r1}]$. These measures have the same form for the principal component rotation and the new rotation, and are used to supplement the OP measures in Eqs. 3.12 and 3.25.

2. The proportion of variance in common factor $r$ explained by known covariate $h$ is:

$$ v_{rh} = (s_h^\top [S\lambda_{sr} + \lambda_{rr}])^2, $$

such that the variance explained by all known covariates is $v_r = \lambda_{sr}^\top S^\top [S\lambda_{sr} + \lambda_{rr}]$. These measures also have the same form for the principal component rotation and the new rotation. However, when at least one covariance in $\Sigma_{21}$ from Eq. 3.17 is negative, a covariance is induced in the new rotation between the first factor and the higher order factors (see Eq. 3.24).

The proportion of covariance between factor 1 and factor $r$ explained by known covariate $h$ is then given by:

$$ v_{1rh} = \frac{(\lambda_{s1h}^\top s_h^\top [S\lambda_{sr} + \lambda_{rr}])^2 + (\lambda_{srh}^\top s_h^\top [S\lambda_{s1} + \lambda_{r1}])^2}{2(S\lambda_{s1} + \lambda_{r1})^\top [S\lambda_{sr} + \lambda_{rr}]}, $$

such that the covariance explained by all known covariates is $v_{1r} = (\lambda_{s1}^\top S^\top [S\lambda_{sr} + \lambda_{rr}] + \lambda_{sr}^\top S^\top [S\lambda_{s1} + \lambda_{r1}])/2[S\lambda_{s1} + \lambda_{r1}]^\top [S\lambda_{sr} + \lambda_{rr}].$
3.8 Application to example dataset

The overall proportion of variance associated with factor $r$ explained by known covariate $h$ is therefore given by:

\[
\bar{v}_{rh} = \frac{v_{rh} l_r + 2 v_{1rh} l_{1r} \left[ S\lambda_{s1} + \lambda_{r1} \right]^T \left[ S\lambda_{sr} + \lambda_{rr} \right]}{l_r + 2 l_{1r} \left[ S\lambda_{s1} + \lambda_{r1} \right]^T \left[ S\lambda_{sr} + \lambda_{rr} \right]},
\]

(3.36)

such that the overall variance explained by all known covariates is $\bar{v}_p$, which is obtained by replacing $v_{rh}$ and $v_{1rh}$. When all covariances in $\Sigma_{21}$ are positive, the first common factor will be independent of the higher order factors and $l_{1r}$ and $v_{1rh}$ will reduce to zero (also see Eq. 3.24). The measures above supplement the responsiveness measure in Eq. 3.26. A similar responsiveness measure is developed for the FAR$k$ model in Appendix C which quantifies the response of genotypes to individual known covariates.

3. The overall proportion of variance in the higher order factors explained by known covariate $h$ is given by:

\[
\bar{v}_h = \frac{\sum_{r=2}^{k} v_{rh} l_r + 2 v_{1rh} l_{1r} \left[ S\lambda_{s1} + \lambda_{r1} \right]^T \left[ S\lambda_{sr} + \lambda_{rr} \right]}{\sum_{r=2}^{k} l_r + 2 l_{1r} \left[ S\lambda_{s1} + \lambda_{r1} \right]^T \left[ S\lambda_{sr} + \lambda_{rr} \right]},
\]

(3.37)

such that the overall variance explained by all known covariates is $\bar{v}_r$. This is obtained by replacing $v_{rh}$ and $v_{1rh}$. Similar measures can be obtained for the principal component rotation by setting $l_{1r} = 0$ for $r \geq 2$. The measures above are used to supplement the RMSD measures in Eqs. 3.27 to 3.30.

### 3.8 Application to example dataset

#### 3.8.1 Data description

The example dataset is the Bayer CropScience cotton breeding dataset previously used in Chapter 2. A total of 208 genotypes were evaluated in 24 environments across eight states in Southeast, MidSouth and Texas, USA. Environmental covariate data were generated for all $p = 24$ environments using $q = 18$ known covariates, with 11 covariates derived from daily weather data and 5 covariates derived from daily soil data. Marker data were generated for $v = 204$ of the 208 genotypes using a high confidence set of 36,009 single-nucleotide polymorphisms, coded as either $-1$, $0$ or $1$ for the homozygous minor, heterozygous or homozygous major allele. Markers were filtered using the default settings in pedicure (Butler, 2019), with minor allele frequency $> 0.002\%$ and missing values $< 0.998\%$. Missing values were then imputed using the $k$-nearest neighbour approach of Troyanskaya et al. (2001), with $k = 10$. A total of $r = 24,265$ markers were retained using this criteria.
3.8.2 Model estimation

All models considered in this chapter were fitted using ASReml-R (Butler et al., 2017), with known environmental covariates included using the \textit{mbf} argument. ASReml-R obtains REML estimates of the factor analytic variance parameters using an extension of the sparse formulation of the average information algorithm (Thompson et al., 2003). The extension considered in this chapter fits the common GE effects and regression residuals as separate components, each with their own variance structure. Let the REML estimates of the key variance parameters in the FA$k$ model be given by $\hat{\Lambda}$ and $\hat{\Psi}$, and let the empirical BLUPs of the key random effects be given by $\tilde{f}$ and $\tilde{\delta}$. The environmental loadings in the IFA$k$ model can be obtained from the FA$k$ model as:

$$
\begin{bmatrix}
\hat{\Lambda}_s \\
\hat{\Lambda}_r
\end{bmatrix} = (B^\top B)^{-1} B^\top \hat{\Lambda},
$$

where $(B^\top B)^{-1}$ is any generalised inverse of $B^\top B$ and $B = [S \ I_p]$. The solutions for $\hat{\Lambda}_s$ and $\hat{\Lambda}_r$ are not unique when back-solving from $\hat{\Lambda}$. Alternatively, the loadings can be estimated from fitting the IFA$k$ model directly, which requires constraints during estimation (see Chapter 2).

3.8.3 Results

There are four important results from applying the selection tools to the example dataset:

1. The overall performance and stability measures obtained from the principal component rotation are mostly similar to those from the new rotation developed in this chapter.

2. The new directional stability measure is more informative than the conventional measure for identifying genotypes with specific adaptation.

3. The extension of the selection tools to multiple TPE demonstrate several genotypes with specific adaptation in the different growing regions.

4. The extension of the selection tools to the IFA$k$ model demonstrate several notable environmental covariates within the different growing regions.

Each result is detailed in the following.

The overall performance and stability measures obtained from the principal component rotation are mostly similar to those from the new rotation, with correlation of 0.99 and 0.98, respectively (Figure 3.11). The similarities occur because the proportion of non-crossover variance is high for the example dataset, with $v_n = 0.57$ captured by the latent regression.
Fig. 3.11 Comparison of the overall performance and stability measures obtained from the principal component rotation and the new rotation developed in this chapter. Genotypes G1, G2 and G3 are referred to in text.

However, there will be differences between the two rotations when the proportion of non-crossover variance is low. In Section 3.4.2, it was shown that the new rotation exclusively captures all non-crossover GEI in the first common factor and all crossover GEI in the higher order factors. The principal component rotation does not guarantee this feature. It constructs common factors that are orthogonal and capture a decreasing amount of variation in the GE effects. As a result, the first principal component predominately captures non-crossover GEI for the example dataset because it is the dominant source of variation in the GE effects. The principal component rotation will therefore provide a reasonable approximation to the new rotation when this is the case, but not when crossover GEI is the dominant form of variation, i.e. when the first principal component contains both positive and negative loadings.

There are notable differences between the stability measures for particular genotypes (Figure 3.11). This is highlighted by G1, G2 and G3 which have a higher RMSD for the new rotation than for the principal component rotation, with bias of 0.02 – 0.03. The practical implication is that the principal component rotation transfers some of the crossover variation from the higher order factors to the first common factor. As a result, the three genotypes appear more stable for the principal component rotation than for the new rotation. This is further emphasised by the OP for these genotypes, which are slightly shrunken for the principal component rotation because they capture some of the crossover variation. The new rotation overcomes these issues because it ensures the resulting OP and RMSD measures
Fig. 3.12 Latent regression plot showing the overall performance measures for genotypes G3 and G4, which are denoted by open circles. The stability measures are given by the root mean square of the deviations around the regression lines. The deviations reflect the common GE effects for the higher order factors, with genotypes and regions distinguished by shape.

are based exclusively on non-crossover and crossover GEI, respectively. The new rotation is demonstrated further in the following.

The overall performance and stability measures are based on the first factor regression (Figure 3.12). The OP measure is equal to the fitted value at the mean environmental loading. In Section 3.4.2, it was shown that the OP measure obtained from the new rotation is equivalent to the genotype main effects. This is because the new rotation transfers all variation associated with the genotype main effects to the first common factor, i.e. only non-crossover variation is transferred. It was also shown that genotype intercepts can be obtained by centring the environmental loadings for the first factor. As a result, the OP measure for a genotype becomes equivalent to the $y$-intercept of the regression line. The RMSD measure is then equal to root mean square of the deviations around the regression line, which capture all variation in the common GE effects for the higher order factors, i.e. they only capture crossover variation. This highlights the appealing features of the new rotation developed in this chapter.

The deviations for Texas are consistently higher than those for the Southeast and Midsouth growing regions (Figure 3.12). This result is utilised by the new directional stability measure, which is equal to the root mean square of the deviations above the regression line minus
3.8 Application to example dataset

those below the regression line. The dRMSD measure therefore summarises the variation above and below the regression line so that it can be utilised for selection, with positive values of dRMSD indicating the potential for specific adaptation. The conventional stability measure does not have this feature. For G3 and G4, variation above the regression line corresponds to specific adaptation in Texas, whereas variation below the regression line corresponds to poor adaptation in the Southeast and Midsouth. It was shown in Chapter 2 that the genotype contrasts between the growing regions appear in the second common factor for the example dataset, so that a responsiveness measure based on this factor would summarise similar adaptation patterns. In fact, the dRMSD and RESP$_2$ measures are analogous for any dataset when only two factors are fitted. However, selection using responsiveness will not be straightforward when the genotype contrasts are small or spread over multiple factors. The appealing feature of the dRMSD measure is that it captures all contrasts across all higher order factors, so it captures all the nuances in the GE effects regardless of their structure and magnitude. This is particularly important for cases where the TPE are not well-defined or in the early stages of a breeding programme where the material is not yet targeted toward specific environments.

The new directional stability measure is more informative than the conventional measure for identifying genotypes with specific adaptation (Figure 3.13). This is highlighted by G3 which appears unstable for the conventional RMSD measure but has a large positive dRMSD measure and is therefore a potential candidate for specific adaptation. The important

Fig. 3.13 Overall performance, stability and directional stability measures for all 204 genotypes in the example dataset. Genotypes G3 and G4 are referred to in text.
Fig. 3.14 Comparison of the stability and directional stability measures for all 204 genotypes in the example dataset. Genotypes G3, G4 and G5 are referred to in text.

The difference here is that the dRMSD measure summarises the variation above and below the first factor regression separately, whereas the conventional measure summarises all variation together regardless of its form. As a result, the dRMSD measure can disentangle the favourable and unfavourable adaptation patterns for each genotype, whereas the conventional measure cannot. The only limitation of the dRMSD measure is that when the variation above and below the first factor regression is equal, it will reduce to zero (see, for example, G5 in Figure 3.14). However, this does not mean the genotype has no potential for specific adaptation (nor is it necessarily stable). In such cases, it may be more efficient to examine the separate positive and negative RMSD measures in Eq. 3.29, or pair multiple stability measures together to examine variability and adaptation simultaneously.

The extension of the selection tools to multiple TPE demonstrates several genotypes with specific adaptation in the different growing regions (Figure 3.15). This is highlighted by G4 which performs exceptionally well in Texas and to a lesser extent in the Midsouth, but not in the Southeast. G6 is the best performing genotype in the Southeast but is unstable in the Midsouth and is average in Texas. In contrast, G3 is very poor in the Southeast and Midsouth but is average in Texas. The adaptation patterns of G3 were previously summarised by the dRMSD measure (Figure 3.13). It identified G3 as a potential candidate for specific adaptation because of its favourable performance in Texas compared to the Southeast and
Midsouth. This example demonstrates the effectiveness of the dRMSD measure, and when paired with the OP measure can identify genotypes with specific adaptation of high interest.

The proportion of non-crossover variance is much higher within growing regions than between growing regions, with $v_n = 0.85 - 0.93$ captured by the latent regression compared to 0.57. This indicates that substantially more of the genotype contrasts and re-rankings occur between regions than within regions. The same patterns are observed in the OP and RMSD measures, which summarise the non-crossover and crossover GEI patterns from the perspective of the genotypes. The OP measures have a larger magnitude within regions than between regions, while the RMSD measures have a smaller magnitude (Figures 3.13 and 3.15). This is because a large part of the crossover variation between regions translates to non-crossover variation within regions, i.e. changes in rank translate to changes in scale.

The overall proportion of variance explained is higher for the Midsouth than the Southeast and Texas, with $\bar{v} = 0.85$ compared to 0.67 – 0.70. This indicates that more of the variation in the GE effects is captured by the latent regression for the Midsouth than for the Southeast and Texas. The same patterns are again observed in the OP and RMSD measures, which have a higher magnitude for the Midsouth than for the Southeast and Texas (Figure 3.15). There may be cases where individual growing regions are not well explained by the regression, e.g. because it is driven by other more dominant regions or because the regions are not well correlated. In such cases, it may be more efficient to run separate models on the different growing regions, provided that selection between growing regions is not of interest.

The extension of the selection tools to the IFA$k$ model demonstrate several notable environmental covariates within the different growing regions (Figure 3.15). This is highlighted by maximum precipitation which explains a large proportion of the variance in the higher order factors for the Southeast and Midsouth, and to a lesser extent in Texas. Average soil moisture 3 explains the most variation in Texas and a large proportion in the Midsouth, but not in the Southeast. The remaining highlighted covariates are average windspeed and latitude, which explain a large proportion of the variance in the Southeast and Midsouth, respectively. This example demonstrates how direct biological interpretation can be ascribed to the common factors in the IFA$k$ model.

The proportion of crossover variance explained by all known environmental covariates is much higher than the proportion of non-crossover variance, with $0.32 - 0.52$ captured in the latent regression compared to 0.02 – 0.05. The known covariates therefore predominately capture changes in rank and genotype contrasts between environments in the example dataset, whereas the latent covariates predominately capture changes in scale and the discriminating ability of the environments. This example provides key insights into the nature and extent of non-crossover and crossover GEI in the cotton growing regions of USA.
Fig. 3.15 Overall performance and stability measures for the Southeast, Midsouth and Texas growing regions in the example dataset. The measures are supplemented with the proportion of variance explained by the 18 known environmental covariates. Genotypes G3, G4 and G6 are referred to in text.
3.9 Concluding remarks

This chapter extended a recent set of selection tools for the IFAk model in Chapter 2 to provide measures of overall performance, responsiveness and stability for one or more TPE. The selection tools were built on a new rotation that places all non-crossover variation into the first common factor and all crossover variation into the higher order factors. Previously, a principal component rotation was used that does not guarantee this feature. The rotation developed in this chapter provides a solid theoretical basis for the selection tools which produces OP and RMSD measures that summarise the response of genotypes exclusively in terms of non-crossover and crossover GEI, respectively.

The new rotation avoids the ambiguity of the principal component rotation when the first principal component contains both positive and negative loadings, i.e. when it captures both non-crossover and crossover GEI. This generally occurs when crossover variation is the dominant form of variation in the GE effects, rather than non-crossover variation. Importantly, the new rotation provides a formal way to (i) quantify the amount of non-crossover and crossover variation in the GE effects and (ii) identify when crossover variation becomes the dominant form, i.e. when at least one covariance in $\Sigma_{12}$ from Eq. 3.17 is negative. In this case, the variation associated with the main effects is attributed to both non-crossover and crossover GEI, which precludes disentangling GEI into two completely independent sources.

The chapter also introduced a directional stability measure which partitions the conventional measure into components representing favourable and unfavourable adaptation. It was shown that the dRMSD measure is more informative than the conventional measure for identifying genotypes with specific adaptation patterns. The dRMSD measure is also more efficient than the responsiveness measure, particularly when the genotype contrasts between environments are small or spread over multiple factors.

The methods and concepts developed in this chapter are becoming increasingly important when selecting in the presence of GEI, particularly with high levels of crossover interaction. Generalisations of the new rotation and resulting selection tools were also developed, which provide plant breeding programmes with a wide-array of statistical models for GEI.

Areas of active research include pairing multiple stability and responsiveness measures and a linear directional stability measure, i.e. average responsiveness.
Chapter 4

Association studies I

4.1 Prelude

This chapter contains the unpublished manuscript *Fast exact methods for genome-wide association studies* by Tolhurst and Gorjanc (2024). The chapter develops fast methods for GWAS which produce all required test statistics from the fit of a single linear mixed model, rather than fitting a very large number of models for all markers of interest.

A fundamental relationship is derived between two alternative approaches for GWAS:

1. The marker-in approach which fits the marker effect of interest as fixed and random.
2. The marker-out approach which fits the marker effect of interest as fixed but not random.

It will be shown that double fitting the marker effect as fixed *and* random causes the two effects to be statistically aliased, which results in shrunken test statistics. The difference between the marker-in and marker-out test statistics is then shown to be a matter of scaling, which is particularly important when the number of markers is low. A simple fix is provided for current software packages employing the marker-in approach.

The methods are generalised for any set of fixed and random effects, including modelling population structure, complex genetic and residual covariance structures and spatial modelling. The methods are also extended for testing a set of markers at a time and testing a single marker while excluding a set from the polygenic term. The set is general and may represent a region, linkage group or chromosome containing the marker/s of interest.

The methods and concepts developed in this chapter have great potential to improve biological discovery across a wide-array of genetic studies, particularly with the advent of large-scale phenotypic and genotypic datasets.
4.2 Introduction

Genome-wide association studies are a powerful exploratory tool for biological discovery in plant, animal and human genetic studies (Cortes et al., 2021; Visscher et al., 2012, 2017; Zhu et al., 2008). The most widely adopted approach involves a linear mixed model that sequentially fits and tests each marker of interest (Kang et al., 2008; Yu et al., 2006). The appealing feature of this approach is that population structure, kinship and other known covariance structures can be modelled simultaneously, which leads to a reduction in false positives (Li et al., 2014). Population structure is modelled by a fixed coefficient regression, with covariates generally obtained from either structured association analysis (Pritchard et al., 2000) or principal component analysis (Price et al., 2006). Kinship is then modelled by a random polygenic term which is parameterised by pedigree (Mrode, 2014) or marker data (Yu et al., 2006), with markers fitted directly or indirectly through a genomic relationship matrix (VanRaden, 2008). However, the use of marker data instead of pedigree has raised an important question in the literature; should the marker effect of interest also be included in the polygenic term? This has resulted in two alternative approaches for GWAS; (i) the marker-in approach which fits the marker effect as fixed and random in the polygenic term (Kang et al., 2008; Yu et al., 2006), and (ii) the marker-out approach which fits the marker effect as fixed but not random (Listgarten et al., 2012; Wang et al., 2014).

Despite their popularity, the marker-in and marker-out approaches are inherently inefficient because they involve fitting a separate model for each marker of interest. Zhou and Stephens (2012) addressed this issue for the marker-in approach by applying dimension reduction to the polygenic term (also see Lippert et al., 2011; Meyer and Tier, 2012). At a similar time, Kang et al. (2010) proposed an efficient approximation where the variance parameters are estimated using a baseline linear mixed model and then constrained for all subsequent models (also see Zhang et al., 2010). This has become standard procedure in many software packages (see, for example, Bradbury et al., 2007; Lipka et al., 2012; Yang et al., 2011). Gualdrón Duarte et al. (2014) later proposed that all test statistics can be obtained from the fit of the same baseline linear mixed model, which elevated the need to fit any subsequent models (also see Bernal Rubio et al., 2016; Zhang et al., 2021). This approach was subsequently badged GWAS by GBLUP (Legarra et al., 2018). GWAS by GBLUP has already been employed by current linear mixed model software (Misztal et al., 2018). However, the marker-in approach fits the marker effect of interest as both fixed and random, so it contributes to the mean and variance. This double fitting contradicts quantitative genetics theory which states that the marker effects contribute to the mean whereas the genotype effects contribute to the variance (Falconer and Mackay, 1996; Gianola et al., 2016). GWAS by GBLUP emulates the marker-in approach, so that it also inherits this issue.
In this chapter, it will be shown that double fitting the marker effect of interest as fixed and random causes the two effects to be statistically aliased, which causes the test statistics to be shrunken by the polygenic variance. This limits the feasibility of the marker-in approach for GWAS. Conversely, the marker-out approach avoids these issues because the marker of interest is excluded from the polygenic term, so the corresponding test statistics are not shrunken by the polygenic variance. This makes the marker-out approach favourable from both a quantitative genetics and statistical perspective. However, there is currently no fast method to obtain the marker-out test statistics from the fit of a single linear mixed model, so it involves fitting a very large number of models for all markers of interest. There is need for an equivalent GWAS by GBLUP approach for obtaining the marker-out test statistics.

The aim of this chapter is to develop fast exact methods for GWAS which produce the marker-in and marker-out test statistics from the fit of a single linear mixed model. These methods are extended for testing a set of markers at a time and testing a single marker while excluding a set from the polygenic term. The methods are also generalised for any set of fixed and random effects, including modelling population structure, complex genetic and residual covariance structures and spatial modelling. It will be shown that the test statistics are exactly the same as those obtained from fitting a separate linear mixed model for each marker of interest, provided that the variance components are known or unaffected by fitting the marker effect as fixed. The effect of estimating variance parameters will also be examined empirically using an example dataset. Lastly, it will be shown that a simple rescaling can be applied to the marker-in test statistics to obtain the marker-out statistics, which provides an immediate fix for current software packages employing the marker-in approach.

4.3 Linear mixed models

Assume a dataset comprises $n$ records on $v$ genotypes with $m$ markers, where $n \geq v$ and $m \geq v$. Let the $n$-vector of phenotypic data be given by $y = (y_1, \ldots, y_n)^\top$ and let the $v \times m$ marker design matrix be given by $M = [m_1 \ldots m_m]$, where $m_i$ is the $v$-vector of centred genotype scores for marker $i$.

Three linear mixed models (LMMs) are considered below:

1. A baseline LMM which fits all marker effects as random.

2. An extension of the baseline LMM which fits marker effect $i$ as fixed and random.

3. An extension of the baseline LMM which fits marker effect $i$ as fixed but not random.

The difference between Models 2 and 3 is whether marker $i$ is in or out of the polygenic term,
so they are hereafter referred to as the marker-in and marker-out LMMs. This difference will be highlighted below using subscripts. Note that the baseline LMM is fitted only once, whereas the marker-in and marker-out LMMs are fitted \( m \) times (once for each marker). It will be shown how the marker-in and marker-out test statistics can be obtained from the fit of the baseline LMM.

### 4.3.1 Baseline linear mixed model

The baseline LMM is given by:

\[
y = X\tau + ZM a + e, \tag{4.1}
\]

where \( \tau \) is a vector of fixed effects with design matrix \( X \), \( a \) is the \( m \)-vector of random marker effects with \( v \times m \) marker design matrix \( M \) and \( n \times v \) genotype design matrix \( Z \) and \( e \) is the \( n \)-vector of residuals. The fixed effects may include mean parameters and population structure, while the residuals may include random non-additive and non-genetic effects. The methods developed in this chapter are applicable to any set of fixed and random effects.

It is assumed that:

\[
\begin{bmatrix}
a \\
e
\end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_a D & 0 \\ 0 & \sigma^2_e R \end{bmatrix} \right), \tag{4.2}
\]

where \( \sigma^2_a \) is the additive genetic variance with \( m \times m \) diagonal marker matrix \( D \) and \( \sigma^2_e \) is the residual variance with \( n \times n \) residual variance matrix \( R \), which is assumed to be completely general. Following VanRaden (2008), \( D \) is assumed to have diagonal elements given by

\[
d_i = \left[ \sum_{t=1}^t 2p_i(1 - p_i) \right]^{-1} \quad \text{or} \quad d_i = \left[ m2p_i(1 - p_i) \right]^{-1},
\]

where \( p_i \) is the minor allele frequency at marker \( i \).

It then follows that:

\[
y \sim N(X\tau, H), \tag{4.3}
\]

where \( H = \sigma^2_a Z_m D Z_m^\top + \sigma^2_e R \) and \( Z_m = ZM \). Standard results for the fixed and random effects give \( \hat{\tau} = (X^\top H^{-1}X)^{-1}X^\top H^{-1}y \) and \( \hat{a} = \sigma^2_a D Z_m^\top H^{-1}(y - X\hat{\tau}) \).

The BLUP of marker effect \( i \) is therefore given by:

\[
\hat{a}_i = \sigma^2_a d_i Z_m^\top H^{-1}(y - X\hat{\tau}) = \sigma^2_a d_i Z_m^\top Py, \tag{4.4}
\]
with prediction error variance given by:

\[
\text{PEV}(\tilde{a}_i) = \sigma^2_a d_i - (\sigma^2_a d_i) \frac{2}{\var(\tilde{a}_i)},
\]

where \( P = H^{-1} - H^{-1}X(X^tH^{-1}X)^{-1}X^tH^{-1} \) and \( Z_{m_i} = ZM_i \). In Section 4.5, the test statistics for all markers of interest will be obtained from the components in Eqs. 4.4 and 4.5.

### 4.3.2 Marker-in linear mixed model

The marker-in LMM is an extension of the baseline LMM in Eq. 4.1 which fits marker effect \( i \) as fixed and random in the polygenic term. The marker-in LMM for marker \( i \) is given by:

\[
y_{in} = X\tau_{in} + Z_{m_i}\mu_{in} + ZM a_{in} + e_{in},
\]

where \( \tau_{in} \) is a vector of fixed effects, \( \mu_{in} \) is the fixed effect of marker \( i \), \( a_{in} \) is the \( m \)-vector of random marker effects including marker \( i \) and \( e_{in} \) is the \( n \)-vector of residuals. All other terms are given in Eq. 4.1.

It is assumed that:

\[
\begin{bmatrix}
a_{in} \\
e_{in}
\end{bmatrix}
\sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_a D & 0 \\ 0 & \sigma^2_e R \end{bmatrix}\right),
\]

where all terms are given in Eq. 4.2. It is important to note that since marker effect \( i \) is fitted as fixed and random, it contributes to the mean and variance of \( y_{in} \), i.e. \( \beta_{in} \sim N(\mu_{in}, \sigma^2_a d_i) \), where \( \beta_{in} = \mu_{in} + a_{in} \) is the total effect of marker \( i \). This double fitting has important ramifications on the estimates and subsequent test statistics (also see Gianola et al., 2016).

It then follows that:

\[
y_{in} \sim N(X\tau_{in} + Z_{m_i}\mu_{in}, H),
\]

where \( H \) is given in Eq. 4.3. The data vector \( y_{in} \) therefore has the same variance matrix as \( y \) but a different mean vector.

The BLUE of marker effect \( i \) is given by:

\[
\hat{\mu}_{in} = (Z_{m_i}^tPZ_{m_i})^{-1}Z_{m_i}^tPy_{in},
\]
with error variance given by:

\[
\text{PEV}(\hat{\mu}_{\text{in}}) = (Z_m^TPZ_m)^{-1} = \text{var}(\hat{\mu}_{\text{in}}),
\]

where \( P \) is given in Eq. 4.5. Note that “PEV” is used throughout this chapter for continuity, despite \( \hat{\mu}_{\text{in}} \) being an estimate not a prediction. Since marker effect \( i \) is also fitted as random, a BLUP of this effect can be obtained in addition to the BLUE. It can be shown that:

\[
\tilde{a}_{\text{in}} = 0 \quad \text{and} \quad \text{PEV}(\tilde{a}_{\text{in}}) = \sigma^2_d \tilde{d}_i = \sigma^2_d \tilde{d}_i - \text{var}(\tilde{a}_{\text{in}}), \tag{4.7}
\]

where \( \text{var}(\tilde{a}_{\text{in}}) = 0 \), meaning that the BLUP of marker effect \( i \) and its variance become zero as a result of fitting the same marker effect as fixed. In fact, the two effects are statistically aliased, so that after estimating \( \hat{\mu}_{\text{in}} \) there is no information left to directly predict \( \tilde{a}_{\text{in}} \). An indirect prediction can be obtained using the information in \( D \), which is why \( \tilde{a}_{\text{in}} = 0 \) when \( D \) is diagonal. Note, however, this is not the case for non-diagonal forms of \( D \).

Lastly, the error covariance between the BLUE and BLUP of marker effect \( i \) is given by:

\[
\text{PEV}(\hat{\mu}_{\text{in}}, \tilde{a}_{\text{in}}) = -\sigma^2_d \tilde{d}_i. \tag{4.8}
\]

In Section 4.4, a fundamental relationship between the marker-in and marker-out LMMs will be derived using the components in Eqs. 4.7 and 4.8.

**4.3.3 Marker-out linear mixed model**

The marker-out LMM is an extension of the baseline LMM in Eq. 4.1 which fits marker effect \( i \) as fixed but not random in the polygenic term. Let the \( v \times (m - 1) \) marker design matrix excluding column \( i \) be given by \( M_{-i} \). The marker-out LMM for marker \( i \) is given by:

\[
y_{\text{out}} = X \tau_{\text{out}} + Z_m \mu_{\text{out}} + ZM_{-i}a_{-\text{out}} + e_{\text{out}}, \tag{4.9}
\]

where \( \tau_{\text{out}} \) is a vector of fixed effects, \( \mu_{\text{out}} \) is the fixed effect of marker \( i \), \( a_{-\text{out}} \) is the \( m - 1 \) vector of random marker effects excluding marker \( i \) and \( e_{\text{out}} \) is the \( n \)-vector of residuals. All other terms are given in Eq. 4.1.
4.3 Linear mixed models

It is assumed that:

\[
\begin{bmatrix}
a_{i_{\text{out}}} \\
e_{\text{out}}
\end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_a^2 D_{-i} & 0 \\ 0 & \sigma_e^2 R \end{bmatrix}\right),
\]

where \(D_{-i}\) is the \((m - 1) \times (m - 1)\) diagonal marker matrix excluding marker \(i\) and all other terms are given in Eq. 4.2. It is important to note that the variance parameters are exactly the same as for the marker-in LMM in Eq. 4.6, regardless of whether they are known or require estimation. This is because the fixed effect of marker \(i\) “knocks out” any potential contribution from the random effect anyway (see Section 4.7).

It then follows that:

\[
y_{\text{out}} \sim N(X\tau_{\text{out}} + Z_m \mu_{\text{out}}, H_{-i}),
\]

where \(H_{-i} = H - \sigma_a^2 d_i Z_m Z_m^\top\). Note that since marker effect \(i\) is fitted as fixed but not random, it contributes to the mean of \(y_{\text{out}}\) but not its variance. The data vector \(y_{\text{out}}\) therefore has a different mean vector and variance matrix compared to \(y\) in Eq. 4.2 and \(y_{\text{in}}\) in Eq. 4.6.

The BLUE of marker effect \(i\) is given by:

\[
\hat{\mu}_{i_{\text{out}}} = (Z_m^\top P_{-i} Z_m)^{-1} Z_m^\top P_{-i} y_{\text{out}},
\]

where \(P_{-i} = H_{-i}^{-1} - H_{-i}^{-1} X(X^\top H_{-i}^{-1} X)^{-1} X^\top H_{-i}^{-1}\).

Using Result 2 in Appendix E, it can be shown that:

\[
Z_m^\top P_{-i} = \frac{Z_m^\top P}{1 - \sigma_a^2 d_i Z_m^\top P Z_m},
\]

where \(\omega_i = 1 - \sigma_a^2 d_i Z_m^\top P Z_m\) is a scaling constant, which is equivalent to \(\omega_i = \text{PEV}(\tilde{a}_i)/\sigma_a^2 d_i\) from the baseline LMM. Substituting Eq. 4.11 into Eq. 4.10 gives:

\[
\hat{\mu}_{i_{\text{out}}} = (Z_m^\top P Z_m)^{-1} Z_m^\top P y_{\text{out}},
\]

i.e. \(\hat{\mu}_{i_{\text{out}}} - \hat{\mu}_{i_{\text{in}}} = 0\),

meaning that the BLUE of marker effect \(i\) is invariant to whether the same marker effect is also fitted as random. Note, however, this is not the case for non-diagonal forms of \(D\). The result in Eq. 4.12 can be generalised as \(\hat{\mu}_{i_{\text{out}}} - (\hat{\mu}_{i_{\text{in}}} + \tilde{a}_{i_{\text{in}}}) = 0\), meaning that the BLUE of the fixed effect from the marker-in LMM is adjusted during estimation to make room for the BLUP of the random effect.
The error variance is then given by:

$$\text{PEV}(\hat{\mu}_{\text{out}}) = (Z_m^TP_iZ_m)^{-1}. \quad (4.13)$$

Substituting Eq. 4.11 into Eq. 4.13 gives:

$$\text{PEV}(\hat{\mu}_{\text{out}}) = (Z_m^TPZ_m)^{-1} - \sigma_a^2 d_i$$

i.e. $$\text{var}(\hat{\mu}_{\text{out}}) = \text{var}(\hat{\mu}_{\text{in}}) - \sigma_a^2 d_i, \quad (4.14)$$

meaning that the variance of the BLUE is inflated by $\sigma_a^2 d_i$ when the same marker effect is also fitted as random. The result in Eq. 4.14 can be generalised as $\text{var}(\hat{\mu}_{\text{out}}) = \text{var}(\hat{\mu}_{\text{in}})$ + $\text{PEV}(\hat{\mu}_{\text{in}}, \tilde{a}_{\text{in}})$, meaning that the variance of the BLUE from the marker-in LMM is adjusted during estimation to make room for the variance of the BLUP. The scaling constant can therefore be rewritten as $\omega_i = \text{var}(\hat{\mu}_{\text{out}}) / \text{var}(\hat{\mu}_{\text{in}})$, which is equivalent to $\omega_i = \text{var}(\hat{\mu}_{\text{out}}) / [\text{var}(\hat{\mu}_{\text{in}}) + \sigma_a^2 d_i]$ and $\omega_i = [\text{var}(\hat{\mu}_{\text{in}}) - \sigma_a^2 d_i] / \text{var}(\hat{\mu}_{\text{in}})$. It then follows that $0 \leq \omega_i \leq 1$, assuming $\sigma_a^2 d_i \geq 0$. These results will be utilised below.

### 4.4 Fundamental expressions

There is a fundamental relationship between the marker-in and marker-out LMMs, which can be described by two expressions:

1. The BLUES of marker effect $i$ are proportional:

$$\hat{\mu}_{\text{out}} - (\hat{\mu}_{\text{in}} + \tilde{a}_{\text{in}}) = 0,$$

i.e. $$\hat{\mu}_{\text{out}} - \hat{\mu}_{\text{in}} = 0, \quad (4.15)$$

since $\tilde{a}_{\text{in}} = 0$ when $D$ is diagonal.

