

Genetics and population analysis

A robust AMMI model for the analysis of genotype-by-environment data

Paulo C. Rodrigues^{1,2,*}, Andreia Monteiro³ and Vanda M. Lourenço^{3,4}

¹Department of Statistics, Federal University of Bahia, Salvador, Brazil, ²CAST—Center for Applied Statistics and Data Analytics, University of Tampere, Finland, ³Centro de Matemática e Aplicações (CMA), and ⁴Department of Mathematics, FCT—NOVA University of Lisbon, Lisbon, Portugal

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on April 2, 2015; revised on June 26, 2015; accepted on September 3, 2015

Abstract

Motivation: One of the most widely used models to analyse genotype-by-environment data is the additive main effects and multiplicative interaction (AMMI) model. Genotype-by-environment data resulting from multi-location trials are usually organized in two-way tables with genotypes in the rows and environments (location-year combinations) in the columns. The AMMI model applies singular value decomposition (SVD) to the residuals of a specific linear model, to decompose the genotype-by-environment interaction (GEI) into a sum of multiplicative terms. However, SVD, being a least squares method, is highly sensitive to contamination and the presence of even a single outlier, if extreme, may draw the leading principal component towards itself resulting in possible misinterpretations and in turn lead to bad practical decisions. Since, as in many other real-life studies the distribution of these data is usually not normal due to the presence of outlying observations, either resulting from measurement errors or sometimes from individual intrinsic characteristics, robust SVD methods have been suggested to help overcome this handicap.

Results: We propose a robust generalization of the AMMI model (the R-AMMI model) that overcomes the fragility of its classical version when the data are contaminated. Here, robust statistical methods replace the classic ones to model, structure and analyse GEI. The performance of the robust extensions of the AMMI model is assessed through a Monte Carlo simulation study where several contamination schemes are considered. Applications to two real plant datasets are also presented to illustrate the benefits of the proposed methodology, which can be broadened to both animal and human genetics studies.

Availability and implementation: Source code implemented in R is available in the [supplementary material](#) under the function `r-AMMI`.

Contact: paulocanas@gmail.com

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Multi-environment trials (MET), which comprise experiments across multiple locations, are important tools for testing both broad and narrow genotype adaptation. Here, when two different genotypes show a differential response to a prototypic trait (e.g. yield) across environments, it is said that genotype-by-environment interaction (GEI) is present. Data from MET are often summarized in

two-way tables of means with genotypes in the rows (columns) and environments (location-year combinations) in the columns (rows).

The additive main effects and multiplicative interaction (AMMI) model (Gauch, 1988, 1992) is one of the most widely used tools to analyse and structure GEI. This model works under a fixed-model framework and is fit in two stages. First, the main effects of the model are estimated using the additive two-way analysis of variance

(ANOVA) by least squares. Then, the singular value decomposition (SVD) is applied to the residuals from the ANOVA, i.e. to the interaction, to obtain the estimates for the multiplicative terms of the AMMI model. The AMMI model, in its standard form, implicitly assumes equal weights for all entries of the two-way dataset and that no outliers (leverage points) are present in the data. [Rodrigues et al. \(2014\)](#) proposed a weighted AMMI algorithm where different weights to columns and/or rows and/or particular cells of the two-way data table are given, and used this generalization to study GEI and quantitative trait loci (QTL)-by-environment interactions. However, their approach does not fully solve the problem that arises from contaminated data—a problem that is common in field data such as data resulting from MET—because of the difficulty in defining the weights associated with possible outlying observations. As a consequence, the results from the analysis using standard methods may be biased leading to possible misinterpretations, which in turn may result in bad practical decisions.

Since the reason behind these biases lies on the fact that: (i) least-squares regression/ANOVA method does not perform well in the presence of outlying observations, which often tend to pull the regression line towards them and consequently compromise inference results; (ii) SVD is a technique essentially based on the computation of eigenvalues and eigenvectors of the sample covariance/correlation matrix, and is thus highly sensitive to the presence of atypical observations; it seems natural that an upgrade of this method may be achieved by improving the performance of the intermediate ANOVA and SVD steps.

Robust statistical methods are designed to overcome the handicaps of traditional methods that are felt when one or more assumptions of the model of analysis are not fulfilled. These violations often translate in biased parameter estimation, underpowered association tests, unreliable confidence intervals and so forth ([Huber, 1981](#)). Therefore, and since data contamination is more a rule than an exception when real-life data are considered, there has been growing interest in the use of these statistical methodologies which allow for valid results even if model assumptions are violated ([Copt and Heritier, 2007](#); [Lourenço et al., 2011](#)). Interest that is also explained by the fact that throughout the years the associated computational burden has been substantially reduced.

To overcome the problem of analysing two-way contaminated data and to better understand GEI, we introduce in this study a robust AMMI model where the linear fit, underlying the ANOVA method, is replaced by a robust fit (M-regression; [Huber, 1964](#)) and the use of the standard SVD by a robust SVD approach (see a review in [Filzmoser and Todorov, 2013](#)). We underline that in the AMMI model the main effects are considered to be fixed and the errors normally distributed, although other extensions have already been proposed in the literature ([Piepho, 1997](#); [Smith et al., 2001](#); [Tsujitani, 1992](#)). [Hadi \(2011\)](#) also proposed an alternative procedure of handling outliers in two-way tables by using robust alternating regression, based on factor analytic ANOVA models and robust projection pursuit principal component analysis, to ‘robustify’ the AMMI model. However, instead of dealing with a two-stage model (first fitting the additive part of the AMMI model and then the multiplicative), [Hadi \(2011\)](#) estimates the additive and multiplicative terms simultaneously.

The performance of the proposed robust extensions of the AMMI model is assessed through Monte Carlo simulations where several contamination schemes are considered. The methods are also compared with the results obtained via the classical AMMI model according to several criteria. Here, the good features of the R-AMMI model are displayed. An application of the methodology

to two real datasets is also presented to illustrate the benefits of this approach.

