Bayesian Additive Regression Trees for Genotype by Environment Interaction Models

Danilo A. Sarti[1,*,✉], **Estevão B. Prado**[1,*], **Alan N. Inglis**[1], **Antônia A. L. dos Santos**[1], **Catherine B. Hurley**[1], **Rafael A. Moral**[1], **and Andrew C. Parnell**[1]

[1]Hamilton Institute, Department of Mathematics and Statistics, National University of Ireland Maynooth, Ireland

We propose a new class of models for the estimation of Genotype by Environment (GxE) interactions in plant-based genetics. Our approach, named AMBARTI, uses semi-parametric Bayesian Additive Regression Trees to accurately capture marginal genotypic and environment effects along with their interaction in a fully Bayesian model. We demonstrate that our approach is competitive or superior to the traditional AMMI models widely used in the literature via both simulation and a real world data set. Furthermore, we introduce new types of visualisation to properly assess both the marginal and interactive predictions from the model. An R package that implements our approach is available at `https://github.com/ebprado/ambarti`.

Correspondence: ***daniloasarti@gmail.com***

Introduction

The interaction between genotypes and environments (GxE) is a key parameter in plant breeding (1). Poor understanding of GxE can lead to sub-optimal selection of new genotypes and inbred lines. Understanding the GxE interactions is crucial for germplasm management, having strong genetic and economic impacts on seed production and crop yield (2). Many models have been proposed for studying GxE in the context of Multi-Environmental Experiments (METs) (3). One special case is the Additive Main Effects Multiplicative Interactions Model (AMMI) (4).

The classical AMMI models combine features of Analysis of Variance (ANOVA) with a bilinear term to represent GxE interactions. In addition, AMMI models allow for estimation of main effects of genotypes and environments, and the decomposition of the interaction through a bilinear term. Many extensions to the AMMI models have been proposed, including Robust AMMI (5) and Weighted AMMI (3).

In this paper, we extend the Bayesian AMMI model of Josse et al. (6) to allow for richer GxE interactions, and similarly sidestep the model choice complexity term present in all AMMI-type approaches. We achieve this goal by including a new variant of the Bayesian Additive Regression Trees (BART) (7), which we term 'double-grow' BART. The new proposed method, named AMBARTI, provides a fully Bayesian joint model, where the 'double-grow' BART component is solely responsible for GxE interactions.

BART is a non parametric Bayesian algorithm that generates a set of trees and uses random splits in the range of explanatory variables to produce predictions for a univariate response. Given its flexibility to deal with non linear structures and complex interactions terms, the use of BART and its extensions has increased with applications in many areas including proteomic studies (8), hospital performance evaluation (9), prediction of risk of credit scores (10) among many others.

We compare our newly proposed AMBARTI model with the traditional AMMI approaches, and we show that its performance is superior (judged on out of sample error) in both simulated and real-world example data. The real dataset we use is taken from the Value of Cultivation and Usage (VCU) experiments of the Irish Department of Agriculture, which were conducted in the years between 2010 and 2019. Furthermore, the output of AMBARTI leads us to suggest several new forms of visualisation that we believe are easier to interpret for non-specialists.

The paper is structured as follows. In the next section, we describe the framework used to collect evidence from METs, including the classic genetic equation to describe the relationship between phenotypes, genotypes, and environments. We also outline the formulation of the AMMI model in its classic form. After, we briefly describe the standard Bayesian Additive Regression Trees model and the structure of our novel AMBARTI approach. Following we present the main results from the simulation experiments and real datasets, respectively. Finally, we conclude with a discussion and outline further opportunities.

Methods

GxE interactions and MET. The phenotypic expression of a genetic character can be usually decomposed in terms of genetic factors, environmental factors, and the interactions between them as shown in Equation 1:

$$p = g + e + (ge), \qquad \textbf{(1)}$$

where p is the phenotypic response, g is the genetic factor, e is the environmental factor, and (ge) is the interaction between genotypes and environments. The last term is necessary due to the different response of genotypes across different environments. The presence of GxE interactions implies that each genotype may have different phenotypic responses across a set of environments. If we produce a rank that orders the performance of each genotype into each environment, we will notice that the order of the best to worst genotype is different

[*] Joint first authors.

across the environments. The presence of GxE interactions is known to be capable of having large effects on the phenotypic response (11, 12).

The (ge) terms can be estimated in a MET design, where several environments and genotypes are evaluated for a given phenotype (13). In plant breeding, the need for METs is constant given the fact that the germplasm generates new genotypes every year and the pressure of diseases and other factors are dynamic. Such experiments require a complex set of logistical activities, leading to high costs of implementation. These trials thus have strong regulatory appeal in the seed and biotech industries around the world (2).

Reliable information about GxE can help breeders make decisions on cultivar recommendations. In this sense, models for the study of GxE need to be able to answer questions such as which genotypes can perform well across a set of environments and which are specifically recommended for a given environment. The answers to these questions are crucial both to broad breeding strategies, i.e., to obtain one or more genotypes that perform well in a set of environments, and to target breeding, where we determine the best genotype for a given environment (3).

Traditional AMMI models. A simple statistical linear model can be used to model data from METs. The model can be written as in Equation 2:

$$y_{ij} = \mu + g_i + e_j + (ge)_{ij} + \epsilon_{ij}, \tag{2}$$

where $i = 1,\dots,I$, $j = 1,\dots,J$, g_i is the effect of genotype i, e_j is the effect of environment j, and $(ge)_{ij}$ represents the interaction between genotype i and environment j.

In the specification of the Equation 2, the term (ge) can be thought of as representing a decomposition of the residual from a more basic linear model. In this sense, (14) and (4) proposed a method to decompose the residual term as a sum of multiplicative factors that includes the (ge) term. This yields the decomposition:

$$(ge)_{ij} = \sum_{q=1}^{Q} \lambda_q \gamma_{iq} \delta_{jq}, \tag{3}$$

where Q is the number of components to be considered in the analysis, λ_q is the strength of the interaction of component q, γ_{iq} represents the importance of genotype i in component q, and δ_{jq} represents the importance of environment j in component q; see Appendix B for the restrictions imposed on γ_{iq}, δ_{jq} and λ_q to make the model identifiable. Hence, the complete AMMI model is present in Equation 4.

$$y_{ij} = \mu + g_i + e_j + \sum_{q=1}^{Q} \lambda_q \gamma_{iq} \delta_{jq} + \epsilon_{ij}. \tag{4}$$

The interaction terms in 4 are estimated by a Singular Value Decomposition (SVD) of the matrix of means of genotypes into environments *(GE)*. In this sense, λ_q is the q-th eigenvalue of the matrix *(GE)*, γ_{ik} is the *i*-th element of the left singular vector and δ_{jk} is *j*-th element of the right singular vector obtained in the SVD (15). In practice, the classical AMMI model can be run in R using the package `agricolae` (16) or via functions programmed by the user as in (17).

The protocol for estimation of the terms in a standard AMMI model is given by (18). This involves the following steps:

1. Obtain the grand mean and principal effects of the genotypes and environments using ANOVA with two factors based on a matrix of means containing the means of each genotype into each environment.

2. Obtain the residuals from the model above that will comprise the interaction matrix, where each row is an environment and each column a genotype.

3. Choose an appropriate value for the number of components Q.

4. Form the multiplicative terms that represent the reduced-dimension interactions via an SVD of the matrix of interaction residuals.

The rank of the matrix (GE) is assumed to be $r = min(I-1, J-1)$. Thus, the number of components Q may vary from $1,\dots,r$. The term $min(I-1, J-1)$ establishes the minimum number of non zero eigenvectors to be obtained in the SVD. Taking $Q = r$, the AMMI model would capture all the variance related to the interaction, and it would result in overfitting. This problem is ameliorated by using a limited number of components Q. The number of Q is related to the amount of total variability captured by the principal components and, in general, is recommended to use a number of PCs that captures at least 80% of the total variability. Usually, the value of Q varies from 1 to 3.

AMMI models have been extensively used for evaluation of phenotype performance of cultivars. (19) used AMMI models to assess the performance of wheat germplasm from International Maize and Wheat Improvement Center. (20) used AMMI models to explore Quantitative Trait Loci (QTL) related to adaptation in Wheat. (21) used AMMI to study GxE for Wheat in the context of drought and normal conditions. (22) used AMMI to evaluate the impact of environmental conditions in the stability of winter Wheat. (23) used AMMI to study the stability of early Maize genotypes in Africa. (24) evaluated the performance of experimental maize hybrids using AMMI models, and (3) studied the performance of the AMMI model in the context of simulated data. The applications of AMMI models can also be found in several other species including: a) rice (25), b) barley (25–28), and c) sugarcane (29).