2. The error variances are proportional:

$$\text{PEV}(\hat{\mu}_{\text{out}}) = \text{PEV}(\hat{\mu}_{\text{in}}) + \text{PEV}(\tilde{a}_{\text{in}}) + 2\text{PEV}(\hat{\mu}_{\text{in}}, \tilde{a}_{\text{in}})$$

$$= \text{PEV}(\hat{\mu}_{\text{in}}) - \sigma_a^2 d_i,$$

i.e. $$\text{var}(\hat{\mu}_{\text{out}}) = \text{var}(\hat{\mu}_{\text{in}}) - \sigma_a^2 d_i, \quad (4.16)$$

since $\text{PEV}(\tilde{a}_{\text{in}}) = -\text{PEV}(\hat{\mu}_{\text{in}}, \tilde{a}_{\text{in}}) = \sigma_a^2 d_i$ when $D$ is diagonal.
These expressions show that the BLUE of the fixed effect and its variance from the marker-in LMM are adjusted by the BLUP of the random effect. The aliasing therefore affects estimation of the fixed and random effects. This has important ramifications on the feasibility of the marker-in test statistics for GWAS. In the next section, it will be shown that the resulting test statistics are shrunken due to the aliasing.

There is also a fundamental relationship between the baseline LMM and the marker-in and marker-out LMMs, which can be described by two further expressions:

3. The variances are inversely proportional:

\[
\text{var}(\tilde{a}_i) = \frac{\left(\sigma_a^2 d_i\right)^2}{\text{var}(\hat{\mu}_{in})},
\]

i.e. \(\frac{\text{var}(\tilde{a}_i)}{\omega_i} = \frac{\left(\sigma_a^2 d_i\right)^2}{\text{var}(\hat{\mu}_{out})}\), \(4.17\)

since \(\text{var}(\hat{\mu}_{in}) = \text{var}(\hat{\mu}_{out}) + \sigma_a^2 d_i\) and \(\omega_i = \text{PEV}(\tilde{a}_i)/\sigma_a^2 d_i\) when \(D\) is diagonal. This is an extension of Proposition 1 in Bernal Rubio et al. (2016) for the marker-out LMM.

4. The BLUEs of marker effect \(i\) are proportional:

\[
\tilde{a}_i = \frac{\sigma_a^2 d_i}{\text{var}(\hat{\mu}_{in})} \hat{\mu}_{in},
\]

i.e. \(\tilde{a}_i = \frac{\sigma_a^2 d_i}{\text{var}(\hat{\mu}_{out})} \hat{\mu}_{out}\), \(4.18\)

since \(\hat{\mu}_{in} - \hat{\mu}_{out} = 0\) when \(D\) is diagonal. This is an extension of Proposition 2 in Bernal Rubio et al. (2016) for the marker-out LMM.

### 4.5 Test statistics

The marker-in and marker-out test statistics are given by:

\[
z_{in} = \frac{\hat{\mu}_{in}}{\sqrt{\text{var}(\hat{\mu}_{in})}} \quad \text{and} \quad z_{out} = \frac{\hat{\mu}_{out}}{\sqrt{\text{var}(\hat{\mu}_{out})}},
\]

\(4.19\)

where \(z_{in}\) is the statistic for testing \(H_0: \mu_{in} = 0\) and \(z_{out}\) is the statistic for testing \(H_0: \mu_{out} = 0\), which are assumed to have standard normal distributions.
It then follows that:

\[ z_{in} \leq z_{out}, \]  

(4.20)

since \( z_{in} = \sqrt{\omega_i} z_{out} \) and \( 0 \leq \omega_i \leq 1, \)

meaning that \( z_{in} \) is shrunken compared to \( z_{out} \) according to \( \sigma_d^2 d_i. \) The difference between the two test statistics is therefore due to improper scaling of \( z_{in} \) as a result of the aliasing. This difference will be important when the number of markers is small, but will become negligible when the number of markers is very large, i.e. when \( d_i \approx 0 \) and thence \( \omega_i \approx 1. \)

It also follows that:

\[ z_{out} = \frac{\hat{\mu}_{iin}}{\sqrt{\text{var}(\hat{\mu}_{iin}) - \sigma_d^2 d_i}}, \]  

(4.21)

since \( z_{out} = z_{in} / \sqrt{\omega_i} \) and \( \omega_i = \left[ \text{var}(\hat{\mu}_{iin}) - \sigma_d^2 d_i \right] / \text{var}(\hat{\mu}_{iin}), \)

meaning that the marker-out test statistics can be obtained from a simple rescaling of the marker-in test statistics. This provides an immediate fix for current software packages fitting the marker-in LMM.

Obtaining the test statistics in Eq. 4.19 for all \( m \) markers is slow because it requires fitting the \( m \) models in Eq. 4.6 and 4.11. However, all required test statistics can be obtained directly from the fit of the baseline LMM in Eq. 4.1. The test statistics are now obtained by dividing the components in Eq. 4.18 by the square root of the components in Eq. 4.17, which gives:

\[ z_{in} = \frac{\hat{a}_i}{\sqrt{\text{var}(\hat{a}_i)}} \quad \text{and} \quad z_{out} = \frac{\hat{a}_i}{\sqrt{\omega_i \text{var}(\hat{a}_i)}}, \]  

(4.22)

where \( \omega_i = \text{PEV}(\hat{a}_i) / \sigma_d^2 d_i. \) This provides a fast method to obtain all required test statistics from the fit of a single LMM. When the variance parameters are known, or when their estimation is unaffected by fitting marker effect \( i \) as fixed, the test statistics in Eq. 4.22 are exactly the same as those in Eq. 4.19. The effect of estimating variance parameters will be examined empirically for an example dataset in Section 4.8.

Lastly, the marker-in and marker-out \( p \)-values are given by:

\[ p_{iin} = -\log_{10}(2\Phi_{iin}) \quad \text{and} \quad p_{iout} = -\log_{10}(2\Phi_{iout}), \]  

(4.23)

where \( \Phi_{iin} \) and \( \Phi_{iout} \) are the cumulative distribution functions for \( z_{in} \) and \( z_{out} \) given \( P[Z > |z|]. \) The \(-\log_{10}p\)-values are hereafter referred to as simply “\( p \)-values” for brevity.
4.5.1 A simple fix for current software

It was shown in Eq. 4.21 how the marker-in test statistics, \( z_{in} \), are shrunken by the polygenic variance, \( \sigma_a^2 d_i \). A simple fix was provided to obtain the marker-out test statistics, that is by rescaling \( z_{in} \) by \( 1 / \sqrt{\omega_i} \). Interestingly, this is equivalent to testing the total effect of marker \( i \), given by \( \beta_{iin} = \mu_{iin} + a_{iin} \), rather than just the fixed effect. It can be shown that:

\[
\tilde{\beta}_{iin} = \hat{\mu}_{iin} \quad \text{and} \quad \text{PEV}(\tilde{\beta}_{iin}) = \text{PEV}(\hat{\mu}_{iin}) + \text{PEV}(\tilde{a}_{iin}) + 2 \text{PEV}(\hat{\mu}_{iin}, \tilde{a}_{iin}) = \text{var}(\hat{\mu}_{iin}) - \sigma_a^2 d_i,
\]

since \( \tilde{a}_{iin} = 0 \) and \( \text{PEV}(\tilde{a}_{iin}) = -\text{PEV}(\hat{\mu}_{iin}, \tilde{a}_{iin}) = \sigma_a^2 d_i \) when \( D \) is diagonal.

It then follows that:

\[
z_{iin}^* = \frac{\tilde{\beta}_{iin}}{\sqrt{\text{PEV}(\tilde{\beta}_{iin})}},
\]

(4.24)

where \( z_{iin} \) is the statistic for testing \( H_0 : \beta_{iin} = 0 \), which is now exactly the same as the marker-out test statistic in Eq. 4.19. The practical implication is that current software which employ the marker-in LMM can produce the marker-out test statistics by either rescaling \( z_{iin} \) by \( 1 / \sqrt{\omega_i} \) or by constructing \( z_{iin}^* \) directly in Eq. 4.24. This result holds regardless of whether the variance parameters are known or require estimation (see Section 4.7).

4.6 Genotype to marker

When there are more markers than genotypes, i.e. \( m > v \), it is more efficient to fit the random genetic term in Eq. 4.1 at the genotype level, rather than at the marker level. This feature forms the basis of GWAS by GBLUP (Legarra et al., 2018). The baseline LMM can therefore be rewritten as:

\[
y = X\tau + Zu + e,
\]

(4.25)

where \( u = Ma \) is the \( v \)-vector of random genotype effects, i.e. breeding values, constructed with all \( m \) markers. All other terms are given in Eq. 4.1.

It is assumed that:

\[
\begin{bmatrix}
u \\
a_i
\end{bmatrix} \sim N\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma_a^2 \begin{bmatrix} G & d_i \bar{m}_i \\ d_i \bar{m}_i^\top & d_i \end{bmatrix} \right),
\]

where \( G = MDM^\top \) is the \( v \times v \) genomic relationship matrix (VanRaden, 2008).
The vector of predicted genotype effects is given by:
\[
\tilde{\mathbf{u}} = \sigma_a^2 \mathbf{GZ}^\top \mathbf{P_y},
\]
with prediction error variance matrix given by:
\[
\text{PEV}(\tilde{\mathbf{u}}) = \sigma_a^2 \mathbf{G} - (\sigma_a^2)^2 \mathbf{GZ}^\top \mathbf{PZG}
\]
\[
= \sigma_a^2 \mathbf{G} - \text{var}(\tilde{\mathbf{u}}),
\]
where \(\text{var}(\tilde{\mathbf{u}}) = (\sigma_a^2)^2 \mathbf{GZ}^\top \mathbf{PZG}\).

Following Stranden and Garrick (2009), the BLUP of marker effect \(i\) can be obtained as:
\[
\tilde{a}_i = d_i \mathbf{m}_i^\top \mathbf{G}^{-1} \tilde{\mathbf{u}}
\]
\[
= \sigma_a^2 d_i \mathbf{Z}_m^\top \mathbf{P}_y,
\]
(4.26)

with prediction error variance obtained as:
\[
\text{PEV}(\tilde{a}_i) = \sigma_a^2 d_i - \sigma_a^2 d_i^2 \mathbf{m}_i^\top \mathbf{G}^{-1} \mathbf{m}_i + d_i^2 \mathbf{m}_i^\top \mathbf{G}^{-1} \text{PEV}(\tilde{\mathbf{u}}) \mathbf{G}^{-1} \mathbf{m}_i
\]
\[
= \sigma_a^2 d_i - \sigma_a^2 d_i \mathbf{Z}_m^\top \mathbf{PZ}_m
\]
\[
= \sigma_a^2 d_i - \text{var}(\tilde{a}_i),
\]
(4.27)

and variance obtained as:
\[
\text{var}(\tilde{a}_i) = d_i^2 \mathbf{m}_i^\top \mathbf{G}^{-1} \text{var}(\tilde{\mathbf{u}}) \mathbf{G}^{-1} \mathbf{m}_i
\]
\[
= (\sigma_a^2 d_i)^2 \mathbf{Z}_m^\top \mathbf{PZ}_m.
\]
(4.28)

The test statistics in Eq. 4.22 can now be constructed using the components in Eqs. 4.26, 4.27 and 4.28. Similar results can also be obtained for the marker-in and marker-out LMMs by fitting the polygenic term at the genotype level, rather than at the marker level.

### 4.7 Model estimation

All models considered in this chapter were fitted using the ASReml-R package (Butler et al., 2017) in R (R Core Team, 2023). ASReml-R obtains REML estimates of the variance parameters, as well as empirical BLUEs and BLUPs of the fixed and random effects, respectively. The prediction error variances of the empirical BLUEs and BLUPs can be obtained using an additional call to ASReml-R. Two important results are derived below.
1. The residual log-likelihood for the marker-in and marker-out LMMs are exactly the same. Following Verbyla (1990), the residual log-likelihood for the marker-in LMM is given by (omitting constants):

\[ L_{\text{Rin}} = -\frac{1}{2} \left[ \log |H| + \log \left| X^* H^{-1} X^* \right| + y^\top P^* y \right], \tag{4.29} \]

where \( X^* = [X \ Z_m] \) and \( P^* = P - PZ_m (Z_m^\top PZ_m)^{-1} Z_m^\top P \). Using Results 4, 5 and 6 in Appendix E, it follows that:

\[ \log |H| + \log \left| X^* H^{-1} X^* \right| = \log |H - i| + \log \left| X^* H_{-i}^{-1} X^* \right|. \tag{4.30} \]

Substituting Eq. 4.30 into Eq. 4.29 gives:

\[ L_{\text{Rin}} = -\frac{1}{2} \left[ \log |H - i| + \log \left| X^* H_{-i}^{-1} X^* \right| + y^\top P^* y \right] = L_{\text{Rout}}, \]

where \( L_{\text{Rout}} \) is the residual log-likelihood for the marker-out LMM. This is an expected result since the residual log-likelihood is in fact a marginal log-likelihood based on the error contrasts, which have been adjusted for the fixed effects (Patterson and Thompson, 1971). The fixed effect of marker \( i \) therefore “knocks out” the corresponding random effect from the residual log-likelihood in Eq. 4.29. Consequently, the REML estimates of variance parameters from the marker-in and marker-out LMMs are also exactly the same. Note, however, the full log-likelihood is not the same.

2. The full log-likelihood for the marker-in LMM can be written in terms of the baseline LMM and the squared test statistics. Following Verbyla (2019), the full log-likelihood for the marker-in LMM is given by (omitting constants):

\[ L_{\text{F in}} = -\frac{1}{2} \left[ \log |H| + y^\top P^* y \right] \]
\[ = L_{\text{Rin}} + \frac{1}{2} \log \left| X^* H^{-1} X^* \right|, \tag{4.31} \]

where \( P^* \) is given in Eq. 4.29. It can be shown that:

\[ y^\top P^* y = (y - Z_m \tilde{a}_i)^\top P (y - Z_m \tilde{a}_i) \]
\[ = y^\top P y - \frac{1}{2} z_{\text{in}}^2, \tag{4.32} \]

where \( z_{\text{in}} = (Z_m^\top PZ_m)^{-1/2} Z_m^\top P y \) is the marker-in test statistic. Substituting Eq. 4.32 into
Eq. 4.31 gives:

\[
L_{F_{in}} = -\frac{1}{2} \left[ \log |H| + y^\top P y - \frac{1}{2} \mu_{in}^2 \right]
= L_F + \frac{1}{2} \mu_{in}^2,
\]

(4.33)

where \( L_F = -\frac{1}{2} \left[ \log |H| + y^\top P y \right] \) is the full log-likelihood for the baseline LMM. This is an interesting result.

### 4.8 Application to example dataset

#### 4.8.1 Data description

The example dataset is a publicly available wheat breeding dataset from the BGLR package (Peréz and de los Campos, 2014) in R. This dataset was previously used by Crossa et al. (2010) and Gianola et al. (2016). The phenotypic data were collected in field trials across Mexico. The trait considered here is grain yield in environments 2 and 3. A total of \( v = 599 \) inbred wheat genotypes were evaluated across these environments, with \( n = 1198 \) plots in total. Marker data were generated for all genotypes using \( m = 1447 \) DArT markers (Jaccoud et al., 2001), coded as either 1 or 0 for the presence or absence of the major allele. Markers were filtered for minor allele frequency \(< 5\% \) and missing values were imputed from the marginal distribution of the observed values. A total of 1279 markers were retained using this criteria. Although this dataset does not include marker positions, plot layouts or experimental design features, it provides a good example to demonstrate the concepts and methods developed thus far.

#### 4.8.2 Models fitted

The baseline, marker-in and marker-out LMMs were then fitted to the example wheat breeding dataset. Each model is detailed in the following.

1. The baseline LMM is given by:

\[
y = X \tau + Z u + e,
\]

(4.34)

where \( \tau \) is a vector of fixed environmental mean parameters, \( u \) is the \( v \)-vector of random genotype main effects and \( e \) is the \( n \)-vector of residuals. All other terms are given in Eq. 4.1.
It is assumed that:

\[
\begin{bmatrix}
  \mathbf{u} \\
  \mathbf{e}
\end{bmatrix}
\sim
\mathcal{N}
\left(
\begin{bmatrix}
  \mathbf{0} \\
  \mathbf{0}
\end{bmatrix},
\begin{bmatrix}
  \sigma_a^2 \mathbf{G} & \mathbf{0} \\
  \mathbf{0} & \sigma_e^2 \mathbf{R}
\end{bmatrix}
\right),
\]

(4.35)

where \( \mathbf{G} = \mathbf{MDM}^\top \) is the \( v \times v \) genomic relationship matrix, \( \mathbf{D} \) is the \( m \times m \) diagonal marker matrix with diagonal elements given by \( d_i = \text{tr} \left( \mathbf{MM}^\top \right)^{-1} \) and \( \mathbf{R} = \bigoplus_{j=1}^{2} \gamma_e \mathbf{I}_v \) is the \( n \times n \) residual variance matrix, such that \( \sigma^2_{e_j} = \sigma^2_e \gamma_e \) is the residual variance for environment \( j \).

2. The marker-in LMM is given by:

\[
y_{in} = \mathbf{X} \boldsymbol{\tau}_{in} + \mathbf{Z_m} \mu_{in} + \mathbf{Zu}_{in} + \mathbf{e}_{in},
\]

(4.36)

where \( \boldsymbol{\tau}_{in} \) is a vector of fixed environmental mean parameters, \( \mu_{in} \) is the fixed main effect of marker \( i \), \( \mathbf{u}_{in} \) is the \( v \)-vector of random genotype effects constructed with marker \( i \) and \( \mathbf{e}_{in} \) is the \( n \)-vector of residuals. All other terms are given in Eq. 4.6.

It is assumed that:

\[
\begin{bmatrix}
  \mathbf{u}_{in} \\
  \mathbf{e}
\end{bmatrix}
\sim
\mathcal{N}
\left(
\begin{bmatrix}
  \mathbf{0} \\
  \mathbf{0}
\end{bmatrix},
\begin{bmatrix}
  \sigma_a^2 \mathbf{G} & \mathbf{0} \\
  \mathbf{0} & \sigma_e^2 \mathbf{R}
\end{bmatrix}
\right),
\]

where all terms are given in Eq. 4.35.

3. The marker-out LMM is given by:

\[
y_{out} = \mathbf{X} \boldsymbol{\tau}_{out} + \mathbf{Z_m} \mu_{out} + \mathbf{Zu}_{-out} + \mathbf{e}_{out},
\]

(4.37)

where \( \boldsymbol{\tau}_{out} \) is a vector of fixed environmental mean parameters, \( \mu_{out} \) is the fixed main effect of marker \( i \), \( \mathbf{u}_{-out} \) is the \( v \)-vector of random genotype effects constructed without marker \( i \) and \( \mathbf{e}_{out} \) is the \( n \)-vector of residuals. All other terms are given in Eq. 4.9.

It is assumed that:

\[
\begin{bmatrix}
  \mathbf{u}_{-out} \\
  \mathbf{e}
\end{bmatrix}
\sim
\mathcal{N}
\left(
\begin{bmatrix}
  \mathbf{0} \\
  \mathbf{0}
\end{bmatrix},
\begin{bmatrix}
  \sigma_a^2 \mathbf{G}_{-i} & \mathbf{0} \\
  \mathbf{0} & \sigma_e^2 \mathbf{R}
\end{bmatrix}
\right),
\]

where \( \mathbf{G}_{-i} = \mathbf{M}_{-i} \mathbf{D}_{-i} \mathbf{M}_{-i}^\top \) is the \( v \times v \) genomic relationship matrix constructed without marker \( i \). All other terms are given in Eq. 4.35.
4.8.3 Results

There are three important results from fitting the baseline, marker-in and marker-out LMMs to the example dataset:

1. The marker-in $p$-values are consistently smaller than the marker-out $p$-values.

2. The $p$-values from the marker-in and marker-out LMMs are exactly the same as those from the baseline LMM when the variance parameters are known.

3. Fitting a single baseline LMM is orders of magnitude faster than sequentially fitting $m$ marker-in and marker-out LMMs.

Each result is detailed in the following.

The marker-in $p$-values are consistently smaller than the marker-out $p$-values, with mean $\pm$ se of $0.44 \pm 0.45$ compared to $0.53 \pm 0.56$ (Figure 4.1). This is highlighted by marker effects M1 and M2, which are significant for the marker-out $p$-values but not the marker-in $p$-values. In Section 4.5, it was shown that the marker-in test statistics are improperly

Fig. 4.1 Comparison of marker-in and marker-out $p$-values obtained using estimated variance parameters. The red lines correspond to a 5% significance level with Bonferroni correction for 1279 samples. Significant marker effects M1 and M2 are labelled.
4.8 Application to example dataset

Fig. 4.2 Comparison of marker-in and marker-out test statistics obtained using known and estimated variance parameters. Significant marker effects M1 and M2 are labelled.

scaled, with the amount of shrinkage being proportional to the polygenic variance. Figure 4.1 indicates that the improper scaling has more impact on marker effects with large \( p \)-values than those with small \( p \)-values (also see Figure 4.2). This is an expected result.

The \( p \)-values from the marker-in and marker-out LMMs are exactly the same as those from the baseline LMM when the variance parameters are known, with bias of 0 (Figure 4.3). This is also the case for the corresponding test statistics (not shown). To demonstrate “known” variance parameters, an initial fit of the baseline LMM was used to obtain REML estimates, with \( \hat{\sigma}_a^2 = 1.06, \hat{\sigma}_e^2 = 0.43 \) and \( \hat{\sigma}_e^3 = 0.48 \). The variance parameters were then constrained to these values in each of the \( m \) subsequent marker-in and marker-out LMMs. The variance parameters were then unconstrained and re-estimated in each of the \( m \) marker-in and marker-out LMMs. There are only very minor differences between the \( p \)-values in this case, with bias of 0.0 – 0.2. It is important to note that the \( p \)-values from the baseline LMM are generally conservative, with >70% smaller than those from the marker-in and marker-out LMMs and always smaller when \(-\log_{10} p > 1\).

Fitting a single baseline LMM is orders of magnitude faster than sequentially fitting \( m \) marker-in and marker-out LMMs, with < 1 sec to complete all tasks compared to 9.3 – 9.6 min when the variance parameters are known and 2.5 sec compared to 23.0 – 23.5 min when the variance parameters are estimated (Table 4.1). The tasks included construction of the genomic relationship matrix, model fitting in \textit{ASReml-R} and back-solving for the BLUPs of the marker effects and associated error variances. All computations were completed on a MacBook Pro 2018 with a 2.4 GHz 8-Core Intel Core i9 processor and 64 GB RAM.
Fig. 4.3 Comparison of $p$-values from the marker-in and marker-out linear mixed models with those from the baseline LMM. The $p$-values were obtained using known or estimated variance parameters. Significant marker effects M1 and M2 are labelled.

There is likely software that is faster than $\sim 10$ min for fitting the $m$ marker-in LMMs, e.g. GEMMA (Zhou and Stephens, 2012) and FaST-LMM (Lippert et al., 2011), but note there is no faster method for fitting the $m$ marker-out LMMs, and there is certainly no method that can complete both tasks in under 3 seconds.

Table 4.1 compares the fit of the baseline, marker-in and marker-out LMMs in terms of AIC. This is an application of the full log-likelihood approach of Verbyla (2019), which enables models to be compared with different fixed effects, e.g. the baseline LMM compared to the marker-in and marker-out LMMs. The baseline LMM provides a better fit on average, with AIC of 811.3 compared to 812.2 for the marker-in LMM and 812.0 for the marker-out
Table 4.1 Summary of the linear mixed models fitted to the example wheat breeding dataset, including the number of models fitted, computing time, log-likelihood, AIC and values of the known or estimated variance parameters. The log-likelihood, AIC and estimated variance parameters for the marker-in and marker-out LMMs are given as the average of all 1279 models fitted.

<table>
<thead>
<tr>
<th>Variances</th>
<th>Model</th>
<th>No. fitted</th>
<th>Time</th>
<th>Loglik$^2$</th>
<th>AIC$^2$</th>
<th>$\sigma_a^2$</th>
<th>$\sigma_e^2$</th>
<th>$\sigma_{\epsilon_3}^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>1</td>
<td>&gt; 1 sec</td>
<td>-400.6</td>
<td>811.3</td>
<td>1.06</td>
<td>0.43</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td>Known</td>
<td>Marker-in$^1$</td>
<td>1279</td>
<td>9.3 min</td>
<td>-400.1</td>
<td>812.2</td>
<td>1.06</td>
<td>0.43</td>
<td>0.48</td>
</tr>
<tr>
<td>Estimated</td>
<td>Marker-out$^1$</td>
<td>1279</td>
<td>9.6 min</td>
<td>-400.0</td>
<td>812.0</td>
<td>1.06</td>
<td>0.43</td>
<td>0.48</td>
</tr>
<tr>
<td>Baseline</td>
<td>1</td>
<td>2.5 sec</td>
<td>-400.6</td>
<td>811.3</td>
<td>1.06</td>
<td>0.43</td>
<td>0.48</td>
<td></td>
</tr>
<tr>
<td>Known</td>
<td>Marker-in</td>
<td>1279</td>
<td>23.0 min</td>
<td>-400.1</td>
<td>812.2</td>
<td>1.06</td>
<td>0.43</td>
<td>0.48</td>
</tr>
<tr>
<td>Estimated</td>
<td>Marker-out</td>
<td>1279</td>
<td>23.5 min</td>
<td>-400.0</td>
<td>812.0</td>
<td>1.06</td>
<td>0.43</td>
<td>0.48</td>
</tr>
</tbody>
</table>

$^1$Variance parameters were constrained to the REML estimates from the baseline LMM

$^2$Reported using the full log-likelihood approach of Verbyla (2019)

It is important to note that all $m$ marker-out LMMs provide a better fit than the corresponding marker-in LMMs, regardless of whether the variance parameters are known or estimated. This is because the aliased random effects in the marker-in LMM reduce the full log-likelihood by $\log(1 + \sigma_a^2 d_i Z_m^T H_{\epsilon_i}^{-1} Z_m)$/2 compared to the marker-out LMM.

Lastly, Table 4.1 provides a summary of the variance parameters from the baseline, marker-in and marker-out LMMs. The top half are the “known” variance parameters obtained from an initial fit of the baseline LMM, with $\hat{\sigma}_a^2 = 1.06$, $\hat{\sigma}_e^2 = 0.43$ and $\hat{\sigma}_{\epsilon_3}^2 = 0.48$. The error variances for environments 2 and 3 were similar, which resulted in very similar heritabilities, i.e. $h_2^2 = 0.71$ and $h_3^2 = 0.69$, calculated as the proportion of additive genetic variance of the total phenotypic variance. The bottom half are variance parameters re-estimated in each of the $m$ marker-in and marker-out LMMs. On average, all estimated variance parameters and heritabilities are the same as for the known variance parameters. However, the variance parameters exhibit minor changes for particular models, with $\hat{\sigma}_a^2 = 0.97 − 1.07$, $\hat{\sigma}_e^2 = 0.42 − 0.43$ and $\hat{\sigma}_{\epsilon_3}^2 = 0.47 − 0.48$, which produced heritabilities of $h_2^2 = 0.69 – 0.72$ and $h_3^2 = 0.67 – 0.69$. Readers are referred to Gianola et al. (2016) for further details on how these parameters are affected by removing individual markers from $G$. 
4.8.4 Discussion

GWAS are a powerful exploratory tool for biological discovery in a wide-array of genetic studies (Visscher et al., 2017; Zhu et al., 2008). Many of the current approaches are inherently inefficient, however, because they involve fitting a very large number of models to obtain the required test statistics for all markers of interest. Many are also deficient because they double fit the marker effect of interest as fixed and random, which can result in shrunken test statistics. The methods developed in this chapter enable the appropriate test statistics to be obtained from the fit of a single LMM.

The most widely adopted approach for GWAS involves a marker-in LMM that sequentially fits and tests each marker effect of interest (Kang et al., 2008; Yu et al., 2006). The LMM approach provides a flexible framework for conducting GWAS, but is inherently inefficient because it involves fitting $m$ models to obtain the required test statistics for all $m$ markers of interest. This issue has been the focus of numerous studies (Kang et al., 2010; Lippert et al., 2011; Zhou and Stephens, 2012), but note that all of these still require sequentially fitting a very large number of marker-in LMMs. An exception is Gualdrón Duarte et al. (2014), who proposed that all required test statistics can be obtained from the fit of a single baseline LMM, which fits all marker effects as random through a genomic relationship matrix (also see Bernal Rubio et al., 2016; Zhang et al., 2021). This approach was subsequently badged GWAS by GBLUP (Legarra et al., 2018). However, there remained the need for an equivalent GWAS by GBLUP approach that emulates fitting the $m$ marker-out LMMs. The methods developed in this chapter produce all required test statistics for both approaches from the fit of the same baseline LMM, with the same genomic relationship matrix.

The two GWAS approaches considered in this chapter are based on the marker-in and marker-out LMMs in Eqs 4.6 and 4.9, respectively. The difference between these approaches is whether the marker effect of interest is included or excluded from the polygenic term. It was shown algebraically that double fitting the marker effect of interest as fixed and random causes the two effects to be statistically aliased, which causes the test statistics to be shrunken by the polygenic variance. The marker-out approach avoids these issues because the marker of interest is excluded from the polygenic term. A fundamental relationship was derived which describes the differences between the marker-in and marker-out LMMs. It states that the estimates of marker effect $i$ are equal but their variance differs according to the polygenic variance. This causes the shrinkage observed in the marker-in test statistics. It also states that the fit of the marker-in and marker-out LMMs are exactly the same, regardless of whether the variance parameters are known or estimated. This lead to a fundamental expression which enables the marker-out test statistics to be directly obtained from the fit of the marker-in LMM, and vice versa. The expression provides a simple fix for current software packages.
fitting the marker-in LMM and elevates the need to construct a separate genomic relationship matrix for each marker.

The methods developed in this chapter were demonstrated on a small wheat breeding dataset available from the BGLR package. There were three important results:

1. The marker-in \( p \)-values were consistently smaller than the marker-out \( p \)-values. This highlights the shrinkage in the marker-in test statistics caused by the aliasing. It is important to note that the marker-in and marker-out test statistics will converge when the number of markers becomes very large, or more specifically when the scaling constant, \( \omega_i \), in Eq 4.20 approaches one.

2. The \( p \)-values from the marker-in and marker-out LMMs were exactly the same as those from the baseline LMM when the variance parameters were known. There were negligible differences between \( p \)-values when the variance parameters were estimated. This highlights the effectiveness of the new scaling for obtaining the marker-out test statistics from the fit of the baseline LMM in addition to the marker-in test statistics.

3. Fitting a single baseline LMM was orders of magnitude faster than sequentially fitting \( m \) marker-in and marker-out LMMs (<3 seconds compared to >20 minutes). This highlights the efficiency of leveraging the GBLUP framework to obtain the marker-in and marker-out test statistics from the fit of the same baseline LMM, with the same genomic relationship matrix.

The methods were generalised for multiple sets of fixed and random effects, including complex residual variance structures. The methods will be extended for testing multiple sets of marker effects and modelling population structure in the remainder of this chapter. Fast exact methods for GWAS with complex genetic variance structures will be introduced in the next chapter.
4.9 Testing a set of markers at a time

This section extends the methods for testing a set of markers at a time, rather than just a single marker. The “set” is completely general and may represent a genome region, linkage group or chromosome. Assume the $m$ markers are partitioned into $s$ sets, such that $m_j$ is the number of markers in set $j$. Let the $v \times m$ marker design matrix be partitioned conformably as $M = [M_1 \ldots M_s]$, where $M_j = [m_1 \ldots m_{m_j}]$ is the $v \times m_j$ matrix for set $j$ and $m_i$ is the $v$-vector for marker $i$ in set $j$.

Three linear mixed models are considered below:

1. A baseline LMM which fits all marker effects as random.

2. An extension of the marker-in LMM which fits the marker effects in set $j$ as fixed and random.

3. An extension of the marker-out LMM which fits the marker effects in set $j$ as fixed but not random.

The difference between Models 2 and 3 is whether set $j$ is in or out of the polygenic term, so they are hereafter referred to as the set-in and set-out LMMs. Note that the baseline LMM is fitted only once, whereas the set-in and set-out LMMs are fitted $s$ times (once for each set).

It will be shown how the set-in and set-out test statistics can be obtained from the fit of the baseline LMM in Eq. 4.1.

4.9.1 Baseline linear mixed model

The baseline LMM is given by:

$$y = X\tau + ZMa + e,$$  \hspace{1cm} (4.38)

where all terms are given in Eq. 4.1. The vector of random marker effects can be partitioned as $a = (a_j^\top, a_{-j}^\top)^\top$, where $a_j$ is the $m_j$-vector for set $j$ and $a_{-j}$ is the $(m - m_j)$-vector for all remaining sets.

It is assumed that:

$$\begin{bmatrix} a \\ e \end{bmatrix} \sim N\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_a^2D & 0 \\ 0 & \sigma_e^2R \end{bmatrix} \right),$$  \hspace{1cm} (4.39)

where all terms are given in Eq. 4.2. The diagonal marker matrix is partitioned conformably.
4.9 Testing a set of markers at a time

with \( \mathbf{a} \) as:

\[
\mathbf{D} = \begin{bmatrix}
\mathbf{D}_j & \mathbf{0} \\
\mathbf{0} & \mathbf{D}_{-j}
\end{bmatrix},
\]

(4.40)

where \( \mathbf{D}_j \) is the \( m_j \times m_j \) matrix for set \( j \) and \( \mathbf{D}_{-j} \) is the \( (m - m_j) \times (m - m_j) \) matrix for all remaining sets.

It then follows that

\[
\mathbf{y} \sim \mathcal{N}(\mathbf{X} \tau, \mathbf{H}),
\]

(4.41)

where \( \mathbf{H} = \sigma_a^2 \mathbf{Z}_m \mathbf{D} \mathbf{Z}_m^\top + \sigma_e^2 \mathbf{R} \) and \( \mathbf{Z}_m = \mathbf{ZM} \).

