Latent Tensor Bayesian Models for Estimating Complex Interactions in Plant Variety Testing

A dissertation submitted for the degree of
Doctor of Philosophy

By:

Antonia Alessandra Lemos dos Santos

Under the supervision of:

Prof. Andrew C. Parnell

Dr. Rafael A. Moral

Department of Mathematics and Statistics
National University of Ireland Maynooth
Ollscoil na hÉireann, Má Nuad

October 2023

To Marta and Francisco.

Declaration

I hereby declare that I have produced this manuscript without the prohibited assistance of any third parties and without making use of aids other than those specified.

The thesis work was conducted from November 2019 to October 2023 under the supervision of Professor Andrew C. Parnell and Dr. Rafael Moral in the Hamilton Institute, National University of Ireland Maynooth.

Antonia Alessandra Lemos dos Santos

Maynooth, Ireland,

October 2023

Sponsor

This work was supported by European Union's Horizon 2020 research and innovation programme under grant agreement No 818144.

Collaborations

Andrew C. Parnell: As my supervisor, Professor Parnell (Maynooth University) supervised and collaborated on the work of all chapters.

Rafael Moral: As my supervisor, Dr. Moral (Maynooth University) supervised and collaborated on the work of all chapters.

Danilo Sarti: As a postdoc working on the same project, Dr. Sarti (Maynooth University) collaborated on the work of all chapters.

Publications

The chapters contained in this thesis have been submitted to peer-reviewed journals. Chapter 3 has been submitted to the *Journal of Agricultural, Biological and Environmental Statistics*. Chapter 4 has been submitted to the *Biometrical Journal*. Chapter 5 has also been submitted to the *Journal of Agricultural, Biological and Environmental Statistics*.

In parallel with the publications presented in this thesis, the author has also collaborated on two other projects. In the first project, the author is a co-author on a paper published in the journal *The Annals of Applied Statistics*. In the second project, where the author is the first author, the paper has been submitted to the journal *Ecological Modelling*.

Peer-reviewed journal article:

- Sarti, D. A., Prado, E. B., Inglis, A. N., Dos Santos, A. A., Hurley, C. B., Moral, R. A., and Parnell, A. C. (2023). Bayesian additive regression trees for genotype by environment interaction models. *The Annals of Applied Statistics*, 17(3):1936–1957.

Submitted articles (under review):

- Dos Santos, A.A.L, Parnell, A. C., Moral, R. A., Sarti, D. A. Variational Inference for Additive Main and Multiplicative Interaction Effects Models. Submitted to the *Journal of Agricultural, Biological and Environmental Statistics*.

- Dos Santos, A.A.L, Parnell, A. C., Moral, R. A., Sarti, D. A. Bayesian Additive Main Effects and Multiplicative Interaction Models Using Tensor

Regression for Multi-environmental Trials. Submitted to the *Biometrical Journal*.

- Dos Santos, A.A.L, Parnell, A. C., Moral, R. A., Sarti, D. A. Bayesian Clustering Additive Main Effects and Multiplicative Interaction Models for Multi-environmental Trials. .Submitted to the *Journal of Agricultural, Biological and Environmental Statistics*.

- Dos Santos, A.A.L., Flood, C., Moral, R. A., C., Parnell, A. C., Griffin, C.T., Fealy, R. Machine learning corrected life cycle model of the large pine weevil. Submitted to the journal *Ecological Modelling*.

Contents

Abstract

This thesis addresses the critical challenge of predicting crop yield. As the demand for food surges due to population growth, accurate and efficient predictive models become increasingly important. Leveraging a comprehensive database from the Horizon 2020 InnoVar project, which includes variables across phenomics, genomics, soil, and weather conditions, we aim to extend the existing Additive Main Effects and Multiplicative Interaction (AMMI) modelling framework.

Our work is divided into three key contributions. First, we propose a computationally efficient Bayesian AMMI model using variational inference, addressing the high computational costs often associated with traditional Markov chain Monte Carlo methods. Second, we introduce the Bayesian Additive Main effects and Multiplicative Interaction Tensor (BAMMIT) model, which extends the AMMI model to accommodate multiple categorical variables. Third, we present the Clustered Bayesian Additive Main Effects and Multiplicative Interaction Tensor (CBAMMIT) model, incorporating Gaussian Mixture Models to allow for the inclusion of categorical representations of numerical variables.

Our findings show that these extensions not only improve predictive accuracy but also offer probabilistic assessments of predictions. They have real-world applicability, as demonstrated using data from Ireland, and can potentially guide stakeholders in agriculture – from farmers to policymakers – in making informed decisions.

Acknowledgements

This thesis is the result of a collective effort that involved the support of numerous people, for whom I am extremely grateful.

First and foremost, I thank my family. To my parents, Marta and Francisco, who gave me much more than life — they taught me how to live it. They provided for me what they themselves could not have and allowed me to achieve a dream I never thought possible. To my beloved sister Sandra, my grandmother Antonia, and my cousin Patrícia, I thank you for all the love.

I am grateful to Professor Andrew Parnell, my supervisor, for his guidance, encouragement, support and extreme patience. I thank my co-supervisor, Professor Rafael Moral, for guiding me on this journey and for all the friendly conversations. Working with them has inspired me and taught me how to conduct research. For this, I will be eternally grateful.

I extend my gratitude to the Horizon 2020 InnoVar project, of which I was part of and which made this work possible. A special thanks to Danilo Sarti, the post-doc of our group, for helping me, particularly in understanding concepts that were previously new to me.

I also express my gratitude for a group of people I met in Ireland, whom I hope will remain in my life forever. Alan, for all the affection, patience with my English, and for telling me that everything would be alright. Estevão, for being my friend from day one, becoming a true brother. To HoChan, Mateus and André for all the friendship. I also thank all my other friends at the Hamilton Institute, with whom I shared special moments.

Finally, I thank Kate and Rosemary, for their kindness and support. To the Hamilton Institute and everyone at Maynooth University, who directly or indirectly contributed to the realisation of this project, thank you so much!

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Motivation

Staple foods are basic dietary items that are consumed regularly and in such quantities as to constitute the dominant part of the diet (UNICEF et al., 2023). Their production plays a crucial role in global food security and economy. In particular, wheat, rice and maize collectively supply nearly half of the world's calories and contributing to approximately 40% of protein intake (Erenstein et al., 2022). According to the Food and Agriculture Organisation (FAO), wheat alone contributes approximately $50 billion annually to the world economy and provides around 20% of all calories consumed by humans (UNICEF et al., 2020). To meet the needs of the growing world population, projected to reach 10 billion by 2050, the FAO estimates that the world will need to produce about 50% more food by then (UNICEF et al., 2020). This surge in demand places additional pressure on the agricultural industry to improve crop yields, thereby ensuring food security, sustainability and economic stability (Godfray et al., 2010; Tilman et al., 2011).

However, the production of these crops faces numerous challenges ranging from climate change to limitations in technological advancements. (Lobell and Gourdji, 2012; Rosenzweig et al., 2014). Given this context, predictive modelling in crop

yield, particularly for the staple cereals, becomes vital. Efficient and accurate models can guide strategies for sustainable agriculture, providing a buffer against the volatility caused by environmental, economic, and demographic factors (Antle and Capalbo, 2010). Accurate forecasts can be an instrument for many purposes. It can guide stakeholders (ranging from farmers to policymakers) in crucial decisions related to storage, distribution, and market pricing (Lobell et al., 2015). Moreover, precise predictions contribute to sustainable resource management, such as the optimised use of water, fertilisers, and land (Zhang et al., 2002; Khosla, 2010).

Predicting crop yield is a complex task for several reasons. First, the challenge of securing high-quality data frequently poses a hurdle (Lobell and Asseng, 2017). This can be due to initial conditions for data collection, incomplete or inconsistent datasets, or issues related to data privacy, which limit the availability of critical information. Second, the inclusion of multiple factors, such as local conditions such as soil and weather can change rapidly over short distances and time, making it difficult to make accurate predictions (Mkhabela et al., 2011; Asseng et al., 2013; Rosenzweig et al., 2014). Additionally, external economic factors such as market demand and price fluctuations further add to the complexity (Antle and Capalbo, 2010). Finally, the intricacy of the models themselves adds another layer of complexity. These models require the integration of various factors including genetics, environment, and management, which is a challenging task, therefore making it difficult to create generalisable models for crop yield prediction.

In this context, the Horizon 2020 project InnoVar[1] focuses on advancing crop variety testing across European farmlands, specifically targeting the needs of an expanding global population amid climate change. It aims to improve the efficiency and accuracy of variety testing in crops, using wheat as the primary test subject. By integrating genomics, phenomics, and machine learning, the project seeks to optimise and modernise the Decision of Uniformity, Stability (DUS), and Value for Cultivation and Use (VCU) testing processes. Involving 21 partners in 10 countries (Czechia, Denmark, Hungary, Ireland, Italy, Lebanon, Netherlands, Portugal, Spain, and the United Kingdom), InnoVar runs trials in 15 locations across Europe and creates tools for variety recommendations. The ultimate goal

[1] https://www.h2020innovar.eu

is to pave the way for improved agricultural practices for other vital crops like perennial ryegrass and maize.

The project has generated a large database, comprising more than 200 variables with approximately 100,000 observations and includes two species of wheat: bread wheat (*Triticum aestivum L.*), primarily used for making bread, and durum wheat (*Triticum durum*), commonly used for pasta, semolina, and similar products. This database is divided into four main modules: phenomics, genomics, soil, and weather. Each of these categories contains numerous associated attributes, ranging from average air temperature and fertiliser application to soil nutrients, genetic markers, and even wheat yield and grain weight. The genomics module is particularly noteworthy, containing around 40,000 Single Nucleotide Polymorphisms (SNPs) for each species of wheat. A standout feature of the project is the inclusion of nearly 300 variables obtained through drone technology, enabling high-throughput phenotyping (HTP) analyses (Gill et al., 2022). The integration of these diverse attributes is specifically aimed at constructing optimal prediction models for DUS and VCU characteristics and optimisation of such experiments.

The InnoVar project also has a historical database for Ireland and Italy. These two datasets contain a series of variables related to location and genotype. For Italy, the yield of durum wheat, measured in tonnes per hectare (t/ha), was recorded for 32 genotypes across seven locations in a randomised complete block design with three replicates over four years (from 2015 to 2018), totalling 1,269 observations. For Ireland, data was collected over ten years (from 2010 to 2019) concerning the production of bread wheat, also measured in tonnes per hectare. The trials were conducted in the counties of Dublin, Cork, Carlow, Kildare, Kilkenny, Louth, Meath, Laois, Tipperary, and Waterford, using a randomised complete block design with four replicates. The dataset includes 85 genotypes and 17 environments, resulting in a total of 6,368 measurements. The historical data for Ireland is the data used throughout this thesis.

Building on the important role of staple foods like wheat in global food security and the burgeoning demand posed by population growth, this thesis aims to advance the predictive modelling in wheat yield. Drawing upon the comprehensive

database generated by the Horizon 2020 InnoVar project, which offers a rich spectrum of variables including phenomics, genomics, soil, and weather conditions, the thesis seeks to extend the existing model Additive Main Effects and Multiplicative Interaction (AMMI) framework, keeping the structure of the model and addressing the main limitations. Given the multi-layered complexity involved in crop yield prediction, from the intricate interplay of genetic, environmental, and management factors to rapidly changing local conditions, we use the Bayesian approach to not only improve the accuracy but also to provide probabilistic assessments of the model's predictions. Specifically, our research addresses the following questions:

(a) The Bayesian AMMI model encounters a significant challenge when handling large volumes of data, given that Markov chain Monte Carlo (MCMC) methods (Robert et al., 1999) have high computational costs. Can we adopt an alternative approach for obtaining the posteriors distributions for the model parameters that reduces computational cost, while still capturing main and interaction effects, and retaining predictive power?

(b) The traditional AMMI model is constrained to studying only the effects of genotype and environment, as well as their interaction, overlooking the influence of other variables that could be associated with prediction. Can we develop a model that incorporates multiple variables, both numerical and categorical, and captures their interactions at different levels?

(c) Often, the results and predictions from these models are implemented and frequently used by farmers and non-technical individuals. Is it possible to present the findings in a manner that is not only statistically robust, but also comprehensible and actionable for individuals from other disciplines?

(d) When models are required to accommodate both categorical and continuous predictors, how do we adapt the current generation of models so that we retain the simple explanatory understanding of interactions that is commonly used in AMMI models?

Throughout this manuscript, we answer these questions through independent chapters as described next.

1.2 Outline and Contributions

In Chapter 2, we introduce the preliminary concepts underlying this research. This chapter starts off with a review of the major statistical models commonly employed in crop prediction, with a particular focus on the additive main-effects and multiplicative interaction (AMMI) model, which serves as the foundation for the methodologies developed in this thesis. We also present an overview of variational inference, tensor regression, and clustering techniques, all of which have been applied in subsequent chapters of this thesis.

In Chapter 3, we propose an alternative to reduce the computational time required for the Bayesian AMMI model by adopting variational inference. We introduce the Bayesian formulation of the model, specifying the priors for each parameter and how the model's inherent constraints are imposed in this approach. Subsequently, we derive the variational distributions for each parameter. Through a simulation study and application to a real-world case study using historical data from the InnoVar project, we compare the results obtained via the variational approach with those obtained via Markov chain Monte Carlo (MCMC). We demonstrate the model's performance in terms of accuracy, prediction and computational speed.

In Chapter 4, we propose an extension of the traditional Bayesian AMMI model, using a tensor structure that allows for multiple interacting categorical variables, in order to understand the phenotypic effect beyond the usual interaction of genotype and environment. In addition to incorporating these new effects, our proposal, named the Bayesian Additive Main effects and Multiplicative Interaction Tensor (BAMMIT) model, allows for the identification of important variables that make meaningful contributions to interaction terms. Specifically, we present the mathematical formulation of this generalised approach, along with the new notation adopted. We discuss the priors used in the hierarchical model and the spike-and-slab structure, the former to ensure model identifiability and the latter to capture both lower and higher-order interactions between variables. Through simulations, we demonstrate that the model remains identifiable and effectively captures interactions of different orders. We also show that the inclusion of new variables in the model enhances its predictive power. We conclude the chapter with an applica-

tion to historical real-world data from Ireland, now with additional variables. We suggest a visualisation approach to present the results more simply and assist in the decision-making process.

In Chapter 5, we introduce the Clustered Bayesian Additive Main Effects and Multiplicative Interaction Tensor (CBAMMIT) model, extending the BAMMIT framework through the integration of Gaussian Mixture Models (GMM). The idea is to use GMM to incorporate categorical representations of numerical variables into the existing BAMMIT model, as its current framework does not allow for the inclusion of such variables. The proposed methodology distinguishes itself from current literature by performing the clustering process simultaneously with estimation, aiming to improve the predictive performance of the model. Initially, we provide context through a brief review of GMM, followed by a conceptualisation of the AMMI and BAMMIT models, as these form the basis for the new CBAMMIT model. We then introduce the likelihoods for this new model, defining its priors and discussing their contextual implications within the chapter. We illustrate the rationale behind the adopted clustering process through a motivational example using toy data, where we use a simplified version of the model to show the advantages of simultaneous clustering and estimation. With the results obtained in the simulation study, we demonstrate the predictive abilities of the CBAMMIT model. Additionally, we indicate scenarios where the model excels, as well as those where it has its shortcomings. Finally, the real-world applicability of the model is tested using the same dataset from Ireland, including two numerical variables to be clustered: average annual temperature and average annual rainfall for the years 2010 to 2019.

In Chapter 6, we present a discussion of the main findings of this research, the broader impact and the outline for future works.

CHAPTER 2

Preliminaries

In this chapter, we provide preliminaries on topics that will be used in the subsequent chapters of this thesis. In Section 2.1, we discuss the concepts of phenotype, genotype and environment, as well as the role of the interaction between genotype and environment on the phenotype performance. We also introduce the idea of multi-environment trials. Following that, we review in Section 2.2 some commonly used models in the literature to capture the effects of genotype and environment variables, as well as the interaction between them, on the phenotypic response. In Section 3.2.1, we focus on the AMMI model, discussing in Section 2.3.1 the biplot graph, which is the most common way to visualise the AMMI model estimates. Having presented the applied context of the thesis, we proceed to introduce the methods applied to the AMMI model. Specifically, in Section 2.4, we introduce the concept of variational inference, which is employed in Chapter 3. In Section 2.5, we present the idea of tensor regression, introducing some concepts that will be applied later in Chapter 4. We conclude the preliminaries with Section 5.1, which contains an overview of Gaussian mixture distributions, a concept used in Chapter 5 of this work.

2.1 Phenotype, Genotype and Environment

As discussed by Brown and Caligari (2011), the primary goal of a plant breeding program is to create superior crop varieties that can be cultivated both efficiently and economically. This objective is rooted in the intricate relationship between phenotype, genotype and environment. As Kang (1988) points out, the performance of a crop is strongly influenced by the environmental conditions in which it is cultivated, shaping the way genetic characteristics show up as the visible traits.

The definitions of the concepts of phenotype, genotype, and environment are broad. In a plant breeding context, genotype is the genetic makeup of the plant, the set of genes inherited by the plant that serves as the genetic blueprint for its potential traits (Allard, 1999). Meanwhile, Yan and Rajcan (2002) defines environment as the sum total of circumstances surrounding or affecting the plant. These conditions include soil quality, climate, and management practices like irrigation and fertilisation. Finally, phenotype refers to the observable traits or characteristics of the plant, such as height, yield, or resistance to disease. A specific example would be a wheat genotype with a genetic predisposition for high yield may not achieve its potential if grown in nutrient-poor soil, demonstrating that both genotype and environment contribute to the final phenotype.

As a result, the observed phenotype can be a function of genotype, environment and the interaction between them:

$$p = g + e + (ge), \tag{2.1}$$

where p is the phenotypic response, g is the genotype factor, e is the environmental factor and (ge) is the interaction between genotype and environment. The equation presented in 2.1 is a conceptual model rather than a strict mathematical formula. While these components are symbolised in a linear additive fashion, the relationship in biological terms involves many interactions that are not directly quantifiable by simple arithmetic.

In Figure 2.1 we show two simple types of interactions. In the left panel, the additive interaction the effects of genotype and environment are independent of each other. It implies, for example, that a particular genotype that performs well in one

set of environmental conditions would also perform well in another. On the other hand, in a crossover interaction, as shown in the right panel, the relative performance of different genotypes changes depending on the environmental conditions. That is, a genotype that performs well under one set of environmental conditions may perform poorly under another. For example, the phenotypic response of g_1 performs better than of g_2 in an arid climate, while in a rainy climate, the opposite occurs. A detailed graphical representation of genotype-by-environment interactions is given by Yan and Kang (2002).

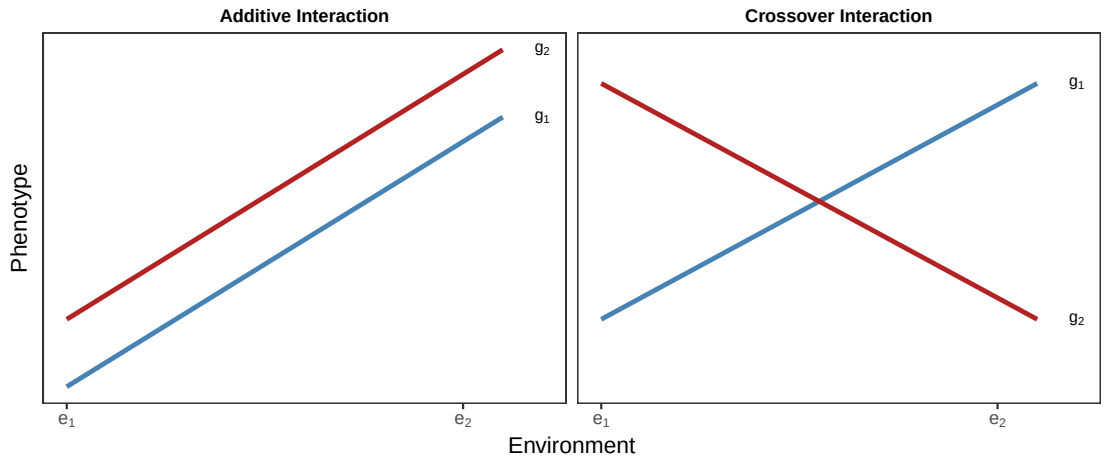


Figure 2.1: Phenotypic response on two types of interaction between two genotypes (g_1 and g_2) in two environmental conditions (e_1 and e_2).

The empirical evaluation of the genotype-by-environment interaction can be performed in a multi-environment trials (METs). METs are experiments in which a range of genotypes are evaluated across multiple environments to assess their performance and stability under varying conditions (DeLacy et al., 1996). The aim of METs is to identify genotypes that consistently perform well across diverse environmental conditions, thereby supporting in the selection of superior plant varieties for widespread cultivation.

2.2 Genotype-by-environment Interaction Models

Statistical models to characterise the genotype-by-environment ($G \times E$) interaction have been studied since the early 20^{th} century (Cooper et al., 1996; Kang et al., 2004). One of the earliest methods employed was Analysis of Variance (ANOVA), which separates the total variability in the data into components due to genotype, environment, and their interaction (Yates and Cochran, 1938). While ANOVA is straightforward and widely used, it primarily serves as an exploratory tool and lacks the capacity to model complex relationships between the variables (Piepho et al., 2008).

A linear regression approach was first used by Finlay and Wilkinson (1963) to describe $G \times E$ interaction. This method is simple and interpretable but assumes a linear relationship not only between genotype and environment but also implicitly assumes a specific form of the response to these factors, which may not adequately capture the complexity of $G \times E$ interactions. Moreover, the regression on the mean model is limited because it characterises environmental quality in a single dimension, making it difficult to capture the complexity of $G \times E$, leaving a substantial amount of interaction unexplained (Malosetti et al., 2013).

Factorial regression extends linear regression models by incorporating terms that model the combined effects of genotypes and environmental variables, thereby providing a more nuanced understanding of $G \times E$ interactions (Comstock and Moll, 1963). Factorial regression models offer a means to understanding $G \times E$ interactions by integrating both genotypic and environmental information within a single analytical framework. They provide insight into how different genotypes might respond or be sensitive to changes in specific environmental conditions or factors (Vargas et al., 1998). These models can be used to offset the modelling of noise from having many variables by being reduced to a single variable for a reduced-rank factorial regression model (Vargas et al., 1998). This method allows for more complex relationships but requires a large data set for accurate estimation.

The Genotype plus Genotype-by-Environment interaction model is another pop-

ular tool for analysing $G \times E$ (Yan et al., 2000). It focuses on the genotypic main effects and the $G \times E$ interaction effects, excluding the main effect of the environment. An advantage of this approach is that a biplot (Gabriel, 1971) (described in more detail in Section 2.3.1), derived from the model, provides a visual representation of the genotypes' main effects and their interactions with environments. However, it is important to note that the main effects of environments are not directly represented in this visualization. By focusing only on the genotype and $G \times E$ interaction, the model excludes the main effects of environments. This can be a limitation when the primary interest is in understanding specific environmental effects.

Multiplicative models that combine univariate (ANOVA) and multivariate techniques have been proposed for decomposing the genotype, environment and $G \times E$ components (De Resende and Thompson, 2004). The most common of these methods is the additive main effect and multiplicative interaction (AMMI) model (Gauch Jr et al., 1992). The AMMI model is discussed in greater detail in the next section.

It is worth noting that in many of these models, the term environment is often simplified to signify either the location or a combination of location and year/season. There are several practical reasons behind this simplification. One of the primary reasons is the ease of modelling and interpretation. Reducing the complexity of environmental factors to location or location/season enables a more straightforward analysis and interpretation of results, a point emphasised in various studies (Piepho et al., 2008; Gauch Jr, 2006). Another key reason is the availability of data. Gathering detailed data on multiple environmental factors such as soil quality, micro-climate conditions, and pest pressures could be both resource-intensive and logistically challenging (Yan et al., 2007). Using location as a proxy offers a convenient way to generalise these numerous variables. While this approach might not capture all the nuances of the genotype-environment interaction, it often serves as a reasonable approximation for first-order environmental effects and provides a balance between model complexity and interpretability (Malosetti et al., 2013).

2.3 AMMI model

The AMMI model is composed of both additive and multiplicative terms, with the latter measuring the interaction between the two additive effects. Specifically, the model can be written as:

$$y_{ij} = \mu + g_i + e_j + \sum_{q=1}^{Q} \lambda_q \gamma_{iq} \delta_{jq} + \epsilon_{ij}, \quad i = 1, \dots I, \text{ and } j = 1, \dots, J, \tag{2.2}$$

where y_{ij} is the phenotypic response and μ is the grand mean across all genotypes and environments. The terms g_i and e_j capture the additive effects of the i^{th} genotype and j^{th} environment, respectively. The noise component ϵ_{ij} is assumed to be normally distributed with a mean of zero and a variance of σ^2.

The interaction between genotypes and environments is represented by the multiplicative term of the equation, also known as the bilinear component. This term captures the $G \times E$ interaction by decomposing it into Q components. The parameter λ_q, usually arranged in descending order, measures the strength of the q^{th} component, while the parameters γ_{iq} and δ_{jq} represent the effect of the i^{th} genotype and j^{th} environment in the q^{th} component of the interaction, respectively.

Given the shared goal of decomposing data into underlying structures, the AMMI model is sometimes mistaken for factor analysis (FA). However, while FA focuses on explaining the covariance structure of the observed variables using latent factors, uncovering latent constructs, the AMMI model focuses on decomposing a data matrix into the interaction of two sets of factors, modelling interactions and decomposing structured data matrices. One advantage of the AMMI model over FA is its ability to model and interpret the interactions between these variables directly. This makes the AMMI model especially useful in agronomic studies.

To ensure the identifiability of the bilinear term, orthonormalisation constraints are assumed:

1. $\sum_i \gamma_{iq}\gamma_{iq'} = \sum_j \delta_{jq}\delta_{jq'} = 0$, for $q \neq q'$.
2. $\sum_i (\gamma_{iq})^2 = \sum_j (\delta_{jq})^2 = 1$, $\forall\ i, j$.

These constraints are necessary to obtain unique estimation of the parameters (Josse et al., 2014). However, they do not universally guarantee identifiability across all scenarios (Guhaniyogi et al., 2017; Miller and Carter, 2020). Additional considerations might be necessary, such as the number of interaction terms included (Q) and the model's overall parameterisation. Specifically, choosing an appropriate number of principal components (Q) is essential because too many components might fit the noise rather than the underlying interaction structure, potentially leading to overfitting and loss of model interpretability.

The frequentist approach for estimating the AMMI model's parameters involves two major steps (Gauch Jr, 2013). The first step is to fit a two-way ANOVA model to estimate the main effects g_i and e_j as well as the grand mean μ. Then, use principal component analysis (PCA) on the residuals from this initial model to construct an interaction matrix (Gilbert, 1963; Gollob, 1968; Gabriel, 1978). In this approach, it is also assumed that $\sum_i g_i = \sum_j e_j = 0$. It ensures that the main effects are centred, separating them from the interactions. For practical examples of how these constraints and frequentist estimations work, see Yan et al. (2007).

In the Bayesian framework, various methodologies have been proposed for estimating the parameters of the AMMI model. Viele and Srinivasan (2000) introduced the use of Markov chain Monte Carlo (MCMC) techniques, with a focus on adhering to the model's identifiability constraints. Liu (2001) improved upon this by developing a more computationally efficient and stable Gibbs sampling method. Further enhancements to Gibbs sampling were made by Crossa et al. (2011) and Perez-Elizalde et al. (2012), who not only stabilised the algorithm but also incorporated statistical inference into biplot visualisations (Gabriel, 1971), adding credibility regions for interaction effects.

In particular, Josse et al. (2014) suggests an alternative approach to handle with the model constraints, by ignoring them when defining priors for all parameters. Different from previous works, this approach does not ensure orthonormality constraints for interaction terms at the prior level. The authors propose working with functions of parameters that are identifiable. After obtaining samples from these posterior distributions using MCMC, a post-processing step is applied, where

each sample matrix is centred by row and column, and then subjected to Singular Value Decomposition (SVD). This author's claim is that this procedure gives new parameter sets that do meet the model constraints, making the results easier to interpret.

Regarding the number of multiplicative components Q in the AMMI model, there are different methods to determine it. Typically, $Q \leq \min(I-1, J-1)$. By setting Q to this maximum value, the model could theoretically capture all the variance in the interaction. However, in practice, Q often takes on a value between 1 and 3 to facilitate easier interpretation and visualisation through biplots. However, more sophisticated proposals have been applied. For example, Cornelius (1993) used parametric significance tests, while other studies have leveraged cross-validation techniques (dos S. Dias and Krzanowski, 2003; Gabriel, 2002; Hadasch et al., 2017). Resampling methods have also been applied to determine the optimal Q (Malik et al., 2018, 2019). On the Bayesian paradigm, approaches for setting Q include using prior choices and Bayes factors (Perez-Elizalde et al., 2012; da Silva et al., 2015). Additionally, a non-parametric Bayesian approach, as presented by Sarti et al. (2021), eliminates the need to specify Q altogether.

There are several reasons to for the AMMI model's popularity. It has strong predictive performance (Gauch Jr, 2006; Gauch Jr et al., 2008), accuracy (Gauch and Moran, 2019) and stability evaluation system (Gauch Jr, 1988; Yue et al., 2022). The AMMI model's flexibility is another advantage. It can be applied to a wide variety of data types and fields, not just in plant breeding but wherever the interaction between two main factors is a point of interest. Also, this model often serves as a baseline model against which more complex models can be compared. Finally, the model enhances interpretability, allowing, for example, for the generation of biplots.

Having explained its advantages and considering the various possible extensions that can be made, we use the AMMI model in this work as a baseline for new implementations and generalisations.

2.3.1 Visualisation of the AMMI model

First introduced by Gabriel (1971), the biplot is a graphical representation that displays both the observations and the variables of a data set in a reduced dimensional space. The biplot shows how each observation can be expressed as a linear combination of the original variables and how each original variable contributes to the principal components.

The biplot has seen increasing popularity in recent years, particularly among agricultural researchers due to its ability to effectively visualise $G \times E$ interactions (Yan et al., 2000). However, there have been many studies examining both the advantages and disadvantages of biplots (for example, see Yan and Tinker (2006), Yan et al. (2007), or Yang et al. (2009)). Some of the main advantages of biplots are that they can provide a concise visual summary of the relationships in a data set and that they can help in identifying patterns, clusters, and outliers. A limitation of biplots includes interpretation issues when there are many variables (Crossa et al., 2004). In data sets with a large number of observations and/or variables, the biplot can become overcrowded, making it difficult to interpret and identify patterns. Additionally, when derived from techniques like PCA (which is inherently a linear dimension reduction technique), biplots reflect the underlying assumption of linear relationships among variables. This can provide misleading interpretations if the true relationships are nonlinear.

In relation to AMMI model, both the γ_{iq} and δ_{jq} are plotted in the same biplot, making it a useful tool for interpreting genotype by environment interactions. Figure 2.2 shows a biplot example, illustrating the wheat yield from the 2015 historical Irish data set. To run the AMMI model and to create the plot, we use the R package `metan` (Olivoto and Lúcio, 2020). In the biplot, each point's coordinates are determined by the first two principal components. In this case, the plot shows two principal components, PC1 and PC2, which explain 31.8% and 28% of the variability in the data, respectively.

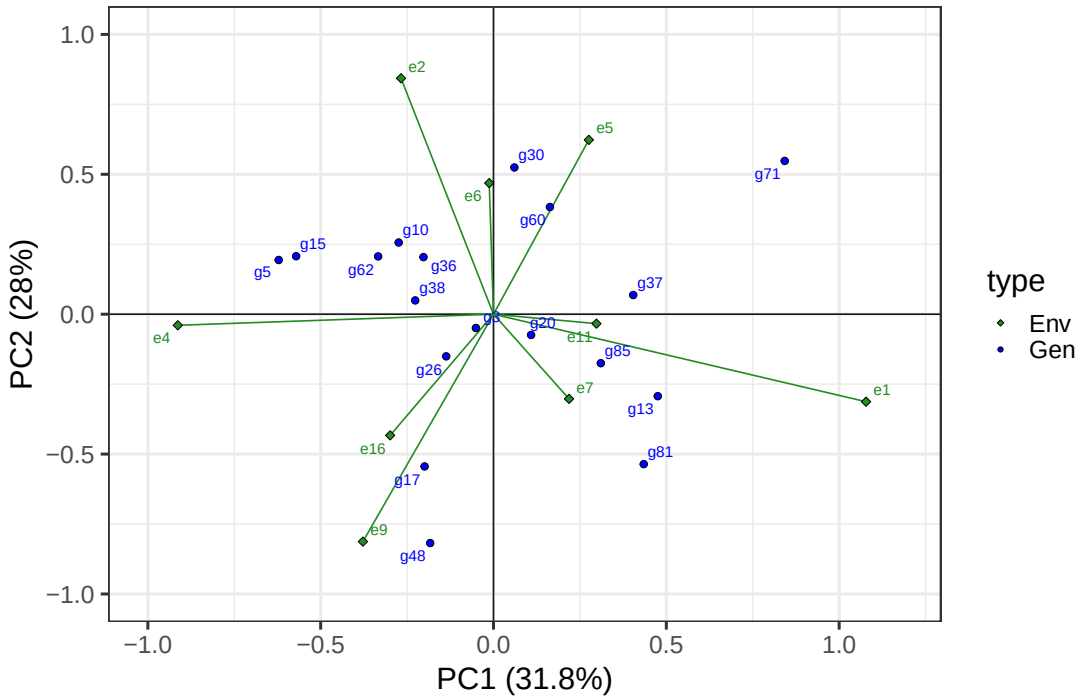


Figure 2.2: Example of a biplot, using the historical Irish data from 2015. The legends represent the environments (Env) and the genotypes (Gen).