Tree-based methods and BART. Introduced by (7), BART is a Bayesian model that uses a sum of trees to approximate a univariate response. In BART, each tree works as a weak learner that yields a small contribution to the final prediction. Based on a design matrix $\mathbf{X}$, BART is able to capture interactions and non-linear relations. The BART model can be written as

$$y_i | \mathbf{x}_i, \mathcal{M}_t, \mathcal{T}_t, \sigma^2 \sim \mathrm{N}\left(\sum_{t=1}^{T} h(\mathbf{x}_i, \mathcal{M}_t, \mathcal{T}_t), \sigma^2 \right), \; i = 1,\dots,n,$$

where $\mathbf{x}_i$ is the i-th row of the design matrix $\mathbf{X}$, $\mathcal{M}_t$ denotes the set of terminal node parameters of tree t, $\mathcal{T}_t$ is the set of binary splitting rules that define the tree t, and $h(\cdot) = \mu_{t\ell}$ is a function that assigns the predicted values $\mu_{t\ell} \in \mathcal{M}_t$ based on the design matrix $\mathbf{X}$ and tree structure $\mathcal{T}_t$. The number of trees T can be chosen so that non-linear and interaction effects are properly estimated, but it can also be selected through cross-validation; Chipman et al. (7) recommends $T = 200$ as a default.

Unlike other tree-based methods where a loss function is optimised to grow the trees, in BART the trees are grown using Markov Chain Monte Carlo (MCMC) iterations (30, 31). The trees are either accepted or rejected via a Metropolis-Hastings step. In addition, the trees can be modified by four moves: grow, prune, change or swap. In the grow move, a terminal node is randomly selected and then two children nodes are created below it. When pruning, a parent of two terminal nodes is selected at random and its children nodes are removed. During the change process, a parent of two terminal nodes is randomly picked and its splitting rule (covariate and split point) are redefined. In the swap move, a pair of parents of terminal nodes is chosen at random and their splitting rules are swapped. It is important to highlight that in all moves the splitting rule is defined by randomly selecting one covariate and one split point.

As a fully Bayesian model, BART assumes prior distributions on all quantities of interest. First, the node-level parameters $\mu_{t\ell}$ are assumed to be i.i.d $\mathrm{N}(0, \sigma^2_\mu)$, where $\sigma_\mu = 0.5/k\sqrt{T}$ and $1 \leq k \leq 3$. Second, the sample variance σ^2 is assumed to be distributed as $\mathrm{IG}(\nu/2, \nu\lambda/2)$, where $\mathrm{IG}(\cdot)$ denotes an Inverse Gamma distribution. Third, to control how shallow/deep a tree may be, each non-terminal node has a prior probability of $\alpha(1+d)^\beta$ of being observed, where $\alpha \in (0,1)$, $\beta \geq 0$, and d corresponds to the depth of the node; (7) recommends $\alpha = 0.95$ and $\beta = 2$ as default values. These hyperparameter values tend to select trees which are not too deep so as to avoid over-fitting.

Finally, the structure of the BART model for a continuous response can be summarised as follows. First, all $\mathcal{T}_t$ are initialised as stumps. Then, each tree is modified, one at a time, using one of the four moves previously described (grow, prune, change or swap). Next, the newly proposed tree $\mathcal{T}_t^*$ is compared to its previous version $\mathcal{T}_t$ via a Metropolis-Hastings step taking into account the partial residuals $r_t = \mathbf{y} - \sum_{k \neq t}^{T} h(\mathbf{X}, \mathcal{M}_k, \mathcal{T}_k)$ and the structure/depth of $\mathcal{T}_t$ and $\mathcal{T}_t^*$. After that, the predicted values for each terminal node ℓ of the tree t are generated and then σ^2 is updated. For a binary outcome, the idea of data augmentation (32) can be used; see (10) and (33) for more details.

Due to its flexibility and excellent performance on regression and classification problems, BART has been applied and extended to credit modelling (34), survival analysis (35), proteomic biomarker analysis (36), polychotomous response (37) and large datasets (8, 38, 39). More recently, works exploring the theoretical aspects of BART have been developed by Linero and Yang (39), Ročková and van der Pas (40), Ročková and Saha (41). In practice, there are many R packages (42) to fit BART, such as `BartMachine` (43), `BART` (44) and `dbarts` (7).

AMBARTI. To insert the BART model inside an AMMI approach, we make some fundamental changes to the way the trees are grown and structured. As a first step, we can write the sum of trees inside the Bayesian version of the AMMI model

$$y_{ij} | \mathbf{x}_{ij}, \Theta \sim \mathrm{N}\left(\mu + g_i + e_j + \sum_{t=1}^{T} h(\mathbf{x}_{ij}, \mathcal{M}_t, \mathcal{T}_t), \sigma^2\right), \tag{5}$$

where y_{ij} denotes the response variable for genotype i and environment j, $\Theta = (\mu, g_i, e_j, \mathcal{M}_t, \mathcal{T}_t, \sigma^2)$, μ is the grand mean, and g_i and e_j denote the effects of genotypes and environments, respectively. The component $\sum_{t=1}^{T} h(\mathbf{x}_i, \mathcal{M}_t, \mathcal{T}_t)$ is the same as presented in the previous section and $\mathbf{x}_{ij}$ contains dummy variables that represent the levels of g_i and e_j. In order to get the posterior distribution of the new parameters, we assume that $\mu \sim \mathrm{N}(m, \sigma^2_m)$, $g_i \sim \mathrm{N}(0, \sigma^2_g)$ and $e_j \sim \mathrm{N}(0, \sigma^2_e)$ as well as that $\sigma^2_g \sim \mathrm{IG}(a_g, b_g)$ and $\sigma^2_e \sim \mathrm{IG}(a_e, b_e)$. At first look our model is similar to the semi-parametric BART proposed by (45). However, our approach differs in that i) we do not partition the covariates into two distinct subsets, as the dummy variables (g_i and e_j) that are used in the linear predictor are also contained in $\mathbf{x}_{ij}$; ii) most importantly, we replace the growing move with a 'double grow' (and equivalently 'double prune') step so that we guarantee the trees will include at least one g_i and one e_j as splitting criteria and; iii) unlike (45), we do not use the residuals $\mathbf{r} = \mathbf{y} - \sum_{t=1}^{T} h(\mathbf{X}, \mathcal{M}_t, \mathcal{T}_t)$ to update the linear predictor estimates, but rather the response variable itself, which is analogous to the two-stage estimation idea from the classical AMMI model. In (45), the trees use the partial residuals rather than the response variable to both generate the node-level predictions and update the main effects. However, in the classical AMMI, the estimation of the main effects (g and e) is carried out taking into account the response variable and then the bilinear/interaction term is estimated via an SVD by using the residuals. That is, in the AMMI model the estimation of the model components is carried out in two stages, which does not occur in (45). In this sense, as AMBARTI is a combination of BART and AMMI, we estimate the main effects as AMMI and the interactions via BART also using a two-stage estimation.

The idea of the 'double' moves is to force the trees to exclusively work on the interactions between g_i and e_j. This removes the chance that they split on a single g_i or e_j variable, which would lead to confounding with the main marginal genomic or environment effects. For example, in the double grow, rather than randomly selecting one covariate and one split point when growing a tree, a variable g^* is chosen and then another variable e^* is randomly selected and both define the splitting rules of the corresponding tree. The dummy variables g^* and e^* represent the sets of all possible combinations of g_i and e_j, respectively. To illustrate this, suppose that $I = 10$. In this case, during a 'double grow' move, there will be $2^{I-1}-1 = 511$ dummy variables g^*, as there are 10, 45, 120, 210 and 126 possible combinations of g_i's when choosing them in sets of length 1, 2, 3, 4 and 5, respectively.

An appealing advantage of AMBARTI over AMMI is that it does not require the specification of the number of components Q in the bilinear sum and does not require complex orthonormality constraints on the interaction structure; see Appendix B for the constraints of the AMMI model. In a Bayesian context, these constraints can lead to complex prior distributions choices for implementation of AMMI (as in [6, 46]). Furthermore, although AMBARTI adds a computational cost to the BART model, we have found this to be negligible for standard MET datasets that usually have values of I and J up to the low tens or hundreds.