The vector of BLUPs for set \( j \) is given by:

\[
\hat{\mathbf{a}}_j = \sigma_a^2 \mathbf{D}_j \mathbf{Z}_m^\top \mathbf{P} \mathbf{y},
\]

(4.42)

with prediction error variance matrix given by:

\[
\text{PEV}(\hat{\mathbf{a}}_j) = \sigma_a^2 \mathbf{D}_j - (\sigma_a^2)^2 \mathbf{D}_j \mathbf{Z}_m^\top \mathbf{P} \mathbf{Z}_m \mathbf{D}_j
\]

\[
= \sigma_a^2 \mathbf{D}_j - \text{var}(\hat{\mathbf{a}}_j),
\]

(4.43)

where \( \mathbf{P} = \mathbf{H}^{-1} - \mathbf{H}^{-1} \mathbf{X} (\mathbf{X}^\top \mathbf{H}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{H}^{-1} \) and \( \mathbf{Z}_m = \mathbf{ZM} \). Later in this section, the test statistics for all sets of interest will be obtained from the components in Eqs. 4.42 and 4.43.

4.9.2 Set-in linear mixed model

The set-in LMM is an extension of the marker-in LMM in Eq. 4.6 which fits the marker effects in set \( j \) as fixed, rather than just marker \( i \). The set-in LMM for set \( j \) is given by:

\[
\mathbf{y}_{in} = \mathbf{X} \tau_{in} + \mathbf{Z}_j \mu_{jin} + \mathbf{ZM} \mathbf{a}_{in} + \mathbf{e}_{in},
\]

(4.44)

where \( \tau_{in} \) is a vector of fixed effects, \( \mu_{jin} \) is the \( m_j \)-vector of fixed effects for set \( j \), \( \mathbf{a}_{in} \) is the \( m \)-vector of random marker effects including set \( j \) and \( \mathbf{e}_{in} \) is the \( n \)-vector of residuals. All other terms are given in Eq. 4.6.

It is assumed that:

\[
\begin{bmatrix}
\mathbf{a}_{in} \\
\mathbf{e}_{in}
\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}
\mathbf{0} \\
\mathbf{0}
\end{bmatrix}, \begin{bmatrix}
\sigma_a^2 \mathbf{D} & \mathbf{0} \\
\mathbf{0} & \sigma_e^2 \mathbf{R}
\end{bmatrix}\right),
\]

where all terms are given in Eq. 4.2. It is important to note that since the marker effects in
set \( j \) are fitted as fixed \textit{and} random, they contribute to the mean \textit{and} variance of \( y_{in} \), i.e. \( \beta_{jin} \sim N(\mu_{jin}, \sigma^2_a D_j) \), where \( \beta_{jin} = \mu_{jin} + a_{jin} \) is the vector of total effects for set \( j \). This doubling fitting has important ramifications on the estimates and subsequent test statistics.

It then follows that:

\[
y_{in} \sim N(\mathbf{X} \tau_{in} + \mathbf{Z}_m \mu_{jin}, \mathbf{H}) ,
\]

where \( \mathbf{H} \) is given in Eq. 4.41. The data vector \( y_{in} \) therefore has the same variance matrix as \( y \) but a different mean vector.

The vector of BLUEs for set \( j \) is given by:

\[
\hat{\mu}_{jin} = (\mathbf{Z}_m^T \mathbf{PZ}_m)^{-1} \mathbf{Z}_m^T \mathbf{Py}_{in},
\]

with prediction error variance matrix given by:

\[
\text{PEV}(\hat{\mu}_{jin}) = (\mathbf{Z}_m^T \mathbf{PZ}_m)^{-1}
= \text{var}(\mu_{jin}),
\]

where \( \mathbf{P} \) is given in Eq. 4.43. Since the marker effects in set \( j \) are also fitted as random, a vector of BLUPs can be obtained in addition to the vector of BLUEs. It can be shown that:

\[
\tilde{a}_{jin} = \mathbf{0} \quad \text{and} \quad \text{PEV}(\tilde{a}_{jin}) = \sigma^2_a D_j
= \sigma^2_a D_j - \text{var}(\tilde{a}_{jin}),
\]

where \( \text{var}(\tilde{a}_{jin}) = \mathbf{0} \), meaning that the vector of BLUPs for set \( j \) and its variance matrix become zero as a result of fitting the same set of marker effects as fixed. In fact, the two vectors are statistically aliased, so that after estimating \( \hat{\mu}_{jin} \) there is no information left to directly predict \( \tilde{a}_{jin} \). An indirect prediction can be obtained using the information in \( D \), which is why \( \text{var}(\tilde{a}_{jin}) = \mathbf{0} \) when \( D \) is diagonal. Note, however, this is not the case for non-diagonal forms of \( D \).

Lastly, the error covariance matrix between the BLUEs and BLUPs for set \( j \) is given by:

\[
\text{PEV}(\hat{\mu}_{jin}, \tilde{a}_{jin}) = -\sigma^2_a D_j.
\]

In the next section, a fundamental relationship between the set-in and set-out LMMs will be derived using the components in Eqs. 4.46 and 4.47.
4.9 Testing a set of markers at a time

4.9.3 Set-out linear mixed model

The set-out LMM is an extension of the marker-out LMM in Eq. 4.9 which fits the marker effects in set \( j \) as fixed, rather than just marker \( i \). Let the \( v \times ( m - m_j ) \) marker design matrix excluding set \( j \) be given by \( M_{-j} \). The set-out LMM for set \( j \) is given by:

\[
y_{\text{out}} = X\tau_{\text{out}} + Z_m\mu_{\text{out}j} + ZM_{-j}a_{\text{out}j} + e_{\text{out}},
\]

(4.48)

where \( \tau_{\text{out}} \) is a vector of fixed effects, \( \mu_{\text{out}j} \) is the \( m_j \)-vector of fixed effects for set \( j \), \( a_{\text{out}j} \) is the \( (m - m_j) \)-vector of random marker effects excluding set \( j \) and \( e_{\text{out}} \) is the \( n \)-vector of residuals. All other terms are given in Eq. 4.9.

It is assumed that:

\[
\begin{bmatrix}
a_{\text{out}j} \\
e_{\text{out}}
\end{bmatrix} \sim N\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_a^2 D_{-j} & 0 \\ 0 & \sigma_e^2 R \end{bmatrix} \right),
\]

where \( D_{-j} \) is the \( (m - m_j) \times (m - m_j) \) diagonal marker matrix excluding set \( j \) and all other terms are given in Eq. 4.2. It is important to note that the variance parameters are exactly the same as for the set-in LMM in Eq. 4.44, regardless of whether they are known or require estimation. This is because the fixed effects of set \( j \) “knock out” any potential contribution from the random effects anyway.

It then follows that:

\[
y_{\text{out}} \sim N(X\tau_{\text{out}} + Z_m\mu_{\text{out}j}, H_{-j}),
\]

(4.49)

where \( H_{-j} = H - \sigma_a^2 Z_m D_j Z_m^\top \). Note that since the marker effects in set \( j \) are fitted as fixed but not random, they contribute to the mean of \( y_{\text{out}} \) but not its variance. The data vector \( y_{\text{out}} \) therefore has a different mean vector and variance matrix compared to \( y \) in Eq. 4.41 and \( y_{\text{in}} \) in Eq. 4.45.

The vector of BLUES for set \( j \) are given by:

\[
\hat{\mu}_{\text{out}j} = (Z_{m_j}^\top P_{-j} Z_{m_j})^{-1} Z_{m_j}^\top P_{-j} y_{\text{out}},
\]

(4.50)

where \( P_{-j} = H_{-j}^{-1} - H_{-j}^{-1} X (X^\top H_{-j}^{-1} X)^{-1} X^\top H_{-j}^{-1} \).

Using Result 2 in Appendix E, it can be shown that:

\[
Z_{m_j}^\top P_{-j} = (I_{m_j} - \sigma_a^2 D_j Z_{m_j}^\top P Z_{m_j})^{-1} Z_{m_j}^\top P,
\]

(4.51)

where \( \Omega_j = I_{m_j} - \sigma_a^2 D_j Z_{m_j}^\top P Z_{m_j} \) is a \( m_j \times m_j \) scaling matrix, which is equivalent to
\( \Omega_j = \text{PEV}(\bar{a}_i)D_j^{-1}/\sigma_a^2 \) from the baseline LMM. Substituting Eq. 4.51 into Eq. 4.50 gives:

\[
\hat{\mu}_{j\text{ out}} = (Z_{m_j}^\top PZ_{m_j})^{-1}Z_{m_j}^\top Py_{\text{ out}},
\]

i.e. \( \hat{\mu}_{j\text{ out}} - \hat{\mu}_{j\text{ in}} = 0 \)

(4.52)

meaning that the vector of BLUEs for set \( j \) is invariant to whether the same set of marker effects is also fitted as random. Note, however, this is not the case for non-diagonal forms of \( D \). The result in Eq. 4.52 can be generalised as \( \hat{\mu}_{j\text{ out}} - (\hat{\mu}_{j\text{ in}} + \bar{a}_j) = 0 \), meaning that the BLUEs of the fixed effects from the set-in LMM are adjusted during estimation to make room for the BLUPs of the random effects.

The error variance is then given by:

\[
\text{PEV}(\hat{\mu}_{j\text{ out}}) = (Z_{m_j}^\top P - jZ_{m_j})^{-1}.
\]

(4.53)

Substituting Eq. 4.51 into Eq. 4.53 gives:

\[
\text{PEV}(\hat{\mu}_{j\text{ out}}) = (Z_{m_j}^\top PZ_{m_j})^{-1} - \sigma_a^2D_j = \text{PEV}(\hat{\mu}_{j\text{ in}}) - \sigma_a^2D_j,
\]

i.e. \( \text{var}(\hat{\mu}_{j\text{ out}}) = \text{var}(\hat{\mu}_{j\text{ in}}) - \sigma_a^2D_j \)

(4.54)

meaning that the variance matrix of the BLUEs is inflated by \( \sigma_a^2D_j \) when the same set of marker effects is also fitted as random. The result in Eq. 4.54 can be generalised as \( \text{var}(\hat{\mu}_{j\text{ out}}) = \text{var}(\hat{\mu}_{j\text{ in}}) + \text{PEV}(\bar{a}_j) + 2\text{PEV}(\hat{\mu}_{j\text{ in}}, \bar{a}_j) \), meaning that the variance matrix of the BLUEs from the set-in LMM is adjusted during estimation to make room for the variance matrix of the BLUPs. The scaling matrix can therefore be written as \( \Omega_j = \text{var}(\hat{\mu}_{j\text{ out}})\var(\hat{\mu}_{j\text{ in}})^{-1} \), which is equivalent to \( \Omega_j = \text{var}(\hat{\mu}_{j\text{ out}})[\text{var}(\hat{\mu}_{j\text{ out}}) + \sigma_a^2D_j]^{-1} \) and \( \Omega_j = [\text{var}(\hat{\mu}_{j\text{ in}}) - \sigma_a^2D_j]\text{var}(\hat{\mu}_{j\text{ in}})^{-1} \). These results will be utilised below.

### 4.9.4 Test statistics

The set-in and set-out test statistics are given by:

\[
\chi^2_{j\text{ in}} = \hat{\mu}_{j\text{ in}}^\top \text{var}(\hat{\mu}_{j\text{ in}})^{-1}\hat{\mu}_{j\text{ in}} \quad \text{and} \quad \chi^2_{j\text{ out}} = \hat{\mu}_{j\text{ out}}^\top \text{var}(\hat{\mu}_{j\text{ out}})^{-1}\hat{\mu}_{j\text{ out}},
\]

(4.55)

where \( \chi^2_{j\text{ in}} \) is the statistic for testing \( H_0: \mu_{j\text{ in}} = 0 \) and \( \chi^2_{j\text{ out}} \) is the statistic for testing...
4.10 Testing a single marker while excluding a set

\[ H_0 : \mu_{jout} = 0, \] which are assumed to have chi-square distributions with \( m_j \) degrees of freedom.

It can be shown that:

\[ \chi^2_{j in} \leq \chi^2_{j out}, \quad (4.56) \]

meaning that \( \chi^2_{j in} \) is shrunken compared to \( \chi^2_{j out} \) and the amount of shrinkage is proportional to \( \sigma^2_a D_j \). The difference between the two test statistics is therefore due to improper scaling of \( \chi^2_{j in} \) as a result of the aliasing. This difference will be important when the number of markers is small, but will become negligible when the number of markers is very large, i.e. when \( d_i \approx 0 \) and hence \( \Omega_j \approx I_{m_j} \).

It also follows that:

\[ \chi^2_{j out} = \hat{\mu}_{j in}^\top \left[ \text{var}(\hat{\mu}_{j in}) - \sigma^2_a D_j \right]^{-1} \hat{\mu}_{j in}, \quad (4.57) \]

since \( \hat{\mu}_{j out} - \hat{\mu}_{j in} = 0 \) and \( \text{var}(\hat{\mu}_{j out}) = \text{var}(\hat{\mu}_{j in}) - \sigma^2_a D_j \),

meaning that the set-out test statistics can be obtained from a simple rescaling of the set-in test statistics. This provides a general approach to obtain the set-out test statistics from the set-in LMM.

Obtaining the test statistics in Eq. 4.55 for all \( s \) sets is slow because it requires fitting the \( s \) models in Eq. 4.44 and 4.48. However, all required test statistics can be obtained directly from the fit of the baseline LMM in Eq. 4.38. The test statistics are now obtained as:

\[ \chi^2_{j in} = \tilde{a}_j^\top \text{var}(\bar{a}_j)^{-1} \bar{a}_j \quad \text{and} \quad \chi^2_{j out} = \bar{a}_j^\top \left[ \text{var}(\bar{a}_j) \Omega_j \right]^{-1} \bar{a}_j, \quad (4.58) \]

where \( \Omega_j = \text{PEV}(\bar{a}_i) D_j^{-1} / \sigma^2_a \). This provides a fast method to obtain all required test statistics from the fit of a single LMM. When the variance parameters are known, or when their estimation is unaffected by fitting the marker effects in set \( j \) as fixed, the test statistics in Eq. 4.58 are exactly the same as those in Eq. 4.55.

**4.10 Testing a single marker while excluding a set**

This section extends the methods for testing a single marker while excluding a set from the polygenic term. The set is again completely general but will be referred to as a “chromosome” to distinguish the methods developed here from the previous section. Assume the \( m \) markers are now partitioned into \( s \) chromosomes, such that chromosome \( j \) contains marker \( i \).
Two linear mixed models are considered below:

1. An extension of the marker-out LMM which fits marker effect $i$ as fixed and excludes all marker effects on chromosome $j$ from the random term.

2. A baseline LMM which fits all marker effects as random except those on chromosome $j$.

Model 1 has chromosome $j$ out of the polygenic term, so it is hereafter referred to as the chromosome-out LMM. Note that the chromosome-out LMM is fitted $m$ times (once for each marker), whereas the the baseline LMM is fitted only $s$ times (once for each chromosome). It will be shown how the chromosome-out test statistics can be obtained from the fit of the baseline LMM.

### 4.10.1 Chromosome-out linear mixed model

The chromosome-out LMM is an extension of the marker-out LMM in Eq. 4.9 which excludes chromosome $j$ from the polygenic term, rather than just marker $i$. Let the $v \times (m - m_j)$ marker design matrix excluding chromosome $j$ be given by $M_{-j}$, where $m_j$ is the number of markers on chromosome $j$. The chromosome-out LMM for marker $i$ on chromosome $j$ is given by:

$$y_{out} = X\tau_{out} + Z_m\mu_{iout} + ZM_{-j}a_{-jout} + e_{out}, \quad (4.59)$$

where $\tau_{out}$ is the vector of fixed effects, $\mu_{iout}$ is the fixed effect of marker $i$, $a_{-jout}$ is the $(m - m_j)$-vector of random marker effects excluding chromosome $j$ and $e_{out}$ is the $n$-vector of residuals. All other terms are given in Eq. 4.9.

It is assumed that:

$$\begin{bmatrix} a_{-jout} \\ e_{out} \end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_a^2 D_{-j} & 0 \\ 0 & \sigma_e^2 R \end{bmatrix} \right), \quad (4.60)$$

where $D_{-j}$ is the $(m - m_j) \times (m - m_j)$ diagonal marker matrix excluding chromosome $j$. The variance parameters are assumed to be unaffected by excluding chromosome $j$ from the polygenic term, but note that the results in this section hold regardless of whether they are affected or not.

It then follows that:

$$y_{out} \sim N(X\tau_{out} + Z_m\mu_{iout}, H_{-j}), \quad (4.61)$$

where $H_{-j} = H - \sigma_a^2 Z_m D_j Z_m^\top$ and $Z_m = ZM_j$. Note that since the marker effects on chromosome $j$ are excluded from the random polygenic term, they do not contribute to the
mean or variance of \( y_{out} \). The exception is marker effect \( i \), which contributes to the mean but not the variance.

The BLUE of marker effect \( i \) is given by:

\[
\hat{\mu}_{iout} = (Z_m^\top P_{-j}Z_m)^{-1}Z_m^\top P_{-j}y_{out},
\]

with prediction error variance given by:

\[
\text{PEV}(\hat{\mu}_{iout}) = (Z_m^\top P_{-j}Z_m)^{-1} = \text{var}(\hat{\mu}_{iout}),
\]

where \( P_{-j} = H_{-j}^{-1} - H_{-j}^{-1}X(X^\top H_{-j}^{-1}X)^{-1}X^\top H_{-j}^{-1} \).

The chromosome-out test statistic is then given by:

\[
z_{iout} = \frac{\hat{\mu}_{iout}}{\sqrt{\text{var}(\hat{\mu}_{iout})}},
\]

where \( z_{iout} \) is the statistic for testing \( H_0 : \mu_{iout} = 0 \), which is assumed to have a standard normal distribution. Obtaining the test statistics in Eq. 4.63 for all \( m \) markers is slow because it requires fitting the \( m \) models in Eq. 4.59. However, the test statistics can be obtained from the fit of just \( s \) baseline LMMs, where \( s \ll m \) is the number of chromosomes.

### 4.10.2 Baseline linear mixed model

The baseline LMM is an extension of Eq. 4.1 which excludes chromosome \( j \) from the random marker effects. The baseline LMM for chromosome \( j \) is given by:

\[
y = X\tau + ZM_{-j}a_{-j} + e,
\]

where \( a_{-j} \) is the \((m - m_j)\)-vector of random marker effects excluding chromosome \( j \). All other terms are given in Eq. 4.1.

It is assumed that:

\[
\begin{bmatrix}
a_{-j} \\
e
\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}
0 \\
0
\end{bmatrix}, \begin{bmatrix}
\sigma_a^2 D_{-j} & 0 \\
0 & \sigma_e^2 R
\end{bmatrix}\right),
\]

where all terms are given in Eq. 4.60.
It then follows that:
\[ y \sim N(\mathbf{X}\tau, \mathbf{H}_{-j}) , \]

where \( \mathbf{H}_{-j} \) is given in Eq. 4.61. The data vector \( y \) therefore has the same variance matrix as \( y_{out} \) but a different mean vector.

The vector of BLUPs for chromosome \( j \) is given by:
\[ \tilde{\mathbf{a}}_{-j} = \sigma_a^2 \mathbf{D}_{-j} \mathbf{Z}^\top \mathbf{P}_{-j} y , \]

with prediction error variance matrix given by:
\[
\text{PEV}(\tilde{\mathbf{a}}_{-j}) = \sigma_a^2 \mathbf{D}_{-j} - \left( \sigma_a^2 \right)^2 \mathbf{D}_{-j} \mathbf{Z}^\top \mathbf{P}_{-j} \mathbf{Z} \mathbf{D}_{-j} \\
= \sigma_a^2 \mathbf{D}_{-j} - \text{var}(\tilde{\mathbf{a}}_{-j}) ,
\]

where \( \mathbf{P}_{-j} \) is given in Eq. 4.62.

An indirect prediction of marker effect \( i \) can also be obtained. Since chromosome \( j \) contains marker \( i \), it follows that:
\[
\begin{bmatrix}
    a_i \\
    \mathbf{a}_{-j}
\end{bmatrix}
\sim N \left( 
\begin{bmatrix}
    \mathbf{0} \\
    \mathbf{0}
\end{bmatrix} , \sigma_a^2 \begin{bmatrix}
    d_i & 0 \\
    0 & \mathbf{D}_{-j}
\end{bmatrix} \right) ,
\]

(4.65)

meaning that the BLUP of marker effect \( i \) is obtained as \( \tilde{a}_i = 0 \), with \( \text{var}(\tilde{a}_i) = 0 \). However, it is important to note that a prediction of \( a_i \) commensurate with the assumptions in Eq. 4.65 is not actually required. What is required, is a prediction that produces the chromosome-out test statistic in Eq. 4.63. Importantly, the test statistic can be constructed from other components in the baseline LMM.

The required BLUP of marker effect \( i \) is constructed as:
\[
\tilde{a}_i^* = d_i \mathbf{m}_i^\top \mathbf{D}_{-j}^{-1} \tilde{\mathbf{a}}_{-j} \\
= \sigma_a^2 d_i \mathbf{Z}_m^\top \mathbf{P}_{-j} y ,
\]

(4.66)

with prediction error variance constructed as:
\[
\text{PEV}(\tilde{a}_i^*) = \sigma_a^2 d_i - \sigma_a^2 d_i^2 \mathbf{m}_i^\top \mathbf{D}_{-j}^{-1} \mathbf{m}_i + d_i^2 \mathbf{m}_i^\top \mathbf{D}_{-j}^{-1} \text{PEV}(\tilde{\mathbf{a}}_{-j}) \mathbf{D}_{-j}^{-1} \mathbf{m}_i \\
= \sigma_a^2 d_i - \left( \sigma_a^2 d_i \right)^2 \mathbf{Z}_m^\top \mathbf{P}_{-j} \mathbf{Z}_m \\
= \sigma_a^2 d_i - \text{var}(\tilde{a}_i^*) ,
\]
and variance constructed as:

\[
\text{var}(\tilde{a}_i^*) = d_i^2 m_i^\top D_{-j}^{-1} \text{var}(\tilde{a}_{-j}) D_{-j}^{-1} m_i
\]

\[
= (\sigma_a^2 d_i)^2 Z_m^\top P_{-j} Z_m.
\]

(4.67)

The chromosome-out test statistic is then obtained by dividing the BLUP in Eq. 4.66 by the square root of the variance in Eq. 4.67, which gives:

\[
z_{i_{\text{out}}} = \frac{\tilde{a}_i^*}{\sqrt{\text{var}(\tilde{a}_i^*)}}.
\]

(4.68)

This provides a fast method to obtain all required test statistics from the fit of just \(s\) models, where \(s\) is the number of chromosomes. When the variance parameters are known, or when their estimation is unaffected by fitting marker effect \(i\) as fixed, the test statistics in Eq. 4.68 are exactly the same as those in Eq. 4.63.

### 4.11 Modelling population structure

This section extends the methods for modelling population structure in the data. Population structure is often modelled by fitting the first \(k\) principal components in the genomic relationship matrix as fixed. Let the genomic relationship matrix be written as:

\[
G = ULU^\top,
\]

(4.69)

where \(U\) is a \(v \times v\) orthonormal matrix with columns given by the eigenvectors of \(G\) and \(L\) is a \(v \times v\) diagonal matrix with diagonal elements given by the eigenvalues of \(G\) sorted in decreasing order. The first \(k\) principal components in \(G\) can be fitted as fixed by setting \(X = [X_1 \ ZU_k]\) and \(\tau = (\tau_1^\top, \tau_k^\top)^\top\) in Eq 4.1, where \(U_k\) is given by the first \(k\) columns in \(U\). The baseline LMM can therefore be rewritten as:

\[
y = X_1 \tau_1 + ZU_k \tau_k + ZMa + e,
\]

(4.70)

where all terms are given in Eq 4.1. When the variance parameters are known, or when their estimation is unaffected by fitting marker effect \(i\) as fixed, the marker-in and marker-out test statistics obtained from the baseline LMM in Eq. 4.70 are exactly the same as those obtained from extensions of the marker-in and marker-out LMMs with the same \(k\) principal components fitted as fixed.
4.12 Concluding remarks

This chapter developed fast exact methods for GWAS which produce all required test statistics from the fit of a single linear mixed model. The methods were built on an existing approach for GWAS which leverages the efficiency of the GBLUP framework, i.e. GWAS by GBLUP. The objective here was to extend that approach for obtaining the marker-out test statistics, in addition to the marker-in test statistics. The objective was traditionally addressed by sequentially fitting a very large number of linear mixed models, with a different genomic relationship matrix constructed for each model without the marker of interest. The methods developed in this chapter enable the marker-in and marker-out test statistics to be obtained from the fit of the same linear mixed model, with the same genomic relationship matrix.

The chapter also addressed an important question in the literature; should the marker effect of interest be included in the polygenic term? It was shown algebraically that double fitting the marker effect as fixed and random causes the two effects to be statistically aliased, which causes the test statistics to be shrunken by the polygenic variance. A fundamental expression was derived which provides an efficient way to obtain the marker-out test statistics directly from the fit of the marker-in LMM, and vice versa. This provides a simple fix for current software packages employing the marker-in approach.

The GWAS by GBLUP approach was generalised for any set of fixed and random effects, including modelling population structure and complex residual covariance structures. The approach was also extended for testing a set of markers at a time and testing a single marker while excluding a set from the polygenic term. The set was completely general and may represent a region, linkage group or chromosome containing the marker/s of interest.

The methods and concepts developed in this chapter have great potential to improve biological discovery in a wide-array of genetic studies, particularly with the advent of large-scale phenotypic and genotypic datasets. The methods provide a general framework for conducting GWAS which leverages the efficiency of GBLUP and the flexibility of linear mixed models. This provides a solid basis to introduce complex models for GEI, which is the focus of the next chapter.

Areas of active research include efficient methods for simultaneously testing multiple sets, removing population structure from the genomic relationship matrix and re-estimating variance parameters with complex genetic and residual variance structures.
Chapter 5

Association studies II

5.1 Prelude

This chapter contains the unpublished manuscript *Genome-wide association studies built on multiplicative models for genotype by environment interaction* by Tolhurst et al. (2024b). Methods are developed that combine phenotypic, genotypic and environmental data to elucidate complex phenotype to genotype to environment associations underlying complex traits such as genotype adaptation and disease resistance in plants. The chapter builds on the fast exact methods for GWAS in Chapter 4 using the genomic prediction models in Chapter 2. Three models for GEI are considered:

1. Compound symmetry, which involves testing the contribution of each marker to the average genotype response across all environments and to the specific genotype response within individual environments.

2. Random regression, which involves testing the contribution of each marker to the average genotype response across all environments and to the specific genotype response to individual environmental covariates.

3. Factor analytic, which involves testing the contribution of each marker to the specific genotype response to individual environmental factors.

Methods are also developed that produce statistics for testing the contribution of each marker to the selection tools in Chapter 3, which include measures of overall performance, responsiveness and stability for each genotype.

The methods developed in this chapter have great potential to improve discovery across a wide-array of genetic studies, particularly with the advent of large-scale datasets and complex genotype by environment interactions.
5.2 Linear mixed models

Assume a MET dataset comprises \( n \) records on \( v \) genotypes with \( m \) markers and \( p \) environments, where \( m \geq v \) and often \( n \leq vp \). Let the \( n \)-vector of phenotypic data be given by \( y = (y_1^\top, \ldots, y_p^\top)^\top \), where \( y_j \) is the \( n_j \)-vector for environment \( j \). Also let the \( v \times m \) marker design matrix be given by \( M = [m_1 \ldots m_m] \), where \( m_i \) is the \( v \)-vector of centred genotype scores for marker \( i \). The methods developed in this chapter involve fitting and testing marker effects inherent to different models for GEI, such as marker main effects and marker by environment interaction effects or marker intercepts and slopes.

Three linear mixed models are considered below:

1. A baseline LMM which fits all marker effects as random.
2. An extension of the baseline LMM which fits the marker effect of interest as fixed and random.
3. An extension of the baseline LMM which fits the marker effect of interest as fixed but not random.

The difference between Models 2 and 3 is whether the marker effect of interest is in or out of the polygenic term, so they are referred to as the marker-in and marker-out LMMs following Chapter 4. For brevity, the results in this chapter focus on the baseline and marker-out LMMs. Methods to obtain the marker-in test statistics from the marker-out LMM will be highlighted in text. Note that the baseline LMM is only fitted once, whereas the marker-out LMM is fitted at least \( m \) times (once for each marker) - the exact number will depend on the type of marker effects being tested.

5.3 Preliminary baseline linear mixed model

The preliminary baseline LMM is given by:

\[
y = X\tau + Z(I_p \otimes M)a + e ,
\]

where \( \tau \) is a vector of fixed effects with design matrix \( X \), \( a \) is the \( mp \)-vector of random marker by environment (ME) effects with \( v \times m \) marker design matrix \( M \) and \( n \times vp \) genotype design matrix \( Z \) and \( e \) is the \( n \)-vector of residuals. The fixed effects include the overall mean and environmental main effects, while the residuals may include random non-additive and non-genetic effects. The methods developed in this chapter are applicable to any set of fixed and random effects.
It assumed that:

\[
\begin{bmatrix}
    \mathbf{a} \\
    \mathbf{e}
\end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix}
    0 \\
    0
\end{bmatrix}, \begin{bmatrix}
    \mathbf{G}_a \otimes \mathbf{D} & 0 \\
    0 & \mathbf{R}
\end{bmatrix} \right),
\]

(5.2)

where \( \mathbf{G}_a \otimes \mathbf{D} \) is a \( mp \times mp \) marker by environment additive genetic variance matrix and \( \mathbf{R} \) is a \( n \times n \) residual variance matrix, which is assumed to be completely general.

The additive genetic variance matrix is given by:

\[
\text{var}(\mathbf{a}) = \mathbf{G}_a \otimes \mathbf{D},
\]

where \( \mathbf{G}_a \) is a \( p \times p \) between-environment variance matrix and \( \mathbf{D} \) is a \( m \times m \) diagonal marker matrix with diagonal elements given by \( d_i \) (see Chapters 2, 3 and 4). The preliminary baseline LMM in Eq. 5.1 can be updated based on the structure for \( \mathbf{G}_a \).

Three structures for \( \mathbf{G}_a \) are considered below:

1. Compound symmetry
2. Random regression
3. Factor analytic.

The form of the marker-out LMM depends on the structure for \( \mathbf{G}_a \) and the type of marker effects being tested, e.g. whether main effects or interaction effects are being tested. The relevant baseline and marker-out LMMs are derived for each structure below.

### 5.4 Compound symmetry model

The compound symmetry model assumes all environments have the same genetic variance and all pairs of environments have the same genetic covariance, i.e. it assumes a completely homogeneous model for GEI. This model was previously demonstrated for the example dataset in Chapter 2 as a basis for comparing different genomic prediction models.

The compound symmetry model is given by:

\[
\mathbf{a} = (\mathbf{1}_p \otimes \mathbf{a}_1) + \mathbf{a}_2,
\]

(5.3)

where \( \mathbf{a}_1 = (a_{11}, \ldots, a_{1r})^\top \) is the \( m \)-vector of marker main effects and \( \mathbf{a}_2 = (\mathbf{a}_2^\top_1, \ldots, \mathbf{a}_2^\top_p)^\top \) is the \( mp \)-vector of marker by environment interaction effects, which are specific to individual environments. The vector \( \mathbf{a} \) is hereafter referred to as the vector of total ME effects.
It is assumed that:
\[
\begin{bmatrix}
a_1 \\
a_2
\end{bmatrix} \sim N\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_1 & 0 \\ 0 & \sigma^2_2 \end{bmatrix} \otimes D \right),
\]
(5.4)
where \( \sigma^2_1 \) is the main effect variance and \( \sigma^2_2 \) is the interaction variance.

It then follows that:
\[\text{var}(a) = (\sigma^2_1 I_p + \sigma^2_2 I_p) \otimes D,\]
where \( G_a = \sigma^2_1 I_p + \sigma^2_2 I_p \) is the between-environment variance matrix.

### 5.4.1 Baseline linear mixed model

The baseline LMM for the compound symmetry model is obtained by substituting Eq. 5.3 into Eq. 5.1, which gives:
\[
y = X\tau + Z(I_p \otimes M) a_1 + Z(I_p \otimes M) a_2 + e,
\]
(5.5)
where all terms are given in Eq. 5.1.

It then follows that:
\[
y \sim N(X\tau, H),
\]
(5.6)
where \( H = Z_m(G_a \otimes D)Z_m^\top + R \) and \( Z_m = Z(I_p \otimes M) \). Standard results for the random effects give \( \tilde{a}_1 = (\sigma^2_1 I_p \otimes D)Z_m^\top P\tilde{y} \) and \( \tilde{a}_2 = (\sigma^2_2 I_p \otimes D)Z_m^\top P\tilde{y} \).

The BLUP of marker main effect \( i \) is then given by:
\[
\tilde{a}_{1i} = \sigma^2_1 d_i Z_{m_{1i}}^\top P\tilde{y} \quad \text{and} \quad \text{PEV}(\tilde{a}_{1i}) = \sigma^2_1 d_i - (\sigma^2_1 d_i)^2 Z_{m_{1i}}^\top PZ_{m_{1i}} = \sigma^2_1 d_i - \text{var}(\tilde{a}_{1i}),
\]
(5.7)
where \( P = H^{-1} - H^{-1}X(X^\top H^{-1}X)^{-1}X^\top H^{-1} \) and \( Z_{m_{1i}} = Z(I_p \otimes m_i) \).

The vector of BLUPs for the interaction effects involving marker \( i \) is given by:
\[
\tilde{a}_{2i} = \sigma^2_2 d_i Z_{m_{2i}}^\top P\tilde{y} \quad \text{and} \quad \text{PEV}(\tilde{a}_{2i}) = \sigma^2_2 d_i - (\sigma^2_2 d_i)^2 Z_{m_{2i}}^\top PZ_{m_{2i}} = \sigma^2_2 d_i - \text{var}(\tilde{a}_{2i}),
\]
(5.8)
where \( Z_{m_{2i}} = Z(I_p \otimes m_i) \).
Lastly, the vector of BLUPs for the total effects of marker $i$ is given by:

$$\tilde{a}_i = d_i G_a Z_{m_2}^T P y$$

and

$$\text{PEV}(\tilde{a}_i) = d_i G_a - d_i^2 G_a Z_{m_2}^T P Z_{m_2} G_a$$

$$= d_i G_a - \text{var}(\tilde{a}_i),$$

where $G_a = \sigma_1^2 J_p + \sigma_2^2 I_p$. In the next section, the test statistics for all marker effects of interest will be obtained from the components in Eqs. 5.7 and 5.8.

### 5.4.2 Marker-out linear mixed model

The marker-out LMM is an extension of Eq. 4.9 which fits a compound symmetry model to the polygenic term. The objective of this model is to sequentially fit and test the main effect of each marker or the interaction effects involving each marker.

Two linear mixed models are considered below:

1. An extension of the marker-out LMM which fits the main effect of marker $i$ as fixed but not random.

2. An extension of the set-out LMM which fits the main effect of marker $i$ and all interaction effects involving marker $i$ as fixed but not random.