The proximity of each pair of points in the biplot represents similarity. For example, points that are closer together represent observations with similar mean-centred values and vice-versa. The arrows in the biplot represent the original variables, with their length reflecting the extent of the variable's contribution to the variance captured by the displayed principal components. The direction and angle of an arrow in relation to the axes of the principal components indicate the strength of the variable's correlation: a tighter angle reveals a stronger correlation. Additionally, the angle between two arrows in the biplot approximates the correlation between the corresponding variables; a smaller angle suggests a positive correlation, while a larger angle, especially one exceeding 90 degrees, implies a negative correlation.

In Figure 2.2, the plot is divided into quadrants by PC1 and PC2. Points that are close together represent genotypes or environments with similar performance, while those farther apart differ more. For example, genotype $g5$ and environment $e1$ are on opposite ends of PC1, suggesting a significant difference in their impact on yield. Conversely, genotype $g85$ and environment $e7$ are positioned closely on the plot, suggesting they have similar scores on the principal components. Environment $e4$, located in the bottom-left quadrant, is close to zero on the PC2-axis and has a score near -1 on the PC1-axis. This positioning suggests that the variability or differences in yield for $e4$ are not strongly explained by PC2, but are more influenced by PC1. The length of the arrow for $e4$ indicates that this environment has a significant impact on the overall variance captured by the principal components, especially in terms of the dimension associated with PC1. One drawback of this type of plot is interpretability. The concepts involved in understanding the plot require some mathematical knowledge which can hinder its usefulness to a wider audience.

An alternative to visualising the AMMI model through biplots was suggested by Sarti et al. (2023) using heatmaps. In Chapter 4 of this work, we extend this visualisation approach by incorporating uncertainty into the heatmap legends, as proposed by Inglis et al. (2022).

2.4 Variational Inference

Let $\boldsymbol{x} = (x_1, \ldots, x_N)$ be a set of N observations and $\boldsymbol{z} = \{z_1, \ldots, z_M\}$ a set of M latent variables that we need to infer. The Bayes Theorem give us

$$p(\boldsymbol{z}|\boldsymbol{x}) \quad = \quad p(\boldsymbol{z})\frac{p(\boldsymbol{x}|\boldsymbol{z})}{\int p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}},$$

where the numerator $p(\boldsymbol{z}, \boldsymbol{x}) = p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z})$ represents the joint density. The denominator is known as the marginal likelihood or the *evidence*, defined as

$$p(\boldsymbol{x}) \quad = \quad \int p(\boldsymbol{z}, \boldsymbol{x})d\boldsymbol{z}.$$

In general, it is difficult to calculate $p(\boldsymbol{x})$. Traditionally, Markov chain Monte Carlo (MCMC) methods have been used to generate samples from the posterior

distribution $p(\boldsymbol{z}|\boldsymbol{x})$, which, over many iterations, approximate the true posterior distribution. This sampling approach provides estimates of the posterior characteristics but can be computationally intensive and slow to converge, particularly in complex models. To address this challenge, we can reformulate the integration problem as an optimisation problem.

Variational Inference (VI) or variational Bayes is a method for approximating probability densities (Blei et al., 2017). Although VI has its roots in statistical physics, its application has been popularised in machine learning and probabilistic modelling (Jordan et al., 1999; Wainwright et al., 2008; Blei et al., 2017; Bishop and Nasrabadi, 2006). Particularly, VI has been employed for topic modelling using Latent Dirichlet Allocation, Gaussian Process Regression (Titsias, 2009), and deep generative models like Variational Autoencoders (Kingma and Welling, 2013). In natural language processing, it has been applied to sentiment analysis, topic categorisation, and machine translation (Blei et al., 2003; Hoffman et al., 2013). In an applied context, Carbonetto and Stephens (2012) used VI to model genetic variation and gene expression. In neuroscience, (Archer et al., 2013) applied VI methods to decode neural signals.

The main idea behind VI is to use optimisation methods rather than sampling to approximate the true posterior distribution. First, we choose a family of approximate densities $\mathcal{Q}$ parameterised by $\boldsymbol{\theta}$, denoted as $q(\boldsymbol{z}|\boldsymbol{\theta})$. The goal is to find the set of parameters $\boldsymbol{\theta}$ that minimises the divergence between our approximation q and the true posterior distribution. This idea is illustrated in Figure 2.3.

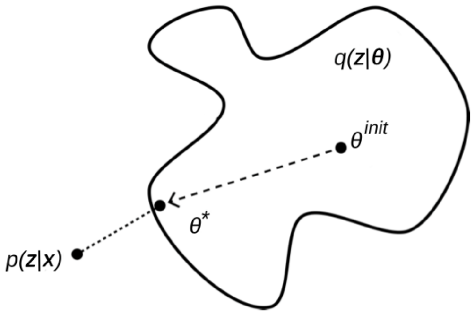


Figure 2.3: Illustration of the idea behind variational inference. This figure was produced by the author but inspired by Kejzlar and Hu (2023).

In Figure 2.3, the true posterior distribution $p(\boldsymbol{z}|\boldsymbol{x})$ represents the distribution we aim to approximate through variational inference. The contour outlined represents the family of approximate densities $q(\boldsymbol{z}|\boldsymbol{\theta})$, within which we seek the optimal parameter set $\boldsymbol{\theta}^*$. The initial parameter set is indicated by $\boldsymbol{\theta}^{\text{init}}$, showing the starting point for the optimisation. The dotted line traces the path of the optimisation process towards $\boldsymbol{\theta}^*$, which is the parameter set that minimises the divergence between $q(\boldsymbol{z}|\boldsymbol{\theta})$ and the true posterior $p(\boldsymbol{z}|\boldsymbol{x})$, effectively approximating it as closely as possible within the chosen family $\mathscr{Q}$.

To measure the closeness between the distributions $q(\boldsymbol{z}|\boldsymbol{\theta})$ and $p(\boldsymbol{z}|\boldsymbol{x})$, we use the Kullback-Leibler (KL) divergence. This quantifies the amount of information lost when $q(\boldsymbol{z}|\boldsymbol{\theta})$ is used to approximate $p(\boldsymbol{z}|\boldsymbol{x})$. Specifically, the KL divergence is a measure of how one probability distribution diverges from a second expected probability distribution:

$$q^*(\boldsymbol{z}|\boldsymbol{\theta}^*) \quad = \quad \arg\min_{q(\boldsymbol{z}|\boldsymbol{\theta})\in\mathscr{Q}} \mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{\theta})||p(\boldsymbol{z}|\boldsymbol{x})).$$

The aim is to find q^* which serves as the optimal variational distribution. As it is not possible to minimise the KL divergence directly, we optimise another quantity, the evidence lower bound (ELBO) (Blei et al., 2017):

$$\mathrm{ELBO}(q) \quad = \quad E_q[\log p(\boldsymbol{x}, \boldsymbol{z})] - E_q[\log q(\boldsymbol{z})].$$

By maximising the ELBO, we indirectly minimise the KL divergence between the variational distribution and the true posterior. This is because the KL divergence can be rewritten as:

$$\mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{\theta})||p(\boldsymbol{z}|\boldsymbol{x})) \quad = \quad \log p(\boldsymbol{x}) - \mathrm{ELBO}(q).$$

Since $\log p(\boldsymbol{x})$ is constant with respect to the variational parameters $\boldsymbol{\theta}$, maximising the ELBO is equivalent to minimising the KL divergence, thereby improving the approximation $q(\boldsymbol{z}|\boldsymbol{\theta})$ towards $p(\boldsymbol{z}|\boldsymbol{x})$.

A common choice for the variational family $\mathscr{Q}$ is the mean-field variational family, where we assume that the latent variables $\boldsymbol{z}$ are mutually independent when

conditioned on the variational parameters $\boldsymbol{\theta}$. Mathematically, this is expressed by factorising $q(\boldsymbol{z}|\boldsymbol{\theta})$ as

$$q(\boldsymbol{z}|\boldsymbol{\theta}) \quad = \quad \prod_{m=1}^{M} q_m(z_m|\theta_m).$$

To optimise the variational parameters $\boldsymbol{\theta}$ under the mean-field assumption, one popular algorithm is the coordinate ascent variational inference (CAVI) (Bishop and Nasrabadi, 2006). In CAVI, each variational parameter θ_m is updated iteratively while keeping the other variational parameters fixed. Specifically, the optimal $q_m(\boldsymbol{\theta}_m)$ are proportional to

$$q_m(\boldsymbol{\theta}_m) \quad \propto \quad \exp\{E_{-\boldsymbol{\theta}_m} \log p(\boldsymbol{\theta}_m|\boldsymbol{y}, \boldsymbol{\theta}_{-\boldsymbol{\theta}_m})\}.$$

where $E_{-\boldsymbol{\theta}_m}$ denotes the expectation with respect to all elements of the vector $\boldsymbol{\theta}$ except $\boldsymbol{\theta}_m$.

Although the mean field approach is computationally more tractable, it is worth noting that the assumption that latent variables are mutually independent might not be suitable for many problems, particularly those where correlations between latent variables are important (Blei et al., 2017).

In Chapter 3, the variational inference approach is applied as an alternative to MCMC methods for obtaining the posterior distributions of the Bayesian AMMI model, with the aim of reducing computational time cost.

2.5 Tensor Regression

Let $I_1, \ldots, I_N$ be positive integers. A tensor of order N (or N-way tensor) is a mathematical object represented as a multi-dimensional array of size $I_1 \times I_2 \times \ldots \times I_N$ (Kolda and Bader, 2009). Tensors generalise scalars, vectors and matrices. A vector is a tensor of order 1, while a matrix is a tensor of order 2 (see Figure 2.4).

Tensors have found applications in a wide range of disciplines (Cichocki et al., 2015). In chemistry, tensors are often used in the context of quantum chemistry, molecular dynamics, and spectroscopy (Khoromskaia and Khoromskij, 2018). Tensors are used in multiple areas of physics. For example, in general relativity, tensors

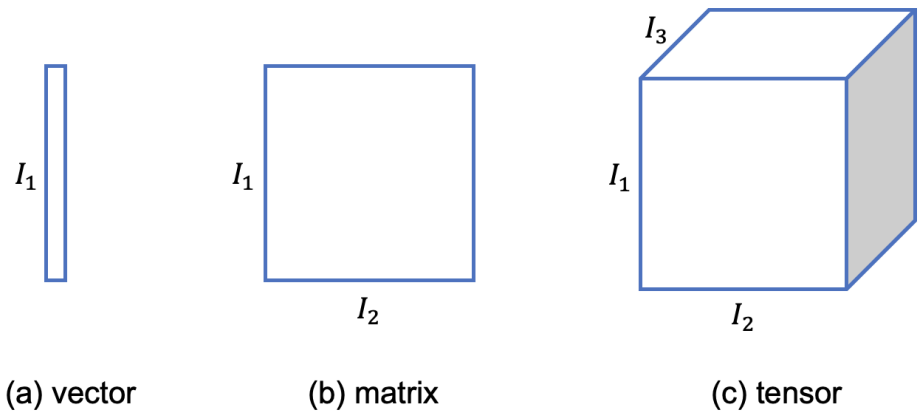


Figure 2.4: Graphical representation of tensors of order 1, 2 and 3. This figure was produced by the author, but inspired by a similar one in Liu et al. (2022).

are used in formulating the gravitational field equations (Einstein, 1916). Similarly, in quantum mechanics, the concept of tensors is used to depict the statistical state of a quantum system (Von Neumann, 2018). In the field of computer science, tensors are regularly used in areas such as natural language processing (Bouchard et al., 2015), or when using deep learning for image recognition to enable machines to interpret and make decisions based on visual data (Panagakis et al., 2021). In statistics, tensors are applied to extend traditional linear regression models for multi-dimensional data (Zhou et al., 2013).

Mathematically, let be N be the number of observations, $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N$ a collection of predictor tensors and $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_N$ a collection of response tensors, where each $\boldsymbol{X}_i$ would be of dimension $I_1, \ldots, I_P$ and and each $\boldsymbol{Y}_i$ would be of dimension $I_1, \ldots, I_R$. For each observation $i = 1, \ldots, N$, the tensor regression model can be formulated as:

$$\mathbf{Y}_i = f(\mathbf{X}_i, \mathbf{W}) + \mathbf{E}_i,$$

where, $f(\mathbf{X}_i, \mathbf{W})$ is the regression function mapping the predictor tensor $\mathbf{X}_i$ and coefficient tensor $\mathbf{W}$ to the response tensor $\mathbf{Y}_i$. $\mathbf{E}_i$ is the error tensor corresponding to the i^{th} observation.

Depending on the assumptions made about the functions $f(\mathbf{X}_i, \mathbf{W})$ various feature selection strategies, like tensor decomposition, can be used (Liu et al., 2022). There are several ways to decompose a tensor, but two of the most commonly used methods are the Canonical Polyadic (CP) decomposition and the Tucker decomposition. These decompositions aim to reduce the dimensionality of the tensor, thereby making it computationally more manageable (Kolda and Bader, 2009).

It is interesting to note a conceptual resemblance between the multiplicative interaction terms in the AMMI model (Equation 2.2) and tensor regression models. The bilinear terms $\lambda_q \gamma_{iq} \delta_{jq}$, which capture the interactions between genotypes and environments, can be viewed as a latent tensor in a product form. This comparison is particularly relevant in the context of tensor regression, where such a structure is explicitly used to model multi-way interactions between factors.

In this work, we employ tensor regression in Chapter 4 to generalise the AMMI model. We follow the Bayesian framework proposed by Guhaniyogi et al. (2017), where the a Bayesian regression model using vector or tensor covariates is used to predict a univariate response via a category of multiway shrinkage priors.

2.6 Bayesian Mixture of Gaussian Distributions

Clustering analysis aims to segment observations into distinct groups, or clusters. Items in the same cluster are more closely related to each other than to items in other clusters. The clustering is accomplished by employing various metrics for measuring similarity or distance between data points.

The clustering technique studied in this work is based on probabilistic models and uses the principle of mixture of distributions. This stands in contrast to conventional clustering methods such as K-means, which rely on heuristic criteria for similarity or distance. The objective is to identify the statistical model that is most likely to have generated the data. Each cluster is essentially a component of this mixture model, where each observation is assigned to a particular cluster based on the likelihood that it was produced by that specific component model.

For a comprehensive understanding of model-based clustering, see the review by

McNicholas (2016) and a more recent work by Gormley et al. (2023). Additionally, Bouveyron and Brunet-Saumard (2014) provides an extensive review focused on the application of Model-Based Clustering to high-dimensional data. McLachlan et al. (2019) also present a review of theory and methodological developments in finite mixture models.

Frequently, a Gaussian mixture model (GMM) it is used, which assumes that each component follows a Gaussian distribution. Formally, let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ be N independent and identically distributed random variables. Assume that $\boldsymbol{x}_i$ is a p-dimensional random vector following a multivariate Normal distribution. The density of $\boldsymbol{x}_i$ is modelled as a mixture of K components:

$$f(\boldsymbol{x}_i|\boldsymbol{\theta}_k) \quad = \quad \sum_{k=1}^{K} \pi_k \mathrm{N}_p(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{2.3}$$

where $\boldsymbol{\theta}_k$ is the vector of parameters of the mixture model, with $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ the mean vector and the covariance matrix associated with the k^{th} component, respectively. The parameters $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ are the cluster mixing proportion, such that $\pi_k > 0$ and $\sum_{k=1}^{K} \pi_k = 1$. The model in (2.3) can be expressed through latent allocation variables $\boldsymbol{c} = (\boldsymbol{c}_1, ..., \boldsymbol{c}_n)^\top$, where $\boldsymbol{c}_i \in \{1, \ldots K\}$ identifies the cluster in which the respective $\boldsymbol{x}_i$ belongs to. We can assume that $\boldsymbol{c}_i \sim \mathrm{Cat}(\boldsymbol{\pi})$ and $\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha})$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K) = 1$.

In the nonparametric Bayesian approach, the structure of π fundamentally determines the allocation of data points to different clusters within the model, directly influencing its ability to model the underlying data distribution. The flexibility of the Dirichlet process allows for an unbounded number of potential components, with the π's governing the richness of the mixture and its capacity to adapt to the complexity of the data. Thus, the inherent structure of the π's is pivotal to the performance and adaptability of the nonparametric Bayesian models. For more details, see Murugiah (2010).

As discussed by Gormley et al. (2023), GMM is sensitive to the number of parameters p. This means that as the number of parameters increases, challenges may arise in interpreting the results and in the estimations, due to the increasing dimension of the $\boldsymbol{\Sigma}_k$ matrix. One proposed solution to address this issue is to

consider an eigen-decomposition of the form

$$\boldsymbol{\Sigma}_k = \rho_k \boldsymbol{D}_k \boldsymbol{A}_k \boldsymbol{D}_k^\top. \tag{2.4}$$

Each component in Equation 2.4 is responsible for a specific geometric property of volume, shape and orientation. The parameter ρ_k is a proportionality constant that determines the volume occupied by the k^{th} component, while the matrix of eigenvectors $\boldsymbol{D}_k$ defines the orientation. The diagonal matrix of eigenvalues $\boldsymbol{A}_k$ establishes the shape. Through this decomposition, it is possible to obtain 14 different models using finite mixtures of multivariate Gaussian, with different geometric characteristics. This decomposition is used in the implementation of the model-based clustering R package `mclust` (Scrucca et al., 2016).

One of the questions to be answered concerns how many clusters K should be included in the model. If the value of K is too small, some clusters may not be included, whereas if K is too large, the model will overfit. Therefore, there is a trade-off between the model's complexity and the number of groups. Extensive research has been conducted on the selection of K, by both Bayesian and non-Bayesian approaches.

A conventional method is fitting models of various complexities and using criteria like Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC) for model selection (Alamichel et al., 2022). Another approach is to consider $K \to \infty$, then the model becomes a non-parametric mixture, also known as an infinite mixture model (Rasmussen, 1999). The most commonly used prior for this type of mixture is the Dirichlet process (Ferguson, 1973; Lo, 1984).

Another challenge that arises in the Bayesian framework, it is the label switching problem (Diebolt and Robert, 1994; Richardson and Green, 1997) where the labels of the clusters may change upon model refitting or resampling, although the model itself remains the same (Celeux et al., 2000; Stephens, 2000). As pointed out by Stephens (2000), this changing can result in a highly symmetric and multimodal posterior distribution of the parameters, which complicates the task of summarisation. Papastamoulis (2016) and Jasra et al. (2005) present a detailed review on solutions to this label switching problem.

In this work, the GMM-based clustering process is applied in Chapter 5. The method is considered to enable the inclusion of continuous variables in the AMMI model.

CHAPTER 3

Variational Inference for Additive Main and Multiplicative Interaction Effects Models

In plant breeding the presence of a genotype by environment ($G \times E$) interaction has a strong impact on cultivation decision making and the introduction of new crop cultivars. The combination of linear and bilinear terms has been shown to be very useful in modelling this type of data. A widely-used approach to identify $G \times E$ is the Additive Main Effects and Multiplicative Interaction Effects (AMMI) model. However, as data frequently can be high-dimensional, Markov chain Monte Carlo (MCMC) approaches can be computationally infeasible. In this article, we consider a variational inference approach for such a model. We derive variational approximations for estimating the parameters and we compare the approximations to MCMC using both simulated and real data. The new inferential framework we propose is on average two times faster whilst maintaining the same predictive performance as MCMC.

3.1 Introduction

In plant breeding, it is often of interest to identify which genotypes perform best in different environments. The presence of genotype × environment ($G \times E$) interactions is an important factor, and has a strong impact on the yield. Furthermore, it contributes to the improvement of breeding programs (McLaren and Chaudhary, 1994). Many authors deal with modelling interactions in plant genomic data. In particular, Crossa et al. (2010) review various statistical models for analysing $G \times E$ interactions. A common approach is to create models that combine linear and bilinear (interaction) terms. Often, these bilinear interaction terms are not simple multiplicative combinations of genotype-environment pairs, but rather latent parameters to be estimated. For papers on linear-bilinear models, see Gauch Jr et al. (2008), Crossa et al. (2010), Poland et al. (2012) and Gauch Jr (2013).

One of the most widely used linear-bilinear models is the Additive Main and Multiplicative Interaction (AMMI) effects model (Gauch Jr, 1988; Gauch Jr et al., 1992). AMMI incorporates both additive and multiplicative components from the two-way data structure, first educing the principal additive components, and then investigating the $G \times E$ component with principal components analysis. The results of an AMMI-style analysis are usually displayed graphically in biplots (Gauch Jr and Zobel, 1997; Yan et al., 2000; Yan and Rajcan, 2002), that help to interpret the $G \times E$ interactions. Further, these models typically perform well with regard to their predictive properties (Gauch Jr, 2006; Gauch Jr et al., 2008).

Several approaches have been made in estimating the parameters of the AMMI model, mostly from the point of view of classical inference (Gilbert, 1963; Gabriel, 1978; Gauch Jr, 1988; Van Eeuwijk, 1995). However, the desirable characteristics for evaluating uncertainty and including expert knowledge, as a priori information, led the Bayesian approach to be applied in $G \times E$ modelling (Foucteau et al., 2001; Theobald et al., 2002; Cotes et al., 2006), and consequently, in the AMMI model. In the latter case, it is necessary to deal with the constraints associated with the model, and the problems caused by the orthonormal bases used in the decomposition of the bilinear term. Due to these restrictions, complicated prior distribution decisions for AMMI inference may be necessary (Viele and Srinivasan,

2000; Crossa et al., 2011). An alternative approach proposed by Josse et al. (2014) ignores these restrictions on the prior distributions and applies later-level processing of the posterior.

When the data sets are particularly large or when models become complex, traditional methods for Bayesian inference can fail to sample from the posterior distribution. An alternative way to deal with these issues is to use methods such as Variational Inference (VI); also called Variational Bayes (Beal, 2003; Ormerod and Wand, 2010; Blei et al., 2017). This approach has been widely used in recent years because, whilst the variational approximations do not always converge to the exact posterior distribution, they are computationally much faster. In some agricultural models, for example, VI has been demonstrated to perform well (Montesinos-López et al., 2017; Gillberg et al., 2019).

To date we have not found a paper that applies VI to AMMI models, and this is the focus of our paper. We follow Josse et al. (2014)'s approach for constructing the AMMI model, but fit the model using VI in order to enjoy the characteristics of the Bayesian approach and reduce the computational cost in the approximation of the posteriors of the model. In a set of simulation studies, we find that VI performs well in terms of speed and, in comparison with the results obtained via Markov chain Monte Carlo (MCMC), our results were similar in terms of accuracy. Unsurprisingly we find the speed boost to be greater when the data set grows larger.

Our paper is structured as follows. We begin in Section 3.2.1 by reviewing the AMMI model, which is formulated in the Bayesian style. Section 3.2.2 provides a review of VI. In Section 3.2.3 we present the mathematics of the variational updates for the AMMI model. In Section 3.3 we evaluate our methods and make comparisons with MCMC using simulated and real data. We present graphs comparing the results of the two methods, as well as the prediction results. In Section 4.5 we provide a discussion of our findings and point to future research. An R implementation of the AMMI models and variational inference of the model used to run all experiments and generate the results of the paper is available on GitHub here (`https://github.com/Alessandra23/vammi`).

3.2 Theoretical Background

3.2.1 AMMI Model

In this section, we define the AMMI model and present the restrictions that guarantee identifiability of the model parameters. The AMMI model for an outcome variable (e.g. log yield per hectare), $y_{ij} \in \mathbb{R}$, is formulated as

$$y_{ij} = \mu + g_i + e_j + \sum_{q=1}^{Q} \lambda_q \gamma_{iq} \delta_{jq} + \epsilon_{ij}, \tag{3.1}$$

where μ is the grand mean, g_i, $i \in \{1, \ldots, I\}$ and e_j, $j \in \{1, \ldots, J\}$, represent the effect of the i-th genotype and j-th environment, respectively, and ϵ_{ij} is noise distributed as $\mathcal{N}(0, \sigma^2)$. The term in the summation is called the bilinear component and captures the interaction effects, where λ_q is the singular value of the q-th bilinear component, and γ_{iq} and δ_{jq} are the left and right singular vectors of the interaction, respectively. The upper term of the summation Q represents the number of bilinear terms to include and is fixed before running the model. For visualization and interpretability issues it is common to set Q less than three. All the terms on the right hand side of the equation (with the exception of Q) are parameters to be estimated. The model can be trivially extended to account for block or replicate effects, but we do not explore such extensions in this paper.

In matrix form, denoting $\boldsymbol{Y} = (y_{ij}) \in \mathbb{R}^{I \times J}$, Equation (3.1) is equivalent to

$$\boldsymbol{Y} = \mu \mathbf{1}_I \mathbf{1}_J' + \boldsymbol{g} \mathbf{1}_J' + \mathbf{1}_I \boldsymbol{e}' + \boldsymbol{\Gamma} \boldsymbol{\Lambda} \boldsymbol{\Delta}' + \boldsymbol{E}. \tag{3.2}$$

where $\boldsymbol{\Gamma} \in \mathbb{R}^{I \times Q}$, $\boldsymbol{\Delta} \in \mathbb{R}^{J \times Q}$, $\boldsymbol{\Lambda} \in \mathbb{R}^{Q \times Q}$, $\mathbf{1}_m$ is a column vector of ones of size m. $\boldsymbol{\gamma}_q$ and $\boldsymbol{\delta}_q$ arise from the q-th column of matrices $\boldsymbol{\Gamma}$ and $\boldsymbol{\Delta}$ respectively. $\boldsymbol{\Lambda}$ is a diagonal matrix consisting of the terms $\lambda_1, \ldots, \lambda_Q$. The other bold terms indicate vector stacking of the individual main effects.

The nature of the over-parameterisation means that it is necessary to establish some conditions of identifiability and interpretability for the parameters. It is thus common to assume the constraints below:

(i) $\mathbf{1}_I' \boldsymbol{g} = \mathbf{1}_J' \boldsymbol{e} = 0$,

(ii) $\mathbf{1}_I'\boldsymbol{\gamma}_q = \mathbf{1}_J'\boldsymbol{\delta}_q = 0$ for all $q \in \{1, \ldots, Q\}$,

(iii) $\boldsymbol{\Gamma}'\boldsymbol{\Gamma} = \boldsymbol{\Delta}'\boldsymbol{\Delta} = \mathbf{I}_Q$,

(iv) the diagonal terms of $\boldsymbol{\Lambda}$ are ordered such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_Q \geq 0$,

(v) the first entry of each column of $\boldsymbol{\Gamma}$ is positive.

The constraints (i) and (ii) are interpretability constraints, whereas (iii), (iv) and (v) are necessary constraints to ensure identifiability of the model.

From a frequentist perspective, a multi-stage least squares method can be used to estimate the parameters of the model, first estimating the linear terms $\mu + g_i + e_j$ and then using Principal Components Analysis (PCA), equivalent to maximum likelihood estimation, to estimate the residuals of the nonlinear term and so create $\sum_{q=1}^{Q} \lambda_q \gamma_{iq} \delta_{jq}$ (Gilbert, 1963; Gollob, 1968; Gabriel, 1978).

From a Bayesian perspective, the restrictions can be dealt with via the prior distributions. In Cornelius and Crossa (1999) a Bayesian shrinkage estimator is proposed. Crossa et al. (2011) introduce proper prior distributions and use a Gibbs sampler for inference on the parameters to target the exact posterior distribution. Perez-Elizalde et al. (2012) propose the von Mises–Fisher distribution as a prior for the orthonormal matrices $\boldsymbol{\Gamma}$ and $\boldsymbol{\Delta}$. By contrast, Josse et al. (2014) propose that the over-parameterisation be handled subsequent to the model fitting, initially ignoring the problems at the prior level and applying an appropriate post-processing of the posterior distribution. They argue that this approach allows easily interpretable inferences to be obtained, in addition to simpler implementation in statistical software. Mendes et al. (2020) and Omer and Singh (2017) perform a comparison between the classical and Bayesian approaches of the AMMI model, showing that the Bayesian modelling is more flexible than the classical one. More recently, Sarti et al. (2023) propose the use of semi-parametric Bayesian Additive Regression Trees (BART) to capture genotypic by environment interactions.

One of the main problems when using the Bayesian methodology in linear-bilinear models is the associated computational cost, given the complex structure of the parameters. Taking this into account, we follow the proposal of Josse et al. (2014),

applying the prior distributions suggested in the matrices of the linear term, and using the variational inference to obtain the estimates of the AMMI model parameters, thus reducing the computational time of the fitting process.

3.2.2 Variational Inference

The basic idea of VI, as an alternative to Markov chain Monte Carlo, is to approximate the posterior distribution via optimisation, thereby making the estimation process computationally faster. First, we choose a family of approximate densities over a set of latent variables which we believe will provide a good approximation to the true posterior. Then we find the set of parameters that make our approximation as close as possible to the posterior distribution. Blei et al. (2017) reviews the method, presenting examples, applications, and a discussion of problems and current research on the topic.

Let $\boldsymbol{y} = y_1, ..., y_n$ a set of observations, $\boldsymbol{\theta}$ a vector of parameters, $p(y)$ the marginal distribution of observations, and $p(\boldsymbol{y}, \boldsymbol{\theta})$ the joint density of the model and the parameters. The density transform approach, one of the most common variants of VI (Ormerod and Wand, 2010), consists of approximating the posterior distribution of $p(\boldsymbol{\theta}|\boldsymbol{y}) = p(\boldsymbol{y}, \boldsymbol{\theta})/p(\boldsymbol{y})$ by a distribution $q(\boldsymbol{\theta})$ from a set of tractable distributions $\mathscr{Q}$, for which we minimise the Kullback-Leibler (KL) divergence. Since it is not possible to calculate KL directly, the optimisation is done over an equivalent quantity, called the evidence lower bound (ELBO) (Blei et al., 2017), where maximising it is equivalent to minimising the KL divergence:

$$\text{ELBO}[q(\boldsymbol{\theta})] \quad = \quad -\text{KL}[q(\boldsymbol{\theta}) \| p(\boldsymbol{\theta} \mid \mathbf{y})] + \log p(\mathbf{y}). \tag{3.3}$$

Maximising the ELBO, comprising the expected log-likelihood of the observed data and the entropy of the variational distribution, is as an effective way by indirectly minimising the KL divergence. It allows for efficient and scalable inference in models where exact Bayesian inference is unfeasible. It is computationally simpler and adapts well to large data sets and complex models, such as those found in this work.

A broad class of distributions to describe the variational family $\mathscr{Q}$ is the mean-field approximation, which assumes that the elements of $\boldsymbol{\theta}$ are mutually independent,

then $q(\boldsymbol{\theta})$ can be factored into $\prod_{m=1}^{M} q_m(\boldsymbol{\theta}_m)$, where M is the number of partitions of the vector of parameters $\boldsymbol{\theta}$. Therefore, each variable $\boldsymbol{\theta}_m$ can be governed by its own variational factor. Bringing together the ELBO and the mean-field family, the optimal $q_m(\boldsymbol{\theta}_m)$ are proportional to

$$q_m(\boldsymbol{\theta}_m) \quad \propto \quad \exp\{E_{-\boldsymbol{\theta}_m} \log p(\boldsymbol{\theta}_m | \boldsymbol{y}, \boldsymbol{\theta}_{-\boldsymbol{\theta}_m})\}, \tag{3.4}$$

where $E_{-\boldsymbol{\theta}_m}$ denotes the expectation with respect to all $\boldsymbol{\theta}$ variational distributions except $\boldsymbol{\theta}_m$. If the prior distributions are conjugate it is possible to obtain explicit expressions for each component leading to simple parameter updates, which is one of the advantages of the approach. As a disadvantage, Montesinos-López et al. (2017) comments that the restrictions imposed by the mean-field approximation can lead to underestimation of the variability of parameter estimates, and further emphasises that it is not a good option if there is strong dependency between the parameters.