We highlight that the range of possible applications of a robust AMMI model is wide, spanning from plant genetics such as QTL detection and QTL-by-environment interaction studies ([Gauch et al., 2011](#); [Rodrigues et al., 2014](#); [Romagosa et al., 1996](#)), to microarray experiments ([Crossa et al., 2005](#)), ribosomal DNA studies ([Adams et al., 2002](#)), plant and microbial populations’ growth across several environmental conditions ([Culman et al., 2009](#)), animal sciences ([Barhadi and Dube, 2010](#)) and human genetics ([Mukherjee et al., 2012](#)).

2 Materials and methods

2.1 AMMI model

The AMMI model combines the features of ANOVA and SVD as follows: first the ANOVA estimates the additive main effects of the two-way data table; then the SVD is applied to the residuals from the additive ANOVA model, estimating $N \leq \min(I - 1, J - 1)$ interaction principal components (IPCs). Here, I represents the number of genotypes (rows) and J the number of environments (columns) considered in the study and described in the two-way data table. Assuming for simplicity a completely randomized design for individual trials, the model can be written as ([Gauch, 1992](#)):

$$y_{i,j,k} = \mu + \alpha_i + \beta_j + \sum_{n=1}^N \lambda_n \gamma_{i,n} \delta_{n,j} + \rho_{i,j} + \epsilon_{i,j,k}, \quad (1)$$

where $y_{i,j,k}$ is the phenotypic trait (yield or some other quantitative trait of interest) of the i th genotype in the j th environment for replicate k ; μ is the grand mean; α_i are the genotype deviations from μ ; β_j are the environment deviations from μ ; λ_n is the singular value of the IPC analysis axis n ; $\gamma_{i,n}$ and $\delta_{n,j}$ are the i th and j th genotype and environment IPC scores (i.e. the left and right singular vectors, scaled as unit vectors) for axis n , respectively; $\rho_{i,j}$ is the residual containing all multiplicative terms not included in the model; $\epsilon_{i,j,k}$ is the experimental error; and N is the number of principal components retained in the model.

In matrix formulation the AMMI model can be written as:

$$\mathbf{Y} = \mathbf{1}_I \mathbf{1}_J^T \mu + \alpha_I \mathbf{1}_J^T + \mathbf{1}_I \beta_J^T + \mathbf{U} \mathbf{D} \mathbf{V}^T + \epsilon, \quad (2)$$

where \mathbf{Y} is the $(I \times J)$ two-way table of genotypic means across trials or environments. Each column of \mathbf{Y} represents the vector of genotypic means as obtained from the phenotypic analysis of a corresponding trial by an appropriate mixed model analysis that accounts for experimental design features and/or spatial trends. $\mathbf{1}_I \mathbf{1}_J^T \mu$ is a $(I \times J)$ matrix with the grand mean μ in all positions, $\alpha_I \mathbf{1}_J^T$ is a $(I \times J)$ matrix of genotype main effects (equal rows), and $\mathbf{1}_I \beta_J^T$ is a $(I \times J)$ matrix of environmental main effects (equal columns). The interaction part of the model $\mathbf{Y}^* = \mathbf{Y} - \mathbf{1}_I \mathbf{1}_J^T \mu - \alpha_I \mathbf{1}_J^T - \mathbf{1}_I \beta_J^T$ is approximated by the product of matrices $\mathbf{U} \mathbf{D} \mathbf{V}^T$, with \mathbf{U} an $(I \times N)$ matrix whose columns contain the left singular vectors of the interaction, \mathbf{D} a $(N \times N)$ diagonal matrix containing the singular values of \mathbf{Y}^* , and \mathbf{V} a $(J \times N)$ matrix whose columns contain the right singular vectors of \mathbf{Y}^* . The residual term in Equation (2), the $(I \times J)$ matrix ϵ , includes both the lack of fit term and the error term of model (1).

The choice of the number of multiplicative terms in the model, N , is of key importance and must be done carefully as it will affect the subsequent results ([Gauch, 2013](#)). [Dias and Krzanowski \(2003, 2006\)](#) generalized a cross-validation-based method originally developed for principal component analysis ([Eastment and](#)

Krzanowski, 1982) to be used in the AMMI model. A more recent approach on how to choose the number of components in the AMMI model was proposed by Forkman and Piepho (2014), which was based on parametric bootstrap resampling.

2.2 Robust AMMI model

To overcome the problem of data contamination with outlying observations, we propose a robust AMMI model, which can be obtained in two stages as follows: (i) use the robust regression based on the M-Huber estimator (Huber, 1981) to replace the ANOVA model; and (ii) use a robust SVD/PCA procedure to replace the standard SVD. In the second stage, we will consider several methods, briefly described below, thus conveying a total of five robust AMMI candidate models:

R-AMMI: robust SVD (Hawkins et al., 2001): uses the L1 norm instead of the more usual least squares L2 norm, to compute a robust approximation to the SVD of a rectangular matrix. This method is not robust in the presence of leverage points.

H-AMMI: robust PCA (Hubert et al., 2005): Combines projection-pursuit (PP) and robust covariance estimation (minimum covariance determinant) techniques to compute the robust loadings. It is most adequate for high-dimensional data.

G-AMMI: robust grid algorithm (Croux et al., 2007): Uses PP to compute PCA estimators. The optimization is done via the grid search algorithm in the plane instead of the p-dimensional space.

L-AMMI: robust spherical PCA (Locantore et al., 1999): Uses a spherical principal components procedure. The idea behind this approach is to perform classical PCA on the data but projected onto a unit sphere. When the data are elliptically distributed the estimates of the eigenvectors are consistent. In addition, this is an extremely fast procedure.

PP-AMMI: robust PP (Croux and Ruiz-Gazen, 2005): Uses PP calculating the robust eigenvalues and eigenvectors without going through robust covariance estimation. The principal components can be sequentially computed and thus this method is very appealing in $p \gg n$ problems, which, in the case of MET and AMMI modelling, are found when few genotypes are evaluated under a wide range of environmental and/or experimental conditions.

The robust methods described are available in the R software in packages MASS, pcaMethods and rrcov via functions `rlm()`, `robustSVD()`, `PcaHubert()`, `PcaGrid()`, `PcaLocantore()` and `PcaProj()`, respectively. More details on the methods can be found in Maronna (2005), Todorov and Filzmoser (2009) and Filzmoser and Todorov (2013).