The difference between Models 1 and 2 is whether a single marker effect or a set of marker effects are fitted and tested, so they are hereafter referred to as the **marker-out** and **set-out** LMMs. Note that the marker-out and set-out LMMs are each fitted $m$ times (once for each marker), whereas the baseline LMM in Eq. 5.5 is fitted just once. It will be shown how the test statistics for both models can be obtained directly from the fit of the baseline LMM.

**Testing marker main effects**

The marker-out LMM is an extension of Eq. 4.9 which fits the main effect of marker $i$ as fixed but not random in the polygenic term. Let the $v \times (m-1)$ marker design matrix excluding column $i$ be given by $M_{-i}$. The marker-out LMM for marker $i$ is given by:

$$y_{out} = X \tau_{out} + Z_{m_{i out}} \mu_{i out} + Z(I_p \otimes M_{-i}) a_{1-iout} + Z(I_p \otimes M) a_{2out} + e_{out}, \quad (5.9)$$

where $\tau_{out}$ is a vector of fixed effects, $\mu_{i out}$ is the fixed main effect of marker $i$, $a_{1-iout}$ is the $m-1$ vector of random main effects excluding marker $i$, $a_{2out}$ is the $mp$-vector of random interaction effects including marker $i$ and $e_{out}$ is the $n$-vector of residuals. All other terms are given in Eq. 5.1.
It is assumed that:

\[
\begin{bmatrix}
a_{1-out} \\
a_{2-out}
\end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_i^2 D_{-i} & 0 \\ 0 & \sigma_i^2 I_p \otimes D \end{bmatrix} \right),
\]

where \( D_{-i} \) is the \((m-1) \times (m-1)\) diagonal marker matrix excluding marker \( i \).

It then follows that:

\[
y_{out} \sim N \left( X \tau_{out} + Z_{m_{1i}} \mu_{1iout}, H_{-1i} \right),
\]

(5.10)

where \( H_{-1i} = H - \sigma_i^2 d_i Z_{m_{1i}} Z_{m_{1i}}^T \). Note that since the main effect of marker \( i \) is fitted as fixed but not random, it contributes to the mean of \( y_{out} \) but not its variance. The data vector \( y_{out} \) therefore has a different mean vector and variance matrix compared to \( y \) in Eq. 5.6.

The BLUE of marker main effect \( i \) is given by:

\[
\hat{\mu}_{1iout} = (Z_{m_{1i}} P_{-1i} Z_{m_{1i}})^{-1} Z_{m_{1i}} P_{-1i} y_{out},
\]

(5.11)

where \( P_{-1i} = H_{-1i}^{-1} - H_{-1i}^{-1} X (X^T H_{-1i}^{-1} X)^{-1} X^T H_{-1i}^{-1} \).

Using Result 2 in Appendix E, it can be shown that:

\[
Z_{m_{1i}}^T P_{-1i} = \frac{Z_{m_{1i}}^T P}{1 - \sigma_i^2 d_i Z_{m_{1i}}^T P Z_{m_{1i}}},
\]

(5.12)

where \( \omega_i = 1 - \sigma_i^2 d_i Z_{m_{1i}}^T P Z_{m_{1i}} \) defines a scaling constant, which is equivalent to \( \omega_i = \text{PEV}(\tilde{a}_{1i}) / \sigma_i^2 d_i \) from the baseline LMM. Substituting Eq. 5.12 into Eq. 5.11 gives:

\[
\hat{\mu}_{1iout} = (Z_{m_{1i}} P Z_{m_{1i}})^{-1} Z_{m_{1i}}^T P y_{out}.
\]

The error variance is then given by:

\[
\text{PEV}(\hat{\mu}_{1iout}) = (Z_{m_{1i}} P_{-1i} Z_{m_{1i}})^{-1}.
\]

(5.13)

Substituting Eq. 5.11 into Eq. 5.13 gives:

\[
\text{PEV}(\hat{\mu}_{1iout}) = (Z_{m_{1i}} P Z_{m_{1i}})^{-1} - \sigma_i^2 d_i = \text{var}(\hat{\mu}_{1iout}).
\]

The marker-in LMM is similar to the marker-out LMM in Eq. 5.9, with the key difference
that the main effect of marker \(i\) is included in the polygenic term. Using the results from Chapter 4, it can be shown that:

\[
\hat{\mu}_{1iin} - \hat{\mu}_{1iout} = 0 \quad \text{and} \quad \text{PEV}(\hat{\mu}_{1iin}) = \text{PEV}(\hat{\mu}_{1iout}) + \sigma_i^2 d_i = \text{var}(\hat{\mu}_{1iin}),
\]

(5.14)

where \(\hat{\mu}_{1iin}\) is the BLUE of marker main effect \(i\) from the marker-in LMM. This provides an efficient way to obtain the marker-in components from the fit of the marker-out LMM. Note, however, it is more practical to fit the marker-in LMM and then obtain the marker-out components from that model. This is because fitting the marker-in LMM elevates the need to construct a separate genomic relationship for each marker of interest.

**Test Statistics**

The marker-out test statistic is given by:

\[
z_{1iout} = \frac{\hat{\mu}_{1iout}}{\sqrt{\text{var}(\hat{\mu}_{1iout})}},
\]

(5.15)

where \(z_{1iout}\) is the statistic for testing \(H_0: \mu_{1iout} = 0\), which is assumed to have a standard normal distribution. The marker-in test statistic can be obtained from the components in Eq. 5.14, or by rescaling the marker-out test statistic as:

\[
z_{1iin} = \sqrt{\omega_{1i} z_{1iout}},
\]

(5.16)

where \(\omega_{1i} = \text{var}(\hat{\mu}_{1iout})/[\text{var}(\hat{\mu}_{1iout}) + \sigma_i^2 d_i]\), which is equivalent to the scaling constant in Eq. 5.12. The marker-in test statistic is therefore shrunken compared to the marker-out test statistic. This difference will be important when the number of markers is small, but will become negligible when the number of markers is very large, i.e. when \(\omega_{1i} \approx 1\).

Obtaining the test statistics in Eqs. 5.15 and 5.16 for all \(m\) markers is slow because it requires fitting \(m\) linear mixed models. However, all required test statistics can be obtained directly from the fit of the baseline LMM in Eq. 5.5, with compound symmetry model given in Eq. 5.3. The test statistics are now obtained as:

\[
z_{1iin} = \frac{\tilde{a}_{1i}}{\sqrt{\text{var}(\tilde{a}_{1i})}} \quad \text{and} \quad z_{1iout} = \frac{\tilde{a}_{1i}}{\sqrt{\omega_{1i} \text{var}(\tilde{a}_{1i})}},
\]

(5.17)

where \(\omega_{1i} = \text{PEV}(\tilde{a}_{1i})/\sigma_i^2 d_i\). This provides a fast method to obtain all required test statistics from the fit of a single LMM. When the variance parameters are known, or when their
estimation is unaffected by fitting the main effect of marker $i$ as fixed, the test statistics in Eq. 5.17 are exactly the same as those in Eqs. 5.15 and 5.16. This is a trivial extension of the methods developed in Chapter 4.

**Testing marker by environment interaction effects**

The set-out LMM is an extension of Eq. 4.48 which fits the main effect of marker $i$ and all interaction effects involving marker $i$ as fixed but not random in the polygenic term. Let the $p$-vector of fixed total effects for marker $i$ be given by:

$$
\mu_{i\text{out}}^* = (1_p \otimes \mu_{1\text{out}}) + \mu_{2\text{out}}, \tag{5.18}
$$

where $\mu_{i\text{out}}$ is the fixed main effect of marker $i$ and $\mu_{2\text{out}}$ is the $p$-vector of fixed interaction effects involving marker $i$. When $\mu_{i\text{out}}$ and $\mu_{2\text{out}}$ are both fitted as fixed, the joint design matrix becomes less than full rank and the effects must be constrained to enable estimation (see Smith, 1999). This creates non-estimable functions underlying the fixed effects, which prevents testing the main effects and interaction effects separately. The main effects and interaction effects can be tested together by constructing $\mu_{2\text{out}}^*$ after estimation, however, it is more efficient to fit and test $\mu_{2\text{out}}^*$ directly. This approach produces exactly the same results but avoids the issues regarding estimability.

The marker-out LMM for marker $i$ is therefore given by:

$$
y_{\text{out}} = X\tau_{\text{out}} + Z_{m_2}^i \mu_{2\text{out}}^* + Z(1_p \otimes M_{-i}) a_{1-i\text{out}} + Z(1_p \otimes M_{-i}) a_{2-i\text{out}} + e_{\text{out}}, \tag{5.19}
$$

where $\tau_{\text{out}}$ is a vector of fixed effects, $\mu_{2\text{out}}^*$ is the $p$-vector of fixed total effects for marker $i$, $a_{1-i\text{out}}$ is the $m-1$ vector of random main effects excluding marker $i$, $a_{2-i\text{out}}$ is the $mp-p$ vector of random interaction effects excluding marker $i$ and $e_{\text{out}}$ is the $n$-vector of residuals. All other terms are given in Eq. 5.1.

It is assumed that:

$$
\begin{bmatrix}
a_{1-i\text{out}} \\
a_{2-i\text{out}}
\end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 D_{-i} & 0 \\ 0 & \sigma_2^2 I_p \otimes D_{-i} \end{bmatrix}\right),
$$

where $D_{-i}$ is the $(m-1) \times (m-1)$ diagonal marker matrix excluding marker $i$.

It then follows that:

$$
y_{\text{out}} \sim N\left(X\tau_{\text{out}} + Z_{m_2}^i \mu_{2\text{out}}^*, H_{-2i}\right),
$$

where $H_{-2i} = H - d_i Z_{m_2} G_a Z_{m_2}^\top$ and $G_a = \sigma_1^2 J_p + \sigma_2^2 I_p$. Note that since the total effects
of marker $i$ are fitted as fixed but not random, they contribute to the mean of $\mathbf{y}_{\text{out}}$ but not its variance. The data vector $\mathbf{y}_{\text{out}}$ therefore has a different mean vector and variance matrix compared to $\mathbf{y}$ in Eq. 5.6.

The vector of BLUEs for the total effects of marker $i$ is given by:

$$
\hat{\mu}^{\ast}_{2i_{\text{out}}} = (\mathbf{Z}_{m_{2i}}^\top \mathbf{P}^{-1} \mathbf{Z}_{m_{2i}})^{-1} \mathbf{Z}_{m_{2i}}^\top \mathbf{P}^{-1} \mathbf{y}_{\text{out}},
$$

(5.20)

where $\mathbf{P}^{-1} = \mathbf{H}^{-1} - \mathbf{H}^{-1} \mathbf{X} (\mathbf{X}^\top \mathbf{H}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{H}^{-1}$. Using Result 2 in Appendix E, it can be shown that:

$$
\mathbf{Z}_{m_{2i}}^\top \mathbf{P}^{-1} \mathbf{Z}_{m_{2i}} = \mathbf{I}_{p} - d_i \mathbf{G}_a \mathbf{Z}_{m_{2i}}^\top \mathbf{P} \mathbf{Z}_{m_{2i}}^{-1} \mathbf{P} \mathbf{Z}_{m_{2i}}^\top, \tag{5.21}
$$

where $\mathbf{\Omega}_{2i} = \mathbf{I}_{p} - d_i \mathbf{G}_a \mathbf{Z}_{m_{2i}}^\top \mathbf{P} \mathbf{Z}_{m_{2i}}^{-1} \mathbf{P}$, which is equivalent to $\mathbf{\Omega}_{2i} = \text{PEV}(\tilde{\mathbf{a}}_i) \mathbf{G}_a^{-1} / d_i$ from the baseline LMM. Substituting Eq. 5.21 into Eq. 5.20 gives:

$$
\hat{\mu}^{\ast}_{2i_{\text{out}}} = (\mathbf{Z}_{m_{2i}}^\top \mathbf{P} \mathbf{Z}_{m_{2i}})^{-1} \mathbf{Z}_{m_{2i}}^\top \mathbf{P} \mathbf{y}_{\text{out}}.
$$

The error variance is then given by:

$$
\text{PEV}(\hat{\mu}^{\ast}_{2i_{\text{out}}}) = (\mathbf{Z}_{m_{2i}}^\top \mathbf{P}^{-1} \mathbf{Z}_{m_{2i}})^{-1}.
$$

(5.22)

Substituting Eq. 5.21 into Eq. 5.22 gives:

$$
\text{PEV}(\hat{\mu}^{\ast}_{2i_{\text{out}}}) = (\mathbf{Z}_{m_{2i}}^\top \mathbf{P} \mathbf{Z}_{m_{2i}})^{-1} - d_i \mathbf{G}_a \\
= \text{var}(\hat{\mu}^{\ast}_{2i_{\text{out}}}).
$$

It is important to note that $\hat{\mu}^{\ast}_{2i_{\text{out}}}$ obtained here is exactly the same as when the separate terms in Eq. 5.18 are fitted and then summed, i.e. $\hat{\mu}^{\ast}_{2i_{\text{out}}} = (\mathbf{1}_p \otimes \hat{\mathbf{\mu}}_{1_{\text{out}}}) + \hat{\mu}^{\ast}_{2i_{\text{out}}}$. This also holds for their respective error variance matrices.

The marker-in LMM is similar to the marker-out LMM in Eq. 5.19, with the key difference that the effects of marker $i$ are included in the polygenic term. Using the results from Chapter 4, it can be shown that:

$$
\hat{\mu}^{\ast}_{2i_{\text{in}}} - \hat{\mu}^{\ast}_{2i_{\text{out}}} = 0 \quad \text{and} \quad \text{PEV}(\hat{\mu}^{\ast}_{2i_{\text{in}}}) = \text{PEV}(\hat{\mu}^{\ast}_{2i_{\text{out}}}) + d_i \mathbf{G}_a \\
= \text{var}(\hat{\mu}^{\ast}_{2i_{\text{in}}}),
$$

(5.23)

where $\hat{\mu}^{\ast}_{2i_{\text{in}}}$ is the $p$-vector of BLUEs for the total effects of marker $i$ from the marker-in LMM. This provides an efficient way to obtain the marker-in components from the fit of the
marker-out LMM. Note, however, it is more practical to fit the marker-in LMM and then obtain the marker-out components from that model. This is because fitting the marker-in LMM elevates the need to construct a separate genomic relationship for each marker.

**Test Statistics**

The marker-out test statistic is given by:

\[
\chi^2_{2iout} = \hat{\mu}_{2iout}^\top \text{var}(\hat{\mu}_{2iout})^{-1} \hat{\mu}_{2iout},
\]

where \(\chi^2_{2iout}\) is the statistic for testing \(H_0 : \mu_{2iout} = 0\), which is assumed to have a chi-square distribution with \(p\) degrees of freedom. The marker-in test statistic can be obtained from the components in Eq. 5.23, or by rescaling the marker-out test statistic as:

\[
\chi^2_{2iin} = \hat{\mu}_{2iout}^\top \left[ \text{var}(\hat{\mu}_{2iout}) + d_i G_a \right]^{-1} \hat{\mu}_{2iout},
\]

where \(\chi^2_{2iin}\) is the statistic for testing \(H_0 : \mu_{2iin} = 0\), which is also assumed to have a chi-square distribution with \(p\) degrees of freedom. The marker-in test statistic is again shrunken compared to the marker-out test statistic. This difference will be important when the number of markers is small, but will become negligible when the number of markers is very large, i.e. when \(\Omega_2i \approx I_p\).

Obtaining the test statistics in Eqs. 5.24 and 5.25 for all \(m\) markers is slow because it requires fitting \(m\) linear mixed models. However, all required test statistics can be obtained directly from the fit of the baseline LMM in Eq. 5.5, with compound symmetry model given in Eq. 5.3. The test statistics are now obtained as:

\[
\chi^2_{2iin} = \bar{a}_2i^\top \text{var}(\bar{a}_2i)^{-1} \bar{a}_2i \quad \text{and} \quad \chi^2_{2iout} = \bar{a}_2i^\top \left[ \text{var}(\bar{a}_2i) \Omega_{2i} \right]^{-1} \bar{a}_2i,
\]

where \(\Omega_{2i} = \text{PEV}(\bar{a}_i) G_a^{-1}/d_i\). This provides a fast method to obtain all required test statistics from the fit of a single LMM. When the variance parameters are known, or when their estimation is unaffected by fitting the total effects of marker \(i\) as fixed, the test statistics in Eq. 5.26 are exactly the same as those in Eqs. 5.24 and 5.25. Test statistics for the total effects in each environment can also be obtained where required. This is a practical extension of the methods developed in Chapter 4.
5.5 Random regression model

Assume all $p$ environments have $q$ covariates available, such that $p \geq q$. Let the $p \times q$ matrix of known environmental covariates be given by $S = [s_1 \ldots s_q]$, where $s_h$ is the $p$-vector of centred and scaled environmental scores for covariate $h$, such that $s_h^\top s_h = 1$. The random regression model assumes a different genetic variance for each environment and a different genetic covariance for each pair of environments, which are based on the known covariates.

The random regression model is given by:

$$ a = (1_p \otimes a_1) + (S \otimes I_m) a_s + \delta, $$

where $a_1 = (a_{11}, \ldots, a_{1m})^\top$ is the $m$-vector of marker main effects (intercepts), $a_s = (a_{s1}, \ldots, a_{sq})^\top$ is the $mq$-vector of marker slopes and $\delta = (\delta_1^\top, \ldots, \delta_p^\top)^\top$ is a $mp$-vector of regression residuals (deviations), which are specific to individual environments. Note that since the covariates are centred, the marker intercepts reflect main effects taken at average (zero) values of the covariates.

Chapter 2 developed an extension of the random regression model which includes a reduced rank factor analytic variance matrix for the main effects and slopes. This extension is referred to as the FAR$k$ model, where $k$ denotes the number of common factors. The FAR$k$ model for the marker main effects and slopes is:

$$ a_1 = (\Lambda_1 \otimes I_m) f \quad \text{and} \quad a_s = (\Lambda_s \otimes I_m) f, $$

(5.28)

where $\Lambda_1 = [\lambda_{11} \ldots \lambda_{kk}]$ is a $k$ row-vector of loadings corresponding to the marker main effects, $\Lambda_s = [\lambda_{s1} \ldots \lambda_{sk}]$ is a $p \times k$ matrix of loadings corresponding to the marker slopes and $f = (f_1^\top, \ldots, f_k^\top)^\top$ is a $vk$-vector of marker scores. The FAR$k$ model for the total ME effects is then obtained by substituting Eq. 5.28 into Eq. 5.27, which gives:

$$ a = ([1_p \Lambda_1 + SA_s] \otimes I_m) f + \delta, $$

(5.29)

where all terms are previously defined, with subscripts not applied to $f$ and $\delta$ for brevity.

It is assumed that:

$$ \begin{bmatrix} f \\ \delta \end{bmatrix} \sim N\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} L & 0 \\ 0 & \Psi \end{bmatrix} \otimes D \right), $$

where $L = \oplus_{r=1}^k l_r$ is a $k \times k$ diagonal variance matrix comprising a separate variance for each common factor and $\Psi = \oplus_{j=1}^p \psi_j$ is a $p \times p$ diagonal variance matrix comprising a separate variance for each environment.
It then follows that:

\[
\begin{bmatrix}
a_1 \\
a_s
\end{bmatrix}
\sim \mathcal{N}\left(\begin{bmatrix}0 \\
\Lambda_1 \Lambda_1^\top \\
\Lambda_s \Lambda_s^\top
\end{bmatrix} \otimes D\right).
\]

The variance matrix can be generalised as (ignoring the relationship matrix):

\[
\begin{bmatrix}
\sigma_1^2 & \Sigma_{1s} \\
\Sigma_{s1} & \Sigma_{ss}
\end{bmatrix}
= \begin{bmatrix}
\Lambda_1 \Lambda_1^\top & \Lambda_1 \Lambda_s^\top \\
\Lambda_s \Lambda_1^\top & \Lambda_s \Lambda_s^\top
\end{bmatrix},
\]

where \(\sigma_1^2\) is the marker main effect variance, \(\Sigma_{s1} = \Sigma_{1s}^\top\) is a \(q\)-vector with elements given by the covariance between the main effects and slopes for each known environmental covariate and \(\Sigma_{ss}\) is a \(q \times q\) variance matrix for the slopes with diagonal elements, \(\sigma_{sh}^2\), given by the variances for each covariate. It is important to note that the variance matrix in Eq. 5.30 will be equivalent to an unstructured variance matrix when the number of common factors equals the number of environments, i.e. when \(k = p\). Setting \(k\) in this manner allows the methods to be generalised for any random regression type model.

It therefore follows that:

\[
\text{var}(a) = \left(\begin{bmatrix}1_p & S\end{bmatrix} \begin{bmatrix}\sigma_1^2 & \Sigma_{1s} \\
\Sigma_{s1} & \Sigma_{ss}\end{bmatrix} \begin{bmatrix}1_p \\
S\end{bmatrix}^\top + \Psi\right) \otimes D,
\]

where \(G_a = \sigma_1^2 I_p + \Sigma_{1s} S^\top + S \Sigma_{s1} 1_p^\top + S \Sigma_{ss} S^\top + \Psi\) is the between-environment variance matrix.

### 5.5.1 Baseline linear mixed model

The baseline LMM for the random regression model is obtained by substituting Eq. 5.27 into Eq. 5.1, which gives:

\[
y = X\tau + Z(1_p \otimes M)a_1 + Z(S \otimes M)a_s + Z(1_p \otimes M)\delta + e,
\]

where all terms are given in Eq. 5.1.

It then follows that:

\[
y \sim \mathcal{N}(X\tau, H),
\]

where \(H = Z_m(G_a \otimes D)Z_m^\top + R\) and \(Z_m = Z(1_p \otimes M)\). Standard results for the random effects give \(a_1 = ([\sigma_1^2 1_p^\top + \Sigma_{1s} S^\top] \otimes D)Z_m Py\) and \(a_s = ([\Sigma_{s1} 1_p^\top + \Sigma_{ss} S^\top] \otimes D)Z_m Py\). It is important
to note that $\tilde{a}_1$ involves the covariances in $\Sigma_{1s}$ while $\tilde{a}_s$ involves the covariances in $\Sigma_{s1}$ and $\Sigma_{ss}$. However, in order for the test statistics obtained from the baseline LMM to match those from the marker-in and marker-out LMMs, the covariance terms must be removed from both vectors. Using the new rotation developed in Chapter 3, further adjustments can be made to the BLUPs of the marker main effects and slopes where required to transfer all non-crossover variation from the slopes to the main effects.

The BLUP of marker main effect $i$ is then given by (after removing the covariances):

$$\tilde{a}^*_1i = \sigma^2_1 d_i Z_{m_{1i}}^t P y$$
and
$$\text{PEV}(\tilde{a}^*_1i) = \sigma^2_1 d_i - (\sigma^2_1 d_i)^2 Z_{m_{1i}}^t P Z_{m_{1i}} = \sigma^2_1 d_i - \text{var}(\tilde{a}^*_1i), \quad (5.33)$$

where $P = H^{-1} - H^{-1}X(X^t H^{-1}X)^{-1}X^t H^{-1}$ and $Z_{m_{1i}} = Z (1_p \otimes m_i)$.

The BLUP of marker slope $i$ for known covariate $h$ is (after removing the covariances):

$$\tilde{a}^*_si = \sigma^2_s h d_i Z_{m_{2hi}}^t P y$$
and
$$\text{PEV}(\tilde{a}^*_si) = \sigma^2_s h d_i - (\sigma^2_s h d_i)^2 Z_{m_{2hi}}^t P Z_{m_{2hi}} = \sigma^2_s h d_i - \text{var}(\tilde{a}^*_si), \quad (5.34)$$

where $\sigma^2_{sh}$ is the variance of known covariate $h$ and $Z_{m_{2hi}} = Z (s_h \otimes m_i)$. In the next section, the test statistics for all marker effects of interest will be obtained from the components in Eqs. 5.33 and 5.34.

### 5.5.2 Marker-out linear mixed model

The marker-out LMM is an extension of Eq. 4.9 which fits a random regression model to the polygenic term. The objective of this model is to sequentially fit and test the main effect and slopes of each marker.

Two linear mixed models are considered below:

1. An extension of the marker-out LMM which fits the main effect of marker $i$ as fixed but not random.

2. An extension of the marker-out LMM which fits the slope of marker $i$ for covariate $h$ as fixed but not random.

Model 1 is fitted $m$ times (once for each marker) and Model 2 is fitted $mq$ times (once for each marker by covariate combination). It will be shown how the test statistics for both models can be obtained directly from the fit of the baseline LMM in Eq. 5.31.
Testing marker main effects

The first marker-out LMM fits the main effect of marker \( i \) as fixed but not random in the polygenic term. Let the \( v \times (m - 1) \) marker design matrix excluding column \( i \) be given by \( M_{-i} \). The first marker-out LMM for marker \( i \) is given by:

\[
y_{\text{out}} = X\tau_{\text{out}} + Z_{m,i}\mu_{1i\text{out}} + Z\left(1_p \otimes M_{-i}\right)a_{1-i\text{out}} + Z(S \otimes M)a_{\text{sout}} + Z(I_p \otimes M)\delta_{\text{out}} + e_{\text{out}},
\]

where \( \tau_{\text{out}} \) is a vector of fixed effects, \( \mu_{1i\text{out}} \) is the fixed main effect of marker \( i \), \( a_{1-i\text{out}} \) is the \( (m - 1) \) vector of random main effects excluding marker \( i \), \( a_{\text{sout}} \) is the \( mq \)-vector of random slopes including marker \( i \), \( \delta_{\text{out}} \) is the \( mp \)-vector of regression residuals and \( e_{\text{out}} \) is the \( n \)-vector of residuals. All other terms are given in Eq. 5.1.

It is assumed that:

\[
\begin{bmatrix}
    a_{1-i\text{out}} \\
    a_{\text{sout}}
\end{bmatrix} 
\sim N\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma^2_{1}D_{-i} & \Sigma_{1s} \otimes \Sigma_{ss} \\ \Sigma_{1s} \otimes \Sigma_{ss} & \Sigma_{ss} \otimes D_{-i} \end{bmatrix} \right),
\]

where \( D_{-i} \) is the \( (m - 1) \times m \) diagonal marker matrix excluding row \( i \) and \( D_{-i} \) is the \( m \times (m - 1) \) matrix excluding column \( i \).

It then follows that:

\[
y_{\text{out}} \sim N\left(X\tau_{\text{out}} + Z_{m,i}\mu_{1i\text{out}}, \mathbf{H}_{-12i}\right),
\]

where \( \mathbf{H}_{-12i} = \mathbf{H} - d_i Z_{m,i}(\sigma^2_{1} J_p + I_p \Sigma_{1s}^T + S \Sigma_{ss} I_p^T)Z_{m,i}^T \), such that the variance matrix is adjusted for all variances and covariances associated with the main effect of marker \( i \). Note that since the main effect is fitted as fixed but not random, it contributes to the mean of \( y_{\text{out}} \) but not its variance. The data vector \( y_{\text{out}} \) therefore has a different mean vector and variance matrix compared to \( y \) in Eq. 5.32.

The BLUE of marker main effect \( i \) is given by:

\[
\hat{\mu}_{1i\text{out}} = (Z_{m,i}^T \mathbf{P}_{-12i} Z_{m,i})^{-1}Z_{m,i}^T \mathbf{P}_{-12i} y_{\text{out}},
\]

where \( \mathbf{P}_{-12i} = \mathbf{H}_{-12i}^{-1} - \mathbf{H}_{-12i}^{-1} X(X^T \mathbf{H}_{-12i}^{-1} X)^{-1}X^T \mathbf{H}_{-12i}^{-1} \).

Using Result 1 in Appendix E, it can be shown that:

\[
Z_{m,i}^T \mathbf{P}_{-12i} = Z_{m,i}^T \mathbf{P} \left[ I_n + Z_{m,2i}(G_a^* - Z_{m,2i}^T \mathbf{P} Z_{m,2i})^{-1}Z_{m,2i}^T \mathbf{P} \right]^{-1} Z_{m,2i}^T \mathbf{P},
\]

where \( G_a^* = \sigma^2_{1} J_p + I_p \Sigma_{1s}^T + S \Sigma_{ss} I_p^T \), which is equivalent to \( G_a^* = G_a - S \Sigma_{ss} S^T - \Psi \).
Substituting Eq. 5.36 into Eq. 5.35 gives:

\[
\hat{\mu}_{1out} = (Z_{m_i}^T P Z_{m_i})^{-1} Z_{m_i}^T P Y_{out} + d_i \Sigma_{1s} \Sigma_{2s}^T \hat{a}_{1out},
\]

where \(\hat{a}_{1out} = d_i \Sigma_{1s} \Sigma_{2s}^T P^* Z_{m_i} \). Note that \(\hat{a}_{1out}\) is an indirect prediction which can be obtained using the information in \(\Sigma_{1s}\).

The error variance is then given by:

\[
\text{PEV}(\hat{\mu}_{1out}) = (Z_{m_i}^T P_{-12} Z_{m_i})^{-1}. \tag{5.37}
\]

Substituting Eq. 5.36 into Eq. 5.37 gives:

\[
\text{PEV}(\hat{\mu}_{1out}) = (Z_{m_i}^T P Z_{m_i})^{-1} \\
+ \sigma_1^2 d_i - d_i^2 \Sigma_{1s} \Sigma_{2s}^T \hat{a}_{1out} \\
- 2d_i (Z_{m_i}^T P Z_{m_i})^{-1} Z_{m_i}^T P Z_{m_i} \Sigma_{2s} \Sigma_{1s} \\
= \text{var}(\hat{\mu}_{1out}). \tag{5.39}
\]

The marker-in LMM is similar to the marker-out LMM, with the key difference that the main effect of marker \(i\) is included in the polygenic term. Using the results from Chapter 4, it can be shown that:

\[
\hat{\mu}_{1in} = \hat{\mu}_{1out} - \hat{a}_{1out}, \tag{5.38}
\]

where \(\hat{\mu}_{1in}\) is the BLUE of marker main effect \(i\) from the marker-in LMM.

It can also be shown that:

\[
\text{PEV}(\hat{\mu}_{1in}) = \text{PEV}(\hat{\mu}_{1out}) \\
- \sigma_1^2 d_i + d_i^2 \Sigma_{1s} \Sigma_{2s}^T \hat{a}_{1in} \\
+ 2d_i (Z_{m_i}^T P Z_{m_i})^{-1} Z_{m_i}^T P Z_{m_i} \Sigma_{2s} \Sigma_{1s} \\
= \text{var}(\hat{\mu}_{1in}). \tag{5.39}
\]

This provides an efficient way to obtain the marker-in components from the fit of the marker-out LMM. Note, however, it is more practical to fit the marker-in LMM and then obtain the marker-out components from that model. This is because \(P^*\) is not a natural component of the marker-out LMM and fitting the marker-in LMM also elevates the need to construct a separate genomic relationship for each marker of interest.
Test Statistics

The marker-out test statistic is given by:

$$z_{i\text{out}} = \frac{\hat{\mu}_{i\text{out}}}{\sqrt{\text{var}(\hat{\mu}_{i\text{out}})}}, \quad (5.40)$$

where $z_{i\text{out}}$ is the statistic for testing $H_0: \mu_{i\text{out}} = 0$, which is assumed to have a standard normal distribution. The marker-in test statistic can be obtained from the components in Eqs. 5.38 and 5.39, or by rescaling the marker-out test statistic as:

$$z_{i\text{in}} = \sqrt{\omega_{ii}} z_{i\text{out}} - \sqrt{\omega_{ii}/\text{var}(\hat{\mu}_{i\text{out}})} \hat{a}_{i\text{out}}, \quad (5.41)$$

where $\omega_{ii} = \text{var}(\hat{\mu}_{i\text{out}})/[\text{var}(\hat{\mu}_{i\text{out}}) - \sigma_1^2 d_i + d_1^2 \Sigma_{is}^T Z_{m2}^T P^* Z_{m2} s_1 + 2d_1(Z_{m1}^T P Z_{m1})^{-1} Z_{m1}^T P Z_{m2} (\sigma_1^2 I_p + S \Sigma_{s1})]$.

Obtaining the test statistics in Eqs. 5.40 and 5.41 for all $m$ markers is slow because it requires fitting $m$ linear mixed models. However, the marker-in test statistics can be obtained directly from the fit of the baseline LMM in Eq. 5.31, with random regression model given in Eq. 5.27. The test statistics are now obtained as:

$$z_{i\text{in}} = \frac{\hat{a}_{i\text{in}}^*}{\sqrt{\text{var}(\hat{a}_{i\text{in}}^*)}}. \quad (5.42)$$

This provides a fast method to obtain the marker-in test statistics from the fit of a single LMM. When the variance parameters are known, or when their estimation is unaffected by fitting the main effect of marker $i$ as fixed, the test statistics in Eq. 5.42 are exactly the same as those in Eq. 5.41. A fast exact method for obtaining the marker-out test statistics is not readily available because $P^*$ is not a natural component of the baseline LMM.

It is important to note that the equivalence between the test statistics arises from the fact that the covariance terms have been removed from the BLUPs of the marker main effects in Eq. 5.33. The covariance terms can be included where required, but note that this will result in different test statistics. Using the new rotation developed in Chapter 3, further adjustments can be made to the BLUPs of the marker main effects in order to exclusively capture all available non-crossover variation.
Testing marker slopes

The second marker-out LMM fits the slope of marker \( i \) on known covariate \( h \) as fixed but not random in the polygenic term. The second marker-out LMM for marker \( i \) is given by:

\[
y_{out} = X\tau_{out} + Z_{m2hi}\mu_{shiout} + Z(1_p \otimes M)a_{1out} + Z(S \otimes M)a_{s-hiout} + Z(1_p \otimes M)\delta_{out} + e_{out},
\]

where \( \tau_{out} \) is a vector of fixed effects, \( \mu_{shiout} \) is the fixed slope of marker \( i \) for covariate \( h \), \( a_{1out} \) is the \( m \)-vector of random main effects including marker \( i \), \( a_{s-hiout} \) is the \( mq - 1 \) vector of random slopes excluding marker \( i \) for covariate \( h \), \( \delta_{out} \) is the \( mp \)-vector of regression residuals and \( e_{out} \) is the \( n \)-vector of residuals. Note that the fixed main effect of marker \( i \) is not fitted in order to obtain equivalent test statistics from the baseline LMM.

It is assumed that:

\[
\begin{bmatrix}
  a_{1out} \\
  a_{s-hiout}
\end{bmatrix}
\sim N\left(\begin{bmatrix}
  0 \\
  0
\end{bmatrix}, \begin{bmatrix}
  \Sigma_{1s}D & (\Sigma_{1s} \otimes D)_{-hi} \\
  (\Sigma_{s1} \otimes D)_{-hi} & (\Sigma_{ss} \otimes D)_{-hi}
\end{bmatrix}\right),
\]

where the operator \((\cdot)_{-hi}\) removes the combination of marker \( i \) and covariate \( h \).