An additional advantage of using a fully Bayesian approach as in AMBARTI is that we have access to full posterior distributions of each parameter. As the model is fitted jointly, we are thus able to ascertain the general levels of uncertainty in each g_i or e_j component, which may assist with future experimental designs. Similarly, the interaction term is also estimated probabilistically, and so may avoid interpretation errors associated with, e.g., biplots from a traditional AMMI model.

The AMBARTI model can be fitted as follows. First, the parameter estimates g_i and e_j are sampled taking into account the response variable $\mathbf{y}$ (not the residuals). Then, one at a time, the trees are updated via partial residuals $r_t = \mathbf{y} - \hat{\mu} - \hat{g}_i - \hat{e}_j - \sum_{k \neq t}^{T} h(\mathbf{X}, \mathcal{M}_k, \mathcal{T}_k)$. Hence, the terminal node parameters are generated and the sample variance is updated. In the end, posterior samples associated with μ, g_i, e_i, σ_g, σ_e, $\mathcal{T}_t$, $\hat{\mathbf{y}}$ are available, which allow for the calculation of credible intervals and evaluation of the significance of the parameter estimates; see Algorithm 1 in Appendix A for more details.

Simulation Study

Here we compare AMMI and AMBARTI using the Root Mean Square Error (RMSE) for predicted values $\hat{y}$ and for the interaction term on out of sample data. Our simulation experiment was carried out considering two scenarios. In the first, we simulated from the AMMI model with $Q = \{1,2,3\}$, and then fitted AMBARTI and AMMI. In the second scenario, we simulated from the AMBARTI equation and then fitted three AMMI models with different number of components to describe the interactions (i.e., $Q = \{1,2,3\}$) and AMBARTI. In both scenarios, we fitted the models to a training set with $I \times J$ observations and evaluated the performance on an out-of-sample test set of the same size.

For both scenarios, we set $I = J = \{10\}$ or $I = J = \{25\}$, $\mu = 100$ and generated g_i and e_j from $\text{N}(0, \sigma_g^2)$ and $\text{N}(0, \sigma_e^2)$, respectively, where $\sigma_g = \sigma_e \in \{1, 5\}$. The parameters γ_{ik} and δ_{jk} were generated from $\text{N}(0,1)$ and then the orthornormality constraints were applied following the idea presented in Appendix B. In addition, for $Q = 1$, we consider $\lambda = (\{8\}, \{12\})$; for $Q = 2$, $\lambda = (\{12, 8\}, \{12, 10\})$ and; for $Q = 3$, $\lambda = \{12, 10, 8\}$. In the simulation from the AMBARTI equation, we set $T = 200$ trees and generated each tree by using the 'double grow' move considering $2^{I-1}-1$ possible covariates for g_i and $2^{J-1}-1$ for e_j

Finally, the AMMI model used in the simulations is presented in Equation 4, which represents a Completely Randomised Trial Design (CRTD). In constrast, the AMBARTI model used is shown in Equation 5.

Simulation results. We start with the harshest test for the AMBARTI model. Figure 1 shows the RMSE values for $\hat{y}$ based on the out-of-sample sets of both models considering 10 Monte Carlo repetitions. The datasets considered in this Figure were simulated considering $I = 10$ genotypes and $J = 10$ environments, with different values of $Q \in \{1, 2, 3\}$, and two values for the genotypic and environmental variances $\sigma_g \in \{1, 5\}$ and $\sigma_e \in \{1, 5\}$, respectively.

As the data were simulated from the AMMI equation, we would expect the AMMI model performs exceedingly well, and this is what we see in general considering all the results of Figure 1. More specifically, we can see in the first upper panel that AMBARTI has higher RMSEs compared to AMMI for all values of Q. In addition, it is possible to note that there is no clear effect of σ_g or σ_e on the RMSEs. However, even with the AMMI presenting the best results, AMBARTI demonstrates highly competitive performance, with RMSE values around 17% higher than that of the AMMI model.

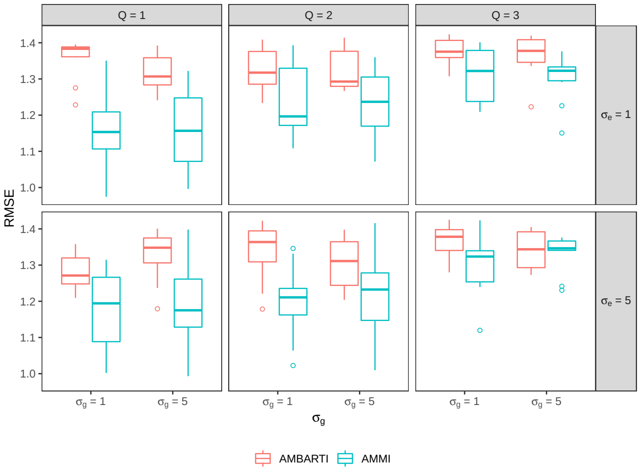

Fig. 1. Out-of-sample RMSE for $\hat{y}$ based on the results of AMMI and AMBARTI for data simulated from the AMMI model with $I = J = 10$. The different panels contain 10 Monte Carlo repetitions and represent different combinations of the simulated parameters for the creation of the dataset. Unsurprisingly, AMMI performs very well here, with AMBARTI having RMSE values around 17% higher.

Figure 2 shows the results of the second simulation scenario, where the data were simulated from the AMBARTI equation. Again, different combinations of parameters were used in the simulation of the training and out-of-sample sets. The upper panels show results for $I = 10$ genotypes and $J = 10$ environments; the lower ones for $I = 25$ and $J = 25$. Furthermore, three AMMI models were fitted considering $Q = 1$, $Q = 2$ and $Q = 3$. In this case, the AMMI model, even with high values of Q, performs very poorly with RMSE values 3 times higher on average than that of AMBARTI. In this comparison, it is worth mentioning that more complex possibilities of interactions may be obtained when simulating from

AMBARTI compared to AMMI.

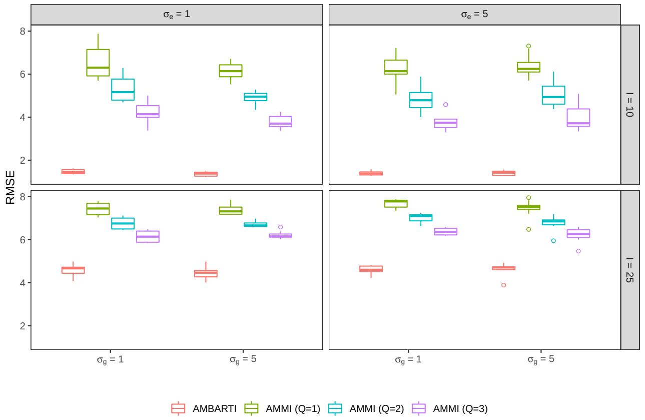

Fig. 2. Out-of-sample RMSE for $\hat{y}$ based on the results of AMMI (with varying Q) and AMBARTI for data simulated from the AMBARTI model with $I = J = 10$ and $I = J = 25$. The different panels contain 10 Monte Carlo repetitions and represent different combinations of the simulated parameters for the creation of the dataset. The AMMI RMSE values are on average 3 times higher than that of AMBARTI.

The next important comparison to be made between AMBARTI and AMMI is related to the interaction term (i.e., the bilinear term for AMMI and the BART component for AMBARTI). Such tests are shown in Figures 3 and 4, where we show the RMSE performance just on the interaction component.

Figure 3 presents the RMSEs associated solely with the interaction terms from AMMI and AMBARTI when the data are simulated from AMMI (which has a bilinear structure for the interactions). The results are presented considering 10 genotypes and 10 environments with different combinations of genotypic and environmental variances. The performance of AMMI is optimal compared to AMBARTI, though the difference between the two is lessened with more complex AMMI model structures (i.e. $Q = 3$).

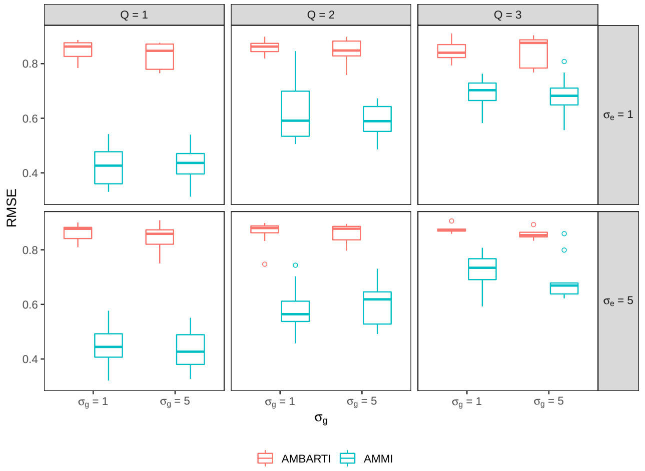

Fig. 3. Out of sample RMSE related to the interaction term of AMMI models for data simulated from AMMI. The different panels show the different parameter values used in the simulation. The performance of AMMI here is optimal, with AMBARTI performing slightly worse than AMMI when $Q = 3$.