3.2.3 VI applied to the AMMI Model

Exact inference for the model is intractable and we use the variational approximation for computational efficiency. Following Josse et al. (2014), we list the priors defined for the complete set of parameters without considering any hard constraints (Table 3.1). To simplify notation, let $\Theta = \{\mu, \boldsymbol{g}, \boldsymbol{e}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\sigma^2}\}$ be the set of parameters and let $\boldsymbol{\theta} = \{\mu_\mu, \sigma^2_\mu, \sigma^2_g, \sigma^2_e, \sigma^2_\lambda, a, b\}$ be all the hyper-parameters. Omitting the dependency on $\boldsymbol{\theta}$, we state the joint posterior density of the parameter vector of the variational AMMI model as

$$p(\boldsymbol{Y}, \mu, \boldsymbol{g}, \boldsymbol{e}, \boldsymbol{\Lambda}, \boldsymbol{\Gamma}, \boldsymbol{\Delta}, \sigma^2) \propto p(\boldsymbol{Y} | \mu, \boldsymbol{g}, \boldsymbol{e}, \boldsymbol{\Lambda}, \boldsymbol{\Gamma}, \boldsymbol{\Delta}, \sigma^2) p(\mu) p(\boldsymbol{g}) p(\boldsymbol{e}) p(\boldsymbol{\Lambda}) p(\boldsymbol{\Gamma}) p(\boldsymbol{\Delta}) p(\sigma^2),$$

and assume that the posterior distribution is approximated by factorising the variational approximation,

$$p(\mu, \boldsymbol{g}, \boldsymbol{e}, \boldsymbol{\Lambda}, \boldsymbol{\Gamma}, \boldsymbol{\Delta}, \sigma^2) \quad \propto \quad q(\mu) q(\boldsymbol{g}) q(\boldsymbol{e}) q(\boldsymbol{\Lambda}) q(\boldsymbol{\Gamma}) q(\boldsymbol{\Delta}) q(\sigma^2).$$

For our model, the resulting approximate posterior distribution of each factor follows the same distribution as the corresponding prior. Table 3.1 presents the

assumed prior distributions as well as the approximations of the posterior distribution, where $\mathcal{N}$, $\mathcal{N}^+$ and $\mathcal{G}$ denote the normal, truncated normal and gamma distributions, respectively. For brevity, we do not list here the variational updates of all these parameters. We provide the full mathematical derivation in Appendix 3.A.

A key point in variational inference is the dependence of the estimates on the initial values of the optimisation (Rossi et al., 2019; Airoldi et al., 2008; Lim and Teh, 2007). In our variational approach, the initial values proved to be especially important, perhaps due to the dependence on the parameters of the bilinear term, which is ignored by the coordinate ascent. In view of this, our way to get around this problem was to initialise the Θ set in the algorithm with the frequentist estimates of the parameters, then we recursively update the variational parameters $\mu_{q(\mu)}$, $\Sigma_{q(\mu)}$, $\mu_{q(g_i)}$, $\Sigma_{q(g_i)}$, $\mu_{q(e_j)}$, $\Sigma_{q(e_j)}$, $\mu_{q(\lambda_q)}$, $\Sigma_{q(\lambda_q)}$, $\mu_{q(\gamma_i)}$, $\Sigma_{q(\gamma_i)}$, $\mu_{q(\delta_j)}$, $\Sigma_{q(\delta_j)}$, $a_{q(\sigma^{-2})}$, and $b_{q(\sigma^{-2})}$, for all i, j, q, until convergence. For more discussion of this initialisation see Section 3.3.

Table 3.1: Overview of prior distributions and the respective variational distributions of each parameter of the model.

Parameter	Prior Distribution	Variational Distribution
μ	$\mathcal{N}(\mu_\mu; \sigma_\mu^2)$	$\mathcal{N}(\mu_{q(\mu)}, \Sigma_{q(\mu)})$
$\boldsymbol{g}$	$\prod_{i=1}^{I} \mathcal{N}(0; \sigma_g^2)$	$\prod_{i=1}^{I} \mathcal{N}(\mu_{q(g_i)}, \Sigma_{q(g_i)})$
$\boldsymbol{e}$	$\prod_{j=1}^{J} \mathcal{N}(0; \sigma_e^2)$	$\prod_{j=1}^{J} \mathcal{N}(\mu_{q(e_j)}, \Sigma_{q(e_j)})$
$\boldsymbol{\lambda}$	ordered sample of $\prod_{q=1}^{Q} \mathcal{N}^+(0; \sigma_\lambda^2)$	$\prod_{q=1}^{Q} \mathcal{N}^+(\mu_{q(\lambda_q)}, \Sigma_{q(\lambda_q)})$
$\boldsymbol{\gamma}$	$\mathcal{N}^+(0;1) \prod_{q=1}^{Q} \prod_{i=2}^{I} \mathcal{N}(0;1)$	$\prod_{i=1}^{I} \mathcal{N}(\mu_{q(\gamma_i)}, \Sigma_{q(\gamma_i)})$
$\boldsymbol{\delta}$	$\prod_{q=1}^{Q} \prod_{j=1}^{J} \mathcal{N}(0;1)$	$\prod_{j=1}^{J} \mathcal{N}(\mu_{q(\delta_j)}, \Sigma_{q(\delta_j)})$
σ^{-2}	$\mathcal{G}(a; b)$	$\mathcal{G}(a_{q(\sigma^{-2})}, b_{q(\sigma^{-2})})$

The variational distributions listed are a consequence of the variational approximation method used. We employed a mean-field variational Bayes approach, which assumes that the variational distribution factorises over the parameters, leading to the specific forms shown.

3.3 Results

3.3.1 Simulation Study

In this section, we describe the simulation study to investigate the performance of the proposed algorithm. The data were generated as follows:

1. The number of environments and genotypes were $I \in \{6, 10, 12, 25, 50, 100, 200\}$ and $J \in \{10, 12, 20, 30, 50, 100\}$, respectively, and we used $Q \in \{1, 2\}$.
2. We simulated the additive term as $\mu \sim \mathcal{N}(90, 10)$, $g_i \sim \mathcal{N}(0, \sigma_g^2)$, and $e_j \sim \mathcal{N}(0, \sigma_e^2)$. We fixed the standard deviations $\sigma_g^2 = 10$ and $\sigma_e^2 = 10$.
3. For the multiplicative term, we fixed $\lambda_q = \{12, 20, 25, 60\}$, and obtained the matrices $\mathbf{\Gamma}$ and $\mathbf{\Delta}$ through orthonormalization, ensuring their columns are orthogonal with unit norm. The process is described in Sarti et al. (2023).
4. We simulated the values y_{ij} from $\mathcal{N}(\mu + g_i + e_j + \sum_{q=1}^{Q} \lambda_q \gamma_{iq} \delta_{jq}, \sigma_y^2)$, where we fixed $\sigma_y^2 = 1$.

We performed several experiments to determine how the Root Mean Square Error (RMSE) is influenced by the initialisation of the $\mathbf{\Theta}$ vector in the algorithm. To achieve this, we altered the initial values of the $\mathbf{\Theta}$ vector in three different scenarios. That is, for each scenario we supplied as initial values: (a) random values from a normal distribution; (b) frequentist estimates from the AMMI model; and (c) Bayesian estimates of the model, considering 25% of the simulated data.

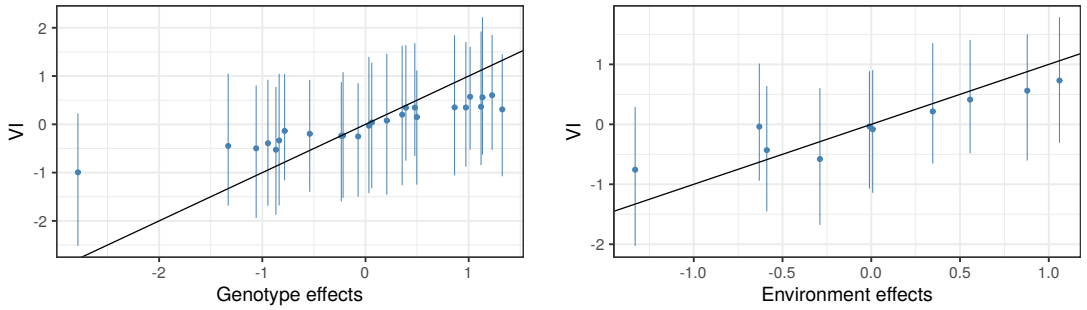
We evaluated the performance by calculating the accuracy of the predicted values, computational time and compatibility of the estimates compared to those obtained via MCMC. We used RMSE as a measure to assess the quality of the estimates in the predicted values. In the evaluation of the criteria for accuracy and comparison

of estimates, we used the results from the implementation of MCMC using the package 'R2jags' (Plummer et al., 2003; Su and Yajima, 2012). The VI algorithm was implemented in R (Team, 2021), on a MacBook Pro 1.4GHz Quad-Core Intel Core i5 with 8GB memory. The computational time was measured in minutes.

The main effects parameters and the mean behave well, in the sense that the respective estimates converge to their true values, regardless of the Q value taken. The variational distributions of these parameters are in line with those obtained via MCMC. The variational estimates of the main effects of genotypes and environments versus the true values are shown in Figure 3.1 as well as the respective subsequent MCMC, considering Q = 1, $\lambda = 60$, 25 genotypes and 10 environments.

Although the results for the main effects behave well, our main interest is to analyse the quality of the estimation of the bilinear interaction term. To achieve this, we analysed three scenarios in particular, one when λ was close to zero, another when λ was close to 20, and finally when λ was 60, with $Q = 1$ for each scenario. The first case performs the worst, as the model cannot capture the interaction and, consequently, the estimates are poor. However, as λ grows, the simulation results showed that the VI algorithm is able to estimate the bilinear parameters reasonably well, see Figure 3.2a.

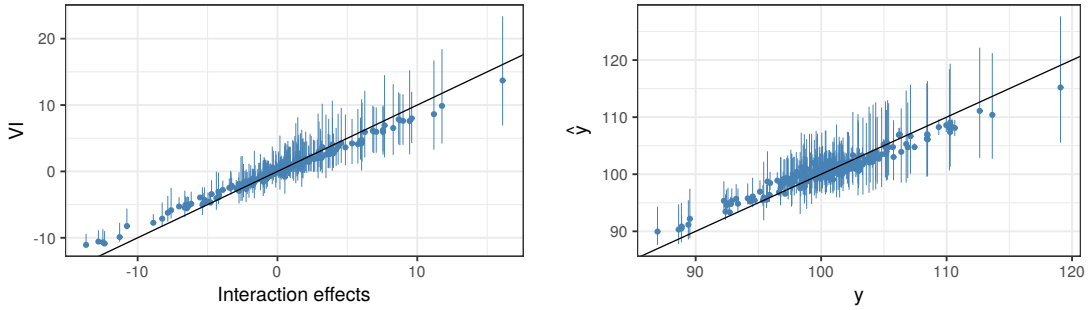
Note that in the main effects parameters, interaction term and predicted $\hat{y}$, a systematic bias is observed. This bias arises due to the approximation quality of VI, which uses a simpler distribution to approximate the posterior, and the choice of the initial values. Additionally, the convergence is to a local optima rather than the true posterior. The constraints of the AMMI model are also affecting the estimates.



(a) True (horizontal axis) versus estimated (vertical axis) values of genotype effects.

(b) True (horizontal axis) versus estimated (vertical axis) of environment effects.

Figure 3.1: Variational estimates of genotype and environment effects considering $Q = 1$, $\lambda = 60$, $I = 25$, $J = 10$, and the the respective quantile intervals, 5% and 95%.



(a) True bilinear effects (horizontal axis) versus variational estimates (vertical axis).

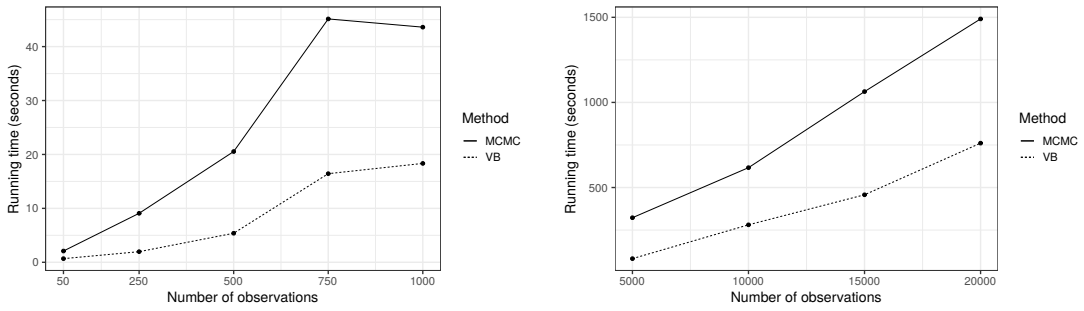
(b) True values of y (horizontal axis) versus predicted values (vertical axis).

Figure 3.2: Variational estimates of bilinear term and predicted values considering $Q = 1$, $\lambda = 60$, $I = 25$, $J = 10$, and the the respective quantile ranges, 5% and 95%.

To evaluate the computational performance of the VI algorithm, we considered different scenarios, and separated them into two different groups: smaller versus larger sample sizes, which consisted of $\{100, 250, 500, 1000\}$ and $\{5000, 10000, 15000, 20000\}$ observations respectively. When the number of observations is smaller ($I \times J \leq 1000$) the computational cost of MCMC and VI are similar (Figure 3.3a). However, when there is an increase in the number of genotypes and environments ($I \times J > 5000$), it is possible to observe a more significant difference

between the two procedures (Figure 3.3b). The value of Q also naturally changes the computational time of the algorithms, and as it grows, the time increases.

Note that there is a theoretical dependency between the number of genotypes, the number of environments, and the total number of observations. As the total number of observations is generally defined as the product of the number of genotypes and the number of environments, assuming each genotype is tested in each environment, the model's complexity can be influenced by the values of I and J. If either I or J is very small, the power to detect significant interactions decreases, and the estimates of interaction effects become less reliable. Similarly, a small N relative to the complexity of the model (e.g., when many genotypes and environments are involved but few observations per cell) might lead to unreliable estimates.



(a) Comparison of simulation times for sample sizes $\{100, 250, 500, 1000\}$.

(b) Comparison of simulation times for sample sizes $\{5000, 10000, 15000, 20000\}$.

Figure 3.3: Computational fitting time of AMMI model VI versus MCMC. For smaller experiments where the number of genotypes/environments is in the 10s then the time difference is small. For large experiments where $I, J > 100$ then the benefits of the VI model are clearer.

3.3.2 Real Data Set

We now illustrate our methods applied to a real data set from the Horizon2020 EU InnoVar project (www.h2020innovar.eu) that aims to build new cultivation tools from genomic, phenomic and environmental data. For this study, we consider data spanning ten years (2010 - 2019) concerning the production of a common species of wheat called *Triticum aestivum L.*, in Ireland, with the response being

the yield of wheat measured in tonnes per hectare. The data were supplied by the Irish Department of Agriculture, Food, and Marine. The experiments were conducted using a block design with four replicates, with the yield averaged across the replicates. For our study, we considered a single data set with all years, taking the averages of the years of genotypes and environments, resulting in a final data set containing 85 genotypes, 17 environments, and a total of 810 observations.

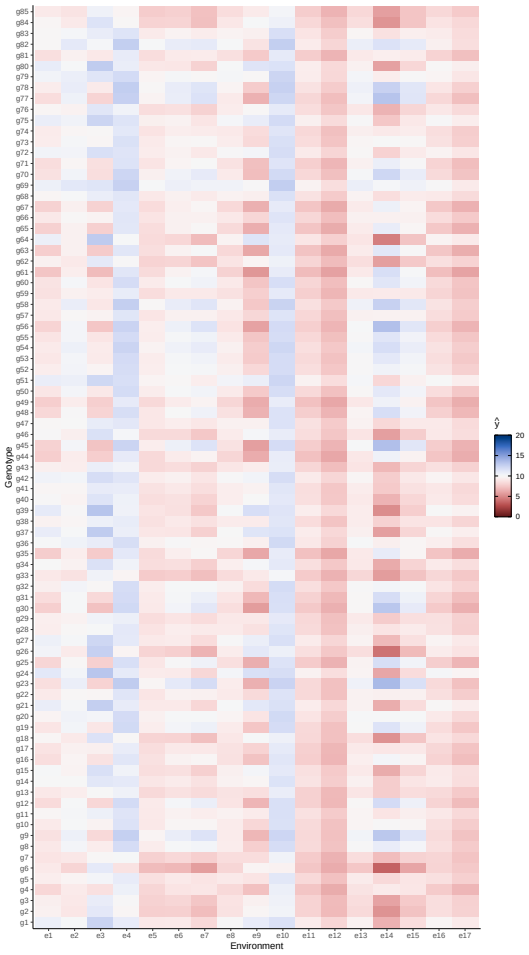
Initially, our objective is to perform a comparison of VI to MCMC, evaluating the computational cost of the two methods, contrasting the posterior distributions and the accuracy of the predictions. Then, we use the posterior variational distributions to make inferences about the genotypes and environments under study, identifying which genotypes perform better in each environment. In both algorithms, we fit the model by taking $Q = 1$ and $Q = 2$.

In terms of computational time, both methods took approximately the same time to run, with VI being seconds faster. Regarding the accuracy of the predictions, calculated in sample, although it is known that the VI approach is less accurate than MCMC, in the data set under study the VI was quite satisfactory, since it had an RMSE of 1.08 for $Q = 1$, while for the MCMC it was 0.68.

The second point to be addressed concerns the effect of genotypes in respective environments. Josse et al. (2014) discuss how the Bayesian methodology could be used to provide additional insights into the analysis of $G \times E$ interactions. For example, which genotype has the best performance across environments, and which genotype has the best performance in a given environment. In the classical methodology, this type of question is answered using a biplot (Gabriel, 1978). However, some authors have already discussed how careful the researcher should be when using this tool. Josse et al. (2014) presents a graphical way in which credibility boxes are created using the quantiles of the posterior distributions. Sarti et al. (2021) propose a heatmap plot to observe which genotype and environment are best for producing wheat. This plot, when compared with the traditional biplot, provides more complete and objective information about which genotypes and environments interact most effectively, in an intuitive and interpretable way.

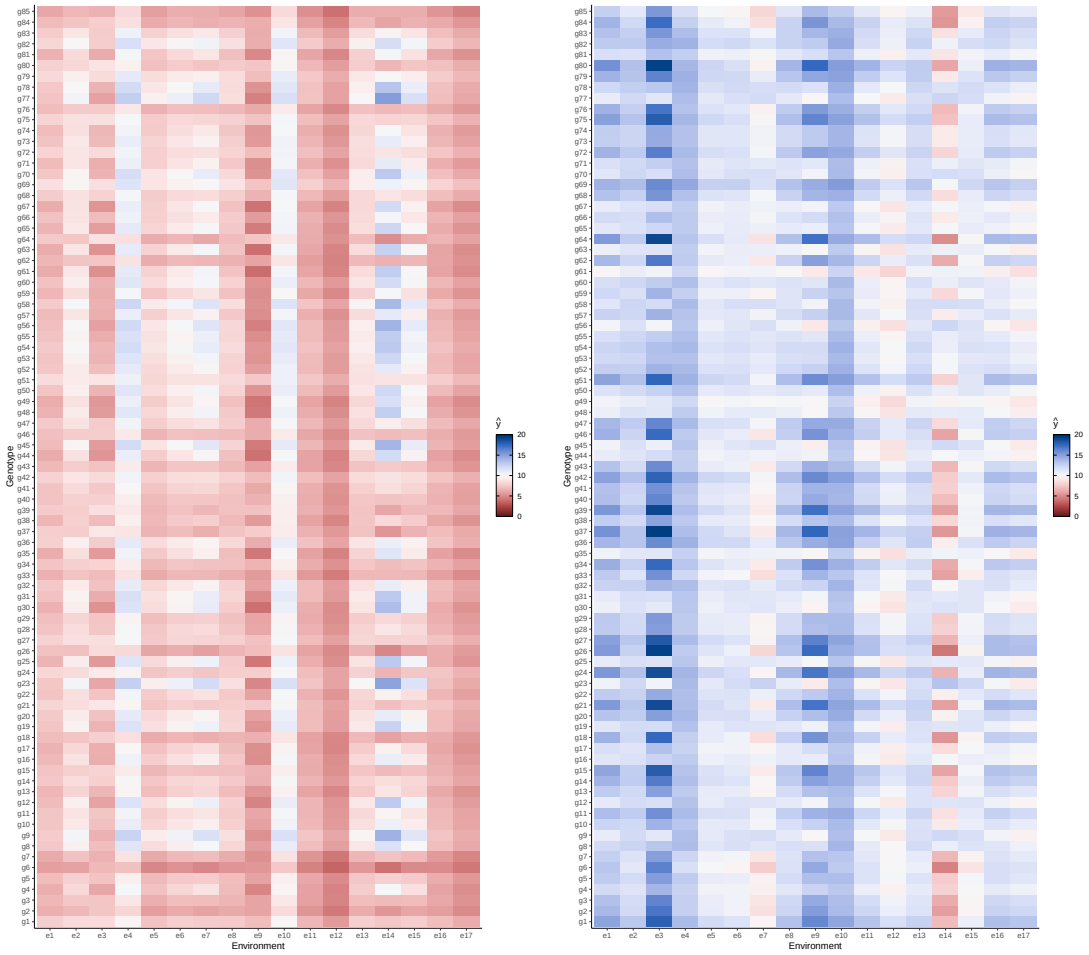
Figures 3.4 and 3.5 shows the visualisation proposed by Sarti et al. (2021), con-

sidering the variational posterior distributions setting $Q = 1$. Figure 3.4 shows that environment e_{12} produces a higher predicted yield with genotypes g_{17} and g_{21}. Additionally, we can see that environments e_5 and e_{10} have similar fits across the genotypes, as do e_4 and e_{11}. Conversely, we can also identify certain genotypes that display similar behaviour across all environments, such as g_{17}, g_{36}, or g_{81}. Genotypes g_{21} and g_{73} have highest predicted yeild in environment e_7, whereas, genotype g_{74} has the lowest predicted yield in environment e_8. In general, environment e_8 has the lowest predicted wheat yields, while environments e_{12}, e_{17}, e_1, e_5, and e_9 have the highest predicted yields. Genotypes g_{17}, g_{21}, and g_{73} performed well in practically all environments in which it was present, while genotypes g_{57}, g_{19} and g_{64} showed lower values in all environments. Figure 3.5 shows the 5% and 95% quantiles, which are uncertainties associated with the predicted yields. Although it can be useful, to examine the interaction terms and their credible intervals, the main goal of this work was to investigate how the genotype by environments affect the predicted response.



(a) 50% quantile.

Figure 3.4: Quantile 50% of the predicted wheat yields, from the posterior variational distribution ($Q = 1$).



(a) 5% quantile.

(b) 95% quantile.

Figure 3.5: Quantiles 5% and 95% of the predicted wheat yields, from the posterior variational distribution ($Q = 1$).

3.4 Discussion

In this work we propose a variational inference on the Additive Main Effects and Multiplicative Interaction Effect models, which is widely used in analysing $G \times E$ interactions. Our main contribution is the formulation of an efficient variational approximation scheme that satisfies the model constraints while simultaneously obtaining a computationally faster algorithm. To satisfy the model constraints, we adopted the priors suggested by Josse et al. (2014), where the authors address

the over-parameterisation problem at a posterior level, instead of at a prior level.

As demonstrated in our simulation study, our inferential approach performs well in various scenarios, outperforming other Bayesian methods previously described in the literature and the results presented by Josse et al. (2014). This holds true for both small and larger sample sizes. The variational scheme was able to satisfactorily estimate the main effects and the interaction effect, as well as provide good predictions. This performance was carried over into the real InnoVar data set, where the predictive performance was comparable to that of MCMC methods.

The computational time for our method proved to be superior when compared to both the algorithm when run in JAGS. In the real data scenario, VI was slightly faster than MCMC. In the real data scenario, VI was considerably faster than MCMC methods. In this case, the difference in computational time would be even more evident if the sample size were larger.

For future work, we aim to compare the variational proposal with the Bayesian AMMI model implemented in NIMBLE (de Valpine et al., 2017). In real-world scenarios, we plan to consider data sets with a larger number of genotypes and environments, and consequently, a greater number of observations. Given the method's sensitivity to the initial values of optimisation, an analysis of variational methods for multimodal surfaces may also be undertaken (Morningstar et al., 2021; Nedelkoski et al., 2020).

Appendix

In this chapter, we provide the variational updates for the model's parameters presented in Table 3.1. For simplicity, we present the derivations only for Q = 1. Additionally, we present the heatmaps plots for the interaction term of the real data presented in Section 3.3.2.

3.A Variational updates

3.A.1 Likelihood

$$
\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{\Theta}) &= \prod_{i=1}^{I}\prod_{j=1}^{J}\mathcal{N}(\mu + g_i + e_j + \lambda_1\gamma_i\delta_j; \tau^{-1}) \\
&= \prod_{i=1}^{I}\prod_{j=1}^{J}\frac{\tau^{-1}}{\sqrt{2\pi}}\exp\left\{-\frac{\tau}{2}\left(y - (\mu + g_i + e_j + \lambda_1\gamma_i\delta_j)^2\right)\right\}
\end{aligned}
$$

$$
\begin{aligned}
\log p(\boldsymbol{y}|\boldsymbol{\Theta}) &\propto \sum_{i=1}^{I}\sum_{j=1}^{J}\left\{\frac{\log\tau}{2} - \frac{\tau}{2}\left(y - (\mu + g_i + e_j + \lambda_1\gamma_i\delta_j)^2\right)\right\} \\
&\propto \sum_{i=1}^{I}\sum_{j=1}^{J}\left\{\frac{\log\tau}{2} - \frac{\tau}{2}\Big(y^2 - 2\mu y_{ij} - 2y_{ij}g_i - 2y_{ij}e_j - 2y_{ij}\lambda_1\gamma_i\delta_j + \mu^2 + \right. \\
&\quad \left. 2\mu g_i + 2\mu e_j + 2\mu\lambda_1\gamma_i\delta_j + g_i^2 + 2g_ie_j + 2g_i\lambda_1\gamma_i\delta_j + e_j^2 + 2e_j\lambda_1\gamma_i\delta_j + \lambda^2\gamma_i^2\delta_j^2\Big)\right\}.
\end{aligned}
$$

3.A.2 Variational distribution of μ.

$$
\begin{aligned}
q(\mu) &\propto \exp\{E_{-\mu}[\log p(\boldsymbol{y}|\boldsymbol{\Theta}) + \log p(\mu)]\} \\
\exp q(\mu) &\propto E_{-\mu}\left\{\sum_{i=1}^{I}\sum_{j=1}^{J} -\frac{\tau}{2}\left[-2y_{ij}\mu + \mu^2 + 2\mu g_i + 2\mu e_j + 2\mu\lambda_1\gamma_{i1}\delta_{j1}\right] - \frac{1}{2\sigma_\mu^2}(\mu^2 - 2\mu\mu_\mu)\right\} \\
&\propto \mu^2\left[-\frac{1}{2}\left(n\tilde{\tau} + \frac{1}{\sigma_\mu^2}\right)\right] + \mu\left[\tilde{\tau}\left(\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij} - J\sum_{i=1}^{I}\tilde{g}_i - I\sum_{j=1}^{J}\tilde{e_j} - \sum_{i=1}^{I}\sum_{j=1}^{J}\tilde{\lambda}_1\tilde{\gamma}_{i1}\tilde{\delta}_{j1}\right) + \frac{\mu_\mu}{\sigma_\mu^2}\right] \\
q(\mu) &\sim \mathcal{N}(\mu_{q(\mu)}, \Sigma_{q(\mu)}^{-1}); \\
\mu_{q(\mu)} &= \Sigma_{q(\mu)}^{-1}\left[\tilde{\tau}\left(\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij} - J\sum_{i=1}^{I}\tilde{g}_i - I\sum_{j=1}^{J}\tilde{e_j} - \sum_{i=1}^{I}\sum_{j=1}^{J}\tilde{\lambda}_1\tilde{\gamma}_{i1}\tilde{\delta}_{j1}\right) + \frac{\mu_\mu}{\sigma_\mu^2}\right] \\
\Sigma_{q(\mu)} &= n\tilde{\tau} + \frac{1}{\sigma_\mu^2}
\end{aligned}
$$

3.A.3 Variational distribution of g.

$$
\begin{aligned}
q(\boldsymbol{g}) &\propto \exp\{E_{-\boldsymbol{g}}[\log p(\boldsymbol{y}|\boldsymbol{\Theta}) + \log p(\boldsymbol{g})]\} \\
\exp q(\boldsymbol{g}) &\propto E_{-g_i}\left\{\sum_{i=1}^{I}\sum_{j=1}^{J} -\frac{\tau}{2}\left[-2y_{ij}g_i + 2\mu g_i + g_i^2 + 2g_i e_j + 2g_i\lambda_1\gamma_{i1}\delta_{j1}\right] - \sum_{i=1}^{I}\frac{1}{2\sigma_g^2}g_i^2\right\} \\
&\propto \sum_{i=1}^{I}\left\{g_i^2\left[-\frac{1}{2}\left(n\tilde{\tau} + \frac{1}{\sigma_g^2}\right)\right] + g_i\left[\tilde{\tau}\left(\sum_{j=1}^{J} y_{ij} - J\tilde{\mu} - \sum_{j=1}^{J}\tilde{e_j} - \sum_{j=1}^{J}\tilde{\lambda}_1\tilde{\gamma}_{i1}\tilde{\delta}_{j1}\right)\right]\right\} \\
q(\boldsymbol{g}) &\sim \prod_{i=1}^{I}\mathcal{N}(\mu_{q(g_i)}, \Sigma_{q(g_i)}^{-1}); \\
\mu_{q(g_i)} &= \Sigma_{q(g_i)}^{-1}\left[\tilde{\tau}\left(\sum_{j=1}^{J} y_{ij} - J\tilde{\mu} - \sum_{j=1}^{J}\tilde{e_j} - \sum_{j=1}^{J}\tilde{\lambda}_1\tilde{\gamma}_{i1}\tilde{\delta}_{j1}\right)\right] \\
\Sigma_{q(g_i)} &= J\tilde{\tau} + \frac{1}{\sigma_g^2}
\end{aligned}
$$

3.A.4 Variational distribution of e.

$$
\begin{aligned}
q(\boldsymbol{e}) &\propto \exp\{E_{-\boldsymbol{e}}[\log p(\boldsymbol{y}|\boldsymbol{\Theta}) + \log p(\boldsymbol{e})]\} \\
\exp q(\boldsymbol{e}) &\propto E_{-e_j}\left\{\sum_{i=1}^{I}\sum_{j=1}^{J} -\frac{\tau}{2}\left[-2y_{ij}e_j + 2\mu e_j + e_j^2 + 2g_i e_j + 2e_j\lambda_1\gamma_{i1}\delta_{j1}\right] - \sum_{i=1}^{I}\frac{1}{2\sigma_e^2}e_j^2\right\} \\
&\propto \sum_{j=1}^{J}\left\{e_j^2\left[-\frac{1}{2}\left(n\tilde{\tau} + \frac{1}{\sigma_e^2}\right)\right] + e_j\left[\tilde{\tau}\left(\sum_{i=1}^{I} y_{ij} - I\tilde{\mu} - \sum_{i=1}^{I}\tilde{g}_i - \sum_{i=1}^{I}\tilde{\lambda}_1\tilde{\gamma}_{i1}\tilde{\delta}_{j1}\right)\right]\right\} \\
q(\boldsymbol{e}) &\sim \prod_{j=1}^{J}\mathcal{N}(\mu_{q(e_j)}, \Sigma^{-1}_{q(e_j)}); \\
\mu_{q(e_j)} &= \Sigma^{-1}_{q(e_j)}\left[\tilde{\tau}\left(\sum_{i=1}^{I} y_{ij} - I\tilde{\mu} - \sum_{i=1}^{I}\tilde{g}_i - \sum_{i=1}^{I}\tilde{\lambda}_1\tilde{\gamma}_{i1}\tilde{\delta}_{j1}\right)\right] \\
\Sigma_{q(e_j)} &= I\tilde{\tau} + \frac{1}{\sigma_e^2}
\end{aligned}
$$

3.A.5 Variational distribution of λ_1.

$$
\begin{aligned}
q(\lambda_1) &\propto \exp\{E_{-\lambda_1}[\log p(\boldsymbol{y}|\boldsymbol{\Theta}) + \log p(\lambda_1)]\} \\
\exp q(\lambda_1) &\propto E_{-\lambda_1}\left\{\sum_{i=1}^{I}\sum_{j=1}^{J} -\frac{\tau}{2}\left[-2y_{ij}\lambda_1\gamma_i\delta_j + 2\mu\lambda_1\gamma_i\delta_j + 2g_i\lambda_1\gamma_i\delta_j + 2e_j\lambda_1\gamma_i\delta_j + (\lambda_1\gamma_i\delta_j)^2\right] - \right. \\
&\qquad \left. \frac{\lambda_1^2}{2\sigma_\lambda^2(1-\Phi(0))}\right\} \\
q(\lambda_1) &\sim \mathcal{TN}(\mu_{q(\lambda_1)}, \Sigma^{-1}_{q(\lambda_1)}); \\
\mu_{q(\lambda_1)} &= \Sigma^{-1}_{q(\lambda_1)}\left[\tilde{\tau}\sum_{i=1}^{I}\sum_{j=1}^{J}\gamma_i\delta_j(-2y_{ij} + 2\mu + 2g_i + 2e_j)\right] \\
\Sigma_{q(\lambda_1)} &= \tilde{\tau}\sum_{i=1}^{I}\sum_{j=1}^{J}(\widetilde{\gamma^2}\widetilde{\delta^2}) + \frac{1}{\sigma_\lambda^2(1-\Phi(0))}
\end{aligned}
$$