It then follows that:

\[
y_{out} \sim N(XX\tau_{out} + Z_{m2hi}\mu_{shiout}, H_{-2hi}),
\]

where \( H_{-2hi} = H - d_i Z_{m2i} (1_p \Sigma_{1sh}S_h^T + s_h \Sigma_{ssh}1_p + S\Sigma_{ssh}^*S^T)Z_{m2i}^T \) and \( \Sigma_{ssh}^* \) is the variance matrix for the marker slopes in Eq. 5.30 with zeros everywhere except column and row \( h \).

The variance matrix is therefore adjusted for all variances and covariances associated with the slope of marker \( i \) for covariate \( h \). Note that since the slope is fitted as fixed but not random, it contributes to the mean of \( y_{out} \) but not its variance. The data vector \( y_{out} \) therefore has a different mean vector and variance matrix compared to \( y \) in Eq. 5.32.

The BLUE of marker slope \( i \) for known covariate \( h \) is given by:

\[
\hat{\mu}_{shiout} = (Z_{m2hi}^T P_{-2hi} Z_{m2hi})^{-1} Z_{m2hi}^T P_{-2hi} y_{out},
\]

where \( P_{-2hi} = H_{-2hi}^{-1} - H_{-2hi}^{-1}X(X^\top H_{-2hi}^{-1}X)^{-1}X^\top H_{-2hi}^{-1} \).

Using Result 1 in Appendix E, it can be shown that:

\[
Z_{m2hi} P_{-2hi} = Z_{m2hi} P[I_n + Z_{m2i} (G_a^* - Z_{m2i} PZ_{m2i})^{-1} Z_{m2i}^T P],
\]

where \( G_a^* = 1_p \Sigma_{1sh}S_h^T + s_h \Sigma_{ssh}1_p + S\Sigma_{ssh}^*S^T \).
Substituting Eq. 5.44 into Eq. 5.43 gives:

\[
\hat{\mu}_{\text{shiout}} = \left( \mathbf{Z}_{\text{m2hi}}^\top \mathbf{P} \mathbf{Z}_{\text{m2hi}} \right)^{-1} \mathbf{Z}_{\text{m2hi}}^\top \mathbf{P} \mathbf{y}_{\text{out}} + d_i \left( \mathbf{\Sigma}_{\text{sh1}} \mathbf{1}_p + \mathbf{\Sigma}_{\text{ssh}} - \mathbf{h} \mathbf{S}_{\text{sh}}^\top \mathbf{h} \right) \mathbf{Z}_{\text{m2hi}}^\top \mathbf{P}^\top \mathbf{y}_{\text{out}},
\]

where \( \bar{a}_{\text{shiout}} = d_i \left( \mathbf{\Sigma}_{\text{sh1}} \mathbf{1}_p + \mathbf{\Sigma}_{\text{ssh}} - \mathbf{h} \mathbf{S}_{\text{sh}}^\top \mathbf{h} \right) \mathbf{Z}_{\text{m2hi}} \mathbf{P}^\top \mathbf{y}_{\text{out}}, \mathbf{S}_{\text{sh}} \) is the \( p \times (q - 1) \) matrix of known environmental covariates excluding column \( h \) and \( \mathbf{P}^\star = \mathbf{P} - \mathbf{P} \mathbf{Z}_{\text{m2hi}} \left( \mathbf{Z}_{\text{m2hi}}^\top \mathbf{P} \mathbf{Z}_{\text{m2hi}} \right)^{-1} \mathbf{Z}_{\text{m2hi}}^\top \mathbf{P} \).

Note that \( \bar{a}_{\text{shiout}} \) is an indirect prediction which can be obtained after estimation using the information in \( \mathbf{\Sigma}_{\text{sh}} \) and \( \mathbf{\Sigma}_{\text{ss}} \).

The error variance is then given by:

\[
\text{PEV}(\hat{\mu}_{\text{shiout}}) = \left( \mathbf{Z}_{\text{m2hi}}^\top \mathbf{P} - 2d_i \mathbf{Z}_{\text{m2hi}} \right)^{-1}. \tag{5.45}
\]

Substituting Eq. 5.44 into Eq. 5.45 gives:

\[
\text{PEV}(\hat{\mu}_{\text{shiout}}) = \left( \mathbf{Z}_{\text{m2hi}}^\top \mathbf{P} \mathbf{Z}_{\text{m2hi}} \right)^{-1} + \sigma_{d_i}^2 d_i - d_i^2 \left( \mathbf{\Sigma}_{\text{sh1}} \mathbf{1}_p + \mathbf{\Sigma}_{\text{ssh}} - \mathbf{h} \mathbf{S}_{\text{sh}}^\top \mathbf{h} \right) \mathbf{Z}_{\text{m2hi}}^\top \mathbf{P}^\star \mathbf{Z}_{\text{m2hi}} \left( \mathbf{1}_p \mathbf{\Sigma}_{\text{sh1}} + \mathbf{S}_{\text{sh}} \mathbf{\Sigma}_{\text{ssh}} \right) - 2d_i \left( \mathbf{Z}_{\text{m2hi}}^\top \mathbf{P} \mathbf{Z}_{\text{m2hi}} \right)^{-1} \mathbf{Z}_{\text{m2hi}}^\top \mathbf{P} \mathbf{Z}_{\text{m2hi}} \left( \mathbf{1}_p \mathbf{\Sigma}_{\text{sh1}} + \mathbf{S} \mathbf{\Sigma}_{\text{ssh}} \right) = \text{var}(\hat{\mu}_{\text{shiout}}).
\]

The marker-in LMM is similar to the marker-out LMM, with the key difference that the slope of marker \( i \) for covariate \( h \) is included in the polygenic term. Using the results from Chapter 4, it can be shown that:

\[
\hat{\mu}_{\text{shin}} = \hat{\mu}_{\text{shiout}} - \bar{a}_{\text{shiout}}, \tag{5.46}
\]

where \( \hat{\mu}_{\text{shin}} \) is the BLUE of marker slope \( i \) for covariate \( h \) from the marker-in LMM.

It can also be shown that:

\[
\text{PEV}(\hat{\mu}_{\text{shin}}) = \text{PEV}(\hat{\mu}_{\text{shiout}}) - \sigma_{d_i}^2 d_i + d_i^2 \left( \mathbf{\Sigma}_{\text{sh1}} \mathbf{1}_p + \mathbf{\Sigma}_{\text{ssh}} - \mathbf{h} \mathbf{S}_{\text{sh}}^\top \mathbf{h} \right) \mathbf{Z}_{\text{m2hi}}^\top \mathbf{P}^\star \mathbf{Z}_{\text{m2hi}} \left( \mathbf{1}_p \mathbf{\Sigma}_{\text{sh1}} + \mathbf{S}_{\text{sh}} \mathbf{\Sigma}_{\text{ssh}} \right) + 2d_i \left( \mathbf{Z}_{\text{m2hi}}^\top \mathbf{P} \mathbf{Z}_{\text{m2hi}} \right)^{-1} \mathbf{Z}_{\text{m2hi}}^\top \mathbf{P} \mathbf{Z}_{\text{m2hi}} \left( \mathbf{1}_p \mathbf{\Sigma}_{\text{sh1}} + \mathbf{S} \mathbf{\Sigma}_{\text{ssh}} \right) = \text{var}(\hat{\mu}_{\text{shin}}). \tag{5.47}
\]

This provides an efficient way to obtain the marker-in components from the fit of the marker-out LMM. Note, however, it is more practical to fit the marker-in LMM and then obtain the marker-out components from that model. This is because \( \mathbf{P}^\star \) is not a natural component of
5.5 Random regression model

the marker-out LMM and fitting the marker-in LMM also elevates the need to construct a separate genomic relationship for each marker of interest.

Test Statistics

The marker-out test statistic is given by:

\[ z_{sh\text{out}} = \frac{\hat{\mu}_{sh\text{out}}}{\sqrt{\text{var}(\hat{\mu}_{sh\text{out}})}}, \]  

(5.48)

where \( z_{sh\text{out}} \) is the statistic for testing \( H_0: \mu_{sh\text{out}} = 0 \), which is assumed to have a standard normal distribution. The marker-in test statistic can be obtained from the components in Eqs. 5.46 and 5.47, or by rescaling the marker-out test statistic as:

\[ z_{sh\text{in}} = \sqrt{\omega_{sh}} z_{sh\text{out}} - \sqrt{\omega_{sh}/\text{var}(\hat{\mu}_{sh\text{out}})} \tilde{a}_{sh\text{out}}, \]  

(5.49)

where \( \omega_{sh} = \text{var}(\hat{\mu}_{sh\text{out}})/[\text{var}(\hat{\mu}_{sh\text{out}}) - \sigma_{sh}^2 d_i + d_i^2 (\Sigma_{ssh} - h S_{hh}^{-1}) Z_{m2}^T P^* Z_{m2}] \). Obtaining the test statistics in Eqs. 5.48 and 5.49 for all \( m \) markers is slow because it requires fitting \( m \) linear mixed models. However, the marker-in test statistics can be obtained directly from the fit of the baseline LMM in Eq. 5.31, with random regression model given in Eq. 5.27. The test statistics are now obtained as:

\[ z_{sh\text{in}} = \frac{\tilde{a}_{sh}^*}{\sqrt{\text{var}(\tilde{a}_{sh}^*)}}, \]  

(5.50)

This provides a fast method to obtain the marker-in test statistics from the fit of a single LMM. When the variance parameters are known, or when their estimation is unaffected by fitting the slope of marker \( i \) for covariate \( h \) as fixed, the test statistics in Eq. 5.50 are exactly the same as those in Eq. 5.49. A fast exact method for obtaining the marker-out test statistics is not readily available because \( P^* \) is not a natural component of the baseline LMM.

It is important to note that the equivalence between the test statistics arises from the fact that the covariance terms have been removed from the BLUPs of the marker slopes in Eq. 5.34 and main effects are not fitted in the marker-in and marker-out LMMs. The covariance terms and main effects can be included where required, but note that this will result in different test statistics. Using the new rotation developed in Chapter 3, further adjustments can be made to the BLUPs of the marker slopes in order to remove all non-crossover variation associated with the main effects.
5.6 Factor analytic model

The factor analytic model is similar to the random regression model in Eq.5.28, with the key difference that the genetic variances and covariances are based on a latent regression. This model is referred to as the FA\(k\) model, where \(k\) denotes the number of latent common factors.

The FA\(k\) model is given by:

\[
a = (\Lambda \otimes I_m) f + \delta, \quad (5.51)
\]

where \(\Lambda = [\lambda_1 \ldots \lambda_k]\) is a \(p \times k\) orthogonal matrix of latent environmental loadings (covariates), \(f = (f_1^\top, \ldots, f_k^\top)^\top\) is a \(mk\)-vector of marker scores (slopes) and \(\delta = (\delta_1^\top, \ldots, \delta_p^\top)^\top\) is a \(mp\)-vector of regression residuals (deviations), which are specific to individual environments. This specification highlights the analogy to the ordinary random regression, with the difference that the covariates are estimated from the data as well as the slopes.

The vector of total ME effects can be rewritten as:

\[
a = (\Lambda \otimes I_m) f + \delta = \beta + \delta,
\]

where \(\beta = (\Lambda \otimes I_m) f\) is a \(mp\)-vector of common ME effects and \(\delta\) is a \(mp\)-vector of specific ME effects. The common ME effects will be used to construct the selection tools in Section 5.6.3, but note that the specific ME effects can also be included where appropriate.

It is assumed that:

\[
\begin{bmatrix} f \\ \delta \end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} L & 0 \\ 0 & \Psi \end{bmatrix} \otimes D \right),
\]

where \(L = \bigoplus_{r=1}^k l_r\) is a \(k \times k\) diagonal variance matrix comprising a separate variance for each common factor sorted in decreasing order and \(\Psi = \bigoplus_{j=1}^p \psi_j\) is a \(p \times p\) diagonal variance matrix comprising a separate variance for each environment.

It then follows that:

\[
\text{var}(a) = (\Lambda \Lambda^\top + \Psi) \otimes D,
\]

where \(G_a = \Lambda \Lambda^\top + \Psi\) is the between-environment variance matrix.
5.6 Factor analytic model

5.6.1 Baseline linear mixed model

The baseline LMM for the factor analytic model is obtained by substituting Eq. 5.51 into Eq. 5.1, which gives:

\[ y = X\tau + Z(\Lambda \otimes M)f + Z(I_p \otimes m)\delta + e, \]

where all other terms are given in Eq. 5.1.

It then follows that:

\[ y \sim N(X\tau, H), \]

where \( H = Z_m(G_a \otimes D)Z_m^\top + R \) and \( Z_m = Z(I_p \otimes M) \).

Constraints are required in factor analytic models to ensure unique solutions for \( \Lambda \) and \( L \) during estimation. Following Smith et al. (2021b), the upper right elements of \( \Lambda \) are set to zero when \( k > 1 \) and \( L \) is set to \( I_k \). Let the environmental loadings and marker scores with these constraints be denoted by \( \Lambda^* \) and \( f^* \), such that \( f^* \sim N(0, I_k \otimes D) \). The loadings and scores are generally rotated after estimation, which can be achieved using a principal component rotation or the new rotation developed in Chapter 3. Each rotation is applied to the REML estimates of the loadings and empirical BLUPs of the scores. For simplicity, the loadings are assumed to be known, so that the predictions of the random effects are hereafter referred to as simply “BLUPs”.

Standard results for the random effects give:

\[ \tilde{f}^* = (A^* \otimes D)Z_m^\top P_y, \]

where \( P = H^{-1} - H^{-1}X(X^\top H^{-1}X)^{-1}X^\top H^{-1} \).

The BLUP of (unrotated) marker score \( i \) for common factor \( r \) is given by:

\[ \tilde{f}_{ri} = d_i Z_{m2r_i}^* P_y \quad \text{and} \quad \text{PEV}(\tilde{f}_{ri}^*) = d_i - d_i^2 Z_{m2r_i}^* P Z_{m2r_i}^* = d_i - \text{var}(\tilde{f}_{ri}^*), \]

where \( Z_{m2r_i}^* = Z(\lambda^*_r \otimes m_i) \) and \( \lambda^*_r \) is the \( p \)-vector of (unrotated) loadings for factor \( r \).

5.6.2 Rotating the loadings and scores

The environmental loadings and marker scores are rotated after estimation to facilitate interpretation and application of the selection tools, which are used to test the contribution of each marker to the measures of overall performance, responsiveness and stability in Chapter 3.
The choice of rotation is important because it dictates the efficiency of using the selection tools in this manner.

Two rotations are considered below:

1. A principal component rotation, which ensures the common factors are orthogonal and sorted in decreasing order.

2. The new rotation developed in Chapter 3, which ensures the first common factor exclusively captures non-crossover GEI and the higher order factors are orthogonal, sorted in decreasing order and exclusively capture crossover GEI.

Each rotation is detailed in the following.

**Principal component rotation**

The principal component rotation is given by:

\[
\Lambda = \Lambda^* V L^{-1/2} \quad \text{and} \quad \tilde{f} = (L^{1/2} V^T \otimes I_m) \hat{f}^*,
\]

which is based on the singular value decomposition given by:

\[
\Lambda^* = UL^{1/2} V^T,
\]

where \(U = [u_1 \ldots u_k]\) is a \(p \times k\) orthogonal matrix of left singular vectors, \(V = [v_1 \ldots v_k]\) is a \(k \times k\) orthogonal matrix of right singular vectors and \(L = \bigoplus_{r=1}^{k} l_r I_k\) is a \(k \times k\) diagonal matrix of singular values sorted in decreasing order, i.e. \(l_1 > \ldots > l_k\).

It then follows that:

\[
\Lambda \equiv U \quad \text{and} \quad \tilde{f} \equiv (L^{1/2} V^T \Lambda^* \otimes D) Z_{m}^T P_y,
\]

such that the common factors are now orthogonal, i.e. \(\Lambda^T \Lambda = I_k\).

The BLUP of (rotated) marker score \(i\) for common factor \(r\) is given by:

\[
\tilde{f}_{ri} = d_i \sqrt{l_r} v_r^T Z_{m2i}^* P_y \quad \text{and} \quad \text{PEV}(\tilde{f}_{ri}) = d_il_r - d_i^2 l_r v_r^T Z_{m2i}^* P Z_{m2i}^* v_r
\]

\[
= d_i l_r - \text{var}(\tilde{f}_{ri}),
\]

where \(Z_{m2i}^* = Z(\Lambda^* \otimes m_i)\). In the next section, the statistics for testing the contribution of markers to the selection tools will be obtained from the components in Eqs. 5.53 and 5.54.
The new rotation

The new rotation places all non-crossover variation into the first common factor and all crossover variation into the higher order factors. This feature ensures that the measure of overall performance is based on non-crossover GEI and that the measures of responsiveness and stability are based on crossover GEI.

The first common factor is given by:

\[ \lambda_1 = \frac{\Sigma_{21}}{\sqrt{\Sigma_{12} \Sigma_{21}}} \quad \text{and} \quad \tilde{f}_1 = \frac{\sqrt{\Sigma_{12} \Sigma_{21}} (\Lambda^* \otimes I_m) \tilde{f}^* / \sigma_1^2}, \tag{5.55} \]

which is based on the variance matrix given by:

\[ \begin{bmatrix} \sigma_1^2 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = \begin{bmatrix} \Lambda^* \tilde{\Lambda}^* & \Lambda^* \tilde{\Lambda}^* \\ \Lambda^* \tilde{\Lambda}^* & \Lambda^* \tilde{\Lambda}^* \end{bmatrix}, \]

where \( \sigma_1^2 \) is the marker main effect variance, \( \Sigma_{21} = \Sigma_{12} \) is a \( p \)-vector with elements given by the covariance between the marker main effects and common ME effects for each environment and \( \Sigma_{22} \) is a \( p \times p \) variance matrix for the common ME effects. Note that constraints are required when at least one covariance in \( \Sigma_{21} \) (and \( \Sigma_{12} \)) is negative (see Chapter 3).

It then follows that:

\[ \lambda_1 = \frac{\Sigma_{21}}{\sqrt{\Sigma_{12} \Sigma_{21}}} \quad \text{and} \quad \tilde{f}_1 \equiv \frac{\sqrt{\Sigma_{12} \Sigma_{21}} (\Lambda^* \Lambda^* \otimes D) Z_m^* P y / \sigma_1^2}, \]

such that the first common factor now has unit length, i.e. \( \lambda_1^* \lambda_1 = 1 \).

The BLUP of (rotated) marker score \( i \) for the first common factor is given by:

\[ \tilde{f}_{1i} = \frac{d_i \sqrt{\sigma_1^2 l_1} \tilde{\Lambda}^* Z_{m_{2i}}^* P y / \sigma_1^2} \quad \text{and} \quad \text{PEV}(\tilde{f}_{1i}) = d_i l_1 - d_i^2 l_1 \tilde{\Lambda}^* Z_{m_{2i}}^* P Z_{m_{2i}}^* Z_{m_{2i}}^* / \sigma_1^2 \]

\[ = d_i l_1 - \text{var}(\tilde{f}_{1i}), \tag{5.56} \]

where \( l_1 = \Sigma_{12} \Sigma_{21} / \sigma_1^2 \) and \( Z_{m_{2i}} = Z(\Lambda^* \otimes m_i) \).

The higher order factors are given by:

\[ \Lambda_{22} = [\Lambda^* - \Sigma_{21} \tilde{\Lambda}^* / \sigma_1^2] V_{22} L_{22}^{-1/2} \quad \text{and} \quad \tilde{f}_{22} = (L_{22}^{1/2} V_{22}^* \otimes I_m) \tilde{f}^*, \tag{5.57} \]

which is based on the singular value decomposition given by:

\[ [\Lambda^* - \Sigma_{21} \tilde{\Lambda}^* / \sigma_1^2] = U_{22} L_{22}^{1/2} V_{22}, \]

where \( U_{22} \) is an orthogonal matrix of left singular vectors, \( V_{22} \) is an orthogonal matrix of
right singular vectors and $L_{22}$ is a diagonal matrix of singular values sorted in decreasing order. Note that the rank of $U_{22}$ and $V_{22}$ is either $k - 1$ when all covariances in $\Sigma_{21}$ are positive or $k$ when at least one covariance is negative (see Chapter 3). For simplicity, let the rank of both matrices equal $k - 1$, so the number of higher order factors is also $k - 1$.

It then follows that:

$$\Lambda_{22} = U_{22} \quad \text{and} \quad \tilde{f}_{22} = (L_{22}^{1/2}V_{22}^T \Lambda^{*T} \otimes D)Z_m^P y,$$

such that the higher order factors are orthogonal and sum to zero, i.e. $\Lambda_{22}^T \Lambda_{22} = I_{k-1}$ and $I_p^T \Lambda_{22} = 0$. The loadings and scores for all factors are then constructed from the components in Eqs. 5.55 and 5.57 as $A = [\Lambda_1 \; \Lambda_{22}]$ and $\tilde{f} = (\tilde{f}_1^T, \tilde{f}_{22}^T)^T$.

The BLUP of (rotated) marker score $i$ for higher order factor $r$ is given by:

$$\tilde{f}_{ri} = d_i \sqrt{l_r} v_r Z_{m_l}^T P y \quad \text{and} \quad \text{PEV}(\tilde{f}_{ri}) = d_i l_r - d_i^2 l_r v_r Z_{m_l}^{*T} P Z_{m_l}^* V_r = d_i l_r - \text{var}(\tilde{f}_{ri}),$$

which has the same form as the principal component rotation in Eq. 5.53, but is based on a different singular value decomposition. In the next section, the statistics for testing the contribution of each marker to the selection tools will be obtained from the components in Eqs. 5.56 and 5.58.

### 5.6.3 Selection tools

The selection tools are used to test the contribution of each marker to the measures of overall performance, responsiveness and stability. The selection tools can be constructed based on the principal component rotation or the new rotation. The choice of rotation is important because it dictates the efficiency of using the selection tools in this manner, i.e. it dictates the ability to detect the contribution of each marker to the response of genotypes with regards to non-crossover and crossover GEI.

There are three measures of interest:

1. Overall performance, representing the contribution of markers to the average genotype response across environments.
2. Responsiveness, representing the contribution of markers to the genotype response to different environmental factors.
3. Stability, representing the contribution of markers to the variability in genotype response across environments.
5.6 Factor analytic model

Testing overall marker performance

The contribution of marker $i$ to overall genotype performance is given by:

$$\text{OP}_i = \tilde{\lambda}_1 \tilde{f}_{1i},$$

where $\tilde{\lambda}_1$ is the mean environmental loading for the first common factor and $\tilde{f}_{1i}$ is the BLUP of rotated marker score $i$. The contribution to OP is therefore equal to the fitted value of the first factor regression at the mean loading of $\tilde{\lambda}_1$. When all environmental loadings for the first factor are positive, the latent regression will capture non-crossover GEI. The choice of rotation is important because the new rotation guarantees this feature whereas the principal component rotation does not. This makes the new rotation particularly favourable in the presence of high levels of crossover GEI, i.e. when the first principal component contains both positive and negative loadings. The contribution to the adjusted OP measure in Eq. 3.14 can also be obtained, but note that this measure is exactly the same as the conventional measure for the new rotation because the higher order factors sum to zero by design.

The marker-in and marker-out test statistics are given by:

$$z_{\text{in}} = \frac{\tilde{f}_{1i}}{\sqrt{\text{var}(\tilde{f}_{1i})}} \quad \text{and} \quad z_{\text{out}} = \frac{\tilde{f}_{1i}}{\omega_{ii} \text{var}(\tilde{f}_{1i})},$$

where $\omega_{ii} = \text{PEV}(\tilde{f}_{1i})/d_{1l1}$, and all other terms are given in Eq. 5.54 for the principal component rotation or Eq. 5.56 for the new rotation. It can be shown that $0 \leq \omega_{ii} \leq 1$, which means that $z_{\text{in}}$ is shrunken compared to $z_{\text{out}}$ and the amount of shrinkage is proportional to $d_{1l1}$. This difference will be important when the number of markers is small, but will become negligible when the number of markers is very large, i.e. when $\omega_{ii} \approx 1$.

Testing marker responsiveness

The contribution of marker $i$ to genotype responsiveness is given by:

$$\text{RESP}_{ri} = (\tilde{\lambda}_r^+ - \tilde{\lambda}_r^-) \tilde{f}_{ri},$$

where $\tilde{\lambda}_r^+$ is the mean of the positive environmental loadings for factor $r$, $\tilde{\lambda}_r^-$ is the mean of the negative environmental loadings and $\tilde{f}_{ri}$ is the BLUP of rotated marker score $i$. The contribution to responsiveness is therefore equal to the fitted value of the regression at the mean contrast of $\tilde{\lambda}_r^+ - \tilde{\lambda}_r^-$. When the environmental loadings are a mixture of positive and negative values, the latent regression will capture at least some crossover GEI.
The marker-in and marker-out test statistics are given by:

\[ z_{riin} = \frac{\tilde{f}_i}{\sqrt{\text{var}(\tilde{f}_i)}} \quad \text{and} \quad z_{riout} = \frac{\tilde{f}_i}{\sqrt{\omega_{ri}\text{var}(\tilde{f}_i)}} \]

where \( \omega_{ri} = \text{PEV}(\tilde{f}_i) / d_i \) such that \( z_{riin} \leq z_{riout} \), and all other terms are given in Eq. 5.54 for the principal component rotation or Eq. 5.58 for the new rotation.

### Testing marker stability

The contribution of marker \( i \) to genotype stability is given by:

\[ \text{RMSD}_i = \sqrt{\tilde{E}_i\tilde{E}_i^\top} / p, \]

where \( \tilde{E}_i = \tilde{F}_i \Lambda - \tilde{f}_1 \lambda_1 \) is a \( p \) row-vector of BLUPs for the common ME effects adjusted for the first rotated factor and \( \tilde{F}_i = [\tilde{f}_1 \ldots \tilde{f}_k] \) is a \( k \) row-vector of BLUPs for marker \( i \).

The contribution to RMSD is therefore equal to the root mean square of the deviations around the first factor regression. Note that the contribution can be written succinctly as \( \text{RMSD}_i = \sqrt{\sum_{k=2}^{k} \tilde{f}_i^2} / p \), since the higher order factors are orthogonal for the principal component rotation and the new rotation.

The RMSD measure involves a quadratic function of the loadings and scores for all higher order factors. As a result there is no closed form of the variance matrix, which prevents testing the contribution of each marker to genotype stability in a straightforward manner. However, test statistics can be constructed for the vector of adjusted ME effects, \( E_i \).

It follows that:

\[ E_i^\top \sim \mathcal{N}(0, d_i G_a^*), \]

where \( G_a^* = \Lambda \Lambda^\top - l_1 \lambda_1 \lambda_1^\top \), which is equivalent to \( G_a^* = G_a - l_1 \lambda_1 \lambda_1^\top - \Psi \).

The vector of BLUPs for marker \( i \) is given by:

\[ \tilde{E}_i = \Lambda \tilde{f}_i - \lambda_1 \tilde{f}_1 \] and \( \text{PEV}(\tilde{E}_i) = d_i G_a^* - \text{var}(\tilde{E}_i) \),

where \( \tilde{f}_i = \text{vec}(\tilde{F}_i) \) and all other terms are given in Eq. 5.54 or Eqs. 5.56 and 5.58.

The marker-in and marker-out test statistics are then given by:

\[ \chi^2_{iin} = \tilde{E}_i \text{var}(\tilde{E}_i)^{-1} \tilde{E}_i^\top \quad \text{and} \quad \chi^2_{iout} = \tilde{E}_i \text{var}(\tilde{E}_i) \Omega_i^{-1} \tilde{E}_i^\top, \]

where \( \Omega_i = \text{PEV}(\tilde{E}_i) G_a^*^{-1} / d_i \), such that \( \chi^2_{iin} \leq \chi^2_{iout} \).
5.7 Concluding remarks

This chapter developed fast exact methods for GWAS using complex models for GEI. The methods were built on the GWAS by GBLUP approach in Chapter 4, which leverages the efficiency of the GBLUP framework and the flexibility of linear mixed models. The objective here was to extend that approach for obtaining the marker-in and marker-out test statistics for genotype by environment data. The objective was traditionally addressed by testing the main effects while ignoring any potential interaction effects. The methods developed in this chapter enable marker effects from the compound symmetry, random regression and factor analytic model to be tested.

The extension to the compound symmetry model involved testing marker main effects and marker by environment interaction effects. It was shown that fitting the main effects and interaction effects together causes issues with estimability, which prevents testing the contribution of markers to each source separately. The GWAS by GBLUP approach can be used to overcome this issue by leveraging the random effects framework, i.e. by adjusting the interaction effects in the baseline LMM for the main effects.

The extension to the random regression model involved testing the marker main effects and slopes. It was shown how the marker-in test statistics can be obtained from the baseline LMM provided that (i) the covariance terms between the main effects and slopes are removed from the BLUPs in the baseline LMM and (ii) the marker main effects are not fitted in the marker-in LMMs. The equivalent marker-out test statistics were not readily available because some of the required terms are not natural components of the baseline LMM. The marker-out test statistics are readily available when the covariances in $\Sigma_{s1}$ and $\Sigma_{ss}$ from Eq. 5.30 are zero, i.e. when the main effects and slopes are independent.

The chapter also extended the GWAS by GBLUP approach for testing the contribution of markers to the selection tools in Chapter 3, which provide measures of overall performance, responsiveness and stability for each genotype. The methods were generalised for the principal component rotation and the new rotation developed in Chapter 3, which places all non-crossover variation into the first common factor and all crossover variation into the higher order factors. This provides a way forward to obtain all required components for the selection tools and GWAS from the fit of a single factor analytic linear mixed model.

The methods developed in this chapter have great potential to improve discovery across a wide-array of genetic studies, particularly with the advent of large-scale datasets and complex genotype by environment interactions. The methods provide a general framework for conducting GWAS for genotype by environment data and provide a solid theoretical basis to implement a wide array of models for GEI.
Chapter 6

Simulation studies

6.1 Prelude

This chapter contains a condensed version of the published manuscript *A framework for simulating genotype by environment interaction using multiplicative models* by Bancic et al. (2024). The chapter develops a general framework for simulating GEI using the class of multiplicative models. The motivation for the framework has stemmed from the importance of GEI in plant breeding and the lack of appropriate methods in current simulations.

The framework can be used to simulate genotype by environment data with features relevant to many plant breeding programmes, including:

1. Low, moderate and high GEI
2. Multiple target populations of environments and multiple phenotypic traits
3. Correlated additive, dominance and epistatic genetic effects
4. Correlated plot errors capturing spatial variation
5. Unbalanced and incomplete experimental designs.

The framework utilises the measures of non-crossover and crossover variance explained developed in Chapter 3 to tune the simulation of GEI. It also develops measures of expected accuracy and MET-TPE alignment to summarise the generation of new MET datasets.

The framework will be demonstrated using two working examples which compare different statistical models and breeding strategies in the presence of low, moderate and high GEI. The examples demonstrate how the new framework can be used to generate realistic MET datasets and model plant breeding programmes that better reflect the complexity of real-world settings, making it a valuable tool for many research objectives.
6.2 Introduction

Simulation is a fast and cost-efficient tool for comparing different plant breeding strategies over time (Gaynor et al., 2021). Many of the current simulations, however, do not adequately capture the complexity of GEI because they either use unrealistic models to simulate it or they ignore it completely. The framework developed in this chapter simulates GEI using the class of multiplicative models. The framework can be used to simulate realistic MET datasets and model plant breeding programmes that better reflect the complexity of real-world settings.

Many stages of plant breeding are complicated by GEI, from the selection of potential parents to the development of improved genotypes for release to growers. GEI can be broadly categorised as non-crossover and crossover interaction, which reflect changes in the scale of genotype response between environments or changes in genotype rank (Baker, 1988; Gail and Simon, 1985, also see Figure 1.2). Crossover GEI is of particular importance to breeders because their selection decisions are more complicated by changes in genotype rank than changes in scale (Eisemann et al., 1990). Plant breeders gauge the extent of GEI in their programmes by accumulating and analysing MET datasets, which contain a sample of environments that generally span multiple years and locations (Smith et al., 2021a). An important consideration when constructing a MET dataset is the extent to which the sample represents the breeder’s target population of environments (TPE, Comstock, 1977; Cooper et al., 1993). This is referred to as the MET-TPE alignment (Cooper et al., 2023).

Multiplicative models have gained popularity in plant breeding because they are effective at capturing non-crossover and crossover GEI. The most general model for GEI is the unstructured model, which fits a separate genetic variance for each environment and a separate genetic covariance between each pair of environments. The unstructured model captures the maximum amount of GEI in the data, however, it becomes computationally prohibitive and unnecessarily complicated as the number of environments increases. These issues can be overcome using multiplicative models. The appealing feature of multiplicative models is that they capture a large proportion of GEI with a small number of multiplicative terms, where each term is the product of an environmental effect and a genotype effect (Mandel, 1971). Some traditional examples include AMMI (Gauch, 1992; Kempton, 1984), GGE (Cornelius et al., 1996; Yan et al., 2000) and factor analytic models (Piepho, 1997; Smith et al., 2001). These models have been shown to provide an informative model for non-crossover and crossover GEI and provide a good fit to MET datasets in general (Gauch et al., 2008; Kelly et al., 2007).

Despite the importance of GEI in plant breeding, current simulations do not adequately capture the complexity observed in real-world settings. Numerous simulation packages have been developed to model plant breeding programmes, such as AlphaSimR (Gaynor et al.,
6.3 Simulation Framework

The framework can be used to simulate two forms of genotype by environment data:

1. Genotype by environment (GE) effects, which have desirable levels of non-crossover and crossover GEI. This process is tuned using measures of variance explained.

2. Multi-environment phenotypes, which are constructed by combining the GE effects with appropriate non-genetic effects. This process is aided with measures of accuracy.

Each process is detailed in the following.

6.3.1 Simulating genotype by environment effects

The framework is built on multiplicative models, which provide a general and scalable approach for simulating a wide range of GEI patterns. Assume the GE effects are simulated
for \( v \) genotypes in \( p \) environments. Let the \( vp \)-vector of GE effects be given by \( \mathbf{u} = (\mathbf{u}_1^\top, \ldots, \mathbf{u}_p^\top)^\top \), where \( \mathbf{u}_j \) is the \( v \)-vector for environment \( j \).

The GE effects are simulated as:

\[
\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}_a \otimes \mathbf{G}),
\]

where \( \mathbf{G}_a \) is a \( p \times p \) between-environment genetic variance matrix and \( \mathbf{G} \) is a \( v \times v \) between-genotype relationship matrix, which is assumed to be completely general. The matrix \( \mathbf{G}_a \) is initially formulated as an unstructured model and then reformulated as a reduced rank multiplicative model. The matrix \( \mathbf{G} \) may be known, simulated or previously estimated from empirical analyses. Both matrices are assumed to be positive (semi)-definite.