In Figure 4, the values of RMSE are presented for datasets simulated from AMBARTI. In the margins of the figure, the parameters used in the simulations can be found. The RMSE values show that AMMI performs worse than AMBARTI in all scenarios, and in the same cases AMMI RMSEs are three times higher on average than those of AMBARTI.

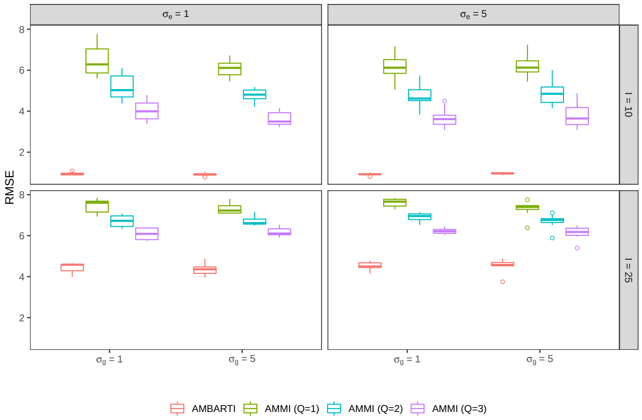

Fig. 4. Out of sample RMSE related to interaction term of AMBARTI and AMMI models for data simulated from AMBARTI. The different panels show the different parameter values used for the simulation. It appears that the AMMI structure, even with $Q = 3$ cannot capture the interaction behaviour present in the AMBARTI model.

In summary, the information presented in Figures 3 and 4 shows that the AMMI model fails for the complex interactions that can be obtained in the AMBARTI simulated datasets. From a quantitative genetics/biological perspective, there is no reason for the structure of interactions between genotype and environments be modelled strictly by a bilinear structure, as more complex structures can be assumed to be present in nature. In this sense, AMBARTI may be a more suitable model to estimate the interaction structure in the real-world applications.

Case study: Innovar wheat data

In addition to the simulated datasets, real datasets were used to evaluate the performance of AMMI and AMBARTI. A set of Value of Cultivation and Usage (VCU) experiments conducted in Ireland between the years of 2010 and 2019 were considered, and such experiments evaluated the performance of genotypes of wheat *Triticum aestivum L.* across the country for regulatory purposes (i.e., registration of new varieties). Here, our phenotypic response variable is the production of wheat in tonnes per hectare. The design of experiments used was that of a block design with 4 replicates. VCUs alongside Distinctness, Uniformity and Stability (DUS) are the most important kind of regulatory Multi Environmental Trials conducted around the world.

The data were kindly provided by the Irish Department of Agriculture, Food, and Marine. Both genotypes and environments were anonymised. These historical VCUs form part of the Horizon2020 EU project Innovar project database (`www.h2020innovar.eu`). The project aims to build and improve technical solutions for cultivar recommendation based on genomic and phenomic parameters. The models were fitted for all years available (summarised in Table 1), but for brevity we show detailed plots only for the year 2015.

We compare the models by evaluating the estimated values of the genotype and environment effects, and the predictions of interaction behaviour evaluated as: $\widehat{(ge)}_{ij} = y_{ij} - \hat{g}_i - \hat{e}_j$. To fit both models, we average the response variable across the replicates. This is a common practice in the analysis of GxE experiments with AMMI models (6, 46). However, this preprocessing is not needed for AMBARTI. To validate the models, we sample at random two replicates and then calculate the RMSE for $\hat{y}$.

The estimates of the genotype effects g_i and environment effects e_j are shown in Figure 5 and 6, respectively. Here, both models provided similar parameter estimates when we compare the posterior means and the point estimates from the classical AMMI. The advantages of the posterior distribution can be highlighted here once they indicate that the range of possible true values for the main effect of genotypes or environments can be much different from the point estimated obtained by AMMI models (that consider main effects as fixed effects).

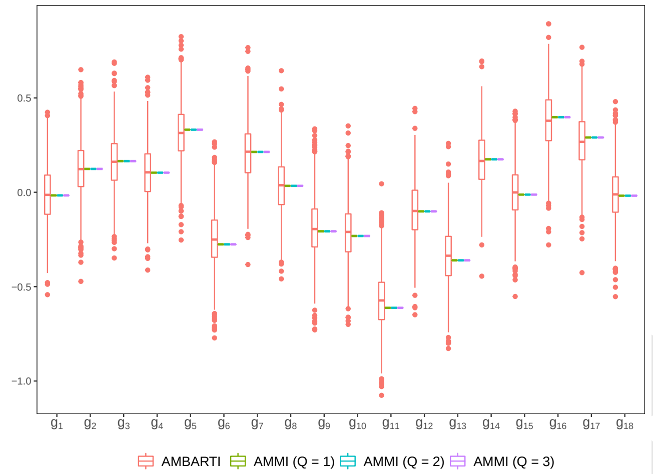

Fig. 5. Parameter estimates of genotype effects for the Irish dataset for 2015. The boxplots represent the posterior distribution obtained via AMBARTI. The point estimates obtained via AMMI are given for different values of Q.

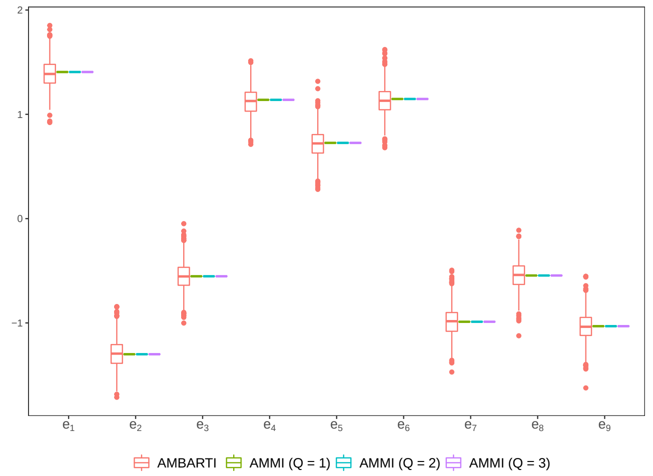

Fig. 6. Parameter estimates of environment effects for the Irish dataset for 2015. The boxplots represent the posterior distribution obtained via AMBARTI. The point estimates obtained via AMMI are given for different values of Q.

Year	AMMI	AMBARTI
2010	0.50	0.48
2011	0.45	0.42
2012	0.37	0.30
2013	0.44	0.42
2014	0.46	0.43
2015	0.38	0.33
2016	0.42	0.38
2017	0.41	0.39
2018	0.56	0.55
2019	0.45	0.40

Table 1. RMSE for $\hat{y}$ on out-of-sample data considering all years in the historical Innovar data. The values of RMSE obtained with AMBARTI are smaller than the ones obtained via AMMI models ($Q = 3$) for all years considered.

A more complete comparison across all years is shown in Table 1. In this table, we calculated the predicted values $\hat{y}$ on the 'out-of-sample' data (i.e., on the two replicates randomly selected). We can see that RMSEs obtained with AMBARTI are smaller than the ones returned by the AMMI model for all years, thus highlighting that the AMBARTI model can more accurately estimate the marginal effects along with interaction component.

Regarding the computational time, AMBARTI took about 6 minutes on average to run, considering 50 trees, 1000 iterations as burn-in and 1000 iterations as post burn-in. This time was registered in a MacBook Pro 2.3 GHz Dual-Core Intel Core i5 with 8GB memory. AMMI took just seconds. This difference could be reduced by optimising the AMBARTI implementation using routines in C++ similar to those for BART implementations in R packages `BART` (44) and `dbarts` (7). However, we believe AMBARTI's superior performance and posterior estimation of uncertainties outweighs the longer computational time.

New visualisations for AMBARTI main effects and interactions. One of the key outputs of the standard AMMI model is the biplot (47), which assists in the determination of key GxE interactions and may be used for cultivar recommendation. However, these plots display only the interaction measure, thus missing the key marginal effects that may also come into play. For example, a certain genotype and environment may have a strong positive interaction, but if this genotype is consistently poor in all environments this may not be clear in the biplot. Instead, we introduce new types of plots that enables this full consideration.