3.A.6 Variational distribution of $\boldsymbol{\gamma}$

$$
\begin{aligned}
q(\boldsymbol{\gamma}) &\propto \exp\{E_{-\boldsymbol{\gamma}}[\log p(\boldsymbol{y}|\boldsymbol{\Theta}) + \log p(\boldsymbol{\gamma})]\} \\
\exp q(\gamma) &\propto E_{-\gamma}\left\{\sum_{i=1}^{I}\sum_{j=1}^{J} -\frac{\tau}{2}\left[-2y_{ij}\lambda_1\gamma_i\delta_j + 2\mu\lambda_1\gamma_i\delta_j + 2g_i\lambda_1\gamma_i\delta_j + 2e_j\lambda_1\gamma_i\delta_j + (\lambda_1\gamma_i\delta_j)^2\right] - \right. \\
&\qquad \left. \frac{1}{2}\gamma_1^2 - \sum_{i=2}^{I}\frac{1}{2}\gamma_i^2\right\} \\
q(\gamma) &\sim \prod_{i=1}^{I}\mathcal{N}(\mu_{q(\gamma_i)}, \Sigma^{-1}_{q(\gamma_i)}); \\
\mu_{q(\gamma)} &= \Sigma^{-1}_{q(\gamma)}\left[\tilde{\tau}\tilde{\lambda_1}\sum_{j=1}^{J} y_{ij}\tilde{\delta_j} + \tilde{\mu}\tilde{\lambda_1}\sum_{j=1}^{J} + \tilde{g_i}\tilde{\lambda_1}\sum_{j=1}^{J}\tilde{\delta_j} + \tilde{\lambda_1}\sum_{j=1}^{J}\tilde{e_j}\tilde{\delta_j}\right] \\
\Sigma_{q(\gamma)} &= \tilde{\tau}\tilde{\lambda_1^2}\sum_{j=1}^{J}\tilde{\delta_j^2} + 1.
\end{aligned}
$$

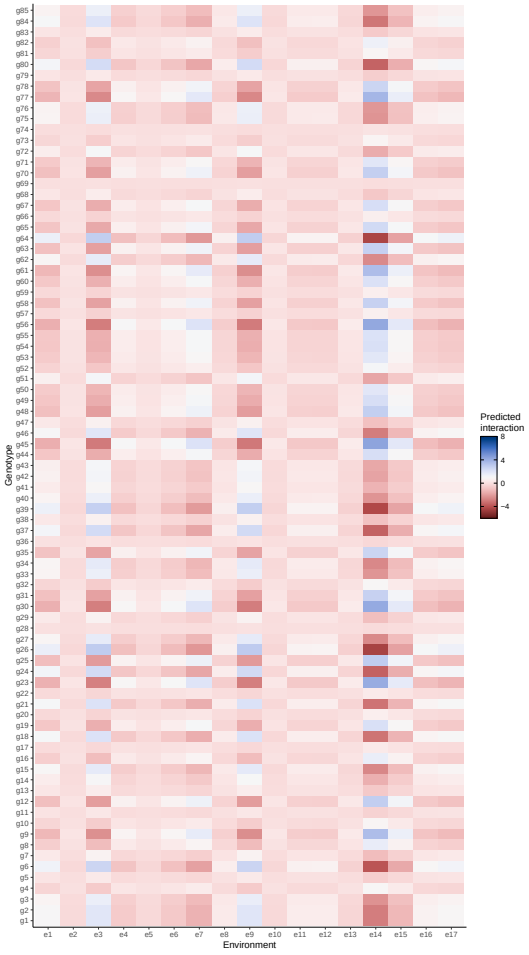
3.A.7 Variational distribution of $\boldsymbol{\delta}$

$$
\begin{aligned}
q(\boldsymbol{\delta}) &\propto \exp\{E_{-\boldsymbol{\delta}}[\log p(\boldsymbol{y}|\boldsymbol{\Theta}) + \log p(\boldsymbol{\delta})]\} \\
\exp q(\delta) &\propto E_{-\delta}\left\{\sum_{i=1}^{I}\sum_{j=1}^{J} -\frac{\tau}{2}\left[-2y_{ij}\lambda_1\gamma_i\delta_j + 2\mu\lambda_1\gamma_i\delta_j + 2g_i\lambda_1\gamma_i\delta_j + 2e_j\lambda_1\gamma_i\delta_j + (\lambda_1\gamma_i\delta_j)^2\right] - \right. \\
&\qquad \left. \frac{1}{2}\sum_{j=1}^{J}\delta_j^2\right\} \\
q(\delta) &\sim \prod_{j=1}^{J}\mathcal{N}(\mu_{q(\delta_j)}, \Sigma^{-1}_{q(\delta_j)}); \\
\mu_{q(\delta_j)} &= \Sigma^{-1}_{q(\delta)}\left[\tilde{\tau}\tilde{\lambda_1}\sum_{i=1}^{I} y_{ij}\tilde{\gamma_i} + \tilde{\mu}\tilde{\lambda_1}\sum_{i=1}^{I}\tilde{\gamma_i} + \tilde{\lambda_1}\sum_{i=1}^{I}\tilde{\gamma_i}\tilde{g_i} + \tilde{\lambda_1}\tilde{e_j}\sum_{i=1}^{I}\tilde{\gamma_i}\right] \\
\Sigma_{q(\delta_j)} &= \tilde{\tau}\tilde{\lambda_1^2}\sum_{i=1}^{I}\tilde{\gamma_i^2} + 1.
\end{aligned}
$$

3.A.8 Variational distribution of τ.

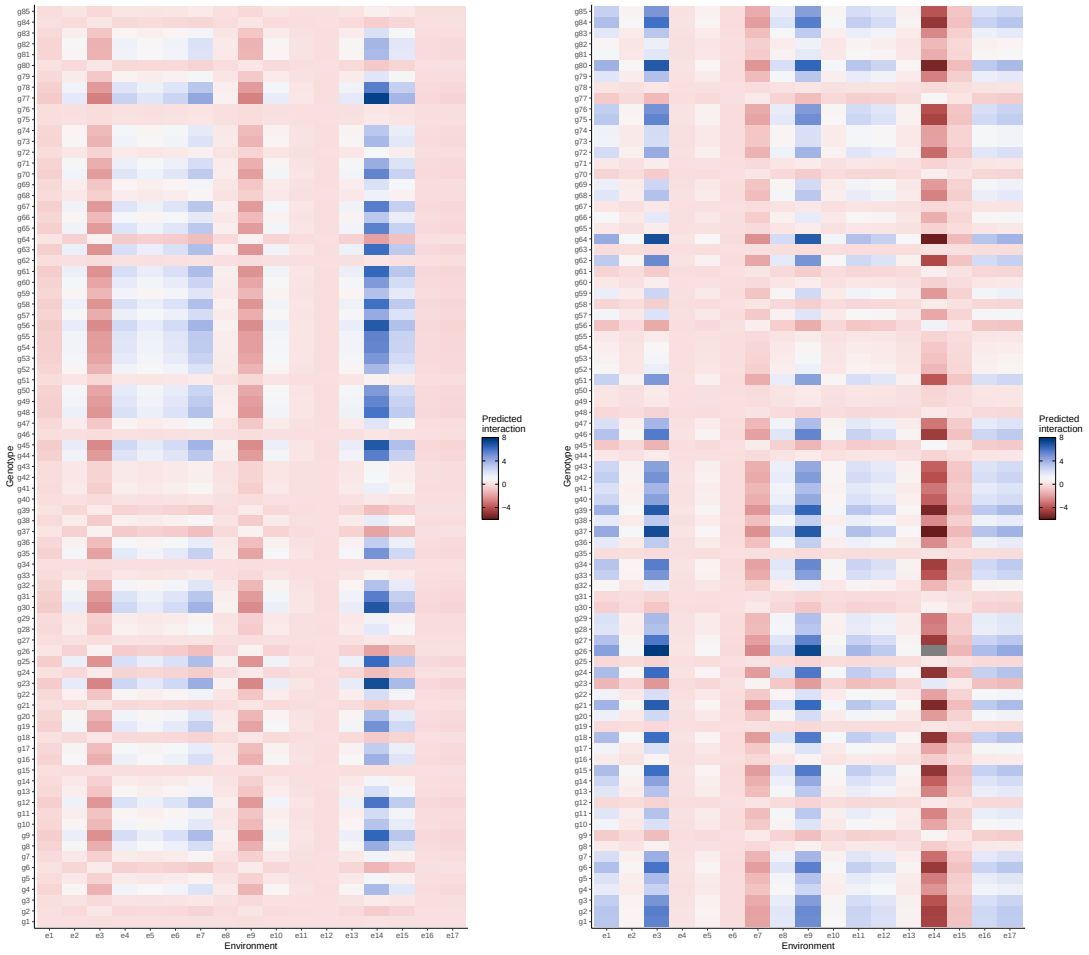
$$
\begin{aligned}
q(\boldsymbol{\tau}) &\propto \exp\{E_{-\boldsymbol{\tau}}[\log p(\boldsymbol{y}|\boldsymbol{\Theta}) + \log p(\boldsymbol{\tau})]\} \\
q(\tau) &\propto \Gamma(a_{q(\tau)}, a_{q(\tau)}) \\
a_{q(\tau)} &= a + \frac{N}{2} \\
b_{q(\tau)} &= \frac{1}{2}\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij}^2 - 2\tilde{\mu}y_{ij} - 2y_{ij}\tilde{g}_i - 2y_{ij}\tilde{e_j} - 2y_{ij}\tilde{\lambda}\tilde{\gamma_i}\tilde{\delta_j} + \tilde{\mu^2} + 2\tilde{\mu}\tilde{g}_i + 2\tilde{\mu}\tilde{e_j} + 2\tilde{\mu}\tilde{\lambda_1}\tilde{\gamma_i}\tilde{\delta_j} + \tilde{g_i^2} + \\
&\quad 2\tilde{g}_i\tilde{e_j} + 2\tilde{g}_i\tilde{\lambda_1}\tilde{\gamma_i}\tilde{\delta_j} + \tilde{e_j^2} + 2\tilde{e_j}\tilde{\lambda_1}\tilde{\gamma_i}\tilde{\delta_j} + \tilde{\lambda^2}\tilde{\gamma_i^2}\tilde{\delta_j^2}) + b
\end{aligned}
$$

3.B Interaction term heatmaps - real data set



(a) 50% quantile.

Figure 3.B.1: Quantile 50% of the interaction term, from the posterior variational distribution ($Q = 1$).



(a) 5% quantile.

(b) 95% quantile.

Figure 3.B.2: Quantiles 5% and 95% of the interaction term, from the posterior variational distribution ($Q = 1$).

CHAPTER 4

Bayesian Additive Main Effects and Multiplicative Interaction Models Using Tensor Regression for Multi-environmental Trials

We propose a Bayesian tensor regression model to accommodate the effect of multiple factors on phenotype prediction. We adopt a set of prior distributions that resolve identifiability issues that may arise between the parameters in the model. Further, we incorporate a spike-and-slab structure that identifies which interactions are relevant for inclusion in the linear predictor, even when they form a subset of the available variables. Simulation experiments show that our method outperforms previous related models and machine learning algorithms under different sample sizes and degrees of complexity. We further explore the applicability of our model by analysing real-world data related to wheat production across Ireland from 2010 to 2019. Our model performs competitively and overcomes key limitations found in other analogous approaches. Finally, we adapt a set of visualisations for the posterior distribution of the tensor effects that facilitate the identification of optimal interactions between the tensor variables, whilst accounting for the uncertainty in the posterior distribution.

4.1 Introduction

The phenotypic performance of a cultivar is associated with many potentially interacting variables (Hara et al., 2021). These may include, but are not limited to: genetic factors, environmental exposure, soil type, climatic conditions, and season. Any combinations of these factors may contribute, either positively or negatively, to the variability of the production of the crop of interest (Kross et al., 2020). Statistical modelling of the effect of these variables, both singly and jointly, is an important decision-making tool for farmers and those in the agricultural sector for predicting e.g. yield (Adisa et al., 2019).

One of the main interactions believed to impact most on crop production is between genotype and environment, which we denote for notational convenience as $G \times E$. This sort of interaction is characterised by cultivars that do not behave consistently in differing environments. Therefore, it is necessary to estimate the amount of variation in crop yield that is caused by it. Many models have been proposed to estimate $G \times E$ (Gauch Jr et al., 2008; Crossa et al., 2010; Gauch Jr, 2013). The most popular is the additive main effects and multiplicative interaction (AMMI) model (Gauch Jr, 1988), which consists of two components. The first term is the additive component, which contains the main effects of categorically structured genotype and environmental factors. The second term involves a sum of a multiplication of parameters, which are constrained to an orthonormal space and represent how strong/weak the interactions between the genotypes and environments are. The AMMI model is restricted to using only these two covariates. While genotype and environment interactions are key components in predicting crop yield, a comprehensive understanding requires the consideration of additional factors.

Our approach allows for more components beyond genotype and environment to be included in the AMMI model. We follow the Bayesian tensor regression technique of Guhaniyogi et al. (2017) to allow for any number of interacting categorical factors. Tensors are algebraic structures that generalise matrices and provide a generic way of describing multidimensional arrays on a given number of axes. Tensor decomposition methods have the advantage of capturing the information

in the data with a multi-linear structure and bring a unique representation without the requirement for additional constraints like sparsity or statistical independence (Jørgensen et al., 2018). The two main tensor decompositions are the PARAFAC (Carroll and Chang, 1970; Harshman et al., 1970) and Tucker models (Tucker, 1963). Tensors have been used in many fields of study, including physics (Gaillac et al., 2016), chemistry (Facelli, 2011), medicine (Peyrat et al., 2007), and data mining (Mørup, 2011). Guhaniyogi et al. (2017) propose a tensor-based Bayesian regression model where vector/tensor covariates are used to estimate a univariate response through a class of multiway shrinkage priors. They illustrate the model on real-world data from the brain connectome and provide theoretical results concerning the speed at which the posterior distribution converges to the true posterior (i.e., the contraction rate). Similarly, Papadogeorgou et al. (2021) propose a soft tensor regression to investigate the connection between human traits and brain structural connectomics.

In this paper, we propose the Bayesian Additive Main effects and Multiplicative Interaction Tensor (BAMMIT) model, which generalises the AMMI model to contain a tensor of interacting terms. We extend the standard AMMI model to include new parameters to the additive and multiplicative terms of the model, taking into account factors other than genotype and environment on the phenotype of a given cultivar. Common extra factors might include soil types, time points, or growth stages. Our methodology captures both minor and major interactions of these factors and identifies which of these are important and need to be included in the model. We present our new model in a Bayesian hierarchical format where we place prior distributions on the main and tensor product terms so as to guarantee the model's identifiability and impose orthonormality constraints, which are an essential part of both the original AMMI and the BAMMIT models.

We evaluate our proposed approach through a set of simulation experiments. Our interest is to investigate the model's performance under different sample sizes and degrees and complexity of interaction structures. We compare the predictions from our model with other machine learning models and standard linear mixed effects models in terms of the root mean squared error (RMSE) and the coefficient of determination (R^2).

The remainder of this paper is structured as follows. In Section 4.2.1, we review the AMMI model and present the constraints imposed on its two components. In Section 4.2.2, we introduce our BAMMIT model with its extended additive and multiplicative terms. We outline the interpretability and identifiability constraints, as well as the prior distributions considered for the parameters and a description of how we obtain the posterior distributions. In Section 5.4, we compare the results from BAMMIT with other relevant models based on synthetic data. In Section 4.4, we analyse real-world data on wheat production in Ireland from 2010 to 2019. We also utilise a new set of visualisations to assess the posterior distributions of the components of the BAMMIT model, and identify optimal interactions and their associated uncertainty. Finally, we review and discuss the findings of the work in Section 4.5.

4.2 Methods

In this section, we review the original AMMI model and define terminology and notation. We then introduce the BAMMIT model detailing the necessary constraints to ensure identifiability as well as the prior distributions and inferential scheme. In real-world scenarios the models we introduce below often include additional replicate or block effects that complicate the hierarchical structure of the data. However we remove these for simplicity of exposition in our model definitions, and include them only in our case studies.

4.2.1 AMMI model

The traditional AMMI model takes into account only two categorical factors, commonly genotype and environment, and is given by a combination of two parts, one additive and one multiplicative. Let y_{ij} be the outcome variable with i representing genotype and j environment. We write the model as:

$$y_{ij} = \mu + b_i^{(1)} + b_j^{(2)} + \sum_{q=1}^{Q} \lambda_q \beta_{iq}^{(1)} \beta_{jq}^{(2)} + \varepsilon_{ij},\ \varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2), \tag{4.1}$$

where $b_i^{(1)}$ and $b_j^{(2)}$ represent the marginal effect of the i^{th} genotype and j^{th} environment, respectively, $i = 1, ..., B_1$ and $j = 1, ..., B_2$. The bilinear term (i.e. the

summation) is composed of Q components, each of which contains a variable λ_q and the scores $\beta_{iq}^{(1)}$ and $\beta_{jq}^{(2)}$. The parameter λ_q measures the interaction strength of the q^{th} component and is usually ordered such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_Q$. The scores $\beta_{iq}^{(1)}$ and $\beta_{jq}^{(2)}$ represent the importance of the i^{th} genotype and the j^{th} environment in the interaction. To ensure identifiability, the bilinear term is constrained so that $\sum_i \beta_{iq}^{(1)} \beta_{iq'}^{(1)} = \sum_j \beta_{jq}^{(2)} \beta_{jq'}^{(2)} = 0$, for $q \neq q'$ and $\sum_i (\beta_{iq}^{(1)})^2 = \sum_j (\beta_{jq}^{(2)})^2 = 1$.

There are a range of approaches for estimating the parameters of the AMMI model. In the frequentist paradigm, the additive term of Equation (4.1) is estimated by ordinary least squares ignoring the interaction term, and subsequently a singular value decomposition (SVD) of the matrix of residuals is used to estimate the multiplicative terms (Gabriel, 1978). In this case, in addition to the constraints applied to the multiplicative term, it is generally ensured that $\sum_i b_i^{(1)} = \sum_j b_j^{(2)} = 0$, in order to maintain orthogonality between the main effects and interaction terms, as well as to directly interpret the parameters.

Within the Bayesian context, Viele and Srinivasan (2000) proposed the use of Markov chain Monte Carlo (MCMC) to estimate the parameters of the AMMI model ensuring that the inherent constraints of the model are not violated. Liu (2001) formulated a more stable and computationally faster Gibbs sampler. Crossa et al. (2011) and Perez-Elizalde et al. (2012) proposed a Gibbs sampler such that the algorithm was stabilised and incorporated statistical inference in the visualisation of biplots (Gabriel, 1971), drawing credibility regions for the interaction effects. By contrast, Josse et al. (2014) introduced an approach to deal with the over-parametrisation issue of the model by defining priors for the complete set of parameters ignoring the constraints, then applying a post-processing on the posterior samples of each parameter. Sarti et al. (2023) used Bayesian additive regression trees (BART) in which a 'double-grow' BART is responsible for capturing the interaction term.

The number of terms in the summation, Q, is crucial for modelling genotype-by-environment interactions in the AMMI model. While theoretically Q can be as large as $Q \leq \min(B_1 - 1, B_2 - 1)$, allowing the model to capture all variance in the interaction, it is commonly limited to an integer between 1 and 3. This range

helps in maintaining the balance between capturing significant interaction patterns and ensuring the model's interpretability, by preventing overfitting and allowing clearer visualisation through biplots.

However, many approaches can be applied to determine the value of Q. Examples include Cornelius (1993) who applied parametric significance tests; other authors who employed cross validation techniques (dos S. Dias and Krzanowski, 2003; Gabriel, 2002; Hadasch et al., 2017), or those using resampling techniques (Malik et al., 2018, 2019). Examples in the Bayesian field include Perez-Elizalde et al. (2012) and da Silva et al. (2015) where the prior choice and Bayes factor deal with determining the number of components of the multiplicative term. The non-parametric Bayesian approach of Sarti et al. (2023) bypasses the need to provide Q completely but, like many BART models, suffers from interpretability problems due to the complexity of the regression trees.

One of the reasons for the popularity of the AMMI model is its strong predictive performance (Gauch Jr, 2006; Gauch Jr et al., 2008), accuracy (Gauch and Moran, 2019) and its stability evaluation system (Gauch Jr, 1988; Yue et al., 2022). Given its desirable properties, many extensions can be found in the literature, as highlighted above. In this work, we aim to maintain the structure of the AMMI model and add the effects of other categorical factors that are commonly available in real-world multi-environmental trials.

4.2.2 BAMMIT model

The model in (4.1) can be extended to include the effect of many factors apart from genotype and environment. Let $y_{ij\ldots v}$ be an outcome variable, in a setting with a total of N observations and V predictors. We define the BAMMIT model as:

$$y_{ij\ldots v} = \mu + b_i^{(1)} + b_j^{(2)} + \cdots + b_v^{(V)} + \sum_{q=1}^{Q} \lambda_q \left(\beta_{iq}^{(1)} \times \beta_{jq}^{(2)} \times \cdots \times \beta_{vq}^{(V)}\right) + \varepsilon_{ij\ldots v}, \quad (4.2)$$

where $\varepsilon_{ij\ldots v} \sim \mathrm{N}(0, \sigma^2)$. This is similar to the AMMI model described in (4.1), however now we have V factors instead of only two. Alternatively, we can rewrite the coefficients of the additive and multiplicative terms of (4.2) in tensor notation. Let $\boldsymbol{b}^{(v)} = (b_1^{(v)}, \ldots, b_{B_v}^{(v)})^\top$ be a B_v-dimensional vector of parameters of the

v^{th} predictor and $\boldsymbol{\beta}_q^{(v)} = (\beta_{1q}^{(v)}, \ldots, \beta_{B_v q}^{(v)})^\top$ be a B_v-dimensional vector of singular values, with $q = 1, ..., Q$. Binding the column vectors $\boldsymbol{\beta}_q^{(v)}$, we get $\boldsymbol{\beta}^{(v)}$, a matrix of dimension $B_v \times Q$. We define $N = \left(\prod_{v=1}^{V} B_v\right)$ as the total number of observations (though, for example, replication may increase N without any need for extra parameters).

For notational convenience, we define a cumulative direct sum and a cumulative Kronecker product resulting in an N-dimensional vector as $\bigoplus_{v=1}^{V} \mathbf{b}^{(v)} = \mathbf{b}^{(1)} \bigoplus \cdots \bigoplus \boldsymbol{b}^{(V)}$ and $\bigotimes_{v=1}^{V} \boldsymbol{\beta}_q^{(v)} = \boldsymbol{\beta}_q^{(1)} \otimes \cdots \otimes \boldsymbol{\beta}_q^{(V)}$, respectively. The direct sum operation is defined such that for vectors $\mathbf{a} = (a_1, a_2)^\top$ and $\mathbf{b} = (b_1, b_2, b_3)^\top$, for example, $\mathbf{a} \bigoplus \mathbf{b} = (a_1 + b_1, a_1 + b_2, a_1 + b_3, a_2 + b_1, a_2 + b_2, a_2 + b_3)^\top$.

Following the tensor notation presented, the BAMMIT model can be written more compactly as:

$$\boldsymbol{y} \;=\; \mu \mathbf{1}_N^\top + \bigoplus_{v=1}^{V} \boldsymbol{b}^{(v)} + \sum_{q=1}^{Q} \lambda_q \left(\bigotimes_{v=1}^{V} \boldsymbol{\beta}_q^{(v)} \right) + \boldsymbol{\varepsilon}, \tag{4.3}$$

where $\boldsymbol{y}$ is an N-dimensional vector, as before μ is the grand mean, λ_q is the strength of the q^{th} component, and $\boldsymbol{\varepsilon}$ is a noise vector such that each entry $\varepsilon_n \sim \mathcal{N}(0, \sigma^2)$, with $n = 1, \ldots, N$. Even though we assume normality, it is possible to assume other distributions for the noise vector. Moreover, it is feasible to consider a heteroscedastic structure for some components of the model, as will be presented in Section 4.4.

In Equation (4.3), each vector $\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}, \ldots, \boldsymbol{b}^{(V)}$ consists of $B_1, B_2, \ldots, B_V$ values, respectively, each of which representing the levels of a factor (e.g., 8 genotypes, 10 environments and 4 soil types would yield $B_1 = 8, B_2 = 10, B_3 = 4$, with $V = 3$). The cumulative direct sum operator $\bigoplus$ then ensures sums of main effects representing all possible combinations between levels, each corresponding to one observation in the data set. The additive term represents the individual effect of each predictor, while the summation captures via Q components the interactions between the individual effects. In the case where there is only the effect of two variables, the model in Equation (4.3) is reduced to the AMMI model. The summation term provides a regularisation on the complexity of the model, with larger Q yielding a more complex set of interactions.

The model in the form presented in Equation (4.3) allows for the inclusion and study of multiple categorical predictors beyond the standard $G \times E$ pair used in AMMI models, and the understanding of their effects in two parts, individually and when interacting. As in the traditional AMMI model, Q is fixed and represents how many multiplicative terms are included in the model. Common extra predictors that might be added to the model include soil type, time, or growth stages, amongst many others. Being able to tractably estimate the effect of each of these on a phenotype would be extremely useful for practitioners, whilst retaining the simple interpretation of the parameters in the AMMI model.

Importantly, our proposed model allows the inclusion of the entirety of potential interactions, from 2-way to V-way. Through the use of appropriate selection priors and by adjusting the value of the hyperparameter Q, we ensure that the model not only accounts for these interactions but also scales efficiently with the complexity introduced by them. Thus interactions between certain factors that are not present in the data are not incorporated. This is a key feature of the model, as by allowing for the modelling of multiple levels of interactions, the model retains predictive and explanatory power. This can be useful when we want the model to be flexible and adapt to the data, or when we have prior knowledge that the effect of an interaction should be negligible.

4.2.3 Prior distributions in the BAMMIT model

In order to ensure the tractability of the coefficients in the model, it is necessary to establish restrictions on the interaction terms, similarly to those applied in the AMMI model. However, it is not trivial to ensure the identifiability of each parameter individually, only the entire product term (Guhaniyogi et al., 2017). In the Bayesian context, these constraints are ensured from the definition at the prior level. For example, in the Bayesian AMMI model proposed by Perez-Elizalde et al. (2012), the von Mises-Fisher distribution is considered for the coefficients of the multiplicative term. In the tensor field, Guhaniyogi et al. (2017) introduce multiway shrinkage priors in their tensor regression model. In our approach, we provide a new method by which the constraints are met by applying the restrictions of the interaction term through parameter transformations which we describe next.

Formally, we frame a hierarchical model in which prior distributions of the grand mean, main additive effects and variance parameters are

$$\begin{aligned}
\mu &\sim \mathrm{N}(\mu_\mu, \sigma_\mu^2), \\
\boldsymbol{b}^{(v)} &\sim \mathrm{N}(0, \sigma^2_{\boldsymbol{b}^{(v)}}), \\
\lambda_q &\sim \mathrm{N}^+(0, \sigma_\lambda^2), \\
\sigma^{-2} &\sim \mathrm{G}(a_0, a_1), \\
\sigma_{\boldsymbol{b}^{(v)}} &\sim \mathrm{t}^+(0, a_2), \\
\sigma_\lambda &\sim \mathrm{t}^+(0, a_3),
\end{aligned}$$

where N, N^+, G, and t^+, are the Normal, truncated Normal, Gamma, and truncated t-Student distributions, respectively. The hyperparameters of the grand mean μ_μ and σ_μ^2 are fixed as are all a_k terms, $k = 0, 1, 2, 3$. We treat the additive effects as random and so estimate $\sigma_{\boldsymbol{b}^{(v)}}$, though a 'fixed effects' version could also be implemented. We express the prior knowledge on the standard deviations of the additive term parameters and the λ parameter using a truncated t distribution. Additionally, we impose that the estimated $\boldsymbol{\lambda}$ vector values are in descending order.

For the product parameters in the interaction term, we use a transformation to ensure the constraints are met and to capture all the interaction orders. Specifically, we generate an auxiliary variable $\theta_{\beta_{iq}^{(v)}}$ from a standard $\mathcal{N}(0, 1)$ distribution (the transformation is invariant to the scale of this distribution), with $i = 1, \ldots, B_v$, $v = 1, \ldots, V$, $q = 1, \ldots, Q$. Then, we centre by the mean $\mu_{\boldsymbol{\beta}_q^{(v)}}$ of the vector $\boldsymbol{\beta}_q^{(v)}$, that is, for each vector $\boldsymbol{\beta}_q^{(v)}$ we calculate its mean and then subtract it from the auxiliary variable $\theta_{\beta_{iq}^{(v)}}$ for the respective value of q. Finally, we get $\beta_{iq}^{(v)}$ via:

$$\beta_{iq}^{(v)} = \left(\theta_{\beta_{iq}^{(v)}} - \mu_{\boldsymbol{\beta}_q^{(v)}}\right) \left[\sum_i \left(\theta_{\beta_{iq}^{(v)}} - \mu_{\boldsymbol{\beta}_q^{(v)}}\right)^2\right]^{-1/2}. \tag{4.4}$$

By applying this procedure to the parameters of the matrix $\boldsymbol{\beta}^{(v)}$ guarantees that the identifiability constraints (2) and (3) of the model are met in the inferential process. To ensure individual variables can be removed from the interaction terms (and hence capture sub V-way interactions) we include extra parameters in the

prior to give:

$$\beta_{iq}^{(v)} = \left[\left(\theta_{\beta_{iq}^{(v)}} - \mu_{\boldsymbol{\beta}_q^{(v)}}\right) + M\gamma_q^{(v)}\right]\left[\left(\sum_i \left(\theta_{\beta_{iq}^{(v)}} - \mu_{\boldsymbol{\beta}_q^{(v)}}\right)^2\right)^{1/2} + M\gamma_q^{(v)}\right]^{-1},$$

where the binary variable $\gamma_q^{(v)}$ acts as a switch in the model, allowing the model to include or exclude the effects of the corresponding variable v depending on the data. The value M is set as a large constant. Thus $\gamma_q^{(v)} = 1 \rightarrow \beta_{iq}^{(v)} \approx 1$ and so that interaction is removed. When $\gamma_q^{(v)} = 0$ the interaction is included as normal in the BAMMIT model. To achieve this, we impose the additional prior distributions:

$$\begin{aligned} \gamma_q^{(v)} &\sim \text{Bernoulli}(p_q^v), \\ p_q^{(v)} &\sim \text{Beta}(1, 10). \end{aligned}$$

Our approach here is akin to a spike-and-slab structure where we can control the inclusion or exclusion of certain variables based on the data, and consequently determine which interactions are involved.

4.3 Simulation Studies

In this section we examine the performance of our proposed methodology in three different ways. First we assess whether the proposed priors are indeed capturing the lower and higher order interactions. Then, we examine whether the inclusion of additional variables in the model truly improves prediction and observe how the model behaves with the insertion of these new variables. Finally, we conduct a comparative analysis between our method and other existing methodologies.

4.3.1 Simulation scenarios

The simulation scenarios were designed considering $V \in \{2, 3, 4\}$. In each scenario, we configure the number of levels $(B_1, ..., B_V)$ to allow for differences in the interaction structures. We set up the values as follows:

(i) $V = 2$ and $N = 120$, with $B_1 = 12$, $B_2 = 10$. This is the classic AMMI approach with only two variables.

(ii) $V = 3$ and $N = 480$, with $B_1 = 12$, $B_2 = 10$, $B_3 = 4$. This is an extension of the AMMI approach with three variables. This scenario is chosen to illustrate the model's ability to capture multiple levels of interaction (i.e. 2 and/or 3-way). Specifically, the interaction term of Equation (4.3) is partitioned into the following equations, where I_{ij} and I_{ijk} are used to represent the 2 and 3-way interactions of interest, respectively:

$$(a) \qquad I_{ij} = \sum_{q=1}^{Q} \lambda_q \left(\beta_{iq}^{(1)} \beta_{jq}^{(2)}\right),$$

$$(b) \qquad I_{ijk} = \sum_{q=1}^{Q_1} \lambda_{1q} \left(\beta_{iq}^{(1)} \beta_{jq}^{(2)}\right) + \sum_{q=1}^{Q_2} \lambda_{2q} \left(\beta_{iq}^{(1)} \beta_{jq}^{(2)} \beta_{kq}^{(3)}\right),$$

$$(c) \qquad I_{ijk} = \sum_{q=1}^{Q} \lambda_q \left(\beta_{iq}^{(1)} \beta_{jq}^{(2)} \beta_{kq}^{(3)}\right).$$

(iii) $V = 4$ and $N = 960$, with $B_1 = 12$, $B_2 = 10$, $B_3 = 4$, $B_4 = 2$. This is a further extension with four variables which allows for potentially many complex and multi-layer interactions. In this case, we are not partitioning the interaction term of Equation (4.3). Therefore, the data is simulated while solely taking into account the 4-way interaction.