The unstructured model provides the most general structure for simulating \( \mathbf{G}_a \) using all \( p(p + 1)/2 \) parameters. This generates a heterogeneous GEI pattern based on a different genetic variance for each environment, \( \sigma_{a_j}^2 \), and a different genetic covariance for each pair of environments, \( \sigma_{a_i a_j} \). Importantly, the unstructured model can be written as a multiplicative model with all \( p \) terms:

\[
\begin{align*}
\mathbf{u} &= (s_1 \otimes \mathbf{I}_v) \mathbf{f}_1 + \ldots + (s_p \otimes \mathbf{I}_v) \mathbf{f}_p \\
&= (\mathbf{S} \otimes \mathbf{I}_v) \mathbf{f},
\end{align*}
\]

where \( \mathbf{S} = [s_1 \ldots s_p] \) is a \( p \times p \) matrix of environmental effects (covariates) and \( \mathbf{f} = (\mathbf{f}_1^\top, \ldots, \mathbf{f}_p^\top)^\top \) is a \( vp \)-vector of genotype effects (slopes).

Consider the eigendecomposition given by:

\[
\mathbf{G}_a = \mathbf{U} \mathbf{L} \mathbf{U}^\top,
\]

where \( \mathbf{U} = [\mathbf{u}_1 \ldots \mathbf{u}_p] \) is a \( p \times p \) orthogonal matrix of eigenvectors and \( \mathbf{L} = \oplus_{r=1}^p l_r \) is a \( p \times p \) diagonal matrix of eigenvalues sorted in decreasing magnitude. The environmental covariates and genotype slopes are then obtained as:

\[
\mathbf{S} = \mathbf{U} \quad \text{and} \quad \mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{L} \otimes \mathbf{G}),
\]

meaning that the environmental covariates are the eigenvectors of \( \mathbf{G}_a \) and the variances of the genotype slopes are the eigenvalues of \( \mathbf{G}_a \). The proportion of variance explained by multiplicative term \( r \) is therefore given by \( l_r/\sum_{r=1}^p l_r \), where the denominator is equivalent to the sum of the diagonal elements of \( \mathbf{G}_a \), i.e. \( \sum_{j=1}^p \sigma_{a_j}^2 \). A large proportion of variance is typically explained by the first few terms, which makes the full rank form in Eq. 6.2 unnecessary as \( p \) increases.
The reduced rank form of Eq. 6.2 arises from taking the first $k$ eigenvectors in Eq. 6.3, which gives:

$$u = (S_k \otimes I_v)f_k,$$

where $S_k$ is a $p \times k$ reduced rank matrix and $f_k$ is a $vk$-vector, such that $f_k \sim N(0, L_k \otimes G)$.

The genotype slopes can be simulated independently or with a breeding simulation package by defining each term as a separate trait, with mean vector set to $0$ and variance matrix set to $L_k$. This formulation requires just $k$ traits (terms) to be simulated, which makes the reduced rank form in Eq. 6.4 even more appealing (see Section 6.4.2).

It then follows that:

$$u \sim N(0, S_k L_k S_k^\top \otimes G),$$

where $G_a \simeq S_k L_k S_k^\top$ is a $p \times p$ reduced rank between-environment genetic variance matrix. Note that the framework simulates $G_a$ first and then obtains the environmental covariates and genotype slopes. The covariates and slopes can be simulated directly, however, this typically leads to an uncontrollable structure for $G_a$ and spurious correlations between environments.

### 6.3.2 Simulating main effects and interaction effects

The reduced rank model in Eq. 6.4 does not have explicit genotype main effects, which instead arise naturally from the form of $G_a$ simulated here. Main effects and interaction effects can be obtained after simulation as:

$$u_1 = (\bar{S}_k \otimes I_v)f_k \quad \text{and} \quad u_2 = (S^*_k \otimes I_v)f_k,$$

where $\bar{S}_k = I_p^\top S_k/p$ is a $k$ row-vector of covariate means and $S^*_k = S_k - (\bar{S}_k \otimes I_p)$ is a $p \times k$ matrix of centred covariates.

The multiplicative model can therefore be rewritten as:

$$u = (1_p \otimes u_1) + u_2,$$

where $u_1$ is the $v$-vector of genotype main effects and $u_2$ is the $vp$-vector of genotype by environment interaction effects.

It then follows that:

$$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \bar{S}_k L_k \bar{S}_k^\top & S_k L_k S_k^\top S_k^* L_k S_k^*^\top \\ S_k^* L_k \bar{S}_k^\top & S_k^* L_k S_k^*^\top \end{bmatrix} \otimes G \right),$$

meaning that the multiplicative model induces a covariance between the main effects and
interaction effects. The variance matrix between the main effects and interaction effects can be rewritten as (ignoring the relationship matrix):

\[
\begin{bmatrix}
\sigma^2_{a1} & \Sigma_{a12} \\
\Sigma_{a21} & \Sigma_{a22}
\end{bmatrix}
= \begin{bmatrix}
\bar{S}_a L_a \bar{S}_k^T & \tilde{S}_a L_a \bar{S}_k^* \\
\bar{S}_k^* L_a \bar{S}_k & \bar{S}_k^* L_a \bar{S}_k^*
\end{bmatrix},
\]

(6.7)

where \(\sigma^2_{a1}\) is the main effect variance, \(\Sigma_{a21} = \Sigma_{a12}^T\) is a \(p\)-vector comprising the covariances between the main effects and interaction effects for each environment and \(\Sigma_{a22}\) is a \(p \times p\) variance matrix comprising a separate interaction variance for each environment and a separate interaction covariance between each pair of environments. These parameters will be used to generate measures of variance explained and accuracy in Sections 6.3.4 and 6.3.7.

### 6.3.3 Simulating a between-environment genetic variance matrix

An important feature of the framework is an approach for simulating \(G_a\) with realistic structure and complexity. This is achieved by simulating heterogeneous genetic variances and correlations through:

\[
G_a = D_a^{1/2} C_a D_a^{1/2},
\]

(6.8)

where \(D_a\) is a \(p \times p\) diagonal genetic variance matrix and \(C_a\) is a \(p \times p\) reduced rank between-environment genetic correlation matrix.

The genetic variances in \(D_a\) are simulated as \(\sigma^2_{a_j} \sim \text{Gamma}(\alpha, \beta)\), where \(\alpha\) is the shape parameter and \(\beta\) is the scale parameter. These parameters are set to \(\alpha = 1.5\) and \(\beta = 1\) for all examples in this chapter (Figure 6.1), but note that other values can be used where required. The gamma distribution was chosen to ensure the genetic variances are positive and that adequate variance heterogeneity is captured by the first multiplicative term in Eq. 6.4. Other distributions can be used where required.

Applying Hardin et al. (2013), the reduced rank between-environment genetic correlation matrix is simulated as:

\[
C_a = \rho I_p + \varepsilon \Lambda^\top \Lambda,
\]

(6.9)

where \(\rho\) is the baseline genetic correlation, \(\varepsilon\) is the variability of the correlations around the baseline and \(\Lambda = [\lambda_1 \ldots \lambda_p]\) is a \((k - 1) \times p\) matrix in which \(\lambda_j\) is the \((k - 1)\)-vector of latent covariates for environment \(j\). Note that the variability of the correlations, \(\varepsilon\), determines the amount of structured noise around the baseline correlation, \(\rho\). The reduced rank form of \(C_a\) arises from the fact that \(J_p\) has rank 1 and \(\Lambda\) has rank \(k - 1\), or more specifically that \(J_p + \Lambda^\top \Lambda\) has rank \(k\). Other base correlation functions can be used where required.
6.3 Simulation Framework

The baseline correlation is subject to the constraint $0 \leq \rho < 1$, which ensures $C_a$ is positive (semi)-definite. If the constraint is not imposed and $-1 < \rho < 0$, indefinite matrices may be generated that require bending. The noise is also subject to the constraint $\varepsilon = 1 - \rho$, which ensures the rank of $C_a$ equals $k$ when $0 < \rho < 1$. If the constraint is not imposed and $\varepsilon < 1 - \rho$, the diagonal must be constrained to one and the rank of $C_a$ will equal $p$. The first $k$ terms in Eq. 6.3 will still capture the majority of variation in $G_a$, but now the remaining $p-k$ terms will each capture a small proportion of variance given by $1 - \rho - \varepsilon$.

An extension of Hardin et al. (2013) is used to simulate the genetic correlations with new functionality to control the skewness of their distribution. The columns of $\Lambda$ in Eq. 6.9 are simulated as $\lambda_j \sim \mathcal{U}(-1, 1 + \gamma)$, where $\gamma$ governs the amount of negative skewness and $-1 \leq \gamma \leq 0$. The $\lambda_j$ are then scaled to unit length, i.e. $\lambda_j^\top \lambda_j = 1$. Note that when $\gamma = 0$, the baseline correlation $\rho$ approaches the mean genetic correlation between environments, but not when $\gamma \neq 0$. Also note that a normal distribution can be used instead of a uniform distribution (see Hardin et al., 2013, for further details).

Fig. 6.1 Simulated genetic variances for 1000 environments, obtained by sampling from a gamma distribution with shape parameter of 1.5 and scale parameter of 1. The vertical blue line represents the mean genetic variance, $\bar{\sigma}_a^2$.
Fig. 6.2 Simulated between-environment genetic correlation matrices with low, moderate and high GEI for 1000 environments. The vertical blue lines on the histograms represent the mean genetic correlation, $\bar{\rho}_a$. Note: All matrices are hierarchically ordered based on separate dendrograms, and have been reduced to the first 100 environments for display.
6.3 Simulation Framework

Table 6.1 Summary of the simulated between-environment genetic variance matrices, $G_a$, with low, moderate and high GEI. Presented are the input parameters for simulating the variances in $D_a$ and correlations in $C_a$, as well as the measures of variance explained.

<table>
<thead>
<tr>
<th>GEI</th>
<th>Input parameters</th>
<th>Variance explained</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha$</td>
<td>$\beta$</td>
</tr>
<tr>
<td>Low</td>
<td>1.5</td>
<td>1</td>
</tr>
<tr>
<td>Moderate</td>
<td>1.5</td>
<td>1</td>
</tr>
<tr>
<td>High</td>
<td>1.5</td>
<td>1</td>
</tr>
</tbody>
</table>

The framework above was used to simulate three examples of $C_a$, which are presented in Figure 6.2 and summarised in Table 6.1. The matrices were constructed with varying $\rho$, $\varepsilon$ and $\gamma$, but applied $k = 7$ for all matrices. All matrices were then multiplied with $D_a$ in Figure 6.1 to create three examples of $G_a$. These matrices form the basis of the low, moderate and high GEI scenarios used throughout the remainder of the chapter.

6.3.4 Measures of variance explained

An important supplement to simulating $G_a$ are measures of variance explained for the GE effects. These measures are used to quantify and tune the proportion of (i) main effect and interaction variances and (ii) non-crossover and crossover variances simulated in $G_a$.

1. The proportions of main effect and interaction variance are given by:

$$v_{a1} = \frac{\sigma^2_{a1}}{\sigma^2_{a1} + \sigma^2_{a2}} \quad \text{and} \quad v_{a2} = \frac{\sigma^2_{a2}}{\sigma^2_{a1} + \sigma^2_{a2}}, \quad (6.10)$$

where $\sigma^2_{a1}$ is the main effect variance and $\sigma^2_{a2}$ is the pooled interaction variance given by:

$$\sigma^2_{a1} = \bar{S}_k^* \textbf{L}_k \bar{S}_k^* \quad \text{and} \quad \sigma^2_{a2} = \bar{S}_k^* \textbf{L}_k \bar{S}_k^*,$$

where $\bar{S}_k^* = \frac{1}{k} \textbf{S}_k^*/p$, such that $\sigma^2_{a1}$ is equivalent to the mean element of $G_a$ and $\sigma^2_{a2}$ is equivalent to the mean diagonal element of $[G_a - \sigma^2_{a1} \textbf{I}_p]$.

The main effect and pooled interaction variances can be rewritten as:

$$\sigma^2_{a1} = \sum_{j=1}^{p} \sigma^2_{a_j}/p^2 + 2 \sum_{i<j}^{p} \sigma_{a_{ij}}/p^2 \quad \text{and} \quad \sigma^2_{a2} = \sum_{j=1}^{p} \sigma^2_{a_j}/p - \sigma^2_{a1}, \quad (6.11)$$

where $\sigma^2_{a_j}$ is the genetic variance of environment $j$ and $\sigma_{a_{ij}}$ is the genetic covariance between environments $i$ and $j$. 
The total genetic variance is therefore given by:

\[ \sigma^2_a = \sigma^2_{a1} + \sigma^2_{a2} = \sum_{j=1}^{p} \frac{\sigma^2_{aj}}{p}, \]

which is equivalent to the mean diagonal element of \( G_a \).

2. The proportions of non-crossover and crossover variance are given by:

\[ v_n = \frac{\sigma^2_n}{\sigma^2_{a1} + \sigma^2_{a2}} \quad \text{and} \quad v_c = \frac{\sigma^2_c}{\sigma^2_{a1} + \sigma^2_{a2}}, \quad (6.12) \]

where \( \sigma^2_n \) is the non-crossover variance and \( \sigma^2_c \) is the crossover variance, which are obtained from the rotation developed in Chapter 3. The examples in Table 6.1 were classified as low, moderate and high GEI based on the proportions of non-crossover and crossover variance.

### 6.3.5 Simulating phenotypes

The framework for simulating the GE effects is embedded within a linear mixed model. Let the \( n \)-vector of phenotypes be given by \( y = (y_1^T, \ldots, y_p^T)^T \), where \( y_j \) is the \( n_j \)-vector for environment \( j \). The linear mixed model used to generate \( y \) is given by:

\[ y = 1_n \mu + X \tau + Zu + e, \quad (6.13) \]

where \( \mu \) is the overall trait mean, \( \tau \) is a \( p \)-vector of environmental main effects with \( n \times p \) design matrix \( X \), \( u \) is the \( v_p \)-vector of GE effects with \( n \times v_p \) design matrix \( Z \) and \( e \) is the \( n \)-vector of plot errors. A randomised complete block design is assumed for each environment, with \( r \) replicates of all \( v \) genotypes. Simulation of other experimental designs and additional (non-genetic) effects is straightforward.

The environmental main effects are simulated as:

\[ \tau \sim N(0, \sigma^2_e I_p), \quad (6.14) \]

where \( \sigma^2_e \) is the environmental main effect variance. This specification will be demonstrated for the MET dataset simulation in Section 6.4.1, and extended to a regression on environmental covariates for the breeding programme simulation in Section 6.4.2.
The GE effects are simulated based on the framework developed above, which can be summarised by:

1. Simulate a between-environment genetic variance matrix as $G_a = D^{1/2}_a C_a D^{1/2}_a$, tuned using the measures of variance explained.

2. Decompose the between-environment genetic variance matrix as $G_a = ULU^\top$, taking the first $k$ terms to obtain $U_k$ and $L_k$.

3. Obtain the environmental covariates by setting $S_k = U_k$ and simulate the genotype slopes as $f_k \sim N(\mathbf{0}, L_k \otimes G)$, either independently or with a breeding simulation package.

4. Construct the GE effects as $u = (S_k \otimes I_v)f_k$.

The framework can be used to generate additive, dominance and epistatic GE effects by appropriately defining $G_a$ and $G$, either explicitly or through simulated population structure and trait architecture in a breeding programme simulation.

Lastly, the plot errors are simulated as:

$$e \sim N(\mathbf{0}, \sigma_e^2 R),$$

(6.15)

where $\sigma_e^2$ is the mean error variance and $R$ is a $n \times n$ block diagonal matrix comprising a separate two-dimensional spatial model for each environment. Correlated plot errors can be generated using the FieldSimR package (Werner et al., 2023), which simulates the main components of spatial variation in plant breeding field trials, that is local and global trend, extraneous variation and random error.

### 6.3.6 Simulating one or more target population of environments

An additional feature of the framework is the ability to simulate a set of environments that represent a breeder’s TPE. Assume that the vector of simulated GE effects, $u$, includes all $p$ environments in the TPE. A subset of $q$ environments is then sampled from the TPE, which may represent a subset observed in a particular year or multiple years in a MET dataset. The total number of environments in the TPE is set to $p = 1000$ for all examples in this chapter, since this produces a sufficiently large distribution for demonstration purposes.

The concepts above can be extended to simulating multiple TPE simultaneously. Assume that each TPE contains $p_i$ environments and that $q_i$ environments are sampled within each, where $p = \sum_{i=1}^s p_i$ and $q = \sum_{i=1}^s q_i$ are the total numbers across all $s$ TPE. The between-
environment genetic correlation matrix in Eq 6.9 is now simulated as:

\[ C_a = \rho J_p + \bigoplus_{i=1}^{s} \delta_i J_{p_i} + \epsilon \Lambda \Lambda^T, \]  

(6.16)

where \( \rho \) is the baseline correlation across all TPE, \( \delta_i \) is the deviation from the baseline for TPE \( i \) and all other parameters are defined in Eq. 6.9. This specification constructs a separate genetic correlation for each TPE, \( \rho_i = \rho + \delta_i \), but the same genetic correlation between TPE, \( \rho \). The between-environment genetic variance matrix is then constructed using Eq. 6.8.

### 6.3.7 Measures of accuracy

An important supplement to simulating \( y \) are measures of accuracy. These measures are used to quantify the expected correlation between the (i) predicted genotype main effects in the MET dataset and the true genotype main effects in the TPE, (ii) true genotype main effects in the MET dataset and TPE, referred to as the MET-TPE alignment (Cooper et al., 2023), and (iii) predicted and true GE effects in the MET dataset.

1. The expected main effect accuracy in the TPE is given by:

\[ r_{a_1} = \sqrt{\frac{\sigma_{a_1}^2}{\sigma_{a_1}^2 + \sigma_{a_2}^2/q + \sigma^2_e/qr}}, \]  

(6.17)

which is equal to the square root of the line-mean heritability (Cooper and DeLacy, 1994).

2. The expected MET-TPE alignment is given by:

\[ r_{mt} = \sqrt{\frac{\sigma_{a_1}^2}{\sigma_{a_1}^2 + \sigma_{a_2}^2/q}}, \]  

(6.18)

which is obtained by setting \( \sigma_e^2 = 0 \) in Eq. 6.17. This measure will be used in Section 6.4.1 as the maximum main effect accuracy in the TPE.

The fundamental relationship between the measures above is then given by:

\[ r_{a_1} = r_m \times r_{mt}, \]  

(6.19)

where \( r_m \) is the expected main effect accuracy in the MET dataset, which is given by:

\[ r_m = \sqrt{\frac{\sigma_{a_1}^2 + \sigma_{a_2}^2/q}{\sigma_{a_1}^2 + \sigma_{a_2}^2/q + \sigma_e^2/qr}}, \]  

(6.20)

where \( \sigma_{a_1}^2 + \sigma_{a_2}^2/q \) is the main effect variance sampled in the MET dataset. The numerator
arises from an inflation of the true main effect variance in Eq. 6.17 by $\sigma^2_{a_2}/q$, which represents the sampling error in the MET dataset. The practical implication is that the expected accuracies observed in the MET dataset will always be higher than in the TPE, but note that the accuracy in the TPE is the measure of interest to breeders.

3. The expected accuracy of the GE effects in the MET dataset is:

$$r_{a_2} = \sqrt{\frac{\sigma^2_{a_1} + \sigma^2_{a_2}}{\sigma^2_{a_1} + \sigma^2_{a_2} + \sigma^2_{e}/r}},$$ (6.21)

which is equal to the square root of the line-mean heritability within environments.

### 6.4 Application

#### 6.4.1 MET dataset simulation

A small example MET dataset was simulated using the framework developed in Section 6.3, and is presented in Table 6.2. The MET dataset comprises 400 genotypes evaluated in field trials across 10 environments. All trials were generated using a randomised complete block design with 2 replicate blocks of 400 plots. The environmental main effects were sampled from a standard normal distribution. The GE effects were simulated based on a hypothetical

<table>
<thead>
<tr>
<th>Design</th>
<th>Trait</th>
</tr>
</thead>
<tbody>
<tr>
<td>Env</td>
<td>Genos</td>
</tr>
<tr>
<td>E1</td>
<td>400</td>
</tr>
<tr>
<td>E2</td>
<td>400</td>
</tr>
<tr>
<td>E3</td>
<td>400</td>
</tr>
<tr>
<td>E4</td>
<td>400</td>
</tr>
<tr>
<td>E5</td>
<td>400</td>
</tr>
<tr>
<td>E6</td>
<td>400</td>
</tr>
<tr>
<td>E7</td>
<td>400</td>
</tr>
<tr>
<td>E8</td>
<td>400</td>
</tr>
<tr>
<td>E9</td>
<td>400</td>
</tr>
<tr>
<td>E10</td>
<td>400</td>
</tr>
<tr>
<td>Overall</td>
<td>400</td>
</tr>
</tbody>
</table>
additive genetic trait with overall mean of 4 and genetic variances sampled from a gamma distribution with shape parameter of 1.5 and scale parameter of 1. The plot errors were simulated assuming independence between plots, based on a plot-level heritability of 0.3, calculated as the proportion of additive genetic variance of the total phenotypic variance.

Model comparison

This section compares various statistical models using MET datasets built on the small example above. Three hypothetical TPEs were simulated based on the low, moderate and high GEI scenarios in Table 6.1, with 1000 environments in each TPE. Twelve MET datasets were then constructed by randomly sampling 5, 10, 20 and 50 environments from each TPE. This process was replicated 1000 times, with eight statistical models fitted to each replicate (see Chapter 2 for model details). The aim of the analyses was to obtain accurate predictions of the genotype main effects and GE effects for each environment. All models were fitted using the ASReml-R package (Butler, 2019) in R (R Core Team, 2023).

Figure 6.3 presents the prediction accuracy of the eight statistical models. This figure also includes the expected main effect accuracy in the MET dataset and TPE (dashed black lines), expected MET-TPE alignment (solid black line) and expect accuracy of the GE effects in the MET dataset. There are five general results:

1. All model predictions become less accurate as the level of GEI increases. The largest differences occur between models at high GEI.

2. All model predictions become more accurate as the number of environments increases. The largest differences occur between models at 5 environments for the main effects and at 50 environments for the GE effects.

3. The main effect accuracies are higher in the MET dataset compared to the TPE. The smallest differences occur between models at 50 environments, where the sampled MET datasets become more aligned to the TPE.

4. The factor analytic models of order 3 and 4 are the most accurate for the main effects in the TPE. The differences between models are negligible except for the factor analytic model of order 1.

5. The factor analytic models of order 3 and 4 are also the most accurate for the GE effects. The largest differences occur between models at high GEI and 50 environments.

The MET-TPE alignments for all 1000 simulation replicates are also presented in Figure 6.3. Not only does the alignment decrease as the level of GEI increases or the number of sampled environments decreases, but the variability around the expected alignment also increases.
6.4 Application

Fig. 6.3 Prediction accuracy of eight statistical models fitted to simulated MET datasets with low, moderate or high GEI and 5, 10, 20 or 50 environments. The genotype main effects in the diagonal and factor analytic models were obtained as averages across environments. The factor analytic models of order 3 and 4 were fitted without the diagonal term for the 5 environment scenario.

6.4.2 Breeding programme simulation

This section integrates the framework developed in Section 6.3 into a breeding programme simulation. The three hypothetical TPE from the previous section are again used for demonstration. The simulation involves 20 years of breeding for a single hypothetical trait, with 20 environments randomly sampled from the 1000 environments in each TPE every year. There are five stages of field evaluation, with an increasing number of environments and a
<table>
<thead>
<tr>
<th>Year</th>
<th>Stage</th>
<th>Genotypes</th>
<th>Envs</th>
<th>Reps</th>
<th>$\sigma_e^2$</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Crossing</td>
<td>$P_1 \times P_2$</td>
<td>100 crosses</td>
<td></td>
<td></td>
<td>Make crosses</td>
</tr>
<tr>
<td>2</td>
<td>F1</td>
<td>100 families</td>
<td></td>
<td></td>
<td></td>
<td>Produce DHs</td>
</tr>
<tr>
<td>3</td>
<td>HDRW</td>
<td>10,000</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>Advance 500 DHs</td>
</tr>
<tr>
<td>4</td>
<td>PYT</td>
<td>500</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>Yield trial</td>
</tr>
<tr>
<td>5</td>
<td>AYT</td>
<td>50</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>Yield trial</td>
</tr>
<tr>
<td>6</td>
<td>EYT</td>
<td>10</td>
<td>20</td>
<td>2</td>
<td>2</td>
<td>Yield trial</td>
</tr>
<tr>
<td>7</td>
<td>Variety</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>Release variety</td>
</tr>
</tbody>
</table>

Fig. 6.4 Key features of the simulated plant breeding programme. Presented are the number of genotypes, environments and replicates per environment as well as the mean error variance and the action taken. DH - double haploid, FS - full-sib, HDRW - headrow, PYT - preliminary yield trial, AYT - advanced yield trial, EYT - elite yield trial

decreasing number of genotypes observed in each stage (Figure 6.4). This produces a subset of 400 environments from each TPE and a maximum of 20 environments observed in each stage, every year.

**Breeding programme comparison**

This section compares phenotypic and genomic selection strategies using a breeding programme simulation in AlphaSimR (Gaynor et al., 2021). The key features of the breeding programme are presented in Figure 6.4. The breeding programme was simulated with no, low, moderate and high GEI, and then phenotypic or genomic selection was applied for 20 years of breeding. This produced eight scenarios that were replicated 20 times. The MET dataset for phenotypic selection comprises the subset of environments for each stage in the current year only (ranging from 1 for HDRW to 20 for EYT), and for genomic selection comprises all stages and sampled environments from the last three years (60 in total). A compound symmetry model was fitted for the genomic selection strategy, with selection performed on the predicted genotype main effects. The aim of the simulation was to track the genetic gain, genetic variance and the measures of accuracy in the HDRW stage during 20 years of breeding. The scenario without GEI will be used as a baseline for comparison (solid black line).
There are three general results for the genetic gain in Figure 6.5:

1. The genetic gain decreases as the level of GEI in the TPE increases. The genetic gain after 20 years is $42 - 87\%$ lower than the baseline for phenotypic selection and $39 - 84\%$ lower for genomic selection. The smallest difference occurs for low GEI while the largest difference occurs for high GEI.

2. The genetic gain from genomic selection is $49 - 73\%$ higher than phenotypic selection in the TPE after 20 years. The smallest difference occurs for low GEI while the largest difference occurs for high GEI.

3. There are minor differences between the genetic gain in the MET dataset and TPE for genomic selection, but there are noticeable differences for phenotypic selection. The largest difference occurs for high GEI where the genetic gain in the MET is $\sim 30\%$ higher than the TPE after 20 years.
Fig. 6.6 Genetic variance in the simulated plant breeding programmes with no, low, moderate or high GEI and phenotypic or genomic selection strategies. Genetic variance is presented as the variance of the genotype main effects in the TPE and MET dataset for the headrow stage.

There are three general results for the genetic variance in Figure 6.6:

1. The loss in genetic variance decreases as the level of GEI in the TPE increases. The loss in genetic variance after 20 years is $31 - 81\%$ lower than the baseline for phenotypic selection and $15 - 39\%$ lower for genomic selection. The smallest difference occurs for low GEI while the largest difference occurs for high GEI.

2. The loss in genetic variance from genomic selection is $2.0 - 5.4$ times higher than phenotypic selection in the TPE after 20 years. The smallest difference occurs for low GEI while the largest difference occurs for high GEI.

3. There are minor differences between the genetic variance in the MET dataset and TPE for genomic selection, but there are noticeable differences for phenotypic selection. The largest difference occurs for high GEI where the genetic variance in the MET is $\sim 10\%$ higher than the TPE after 20 years.
There are three general results for the measures of accuracy in Figure 6.7:

1. The main effect accuracies and MET-TPE alignment decrease as the level of GEI increases. The lowest accuracies and alignments occur for high GEI.

2. The main effect accuracy for genomic selection is much higher in the TPE than phenotypic selection. The smallest difference occurs for low GEI while the largest difference occurs for high GEI.

3. The MET-TPE alignment for high GEI is much lower and more variable for phenotypic selection than genomic selection. This is because the MET dataset for phenotypic selection includes just 1 environment for the HDRW stage whereas for genomic selection it includes 60 environments.
6.5 Concluding remarks

This chapter developed a general framework for simulating GEI using multiplicative models. The new framework was motivated by the importance of GEI in plant breeding and the lack of appropriate methods in current simulations. The objective here was to develop a scalable, reproducible and user-friendly approach for simulating GEI in terms of tangible parameters to plant breeders and researchers. The framework can be used to generate realistic MET datasets and model plant breeding programmes that better reflect the complexity of real-world settings.

The framework can be summarised by four key steps:

1. **Simulate a reduced rank between-environment genetic variance matrix**, $G_a$, with heterogeneous genetic variances and correlations. The measures of variance explained from Chapter 3 were used to tune the simulation of $G_a$ to achieve desirable structure.

2. **Decompose** $G_a$ to obtain the eigenvectors and eigenvalues for the first $k$ terms. This produces a reduced rank set of vectors that capture the structure in $G_a$.

3. **Obtain environmental covariates** as the eigenvectors and **simulate genotype slopes** based on the eigenvalues. This produces a set of environmental covariates and genotype slopes that can be generated using a simulation package.

4. **Construct GE effects** by multiplying the environmental covariates with the genotype slopes. This produces GE effects based on a reduced rank multiplicative model.

The framework was demonstrated for comparing different statistical models fitted to simulated MET datasets with low, moderate and high GEI. The results show that gains in accuracy can be achieved by sampling more environments from the TPE and that the main effect accuracy in the TPE is always lower than in the MET dataset, as a consequence of the imperfect MET-TPE alignment. The framework was also demonstrated for comparing different breeding strategies over time subject to low, moderate and high GEI. The results show that more realistic projections of genetic gain, genetic variance and prediction accuracy can be obtained in terms of the potential GEI patterns in a breeder’s TPE.

Overall, the results highlight the utility of the new framework to produce key metrics of interest to plant breeding, making it a valuable tool for exploring many new research objectives through simulation.

Areas of active research include extending the concepts of MET-TPE alignment to the GE effects within environments as a novel way to quantify differences in genotype stability between the MET dataset and TPE.
Chapter 7

General discussion

Plant breeding is complicated by the fact that genotypes respond differently to different environments, a phenomenon known as genotype by environment interaction (GEI). Despite its importance, however, many plant breeding programmes still use inefficient methods to handle GEI. This thesis developed a wide-array of methods that leverage GEI for efficient prediction, selection and discovery in plant breeding.

- Chapter 2 developed new genomic prediction models that combines known genotypic covariates derived from marker data and known environmental covariates derived from weather and soil data along with latent environmental covariates estimated directly from the phenotypic data. The models constitute the class of integrated factor analytic linear mixed models (IFA-LMMs).

- Chapter 3 developed selection tools that provide measures of overall performance, responsiveness and stability for one or more target population of environments (TPE). The selection tools were built on a new approach that completely disentangles non-crossover and crossover GEI.

- Chapters 4 and 5 developed fast exact methods for conducting genome-wide association studies (GWAS) using genotype by environment data. The methods produce all required test statistics from the fit of a single linear mixed model, rather than fitting a very large number of models for all marker effects of interest.

- Chapter 6 developed a general framework for simulating GEI using multiplicative models. The framework can be used to simulate realistic multi-environment trial (MET) datasets and model plant breeding programmes that better reflect the complexity of real-world settings.

Each chapter is discussed in the following.
7.1 Genomic prediction models

Genomic prediction is a form of marker-assisted prediction that can improve the genetic gain in animal and plant breeding (Meuwissen et al., 2001). Genomic prediction has been widely and successfully adopted in animal breeding, however, key points of difference were not fully considered in the transfer to plant breeding. The objective of Chapter 2 was to re-emphasise appropriate methods for analysing plant breeding MET datasets, with a particular focus on novel genomic prediction models for GEI.

The integrated factor analytic linear mixed model (IFA-LMM) was developed as an extension of an already effective approach for analysing MET datasets, i.e. the factor analytic linear mixed model (Smith et al., 2001). The objective here was to supplement that approach with a way to (i) ascribe direct biological interpretation to the underlying factors, and (ii) obtain forward predictions that were competitive with the conventional random regression approaches. Traditionally, these objectives were addressed using multiple methods of analysis. The IFA-LMM addressed both objectives within a single, unified approach.

The IFA-LMM was demonstrated using an example cotton breeding dataset from Bayer CropScience. The results show that the known environmental covariates predominately captured crossover GEI while the latent environmental covariates predominately captured non-crossover GEI. The known covariates therefore provided a good basis to capture changes in genotype rank between environments but not changes in scale, which involved more complex latent interactions. The results also show that the predictive ability of the IFA-LMM was higher than the conventional random regression approaches, particularly for the more variable environments in the example dataset. Importantly, this is the first example of using factor analytic linear mixed models for forward prediction, which addresses a longstanding objective of many plant breeders and researchers.

An important limitation of the conventional factor analytic linear mixed model is that it does not scale linearly to large MET datasets. The IFA-LMM is an extension of this approach so that it inherits the same limitation. Current research is focused on scaling factor analytic linear mixed models using novel extensions of the sparse formulation (Thompson et al., 2003) and data augmentation (Clayton and Rasbash, 1999). Data augmentation provides a natural framework for estimation by iteratively fitting each common factor separately while adjusting for the other factors. Another limitation of the IFA-LMM, which is inherited from random regression models, is that the forward predictions are susceptible to poor scaling of the known environmental covariates. The poor scaling is generally amplified for combinations of covariates not fitted in the model and further amplified by the presence of covariance terms between the known covariates. This is likely to be overcome by using large MET datasets with many different environments and many different combinations of covariates.
7.2 Selection tools

Selection for broad and specific adaptation is an important breeding decision in the presence of GEI. Traditionally, these selection decisions have been aided by some measure of overall performance and stability for each genotype (Smith and Cullis, 2018). However, current approaches do not completely separate non-crossover and crossover GEI, and this may reduce the efficiency of selection. The objective of Chapter 3 was to develop selection tools for one or more TPE, with a particular focus on disentangling non-crossover and crossover GEI.

The selection tools were built on a new rotation for factor analytic linear mixed models that places all non-crossover variation into the first common factor and all crossover variation into the higher order factors. The new rotation avoids the ambiguity of the principal component rotation when the first principal component contains both positive and negative loadings, i.e. when it captures both non-crossover and crossover GEI. This generally occurs when crossover variation is the dominant form of variation in the data. Importantly, the new rotation provides a formal way to (i) quantify the amount of non-crossover and crossover variation in the data and (ii) identify when crossover variation becomes the dominant form. Generalisations of the new rotation and resulting selection tools were also developed for any random regression type model.