Our first new plot is based on a heat map adapted to display both the GxE interactions (along with the marginal effects) and the predicted yields from the AMBARTI model. In Figure 7, we display the GxE interactions in the centre of the plot and the marginal effects for both environment and genotype as separate bars in the margins. The ordering applied to Figure 7 is in terms of the marginal effects for both environment and genotype, and displays low values in red to high values in blue. As both the GxE interactions and marginal effects are on the same scale and are centred around zero, we display them using only one legend and use a divergent colour palette. This allows for quick identification of the GxE interactions and to observe which of the environments or genotypes are the most or least optimal.

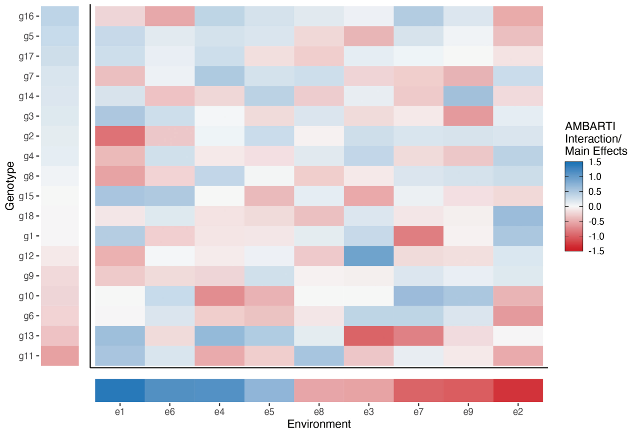

Fig. 7. GxE interactions and main effects for the AMBARTI model sorted by the main effects for the Innovar data in 2015. We can clearly see that environments 1, 4, 5, and 6 provide superior yields for almost all genotypes studied. Furthermore, environment 1, for example, seems to interact particularly strongly in a negative way with genotype 2.

In Figure 8, we show the ordered heat map of the predicted yields (as opposed to their component parts shown in Figure 7) for each combination of environment and genotype for the AMBARTI model. In this case, we use the same ordering as that in Figure 7 with high values being generally displayed in the top left, moving to low values at the bottom right, with the units for the plot being the same as that of the phenotype (i.e. yield/production of grains in tonnes per hectare). For this plot, we use the same diverging colour palette as in Figure 7 as, when combined with the ordering, this gives a clear identification as to which environment and genotype produce high or low yields.

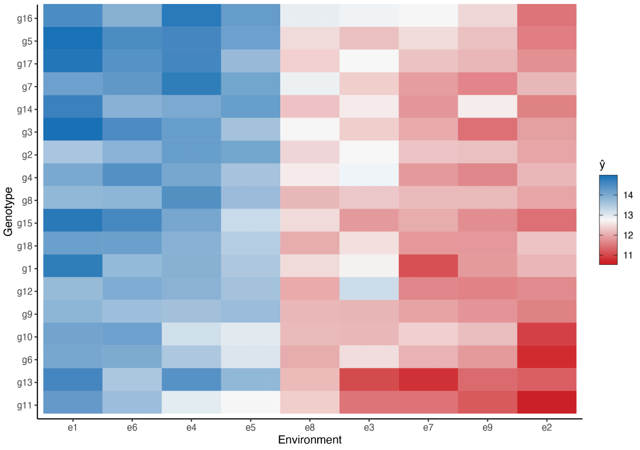

Fig. 8. Predicted yields from the AMBARTI model for the Innovar data in 2015. Values are sorted by the main effects. We can see that the interaction effect of environment 3 with genotype 12 seems particularly strong.

In Figure 9, we show a bipartite plot of the information displayed in Figure 7, but showing only the extremes of the high and low values. In this case, we display just the top 2% and the lowest 2% of the interactions. We employ the same diverging colour palette as Figure 7 except in this case, both the colour and size of each node represents the marginal effects and the colour and width of each edge represents the interaction value. Larger values of the marginal effects will result in larger nodes (and vice-versa), whereas the magnitude of the interactions determine the width of the edges. The aim of this plot is to allow the reader to easily and quickly identify which of the environments are the most and least optimal for each genotype and to also identify where there are clear interactions.

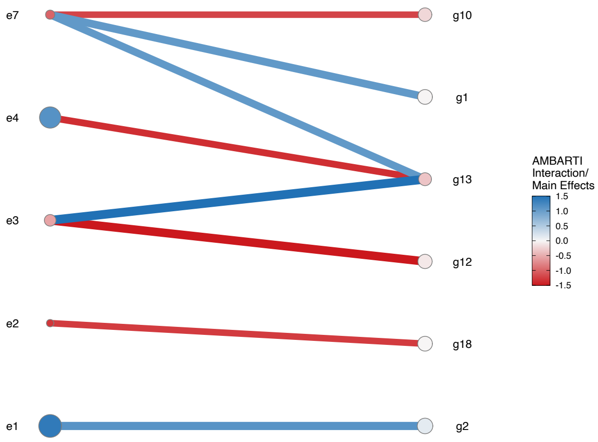

Fig. 9. Bipartite network plot showing the top (in blue) and bottom (in red) 2% GxE interactions and main effects. We can see that environment 3 has strong positive and negative interactions with genotypes 13 and 12 respectively.

The visualisation perspective proposed here helps construct easily interpretable agronomic recommendations. In this sense, Figure 7 can help users with no background in statistics identify that the best genotypes considering yield are the ones in the top left corner: g_{16}, g_5, g_{17}, g_7. These genotypes will have a tendency to have a better acceptance by farmers, considering solely the yield in tonnes per hectare assuming higher yields are economically preferred. Figure 7 shows us that environments e_1, e_6, e_4, e_5 are related to higher marginal effects and should be considered preferential to crop the list of wheat genotypes evaluated.

Figure 7 is also useful to establish combinations of genotypes and environments that should be avoided when the interaction is negative, indicating that a given genotype does not perform well in a given environment. This negative interaction increases the risk of low yield and consequent economic impacts. Combinations to be avoided exist even for environments and genotypes with high marginal effects. For instance, the combination of g_2, e_1 should be avoided even though g_2 and e_1 have high marginal effects. This is an important information for regulators who may be responsible for a variety's commercialisation approval or agents that promote credit or insurance for farmers given to risks that the negative interaction implies. Farmers who produce a genotype not indicated for their environment can end having a worse score or risk. On the other hand, Figure 8 is also useful to spot the combinations of genotypes and environments that should be encouraged once the signal of the interactions is positive.

In adaptability breeding, the breeder seeks to find the best genotype for a specific environment or a small set of environments. In broad target strategies, the aim is to find genotypes that perform well across several environments. For example, in Figure 7, g_5 has high marginal effect and performs well (and interacts positively) with environments $e_1, e_6, e_4, e_5, e_7, e_9$. Similarly g_{16}, the best genotype considering marginal effects, performs well in environments $e_4, e_5, e_8, e_3, e_7, e_9$. Genotypes which present better performance across several environments are classified as high stability genotypes. They tend to be preferred by breeders because they allow optimisation of processes in the chain of seed production.

Discussion

We have introduced a new model named Additive Main Effects Bayesian Additive Regression Trees Interaction (AMBARTI). AMBARTI is a fully Bayesian semi-parametric machine learning approach that estimates main effects of genotypes and environments and interactions with an adapted regression tree-like structure. This approach to interactions allows the treatment of more complex structures than the ones considered by traditional models.

Given the fact that GxE interactions are the result of a tangled myriad of genetics, proteomics, biochemical, environmental and additional factors, the flexibility of AMBARTI in dealing with more complex interactions can be seen as an important improvement in the understanding of the complexities associated to GxE phenomenon. Given the complexities of GxE interactions, a bilinear term is perhaps an oversimplification. AMBARTI allows the possibility of reasoning other than the ones obtained by models which consider the genotypic and environmental effects as linear and the interaction GxE in the maximum as bilinear. This makes AMBARTI a useful candidate to expand understanding of experimental data in quantitative genetics.

The main novelty in AMBARTI comes from its semi-parametric structure which enables the uncertainty to be shared between the main effects and the interaction trees. More specifically, we design the trees so that they can only incorporate the interaction terms by forcing each branch of a tree to split on both a genotype and an environment. We have shown in simulation experiments that this yields similar estimates to traditional models for the marginal effects, and superior estimates for the interaction terms, which are no longer restricted to be linear in a restricted dimensional space. This removes the need for, e.g., the arbitrary selection of the Q parameter in a standard AMMI formulation.