3 Simulation study

We simulate $m = 1000$ two-way data tables with $n = 100$ rows/genotypes and $p = 8$ columns/environments each, where the interaction is explained by two multiplicative terms (i.e. two IPCs; $k = 2$ components to be retained). Without loss of generality, the two-way data tables are simulated in the following way:

- Create a matrix \mathbf{X} with $n = 100$ rows/genotypes and $p = 8$ columns/environments with observations drawn from a $Unif[-0.5, 0.5]$ distribution;
- Do the SVD of \mathbf{X} and obtain the matrices \mathbf{U} , \mathbf{V} and \mathbf{D} , containing, respectively, the left and right singular vectors and the singular values of \mathbf{X} ;
- Simulate the grand mean, the genotypic and environmental main effects, considering: $\mu \sim N(15, 3)$, $\alpha \sim N(5, 1)$ and $\beta \sim N(8, 2)$;

- Simulate a two-way data table \mathbf{Y} with AMMI2 structure: $\mathbf{Y} = \mathbf{1}_n \mathbf{1}_p^T \mu + \alpha_i \mathbf{1}_n^T + \mathbf{1}_n \beta_j^T + 28 \times \mathbf{U}[:, 1] \mathbf{D}[1, 1] \mathbf{V}[:, 1]^T + \mathbf{1}_n \beta_j^T + 15 \times \mathbf{U}[:, 2] \mathbf{D}[2, 2] \mathbf{V}[:, 2]^T$, where $\mathbf{A}[:, i]$ represents the i th, $i = 1, 2$, column of the matrix \mathbf{A} and $\mathbf{A}[i, j]$ represents the cell in the row i and column j , $i = 1, 2$, of the matrix \mathbf{A} .

The number of multiplicative terms is confirmed by the cross-validation procedure generalized by Dias and Krzanowski (2003, 2006) and by the parametric bootstrap-based method proposed by Forkman and Piepho (2014).

Having the good non-contaminated data, a percentage of contamination is introduced in the two-way original data table so as to be consistent with the known shift and point mass outliers described in Rocke and Woodruff (1996): the bad data are generated from the normal distributions $N(\mu_j + k\sigma_j^2, \sigma_j^2)$ (pure shift outliers), $N(\mu_j, \sigma_j^2/100)$ (pure point-mass outliers) and $N(\mu_j + k\sigma_j^2, \sigma_j^2/100)$ (shift-point-mass outliers) where $k = 4, 5, 6, 7$ and 10 units, and μ_j and σ_j^2 are taken as the sample phenotypic mean and sample phenotypic variance according to the correspondent environment j . The bad data replace a percentage of the good data from the two-way table at the assigned positions for three scenarios:

- positions randomly selected in the two-way table thus randomly assigning contamination in different environments for distinct genotypes (2, 5 & 10% *scattered* contamination);
- positions randomly assigned in only one of the environments (5, 10, 15 & 20% *single-environment* contamination); and
- a mix between (i) and (ii), i.e. one of the environments is assigned a percentage of contamination (5, 10, 15 & 20%) and then scattered contamination is considered in the rest of the two-way table, excluding this environment and up to a given total percentage (2, 5 & 10%)—a more realistic scenario (*mixed* contamination).

Method comparison is achieved using four criteria:

- the mean percentage of explained variability (MPEV)

$$\frac{1}{m} \sum_{l=1}^m \frac{\hat{\lambda}_1^{(l)} + \hat{\lambda}_2^{(l)} + \dots + \hat{\lambda}_k^{(l)}}{\lambda_1 + \lambda_2 + \dots + \lambda_p}, \quad (3)$$

- the mean squared error (MSE; Hubert et al., 2005)

$$\text{MSE}(\hat{\lambda}_j) = \frac{1}{m} \sum_{l=1}^m (\hat{\lambda}_j^{(l)} - \lambda_j)^2, \quad (4)$$

where λ_j , $j = 1, \dots, m$, are the true eigenvalues of the two-way data tables (benchmark values considering no outliers in the data), and the $\hat{\lambda}_j^{(l)}$, $j = 1, \dots, m$, are the estimated eigenvalues for each of the replications (obtained under contamination);

- the *maxsub* value (Krzanowski, 1979)

$$\text{maxsub} = \arccos(\sqrt{\lambda_k}) \quad (5)$$

where λ_k is the smallest eigenvalue of $\mathbf{I}'_{k,p} \mathbf{P}_{p,k} \mathbf{P}'_{k,p} \mathbf{I}_{p,k}$, being $\mathbf{I}_{p,k}$ the true orthogonal loadings and $\mathbf{P}_{p,k}$ the estimated loadings. The optimal *maxsub* angle is 0 and it is usual to standardize this value by dividing it by $\frac{\pi}{2}$;

- Biplot interpretation—assessed via a numerical example.

In each run of the simulation, the AMMI and robust AMMI models are used to analyse the data. The best performing method is the one that: (i) minimizes the MSE (meaning that the estimated eigenvalues are closer to the real ones); (ii) gives values of MPEV closest to 100% (the least overestimation the better); and (iii) gives the lowest *maxsub* values (meaning that the estimated components are correct). The best performing methods according to

items (i) to (iii) will then also be compared in terms of the biplots they produce. Biplots more similar, in terms of the position of genotypes (scores) and environments (loadings), to the true ones (i.e. obtained from the data without contamination) are preferred.

Henceforward the notation 1S, 2S, ... will refer to [Supplementary figures and/or tables](#).

4 Results

4.1 Simulation study

[Table 1](#) presents the MSE for the first two IPCs (IPC1 and IPC2), considering the true values those obtained by the standard AMMI model when there is no data contamination. It can be observed that, with the exception of the R-AMMI method, all other robust options show a high MSE for IPC1. As for the IPC2, the methods R-AMMI, H-AMMI and L-AMMI perform rather well. Globally, we conclude that the R-AMMI shows the best performance in terms of MSE when there are no outliers in the data.