The selection tools were demonstrated using an example cotton breeding dataset from Bayer CropScience. The results show that the overall performance and stability measures obtained from the two rotations are mostly similar because the proportion of non-crossover variance is high for the example dataset, so that the first principal component predominately captured non-crossover GEI. However, there will be differences between the two rotations when the proportion of non-crossover variance is low and the first principal component captures both non-crossover and crossover GEI. Current research is focused on determining when selection for broad adaptation becomes inefficient or even inappropriate based on the form and magnitude of GEI.

An important limitation of the conventional selection tools is that the stability measure summarises the variation above and below the first factor regression, so that a selection index will penalise favourable performance as well as unfavourable performance. The directional stability measure overcomes this limitation because it partitions the conventional measure into components representing favourable and unfavourable performance. It was shown that the directional stability measure is more informative than the conventional measure for identifying potential genotypes with specific adaptation. However, the directional stability measure is a quadratic like the conventional measure, which makes breeding for stability difficult. Current research is focused on developing a linear stability measure that has the desirable features of the directional stability measure, without the quadratic complexity.
7.3 Association studies

Genome-wide association studies are a powerful exploratory tool for biological discovery in a wide-array of genetic studies (Visscher et al., 2017; Zhu et al., 2008). Many of the current approaches are inherently inefficient, however, because they involve fitting a very large number of models to obtain the required test statistics for all markers of interest. Many are also deficient because they double fit the marker effect of interest as fixed and random which can result in shrunken test statistics. The objective of Chapter 4 was to develop fast exact methods for GWAS that produce the appropriate test statistics from the fit of a single linear mixed model.

The methods were built on an existing approach for GWAS which leverages the efficiency of the GBLUP framework and the flexibility of linear mixed models, i.e. GWAS by GBLUP (Legarra et al., 2018). The objective here was to extend that approach for obtaining the marker-out test statistics, rather than the marker-in test statistics. The objective was traditionally addressed by sequentially fitting a very large number of linear mixed models, with a different genomic relationship matrix constructed for each model without the marker of interest. The methods developed in Chapter 4 enable the marker-in and marker-out test statistics to be obtained from the fit of the same linear mixed model, with the same genomic relationship matrix. The methods were generalised for any set of fixed and random effects and extended for testing a set of markers at a time.

An important question in the literature was also addressed; should the marker effect of interest be included in the polygenic term? It was shown algebraically that double fitting the marker effect as fixed and random causes the two effects to be statistically aliased, which causes the resulting test statistics to be shrunken by the polygenic variance. A fundamental expression was derived which provides an efficient way to obtain the marker-out test statistics directly from the fit of the marker-in linear mixed model, and vice versa. This provides a simple fix for current software packages employing the marker-in approach.

The main limitation of the GWAS by GBLUP approach is that it assumes the variance parameters are known, or unaffected by fitting the marker effect of interest as fixed. However, the objective of GWAS is to identify potential marker effects that are significant for the observed phenotypes, so that it is only natural to assume fitting said marker effects as fixed should affect variance parameter estimation. It was shown that these assumptions can produce conservative test statistics compared to re-estimating the variance parameters, which can result in higher rates of false negatives. Current research is focused on finding fast exact methods that relax the assumptions on the variance parameters but are still applicable to the extensions for genotype by environment data in Chapter 5.
Extension to genotype by environment data

Many genome-wide association studies in plant breeding use genotype by environment data. Most of these studies, however, do not fully consider the structure of plant breeding MET datasets and the importance of GEI. The objective of Chapter 5 was to introduce appropriate methods for conducting GWAS using genotype by environment data, with a particular focus on applying the genomic prediction models in Chapter 2 and the selection tools in Chapter 3.

The methods were built on the GWAS by GBLUP approach in Chapter 4 and extended for testing marker effects from the compound symmetry, random regression and factor analytic models. The extension to the compound symmetry model involved testing marker main effects and marker by environment interaction effects. It was shown that fitting the main effects and interaction effects together causes issues with estimability, which prevents testing the contribution of markers to each source separately. The GWAS by GBLUP approach can be used to overcome this issue by leveraging the random effects framework.

The extension to the random regression model involved testing the marker main effects and slopes for different known environmental covariates. It was shown that identical marker-in test statistics can be obtained from the baseline linear mixed model provided that (i) the covariance terms between the main effects and slopes were removed from the predictions in the baseline linear mixed model and (ii) the marker main effects were not fitted in the marker-in linear mixed models. The equivalent marker-out test statistics were not readily available because some of the required terms are not natural components of the baseline linear mixed model. Equivalent marker-out test statistics can be obtained when the covariance terms between the main effects and slopes are not fitted, i.e. when they are completely independent.

The extension to the factor analytic model involved testing the marker scores for different latent environmental factors. The extension also involved testing the contribution of markers to the selection tools, which provide measures of overall performance, responsiveness and stability for each genotype. The methods were generalised for the principal component rotation and the new rotation developed in Chapter 3, which places all non-crossover variation into the first common factor and all crossover variation into the higher order factors. This provides a way forward to obtain all required components for genomic prediction, selection and association studies from the fit of a single factor analytic linear mixed model, which addresses a longstanding objective of many plant breeders and researchers.

The methods developed in Chapter 5 have great potential to improve discovery across a wide-array of genetic studies, particularly with the advent of large-scale datasets and complex genotype by environment interactions. Current research is focused on applying the methods to real-world MET datasets and extensions to the IFA-LMM.
7.4 Simulation studies

Simulations are routinely used in plant breeding as a fast and cost-effective way to compare different breeding methodologies over time (Gaynor et al., 2021). Many of the current simulations, however, do not adequately capture the complexity of GEI observed in real-world plant breeding. The objective of Chapter 6 was to develop a general framework for simulating GEI using the class of multiplicative models.

The framework can be summarised by four key steps:

1. Simulate a between-environment genetic variance matrix with heterogeneous genetic variances and covariances. New measures of variance explained were developed to tune the amount of non-crossover and crossover variance.

2. Decompose the between-environment genetic variance matrix to obtain the eigenvectors and eigenvalues for the first $k$ terms. This produces a reduced rank set of vectors that effectively capture the structure of simulated GEI.

3. Set the environmental covariates as the eigenvectors and simulate the genotype slopes based on the eigenvalues. This produces a reduced rank set of environmental covariates and genotype slopes, which can be generated using a breeding simulation package.

4. Construct genotype by environment (GE) effects by multiplying the environmental covariates with the genotype slopes. Construct phenotypes as required by adding appropriate non-genetic effects to the simulated GE effects.

The framework can be embedded within a linear mixed model to generate realistic MET datasets or integrated within a simulation package to model real-world plant breeding programmes. The latter is achieved by utilising the pre-existing features of current simulation packages, i.e. their ability to simulate multiple correlated traits. There are two appealing features of this approach. Firstly, the reduced rank model only requires a small number of traits (multiplicative terms) to be simulated, rather than all environments across all years of breeding. Secondly, the environmental covariates are constructed prior to simulation so they capture the full set of environmental effects for all years of breeding, while the genotype slopes are generated during breeding so they capture the changing genetics in the population over time. This produces realistic genotype by environment data that is a natural function of the changing environments and the changing genetics over time.

The framework provides a solid basis to evaluate and optimise the wide-array of methods developed in this thesis. Current research is focused on applications to hybrid breeding programmes and transition strategies for two-part breeding programmes.
7.5 Concluding remarks

Genotype by environment interaction continues to receive much attention in plant breeding and its importance is becoming more profound in other agricultural and scientific fields. This thesis developed a wide-array of methods that leverage GEI for efficient prediction, selection and discovery. Although the methods and concepts were developed for plant breeding settings, they do have broader applications to animal and human genetics.

- Chapter 2 developed a new genomic prediction approach that combines known genotypic and environmental covariates with latent environmental covariates. The approach has great potential in animal breeding settings that collate large volumes of environmental data, such as dairy breeding and aquaculture.

- Chapter 3 developed selection tools that provide measures of overall performance, responsiveness and stability. The selection tools have relevance to animal breeding settings when selecting high performing and stable parents for key traits of interest.

- Chapters 4 and 5 developed fast exact methods for conducting GWAS. The methods have immediate application to many plant, animal and human genetic studies, particularly with the advent of large-scale phenotypic and genotypic datasets.

- Chapter 6 developed a general framework for simulating GEI using multiplicative models. The framework can be used to model various animal breeding programmes that better reflect the complexity of real-world settings. It also provides a scalable approach to generate high-dimensional data for human association studies.

Genotype by environment interaction was traditionally viewed as an impediment to plant breeding. Over the past 100+ years, however, methods have been developed that leverage GEI for efficient prediction, selection and discovery. The methods and concepts developed in this thesis have great potential to improve predictive plant breeding, particularly with the emergence of rapidly changing and more extreme environments amidst climate change.
References


Appendix A

Sparse formulation

This appendix extends the sparse implementation of the average information (AI) algorithm for the integrated factor analytic (IFA$k$) model developed in Chapter 2. Thompson et al. (2003) introduced the sparse implementation for the conventional factor analytic model as a more efficient approach to estimate the key variance parameters, i.e. the environmental loadings, factor variances and specific variances. This approach also provides a natural way to handle cases where some specific variances are zero and some are non-zero. When more than one specific variance is zero, the factor analytic variance matrix is not positive definite. The reduced rank factor analytic model arises when all specific variances are zero (Kirkpatrick and Meyer, 2004). Note that the notation used below has been aligned with the remainder of the thesis.

A.1 Preliminaries

Assume a MET dataset comprises $v$ genotypes evaluated in $t$ field trials across $p$ environments, where $t = \sum_{j=1}^{p} t_j$ and $t_j$ is the number of trials in environment $j$. Let the $n$-vector of phenotypic data be given by $y = (y_1^\top, \ldots, y_p^\top)^\top$, where $y_j = (y_j^\top_1, \ldots, y_j^\top_{t_j})^\top$ is the $n_j$-vector for environment $j$ and $y_{jk}$ is the $n_{jk}$-vector for trial $k$ in environment $j$. Also assume that all $p$ environments have $q$ known covariates available, with $p \geq q$. Let the $p \times q$ matrix of known covariates be given by $S = [s_1 \ldots s_q]$, where $s_i$ is the vector of centred and scaled environment scores for covariate $i$, i.e. $s_i^\top s_i = 1$.

The linear mixed model for $y$ can be rewritten as:

$$ y = X\tau + Z_{Alf}f + Z\delta + e, \quad (A.1) $$

where $Z_{Alf} = Z(BA_b \otimes I_v)$ is the $n \times vk$ integrated factor analytic design matrix, $B = [S \Gamma']$
is the $p \times p$ matrix of basis functions, $\mathbf{A}_b = \begin{bmatrix} \mathbf{A}_s \\ \mathbf{A}_r \end{bmatrix}$ is the $p \times k$ matrix of joint factor loadings. The IFA-LMM in Eq. A.1 can be extended by partitioning the regression residuals as $\mathbf{\delta} = (\mathbf{\delta}_1^T, \mathbf{\delta}_2^T)^T$, where $\mathbf{\delta}_1$ is a $vp_1$-vector with no zero elements and $\mathbf{\delta}_2$ is a $vp_2$-vector with all zero elements, such that $p = p_1 + p_2$. Two further models can be obtained:

1. When $p_1 = p$ and $p_2 = 0$, no specific variances are zero. This model is the full rank IFA-LMM.

2. When $p_1 = 0$ and $p_2 = p$, all specific variances are zero. This model is the reduced rank IFA-LMM.

The model considered below allows some specific variances to be non-zero and some to be zero. The IFA-LMM can therefore be written as:

$$y = \mathbf{X}\mathbf{\tau} + \mathbf{Z}_{A_b} f + \mathbf{Z}_1 \mathbf{\delta}_1 + \mathbf{Z}_2 \mathbf{\delta}_2 + \mathbf{e},$$

where $\mathbf{Z} = [\mathbf{Z}_1 \mathbf{Z}_2]$ is partitioned conformably with $\mathbf{\delta}$.

It is assumed that $\mathbf{\delta}_2 = \mathbf{0}$, and that:

$$\begin{bmatrix} f \\ \mathbf{\delta}_1 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{L} & 0 \\ 0 & \mathbf{\Psi}_1 \end{bmatrix} \otimes \mathbf{G}\right),$$

where $\mathbf{L} = \bigoplus_{r=1}^k \mathbf{I}_r$ is a $k \times k$ factor variance matrix comprising a separate variance for each common factor sorted in decreasing magnitude, i.e. $l_1 > \ldots > l_k$, and $\mathbf{\Psi}_1 = \bigoplus_{j=1}^{p_1} \mathbf{\psi}_{1j}$ is a $p_1 \times p_1$ specific variance matrix comprising a separate non-zero variance for each environment. It is assumed that the genotype scores and regression residuals are independent. Thompson et al. (2003) also present an equivalent formulation where these effects are assumed to be dependent. This will be extended for the IFA$k$ model in subsequent work.

It then follows that:

$$y \sim \mathcal{N}(\mathbf{0}, \mathbf{H}),$$

where $\mathbf{H} = \mathbf{Z}_{A_b} (\mathbf{L} \otimes \mathbf{G}) \mathbf{Z}_{A_b}^T + \mathbf{Z}_1 (\mathbf{\Psi}_1 \otimes \mathbf{G}) \mathbf{Z}_1^T + \mathbf{R}$. 
A.2 Prediction of genotype scores and regression residuals

The mixed model equations for the IFA-LMM in Eq. A.2 are given by:

\[
\begin{pmatrix}
X^\top R^{-1}X & X^\top R^{-1}Z_{A_b} & X^\top R^{-1}Z_1 \\
Z_{A_b}^\top R^{-1}X & Z_{A_b}^\top R^{-1}Z_{A_b} + L^{-1} \otimes G^{-1} & Z_{A_b}^\top R^{-1}Z_1 \\
Z_1^\top R^{-1}X & Z_1^\top R^{-1}Z_{A_b} & Z_1^\top R^{-1}Z_1 + \Psi_1^{-1} \otimes G^{-1}
\end{pmatrix}
\begin{pmatrix}
\hat{\tau} \\
\tilde{f} \\
\tilde{\delta}_1
\end{pmatrix} =
\begin{pmatrix}
\hat{\tau} \\
\tilde{f} \\
\tilde{\delta}_1
\end{pmatrix}
\]

(A.3)

Following Smith et al. (2019), Eqs. A.2 and A.3 can be written as:

\[
y = W\beta + e \quad \text{and} \quad C\tilde{\beta} = W^\top R^{-1}y,
\]

(A.4)

where \(W = [X Z_{A_b} Z_1]\) and \(C = W^\top R^{-1}W + G_c^{-1}\), with:

\[
\tilde{\beta} = \begin{pmatrix}
\hat{\tau} \\
\tilde{f} \\
\tilde{\delta}_1
\end{pmatrix}
\]

and \(G_c^{-1} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & L^{-1} \otimes G^{-1} & 0 \\ 0 & 0 & \Psi_1^{-1} \otimes G^{-1} \end{pmatrix} \).

The BLUPs of the key random effects are obtained via absorption of \(C\) onto \(y^\top R^{-1}y\), which gives:

\[
\tilde{f} = [L \otimes G] Z_{A_b}^\top P y \quad \text{and} \quad \tilde{\delta}_1 = [\Psi_1 \otimes G] Z_1^\top P y,
\]

where \(P = R^{-1} - R^{-1}W C^{-1} W^\top R^{-1}\) is the \(n \times n\) residual sum of squares matrix. Absorption of \(C\) also produces the prediction error variance matrices, with:

\[
\text{PEV}(\tilde{f}) = C\tilde{f} \quad \text{and} \quad \text{PEV}(\tilde{\delta}_1) = C\tilde{\delta}_1,
\]

where \(C\tilde{f} = [L \otimes G] (I_{v_k} - Z_{A_b}^\top P Z_{A_b} [L \otimes G])\) is a \(v_k \times v_k\) matrix and \(C\tilde{\delta}_1 = [\Psi_1 \otimes G] (I_{v_p} - Z_1^\top P Z_1 [\Psi_1 \otimes G])\) is a \(v_p \times v_p\) matrix. These matrices are equivalent to the diagonal blocks in \(C^{-1}\) corresponding to \(\tilde{f}\) and \(\tilde{\delta}_1\), respectively.

It is important to note that the components in the equations above assume the variance parameters are known. The variance parameters are unknown, however, so they must be estimated. The resulting predictions are therefore referred to as empirical BLUPs.
A.3 Estimation of loadings and specific variances

The REML estimates of the key variance parameters are obtained by maximising the residual log-likelihood, which is given by:

\[ L(y_2) = -\frac{1}{2} \left( \log |H| + \log |X^\top H^{-1} X| + y^\top Py \right) \]
\[ = -\frac{1}{2} \left( \log |H| + \log |X^\top H^{-1} X| + y_r^\top H^{-1} y_r \right), \quad (A.5) \]

where \( y_2 = L_2^\top y \) such that \( L_2^\top X = 0 \) and \( y_r = y - X \hat{\tau} \), with \( \text{var}(y_r) = H \) (Verbyla, 1990). The REML estimates are then obtained by solving a set of (score) equations, which are given by:

\[ s(\kappa) = \frac{\partial L(y_2)}{\partial \kappa^\top} = 0, \quad (A.6) \]

where \( \kappa \) is the \( b \)-vector of variance parameters in the IFA-LMM. The score equations were traditionally solved using numerical approaches based on the observed or expected information. The AI algorithm is based on the average of this information, and computes updates as:

\[ \kappa^{(m+1)} = \kappa^{(m)} + \left[ I_{a}^{(m)} \right]^{-1} s(\kappa^{(m)}), \quad (A.7) \]

where \( \kappa^{(m)} \) is the \( b \)-vector of variance parameters, \( I_{a}^{(m)} \) is the AI matrix and \( s(\kappa^{(m)}) \) contains the score equations for iteration \( m \). Note that the superscript “\( m \)” is removed in the following for brevity. The score equation for \( \kappa_i \) is given by (Gilmour et al., 1995):

\[ s(\kappa_i) = -\frac{1}{2} \left[ \text{tr}(P \dot{H}_i) - y^\top P q_i \right], \quad (A.8) \]

where \( \dot{H}_i = \frac{\partial H}{\partial \kappa_i} \) and \( q_i \) is the working variate for variance parameter \( i \), which is given by:

\[ q_i = \dot{H}_i Py. \quad (A.9) \]

Next, let the vectors of key variance parameters in the IFA-LMM be denoted by:

\[ \lambda_b = \text{vect}(\Lambda_b), \quad 1 = \text{diag}(L) \quad \text{and} \quad \Psi_1 = \text{diag}(\Psi_1), \quad (A.10) \]

where \( \Lambda_b = \begin{bmatrix} \Lambda_s \\ \Lambda_r \end{bmatrix} \).
### A.3 Estimation of loadings and specific variances

The working variates for the key variance parameters can therefore be written as:

\[
q_i = \begin{cases} 
Z \left[ B \left( \Lambda_b L \Lambda_b^\top + \tilde{\Lambda}_b L \Lambda_b^\top \right) B^\top \otimes G \right] Z_i^\top P y, & \kappa_i \in \Lambda_b \\
Z \Lambda_b \left[ I_i \otimes G \right] Z \Lambda_b^\top P y, & \kappa_i \in \Lambda_i \\
Z_i \left[ \Psi_{1i} \otimes G \right] Z_i^\top P y, & \kappa_i \in \Psi_1
\end{cases}
\]

where \( \tilde{\Lambda}_b = \frac{\partial \Lambda_b}{\partial \lambda_b} \), \( I_i = \frac{\partial I}{\partial \lambda_i} \) and \( \Psi_{1i} = \frac{\partial \Psi_{1}}{\partial \psi_{1i}} \). The working variates for the specific variances can be simplified as:

\[
q_i = Z_{1i} G Z_{1i}^\top P_i y, \quad \kappa_i \in \Psi_1
\]

where \( Z_{1i} \) is the \( n_i \times v \) design matrix for environment \( i \) with non-zero specific variance, \( P_i \) are the corresponding \( n_i \) rows of \( P \) and \( Z_{1i} \) are the corresponding \( v \) columns in \( Z_1 \).

The other important components in Eq. A.8 are the trace terms. The trace terms for the key variance parameters can therefore be written as:

\[
\text{tr}(\Phi_i) = \begin{cases} 
2 \text{tr} \left[ \left( \tilde{\Lambda}_b^\top B^\top \otimes I_i \right) Z_i^\top R^{-1} W C^{-1} \tilde{f} \right], & \kappa_i \in \Lambda_b \\
\text{tr} \left[ \left( \tilde{L}_i L^{-1} \otimes I_i \right) Z \Lambda_b^\top R^{-1} W C^{-1} \tilde{f} \right], & \kappa_i \in \Lambda_i \\
\text{tr} \left[ \left( \Psi_{1i} \Psi_{1i}^{-1} \otimes I_i \right) Z_i^\top R^{-1} W C^{-1} \tilde{\delta}_i \right], & \kappa_i \in \Psi_1
\end{cases}
\]

where \( C^{-1} \) are the \( vk \) columns in \( C^{-1} \) corresponding to \( \tilde{f} \) and \( C^{-1} \tilde{\delta}_i \) are the \( vp \) columns corresponding to \( \tilde{\delta}_i \). The trace terms for the specific variances can be further simplified as:

\[
\text{tr}(\Phi_i) = \frac{1}{\psi_{1i}} \text{tr} \left[ Z_{1i} R_{1i}^{-1} W_i C^{-1} \tilde{\delta}_{1i} \right], \quad \kappa_i \in \Psi_1
\]

where \( R_i \) is diagonal block \( i \) of \( R \) corresponding to environment \( i \), \( W_i \) are the \( n_i \) rows in \( W \) and \( C^{-1} \tilde{\delta}_{1i} \) are the \( v \) columns in \( C^{-1} \) corresponding to \( \tilde{\delta}_{1i} \). The trace terms avoid working with the dense genomic relationship matrix. When \( G \) is not prohibitively large, the trace term for the specific variances can also be computed as:

\[
\text{tr}(\Phi_i) = \frac{\psi_{1i}}{\psi_{1}} - \text{tr} \left[ GC^{-1} \eta_{1i} \eta_{1i} \right], \quad \kappa_i \in \Psi_1
\]

where \( C^{-1} \eta_{1i} \) is the prediction error variance matrix of \( \eta_{1i} = Z_{1i}^\top P_i y \).

Lastly, the AI matrix in Eq. A.7 is given by:

\[
I_a = \frac{1}{2} Q^\top P Q, \quad (A.11)
\]

where \( Q = [q_1 \ q_2 \ldots \ q_b] \) is the \( n \times b \) matrix of working variates across all \( b \) variance parameters. The AI matrix is obtained via absorption of \( C \) onto \( Q^\top R^{-1} Q \) (Smith, 1999).
A.4 Trimming the regression residuals

The working variates and trace terms for the specific variances can be further simplified for incomplete data, that is by only considering the regression residuals with phenotypic data. Mazur (2021) refer to this operation as “trimming”.

Let the \( v \)-vector of regression residuals for environment \( i \) with non-zero specific variance be partitioned as \( \delta_{i;1} = (\delta_{i;1}^T, \delta_{i;0}^T)^T \), where \( \delta_{i;1} \) is the \( v_i \)-vector corresponding to genotypes with phenotypic data and \( \delta_{i;0} \) is the \((v - v_i)\)-vector corresponding to genotypes without phenotypic data. Also let the \( n_i \times v \) design matrix be partitioned conformably as \( Z_{1i} = [Z_{1i;1} \ 0] \), where \( Z_{1i;1} \) is a \( n_i \times v_i \) matrix.

It is assumed that:

\[
\begin{bmatrix}
\delta_{i;1} \\
\delta_{i;0}
\end{bmatrix} \sim N\left(\begin{bmatrix}
0 \\
0
\end{bmatrix}, \psi_{i1} \begin{bmatrix}
G_{i;11} & G_{i;10} \\
G_{i;01} & G_{i;00}
\end{bmatrix}\right).
\]

Trimming is possible for the regression residuals since \( \delta_{i;0} \) does not contribute to \( H \), and thence does not contribute to the residual (or full) log-likelihood. It is therefore more efficient to fit \( Z_{1i;1} \delta_{i;1} \) for environment \( i \) instead of \( Z_{1i} \delta_{ii} \). The BLUPs of the trimmed regression residuals are then given by:

\[
\tilde{\delta}_{i;1} = \psi_{i1} G_{i;11} Z_{1i;1}^T P_i y.
\] (A.12)

The working variates and trace terms in Eq. A.8 can be written as:

\[
q_i = Z_{1i;1} G_{i;11} Z_{1i;1}^T P_i y
\]

\[
\text{tr}(\Phi_i) = \frac{1}{\psi_{i1}} \text{tr}\left[ Z_{1i;1}^T R_i^{-1} W_i C \delta_{i;1} \right]
\]

\[
= \frac{v_i}{\psi_{i1}} - \text{tr}\left[ G_{i;11} C^\eta_{i;1} \eta_{i;1} \right], \quad \kappa_i \in \psi_i
\]

where \( Z_{1;1} \) are the \( v_i \) columns in \( Z_1 \) corresponding to \( \delta_{1;1}, \) \( C^\theta_{i;1} \) are the \( v_i \) columns in \( C^\tilde{\delta}_{i;i} \) and \( C^\eta_{i;1} \eta_{i;1} \) is the prediction error variance matrix of \( Z_{1;1}^T P_i y \).

Lastly, the BLUPs of the regression residuals without phenotypic data can be obtained after REML estimation as:

\[
\tilde{\delta}_{i;0} = G_{i;01} G_{i;11}^{-1} \delta_{i;1}
\]

\[
= \psi_{i1} G_{i;01} Z_{1i;1}^T P_i y.
\] (A.13)
Appendix B

Proof of invariance

This appendix provides a mathematical proof of the invariance of the factor analytic regression (FAR$k$) model developed in Chapter 3. Firstly, note there is an error in the published manuscript, which states “Jennrich and Schluchter (1986) proposed an extension of the random regression model which includes a factor analytic model for the known environmental covariates” - they did not propose a factor analytic variance matrix for the intercepts and slopes. Buntaran et al. (2021) did suggest that a full rank factor analytic variance matrix could be used for this purpose, however, they neither developed nor demonstrated fitting this model. The FAR$k$ model developed in Chapter 3 produces a reduced rank factor analytic variance matrix for the intercepts and slopes, with rank equal to $k$.

The most general variance matrix for the intercepts and slopes is the unstructured model, which fits a separate variance for each term and a separate covariance for each pair of terms. The unstructured model is scale, translational and rotational invariant. The reduced rank factor analytic variance matrix provides an approximation to the unstructured variance matrix, and comprises $(q+1)k - k(k-1)/2$ parameters compared to $(q+1)(q+2)/2$ parameters. Despite this approximation, the FAR$k$ model is also scale, translational and rotational invariant. A mathematical proof is provided in the following.

B.1 Preliminaries

Assume all $p$ environments have $q$ known covariates available, such that $p > q$. Let $A = [1_p^* \ S]$ be a $p \times (q+1)$ matrix of basis functions (assumed to have full column rank), where $1_p^* = 1_p/\sqrt{p}$ is a $p$-vector with unit length and $S = [s_1 \ldots s_q]$ is a $p \times q$ matrix of environmental covariates. Also let $\Lambda_0 = [\Lambda_1 \ldots \Lambda_k]$ be a $(q+1) \times k$ matrix of loadings, where $\Lambda_1 = [\lambda_{11} \ldots \lambda_{1k}]$
Proof of invariance

is a $k$ row-vector corresponding to the genotype intercepts and $\Lambda_s = [\lambda_{s1} \ldots \lambda_{sk}]$ is a $q \times k$ matrix corresponding to the genotype slopes on the known covariates.

### B.2 Scale invariance

Let $A^* = [1^* \ S] W$, where $W = \oplus_{i=0}^q w_i$ is a $(q + 1) \times (q + 1)$ diagonal matrix comprising a separate non-zero scaling constant for the intercepts and slopes on each covariate, often with $w_0 = 1$. The matrix $W$ and its inverse are given by:

$$W = \begin{bmatrix}
w_0 & 0 & \cdots & 0 \\
0 & w_1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & w_q
\end{bmatrix} \quad \text{and} \quad W^{-1} = \begin{bmatrix}
w_0^{-1} & 0 & \cdots & 0 \\
0 & w_1^{-1} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & w_q^{-1}
\end{bmatrix}.$$

When $w_0 = 1$, it follows that the top left element of $W$ and $W^{-1}$ will equal one.

The FAR$k$ model is scale invariant provided there exists a $(q + 1) \times k$ matrix, $\Lambda^*_a$, such that:

$$AA_aLL_a^T = A^* \Lambda^*_aL^T A^*, \quad (B.1)$$

where $L = \oplus_{r=1}^k l_r$ is a $k \times k$ diagonal variance matrix comprising a separate variance for each common factor which are sorted in decreasing order.

It follows that:

$$\Lambda^*_a = (A^*A^*)^{-1}A^T A A_a$$

$$= W^{-1} \Lambda_a, \quad (B.2)$$

noting that $A^* \Lambda^*_a = \Lambda A_a$ since $A^*(A^*A^*)^{-1}A^T A = A$ and $WA_a^* = \Lambda_a$.

Substituting Eq. B.2 into Eq. B.1 gives:

$$A^* \Lambda^*_aL^T A^* = AWW^{-1} \Lambda_aL^T A^T W^{-1} W^T A^T$$

$$= AA_aL^T A^T.$$

It also follows that if $\hat{\Lambda}_a$ is the REML estimate of $\Lambda_a$, then $\hat{\Lambda}_a^* = W^{-1} \hat{\Lambda}_a$ is an estimate of $\Lambda^*_a$ which satisfies Eq. B.1. Consequently, when $A^* = [1^* \ S] W$ is fitted instead of $A = [1^* \ S]$, the loadings matrix $\hat{\Lambda}_a^* = W^{-1} \hat{\Lambda}_a$ is obtained instead of $\hat{\Lambda}_a$. Note, however, this solution is not unique as orthogonal rotations of $\hat{\Lambda}_a^*$ also satisfy Eq. B.1, e.g. $\hat{\Lambda}_a^* = -W^{-1} \hat{\Lambda}_a$. 
B.3 Translational invariance

Let \( A^* = [I_p^* S] - x^T \otimes I_p \), where \( x = (x_0, \ldots, x_q)^T \) is a \( q + 1 \) vector comprising a separate translating constant for the intercepts and slopes on each covariate, often with \( x_0 = 0 \). This matrix can be rewritten as \( A^* = [I_p^* S]X \), where \( X = I_{q+1} - (A^T A)^{-1} A^T [x^T \otimes I_p] \), noting that \( A(A^T A)^{-1} A^T [x^T \otimes I_p] = [x^T \otimes I_p] \) and \( X^{-1} \exists \).

The matrix \( X \) and it’s inverse are given by:

\[
X = \begin{bmatrix}
    s_0^* / s_0 & -x_1 / s_0 & \ldots & -x_q / s_0 \\
    0 & 1 & \ldots & 0 \\
    \vdots & \vdots & \ddots & \vdots \\
    0 & 0 & \ldots & 1
\end{bmatrix}
\quad \text{and} \quad
X^{-1} = \begin{bmatrix}
    s_0^* / s_0 & x_1 / s_0^* & \ldots & x_q / s_0^* \\
    0 & 1 & \ldots & 0 \\
    \vdots & \vdots & \ddots & \vdots \\
    0 & 0 & \ldots & 1
\end{bmatrix},
\]

where \( s_0^* = s_0 - x_0 \) and \( s_0 = 1/\sqrt{p} \), i.e. \( s_0^* = 1/\sqrt{p} - x_0 \). When \( x_0 = 0 \), it follows that \( s_0^* = s_0 \), and the top left element of \( X \) and \( X^{-1} \) will equal one.

The FAR\( k \) model is translational invariant provided there exists a \((q+1) \times k\) matrix, \( \Lambda^*_a \), such that:

\[
A \Lambda^*_a L \Lambda^*_a^T A^T = A^* \Lambda^*_a L \Lambda^*_a^T A^*^T,
\]  
(B.3)

where all parameters are as previously defined.

It follows that:

\[
\Lambda^*_a = (A^* A^*)^{-1} A^* A \Lambda_a \\
= X^{-1} \Lambda_a,
\]  
(B.4)

noting that \( A^* \Lambda_a^* = \Lambda \Lambda_a \) since \( A^* (A^* A^*)^{-1} A^* A = A \) and \( X \Lambda_a^* = \Lambda_a \).

Substituting Eq. B.4 into Eq. B.3 gives:

\[
A^* \Lambda^*_a L \Lambda^*_a^T A^*^T = AXX^{-1} \Lambda_a L \Lambda_a^T X^{-1} X^T A^T
\]

\[
= \Lambda \Lambda_a L \Lambda_a^T A^T.
\]

It also follows that if \( \hat{\Lambda}_a \) is the REML estimate of \( \Lambda_a \), then \( \hat{\Lambda}_a^* = X^{-1} \hat{\Lambda}_a \) is an estimate of \( \Lambda_a^* \) which satisfies Eq. B.3. Consequently, when \( A^* = [I_p^* S]X \) is fitted instead of \( A = [I_p^* S] \), the loadings matrix \( \hat{\Lambda}_a^* = X^{-1} \hat{\Lambda}_a \) is obtained instead of \( \hat{\Lambda}_a \). Note, however, this solution is not unique as orthogonal rotations of \( \hat{\Lambda}_a^* \) also satisfy Eq. B.3, e.g. from multiplying \( \hat{\Lambda}_a^* \) by -1 or from constraining the upper right elements of \( \hat{\Lambda}_a^* \) to zero during estimation.
### B.4 Rotational invariance

Let \( \mathbf{A}^* = [\mathbf{1}_p \, \mathbf{S}] \mathbf{V} \), where \( \mathbf{V} \) is a \((q+1) \times (q+1)\) general rotation matrix (assumed to have full rank), which is often non-symmetric. The matrix \( \mathbf{V} \) is given by:

\[
\mathbf{V} = \begin{bmatrix}
  v_{00} & v_{10} & \cdots & v_{q0} \\
  v_{01} & v_{11} & \cdots & v_{q1} \\
  \vdots & \vdots & \ddots & \vdots \\
  v_{0q} & v_{1q} & \cdots & v_{qq}
\end{bmatrix},
\]

with generalised inverse given by \( \mathbf{V}^{-1} = (\mathbf{V}^\top \mathbf{V})^{-1} \mathbf{V}^\top \), noting that \( \mathbf{V}^{-1} = \mathbf{V} \mathbf{V}^{-1} = \mathbf{I}_{q+1} \) since \( \mathbf{V} \) is square with full rank.