A second novelty is that we have introduced new displays that simultaneously allow for interpretation of the marginal and joint effects. We have created both a heatmap and a bipartite network-style plot of the results, which we hope will enable those using the output of AMBARTI models to make more informative decisions about which genotype and environments are most compatible.

We believe that there are many possible extensions of the AMBARTI approach. Other more advanced methods, such as PARAFAC (48, 49), are available for higher dimensional interactions (such as with time). These are different versions of tensor regression (50) and, in theory, there is no reason why the AMBARTI approach cannot be used for higher dimensional tensor-type interactions, though this is not currently possible in our code. Similar enhancements for multivariate outputs and time-series like-structure seem promising, and we hope to explore these in future work.

Appendix A - AMBARTI implementation

In this section, we detail the AMBARTI model. Firstly, the likelihood function associated to y_{ij} is

$$y_{ij}|\mathbf{x}_{ij},\Theta \sim \mathrm{N}\left(\mu+g_i+e_j+\sum_{t=1}^{T} h(\mathbf{x}_{ij},\mathcal{M}_t,\mathcal{T}_t),\sigma^2\right),$$

where y_{ij} denotes the response variable for genotype i and environment j, $\Theta = (\mu, g_i, e_j, \mathcal{M}_t, \mathcal{T}_t, \sigma^2)$, μ is the grand mean, $\mathbf{x}_{ij}$ is the row of the design matrix $\mathbf{X}$ associated to observations with genotype i and environment j, and $h(\cdot) = \mu_{t\ell}$ is a function that assigns the predicted values $\mu_{t\ell} \in \mathcal{M}_t$ to observations that belong to $\mathcal{P}_{t\ell}$, with $\mathcal{P}_{t\ell}$ denoting the set of rules that define the node ℓ of the tree t. In order to obtain the posterior distributions needed for the model, we assume the following prior distributions:

$$\begin{aligned}
\mu &\sim \mathrm{N}(m, \sigma_m^2),\\
\mu_{t\ell}|\mathcal{T}_t &\sim \mathrm{N}(0, \sigma_\mu^2),\\
g_i|\mathcal{T}_t &\sim \mathrm{N}(\mu_g, \sigma_g^2),\\
e_j|\mathcal{T}_t &\sim \mathrm{N}(\mu_e, \sigma_e^2),\\
\sigma_g^2 &\sim \mathrm{IG}(a_g, b_g),\\
\sigma_e^2 &\sim \mathrm{IG}(a_e, b_e),\\
\sigma^2 &\sim \mathrm{IG}(a, b).
\end{aligned}$$

The prior distribution on the tree structure depends on the number of terminal and internal nodes, and is given by

$$p(\mathcal{T}_t) = \prod_{\ell \in \mathcal{A}_I}\left[\alpha(1+d_{t\ell})^{-\beta}\right] \times \prod_{\ell \in \mathcal{A}_T}\left[1-\alpha(1+d_{t\ell})^{-\beta}\right],$$

where $\mathcal{A}_I$ and $\mathcal{A}_T$ denote the sets of indices of the internal and terminal nodes, respectively, and $d_{t\ell}$ represents the depth of the node ℓ of the tree t. Furthermore, let $R_t = \mathbf{y} - \left(\mu + \mathbf{g} + \mathbf{e} + \sum_{k\neq t}^{T} h(\mathbf{X}; \mathcal{T}_k, \mathcal{M}_k)\right)$ denote the vector of the partial residuals, where $\mathbf{g} \in \mathbb{R}^I$ and $\mathbf{e} \in \mathbb{R}^J$ are vectors containing the random effects g_i and e_j. Below, we present the full conditional of μ.

$$\begin{aligned}
p(\mu|-) &\propto p(\mathbf{y}|g_i, e_j, \mathbf{x}_{ij}, \mathcal{M}_t, \mathcal{T}_t, \sigma^2)p(\mu),\\
&\propto \exp\left(-\frac{1}{2\sigma_{m*}^2}(\mu-\mu^*)^2\right),
\end{aligned}$$

which is a

$$\mathrm{N}\left(\frac{\sum_i\sum_j\left[y_{ij}-\hat{y}_{ij}^*\right]/\sigma^2 + m/\sigma_m^2}{n/\sigma^2+1/\sigma_m^2}, \frac{1}{n/\sigma^2+1/\sigma_m^2}\right),$$

where $\hat{y}^*_{ij} = \mathbf{g} + \mathbf{e} + \sum_{t=1}^{T} h(\mathbf{x}_{ij}; \mathcal{T}_t, \mathcal{M}_t)$. Hence, the full conditional of g_i is given by

$$\begin{aligned} p(g_i|-) &\propto p(\mathbf{y}|g_i, e_j, \mathbf{x}_{ij}, \mathcal{M}_t, \mathcal{T}_t, \sigma^2) p(g_i), \\ &\propto \exp\left(-\frac{1}{2\sigma^2_{g*}}(g_i - g_i^*)^2\right), \end{aligned}$$

which is a

$$\mathrm{N}\left(\frac{\sum_j [y_{ij} - \hat{\mu} - \hat{e}_j - \hat{\mu}_{ij}]/\sigma^2}{n_{g_i}/\sigma^2 + 1/\sigma^2_g}, \frac{1}{n_{g_i}/\sigma^2 + 1/\sigma^2_g}\right),$$

where $\hat{\mu}_{ij} = \sum_{t=1}^{T} h(\mathbf{x}_{ij}, \mathcal{T}_t, \mathcal{M}_t)$ and n_{g_i} is the number of observations that belong to g_i; similarly to n_{e_j}. To be able to estimate the linear and interaction components in a two-stage approach like the AMMI model, we assume that $\sum_j \hat{\mu}_{ij} = \sum_i \hat{\mu}_{ij} = 0$, for $i = 1, \ldots, I$ and $j = 1, \ldots, J$. Similarly, the full conditional of e_j can be written as

$$\begin{aligned} p(e_j|-) &\propto p(\mathbf{y}|g_i, e_j, \mathbf{x}_{ij}, \mathcal{M}_t, \mathcal{T}_t, \sigma^2) p(e_j), \\ &\propto \exp\left(-\frac{1}{2\sigma^2_{e*}}(e_j - e_j^*)^2\right), \end{aligned}$$

which is a

$$\mathrm{N}\left(\frac{\sum_i [y_{ij} - \hat{\mu} - \hat{g}_i - \hat{\mu}_{ij}]/\sigma^2}{n_{e_j}/\sigma^2 + 1/\sigma^2_e}, \frac{1}{n_{e_j}/\sigma^2 + 1/\sigma^2_e}\right).$$

The full conditional of σ^2_g is given by

$$p(\sigma^2_g|-) \propto p(\mathbf{g}|\sigma^2_g) p(\sigma^2_g),$$

which is an

$$\mathrm{IG}\left(\frac{I}{2} + a_g, \frac{\sum_{i=1}^{I} g_i^2}{2} + b_g\right).$$

The full conditional of σ^2_e is written as

$$p(\sigma^2_e|-) \propto p(\mathbf{e}|\sigma^2_e) p(\sigma^2_e),$$

which is an

$$\mathrm{IG}\left(\frac{J}{2} + a_e, \frac{\sum_{j=1}^{J} e_j^2}{2} + b_e\right).$$

In addition, we present the full conditional of the trees. This distribution is utilised to compare the previous tree to the current one, as in BART the splitting rules are created by randomly selecting a covariate and a split point. Below, we present the full conditional of $\mathcal{T}_t$ as

$$\begin{aligned} p(\mathcal{T}_t|R_t, \sigma^2) &\propto p(\mathcal{T}_t) \int p(R_t|\mathcal{M}_t, \mathcal{T}_t, \sigma^2) p(\mathcal{M}_t|\mathcal{T}_t) d\mathcal{M}_t, \\ &\propto p(\mathcal{T}_t) p(R_t|\mathcal{T}_t, \sigma^2), \\ &\propto p(\mathcal{T}_t) \prod_{\ell=1}^{b_t} \left[\left(\frac{\sigma^2}{\sigma^2_\mu n_{t\ell} + \sigma^2}\right)^{1/2} \exp\left(\frac{\sigma^2_\mu \left[n_{t\ell} \bar{R}_\ell\right]^2}{2\sigma^2(\sigma^2_\mu n_{t\ell} + \sigma^2)}\right)\right], \end{aligned}$$

where $\bar{R}_\ell = \sum_{(i,j)\in\mathcal{P}_{t\ell}} (r_{ij}^{(t)} - \hat{\mu} - \hat{g}_i - \hat{e}_j)/n_{t\ell}$, $r_{ij}^{(t)} \in R_t$ and $n_{t\ell}$ is the number of observations that belong to $\mathcal{P}_{t\ell}$. To sample from this expression, the Metropolis-Hastings algorithm is used, because a closed-form distribution is not obtained in this case.