In terms of the mean proportion of explained variability, when there is no contamination in the data, the R-AMMI is the most accurate with H-AMMI underestimating the true proportion of explained variability by around 14% and the other overestimating this proportion by more than 15% ([Table 2](#), first row). Regarding the observed *maxsub* values, under the null, R-AMMI attained the

highest median values amongst all methods (0.27; [Fig. 1](#)) thus being the less efficient method.

4.1.1 Pure shift outliers

The observed MSE ratios (classic AMMI MSE over robust AMMI MSE; the greater the ratio the smaller the MSE for the robust method) for the scattered contamination scenarios are displayed in [Figure 2](#). We see that the performance of the R-AMMI model is always better than the other robust alternatives for the IPC1, which is the most relevant component. With one exception (10%, $k = 10$), this is also true in the single-environment contamination scenarios ([Supplementary Fig. 1S](#)). As to the MSE ratios for the IPC2 term, H-AMMI and L-AMMI take the lead with interchangeable performance ([Fig. 2](#) and [Supplementary Fig. 1S](#)). In the case of the mixed scenario ([Supplementary Figs 2S–4S](#)), the performance of the robust R-AMMI, regarding the IPC1 MSE, improves with the increase of total scattered contamination, being the best performing method when a total scattered percentage of contamination of 10% is considered ([Supplementary Fig. 4S](#)). R-AMMI is only overcome by the H-AMMI in the 2% scattered contamination scenario ([Supplementary Fig. 2S](#)) with shift $k = 10$ (all cases) and shift $k = 7$ (15% single-environment contamination only) and in the 5% scattered contamination scenario ([Supplementary Fig. 3S](#)) with shift $k = 10$ (5% single-environment contamination excluded). As to the IPC2, H-AMMI performs best for low levels of total scattered contamination (2 & 5%; [Supplementary Figs 2S](#) and [3S](#)) whereas L-AMMI does better in the total 10% scattered contamination scenarios ([Supplementary Fig. 4S](#)).

With regard to the MPEV values, the R-AMMI model provides overall the best results in the scattered contamination scenarios, being worse than the H-AMMI only in the case of lower percentages of contamination and larger shifts ([Table 2](#) and [Supplementary Table 1S](#)). When considering single-environment contamination, the H-AMMI is in general the best performing method being outperformed by R-AMMI for smaller shifts and lower percentages of contamination ([Table 2](#) and [Supplementary Table 2S](#)). In this scenario of single-environment contamination, H-AMMI performs reasonably well not overestimating the MPEV 100% threshold by >13% in any of the considered scenarios. In the mixed scenarios, the R-AMMI model is the one that provides more acceptable values,

Table 1. MSE for IPCs 1 and 2 when there is no data contamination

Component	R-AMMI	H-AMMI	G-AMMI	L-AMMI	PP-AMMI
IPC1	29.3	768.5	666.1	459.8	717.5
IPC2	29.2	41.4	101.9	26.9	84.3

Table 2. Mean proportion of explained variability when there is no data contamination and in the presence of scattered and single-environment contamination (shift outliers)

	%	AMMI (%)	R-AMMI (%)	H-AMMI (%)	G-AMMI (%)	L-AMMI (%)	PP-AMMI (%)
Shift	0	100.0	100.1	85.8	126.1	115.7	126.9
Scattered							
k = 4	2	111.9	100.2	105.5	141.8	122.9	136.1
	5	131.4	103.3	132.0	165.3	135.5	151.6
	10	164.8	111.9	167.7	207.2	158.2	179.9
k = 7	2	147.7	110.8	106.3	154.7	131.8	147.3
	5	222.7	132.2	144.8	204.8	160.9	182.6
	10	341.9	176.4	244.2	317.2	221.1	262.8
k = 10	2	210.5	130.3	104.9	162.6	137.6	154.0
	5	373.4	187.3	145.6	234.9	179.8	207.9
	10	624.7	284.9	338.7	458.9	280.3	356.2
Single-environment							
k = 4	5	104.7	100.2	92.4	131.3	118.1	130.5
	10	110.9	101.6	98.1	136.6	120.8	135.1
	15	117.5	103.2	102.7	142.9	124.8	140.3
k = 7	20	123.9	106.6	110.9	151.0	129.6	148.0
	5	121.8	103.5	92.0	135.1	121.0	133.9
	10	148.7	114.3	98.9	145.6	127.7	143.3
k = 10	15	173.2	128.3	103.6	159.9	137.4	156.5
	20	195.1	148.2	111.7	179.6	150.9	175.4
k = 10	5	153.6	110.2	92.4	137.1	122.8	135.8
	10	211.7	145.6	99.5	151.0	132.0	148.4
	15	263.2	187.3	104.8	169.2	144.8	166.2
	20	308.6	233.2	111.4	195.5	162.6	192.4