We specifically use scenario (ii) to assess whether the proposed priors are effectively capturing both lower and higher-order interactions. We chose to examine scenario (ii) because it offers a simpler context than scenario (iii) for this evaluation, and it is not limited to a 2-way interaction, as in scenario (i). The rationale behind considering sub-cases (a) and (b) in scenario (ii) is to discern if our model can recover the lower-order interactions even when we fit a model with a higher order interaction term. Similarly, we aim to investigate the effect of the parameter Q when simulating the model with varying orders of interactions. Upon validating that the model successfully captures interactions at multiple levels, we then consider simulation scenarios (i), (ii) part (c), and (iii) to investigate the impact of incorporating additional predictors into the model.

To specify the number of terms for the interaction, we define Q_{sim} as the value of Q used in the simulation, which can take values from the set $\{1, 2, 3\}$. Corre-

spondingly, we set λ such that:

Q_{sim}	λ
1	{10}
2	{8,10}
3	{8,10,12}

Table 4.3.1: Values of λ for each Q_{sim}.

For scenario (ii) part (b) we note that there are two Q values (one for each of the latent interactions), so we set the values of λ for both $Q_{1,\text{sim}}$ and $Q_{2,\text{sim}}$ as they are presented in Table 4.3.1. For the sake of simplicity, we assume $Q_{1,\text{sim}} = Q_{2,\text{sim}}$ in all cases. Specifically, considering the configuration of this scenario, there are 480 observations and three predictors, setting 12 levels for the first predictor, 10 for the second and 4 for the third. Given this number of observations and variables, we generate three configurations of training data sets: one where the value of $Q_{1,\text{sim}} = Q_{2,\text{sim}} = 1$; another where $Q_{1,\text{sim}} = Q_{2,\text{sim}} = 2$; and finally, a data set in which $Q_{1,\text{sim}} = Q_{2,\text{sim}} = 3$; where the λ values correspond to their respective values in Table 4.3.1. The same understanding extends to generate the corresponding test data in all scenarios. In total, we generate 30 simulated data sets (15 for training and 15 for testing). In all scenarios, we set $\mu = 100$, and $\sigma = 1.5$. The settings and model parameterisations for our current simulation study are derived from previous simulation experiments on similar models in the literature (Josse et al., 2014; Sarti et al., 2023).

To fit the BAMMIT model, we run a Markov chain Monte Carlo (MCMC) algorithm through the probabilistic programming language Just Another Gibbs Sampler (JAGS; Plummer et al., 2003) and the R package `R2jags` (Su and Yajima, 2021). We fit five BAMMIT models to each simulated data set, by varying $Q \in \{1, 2, 3, 4, 6\}$. We use $\mu_\mu = 100$, $\sigma^2_\mu = 10$, $a_0 = a_1 = 0.1$, $a_2 = a_3 = 1$, and $M = 10,000$. We use three chains, 4,000 iterations per chain, discarding the first 2,000 as burn-in, and a thinning rate of two. Regarding computational time, a data set with three predictors ($V = 3$), $N = 100$ and $Q = 1$ takes on average three minutes to run, whilst to run a data set with the same configuration, but $Q = 3$

takes 17 minutes. We discuss computational issues further in Section 4.5. All experiments[2] were implemented in `R` on a MacBook Pro with a 1.4 GHz Quad-Core Intel Core i5 and 8 GB of RAM.

To assess the performance of the model when $V > 2$ (and so standard AMMI cannot be applied), we compare BAMMIT with two models extensively employed for prediction purposes, namely Random Forests (RF) and eXtreme Gradient Boosting (XGB). A Bayesian factorial mixed (BFM) model (Rouder et al., 2017) is also fitted for comparison. Ultimately, we compare our approach with the traditional AMMI and the more recent AMBARTI model (Sarti et al., 2023), though these are unavoidably restricted to using only the first two variables. For the RF model, we use the package `randomForest` (Liaw et al., 2002) selecting the default settings, `mtry`= 2 and 500 trees. For the XGB model, we use the package `xgboost` (Chen et al., 2019) setting 50 iterations. For the AMBARTI model we use the package `AMBARTI` [3] setting 50 trees, 500 as burn-in and 1000 iterations as post burn-in. For the BFM model, the main effects are included, along with the insertion of interactions of all orders, and we follow the same priors as applied to the BAMMIT model. All the models were fitted to the training data. We checked the accuracy, using the test data, by comparing the posterior mean estimates with the true parameter values used in the simulations. We use the root mean squared error (RMSE) to measure predictive power (how close $\hat{y}$ is to the true y) and R^2 to assess the proportion of explained variability.

4.3.2 Simulation results

In Figure 4.3.1, we present scatterplots of true versus estimated values in the case where we simulated the interaction structure as given in scenario (ii) part (a), setting $Q_{\text{sim}} = 1$. The first plot corresponds to the fit of the model that follows the same structure as the simulated data, which considers all main effects along with the true simulated 2-way interaction. This model includes solely the step of imposing identifiability constraints, as outlined in Equation (4.4). The second corresponds to the BAMMIT model fit. It can be observed that both fits

[2]The code used is available at https://github.com/Alessandra23/bammit

[3]The code is available at https://github.com/ebprado/AMBARTI.

exhibit similar results, the expected result, indicating that the BAMMIT structure captures the interaction between the two variables.

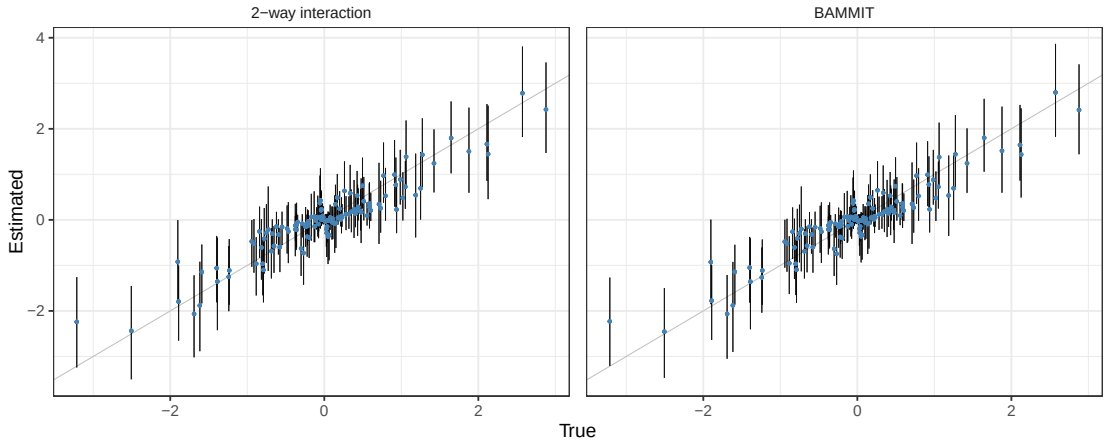


Figure 4.3.1: Scatterplots of true versus estimated interaction term for simulation scenario (ii) part (a), setting $Q_{\text{sim}} = 1$ and $\lambda = 10$. The left panel shows only the estimated interaction effects whilst the right panel shows the full estimated fitted values. The models were fitted with $Q = 1$. The blue points represent the posterior median and the grey bars represent the 95% credible intervals.

Figure 4.3.2 illustrates the results for the case in which the interaction is simulated according to scenario (ii) part (b). In this case, the simulation was conducted with $Q_{1,\text{sim}} = Q_{2,\text{sim}} = 1$, while the BAMMIT model was fitted with $Q = \{1, 2, 3, 4\}$. The graphs clearly highlight the importance of the value assigned to the hyperparameter Q. Specifically, when only one term is present in the interaction ($Q = 1$), the estimation is not accurate. This is because with $Q = 1$, the model cannot capture the complexity of the two different interactions present in the simulated data. However, as Q is increased the BAMMIT model successfully captures the simulated interaction structure. The lesson here is that Q needs to be sufficiently large to capture all possible latent interactions in the data.

To assess the behaviour of the BAMMIT model when data is simulated with a higher value of Q, we present in Table 5.4.2 the RMSE of the interaction for the BAMMIT model fit, varying the number of terms in $Q = \{1, 2, 4, 6\}$. The interaction structure was simulated again as given in scenario (ii) part (b), with

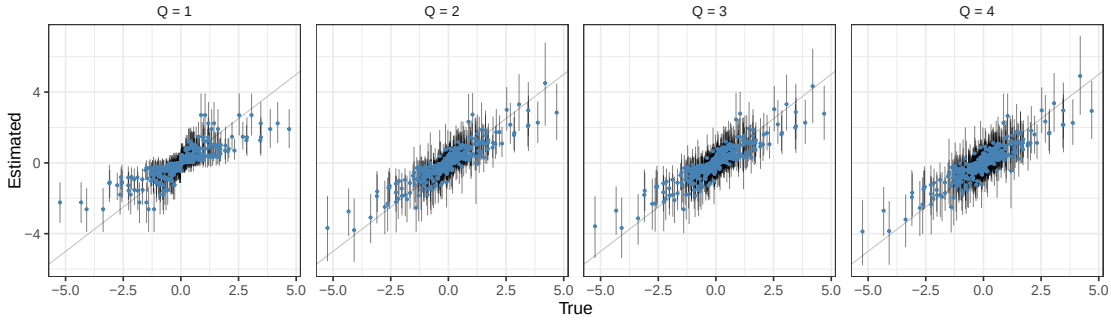


Figure 4.3.2: Scatterplots of true versus estimated interaction term for simulation scenario (ii) and Equation (4.5), setting $Q_{1,\text{sim}} = Q_{2,\text{sim}} = 1$ and $\lambda = 10$. The model was fitted with $Q = \{1, 2, 3, 4\}$. The blue points represent the posterior median and the grey bars represent the 95% credible intervals.

the assumption that $Q_{1,\text{sim}} = Q_{2,\text{sim}} = \{1, 2, 3\}$. Once more, we observe that the incorporation of more terms into the interaction enhances the model's accuracy, leading to a decrease in the RMSE of the interaction as Q increases.

	Fitted Q			
$Q_{1,\text{sim}}$, $Q_{2,\text{sim}}$	1	2	4	6
1	2.75	0.76	0.59	0.58
2	3.38	2.51	1.21	0.98
3	5.65	2.47	0.88	0.84

Table 4.3.2: RMSE for the interaction term of the BAMMIT model, considering that the interaction was simulated from Equation (4.5) and $Q_{1,\text{sim}} = Q_{2,\text{sim}} = \{1, 2, 3\}$ in the simulation. The BAMMIT model was fitted with $Q = \{1, 2, 4, 6\}$. The RMSE values reported are based on the test data, reflecting the model's performance in predicting unseen data. We would expect the model to perform satisfactorily once the fitted Q value is greater than or equal to $Q_1 + Q_2$.

Given that the model is successfully capturing both the lower and higher interactions, we now focus on presenting the results where the data simulation was conducted directly from Equation (4.3), and how the inclusion of new terms contributes to the model's prediction. Initially, the scatterplot in Figure 4.3.3 shows the comparison of the additive term, taking $V = 4$, $Q_{\text{sim}} = 2$, $N = 960$ when the true value of $\lambda = \{8, 10\}$. Each point is an estimated value of the parameters and

the error bars are the 95% credible intervals. By visual inspection, the estimates of the effects of the four main predictors are close to the true values, with narrower intervals for predictors with a greater number of levels.

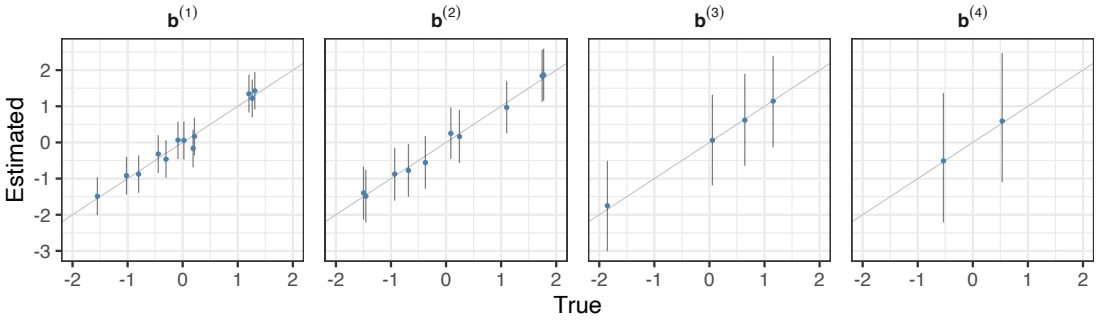


Figure 4.3.3: Scatterplots of true versus estimated additive terms for simulation scenario (iii), setting $Q_{\text{sim}} = 2$, $\lambda = \{8, 10\}$. The bars represent the 95% credible intervals.

In Figure 4.3.4, we compare the estimates against the true values in the case where the number of predictors varies. Each point represents an interaction term estimate in a total of 120 ($V = 2$), 480 ($V = 3$) and 960 ($V = 4$) points, and the bars, again, represent the 95% credible intervals. We observe that when $V = 4$, the dispersion is smaller and the interaction estimates are more concentrated around zero. This can be explained because as more predictors are added to the additive term of the model, the greater the approximation of the response by the predictors and the smaller the amount approximated by the interaction term, despite inserting more variables in both terms of the model. Also, note that the interaction is comprised of all the new variables together, and that this interaction may not be that strong. For example, suppose we are looking at the *genotype* $\times$ *environment* $\times$ *soil type* $\times$ *growth stage* interaction. In this case, the interaction of the four factors together is not as strong as if we were looking only at subsets of these interactions, such as *genotype* $\times$ *environment* $\times$ *growth stage*.

In terms of predictions, Table 4.3.3 shows the prediction RMSE and the R^2 considering the cases where we have three and four predictors in the models (simulation scenarios (ii) and (iii)). To fit BAMMIT and AMMI models we used $Q = 2$. As

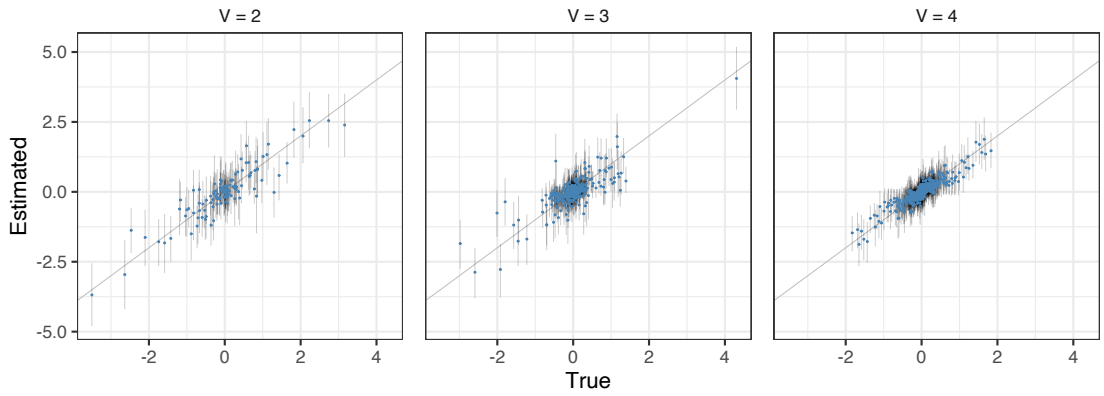


Figure 4.3.4: Scatterplots of true versus estimated interaction terms for simulations scenarios (i), (ii) and (iii) setting $Q_{\text{sim}} = 1$ and $\lambda = 10$. The interaction in scenario (ii) was generated from Equation (4.5). The bars represent the 95% credible intervals.

stated above, the AMBARTI and AMMI models were fitted disregarding the effects of the other variables. Specifically, in scenario (iii), for example, there were three predictors, but the two aforementioned models disregarded the effect of the third predictor. The BAMMIT model clearly performed better than the other two models. In addition to the prediction advantage, our model stands out from RF and XGB as it can provide a simpler structure for the interaction between the variables, while at the same time providing estimates based on full posterior distributions rather than point estimates. Our approach was better than the BFM, that includes all the possible interactions between the variables. Our BAMMIT model was able to satisfactorily explain the variability of the response variable, since the R^2 obtained in all scenarios was above 75%. In a real world scenario where the data were not simulated from the BAMMIT model we might expect that the machine learning approaches would be more competitive in terms of their performance. However, they would still not allow for clear interpretation of the interaction effects.

	V = 3						V = 4					
	BAMMIT	AMBARTI	AMMI	RF	XGB	BFM	BAMMIT	AMBARTI	AMMI	RF	XGB	BFM
RMSE	**1.62**	2.54	2.52	1.68	2.06	1.87	**1.64**	2.74	2.71	1.74	2.05	1.72
R^2	**0.78**	0.02	0.02	0.58	0.69	0.72	**0.81**	0.01	0.01	0.68	0.78	0.79

Table 4.3.3: RMSE and R^2 for $\hat{y}$ on *out-of-sample* data for scenarios (ii) and (iii), setting $Q_{\text{sim}} = Q = 2$. The RMSE values reported are based on the test data.

4.4 Case Study

In this section, we investigate the performance of the model on a real data set. The data was collected over ten years (2010 – 2019) and concerns the production of a common species of wheat (*Triticum aestivum L.*) in Ireland, with the response being the yield of wheat measured in tonnes per hectare (*t/ha*). The data comes from the Horizon2020 EU InnoVar project[4] and was supplied by the Irish Department of Agriculture, Food and the Marine. The experiments were conducted using a randomised complete block design with four replicates. The data set contains 85 genotypes and 17 environments, all anonymised and named as $g_1, \ldots, g_{85}$ and $e_1, \ldots, e_{17}$, respectively. Here, environment refers only to the location, instead of the common interpretation of environment as a location-year combination. Owing to not all genotypes being observed in each location in all seasons, the total number of observations *genotype* $\times$ *environment* $\times$ *year* $\times$ *block* is 6,368, rather than 57,800. It exemplifies one of the advantages of the BAMMIT model, that is able to impute the missing combinations as part of the model fit.

A subsample of this data was previously explored by Sarti et al. (2023), considering only two factors: genotype and environment. However, in our work, we include the additional variables *year* and *block*, present in the Irish data set, as a third and fourth effect in the BAMMIT model. We expect to detect if there is variability between years across environments and genotypes, and to determine such an interaction. Previously Hara et al. (2021) showed that the ability to predict the yield in a certain year can be useful for making decisions such as cultivation

[4] www.h2020innovar.eu

planning and storage. Specifically, the model structure for the data is

$$y_{ijtr} = \mu + b_i^{(1)} + b_j^{(2)} + b_t^{(3)} + b_{jr}^{(4)} + \sum_{q=1}^{Q} \lambda_q \beta_{iq}^{(1)} \beta_{jq}^{(2)} \beta_{tq}^{(3)} + \epsilon_{ijtr}, \qquad (4.5)$$

where the indexes i, j, t and r are associated to genotypic, environmental, time and block effects, respectively. We nest the block effect within environment and allow an environment specific variance for each, thus removing any experimental variation associated with that specific site not related to the response. To avoid numerical under- or overflow we standardise the response before fitting, but convert all predicted values back to the original scale for ease of interpretability.

To fit the model, we partition the data into training and testing sets by selecting three out of the four available blocks for training. The selection of this number of blocks is to ensure a more comprehensive representation of the inherent variability across blocks. The remaining block is used for validation, resulting in 4,776 observations for training and 1,592 for testing. Therefore, we have $V = 4$, $B_1 = 85$ genotypes, $B_2 = 17$ environments, $B_3 = 10$ years and $B_4 = 3$ blocks. Assuming that there is little prior information for hyperparameter specification, we follow a non-informative approach, with $a_0 = a_1 = 0.1$. To input the number of Q terms, we run the model with $Q = \{1, 2, 3, 4\}$. The model is fitted with three Markov chains, 4,000 iterations per chain, discarding 2,000 and a thinning rate of two.

We compare our BAMMIT approach prediction performance with traditional AMMI, AMBARTI and a Bayesian factorial model. For the AMBARTI model, we consider 50 trees, 1000 iterations as burn-in and 1000 iterations as post burn-in. The fitted Bayesian factorial model we use follows the same priors assumed for the BAMMIT model. We define the model structure as:

$$y_{ijtr} = \mu + b_i^{(1)} + b_j^{(2)} + b_t^{(3)} + b_{jr}^{(4)} + b_{ij}^{(1,2)} + b_{it}^{(1,3)} + b_{jt}^{(2,3)} + b_{ijt}^{(1,2,3)} + \epsilon_{ijtr}.$$

We note that this model no longer parameterises the interactions in a latent tensor space (as our models do) and so uses a far greater number of parameters. Since we only evaluate our model on out of sample data the comparison between these approaches remains valid.

In addition to evaluating the model's performance in terms of prediction, we are interested in answering some specific questions. Initially, when fixing the year and block effect, we would like to know which genotype has the best performance, in which environment, and also which environment provides the highest yield and the lowest uncertainty. Subsequently, we explore the interactions between these variables to find potential year/environment/genotype combinations which provide optimal or sub-optimal yield performance.

4.4.1 Results

To assess the performance of the predictions $\hat{y}$, we display the RMSE and the R^2 in Table 4.4.1 for each model. As the AMBARTI and classic AMMI models can only handle the genotype and environment variables, when fitting these models we take into account all rows in the data set and ignore the year and block variables, so that rows corresponding to the same genotype and the same environment are treated as different. We considered $Q = 4$ for the BAMMIT and AMMI models. In terms of convergence, all the models have had their convergence checked, with $\hat{R} \approx 1$. In particular, all parameters of the BAMMIT model converge, except for the individual values of the parameters $\boldsymbol{\beta}_q^{(v)}$, as expected.

	BAMMIT	AMBARTI	AMMI	BFM
RMSE	**0.59**	1.66	1.60	0.72
R^2	**0.92**	0.34	0.38	0.85

Table 4.4.1: Metrics for *out-of-sample* y for the BAMMIT, AMBARTI, AMMI and Bayesian factorial mixed (BFM) models. The AMBARTI and AMMI models ignored the effects of the year and the block variables. Best performance is shown in bold. The RMSE values reported are based on the test data

The results presented in Table 4.4.1 show that the BAMMIT model was superior to the other models, in both RMSE and R^2. Specially, our proposed model outperforms the BFM model, which is the main competitor to BAMMIT, since it directly includes all possible interactions. For the AMBARTI and the classical AMMI model, the inferior performance was expected, since the adjusted models do not take into account the effects of the other variables. Also, for the AMBARTI,

these results can be explained by the high numbers of genotypes and environments. As mentioned above, considering all the 10 years, there are 85 genotypes and 17 environments, and in this case the AMBARTI model is not efficient in the generation of the $2^{B_1-1}-1$ and $2^{B_2-1}-1$ two-partition combinations for genotypes and environments. Thus, due to the high numbers of possible combinations, the interaction component of the AMBARTI model is not able to estimate the interactions between genotypes and environments efficiently and the model performs poorly.

In addition to assessing model performance through RMSE and R^2 metrics, examining the estimates of the β coefficients provides further insights into the efficacy of the modelling approaches. In the BAMMIT model, the β coefficients showed superior performance metrics, highlighting the strength and significance of the interactions between the factors involved. Unlike in the AMBARTI and AMMI models, where the β estimates are constrained by the absence of year and block variables, the BAMMIT and BFM models incorporate these factors, allowing for a more comprehensive understanding of their impact. For the BAMMIT model, despite the non-convergence of the individual values of the parameters β_q^v, the overall model convergence ($\hat{R} \approx 1$) and the high R^2 suggest that the estimates effectively capture the underlying data structure. It is also worth noting that the performance gap between the BAMMIT and the competing models can partly be explained by the efficiency of the β estimates in capturing key interactions, even when some parameters show individual non-convergence. Thus, the robust performance of BAMMIT, is attributable not just to the model structure but also to the effective estimation of interaction effects through the β coefficients, which are crucial for understanding the dynamics among the genotypes, environments, and other variables such as year and block. Further analysis of these β estimates, particularly through graphical visualisation or additional statistical summaries, could provide deeper insights into their specific contributions and interactions within the model.

In Figure 4.4.1, we display the estimated probabilities $\hat{p}_q^{(v)}$ which reflect the likelihood that specific variables are included as part of the interaction terms in the BAMMIT model. The notation $\hat{p}_q^{(v)}$ specifically denotes the estimated probability that the v-th variable interacts in the q-th order interaction. For instance, a lower

$\hat{p}_q^{(v)}$ value near zero suggests a higher probability that the variable v is integral to the q-th order interaction term within the model. This interpretation helps in understanding which variables and their interactions are considered significant in influencing the response variable, which in this case is wheat yield.

Figure 4.4.1 shows the estimated values of $\hat{p}_q^{(v)}$ for each variable involved in the model (*genotype*, *environment*, and *year*), along with their 95% credible intervals. The lower values observed for *environment* and *year* suggest these variables have a higher likelihood of being involved in interaction terms, indicating they are particularly crucial in the prediction of wheat yield. This is consistent with the model's results, which indicate that wheat yield is not solely dependent on any single factor but is significantly influenced by the interactions among these variables. Furthermore, the small uncertainty ranges in these values, compared to the broader $Be(1, 10)$ prior, highlight the model's confidence in these estimates.

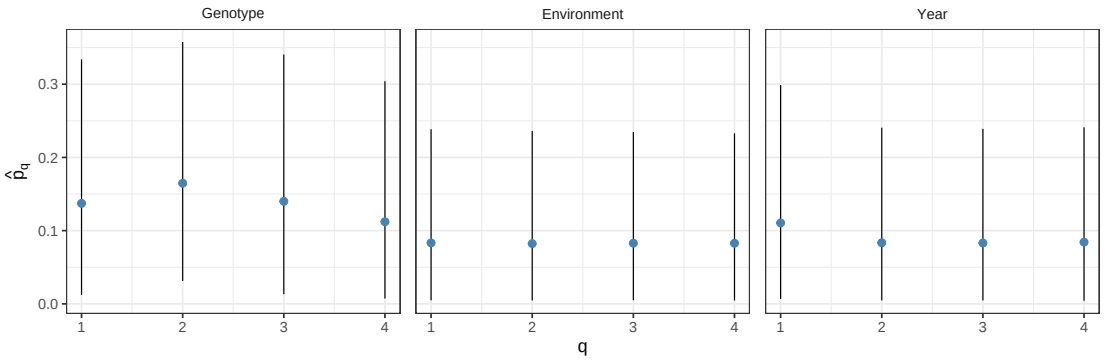


Figure 4.4.1: Estimated values of $\hat{p}_q^{(v)}$ and the 95% credible intervals. Values closer to zero indicate an increasing probability that the variable was included in the interaction term. The general low values indicate a high degree of interaction, with environment being particularly important. We note that the uncertainty ranges in these values are far smaller than that of the $Be(1, 10)$ prior.

4.4.2 Posterior visualisation

A common way to visualise the genotype and environment interactions in an AMMI model is through biplots (Gabriel, 1971). However, Sarti et al. (2023) used a

heatmap to visualise the predictions and interactions. In this visualisation approach, it is possible to identify in a more immediate way the best interactions between genotypes and environments. A shortcoming of their approach concerns the quantification of uncertainty, which cannot be observed directly on the graph. To address this particular issue, in this work we show the prediction for interactions through a heatmap as in Sarti et al. (2023) and the uncertainty is showed as Value-suppressing uncertainty palettes (VSUP) as presented by Inglis et al. (2022).

First introduced by Correll et al. (2018), value-suppressing uncertainty palettes are bivariate colour palettes that represent a measure or value and its uncertainty. The outputs for each combination of value and uncertainty in traditional bivariate palettes are often shown as a 2D square (for example, see Robertson and O'Callaghan, 1986), However, VSUP plots combine cells in the palette using a tree structure to suppress the measure or value at higher levels of uncertainty. In VSUP, when the uncertainty is low, more bins are allocated to the colour space. When increasing in uncertainty, the values are suppressed into fewer bins that blend together their colour value. By doing this, the values will become more distinct as the level of uncertainty reduces, with the intention of making it easier to detect the difference between low and high uncertainty.

In Figure 4.4.2, we display an ordered heatmap and VSUP legend for the Irish data selecting the year 2015 in block four from the test data set in the original scale. This year corresponds to the best average wheat production observed in the data. The remaining years are shown in the Appendix. For these plots we use the standard deviation of the predictions as our measure of uncertainty and the value shown is the median prediction for each variable pair. The ordering of environment in Figure 4.4.2 is based on the *descending* values of $\hat{y}$ (that is, the predicted yield). Meaning, environments with higher $\hat{y}$ values will be placed first. Conversely, the ordering of genotype is based on the *ascending* values of $\hat{y}$. This means that genotypes with lower $\hat{y}$ values will be placed first, and those with higher values will follow. This results in a visualisation where generally high predicted yield values are pushed to the top left of the plot and descend to the bottom right. The environments e_2, e_4, e_9, e_{11} and e_{16} were the worst environments observed, having a small median value compared to the others. On the other

hand, environment e_1 was the best and most stable, presenting a median yield value higher than the others and a lower uncertainty. The environment e_6 had a middling production across some genotypes, but with a higher uncertainty when compared to the others. Applying the same interpretation to the genotypes, we observed that the genotypes g_3 and g_{85} had the best performance. The genotype g_{10} presented a good production of predicted wheat on environment e_1 and e_7, however with a higher standard deviation than the other genotypes, which also produced around 15 *t/ha* in this environment. As an example of the main *genotype* $\times$ *environment* combinations, the worst observed were $g_{81} \times e_{16}$, $g_5 \times e_2$ and $g_{10} \times e_2$, while the best were $g_{85} \times e_1$, $g_3 \times e_1$, $g_5 \times e_7$ and $g_3 \times e_7$. The results are comparable to those of Sarti et al. (2023).

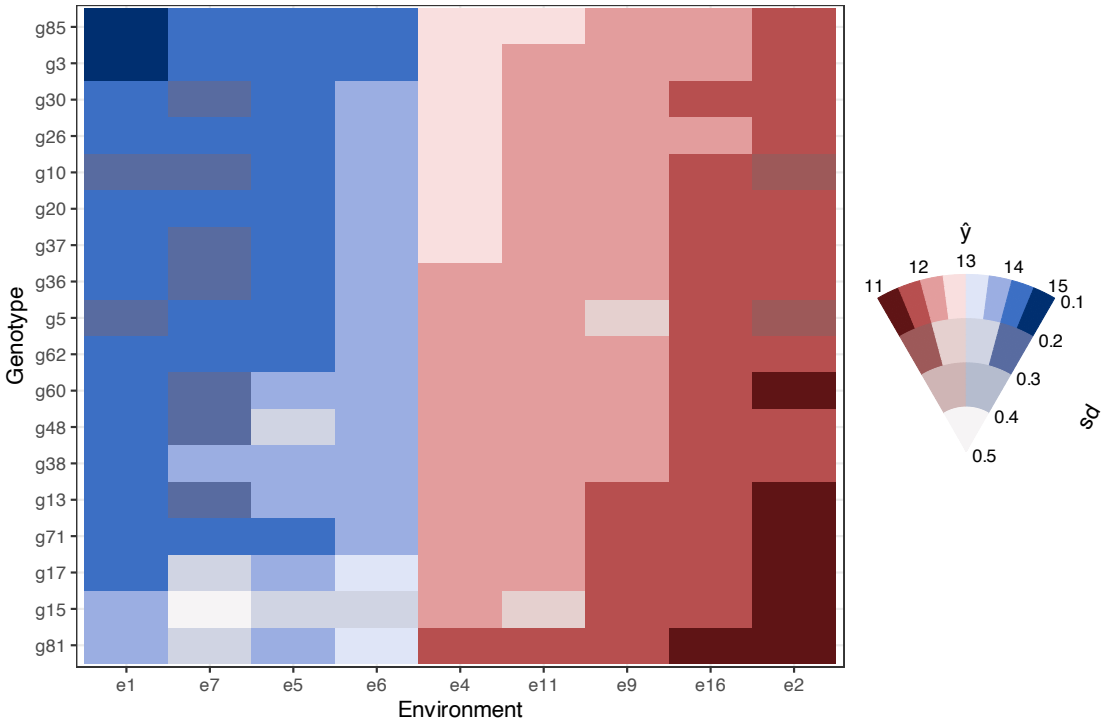


Figure 4.4.2: Predicted yields from the BAMMIT model for the wheat production data set in 2015. Production this year was high, between 11 and 15 tonnes per hectare, with positive emphasis on the environment e_1 and on the combinations $g_3 \times e_1$ and $g_{85} \times e_1$.

The proposed VSUP plots offer a clear advantage over traditional biplots by enabling quicker pattern identification and straightforward interpretation of complex interactions. They are more scalable, handling numerous genotypes and environments without losing readability. This is a notable improvement over biplots, which become cluttered as variables increase and also require a level of expertise to interpret them. VSUP plots additionally include the uncertainty, which is missing from a biplot, and allows for quick identification of variable pairs with high or low confidence.