As all $\mu_{t\ell}$ are i.i.d, it is possible to write $p(\mathcal{M}_t|\mathcal{T}_t, R_t, \sigma^2) = \prod_{\ell=1}^{b_t} p(\mu_{t\ell}|\mathcal{T}_t, R_t, \sigma^2)$. Similarly to the original BART, the full conditional of $\mu_{t\ell}$ in the AMBARTI model also depends only on the information provided by all trees, except by $\mathcal{T}_t$, via partial residual as R_t. Hence, the full conditional of $\mu_{t\ell}$ can be written as

$$\begin{aligned} p(\mu_{t\ell}|-) &\propto p(R_t|\mathcal{M}_t, \mathcal{T}_t, \sigma^2) p(\mu_{t\ell}), \\ &\propto \exp\left(-\frac{1}{2\sigma^2_*}(\mu_{t\ell} - \mu_{t\ell}^*)^2\right), \end{aligned}$$

which is a

$$\mathrm{N}\left(\frac{\sigma^{-2}\sum_{(i,j)\in\mathcal{P}_{t\ell}} r_{ij}^{(t)}}{n_{t\ell}/\sigma^2 + \sigma_\mu^{-2}}, \frac{1}{n_{t\ell}/\sigma^2 + \sigma_\mu^{-2}}\right).$$

Finally, after generating all predicted values for all trees, σ^2 can be updated based on

$$\begin{aligned} p(\sigma^2|-) &\propto p(\mathbf{y}|\mathbf{g}, \mathbf{e}, \mathbf{X}, \mathcal{M}_t, \mathcal{T}_t, \sigma^2) p(\sigma^2) \\ &\propto (\sigma^2)^{-(\frac{n+\nu}{2}+1)} \exp\left(-\frac{S + \nu\lambda}{2\sigma^2}\right), \end{aligned} \qquad \textbf{(6)}$$

where $S = \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \hat{y}_{ij})^2$ and $\hat{y}_{ij} = \mu + \mathbf{g} + \mathbf{e} + \sum_{t=1}^{T} h(\mathbf{x}_{ij}; \mathcal{T}_t, \mathcal{M}_t)$. The expression in Eq. (6) is an $\mathrm{IG}((n+\nu)/2, (S+\nu\lambda)/2)$.

Algorithm 1: AMBARTI Algorithm

```
Update $\mu$, $g_i$ and $e_j$;
for t in 1:T do
    Compute
      $R_t = \mathbf{y} - \left(\mu + \mathbf{g} + \mathbf{e} + \sum_{k \neq t}^{T} h(\mathbf{X}; \mathcal{T}_k, \mathcal{M}_k)\right)$;
    Propose a new tree $\mathcal{T}_t^*$ based on $\mathcal{T}_t$ by growing,
      pruning, changing or swapping;
    Accept the proposed tree with probability
      $\alpha(\mathcal{T}_t, \mathcal{T}_t^*) = \min\left\{1, \frac{p(R_t|\mathcal{T}_t^*, \sigma^2) p(\mathcal{T}_t^*)}{p(R_t|\mathcal{T}_t, \sigma^2) p(\mathcal{T}_t)}\right\}$;
    Update the node-level parameters $\mu_{t\ell}$ from
      $p(\mu_{t\ell}|-)$;
    Update $\sigma^2$ sampling from $p(\sigma^2|-)$.
end
```

Appendix B - Orthonormality constraints of the AMMI model

We recall the AMMI model is overparameterised, so constraints need to be imposed so that the parameters can be estimated (6). In this section, we show how to apply the orthonormality constraints on γ_{iq} and δ_{jq} when simulating from the AMMI model.

Let $\boldsymbol{\gamma}$ be an $I \times Q$ matrix, $\boldsymbol{\delta}$ a $J \times Q$ matrix, and consider that γ_{iq} and δ_{jq} are elements in row i and column q of the corresponding matrices. In this sense, the following constraints are considered: i) $\sum_{i=1}^{I} \gamma_{iq} = \sum_{j=1}^{J} \delta_{jq} = 0$, for $q = 1, \ldots, Q$; ii) $\boldsymbol{\gamma}^{\top} \boldsymbol{\gamma} = \boldsymbol{\delta}^{\top} \boldsymbol{\delta} = I_q$, where I_q represents an identity matrix of dimension q; iii) $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_{q-1} \geq \lambda_q \geq 0$ and; iv) $\gamma_{iq} \geq 0$, for all $q = 1, \ldots, Q$.

To illustrate our the strategy to meet the constraints presented above, we take the γ_{iq} as an example, but this also works for δ_{jq}. First, we create S an $I \times Q$ matrix, where $s_{iq} \sim \mathrm{N}(0, \sigma_x^2 = 1)$. Here, s_{iq} could be sampled from other distributions, such as Gamma or Beta. In addition, we define M as an $I \times Q$ matrix with each element being the mean of the corresponding q column of S. Hence, we know that

$$
\begin{aligned}
B &= S - M \\
&\Rightarrow \mathbf{1}_I^{\top} B = 0 \\
&\nRightarrow B^{\top} B = \mathbb{I},
\end{aligned}
$$

where $\mathbf{1}_I$ is a column vector of dimension I containing ones, B is, by construction, a full rank matrix and $B^{\top} B$ is symmetric. However, we find a matrix A such that $C = BA \Rightarrow C^{\top} C = \mathbb{I}$. That is, we know that

$$
\begin{aligned}
D &= C^{\top} C = \mathbb{I} \\
&\Rightarrow (BA)^{\top} BA = \mathbb{I} \\
&\Rightarrow A^{\top} B^{\top} BA = \mathbb{I} \\
&\Rightarrow B^{\top} B = A^{-\top} A^{-1} \\
&\Rightarrow B^{\top} B = (AA^{\top})^{-1} \\
&\Rightarrow (B^{\top} B)^{-1} = AA^{\top} \\
&\Rightarrow (B^{\top} B)^{-1} = A^2 \text{ (by symmetry)} \\
&\Rightarrow (B^{\top} B)^{-1/2} = A.
\end{aligned}
$$

In the end, we have that $\boldsymbol{\gamma} = B(B^{\top} B)^{-1/2}$.

Acknowledgements

We are very grateful to John Joe Byrne at the Department of Agriculture for providing us with the dataset for our case study.

Funding

Danilo Sarti and Antônia A. L. dos Santos received funding for their work from the European Union's Horizon 2020 research and innovation programme under grant agreement No 818144. Alan Inglis and Estevão Prado's work was supported by the Science Foundation Ireland Career Development Award grant number 17/CDA/4695 and SFI research centre 12/RC/2289_P2. Andrew Parnell's work was supported by: the Science Foundation Ireland Career Development Award (17/CDA/4695); an investigator award (16/IA/4520); the Marine Research Programme funded by the Irish Government, co-financed by the European Regional Development Fund (Grant-Aid Agreement No. PBA/CC/18/01); European Union's Horizon 2020 research and innovation programme under grant agreement No 818144; SFI Centre for Research Training 18CRT/6049, and SFI Research Centre awards 16/RC/3872 and 12/RC/2289_P2. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission'.