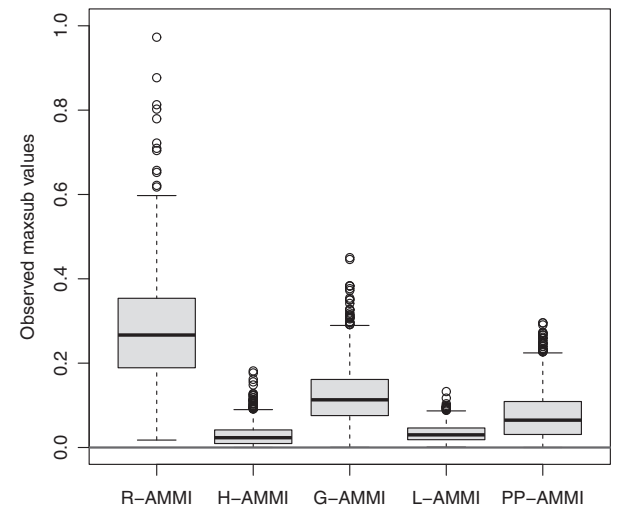


Fig. 1. Observed *maxsub* values under no contamination

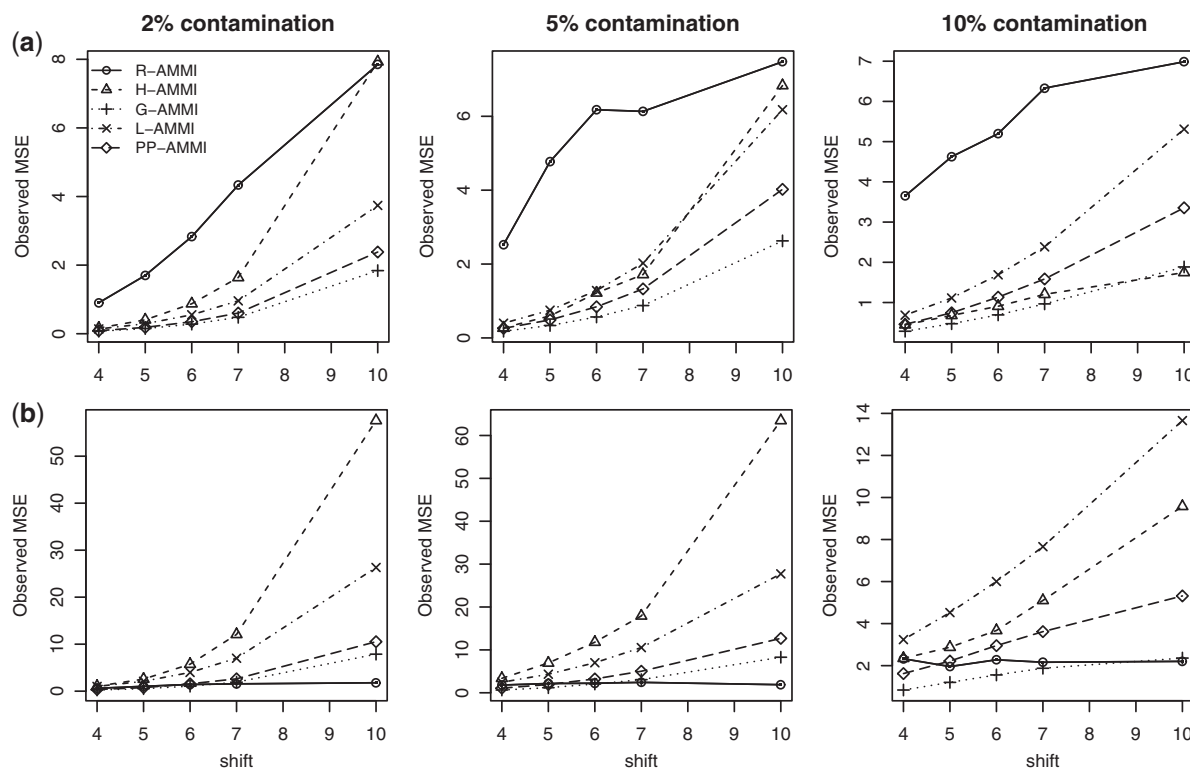


Fig. 2. Observed MSE ratios under scattered contamination for both the first (first row) and second (second row) principal components (shift outliers)

being only surpassed by H-AMMI for lower percentages of contamination and larger shifts (2% with $k=7$; 2 & 5% with $k=10$; Supplementary Table 3S). As to the standard AMMI method, it progressively performs worse and worse as the degree of shift increases. At the $k=7, 10$ shift scenarios it is the poorest performing method (Supplementary Table 3S).

In order to compare the *maxsub* values for the proposed methods, boxplots (based on the 1000 simulations) were obtained for every combination (shift, percentage of contamination, type of contamination). In general, L-AMMI consistently gives the lowest *maxsub* values across scattered contamination scenarios (Fig. 3 and Supplementary Fig. 5S). Even in the worst case scenario (10% shift-10 scattered contamination) its observed median values stay under 0.2. The H-AMMI model gives the overall second best results, being very similar to L-AMMI for lower percentages of contamination but deteriorating with the increase of both the degree of shift and the percentage of contamination (Supplementary Fig. 5S). As for R-AMMI, it is clearly the worst performing method, even worse performing than AMMI in most cases.

In the case of single-environment contamination scenarios, H-AMMI is the best performing method keeping the observed *maxsub* values close to 0 across shifts and levels of contamination (Supplementary Figs 6S and 7S). Here L-AMMI performs as good as H-AMMI for lower percentages of contamination (Supplementary Fig. 6S) but does not keep up with the former when the percentage of contamination increases (Supplementary Fig. 7S). R-AMMI performs in general better than AMMI but still with poor results when compared with the other robust counterparts.

With respect to the mixed scenarios, Supplementary Figures 8S–12S show that L-AMMI performs better than the other robust methods across all shifts and single-environment contamination levels in the case of 5 & 10% overall contamination, giving place to H-AMMI in the 2% overall contamination scenario but with small

differences. In general, in this setting both AMMI and R-AMMI models give the highest *maxsub* values.

4.1.2 Pure point-mass outliers

We observe through the MSE ratio plots (Supplementary Fig. 13S) that R-AMMI provides the smallest observed MSEs amongst all robust methods for the IPC1 in the single-environment scenario and also in the scattered one but for lower levels of contamination. With respect to the IPC2, L-AMMI produces, in general, the highest ratios and thus the lowest MSE values. However, R-AMMI is the second best performing method in terms of observed MSEs for the IPC2 with values quite close to those of L-AMMI. When mixed contamination is considered, Supplementary Figure 14S shows that R-AMMI is the best method in the case of 2 & 5% scattered contamination for the IPC1, across all single-environment contamination levels, and the second best for the IPC2 (10% scattered with 20% single-environment contamination scenario excluded). In the 10% scattered contamination scenario however, the best results for IPC1 are obtained by PP-AMMI. The L-AMMI model shows the best performance regarding the MSE in the case of IPC2.

Supplementary Table 4S shows that, when considering both scattered and single-environment contamination, R-AMMI keeps values of MPEV close to the ones obtained with the AMMI method, always under 100%, whereas all the other methods tend to overestimate this value across all contamination levels.

In the case of the mixed scenario, the classic AMMI and the R-AMMI methods perform similarly (Supplementary Table 5S), keeping the MPEV values under 100% as desired, in contrast to the remaining methods (except for L-AMMI in the case of 10% scattered contamination) which tend to overestimate the 100% threshold. In the case of the 10% scattered contamination scenario L-AMMI is the best performing method.