The FAR\(k\) model is rotational invariant provided there exists a \((q+1) \times k\) matrix, \( \mathbf{\Lambda}^*_a \), such that:

\[
\mathbf{\Lambda} \mathbf{\Lambda}_a \mathbf{L} \mathbf{\Lambda}_a^\top \mathbf{A}^\top = \mathbf{A}^* \mathbf{\Lambda}^*_a \mathbf{L} \mathbf{\Lambda}_a^\top \mathbf{A}^* \mathbf{A}^\top, \quad (B.5)
\]

where all parameters are as previously defined.

It follows that:

\[
\mathbf{\Lambda}^*_a = (\mathbf{A}^* \mathbf{A}^*)^{-1} \mathbf{A}^* \mathbf{A} \mathbf{\Lambda}_a \\
= \mathbf{V}^{-1} \mathbf{\Lambda}_a, \quad (B.6)
\]

noting that \( \mathbf{A}^* \mathbf{\Lambda}^*_a = \mathbf{\Lambda}_a \mathbf{\Lambda}^*_a \) since \( \mathbf{A}^* (\mathbf{A}^* \mathbf{A}^*)^{-1} \mathbf{A}^* \mathbf{A} = \mathbf{A} \) and \( \mathbf{V} \mathbf{\Lambda}^*_a = \mathbf{\Lambda}_a \).

Substituting Eq. B.6 into Eq. B.5 gives:

\[
\mathbf{A}^* \mathbf{\Lambda}_a^* \mathbf{L} \mathbf{\Lambda}_a^* \mathbf{A}^* \mathbf{A}^\top = \mathbf{A} \mathbf{V} \mathbf{V}^{-1} \mathbf{\Lambda}_a \mathbf{L} \mathbf{\Lambda}_a^\top \mathbf{V}^{-\top} \mathbf{V}^\top \mathbf{A}^\top \\
= \mathbf{A} \mathbf{\Lambda}_a \mathbf{L} \mathbf{\Lambda}_a^\top \mathbf{A}^\top.
\]

It also follows that if \( \hat{\mathbf{\Lambda}}_a \) is the REML estimate of \( \mathbf{\Lambda}_a \), then \( \hat{\mathbf{\Lambda}}_a^* = \mathbf{V}^{-1} \hat{\mathbf{\Lambda}}_a \) is an estimate of \( \mathbf{\Lambda}^*_a \) which satisfies Eq. B.5. Consequently, when \( \mathbf{A}^* = [\mathbf{1}_p \, \mathbf{S}] \mathbf{V} \) is fitted instead of \( \mathbf{A} = [\mathbf{1}_p \, \mathbf{S}] \), the loadings matrix \( \hat{\mathbf{\Lambda}}_a^* = \mathbf{V}^{-1} \hat{\mathbf{\Lambda}}_a \) is obtained instead of \( \hat{\mathbf{\Lambda}}_a \). Note, however, this solution is not unique as orthogonal rotations of \( \hat{\mathbf{\Lambda}}_a^* \) also satisfy Eq. B.5, e.g. from multiplying \( \hat{\mathbf{\Lambda}}_a^* \) by -1 or from constraining the upper right elements of \( \hat{\mathbf{\Lambda}}_a^* \) to zero during estimation. The methods above can be applied to cases where \( \mathbf{A} \) is simultaneously scaled, translated and rotated.

Lastly, note that the addition of a specific variance matrix, \( \mathbf{\Psi}_a \), precludes the full rank FAR\(k\) model from being translational and rotational invariant because a diagonal matrix does not exist which satisfies the equivalent forms of Eqs. B.3 and B.5. This was a key motivator to the development of the reduced rank FAR\(k\) model in Chapter 3.
Appendix C

General extensions

The selection tools in Chapter 3 provide measures of overall performance, responsiveness and stability with regards to the conventional FA$k$ model. The selection tools can be extended for any random regression type model, including the FAR$k$ model developed in Chapter 2. The extension requires the environmental covariates and genotype slopes to be rotated, which can be achieved using a principal component rotation or the new rotation developed in Chapter 3. The appealing feature of the new rotation is that it considers non-crossover and crossover GEI associated with all covariates together, rather than each covariate separately. A measure of responsiveness will also be developed for cases where the response of genotypes to individual covariates is of interest.

C.1 Rotating the covariates and scores

Constraints are required in the FAR$k$ model to ensure unique solutions for $\Lambda_1$, $\Lambda_s$ and $L$ during estimation. The upper right elements of $\Lambda_s$ are set to zero when $k > 1$ and $L$ is set to $I_k$. Let the environmental loadings and genotype scores with these constraints be denoted by $\Lambda^*_1$, $\Lambda^*_s$ and $f^*$, such that $f^* \sim N(0, I_k \otimes G)$. The loadings and scores are generally rotated after estimation to enable interpretation.

Two rotations are considered below:

1. A principal component rotation, which ensures the common factors are orthogonal and sorted in decreasing magnitude.

2. The new rotation developed in Chapter 3, which ensures the first common factor exclusively captures non-crossover GEI and the higher order factors are orthogonal, sorted in decreasing magnitude and exclusively capture crossover GEI.
Principal component rotation

The principal component rotation is given by:

$$
\begin{bmatrix}
\Lambda_1 \\
\Lambda_s
\end{bmatrix} =
\begin{bmatrix}
\Lambda_1^* \\
\Lambda_s^*
\end{bmatrix} V L^{-1/2} \quad \text{and} \quad f = (L^{-1/2} V^T \otimes I_v) f^* ,
$$

(C.1)

which is based on the singular value decomposition given by:

$$
1_p A_1^* + S A_s^* = U L^{1/2} V^T ,
$$

where $U = [u_1 \ldots u_k]$ is a $p \times k$ orthogonal matrix of left singular vectors, $V = [v_1 \ldots v_k]$ is a $k \times k$ orthogonal matrix of right singular vectors and $L = \bigoplus_{r=1}^k l_r$ is a $k \times k$ diagonal matrix of singular values sorted in decreasing magnitude, i.e. $l_1 > \ldots > l_k$.

It then follows that:

$$
\begin{bmatrix}
\Lambda_1 \\
\Lambda_s
\end{bmatrix} \equiv U \quad \text{and} \quad f \sim N(0, L \otimes G).
$$

(C.2)

The selection tools in Section 3.12 are then constructed using the rotated loadings and scores in Eq. C.1. Applying the principal component rotation in this manner will be contingent on the ability of the FAR$k$ model to capture non-crossover and crossover GEI within the known environmental covariates. It also assumes that variation associated with all known covariates is of interest, rather than individual covariates. A separate rotation can be obtained for each known covariate where required.

The new rotation

The objective of the new rotation is to transfer all non-crossover variation to the genotype main effects and leave all crossover variation in the regression on known environmental covariates. This is achieved by considering non-crossover and crossover GEI associated with all known covariates together, rather than each covariate separately.

The new rotation can be summarised by three key steps:

1. Obtain genotype main effects and slopes based on the regression.
2. Construct generalised main effects by transferring all non-crossover variation to the genotype main effects, adjust and update the genotype slopes.
3. Construct the first factor based on the generalised main effects and higher order factors based on the genotype slopes, then rotate to their principal component solution.
C.1 Rotating the covariates and scores

Each step is detailed in the following.

1. The genotype main effects and slopes are given by:

\[ u_1 = (\Lambda_1^* \otimes I_v)f^* \quad \text{and} \quad u_s = (\Lambda_s^* \otimes I_v)f^*, \]

where \( u_1 \) is the \( v \)-vector of genotype main effects and \( u_s \) is the \( vq \)-vector of genotype slopes. Note that the main effects are equivalent to intercepts for the FAR\( k \) model because the known environmental covariates in \( S \) are assumed to be centred, i.e. \( 1_p^T S = 0 \).

It then follows that:

\[ \begin{bmatrix} u_1 \\ u_s \end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Lambda_1^* \Lambda_1^{*T} & \Lambda_1^* \Lambda_s^{*T} \\ \Lambda_s^* \Lambda_1^{*T} & \Lambda_s^* \Lambda_s^{*T} \end{bmatrix} \otimes G \right), \]

such that a covariance is present between the main effects and slopes.

The variance matrix can be generalised as (ignoring the relationship matrix):

\[ \begin{bmatrix} \sigma_1^2 & \Sigma_{1s} \\ \Sigma_{s1} & \Sigma_{ss} \end{bmatrix} = \begin{bmatrix} \Lambda_1^* \Lambda_1^{*T} & \Lambda_1^* \Lambda_s^{*T} \\ \Lambda_s^* \Lambda_1^{*T} & \Lambda_s^* \Lambda_s^{*T} \end{bmatrix}, \tag{C.3} \]

where \( \sigma_1^2 \) is the main effect variance, \( \Sigma_{s1} = \Sigma_{1s}^T \) is a \( q \)-vector with elements given by the covariance between the main effects and slopes for each known covariate and \( \Sigma_{ss} \) is a \( q \times q \) variance matrix for the slopes. It is important to note that the covariances in \( \Sigma_{s1} \) (and \( \Sigma_{1s} \)) capture variation associated with the main effects, so they capture non-crossover GEI (see Falconer, 1990; Waters et al., 2023, for a similar argument). Any remaining variation in \( \Sigma_{ss} \), after adjusting for \( \Sigma_{s1} \), will then capture crossover GEI. This result will be utilised in the next step to transfer non-crossover variation from the slopes to the main effects.

2. Following Waters et al. (2023), the genotype main effects and slopes can be adjusted for their covariance, which gives:

\[ u_1^* = (1_p \otimes u_1) + (S \Sigma_{s1} / \sigma_1^2 \otimes u_1) \quad \text{and} \quad u_s^* = u_s - (\Sigma_{s1} / \sigma_1^2 \otimes u_1), \tag{C.4} \]

where \( u_1^* \) is a \( vp \)-vector of generalised main effects and \( u_s^* \) is a \( vq \)-vector of slopes adjusted for the main effects.

It then follows that:

\[ \begin{bmatrix} u_1^* \\ u_s^* \end{bmatrix} \sim N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 I_p + \Omega_{1s} + \Omega_{s1} + S \Sigma_{s1} \Sigma_{1s} S^T / \sigma_1^2 & 0 \\ 0 & \Sigma_{ss} - \Sigma_{s1} \Sigma_{1s} / \sigma_1^2 \end{bmatrix} \otimes G \right), \]

where \( \Omega_{1s} = 1_p \Sigma_{1s} S^T \) and \( \Omega_{s1} = \Omega_{1s}^T \), such that the generalised main effects and genotype
slopes are now independent because the covariances in \( \Sigma_{s1} \) have been transferred entirely to \( u_1^* \). This approach ensures that all variation associated with the genotype main effects is transferred from the slopes to the main effects. However, it does not guarantee that non-crossover variation is exclusively transferred, because this is also a function of the known environmental covariates. In Chapter 3, it was demonstrated when the variation associated with the main effects results in crossover variation, i.e. when at least one covariance in \( \Sigma_{21} \) from Eq. 3.17 is negative. A similar result applies here, and simple constraints can also be imposed so that only variation in \( \Sigma_{s1} \) due to non-crossover GEI is transferred to the generalised main effects.

The generalised main effects and genotype slopes can then be updated as:

\[
\begin{align*}
\mathbf{u}_1^* &= (\mathbf{1}_p \otimes \mathbf{u}_1) + (\mathbf{SS}_{s1}/\sigma_1^{*2} \otimes \mathbf{u}_1) \quad \text{and} \quad \mathbf{u}_s^* = \mathbf{u}_s - (\Sigma_{s1}/\sigma_1^{*2} \otimes \mathbf{u}_1),
\end{align*}
\]

where \( \sigma_1^{*2} = \sigma_1^2 - \sigma_{s1}^* \) is an adjusted main effect variance and \( \sigma_{s1}^* \) is the minimum negative covariance in \( \mathbf{SS}_{s1} + \sigma_1^2 \mathbf{1}_p \), i.e \( \sigma_{s1}^* = \min(\mathbf{SS}_{s1} + \sigma_1^2 \mathbf{1}_p,0) \). The constraints therefore reduce the amount of variation associated with the main effects that is transferred from the slopes to the main effects, so that only non-crossover variation is transferred. Note that setting \( \sigma_{s1}^* \) in this manner enables the methods to be generalised for cases where there are no negative covariances, i.e. where \( \sigma_{s1}^* = 0 \).

The variance matrix for \( \mathbf{u}_1^* \) and \( \mathbf{u}_s^* \) is given by (ignoring the relationship matrix):

\[
\begin{align*}
\begin{bmatrix}
\sigma_1^2(\sigma_1^{*2})^2 \mathbf{I}_p + \Omega_{ls}^* + \Omega_{s1}^* + \sigma_1^2 \mathbf{SS}_{s1} \mathbf{SS}_{ls} \mathbf{S}^T \\
\Sigma_{ls} \sigma_1^2 \mathbf{I}_p + \Sigma_{s1}
\end{bmatrix} &\mathbf{SS}_{ss} - 2\sigma_1^{*2} \mathbf{SS}_{s1} \mathbf{SS}_{ls} + \sigma_1^2 \mathbf{SS}_{s1} \mathbf{SS}_{ls}
\end{align*}
\]

\[
/ (\sigma_1^{*2})^2,
\]

where \( \Sigma_{ss}^* = (\sigma_1^{*2})^2 \mathbf{SS}_{ss}, \Omega_{ls}^* = \sigma_1^{*2} \mathbf{I}_p \mathbf{SS}_{ls} \mathbf{S}^T \) and \( \Omega_{s1}^* = \Omega_{ls}^{*T}, \) such that the generalised main effects and genotype slopes are now dependent because the part of the covariances in \( \Sigma_{s1} \) attributed to crossover GEI remains in the genotype slopes. Note, however, this is not the case when all covariances in \( \mathbf{SS}_{s1} + \sigma_1^2 \mathbf{1}_p \) are positive, and the variance matrix above reverts back to the form where the generalised main effects and genotype slopes are completely independent.

3. The first common factor is based on the generalised main effects, and is constructed as:

\[
\begin{bmatrix}
\lambda_{11} \\
\lambda_{s1}
\end{bmatrix} = \left[ \begin{array}{c}
\sigma_1^{*2} \\
\Sigma_{s1}
\end{array} \right] / \sqrt{\Sigma_{ls} \Sigma_{21}} \quad \text{and} \quad \mathbf{f}_1 = \sqrt{\Sigma_{ls} \Sigma_{21}} \mathbf{u}_1 / \sigma_1^{*2},
\]

with \( \Sigma_{12} = \sigma_1^{*2} \mathbf{I}_p^T + \Sigma_{ls} \mathbf{S}^T \) and \( \Sigma_{21} = \Sigma_{12}^T, \) such that \( [\mathbf{1}_p \lambda_{11} + \mathbf{S} \lambda_{s1}]^T [\mathbf{1}_p \lambda_{11} + \mathbf{S} \lambda_{s1}] = 1 \) and \( \mathbf{f}_1 \sim \mathcal{N}(0, \mathbf{I}_l \mathbf{G}) \), where \( l_1 = p \sigma_1^{*2} + \Sigma_{ls} \mathbf{S}^T \Sigma_{s1} \sigma_1^{*2} / (\sigma_1^{*2})^2. \)
The higher order factors are based on the genotype slopes adjusted for the generalised main effects, and are constructed as:

$$\begin{bmatrix}
\Lambda_{122} \\
\Lambda_{s22}
\end{bmatrix} = 
\begin{bmatrix}
0 \\
\Lambda_s^* - \Sigma_{s1} \Lambda_1^* / \sigma_1^2
\end{bmatrix} V_{22} L_{22}^{-1/2} \text{ and } f_{22} = (L_{22}^{1/2} V_{22}^T \otimes I_v) f^*, \quad (C.7)$$

which is based on the singular value decomposition given by:

$$S[\Lambda_s^* - \Sigma_{s1} \Lambda_1^* / \sigma_1^2] = U_{22} L_{22}^{1/2} V_{22}^T,$$

where $U_{22} = [\mathbf{u}_2 \ldots \mathbf{u}_{k+1}]$ is a $p \times k$ orthogonal matrix of left singular vectors, $V_{22} = [\mathbf{v}_2 \ldots \mathbf{v}_{k+1}]$ is a $k \times k$ orthogonal matrix of right singular vectors and $L_{22} = \oplus_{r=2}^{k+1} l_r$ is a $k \times k$ diagonal matrix of singular values sorted in decreasing magnitude, i.e. $l_2 > \ldots > l_{k+1}$.

It then follows that:

$$S \Lambda_{s22} = U_{22} \quad \text{and} \quad f_{22} \sim N(0, L_{22} \otimes G),$$

such that the higher order factors are orthogonal and sum to zero, i.e. $\Lambda_{s22} = \mathbf{I}_k$ and $1_p^T S \Lambda_{s22} = 0$.

### C.2 Selection tools

The measure of overall performance is given by:

$$\text{OP} = \bar{\lambda}_1 f_1,$$

where $\bar{\lambda}_1$ is the mean environmental loading for the first common factor and $f_1$ is the $v$-vector of genotype scores. The OP measure for a genotype is therefore equal to the fitted value of the first factor regression, which is equivalent to its main effect for the new rotation. The adjusted OP measure in Eq. 3.14 can also be obtained, but note that this measure is exactly the same as the conventional measure for the new rotation because the higher order factors sum to zero.

The measure of responsiveness for common factor $r$ is given by:

$$\text{RESP}_r = (\bar{\lambda}_r^+ - \bar{\lambda}_r^-) f_r,$$

where $\bar{\lambda}_r^+$ is the mean of the positive environmental loadings for factor $r$, $\bar{\lambda}_r^-$ is the mean of the negative environmental loadings and $f_r$ is the $v$-vector of genotype scores.
A similar measure can be obtained for known environmental covariate $h$ given by:

$$\text{RESP}_h = (\bar{s}_h^+ - \bar{s}_h^-)u_{sh}^*,$$

where $\bar{s}_h^+$ is the mean of the positive environmental scores for known covariate $h$, $\bar{s}_h^-$ is the mean of the negative environmental scores and $u_{sh}^*$ is the $v$-vector of genotype slopes which have been adjusted for the non-crossover variation associated with the main effects.

The measure of stability is given by:

$$\text{RMSD} = \sqrt{\text{diag}(\mathbf{EE}^\top)/p},$$

where $\mathbf{E} = \mathbf{FA}^\top - \mathbf{f}_1\mathbf{A}_1^\top$ is a $v \times p$ matrix of common GE effects adjusted for the first factor and $\mathbf{F} = [\mathbf{f}_1 \ldots \mathbf{f}_k]$ is a $v \times k$ matrix of genotype scores. The adjusted RMSD measure in Eq. 3.28 is also a straightforward extension, but note that the adjusted measure is exactly the same as the conventional measure for the new rotation.
Appendix D

Hypothetical examples

This section demonstrates the new rotation in Chapter 3 using four hypothetical examples with \( v = 10 \) genotypes and \( p = 3 \) environments.

Four scenarios are considered below for the GE effects:

1. Entirely non-crossover GEI, no crossover GEI.

2. Entirely crossover GEI, no non-crossover GEI and thence no main effect.

3. Some non-crossover and crossover GEI.

4. Some non-crossover and little crossover GEI.

Each scenario will be demonstrated using either one or two common factors, but note that the results hold for any number of factors. The examples are presented in Figures D.1 to D.4.

1. Entirely non-crossover GEI, no crossover GEI (Figure D.1)

The between-environment genetic variance and genetic correlation matrices are given by:

\[
G_a = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{bmatrix} \quad \text{and} \quad C_a = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}, \quad (D.1)
\]

with environmental loadings and genotype scores initially given by:

\[
\Lambda^* = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad \text{and} \quad f^* \sim N(0, I_v),
\]

and total genetic variance of \( 14/p \), i.e. 4.67.
Fig. D.1 Latent regression plots for Example 1 showing the form of the regression before and after rotation. The genotype main effects are denoted by crosses, which are equivalent to the y-intercepts in the bottom right plot. The common GE effects are distinguished by shape.

The variance matrix between the genotype main effects and common GE effects is:

$$
\begin{pmatrix}
\bar{\Lambda}^* \bar{\Lambda}^{*T} & \bar{\Lambda}^{*T} \\
\Lambda' \bar{\Lambda}^{*T} & \Lambda' \bar{\Lambda}^{*T}
\end{pmatrix}
= 
\begin{bmatrix}
4 & 2 & 4 & 6 \\
2 & 1 & 2 & 3 \\
4 & 2 & 4 & 6 \\
6 & 3 & 6 & 9
\end{bmatrix},
$$

where $\bar{\Lambda}' = 2$. The common GE effects for each environment are completely explained by the genotype main effects, i.e. all variation is due to non-crossover GEI.
After rotation, the loadings and scores become:

\[
\Lambda = \begin{bmatrix}
0.27 \\
0.53 \\
0.80 \\
\end{bmatrix}
\text{ and } f \sim \mathcal{N}(0, 14I_v)
\] (D.2)

where \(\Lambda = \Lambda^*/\sqrt{14}\) and \(f = \sqrt{14}f^*\). The first and only factor exclusively captures non-crossover GEI. This is the same result for the principal component rotation and the new rotation developed in Chapter 3.

Using Eq. 3.23, the generalised main effects can be written as:

\[
u^*_1 = (1_p \otimes u_1) + (\Lambda^*_1 \otimes I_v) f^*_1,
\]

with loadings and scores given by:

\[
[1_p \ \Lambda^*_1] = \begin{bmatrix}
1.00 & -0.71 \\
1.00 & 0.00 \\
1.00 & 0.71 \\
\end{bmatrix}
\text{ and } \begin{bmatrix}
u_1 \\
f^*_1
\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}0 \\
0 \end{bmatrix}, \begin{bmatrix}4.00 & 2.83 \\
2.83 & 2.00
\end{bmatrix} \otimes I_v\right)
\]

where \(\Lambda^*_1 = [\Lambda - 0.53 \ 1_p]/\sqrt{0.14}\) and \(f^*_1 = \sqrt{0.14}f\). The first column in \([1_p \ \Lambda^*_1]\) corresponds to the genotype main effects (intercepts) and the second column corresponds to the deviations around the main effects. The non-crossover variance of 4.67 is partitioned into a main effect variance of 4.00 and an interaction variance of 0.67, which is attributed entirely to heterogeneity of scale.

2. Entirely crossover GEI, no non-crossover GEI and thence no main effect (Figure D.2)

The between-environment genetic variance and genetic correlation matrices are given by:

\[
G_a = \begin{bmatrix}
1 & 2 & -3 \\
2 & 4 & -6 \\
-3 & -6 & 9
\end{bmatrix}
\text{ and } C_a = \begin{bmatrix}
1 & 1 & -1 \\
1 & 1 & -1 \\
-1 & -1 & 1
\end{bmatrix}
\] (D.3)

with environmental loadings and genotype scores initially given by:

\[
\Lambda^* = \begin{bmatrix}
-1 \\
-2 \\
3
\end{bmatrix}
\text{ and } f^* \sim \mathcal{N}(0, I_v)
\]

and total genetic variance of 4.67.
Fig. D.2 Latent regression plots for Example 2 showing the form of the regression before and after rotation. The genotype main effects are denoted by \textit{crosses}, which are zero for this example. The common GE effects for different environments are distinguished by \textit{shape}.

The variance matrix between the genotype main effects and common GE effects is:

$$
\begin{bmatrix}
\tilde{\Lambda}^* \tilde{\Lambda}^{*T} & \tilde{\Lambda}^* \tilde{\Lambda}^{*T} \\
\tilde{\Lambda}^* \tilde{\Lambda}^{*T} & \tilde{\Lambda}^* \tilde{\Lambda}^{*T}
\end{bmatrix}
= 
\begin{bmatrix}
0 & 0 & 0 & 0 \\
0 & 1 & 2 & -3 \\
0 & 2 & 4 & -6 \\
0 & -3 & -6 & 9
\end{bmatrix},
$$

where $\tilde{\Lambda}^* = 0$. There is no main effect variance, i.e. all variation is due to crossover GEI.
After applying a principal component rotation, the loadings and scores become:

\[
\Lambda = \begin{bmatrix}
-0.27 \\
-0.53 \\
0.80 
\end{bmatrix}
\quad \text{and} \quad f \sim N(0, 14I_v),
\]  

(D.4)

where \(\Lambda = \Lambda^*/\sqrt{14}\) and \(f = \sqrt{14}f^*\). The first and only factor captures crossover GEI.

After applying the new rotation, the loadings and scores become:

\[
\Lambda = \begin{bmatrix}
0.00 & -0.27 \\
0.00 & -0.53 \\
0.00 & 0.80 
\end{bmatrix}
\quad \text{and} \quad f \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 14 \end{bmatrix} \otimes I_v\right).
\]  

(D.5)

The first factor reverts to zero for this example because there is no non-crossover GEI. The crossover variance of 4.67 is partitioned into variation due to lack of genetic correlation of 4.00 and heterogeneity of scale of 0.67. This example demonstrates that it is possible to have heterogeneity of scale without non-crossover variance, but it is not possible to have non-crossover variance without a genotype main effect.

3. Some non-crossover and crossover GEI (Figure D.3)

The between-environment genetic variance and genetic correlation matrices are given by:

\[
G_a = \begin{bmatrix}
1 & -2 & -3 \\
-2 & 4 & 6 \\
-3 & 6 & 9 
\end{bmatrix}
\quad \text{and} \quad C_a = \begin{bmatrix}
1 & -1 & -1 \\
-1 & 1 & 1 \\
-1 & 1 & 1 
\end{bmatrix},
\]  

(D.6)

with environmental loadings and genotype scores initially given by:

\[
\Lambda^* = \begin{bmatrix}
-1 \\
2 \\
3 
\end{bmatrix}
\quad \text{and} \quad f^* \sim N(0, I_v),
\]

and total genetic variance of 4.67.

The variance matrix between the genotype main effects and common GE effects is:

\[
\begin{bmatrix}
\tilde{\Lambda}^*\tilde{\Lambda}^{*\top} & \tilde{\Lambda}^*\Lambda^{*\top} \\
\Lambda^*\tilde{\Lambda}^{\top} & \Lambda^*\Lambda^{\top}
\end{bmatrix} = \begin{bmatrix}
1.78 & -1.33 & 2.67 & 4.00 \\
-1.33 & 1.00 & -2.00 & -3.00 \\
2.67 & -2.00 & 4.00 & 6.00 \\
4.00 & -3.00 & 6.00 & 9.00 
\end{bmatrix},
\]

where \(\tilde{\Lambda}^* = 1.33\). The common GE effects for environments 2 and 3 are completely explained.
Fig. D.3 Latent regression plots for Example 3 showing the form of the regression before and after rotation. The genotype main effects are denoted by *crosses* and the common GE effects for different environments are distinguished by *shape*.

by the genotype main effects while the common GE effects for environment 1 are negatively correlated with the main effects, i.e. the variation is due to non-crossover and crossover GEI.

After applying a principal component rotation, the loadings and scores become:

$$\Lambda = \begin{bmatrix} -0.27 \\ 0.53 \\ 0.80 \end{bmatrix} \quad \text{and} \quad f \sim N(0, 14I_v), \quad (D.7)$$

where $\Lambda = \Lambda^*/\sqrt{14}$ and $f = \sqrt{14}f^*$. The first factor captures non-crossover and crossover
GEI because the principal component rotation cannot disentangle these components. After applying the new rotation, the loadings and scores become:

\[
\Lambda = \begin{bmatrix}
0.00 & -0.79 \\
0.60 & 0.23 \\
0.80 & 0.57
\end{bmatrix}
\quad \text{and} \quad
f \sim N\left(\begin{bmatrix}
0 \\
0
\end{bmatrix}, \begin{bmatrix}
8.16 & 3.60 \\
3.60 & 1.59
\end{bmatrix} \otimes I_v\right). \tag{D.8}
\]

The first factor exclusively captures non-crossover GEI and the second factor exclusively captures crossover GEI. Environment 1 has no variation for the first factor because all variation is attributed to crossover GEI. The covariance between the scores for different factors is also attributed to crossover GEI because all covariance attributed to non-crossover GEI has already been transferred to the first factor. The total genetic variance of 4.67 is therefore partitioned into a non-crossover variance of 2.72 and a crossover variance of 1.95.

4. Some non-crossover and little crossover GEI (Figure D.4)

The between-environment genetic variance and genetic correlation matrices are given by:

\[
G_a = \begin{bmatrix}
2.65 & -1.15 & -1.15 \\
-1.15 & 5.00 & 8.00 \\
-1.15 & 8.00 & 13.00
\end{bmatrix}
\quad \text{and} \quad
C_a = \begin{bmatrix}
1.00 & -0.32 & -0.20 \\
-0.32 & 1.00 & 0.99 \\
-0.20 & 0.99 & 1.00
\end{bmatrix}, \tag{D.9}
\]

with environmental loadings and genotype scores initially given by:

\[
\Lambda^* = \begin{bmatrix}
-1.15 & 1.15 \\
2.00 & 1.00 \\
3.00 & 2.00
\end{bmatrix}
\quad \text{and} \quad
f^* \sim N(0, I_{2v}),
\]

and total genetic variance of 6.88.

The variance matrix between the genotype main effects and common GE effects is:

\[
\begin{bmatrix}
\tilde{\Lambda}^* \tilde{\Lambda}^{*\top} & \tilde{\Lambda}^* \Lambda^{s\top} \\
\Lambda^* \tilde{\Lambda}^{*\top} & \Lambda^* \Lambda^{s\top}
\end{bmatrix} = \begin{bmatrix}
3.56 & 0.12 & 3.95 & 6.62 \\
0.12 & 2.65 & -1.15 & -1.15 \\
3.95 & -1.15 & 5.00 & 8.00 \\
6.62 & -1.15 & 8.00 & 13.00
\end{bmatrix},
\]

where \(\tilde{\Lambda}^* = [1.25 \ 1.38]\). The common GE effects for each environment are partially explained by the genotype main effects, i.e. the variation is due to non-crossover \textit{and} crossover GEI.
Fig. D.4 Latent regression plots for Example 4 showing the form of the regression before and after rotation. The genotype main effects are denoted by red crosses and the common GE effects for different environments are distinguished by shape.
After applying a principal component rotation, the loadings and scores become:

\[
\Lambda = \begin{bmatrix}
-0.10 & -0.98 \\
0.52 & 0.07 \\
0.85 & -0.16 \\
\end{bmatrix}
\quad \text{and} \quad f \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 18.11 & 0.00 \\ 0.00 & 2.54 \end{bmatrix} \otimes I_v \right).
\quad (D.10)
\]

The first factor captures both non-crossover and crossover GEI because the principal component rotation cannot disentangle these components any further.

After applying the new rotation, the loadings and scores become:

\[
\Lambda = \begin{bmatrix}
0.01 & -0.82 \\
0.51 & 0.39 \\
0.86 & 0.42 \\
\end{bmatrix}
\quad \text{and} \quad f \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 16.68 & 0.00 \\ 0.00 & 3.96 \end{bmatrix} \otimes I_v \right).
\quad (D.11)
\]

The first factor exclusively captures non-crossover GEI and the second factor exclusively captures crossover GEI. Environment 1 has little variation in the first factor because most of the variation is attributed to crossover GEI. The total genetic variance of 6.88 is partitioned into a non-crossover variance of 5.56 and a crossover of 1.32.
Appendix E

Matrix results

The following results are given in Mardia et al. (1979).

Result 1: Woodbury matrix identity:
\[
(A + BC)^{-1} = A^{-1} - A^{-1}B(I_q + CA^{-1}B)^{-1}CA^{-1},
\]  
(E.1)

where \(A\) is a \(n \times n\) non-singular matrix, \(B\) is a \(n \times q\) matrix and \(C\) is a \(q \times n\) matrix. It therefore follows that:
\[
H^{-1} = H_{-i}^{-1} - \frac{\sigma^2_d i H_{-i}^{-1} Z_m Z_m^\top H_{-i}^{-1}}{1 + \sigma^2_d i Z_m^\top H_{-i}^{-1} Z_m},
\]

where \(H = H_{-i} + \sigma^2_d i Z_m Z_m^\top\).

Result 2: Searle matrix identity:
\[
C(A + BC)^{-1} = (I_q + CA^{-1}B)^{-1}CA^{-1}.
\]  
(E.2)

It therefore follows that:
\[
Z_m^\top H^{-1} = \frac{Z_m^\top H_{-i}^{-1}}{1 + \sigma^2_d i Z_m^\top H_{-i}^{-1} Z_m},
\]

Note that Results 1 and 2 also hold for the projection matrices \(P\) and \(P_{-i}\), where:
\[
P = P_{-i} - \frac{\sigma^2_d i P_{-i} Z_m Z_m^\top P_{-i}}{1 + \sigma^2_d i Z_m^\top P_{-i} Z_m} \quad \text{and} \quad Z_m^\top P = \frac{Z_m^\top P_{-i}}{1 + \sigma^2_d i Z_m^\top P_{-i} Z_m}. 
\]
Using Results 1 and 2, it can be shown that:

\[ P^* = P_{-i} - P_{-i} Z_{m_i} (Z_{m_i}^\top P_{-i} Z_{m_i})^{-1} Z_{m_i}^\top P_{-i} \]
\[ = P - P Z_{m_i} (Z_{m_i}^\top P Z_{m_i})^{-1} Z_{m_i}^\top P, \]

where \( Z_{m_i}^\top P^* = 0 \).

**Result 3: Log-determinant of a product:**

\[ \log |ABC| = \log |A| + \log |BC|, \]  \( \text{(E.3)} \)

where \( BC \) is a \( n \times n \) non-singular matrix product.

**Result 4: Log-determinant of a sum:**

\[ \log |A + BC| = \log |A| + \log |I_q + CA^{-1}B|, \]  \( \text{(E.4)} \)

It therefore follows that:

\[ \log |H| = \log |H_{-i}| + \log (1 + \sigma_a^2 d_i Z_{m_i} H_{-i}^{-1} Z_{m_i}). \]

**Result 5: Log-determinant of a constant multiplied by a matrix:**

\[ \log |dA| = \log |A| + n \log (d), \]  \( \text{(E.5)} \)

where \( d \) is a constant.

**Result 6: Log-determinant of a block matrix:**

\[ \log |A| = \log |A_{11}| + \log |A_{22} - A_{12} A_{11}^{-1} A_{21}|. \]  \( \text{(E.6)} \)

Using Results 5 and 6, it can be shown that:

\[ \log |X^\top H^{-1} X^*| = \log |X^\top H_{-i}^{-1} X^*| - \log (1 + \sigma_a^2 d_i Z_{m_i} H_{-i}^{-1} Z_{m_i}), \]

where \( X^* = [X \ Z_{m_i}] \).