Finally, in Figure 4.4.3 we present the box plot of $\hat{y}$ by year, alongside the actual means observed in the test data. The true means fall within the box plots, what shows the good BAMMIT model's performance in capturing the central tendency of the wheat yield in each year. The predictions suggest that the best production occurred in the years 2015 to 2017, with 2015 having a smaller variation than the others. The analysis of the variable year in this context is particularly important, as it serves as a proxy for multiple factors that change over time, such as climatic conditions and agricultural practices, which are not explicitly captured by the static variable environment (that only represents location).

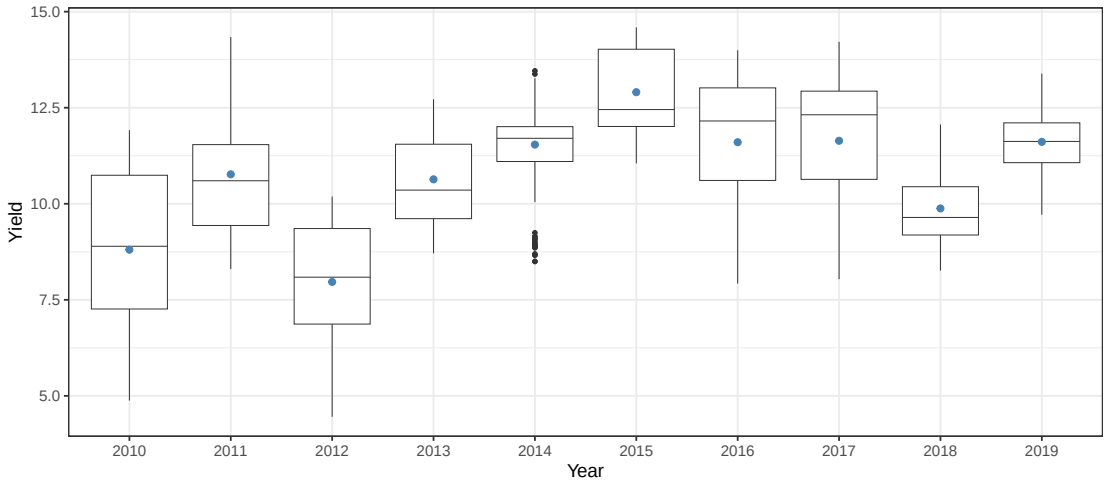


Figure 4.4.3: Box plot of the predicted wheat yield by year. The blue points represent the true mean observed in the test data.

4.5 Discussion

We proposed a generalisation of the AMMI model which extends the tensor regression approach of Guhaniyogi et al. (2017) and Papadogeorgou et al. (2021) to allow for multiple interacting categorical variables. The main idea is to allow for more realistic understanding of phenotypic effects beyond the usually considered pair of genotype and environment. We envisage that in the future such models may be used to further indicate interactions between season, soil, weather conditions, growth stage and other potential predictors. The priors we use on the hierarchical model were built not only to meet the inherent restrictions and ensure identifiability but also to capture both lower and higher-order interactions between the variables.

The simulation results suggest that the model performs well in a variety of prediction tasks whilst retaining simple interpretative output. The model was able to capture the lower and higher-order interactions between the variables regardless of the true number of terms initially specified for the multiplicative term in the simulation. The results indicated that, provided the value of Q was sufficiently large, the model's accuracy in capturing interactions at all levels was satisfactory. Secondly, the model demonstrated robustness when multiple interactions were included with differing levels of complexity, and the model retained the ability to estimate values of the parameters that matched the ground truth. Lastly, our model outperformed other well-known models, such as random forests and factorial mixed models in terms of predictive accuracy. As expected, in all scenarios, when the established number of terms in the interaction increases, the Bayesian model had a better fit, regardless of the true number used in the simulation. All our models were checked to confirm that the algorithm had converged and that any imposed prior constrains were met for all our simulated data sets.

When applied to real data, our approach was superior to other methods when evaluated on predictive performance and RMSE. The model enabled the identification of variables involved in interactions and determine which genotypes, environments and years had the highest wheat production. In this study, understanding the behaviour of the variable year was particularly important because the environment

is limited to location and is not capable of capturing annual changes in climatic conditions or agricultural strategies, for example. Furthermore, we could visualise components of the model, such as the interaction effect, and so it was possible to determine the optimal interactions. The purpose of showing the results in the visualisations presented in this paper is to aid a researcher's ability to interpret the results and improve recommendations.

In relation to the computation time, the cost for the method was high, particularly as the number of predictors V and components Q increased. For example, a data set with three predictors ($V = 3$) and $Q = 1$ took on average one minute to run, whilst for a data set with four predictors ($V = 4$) and $Q = 3$ took around 30 minutes. This drawback was compounded when the data set is large (around 5,000 total observations or more) with the model taking several hours to form a valid posterior distribution. To avoid this computational cost, potential optimisation strategies can be employed, such as parallel processing, variational inference (Blei et al., 2017; Dos Santos et al., 2022) or those as discussed in Papadogeorgou et al. (2021) and Zhang et al. (2020).

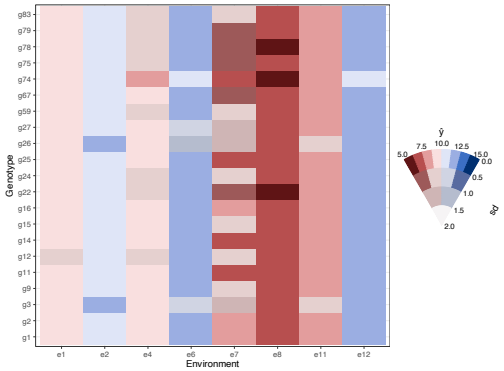
For future work, several extensions can be made to these models. By modifying the prior distributions, new structures can be added to certain predictors, thereby allowing any temporal and spatial components to have their inherent characteristics inserted into the model. Another important extension is the insertion of continuous variables, or latent representations of them, since the current structure does not allow for this type of variable. Also, the choice of the number of components Q is still arbitrary, taking into account only the typical values already mentioned in the literature. Nonetheless, a more sophisticated approach to choosing the rank Q can be applied, as shown by Guhaniyogi et al. (2017).

Appendix

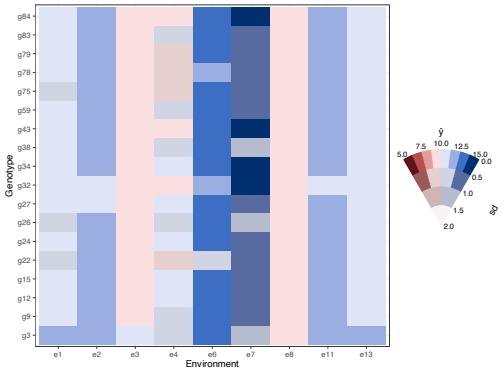
4.A Real data

We complement the results presented in the Section 4.4 by presenting in Figure 4.A.0 the graphs for the predicted yields for the BAMMIT model applied to the Irish time series data set. Analysing only the legend of the figures and looking at the value scale it is possible to see that the forecast of wheat yield for the year 2015 (presented in Figure 4.4.2) was higher than that for all the other years. In order to clearly observe the behaviour of genotype and environment predictors over the years, we scale the estimated value and the uncertainty legend to be equal across plots. Thus, years with high production and low uncertainty have an intense colour while years with contrary behaviour have a washed-out colour.

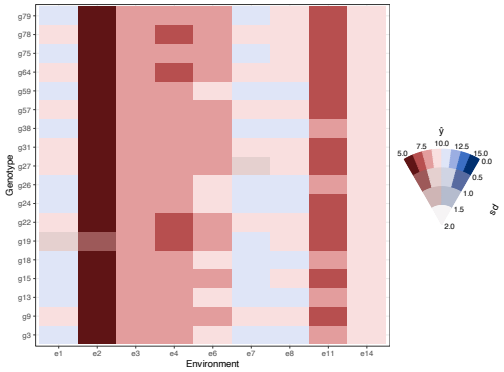
Another way to see to account the year effect, shown in Figure 4.A.1, where the VSUP legend shows the average of the years and the respective sd. Note that the sd is much larger, once the model is predicting combinations that did not exist previously (as mentioned this is one of the features of the BAMMIT model). From this plot, we can see that the best environment was e_5, performing well in all genotypes and across years. Conversely, e_4 and e_{11} performed the worst.



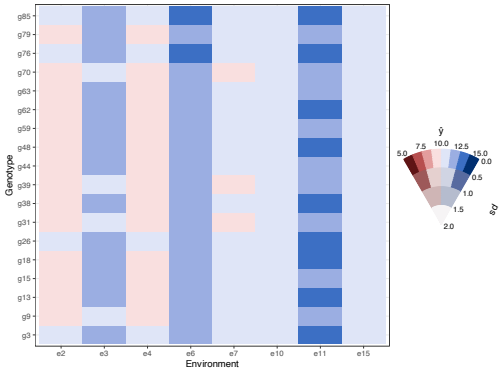
(a) 2010



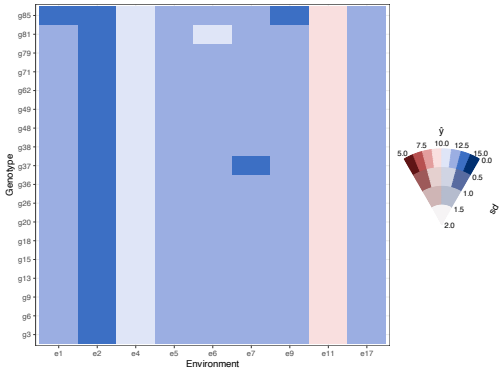
(b) 2011



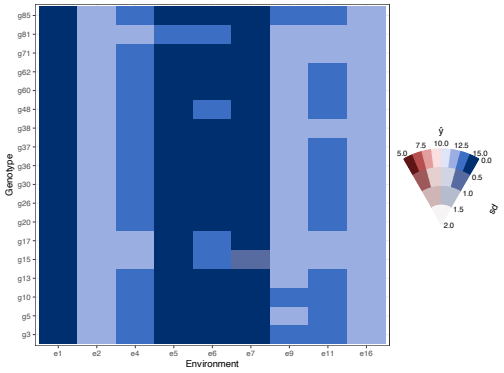
(c) 2012



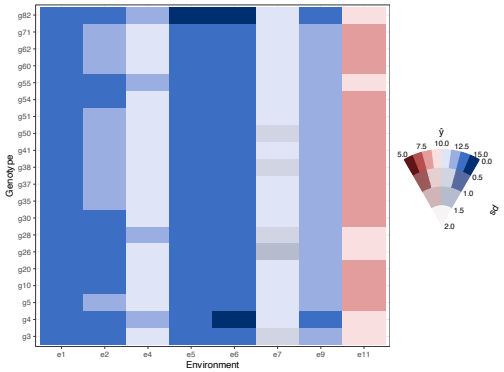
(d) 2013



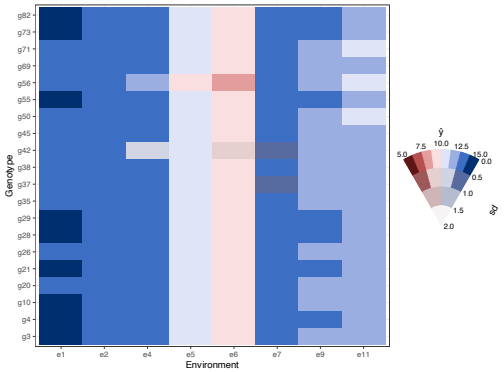
(e) 2014



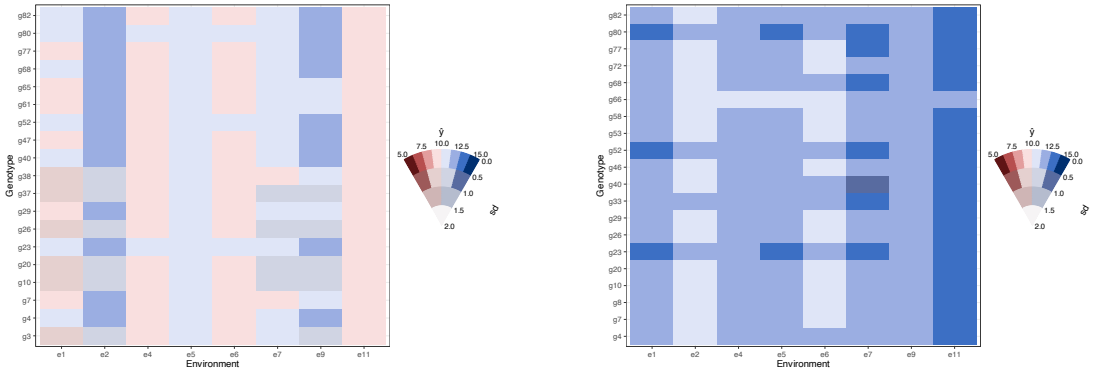
(f) 2015



(g) 2016



(h) 2017



(i) 2018 (j) 2019

Figure 4.A.0: Predicted yields from the BAMMIT model across the Irish dataset, with consistent ordering and colour scales for ease of comparison.

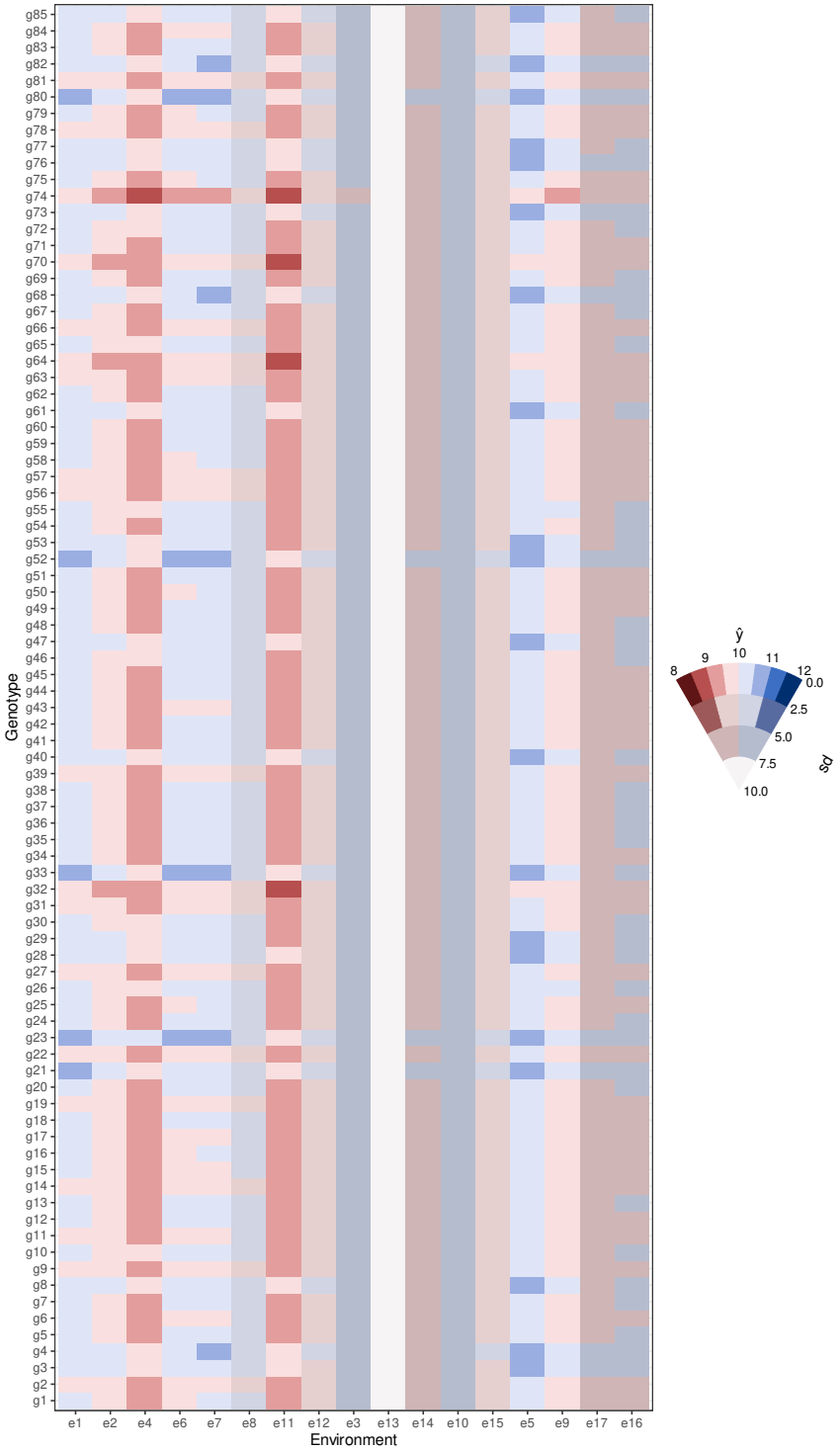


Figure 4.A.1: Predicted yields from the BAMMIT model averaged over all years of the Irish dataset.

4.B Auto-regressive BAMMIT (AR-BAMMIT)

The construction of the BAMMIT model allows new structures to be applied to the parameters, such as spatial or temporal. The inclusion of these new structures brings greater complexity to the model, but the constraints imposed on the multiplicative terms are still needed so as to guarantee the model's identifiability. As an example of this extension on the BAMMIT model, consider the Irish data presented in Section 4.4.

In this particular application, where one of the variables in the BAMMIT model is the year of production, we have the option of extending the model by applying a different structure to the time predictor in both terms, that is, additive and multiplicative. The model is now:

$$\begin{aligned}
y_{ijtr} &= \mu + b_i^{(1)} + b_j^{(2)} + b_t^{(3)} + b_j^{(4)} r + \sum_{q=1}^{Q} \lambda_q \beta_{iq}^{(1)} \beta_{jiq}^{(2)} \beta_{tq}^{(3)} + \epsilon_{ijtr}, \\
b_t^{(3)} &= \alpha_b + \phi_b b_{t-1}^{(3)} + \eta_t, \\
\theta_{tq}^{(3)} &= \alpha_\theta + \phi_\theta \theta_{(t-1)q}^{(3)} + \omega_t, \\
\beta_{tq}^{(3)} &= \left[\left(\theta_{\beta_{iq}^{(3)}} - \mu_{\boldsymbol{\beta}_q^{(3)}} \right) + M\gamma_q^{(3)} \right] \left[\left(\sum_i \left(\theta_{\beta_{iq}^{(3)}} - \mu_{\boldsymbol{\beta}_q^{(3)}} \right)^2 \right)^{1/2} + M\gamma_q^{(3)} \right]^{-1}
\end{aligned} \tag{4.6}$$

where the indexes i, j, t and r are associated to genotypic, environmental, time and block effects, respectively. The priors for the entire model follow the same structure as before (Section 4.4), such that Equation (4.6) is defined to meet the conditions of identifiability of the model. To estimate the auto-regressive parameters ϕ_b and ϕ_θ we use uniform priors in the interval $[-1, 1]$. A normal distribution with mean zero and variance 100 was considered for the parameters α_b and α_θ. The priors for η_t and ω_t were a truncated t-Student distribution with parameters zero and one.

The metrics obtained for the out-of-sample performance of the AR-BAMMIT model were equivalent to those presented in Table 4.4.1 for the BAMMIT model. Specifically, with $Q = 4$, the model that included the temporal structure achieved an RMSE of approximately 0.59. Although it did not yield a better RMSE than the BAMMIT model, the AR-BAMMIT outperformed the others. Additionally, the

application on this data illustrates the possibility of incorporating new structures into the model's predictors.

CHAPTER 5

Bayesian Clustering Additive Main Effects and Multiplicative Interaction Models for Multi-environmental Trials

The Additive Main effects and Multiplicative Interaction (AMMI) model is often used for phenotypic modelling due to its ability to capture interactions between genotypes and environments. However, this model has two primary limitations: it only accepts categorical variables and does not account for numerous other factors that could influence phenotype prediction. In this paper, we propose an extension to the AMMI model that allows for the inclusion of any number of both continuous and categorical variables. This is achieved through model-based clustering, enabling both the insertion of categorical variables and the latent representation of continuous ones. Uniquely, our method performs clustering and estimation simultaneously rather than in two stages, thereby enhancing the model's predictive power. Through both a toy data example and a simulation study, we demonstrate the strong performance of our model in terms of accuracy and predictive capability. We further illustrate the applicability of our proposed method through an analysis of wheat production in Ireland from 2010 to 2019, where annual temperature and rainfall are used to capture between-year variability.

5.1 Introduction

Crop modelling is a vital part of agricultural policy and research. Accurate crop yield prediction has a direct impact on the choices made by stakeholders (which includes farmers, policymakers and commodity traders) regarding crop management, resource allocation, and market price (Lobell et al., 2015). It also helps in cultivating a wider range of crops and in identifying key variables that naturally influence crop production. With data-driven yield predictions, farmers can gain a better understanding of environmental factors affecting the growth of new crops (Pant et al., 2021). Consequently, it is important to understand the interaction between different factors, such as environmental conditions, genotype, and farming practices for improving the prediction accuracy (Mkhabela et al., 2011; Asseng et al., 2013). However, modelling the effect of these variables can be a challenge, particularly the interaction effect between them.

One of the most significant interactions, studied intensively in a variety of fields, including agriculture, medicine, and psychology, is between genotype and environments. This interaction offers crucial insights into how genetic factors and environmental conditions interplay to shape various outcomes (Via and Lande, 1985). In agriculture, it is used for the optimisation of crop yield, sustainability and resilience against stressors like pests, diseases, and climate change (Yan et al., 2000). This interaction is especially important in multi-environment trials (METs), to evaluate how different genotypes perform across a range of environmental conditions (DeLacy et al., 1996). By doing so, researchers can identify crop varieties that are more universally robust or that are specifically tailored to particular environmental conditions (Crossa et al., 2004). However, there are limitations on considering the genotype by environment interaction in isolation (Malosetti et al., 2013; Ebdon and Gauch Jr, 2002). One major limitation is the omission of other key variables, overlooking factors such as soil quality, weather conditions, and even economic factors, which can also impact crop performance (Messina et al., 2011; Cooper et al., 2014). Ignoring these additional variables compromise the generalisability of models and may lead to misleading conclusions. Furthermore, if genotype by environment interaction is not the primary driver of the phenomenon under study, models that focus solely on these interactions may have limited predictive power

(Crossa et al., 2016).

One of the more popular methods proposed for analysing the genotype by environment interaction is the additive main effects and multiplicative interaction (AMMI) model (Gauch Jr, 1988). However, this model is limited in only accommodating these two categorical variables and disregarding other potential influential factors. To address this, the Bayesian Additive Main Effects and Multiplicative Interaction Tensor (BAMMIT) model was proposed Dos Santos et al. (2023). This model extends the AMMI structure by allowing for additional categorical variables and their interactions in a generalised tensor format. As in the AMMI model, the BAMMIT model consists of two primary terms: an additive term for main effects and a multiplicative term for interaction effects. The interaction is captured in both at lower and higher levels. Despite these enhancements, the BAMMIT model still has one main limitation, notably the constraint that all terms must be categorical.

Mixture models, particularly Gaussian Mixture Models (GMM), are widely applied methods for both density estimation and cluster analysis in various fields ranging from computer science to biology (Murphy, 2012; Bishop and Nasrabadi, 2006; Peel and McLachlan, 2000). In agriculture, mixture models have been applied in different contexts, as presented by Mouret et al., 2022; Wu et al., 2021; Bouayad et al., 2021. These models can efficiently approximate an unknown distribution by employing a weighted sum of multiple component distributions and their probabilistic nature allows for soft assignments of data points to clusters, thereby providing a more nuanced understanding of data structure (Fraley and Raftery, 2002). This probabilistic framework also makes mixture models highly adaptable to different forms of data and cluster shapes. They are particularly useful when the data is multidimensional and may be generated from several underlying processes or sub-populations. For example, McLachlan and Peel (2000) and Bishop and Nasrabadi (2006) provide comprehensive insights into the flexibility of mixture models in dealing with high-dimensional data and their ability to capture complex structures. In the context of genetics, Pritchard et al. (2000) highlight how mixture models can be used to identify sub-populations within a given set of genetic data. In machine learning, Hastie et al. (2009) emphasises the strength of

mixture models in handling high-dimensional data with multiple subgroups.

Given the interest in combining the structure of the BAMMIT model with the incorporation of numerical variables through the use of GMM, our study introduces the Clustered Bayesian Additive Main effects and Multiplicative Interaction Tensor (CBAMMIT) model. This new approach allows for the inclusion of continuous variables through clustering, which are still represented categorically in the model. Unlike traditional methods where clustering is performed prior to model inference, our proposed approach aims to carry out both processes simultaneously to achieve better predictions. This grouping process offers several advantages. In addition to improving predictions, the proposed methodology addresses another important aspect: the ability to 'create' new environments. Furthermore, the probabilistic interpretation of Bayesian clustering in this model allows for a nuanced quantification of uncertainty in cluster membership.

The remainder of this paper is organised as follows. Section 5.2 provides a review of Gaussian Mixture Models, the traditional AMMI model, and the BAMMIT model, and introduces our new approach, along with the chosen priors and the inference process. Section 5.3 illustrates the performance of our model in a scenario where simultaneous clustering and inference prove to be more effective than applying the clustering process prior to inference. A simulation study is presented in Section 5.4, followed by an application to real data in Section 5.5. Finally, Section 5.6 offers a discussion and conclusions about the work.

5.2 Methods

We present a synthesis of Gaussian mixture distributions applied to clustering. Then, we review the AMMI and BAMMIT models. Finally, we introduce the new approach, specify the priors, and outline the inferential process.

5.2.1 Bayesian Mixture of Gaussians

Let $\boldsymbol{X} = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_N^\top)^\top$ be a random sample of size N. Assume that $\boldsymbol{x}_i^\top$ is a p-dimensional random vector following a multivariate Normal distribution. We can

model the density of $\boldsymbol{x}_i^\top$ as a mixture of K components, that is

$$f(\boldsymbol{x}_i|\boldsymbol{\theta}_k) \quad = \quad \sum_{k=1}^{K} \pi_k \mathrm{N}_p(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{5.1}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean vector and the covariance matrix associated with the k^{th} component, respectively. The parameters $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^\top$ are the cluster mixing proportions, which satisfies $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$. Equivalently, the mixture model (5.1) can be expressed through latent allocation variables $\boldsymbol{c} = (\boldsymbol{c}_1, ..., \boldsymbol{c}_n)^\top$, where $\boldsymbol{c}_i \in \{1, \ldots K\}$ identifies the cluster that the respective $\boldsymbol{x}_i$ belongs to. Finally, we assume that $\boldsymbol{c}_i \sim \mathrm{Cat}(\boldsymbol{\pi})$ and $\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha})$, where commonly $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K) = 1$.

The choice of K is non-trivial and often depends on the specific application (Fraley and Raftery, 2002). The conventional method for determining the number of components in a mixture model involves fitting models of various complexities and using criteria like Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC) for model selection (Alamichel et al., 2022). As K approaches infinity, the model becomes a non-parametric mixture, also known as an infinite mixture model. While this can provide more flexibility, it can also lead to overfitting and increased computational complexity (Bishop and Nasrabadi, 2006).

In addition to determining the value of K and computational complexity, there are other ongoing issues in working with Gaussian mixture models (or mixture models in general). One of the concerns is identifiability, where multiple sets of parameters can produce the same likelihood, leading to interpretive ambiguity in the model's components (Frühwirth-Schnatter, 2006; Peel and McLachlan, 2000). Another problem is label switching, where the labels of the clusters may change upon model refitting or resampling, although the model itself remains the same (Celeux et al., 2000; Stephens, 2000).

5.2.2 AMMI model

Let y_{ij} be an outcome variable measured at the i^{th} genotype and j^{th} environment, with $i = 1, \ldots, B_1$ and $j = 1, \ldots B_2$. The traditional AMMI model is written as

$$y_{ij} = \mu + b_i^{(1)} + b_j^{(2)} + \sum_{q=1}^{Q} \lambda_q \beta_{iq}^{(1)} \beta_{jq}^{(2)} + \varepsilon_{ij}, \ \varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2), \tag{5.2}$$

where where μ is the overall mean of the outcome variable. The terms $b_i^{(1)}$ and $b_j^{(2)}$ represent the additive effect of the i^{th} genotype and j^{th} environment, respectively. The bilinear term consists of Q terms, with $\lambda_1 > \cdots > \lambda_Q$ as the weights of the interaction. The scores $\beta_{iq}^{(1)}$ and $\beta_{jq}^{(2)}$ represent the effects of genotype and environment in the interaction, respectively, where $\sum_i \beta_{iq}^{(1)} \beta_{iq'}^{(1)} = \sum_j \beta_{jq}^{(2)} \beta_{jq'}^{(2)} = 0$, for $q \neq q'$ and $\sum_i (\beta_{iq}^{(1)})^2 = \sum_j (\beta_{jq}^{(2)})^2 = 1$, $\forall\ i, j$.

Regarding the inference, there are several methods for estimating AMMI model parameters. In the frequentist framework, the additive term of the AMMI model is typically estimated using ordinary least squares, followed by singular value decomposition for the multiplicative terms, as described by Gabriel (1978). In the Bayesian approach, Viele and Srinivasan (2000) used Markov chain Monte Carlo (MCMC) techniques for parameter estimation, ensuring the model's constraints are maintained. Liu (2001) formulated a faster and more stable version of the Gibbs sampler. Further refinements to the Gibbs sampler were made by Crossa et al. (2011) and Perez-Elizalde et al. (2012), who also included statistical inference in biplot visualizations. Josse et al. (2014) tackled the issue of model overparameterisation by defining unconstrained priors and applying post-processing. Most recently, Sarti et al. (2023) employed Bayesian Additive Regression Trees (BART) to effectively capture the interaction term, though with an alternative structure to that of the latent bilinear term.

5.2.3 BAMMIT model

A generalisation of Equation (5.2) was proposed by Dos Santos et al. (2023) to include the effect of factors apart from genotype and environment. The BAMMIT model is defined as

$$y_{ij\ldots v} = \mu + b_i^{(1)} + b_j^{(2)} + \cdots + b_v^{(V)} + \sum_{q=1}^{Q} \lambda_q \left(\beta_{iq}^{(1)} \beta_{jq}^{(2)} \times \cdots \times \beta_{vq}^{(V)} \right) + \varepsilon_{ij\ldots v}, \tag{5.3}$$

where $y_{ij\ldots v}$ is the outcome variable, now observed in V levels. The error term $\varepsilon_{ij\ldots v}$ is normally distributed with a mean of zero and variance σ^2. Each vector $\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}, \ldots, \boldsymbol{b}^{(V)}$ consists of $B_1, B_2, \ldots, B_V$ values, respectively, each of which representing the levels of a factor. The v^{th} predictor is represented by a B_v-dimensional vector of parameters $\boldsymbol{b}^{(v)} = (b_1^{(v)}, \ldots, b_{B_v}^{(v)})^\top$ and the singular values by another B_v-dimensional vector $\boldsymbol{\beta}_q^{(v)} = (\beta_{1q}^{(v)}, \ldots, \beta_{B_v q}^{(v)})^\top$, where $q = 1, \ldots, Q$. By concatenating these column vectors, we obtain the matrix $\boldsymbol{\beta}^{(v)}$ of dimension $B_v \times Q$. The total number of observations N is then defined as $N = \left(\prod_{v=1}^{V} B_v\right)$.

The structure of the BAMMIT model incorporates both additive and multiplicative terms, similar to the AMMI model, but accommodates for the increased complexity introduced by the V predictors. The total number of observations N is determined by the product of the dimensions of all N predictors. Note that the number of observations may increase due to replications without requiring additional parameters. The parameter Q remains fixed, as in traditional AMMI models, signifying the number of multiplicative terms incorporated. Potential additional predictors that can be integrated into the model range from soil type and time to growth stages, among others. Additionally, the model is designed to include all potential interactions, ranging from 2-way to V-way interactions, by employing suitable selection priors and adjusting the hyperparameter Q. This ensures that only meaningful interactions present in the data are incorporated in the model.