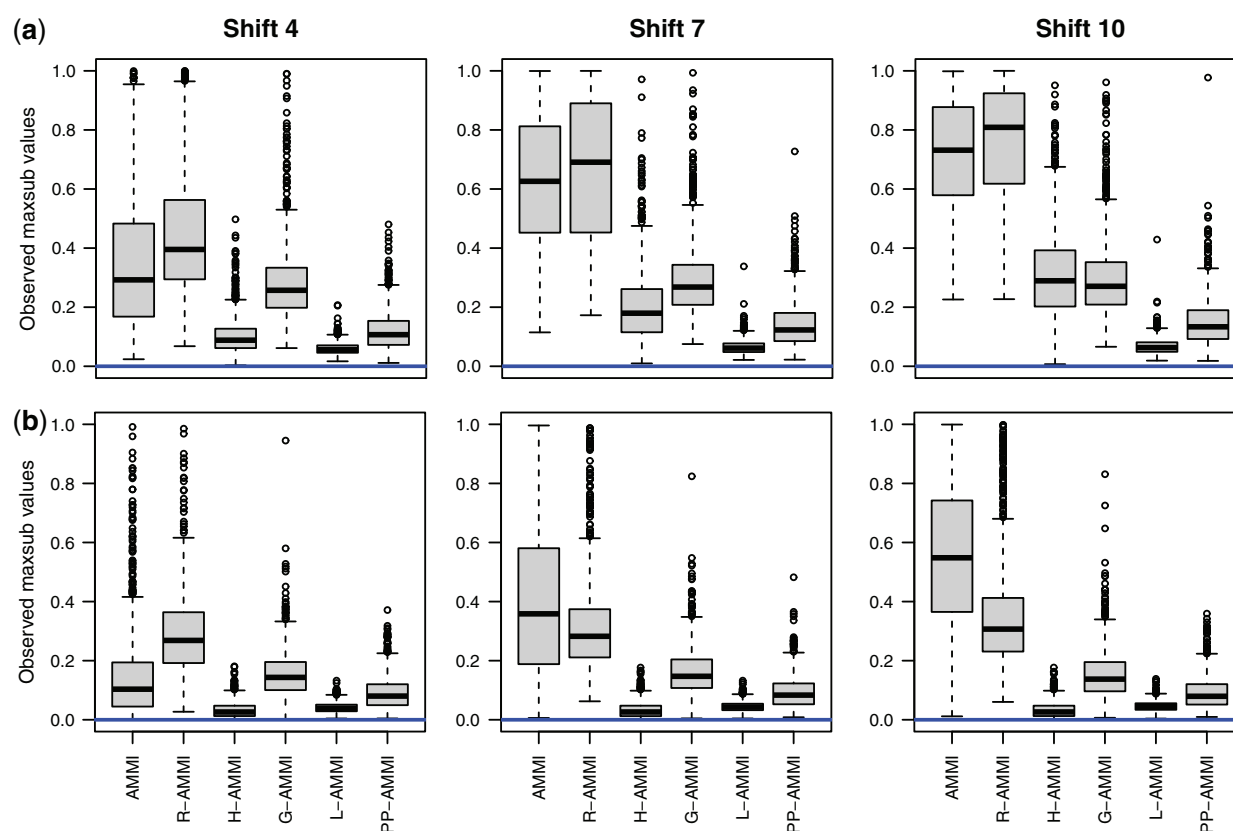


Fig. 3. Observed *maxsub* values under 5% scattered (first row) and single-environment (second row) contamination (shift outliers)

When analyzing the observed *maxsub* values, the best performance is shared by the classic AMMI and the robust H-AMMI and L-AMMI models for all three contamination scenarios (Supplementary Figs 15S–17S). This means that the classic AMMI model is quite unaffected by this type of contamination. Here, R-AMMI has the poorest performance with an overall registered median maximum of ~ 0.28 (Supplementary Figs 15S–17S).

4.1.3 Shift-point-mass outliers

The results regarding the shift plus point mass contamination scenarios did not differ much from the ones obtained in the shift contamination settings and are, therefore, not shown.

4.2 Numerical Biplot example

In order to illustrate the differences between the methods in terms of biplot interpretation, we now consider one dataset from a random run of the previous 1000 simulations. Also, only the classic AMMI model and the two overall best robust models, R-AMMI and H-AMMI, are contemplated in this numerical example. To make the example more realistic, we allocate to the mixed contamination scenario with 10% 5-shift single-environment contamination for a total of 5% contamination, also 5-shift scattered contamination (*mild* contamination in a reasonable amount). This particular situation is usually seen in real field data where a small amount of measurement errors may occur and a small amount of anomalous observations may be observed. These anomalous observations, for instance, when concentrated in a specific-environment, could be the outcome of, e.g. a virus having somehow affected a small part of that environment, and represent low yield values for a given number of individuals/plants.

Supplementary Figure 18S shows the AMMI, R-AMMI and H-AMMI biplots considering two IPCs, obtained for the simulated data set with and without contamination. For the non-contaminated data (top plots of Supplementary Fig. 18S), as expected, the scores for the genotypes and loadings for the environments are located in similar places of the biplots. However, when 5% contamination is considered, the display of genotypes and environments shows a completely different behaviour for the classic AMMI biplot. When considering the robust models to fit the contaminated data, the improvement (results more similar to the biplots for the non-contaminated data) is clear, being the biplot for R-AMMI much more similar to the non-contaminated data. Here, for the R-AMMI model, the impact of the outlying observations is reduced and the position of the scores and loadings become similar to the ‘true’ positions given by the AMMI model without contamination. These same patterns can be observed for the scenarios where scattered (Supplementary Fig. 19S) and mixed (Supplementary Fig. 20S) contamination are considered. When considering single environment contamination, the methods are stable in terms of the biplots they produce (Supplementary Fig. 21S). However, for higher degrees of shift, differences become visible in the AMMI biplot (e.g. shift 10; Supplementary Fig. 21S).