Bibliography

1. R.W Allard and A.D. Bradshaw. Implications of genotype environmental interactions in applied plant breeding. *Crop Science*, 4:503–508, 1992.
2. Danilo Augusto Sarti. *Gerenciamento de incertezas por análise de decisões: aplicações à otimização da produção e demandas incertas.* PhD thesis, Universidade de São Paulo, 2013.
3. Danilo Augusto Sarti. *The statistical paradigm: probabilistic and multivariate analysis applied through computational simulation in the interaction between genotype x environment.* PhD thesis, Universidade de São Paulo, 2019.
4. John Mandel. A new analysis of variance model for non-additive data. *Technometrics*, 13 (1):1–18, 1971.
5. Paulo C Rodrigues, Andreia Monteiro, and Vanda M Lourenço. A robust ammi model for the analysis of genotype-by-environment data. *Bioinformatics*, 32(1):58–66, 2016.
6. Julie Josse, Fred van Eeuwijk, Hans-Peter Piepho, and Jean-Baptiste Denis. Another look at bayesian analysis of ammi models for genotype-environment data. *Journal of Agricultural, Biological, and Environmental Statistics*, 19(2):240–257, 2014.
7. Hugh A Chipman, Edward I George, Robert E McCulloch, et al. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.
8. Belinda Hernández, Adrian E Raftery, Stephen R Pennington, and Andrew C Parnell. Bayesian additive regression trees using bayesian model averaging. *Statistics and computing*, 28(4):869–890, 2018.
9. Yang Liu, Mikhail Traskin, Scott A Lorch, Edward I George, and Dylan Small. Ensemble of trees approaches to risk adjustment for evaluating a hospital's performance. *Health care management science*, 18(1):58–66, 2015.
10. Yaoyuan Vincent Tan and Jason Roy. Bayesian additive regression trees and the general bart model. *Statistics in medicine*, 38(25):5048–5069, 2019.
11. DS Falconer and TFC Mackay. Introduction to quantitative genetics. 1996. *Harlow, Essex, UK: Longmans Green*, 3, 1996.
12. CTS Dias. *Métodos para escolha de componentes em modelo de efeito principal aditivo e interação multiplicativa (AMMI). 2005. 73p.* PhD thesis, Tese (Livre Docência)–Escola Superior de Agricultura Luiz de Queiroz, Piracicaba, 2005.
13. Fikret Isik, James Holland, and Christian Maltecca. Multi environmental trials. In *Genetic data analysis for plant and animal breeding*, pages 227–262. Springer, 2017.
14. Harry F Gollob. A statistical model which combines features of factor analytic and analysis of variance techniques. *Psychometrika*, 33(1):73–115, 1968.
15. Irving John Good. Some applications of the singular decomposition of a matrix. *Technometrics*, 11(4):823–831, 1969.
16. Felipe de Mendiburu. Package 'agricolae'. *R Package, Version*, pages 1–2, 2019.
17. Andrea Onofri and Egidio Ciriciofolo. Using r to perform the ammi analysis on agriculture variety trials. *R News*, 7(1):14–19, 2007.
18. Hugh G Gauch Jr. A simple protocol for ammi analysis of yield trials. *Crop Science*, 53(5): 1860–1869, 2013.
19. MM Nachit, G Nachit, H Ketata, HG Gauch, and RW Zobel. Use of ammi and linear regression models to analyze genotype-environment interaction in durum wheat. *Theoretical and Applied genetics*, 83(5):597–601, 1992.
20. E Farshadfar and J Sutka. Locating qtls controlling adaptation in wheat using ammi model. *Cereal Research Communications*, 31(3):249–256, 2003.
21. MR Naroui Rad, M Abdul Kadir, MY Rafii, Hawa ZE Jaafar, MR Naghavi, and Farzaneh Ahmadi. Genotype environment interaction by ammi and gge biplot analysis in three consecutive generations of wheat (triticum aestivum) under normal and drought stress conditions. *Australian Journal of Crop Science*, 7(7):956, 2013.
22. M Brancourt-Hulmel and C Lecomte. Effect of environmental variates on genotype× environment interaction of winter wheat: A comparison of biadditive factorial regression to ammi. *Crop Science*, 43(2):608–617, 2003.
23. B Badu-Apraku, M Oyekunle, K Obeng-Antwi, AS Osuman, SG Ado, N Coulibay, CG Yallou, M Abdulai, GA Boakyewaa, and A Didjeira. Performance of extra-early maize cultivars based on gge biplot and ammi analysis. *The Journal of Agricultural Science*, 150(4):473, 2012.
24. Bojan Mitroviaã, Sanja Treski, Milisav Stojakkovã, Mile Ivanoviã, Goran Bekavac, et al. Evaluation of experımental maize hybrids tested in multi-location trials using ammi and gge biplot analyses. *Turkish Journal of Field Crops*, 17(1):35–40, 2012.
25. L Mahalingam, S Mahendran, R Chandra Babu, and G Atlin. Ammi analysis for stability of grain yield in rice (oryza sativa l.). *International Journal of Botany*, 2006.
26. Ignacio Romagosa, Steven E Ullrich, Feng Han, and Patrick M Hayes. Use of the additive main effects and multiplicative interaction model in qtl mapping for adaptation in barley. *Theoretical and Applied Genetics*, 93(1-2):30–37, 1996.
27. Kazuhiro Sato and Kazuyoshi Takeda. Pathogenic variation of pyrenophora teres isolates collected from japanese and canadian spring barley. *Report by the Institute of Resource Biological Sciences, Okayama University*, 1(2):147–158, 1993.
28. Yadeta Anbessa, Patricia Juskiw, Allen Good, Joseph Nyachiro, and James Helm. Genetic variability in nitrogen use efficiency of spring barley. *Crop Science*, 49(4):1259–1269, 2009.
29. Luís Cláudio Inácio da Silveira, Volmir Kist, Thiago Otávio Mendes de Paula, Márcio Henrique Pereira Barbosa, Luiz Alexandre Peternelli, and Edelclaiton Daros. Ammi analysis to evaluate the adaptability and phenotypic stability of sugarcane genotypes. *Scientia Agricola*, 70(1):27–32, 2013.

30. Dani Gamerman and Hedibert F Lopes. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press, 2006.
31. Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.
32. James H Albert and Siddhartha Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679, 1993.
33. Estevão B Prado, Rafael A Moral, and Andrew C Parnell. Bayesian additive regression trees with model trees. *Statistics and Computing*, 31(3):1–13, 2021.
34. Junni L Zhang and Wolfgang K Härdle. The bayesian additive classification tree applied to credit risk modelling. *Computational Statistics & Data Analysis*, 54(5):1197–1205, 2010.
35. Rodney A Sparapani, Brent R Logan, Robert E McCulloch, and Purushottam W Laud. Nonparametric survival analysis using bayesian additive regression trees (bart). *Statistics in medicine*, 35(16):2741–2753, 2016.
36. Belinda Hernández, Stephen R Pennington, and Andrew C Parnell. Bayesian methods for proteomic biomarker development. *EuPA Open Proteomics*, 9:54–64, 2015.
37. Bereket P Kindo, Hao Wang, and Edsel A Peña. Multinomial probit bayesian additive regression trees. *Stat*, 5(1):119–131, 2016.
38. Antonio R Linero. Bayesian regression trees for high-dimensional prediction and variable selection. *Journal of the American Statistical Association*, 113(522):626–636, 2018.
39. Antonio R Linero and Yun Yang. Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):1087–1110, 2018.
40. Veronika Ročková and Stéphanie van der Pas. Posterior concentration for bayesian regression trees and forests. *Annals of Statistics*, 48(4):2108–2131, 2020.
41. Veronika Ročková and Enakshi Saha. On theory for bart. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2839–2848. PMLR, 2019.
42. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.
43. Adam Kapelner and Justin Bleich. bartmachine: Machine learning with bayesian additive regression trees. *Journal of Statistical Software, Articles*, 70(4):1–40, 2016. ISSN 1548-7660. doi: 10.18637/jss.v070.i04.
44. Robert McCulloch, Rodney Sparapani, Charles Spanbauer, Robert Gramacy, and Matthew Pratola. *BART: Bayesian Additive Regression Trees*, 2020. R package version 2.8.
45. Bret Zeldow, Vincent Lo Re III, and Jason Roy. A semiparametric modeling approach using bayesian additive regression trees with an application to evaluate heterogeneous treatment effects. *The Annals of Applied Statistics*, 13(3):1989, 2019.
46. José Crossa, Sergio Perez-Elizalde, Diego Jarquin, José Miguel Cotes, Kert Viele, Genzhou Liu, and Paul L Cornelius. Bayesian estimation of the additive main effects and multiplicative interaction model. *Crop Science*, 51(4):1458–1469, 2011.
47. Karl Ruben Gabriel. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467, 1971.
48. KE Basford, PM Kroonenberg, and IH DeLacy. Three-way methods for multiattribute genotype× environment data: an illustrated partial survey. *Field Crops Research*, 27(1-2):131–157, 1991.
49. Richard A Harshman and Margaret E Lundy. Parafac: Parallel factor analysis. *Computational Statistics & Data Analysis*, 18(1):39–72, 1994.
50. Rajarshi Guhaniyogi, Shaan Qamar, and David B Dunson. Bayesian tensor regression. *The Journal of Machine Learning Research*, 18(1):2733–2763, 2017.