5.2.4 CBAMMIT model

We now introduce our new CBAMMIT which aims to capture the behaviour of other factors by representing them as latent categorical structures through their cluster membership. To incorporate the effect of these continuous variables (denoted $\mathbf{x}$) into the BAMMIT model, we rewrite Equation (5.3) as

$$y_i = \mu + b_{c_i^{(1)}}^{(1)} + b_{c_i^{(2)}}^{(2)} + \ldots + b_{c_i^{(V)}}^{(V)} + \sum_{q=1}^{Q} \lambda_q \beta_{q c_i^{(1)}}^{(1)} \times \ldots \times \beta_{q c_i^{(V)}}^{(V)} + \epsilon_i, \tag{5.4}$$

where y_i is the i^{th} response observation, $i = 1, \ldots, N$, and the index $c_i^{(v)}$, previously fixed, becomes random, taking values in the set $\{1, \ldots B_v\}$. This component is now

obtained through GMM, with the following structure:

$$\begin{aligned}
\mathbf{x}^{(v)} &\sim \text{GMM}\left(\boldsymbol{\mu}_{\mathbf{x}^{(v)}}, \boldsymbol{\Sigma}_{\mathbf{x}^{(v)}}, \pi, c_i^{(v)}\right) \\
c_i^{(v)} &\sim \text{Cat}(\pi^{(v)}) \\
\boldsymbol{\mu}_{\mathbf{x}^{(v)}} &\sim \text{N}(0, Z) \\
\boldsymbol{\Sigma}_{\mathbf{x}^{(v)}} &\sim \text{Wishart}_{n_{b^{(v)}}}(R, q) \\
\pi^{(v)} &\sim \text{Dir}(\boldsymbol{\alpha}^{(v)}).
\end{aligned}$$

In this formulation, to each predictor v there are $n_{b^{(v)}}$ associated numerical variables, named $\mathbf{x}^{(v)} = \left[\mathbf{x}_1^{(v)}, \ldots, \mathbf{x}_{n_{b^{(v)}}}^{(v)}\right]^\top$, where each $\mathbf{x}_l^{(v)}$ is a vector of N observations, and $\mathbf{x}^{(v)}$ is a matrix of dimensions $N \times n_{b^{(v)}}$. Model-based clustering is applied to the matrix of continuous data $\mathbf{x}^{(v)}$ to generate the levels of the v^{th} predictor. Importantly these steps are run concomitantly so that uncertainty in the clustering labels and their contribution to the response is accounted for correctly. As in the BAMMIT model, the parameter λ serves as the weight of the interaction in the model. The effect of the v^{th} predictor in the interaction is captured by a B_v-dimensional vector $\boldsymbol{\beta}_q^{(v)}$. The matrix $\boldsymbol{\beta}^{(v)}$ is of dimension $B_v \times Q$, where each column is the vector $\boldsymbol{\beta}_q^{(v)}$.

Note that not all V predictors need to be constructed through clustering. Some effects may already be categorical, obviating the need for grouping. Our interest may lie solely in clustering a specific effect, such as the environment. For example, consider the AMMI model with just two predictors ($V = 2$): genotype and environment. We may assume the genotype variable is already categorical with B_1 levels, while the environment predictor is associated with two numeric variables: air temperature and humidity. These numeric variables are grouped to construct B_2 distinct levels for the environment, each corresponding to a specific cluster. This may be particularly useful when we only have access to climatic, environmental, or genetic continuous variables, rather than simplistic labels that have been applied arbitrarily (such as using location as a stand-in for environment). Our current approach only accounts for clustering of continuous variables, though extensions to to multi-modal model based clustering can be found in the literature (e.g., McParland and Gormley, 2016). These are considerably more complex to

implement and so are left for future work

Regarding the specification of the model's priors, we are guided by the works of Josse et al., 2014; Sarti et al., 2023 and Dos Santos et al. (2023). The remainder of hierarchical model, in addition to those terms given above, is:

$$
\begin{aligned}
y_i &\sim \mathrm{N}\left(\mu + b^{(1)}_{c_i^{(1)}} + b^{(2)}_{c_i^{(2)}} + \cdots + b^{(V)}_{c_i^{(V)}} + \sum_{q=1}^{Q} \lambda_q \beta^{(1)}_{qc_i^{(1)}} \times \cdots \times \beta^{(V)}_{qc_i^{(V)}}\right) \\
\mu &\sim \mathrm{N}(\mu_\mu, \sigma^2_\mu) \\
b^{(v)}_{c_i^{(v)}} &\sim \mathrm{N}(\mu_{b^{(v)}}, \sigma^2_{b^{(v)}}) \\
\lambda_q &\sim \mathrm{N}^+(0, \sigma^2_\lambda) \\
\sigma^{-2} &\sim \mathrm{G}(a_0, a_1) \\
\mu_{b^{(v)}} &\sim \mathrm{N}(0, 1) \\
\sigma_{b^{(v)}} &\sim t^+(0, a_2) \\
\sigma_\lambda &\sim t^+(0, a_3)
\end{aligned}
$$

where N^+, G, and t^+, are the truncated Normal, Gamma, and truncated t-Student distributions, respectively. We assume $\lambda_1 \geq \ldots \lambda_q$ and that all the hyperparameters here are fixed, as well in the GMM structure.

To address model identifiability and include interactions of all orders, we employ a transformation process involving auxiliary variables $\theta_{\beta^{(v)}_{c_i q}}$, generated from a standard $\mathrm{N}(0, 1)$ distribution, and means $\mu_{\boldsymbol{\beta}^{(v)}_q}$, as presented by Dos Santos et al. (2023). This is done for every vector $\boldsymbol{\beta}^{(v)}_q$. The formula incorporates additional parameters M and $\gamma^{(v)}_q$ that act as switches to include or exclude the effects of the v^{th} variable based on the data. These switches are regulated by a Bernoulli prior distribution, guided by a Beta distribution, thus resembling a spike-and-slab structure:

$$
\begin{aligned}
\gamma^{(v)}_q &\sim \text{Bernoulli}(p^v_q), \\
p^{(v)}_q &\sim \text{Beta}(1, 10).
\end{aligned}
$$

This allows for flexibility in choosing which interactions to include or exclude in the BAMMIT model (Dos Santos et al., 2023). Specifically, each parameter $\beta^{(v)}_{qc_i^{(v)}}$

is obtained by

$$\beta^{(v)}_{qc_i^{(v)}} = \left[\left(\theta_{\beta^{(v)}_{qc_i^{(v)}}} - \mu_{\beta_q^{(v)}}\right) + M\gamma_q^{(v)}\right]\left[\left(\sum_i \left(\theta_{\beta^{(v)}_{qc_i^{(v)}}} - \mu_{\beta_q^{(v)}}\right)^2\right)^{1/2} + M\gamma_q^{(v)}\right]^{-1}.$$

5.3 Toy Data Example

In this section, we provide a motivational example to better illustrate the concept behind our proposal, where the clustering process is carried out simultaneously with inference. Consider the simplest version of the Equation (5.4), where there is just one predictor, and consequently, no interaction term:

$$y_i = b_{c_i} + \varepsilon_i, \tag{5.5}$$

where $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$. For simplicity, in this section we drop the superscript of the variable $\boldsymbol{b}$ and of all the parameters related to it. Under this formulation, the hierarchical model is then

$$\begin{aligned}
\mathbf{x} &\sim \mathrm{GMM}\left(\mu_{\mathbf{x}}, \sigma^2_{\mathbf{x}}, c_i, \pi\right)\\
c_i &\sim \mathrm{Cat}(\pi)\\
\pi &\sim \mathrm{Dir}(\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_B) = 1)\\
\mu_{\mathbf{x}} &\sim \mathrm{N}(0, 10^2)\\
\sigma_{\mathbf{x}} &\sim t^+(0, 10^2).
\end{aligned}$$

The predictor b_{c_i} is constructed from a single numeric variable $\boldsymbol{x}$, which is assumed to have B levels after being grouped. Figure 5.3.1 presents a graphical representation of Equation (5.5).

The relationships outlined in the DAG show that y_i is modelled as being affected by the co-variate x and the cluster-specific effects b_i, with σ_y representing the standard deviation of the noise around y_i. The cluster-specific effect b_i is derived from the cluster mean μ_j, which is influenced by the cluster assignments c_j. These assignments are important as they are determined by the GMM clustering process. Additionally, the variability of x and b_i is captured by σ_x and σ_b. We draw the reader's focus in particular to the red arrow which, if included, leads to our modelling approach but, if missing, gives a separable model where the model based

clustering is performed solely on $\boldsymbol{x}$ before the BAMMIT model is run. We refer to this latter two-step approach as C+BAMMIT. When the red arrow is included we name our approach CBAMMIT.

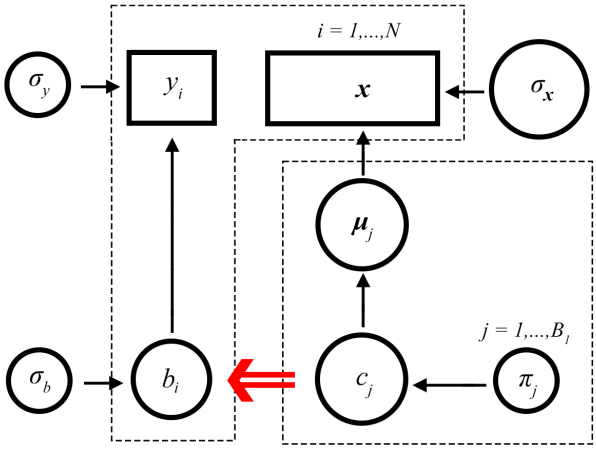


Figure 5.3.1: Directed Acyclic Graph (DAG) for the motivating example following Equation (5.5).

In this toy example, we simulate data to contain two clusters with means $\mu_{\mathbf{x}} = (-2, 2)$ and standard deviation one for each dimension. We set $N = 400$, allocating 300 observations for the training data and 100 for the test data. The clusters are not particularly well-separated, as showed by the histogram in Figure 5.3.2. Although the concept of a cluster may vary between different methods, the configuration in this example presents a challenging scenario for many models that rely heavily on clear cluster separation for accurate predictions.

The CBAMMIT and C+BAMMIT models are fitted assuming the same priors for the model parameters. However, since the latter model consists of two separate processes, the estimates for the levels c_i are obtained as the posterior mode and are then used in the BAMMIT model as the levels for the variable b_{c_i}. We run each model 20 times and calculate the Root Mean Square Error (RMSE) for the predictor b_{c_i} and for the response. The comparative results are displayed in Figure 5.3.3, where box plots illustrate the distribution of RMSE values for the multiple runs of each model. We see that our proposed CBAMMIT outperforms C+BAMMIT,

showing a superior predictive accuracy and robustness in scenarios where clusters are not distinctly separated, particularly for the vector of effects $\boldsymbol{b}$.

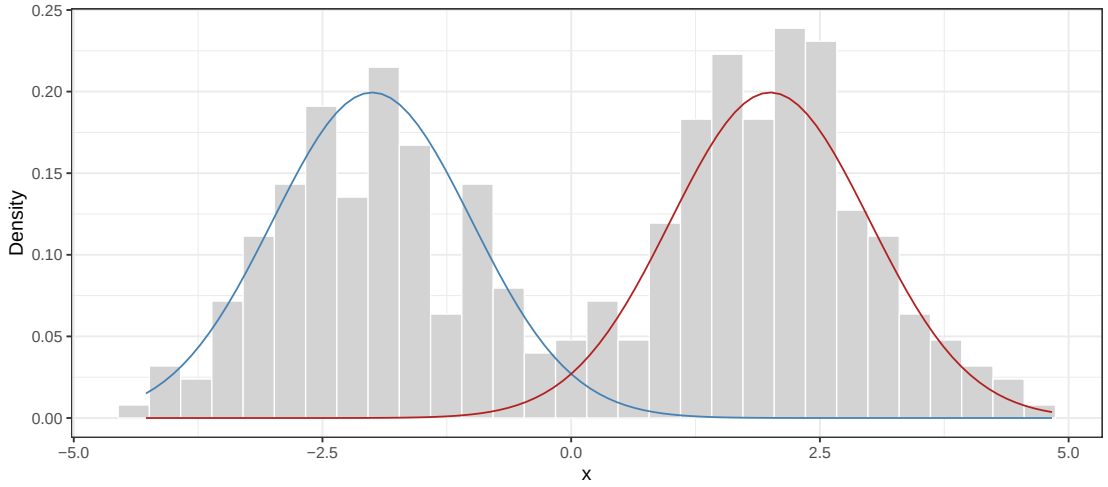


Figure 5.3.2: Histogram of variable x.

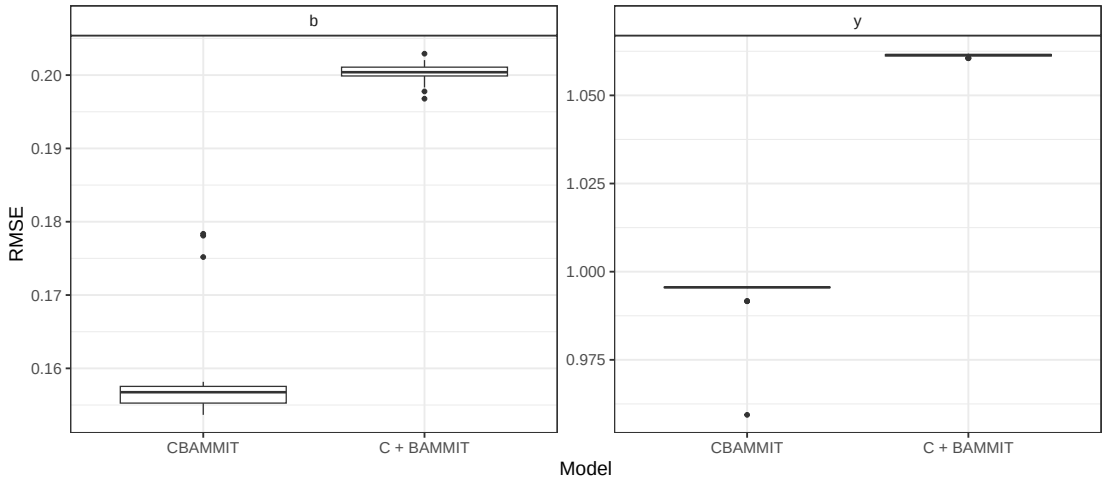


Figure 5.3.3: Box plots of the RMSE of the models BAMMIT and C+BAMMIT on test data for the predictor $\boldsymbol{b}$ and the vector of response $\boldsymbol{y}$.

5.4 Simulation Studies

To evaluate the performance of the new approach, we propose a simulation study. Our main objectives are to observe whether the clusters are estimated correctly and evaluate the prediction performance of the model. We constructed data sets in five different scenarios, presented in Table 5.4.1.

Scenarios	V	N	Variables Clustered	$n_{b^{(v)}}$	B_v^*	$\mu_{\boldsymbol{x}^{(v)}}$
1	2	200	$b^{(2)}$	1	3	{(-2, 5, 10)}
2	3	500	$b^{(2)}$	1	3	{(-2,5, 10)}
3	3	500	$b^{(2)}$	3	3	{(-2,5, 10)}
4	3	500	$b^{(2)}, b^{(3)}$	1,3	3,3	{(-2,5, 10), (-10,0,10)}
5	3	500	$b^{(2)}, b^{(3)}$	1,3	3,10	{(-2,5, 10), U(-50, 50)}

Table 5.4.1: Simulation scenarios settings.

The scenarios were constructed to vary the number of predictors V, the number of numerical variables $n_{b^{(v)}}$ associated with one or more predictors, the corresponding number of clusters for these variables, and consequently, the number of levels B_v^*, where B_v^* is the number of clusters defined in the simulation. We built simulation scenarios considering the following points: observe what happens when clustering is applied on only one of the variables; evaluate when clustering is performed on more than one variable; and finally, evaluate the effect of changing the number of variables used in clustering and the number of groups.

Explicitly, Scenario 1 in Table 5.4.1 serves as a baseline model, representing the AMMI model. In this case, a single numerical variable associated with the predictor $b^{(2)}$ is clustered into three distinct groups with mean values of -2, 5, and 10. Scenarios 2 and 3 build upon Scenario 1 by introducing three additional categorical variables. While $b^{(2)}$ remains a key predictor, the focus of clustering shifts to the number of numerical variables associated with it. Scenarios 4 and 5 introduce further complexity by involving numerical variables associated with both $b^{(2)}$ and $b^{(3)}$ in the clustering process. These scenarios vary in the number of associated numerical variables and their levels. Whenever there is more than one associated

numerical variable, we use the same mean vector $\mu_{\boldsymbol{x}^{(v)}}$ to simulate the data, aiming to maintain consistency in the underlying clustering structure.

In all scenarios, $b^{(1)}$ is kept as a categorical variable with a fixed number of levels, set at 10. The levels of $b^{(3)}$ are also fixed in Scenarios 2 and 3. To evaluate the model's performance, we generated both training and test data sets for each scenario. Specifically, in Scenario 1, we utilised 200 observations for the training set and 100 for the test set. For the remaining scenarios, which have the same total sample size N, we allocate 400 observations for the test set and 100 for the training set.

To fit the model, we run a Markov chain Monte Carlo (MCMC) algorithm through Just Another Gibbs Sampler (JAGS; Plummer et al., 2003) and the R package `R2jags` (Su and Yajima, 2021). We fit the model considering $Q = \{1, 2\}$, with respective $\lambda = \{(10), (8, 10)\}$, in all scenarios, and the number of clusters equal to the number of clusters simulated ($B_v = B_v^*$). We set $\mu_\mu = 100$, $\sigma_\mu^2 = 10$, $a_0 = a_1 = 0.1$, $a_2 = a_3 = 1$, and $M = 10,000$. We use three chains, 4,000 iterations per chain, discarding the first 2,000 as burn-in, and a thinning rate of two. To deal with the label switching problem, we use the package `label.switching` (Papastamoulis, 2015). For all simulation scenarios, we fit the C+BAMMIT model, adhering to the same priors, parameterisation, and settings used in the CBAMMIT model. We compare the two models by calculating the RMSE on both the training and test data. All experiments were implemented in R, on a MacBook Pro with a 1.4 GHz Quad-Core Intel Core i5 and 8 GB of RAM.

5.4.1 Results

We first evaluate whether the model is successfully capturing the effect of the variable $b^{(1)}$ in the additive term. This variable is of particular interest because its levels are fixed and did not undergo the clustering process. In Figure 5.4.1, we present the scatterplots of the true values against the estimated values for each scenario. In this case, we simulate the data with $Q = 1$ and $\lambda = 10$. We observe that the effect of this variable is reasonable captured in all scenarios, with more problems in the scenarios 4 and 5. These deviations could potentially be attributed

to the complexities introduced by the mixture components and the interactions among them.

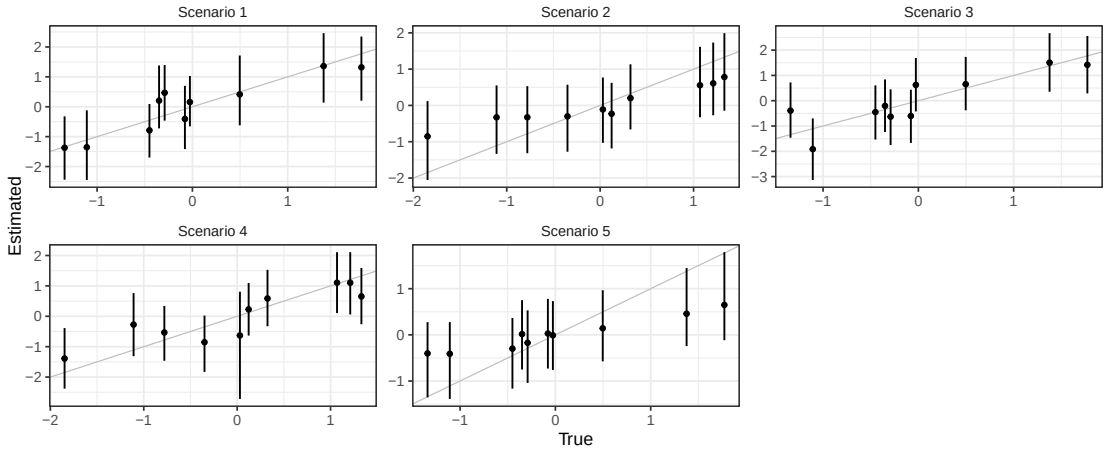


Figure 5.4.1: Scatterplots of true versus estimated additive term $b^{(1)}$, with the respective 95% credible intervals, for all simulation scenarios, setting $Q = 1$ and $\lambda = 10$.

In Figure 5.4.2, we show the estimation of the variable $b^{(2)}$ for all the five simulation scenarios. This predictor is associated with the numerical variables that were clustered into three levels. Despite the increased model complexity in Scenarios 4 and 5, the estimation results for $b^{(2)}$ remained satisfactory. Turning to Figure 5.4.3, we present the scatterplots of true versus estimated values of the variable $b^{(3)}$. Similar to $b^{(2)}$, the results for $b^{(3)}$ were also good across scenarios. To generate both these figures, we once again simulated the data under conditions where $Q = 1$ and $\lambda = 10$.

In Tables 5.4.2 and 5.4.3, we present the RMSE for the interaction term and the response variable, respectively, for both the CBAMMIT and C+BAMMIT models on the test data. The interaction term proved to be the most challenging to predict, showing little difference between the tested methods. This may be influenced by the procedure ensuring model identifiability. Lastly, the prediction of the response variable was better in the CBAMMIT model compared to the C+BAMMIT model. In all scenarios, convergence was measured in terms of $\hat{R}$. Achieving convergence

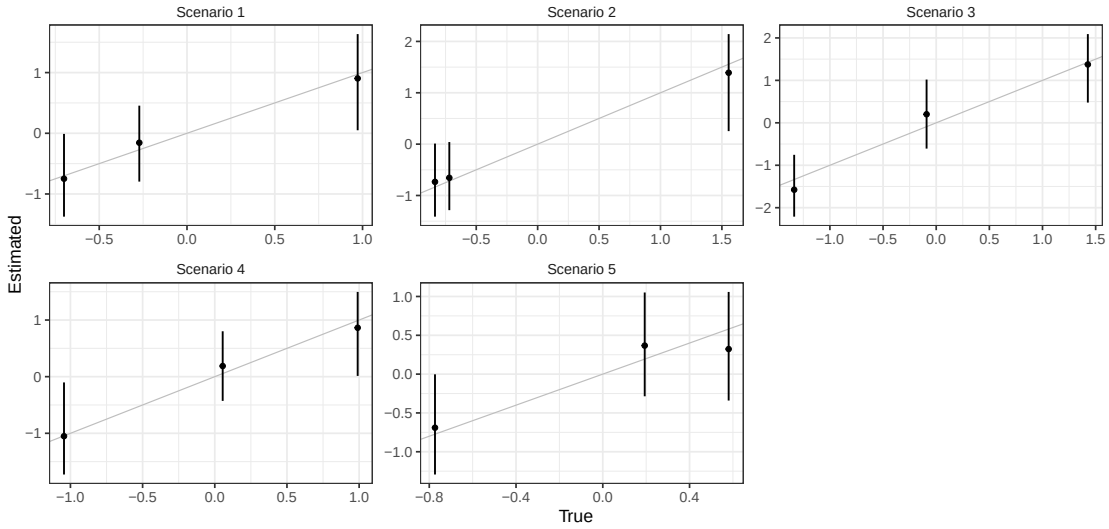


Figure 5.4.2: Scatterplots of true versus estimated additive term $b^{(2)}$, with the respective 95% credible intervals, for all simulation scenarios, setting $Q = 1$ and $\lambda = 10$.

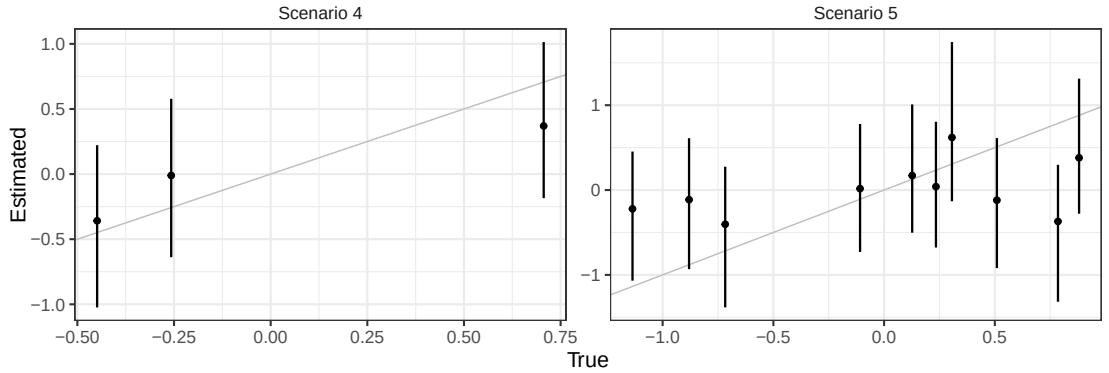


Figure 5.4.3: Scatterplots of true versus estimated additive term $b^{(3)}$, with the respective 95% credible intervals, for simulation scenarios 4 and 5, setting $Q = 1$ and $\lambda = 10$.

($\hat{R} \approx 1$) became more challenging as the number of observations and the value of Q increased. As a result, depending on the value chosen, it was sometimes necessary to increase the number of iterations in the fitting process and to apply

procedures to correct the label-switching problem. Another point to mention is that the created data sets could be imbalanced, in the sense that the response variable was not always measured across all possible combinations of levels. As observed, scenarios 4 and 5 were the most problematic in all evaluated aspects, highlighting the model's difficulty in handling a high number of numerical variables to be clustered.

	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
CBAMMIT	1.12	1.24	1.61	1.94	2.14
C+BAMMIT	1.12	1.23	1.62	1.94	2.14

Table 5.4.2: RMSE for the interaction term in the test data for the five scenarios under study for CBAMMIT and C+BAMMIT.

	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
CBAMMIT	0.98	1.08	1.25	1.80	2.02
C+BAMMIT	1.02	1.18	1.40	1.98	2.24

Table 5.4.3: RMSE for y in the test data for the five scenarios under study for CBAMMIT and C+BAMMIT.

5.5 Application

To demonstrate the applicability of the CBAMMIT model, we consider a historical data set previously analysed in Dos Santos et al. (2023). This data set comprises yield measurements for bread wheat (*Triticum aestivum L.*) in Ireland, expressed in tonnes per hectare (t/ha). The data set, sourced from the Horizon2020 EU InnoVar project[5], was provided by the Irish Department of Agriculture, Food and the Marine. It encompasses various factors: genotype, location, year, and block. The data set is comprehensive, containing 85 genotypes across 17 environments and spanning 10 years (2010-2019) with four blocks, resulting in a total of 6,368 observations. As all combinations at all levels are observed, it offers a robust foundation for our study. To maintain data privacy, all genotypes and locations have been anonymised.

[5] www.h2020innovar.eu

In the study by Dos Santos et al. (2023), the variable 'year' was used as a proxy to represent the unique characteristics of each respective year. In our analysis, we remove the categorical variable year, and then include two additional numerical variables: temperature and rainfall corresponding to each year. These two variables are grouped together and will be responsible for representing the effect of the year. This climate data was obtained from Met Éireann[6], the Irish National Meteorological Service. In Figure 5.5.1, we display these two variables against each other and we can observe that one of the points presented the lowest average temperature and rainfall. We apply a clustering process, using the function `Mclust` from package `mclust` (Scrucca et al., 2016), setting the number of mixture components `G = 2`. The fact that the year 2010 is a group by itself aligns with the results observed by Dos Santos et al. (2023), where the year 2010 had the lowest yield and higher standard deviation range.

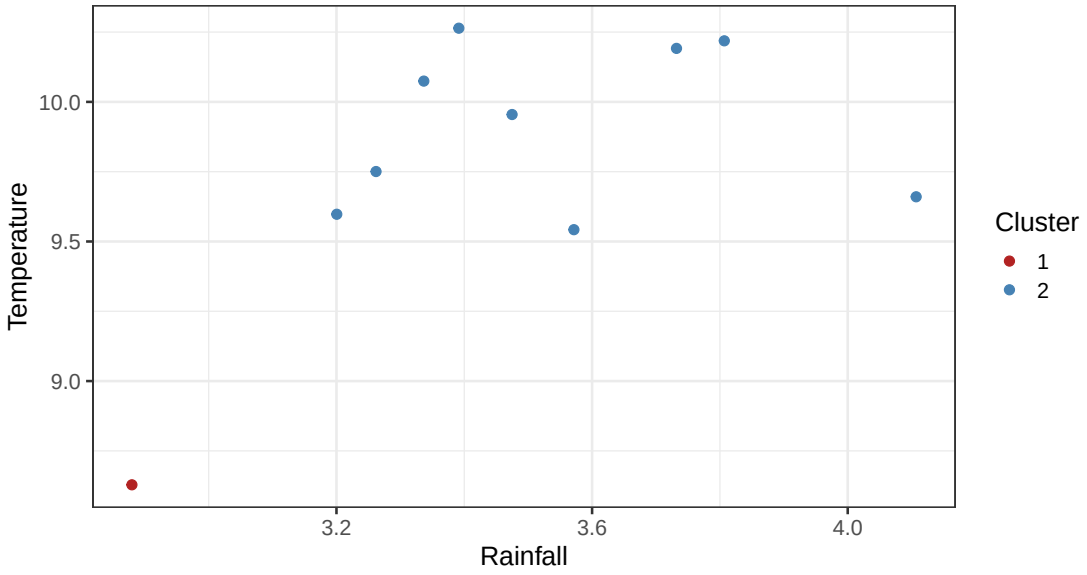


Figure 5.5.1: Variables rainfall against temperature for Ireland in the year 2010-2019.

To fit the CBAMMIT model, we divide the data set into train and test data, in a such way that the data related to the first tree blocks were used for training and

[6] https://www.met.ie

data related to the last block for testing. The CBAMMIT equation is then defined as

$$y_i = \mu + b^{(1)}_{c_i^{(1)}} + b^{(2)}_{c_i^{(2)}} + b^{(3)}_{c_i^{(3)}} + b^{(4)}_{c_i^{(2)}c_i^{(4)}} + \sum_{q=1}^{Q} \lambda_q \beta^{(1)}_{qc_i^{(1)}} \beta^{(2)}_{qc_i^{(2)}} \beta^{(3)}_{qc_i^{(3)}} + \epsilon_i, \quad (5.6)$$

where $b^{(1)}, b^{(2)}, b^{(3)}$ and $b^{(4)}$ are associated to genotypic, environmental, time and block effects, respectively. The indexes $c^{(1)}, c^{(2)}$, and $c^{(4)}$, are fixed, and represent the levels of the variables. The index $c^{(3)}$ are those for which the levels will be obtained through the clustering process.

To fit he model, we set the number of clusters to $B_3 = \{1, 2, 3\}$ and the value of $Q = \{1, 2, 4\}$. To compare the predictive performance, we also fit the C+BAMMIT and just a BAMMIT model, where the third variable is the year. With the last model, we aim to investigate if the inclusion of the variables temperature and rainfall actually improves the yield prediction in this case.

In Table 5.5.1, we present the RMSE of the fitted models. The displayed results correspond to fits with $Q = 4$ and $B_3 = 2$. In terms of predictive performance, the BAMMIT model outperformed its competitors. However, the CBAMMIT model excelled compared to the two-stage process, C+BAMMIT. The poor performance of CBAMMIT relative to its predecessor BAMMIT can be explained by the poor representativeness of the two numerical variables, temperature and rainfall, used to construct the third effect. The fact that the year 2010 had such an atypical behaviour compared to the others, to the extent that it formed a cluster by itself, may have affected the model's outcome.

	CBAMMIT	C + BAMMIT	BAMMIT
RMSE	1.02	1.12	0.59
R^2	0.88	0.87	0.92

Table 5.5.1: RMSE and R^2 for y in the test data for the CBAMMIT, C+BAMMIT and BAMMIT models.

5.6 Discussion

In this paper, we proposed the Clustered Bayesian Additive Main Effects and Multiplicative Interaction Tensor (CBAMMIT) model, which extends the BAM-

MIT approach by incorporating Gaussian Mixture Models. The method aims to incorporate the effect of numerical variables into the model through categorical representations. The levels of each factor, now random, are obtained through a clustering process carried out simultaneously with the estimation process.

Our simulation results demonstrated that the CBAMMIT model offers promising results. Firstly, it enables the inclusion of continuous variables, thereby extending the traditional AMMI model, which was limited to categorical variables. Secondly, even in cases where the clusters were not clearly defined, the model was still capable of providing reasonable predictive results for both the response variable and the additive and interactive terms in the model. However, performance was less robust in these instances compared to when clusters were more distinct. In real-world data application, the model highlighted the importance of setting appropriate values for Q and B_v. The RMSE of the model increased as Q grew and as the correct number of B_v was defined. The importance of the choice of Q is an inherited characteristic from the AMMI and the BAMMIT model, in which in the last one relates to the process of identifying key variables for interaction.

Although the CBAMMIT model shows satisfactory results in terms of accuracy and predictive performance, especially when compared to the two-step process, it does have some significant limitations. The first issue arises from the model's sensitivity to changes in the proposed settings and parameterisation. Additionally, the label-switching problem is not fully resolved, even after applying a process to address it. The computational cost of the method is also high, taking several hours to complete depending on the number of observations, numeric variables, and clusters involved. Another limitation of the CBAMMIT model is that it currently does not allow for the incorporation of both numerical and categorical variables in the clustering process.

Future research should focus on extending the CBAMMIT model to allow for the simultaneous clustering of both categorical and numerical variables. To address the model's sensitivity to parameterisation, the adoption of alternative priors could be explored. Other avenues could involve applying the model to different types of data and further validation through larger and more diverse data sets. It would also

be beneficial to investigate other ways to correct the label-switching problem, as the method presented in this work is solely data-based. Another key consideration would be to employ a process that does not require specifying the number of clusters, such as a Dirichlet process.