4.3 Real data examples

Dataset 1: Steptoe \times Morex barley mapping population

One of the real datasets used in this study is the Steptoe \times Morex (S \times M) barley mapping population produced by the North American barley genome mapping project (Hayes *et al.*, 1993), comprising 150 doubled haploid genotypes evaluated in 16 trials, or

environments, during 1991 and 1992, in the United States and Canada. Following Rodrigues *et al.* (2014), two multiplicative terms were considered in the AMMI model. Supplementary Figure 22S shows the biplots obtained with the AMMI and R-AMMI (the robust model which performed better in the numerical example) approaches considering two IPCs for the analysis.

In the biplot of the classic AMMI model (left-hand side of Supplementary Fig. 22S) the environment OR91 shows a dominant effect over the biplot being non-correlated with most of other environments. This biplot also presents an overlap in the direction of many of the loadings for the environments which makes it difficult to interpret and to analyse. When considering the R-AMMI methodology (right-hand side of Supplementary Fig. 22S) the interpretation of the biplot seems easier, with the environments more spread, different angles between their components loadings and without such dominant influence of a given environment. These results are consistent with Gauch *et al.* (2011) where the environment OR91 was considered to be an outlying environment after conducting a QTL analysis. Moreover, the use of the R-AMMI model made it possible for this particular environment to be included in the analysis without distorting the final results, which was achieved by reducing its influence on the final model. To conclude, this example application further reinforces the usefulness of the proposed methodology.

Dataset 2: Durum wheat mapping population

The second real dataset used in this study is the result of a breeding programme carried out by the Portuguese National Plant Breeding Station in the years of 1992/1993 and 1993/1994 and includes nine genotypes and 11 environments. Following the results of Rodrigues *et al.* (2011), two multiplicative terms were considered to fit the AMMI and R-AMMI models. Supplementary Figure 23S shows the biplots obtained with the two approaches considering the first two IPCs for the analysis. The biplots do not show major differences between themselves. This can be a result of the fact that in this particular study, genotypes with proven strong adaptability were used. As such, no major differences between yields are seen and thus there are no anomalous observations in this sense. Also, this dataset probably does not have a significant number of measurement errors.

5 Discussion

This study outlines the fragility of the classical AMMI model when the two-way data table is contaminated, which often is the case in MET, either, because measurement errors have occurred, some pest/disease have influenced some genotypes in a given environment resulting, e.g. in a lower-than-expected yield, or even due to some inherent characteristics of the genotypes being measured. Although we have considered in the simulation study some extreme contamination scenarios for illustration purposes alone, even in the more plausible contamination scenarios the classical methodology resulted in higher values of MSE and MPEV due to the overestimation of the eigenvalues, greater *maxsub* values and, as seen in the numerical example, biplots harder to interpret. Here, the use of the robust R-AMMI method, besides providing results similar to the classical AMMI model when there was no data contamination (*maxsub* values excluded; observed values not critical), it also gave overall better results when the data were contaminated (*maxsub* values again excluded although very high values were observed in many situations); least biased estimated eigenvalues (i.e. in general lower IPC1 MSE values), less overestimation of the true mean proportion of

explained variability and more parsimonious biplots. We underline that the high observed *maxsub* values of R-AMMI are not at all critical since *maxsub* only assesses subspaces and is only a good measure of comparison when the goal is the selection of the components, which is not the case in this study. The R-AMMI direct competitor in the simulations, H-AMMI, although giving low *maxsub* values, showed high IPC1 MSE values even under the assumption of no contamination, greater tendency to overestimate the MPEV 100% threshold (scattered and mixed scenarios mainly) and failed to provide easy interpretable biplots in many situations.

Despite the fact that this simulation study was made for the AMMI model with two multiplicative terms for simplicity, the results would be similar for a more complex model such as the AMMI3 or AMMI4 models as all the robust procedures allow for more components than two. In practice, the main difference would be the biplot comparison which has a more complex structure (e.g. two biplots for the AMMI3: (i) first and second components; and (ii) first and third components) for the AMMI model with more multiplicative terms. In those cases, the best way to compare the methods would be to use the MSE of the predicted yield values and, in the case of the biplots, either the Procrustes statistic (Dryden and Mardia, 1998) or the root mean square statistic proposed by Shin *et al.* (2008). For real data examples with more than two components, the strategy for the robust AMMI would be the same as for the AMMI model, i.e. to make the comparisons based on the sum of squares provided by the ANOVA tables, based on the model predicted values, and based on biplots with pairs of multiplicative terms.

The employ of the robust version of the AMMI model may be extremely useful for practitioners because a bad representation of genotypes and environments in the biplots (as happens in the Steptoe and Morex study—Supplementary Fig. 22S) can result in a bad decision regarding which genotypes to choose for a given set of environments (i.e. mega-environments; Gauch and Zobel, 1997; Yan *et al.*, 2000). In turn, the choice of the wrong genotypes for the environments due to data contamination might result in big losses in terms of yield production—a problem that also occurs when the choice of the number of IPCs to consider in the AMMI model is not properly chosen (Forkman and Piepho, 2014).

Although the robust AMMI model presented in this article has shown to provide very good results and thus has a huge potential, it is not suitable to analyse every two-way data table. For instance, when the data have only significant main effects (i.e. no interaction), the interaction should be ignored and the family of AMMI models is not useful. When data are replicated and some summary measure such as error variance for the environments can be computed, the weighted AMMI algorithm proposed by Rodrigues *et al.* (2014) might be preferable as the weights can be obtained *a priori*. However, when there is interaction in the data, data contamination is present and there are difficulties in defining the weights associated with possible outlying observations, then the robust AMMI model should be used. It should also be noted that important features such as the ANOVA, showing the proportion of sums of squares due to the main effects, and each successive IPCs are possible to obtain for the robust AMMI model. Moreover, the robust biplots will keep the characteristics and standard interpretation as the ones produced via the AMMI model. Despite all the advantages of the robust AMMI model, as in any other data analysis strategy, caution should be exercised when analysing the data. In particular, cleaning detectable measurement errors from the data and a preliminary descriptive analysis should be made. These missing points can then be imputed before the modelling (Paderewski and Rodrigues, 2014). As an

example of the usefulness of the preliminary descriptive analysis, one can think of the case where a regional trial has both irrigated and no-irrigated plots, which can result in wide yield differences within an environment. In this case it might be more appropriate to analyse such a yield trial as two separate subsets instead of combining all data in one single analysis.