CHAPTER 6

Conclusions

In this thesis, we have introduced novel methods used to model and predict crop yield using machine learning techniques. By exploring the complex relationships that determine the phenotypic response, with a focus on genotype by environment interactions, we aim better understanding these relationships and improve the decision making process for researchers, farmers and plant breeders. In this work, we extended the traditional AMMI model to capture more complex interactions, thereby providing a more accurate and comprehensive insight into factors that influence crop yields.

In Chapter 3 we apply variational inference to the AMMI model. We named our proposed approach VAMMI. This approach was employed to reduce the computational cost associated with Bayesian methods, especially when using large data sets. Our key contribution was to meet the model constraints and produce a computationally faster algorithm by designing an variational approximation scheme for inference based on the priors proposed by Josse et al. (2014). The simulation study showed that the VI approach offered comparable predictive performance when compared with traditional the Markov chain Monte Carlo, at the same time that the computational cost was reduced.

In Chapter 4 we presented the BAMMIT model, an extension of the traditional AMMI model. This new model can accommodate many interacting categorical variables, that builds upon the tensor regression technique of Guhaniyogi et al. (2017) and Papadogeorgou et al. (2021). This approach offers a richer understanding of phenotypic effects and provides the opportunity to model other potential interactions such as season, soil, and weather. The model proved to be efficient and robust in our simulation experiments. It was able to effectively capture both lower-order and higher-order interactions with relative ease as well as out-performing other popular machine learning methods, such as random forests and factorial mixed models. This model performed better when applied to real-world data, especially in terms of RMSE measurements and predictive performance and allowed the identification of which variables interact and to ascertain which genotypes, environments and years produced the greatest yield. Additionally, the model offers visualisation tools that facilitate easier interpretation and analysis. One drawback to this approach is computational time. As the number of predictors increases, so does the computational time, which gets compounded with large data sets. However, to combat the computational time, several optimisation strategies are suggested, including variational inference (Blei et al., 2017; Dos Santos et al., 2022).

In Chapter 5, we introduced the Clustered Bayesian Additive Main Effects and Multiplicative Interaction Tensor (CBAMMIT) model. This method extends the BAMMIT approach by integrating Gaussian Mixture Models, allowing the inclusion of numerical variables via categorical representations, where a clustering process is used to determine the factor level. Our CBAMMIT model successfully incorporates continuous variables, and overcomes the limitations of the traditional AMMI model (which could only include categorical variables). The model was able to produce accurate prediction results for the response variable as well as the additive and interaction components in the model, even in situations where the clusters were not well defined. However, easily identifiable clusters resulted in better performance. Despite the promising results, the CBAMMIT approach has some drawbacks. The model's performance is sensitive to changes in settings and parameterisation. Additionally, the label-switching problem, discussed in this

Chapter, still has to be resolved.

6.0.1 Limitations and Improvements

While the models developed in this thesis represent significant advancements in the field, they come with their own set of limitations. In this section, we discuss some of the limitations of our proposed methods and provide improvements for future research.

6.0.1.1 Computational Time

The primary concern with all the proposed models is the significant computational cost. As the models grew in complexity, so did their demand for computational resources. This is especially true for the BAMMIT and CBAMMIT models presented in Chapter 4 and Chapter 5, which frequently takes several hours to run, which could be problematic for time-sensitive projects or when handling large data sets. Future research efforts may concentrate on optimising and simplifying these algorithms in order to tackle these issues. By using parallel computing or optimisation strategies, processing times may be shortened, increasing the adaptability and efficiency of these models for practical use.

6.0.1.2 Model Sensitivity

As discussed in Chapter 5, the CBAMMIT model's sensitivity to initial settings and parameterisation, is a considerate limitation of this approach. This sensitivity can have a substantial impact on the dependability and performance of the model, necessitating the need for a careful analysis of prior selection. For future work, investigating the underlying causes of this sensitivity may be helpful in identifying areas where the model needs to be adjusted for stability. Alternatively, the use of different priors could be explored.

6.0.1.3 Label Switching

The issue of label switching, discussed in Chapter 5, is an avenue for future work. As this work only presents a data-based solution to solve the label switching problem, it would also be helpful to investigate alternative approaches. In addition to

improving the CBAMMIT model's stability through consistent and comprehensible results across several iterations, finding a robust solution to the label switching problem would also increase the model's usefulness in real-world applications.

6.0.2 Future Work

Aside from the aforementioned limitations and avenues for improvements, there are numerous other directions for research to go.

One interesting direction for future research is the models' incorporation of a time dimension. Time-varying phenomena, such as seasonal variations and climatic shifts, are inherent characteristics of environmental factors. The inclusion of a time component in these models, which takes into account these phenomena, may allow for the identification of complex patterns and trends that could otherwise go missed and could greatly improve the precision and insights obtained from the models.

Under development are standalone `R` packages for the VAMMI, BAMMIT, and CBAMMIT models. The aim of these packages is to offer intuitive and effective tools that will enable the wider implementation and distribution of the approaches associated with each model. However, the current implementation of the proposed methods presented in this work are freely available at `https://github.com/Alessandra23` in separate repositories. As a result, interested practitioners can access the techniques and all studies in this thesis are replicable.

From a visualisation perspective, the heatmaps used to display the genotype by environment interactions in Chapter 4, and the VSUP plots (which include the uncertainty associated with Bayesian models) in Chapter 5, only show bivariate interactions. Examining visualisations for higher dimension interactions that are informative and intuitive may be an interesting line of future work. Additionally, incorporating a temporal dimension into the heatmap visualisations, to show how the interactions vary over time, could be a useful avenue for further research.

Finally, one of the main objectives is to apply the methodologies developed in this thesis to the data from the InnoVar project. Given the diversity of the data in this project, the models developed in Chapters 4 and 5 are the most suitable and

promising. In particular, the visualisation proposal presented in Chapter 4 will be extremely useful for understanding the model results for project partners who are not familiar with the statistical technicalities involved.

Bibliography

Adisa, O. M., Botai, J. O., Adeola, A. M., Hassen, A., Botai, C. M., Darkey, D., and Tesfamariam, E. (2019). Application of artificial neural network for predicting maize production in south africa. *Sustainability*, 11(4):1145. 51

Airoldi, E. M., Blei, D., Fienberg, S., and Xing, E. (2008). Mixed membership stochastic blockmodels. *Advances in neural information processing systems*, 21. 33

Alamichel, L., Bystrova, D., Arbel, J., and King, G. K. K. (2022). Bayesian mixture models (in) consistency for the number of clusters. *arXiv preprint arXiv:2210.14201*. 24, 87

Allard, R. W. (1999). *Principles of plant breeding*. John Wiley & Sons. 8

Antle, J. M. and Capalbo, S. M. (2010). Adaptation of agricultural and food systems to climate change: an economic and policy perspective. *Applied Economic Perspectives and Policy*, 32(3):386–416. 2

Archer, E., Park, I. M., and Pillow, J. W. (2013). Bayesian and quasi-bayesian estimators for mutual information from discrete data. *Entropy*, 15(5):1738–1755. 18

Asseng, S., Ewert, F., Rosenzweig, C., Jones, J. W., Hatfield, J. L., Ruane, A. C., Boote, K. J., Thorburn, P. J., Rötter, R. P., Cammarano, D., et al. (2013). Uncertainty in simulating wheat yields under climate change. *Nature climate change*, 3(9):827–832. 2, 84

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. University of London, University College London (United Kingdom). 28

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer. 18, 20, 85, 87

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877. 18, 19, 20, 28, 31, 76, 105

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022. 18

Bouayad, D., Baroth, J., and Dano, C. (2021). Gaussian mixture model based soil classification using multiple cone penetration tests. In *IOP Conference Series: Earth and Environmental Science*, volume 696, page 012034. IOP Publishing. 85

Bouchard, G., Naradowsky, J., Riedel, S., Rocktäschel, T., and Vlachos, A. (2015). Matrix and tensor factorization methods for natural language processing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 16–18. 21

Bouveyron, C. and Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, 71:52–78. 23

Brown, J. and Caligari, P. (2011). *An introduction to plant breeding*. John Wiley & Sons. 8

Carbonetto, P. and Stephens, M. (2012). Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. 18

Carroll, J. D. and Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 35(3):283–319. 52

Celeux, G., Hurn, M., and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970. 24, 87

Chen, T., He, T., Benesty, M., and Khotilovich, V. (2019). Package 'xgboost'. *R version*, 90:1–66. 62

Cichocki, A., Mandic, D., De Lathauwer, L., Zhou, G., Zhao, Q., Caiafa, C., and Phan, H. A. (2015). Tensor decompositions for signal processing applications: From two-way to multiway component analysis. *IEEE signal processing magazine*, 32(2):145–163. 20

Comstock, R. and Moll, R. H. (1963). Genotype-environment interactions. *Statistical genetics and plant breeding*, 982:164–196. 10

Cooper, M., Byth, D., and Hammer, G. (1996). Understanding plant adaptation to achieve systematic applied crop improvement–a fundamental challenge. *Plant adaptation and crop improvement*, pages 5–23. 10

Cooper, M., Messina, C. D., Podlich, D., Totir, L. R., Baumgarten, A., Hausmann, N. J., Wright, D., and Graham, G. (2014). Predicting the future of plant breeding: complementing empirical evaluation with genetic prediction. *Crop and Pasture Science*, 65(4):311–336. 84

Cornelius, P. (1993). Statistical tests and retention of terms in the additive main effects and multiplicative interaction model for cultivar trials. *Crop science*, 33(6):1186–1193. 14, 55

Cornelius, P. L. and Crossa, J. (1999). Prediction assessment of shrinkage estimators of multiplicative models for multi-environment cultivar trials. *Crop Science*, 39(4):998–1009. 30

Correll, M., Moritz, D., and Heer, J. (2018). Value-suppressing uncertainty palettes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–11. 72

Cotes, J. M., Crossa, J., Sanches, A., and Cornelius, P. L. (2006). A bayesian approach for assessing the stability of genotypes. *Crop Science*, 46(6):2654–2665. 27

Crossa, J., de los Campos, G., Maccaferri, M., Tuberosa, R., Burgueño, J., and Pérez-Rodríguez, P. (2016). Extending the marker x environment interaction model for genomic-enabled prediction and genome-wide association analysis in durum wheat. *Crop Science*, 56(5):2193–2209. 85

Crossa, J., Perez-Elizalde, S., Jarquin, D., Cotes, J. M., Viele, K., Liu, G., and Cornelius, P. L. (2011). Bayesian estimation of the additive main effects and multiplicative interaction model. *Crop Science*, 51(4):1458–1469. 13, 28, 30, 54, 88

Crossa, J., Vargas, M., and Joshi, A. K. (2010). Linear, bilinear, and linear-bilinear fixed and mixed models for analyzing genotype x environment interaction in plant breeding and agronomy. *Canadian Journal of Plant Science*, 90(5):561–574. 27, 51

Crossa, J., Yang, R.-C., and Cornelius, P. L. (2004). Studying crossover genotype x environment interaction using linear-bilinear models and mixed models. *Journal of Agricultural, Biological, and Environmental Statistics*, 9:362–380. 15, 84

da Silva, C. P., de Oliveira, L. A., Nuvunga, J. J., Pamplona, A. K. A., and Balestre, M. (2015). A bayesian shrinkage approach for ammi models. *Plos one*, 10(7):e0131414. 14, 55

De Resende, M. D. V. and Thompson, R. (2004). Factor analytic multiplicative mixed models in the analysis of multiple experiments. *Revista de Matemática e Estatística, São Paulo,*, 22(2):31–52. 11

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., and Bodik, R. (2017). Programming with models: writing statistical algorithms for general model structures with nimble. *Journal of Computational and Graphical Statistics*, 26(2):403–413. 42

DeLacy, I., Basford, K., Cooper, M., Bull, J., McLaren, C., et al. (1996). Analysis of multi-environment trials–an historical perspective. *Plant adaptation and crop improvement*, 39124:39–124. 9, 84

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through bayesian sampling. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(2):363–375. 24

dos S. Dias, C. T. and Krzanowski, W. J. (2003). Model selection and cross validation in additive main effect and multiplicative interaction models. *Crop Science*, 43(3):865–873. 14, 55

Dos Santos, A. A., Moral, R. A., Sarti, D. A., and Parnell, A. C. (2022). Variational inference for additive main and multiplicative interaction effects models. *arXiv preprint arXiv:2207.00011*. 76, 105

Dos Santos, A. A., Sarti, D. A., Moral, R. A., and Parnell, A. C. (2023). Bayesian additive main effects and multiplicative interaction models using tensor regression for multi-environmental trials. *arXiv preprint arXiv:2301.03655*. 85, 88, 91, 99, 100

Ebdon, J. and Gauch Jr, H. (2002). Additive main effect and multiplicative interaction analysis of national turfgrass performance trials: I. interpretation of genotype x environment interaction. *Crop science*, 42(2):489–496. 84

Einstein, A. (1916). The basics of general relativity theory. *Annalen der Physik*, (49):769–822. 21

Erenstein, O., Jaleta, M., Mottaleb, K. A., Sonder, K., Donovan, J., and Braun, H.-J. (2022). Global trends in wheat production, consumption and trade. In *Wheat improvement: food security in a changing climate*, pages 47–66. Springer International Publishing Cham. 1

Facelli, J. C. (2011). Chemical shift tensors: Theory and application to molecular structural problems. *Progress in nuclear magnetic resonance spectroscopy*, 58(3-4):176. 52

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230. 24

Finlay, K. and Wilkinson, G. (1963). The analysis of adaptation in a plant-breeding programme. *Australian journal of agricultural research*, 14(6):742–754. 10

Foucteau, V., Denis, J., Gallais, A., Dillmann, C., and Goldringer, I. (2001). Statistical analysis of successive series of experiments in plant breeding: A bayesian approach. *COLLOQUES-INRA*, pages 49–58. 27

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631. 85, 87

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer. 87

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467. 11, 13, 15, 54, 71

Gabriel, K. R. (1978). Least squares approximation of matrices by additive and multiplicative models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(2):186–196. 13, 27, 30, 38, 54, 88

Gabriel, K. R. (2002). Le biplot-outil d'exploration de données multidimensionnelles. *Journal de la société française de statistique*, 143(3-4):5–55. 14, 55

Gaillac, R., Pullumbi, P., and Coudert, F.-X. (2016). Elate: an open-source online application for analysis and visualization of elastic tensors. *Journal of Physics: Condensed Matter*, 28(27):275201. 52

Gauch, H. G. and Moran, D. R. (2019). Ammisoft for ammi analysis with best practices. *BioRxiv*, page 538454. 14, 55

Gauch Jr, H. et al. (1992). *Statistical analysis of regional yield trials: AMMI analysis of factorial designs*. Elsevier Science Publishers. 11, 27

Gauch Jr, H. G. (1988). Model selection and validation for yield trials with interaction. *Biometrics*, pages 705–715. 14, 27, 51, 55, 85

Gauch Jr, H. G. (2006). Statistical analysis of yield trials by ammi and gge. *Crop science*, 46(4):1488–1500. 11, 14, 27, 55

Gauch Jr, H. G. (2013). A simple protocol for ammi analysis of yield trials. *Crop Science*, 53(5):1860–1869. 13, 27, 51

Gauch Jr, H. G., Piepho, H.-P., and Annicchiarico, P. (2008). Statistical analysis of yield trials by ammi and gge: Further considerations. *Crop science*, 48(3):866–889. 14, 27, 51, 55

Gauch Jr, H. G. and Zobel, R. W. (1997). Identifying mega-environments and targeting genotypes. *Crop science*, 37(2):311–326. 27

Gilbert, N. (1963). Non-additive combining abilities. *Genetics Research*, 4(1):65–73. 13, 27, 30

Gill, T., Gill, S. K., Saini, D. K., Chopra, Y., de Koff, J. P., and Sandhu, K. S. (2022). A comprehensive review of high throughput phenotyping and machine learning for plant stress phenotyping. *Phenomics*, 2(3):156–183. 3

Gillberg, J., Marttinen, P., Mamitsuka, H., and Kaski, S. (2019). Modelling g x e with historical weather information improves genomic prediction in new environments. *Bioinformatics*, 35(20):4045–4052. 28

Godfray, H. C. J., Beddington, J. R., Crute, I. R., Haddad, L., Lawrence, D., Muir, J. F., Pretty, J., Robinson, S., Thomas, S. M., and Toulmin, C. (2010). Food security: the challenge of feeding 9 billion people. *science*, 327(5967):812–818. 1

Gollob, H. F. (1968). A statistical model which combines features of factor analytic and analysis of variance techniques. *Psychometrika*, 33(1):73–115. 13, 30

Gormley, I. C., Murphy, T. B., and Raftery, A. E. (2023). Model-based clustering. *Annual Review of Statistics and Its Application*, 10:573–595. 23

Guhaniyogi, R., Qamar, S., and Dunson, D. B. (2017). Bayesian tensor regression. *The Journal of Machine Learning Research*, 18(1):2733–2763. 13, 22, 51, 52, 57, 75, 76, 105

Hadasch, S., Forkman, J., and Piepho, H.-P. (2017). Cross-validation in ammi and gge models: A comparison of methods. *Crop Science*, 57(1):264–274. 14, 55

Hara, P., Piekutowska, M., and Niedbała, G. (2021). Selection of independent variables for crop yield prediction using artificial neural network models with remote sensing data. *Land*, 10(6):609. 51, 67

Harshman, R. A. et al. (1970). Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis. 52

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer. 85

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*. 18

Inglis, A., Parnell, A., and Hurley, C. (2022). Visualizations for bayesian additive regression trees. *arXiv preprint arXiv:2208.08966*. 17, 72

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. 24

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37:183–233. 18

Jørgensen, P. J., Nielsen, S. F., Hinrich, J. L., Schmidt, M. N., Madsen, K. H., and Mørup, M. (2018). Probabilistic parafac2. *arXiv preprint arXiv:1806.08195*. 52

Josse, J., van Eeuwijk, F., Piepho, H.-P., and Denis, J.-B. (2014). Another look at bayesian analysis of ammi models for genotype-environment data. *Journal of Agricultural, Biological, and Environmental Statistics*, 19(2):240–257. 13, 28, 30, 32, 38, 41, 42, 54, 61, 88, 91, 104

Kang, M. (1988). A rank-sum method for selecting high-yielding, stable corn genotypes. *Cereal Research Communications*, 16(1/2):113–115. 8

Kang, M. S., Balzarini, M. G., and Guerra, J. L. (2004). Genotype-by-environment interaction. *Genetic analysis of complex traits using SAS*, pages 69–96. 10

Kejzlar, V. and Hu, J. (2023). Introducing variational inference in statistics and data science curriculum. *The American Statistician*, pages 1–9. xiii, 18

Khoromskaia, V. and Khoromskij, B. N. (2018). *Tensor numerical methods in quantum chemistry*. Walter de Gruyter GmbH & Co KG. 20

Khosla, R. (2010). Precision agriculture: challenges and opportunities in a flat world. In *19th World Congress of Soil Science, Soil Solutions for a Changing World, Brisbane, Australia*. 2

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*. 18

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, 51(3):455–500. 20, 22

Kross, A., Znoj, E., Callegari, D., Kaur, G., Sunohara, M., Lapen, D. R., and McNairn, H. (2020). Using artificial neural networks and remotely sensed data to evaluate the relative importance of variables for prediction of within-field corn and soybean yields. *Remote Sensing*, 12(14):2230. 51

Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22. 62

Lim, Y. J. and Teh, Y. W. (2007). Variational bayesian approach to movie rating prediction. In *Proceedings of KDD cup and workshop*, volume 7, pages 15–21. Citeseer. 33

Liu, G. (2001). *Bayesian computations for general linear-bilinear models*. University of Kentucky. 13, 54, 88

Liu, Y., Liu, J., Long, Z., and Zhu, C. (2022). *Tensor computation for data analysis*. Springer. xiii, 21, 22

Lo, A. Y. (1984). On a class of bayesian nonparametric estimates: I. density estimates. *The annals of statistics*, pages 351–357. 24

Lobell, D. B. and Asseng, S. (2017). Comparing estimates of climate change impacts from process-based and statistical crop models. *Environmental Research Letters*, 12(1):015001. 2

Lobell, D. B. and Gourdji, S. M. (2012). The influence of climate change on global crop productivity. *Plant physiology*, 160(4):1686–1697. 1

Lobell, D. B., Thau, D., Seifert, C., Engle, E., and Little, B. (2015). A scalable satellite-based crop yield mapper. *Remote Sensing of Environment*, 164:324–333. 2, 84

Malik, W., Hadasch, S., Forkman, J., and Piepho, H.-P. (2018). Nonparametric resampling methods for testing multiplicative terms in ammi and gge models for multienvironment trials. *Crop Science*, 58(2):752–761. 14, 55

Malik, W. A., Forkman, J., and Piepho, H.-P. (2019). Testing multiplicative terms in ammi and gge models for multienvironment trials with replicates. *Theoretical and Applied Genetics*, 132(7):2087–2096. 14, 55

Malosetti, M., Ribaut, J.-M., and van Eeuwijk, F. A. (2013). The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. *Frontiers in physiology*, 4:44. 10, 11, 84

McLachlan, G. J., Lee, S. X., and Rathnayake, S. I. (2019). Finite mixture models. *Annual review of statistics and its application*, 6:355–378. 23

McLachlan, G. J. and Peel, D. (2000). Mixtures of factor analyzers. In *Proceedings of the seventeenth international conference on machine learning*, pages 599–606. 85

McLaren, C. and Chaudhary, R. (1994). Use of additive main effects and multiplicative interaction models to analyze multilocation yield trials. *Philippine Journal of Crop Science*. 27

McNicholas, P. D. (2016). Model-based clustering. *Journal of Classification*, 33:331–373. 23

McParland, D. and Gormley, I. C. (2016). Model based clustering for mixed data: clustmd. *Advances in Data Analysis and Classification*, 10(2):155–169. 90

Mendes, C. T. E., de Oliveira, L. A., da Silva, A. Q., da Silva, C. P., dos Santos, P. M., and Sáfadi, T. (2020). Comparing frequentist and bayesian approaches in ammi analysis in a simulated scenario. 30

Messina, C. D., Podlich, D., Dong, Z., Samples, M., and Cooper, M. (2011). Yield–trait performance landscapes: from theory to application in breeding maize for drought tolerance. *Journal of experimental botany*, 62(3):855–868. 84

Miller, J. W. and Carter, S. L. (2020). Inference in generalized bilinear models. *arXiv preprint arXiv:2010.04896*. 13

Mkhabela, M., Bullock, P., Raj, S., Wang, S., and Yang, Y. (2011). Crop yield forecasting on the canadian prairies using modis ndvi data. *Agricultural and Forest Meteorology*, 151(3):385–393. 2, 84

Montesinos-López, O. A., Montesinos-López, A., Crossa, J., Montesinos-López, J. C., Luna-Vázquez, F. J., Salinas-Ruiz, J., Herrera-Morales, J. R., and Buenrostro-Mariscal, R. (2017). A variational bayes genomic-enabled prediction model with genotype x environment interaction. *G3: Genes, Genomes, Genetics*, 7(6):1833–1853. 28, 32

Morningstar, W., Vikram, S., Ham, C., Gallagher, A., and Dillon, J. (2021). Automatic differentiation variational inference with mixtures. In *International Conference on Artificial Intelligence and Statistics*, pages 3250–3258. PMLR. 42

Mørup, M. (2011). Applications of tensor (multiway array) factorizations and decompositions in data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):24–40. 52

Mouret, F., Albughdadi, M., Duthoit, S., Kouamé, D., Rieu, G., and Tourneret, J.-Y. (2022). Reconstruction of sentinel-2 derived time series using robust gaussian mixture models—application to the detection of anomalous crop development. *Computers and Electronics in Agriculture*, 198:106983. 85

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press. 85

Murugiah, S. (2010). Bayesian nonparametric clustering based on dirichletprocesses. 23

Nedelkoski, S., Bogojeski, M., and Kao, O. (2020). Learning more expressive joint distributions in multimodal variational methods. In *Machine Learning, Optimization, and Data Science: 6th International Conference, LOD 2020, Siena, Italy, July 19–23, 2020, Revised Selected Papers, Part I 6*, pages 137–149. Springer. 42

Olivoto, T. and Lúcio, A. D. (2020). metan: An r package for multi-environment trial analysis. *Methods in Ecology and Evolution*, 11(6):783–789. 15

Omer, S. and Singh, M. (2017). Comparing bayesian and frequentist approaches for gge bi-plot analysis in multi-environment trials in sorghum. *Eur Exp Biol*, 7(6):40. 30

Ormerod, J. T. and Wand, M. P. (2010). Explaining variational approximations. *The American Statistician*, 64(2):140–153. 28, 31

Panagakis, Y., Kossaifi, J., Chrysos, G. G., Oldfield, J., Nicolaou, M. A., Anandkumar, A., and Zafeiriou, S. (2021). Tensor methods in computer vision and deep learning. *Proceedings of the IEEE*, 109(5):863–890. 21

Pant, J., Pant, R., Singh, M. K., Singh, D. P., and Pant, H. (2021). Analysis of agricultural crop yield prediction using statistical techniques of machine learning. *Materials Today: Proceedings*, 46:10922–10926. 84

Papadogeorgou, G., Zhang, Z., and Dunson, D. B. (2021). Soft tensor regression. *J. Mach. Learn. Res.*, 22:219–1. 52, 75, 76, 105

Papastamoulis, P. (2015). label. switching: An r package for dealing with the label switching problem in mcmc outputs. *arXiv preprint arXiv:1503.02271*. 96

Papastamoulis, P. (2016). label.switching: An r package for dealing with the label switching problem in mcmc outputs. *Journal of Statistical Software*, 069(c01). 24

Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and computing*, 10:339–348. 85, 87

Perez-Elizalde, S., Jarquin, D., and Crossa, J. (2012). A general bayesian estimation method of linear–bilinear models applied to plant breeding trials with genotype x environment interaction. *Journal of agricultural, biological, and environmental statistics*, 17(1):15–37. 13, 14, 30, 54, 55, 57, 88

Peyrat, J.-M., Sermesant, M., Pennec, X., Delingette, H., Xu, C., McVeigh, E. R., and Ayache, N. (2007). A computational framework for the statistical analysis of cardiac diffusion tensors: application to a small database of canine hearts. *IEEE transactions on medical imaging*, 26(11):1500–1514. 52

Piepho, H., Möhring, J., Melchinger, A., and Büchse, A. (2008). Blup for phenotypic selection in plant breeding and variety testing. *Euphytica*, 161(1-2):209–228. 10, 11

Plummer, M. et al. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. 35, 61, 96

Poland, J. A., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., Dreisigacker, S., Crossa, J., Sánchez-Villeda, H., Sorrells, M., et al. (2012). Genomic selection in wheat breeding using genotyping-by-sequencing. *Plant Genome*, 5(3):103–113. 27

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959. 85

Rasmussen, C. (1999). The infinite gaussian mixture model. *Advances in neural information processing systems*, 12. 24

Richardson, S. and Green, P. J. (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 59(4):731–792. 24

Robert, C. P., Casella, G., and Casella, G. (1999). *Monte Carlo statistical methods*, volume 2. Springer. 4

Robertson, P. K. and O'Callaghan, J. F. (1986). The generation of color sequences for univariate and bivariate mapping. *IEEE Computer Graphics and Applications*, 6(2):24–32. 72

Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A. C., Müller, C., Arneth, A., Boote, K. J., Folberth, C., Glotter, M., Khabarov, N., et al. (2014). Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison. *Proceedings of the national academy of sciences*, 111(9):3268–3273. 1, 2

Rossi, S., Michiardi, P., and Filippone, M. (2019). Good initializations of variational bayes for deep models. In *International Conference on Machine Learning*, pages 5487–5497. PMLR. 33

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., and Wagenmakers, E.-J. (2017). Bayesian analysis of factorial designs. *Psychological Methods*, 22(2):304. 62

Sarti, D. A., Prado, E. B., Inglis, A., dos Santos, A. A. L., Hurley, C., de Andrade Moral, R., and Parnell, A. (2021). Bayesian additive regression trees for genotype by environment interaction models. *bioRxiv*. 14, 38

Sarti, D. A., Prado, E. B., Inglis, A. N., Dos Santos, A. A., Hurley, C. B., Moral, R. A., and Parnell, A. C. (2023). Bayesian additive regression trees for genotype by environment interaction models. *The Annals of Applied Statistics*, 17(3):1936–1957. 17, 30, 34, 54, 55, 61, 62, 67, 71, 72, 73, 88, 91

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):289–317. 24, 100

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809. 24, 87

Su, Y.-S. and Yajima, M. (2012). R2jags: A package for running jags from r. *R package version 0.03-08, URL http://CRAN. R-project. org/package= R2jags*. 35

Su, Y.-S. and Yajima, M. (2021). R2jags: Using r to run 'jags'. r package version 0.6-1; 2020. *URL https://CRAN. R-project. org/package= R2jags*. 61, 96

Team, R. C. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 35

Theobald, C. M., Talbot, M., and Nabugoomu, F. (2002). A bayesian approach to regional and local-area prediction from crop variety trials. *Journal of Agricultural, Biological, and Environmental Statistics*, 7(3):403–419. 27

Tilman, D., Balzer, C., Hill, J., and Befort, B. L. (2011). Global food demand and the sustainable intensification of agriculture. *Proceedings of the national academy of sciences*, 108(50):20260–20264. 1

Titsias, M. (2009). Variational learning of inducing variables in sparse gaussian processes. In *Artificial intelligence and statistics*, pages 567–574. PMLR. 18

Tucker, L. R. (1963). Implications of factor analysis of three-way matrices for measurement of change. *Problems in measuring change*, 15(122-137):3. 52

UNICEF et al. (2020). The state of food security and nutrition in the world 2020. 1

UNICEF et al. (2023). The state of food security and nutrition in the world 2023. 1

Van Eeuwijk, F. (1995). Linear and bilinear models for the analysis of multi-environment trials: I. an inventory of models. *Euphytica*, 84(1):1–7. 27

Vargas, M., Crossa, J., Sayre, K., Reynolds, M., Ramírez, M. E., and Talbot, M. (1998). Interpreting genotype x environment interaction in wheat by partial least squares regression. *Crop science*, 38(3):679–689. 10

Via, S. and Lande, R. (1985). Genotype-environment interaction and the evolution of phenotypic plasticity. *Evolution*, 39(3):505–522. 84

Viele, K. and Srinivasan, C. (2000). Parsimonious estimation of multiplicative interaction in analysis of variance using kullback–leibler information. *Journal of statistical planning and inference*, 84(1-2):201–219. 13, 27, 54, 88

Von Neumann, J. (2018). *Mathematical foundations of quantum mechanics: New edition*, volume 53. Princeton university press. 21

Wainwright, M. J., Jordan, M. I., et al. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305. 18

Wu, W., Wu, X., Zhang, Y. Y., and Leatham, D. (2021). Gaussian process modeling of nonstationary crop yield distributions with applications to crop insurance. *Agricultural Finance Review*, 81(5):767–783. 85

Yan, W., Hunt, L. A., Sheng, Q., and Szlavnics, Z. (2000). Cultivar evaluation and mega-environment investigation based on the gge biplot. *Crop science*, 40(3):597–605. 11, 15, 27, 84

Yan, W. and Kang, M. S. (2002). *GGE biplot analysis: A graphical tool for breeders, geneticists, and agronomists*. CRC press. 9

Yan, W., Kang, M. S., Ma, B., Woods, S., and Cornelius, P. L. (2007). Gge biplot vs. ammi analysis of genotype-by-environment data. *Crop science*, 47(2):643–653. 11, 13, 15

Yan, W. and Rajcan, I. (2002). Biplot analysis of test sites and trait relations of soybean in ontario. *Crop science*, 42(1):11–20. 8, 27

Yan, W. and Tinker, N. A. (2006). Biplot analysis of multi-environment trial data: Principles and applications. *Canadian journal of plant science*, 86(3):623–645. 15

Yang, R.-C., Crossa, J., Cornelius, P. L., and Burgueño, J. (2009). Biplot analysis of genotype x environment interaction: Proceed with caution. *Crop Science*, 49(5):1564–1576. 15

Yates, F. and Cochran, W. G. (1938). The analysis of groups of experiments. *The Journal of Agricultural Science*, 28(4):556–580. 10

Yue, H., Gauch, H. G., Wei, J., Xie, J., Chen, S., Peng, H., Bu, J., and Jiang, X. (2022). Genotype by environment interaction analysis for grain yield and yield components of summer maize hybrids across the huanghuaihai region in china. *Agriculture*, 12(5):602. 14, 55

Zhang, A. R., Luo, Y., Raskutti, G., and Yuan, M. (2020). Islet: Fast and optimal low-rank tensor regression via importance sketching. *SIAM journal on mathematics of data science*, 2(2):444–479. 76

Zhang, N., Wang, M., and Wang, N. (2002). Precision agriculture—a worldwide overview. *Computers and electronics in agriculture*, 36(2-3):113–132. 2

Zhou, H., Li, L., and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552. 21