The proposed robust R-AMMI model may also be useful in other studies where data contamination is inevitable, e.g. in QTL detection and QTL-by-environment interaction studies. In that case, the robust AMMI model may be used, instead of its classical counterpart, to calculate more accurate predicted values when fitting the model which can then be subject to a QTL analysis in a two-stage procedure, similar to the ones described in Romagosa *et al.* (1996), Gauch *et al.* (2011) and Rodrigues *et al.* (2014). Further comparisons in terms of QTL detection and QTL-by-environment interaction, between standard (Gauch *et al.*, 2011), weighted (Rodrigues *et al.*, 2014) and robust AMMI models, should be done in future work to reinforce the usefulness of the robust AMMI model. Also, the ideas behind this robust approach may also be incorporated in the genotype main effects and genotype-by-environment interaction effects (GGE) model, which, as the AMMI model, is also very popular in MET studies (Yan *et al.*, 2000). The main difference between the two is that in the GGE model SVD is performed on the matrix of residuals from a one-way ANOVA with fixed effects for environments only instead of a two-way ANOVA with fixed effects of genotypes and environments. Further extensions and applications of the proposed methodology can be done, not only for animal and human genetics applications, but also in other fields as the AMMI model is also known in the literature as double-centered principal component analysis.

The use of this robust version of the AMMI model, and others that have shown preliminary good performances (Hadi, 2011) or that may subsequently follow [e.g. based on the methods of Jung (2010) and Zhang *et al.* (2013)], will allow practitioners to make better strategic decisions. Therefore, this discussion is another step forward in the direction of the development and use of more robust and thus more accurate methods, not only in this field of application but in all others where data fail to comply with classical model assumptions.

Acknowledgements

The authors thank Ana M. Pires, Valentin Todorov, Hans-Peter Piepho and two reviewers for providing helpful suggestions which contributed to the improvement of the article.

Funding

Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through the projects PTDC/MAT-STA/0568/2012 and UID/MAT/00297/2013 (Centro de Matemática e Aplicações to V.M.L.). Conselho Nacional de Desenvolvimento Científico e Tecnológico—“National Council of Technological and Scientific Development” grant code 448775/20140, and PRODC program at Federal University of Bahia, Brazil (to P.C.R.).

Conflict of Interest: none declared.

References

Adams, G.C. *et al.* (2002) Virulence and double-stranded RNA in *Sphaeropsis sapinea*. *For. Pathol.*, **32**, 309–329.
Barhdadi, A. and Dube, M.P. (2010) Testing for gene-gene interaction with AMMI models. *Stat. Appl. Genet. Mol. Biol.*, **9**, 2.

Copt, S. and Heritier, S. (2007) Robust alternatives to the F-Test in mixed linear models based on MM-estimates. *Biometrics*, **63**, 1045–1052.
Crossa, J. *et al.* (2005) Using linear-bilinear models for studying gene expression x treatment interaction in microarray experiments. *J. Agric. Biol. Environ. Stat.*, **10**, 337–353.
Culman, S.W. *et al.* (2009) T-REX: software for the processing and analysis of T-RFLP data. *BMC Bioinformatics*, **10**, 171.
Croux, C. and Ruiz-Gazen, A. (2005) High breakdown estimators for principal components: the projection-pursuit approach revisited. *J. Multivariate Anal.*, **95**, 206–226.
Croux, C. *et al.* (2007) Algorithms for projection-pursuit robust principal component analysis. *Chemomet. Intell. Lab. Sys.*, **87**, 218–225.
Dias, C.T.S. and Krzanowski, W.J. (2003) Model selection and cross validation in additive main effect and multiplicative interaction models. *Crop Sci.*, **43**, 865–873.
Dias, C.T.S. and Krzanowski, W.J. (2006) Choosing components in the additive main effect and multiplicative interaction (AMMI) models. *Scientia Agricola*, **63**, 169–175.
Dryden, I.L. and Mardia, K.V. (1998) *Statistical Shape Analysis*. Wiley, Chichester.
Eastment, H.T. and Krzanowski, W.J. (1982) Cross-validatory choice of the number of components from a principal component analysis. *Technometrics*, **24**, 73–77.
Filzmoser, P. and Todorov, V. (2013) Robust tools for the imperfect world. *Inf. Sci.*, **245**, 4–20.
Forkman, J. and Piepho, H.P. (2014) Parametric bootstrap methods for testing multiplicative terms in GGE and AMMI models. *Biometrics*, **70**, 639–647.
Gauch, H.G. (1988) Model selection and validation for yield trials with interaction. *Biometrics*, **44**, 705–715.
Gauch, H.G. (1992) *Statistical Analysis of Regional Yield Trials: AMMI Analysis of Factorial Designs*. Elsevier, Amsterdam.
Gauch, H.G. (2013) A simple protocol for AMMI analysis of yield trials. *Crop Sci.*, **53**, 1860–1869.
Gauch, H.G. and Zobel, R.W. (1997) Identifying mega-environments and targeting genotypes. *Crop Sci.*, **37**, 311–326.
Gauch, H.G. *et al.* (2011) Two new strategies for detecting and understanding QTL x environment interactions. *Crop Sci.*, **51**, 96–113.
Hadi, A.F. (2011) Handling outlier in two-way tables by robust alternating regression of FANOVA models: towards robust AMMI models. *J. ILMU DASAR*, **12**, 123–131.
Hawkins, D.M. *et al.* (2001) *Robust Singular Value Decomposition*. National Institute of Statistical Sciences, Technical Report Number 122.
Hayes, P.M. *et al.* (1993) Quantitative trait locus effects and environmental interaction in a sample of North American barley germplasm. *Theor. Appl. Genet.*, **87**, 392–401.
Huber, P.J. (1964) Robust estimation of a location parameter. *Ann. Math. Stat.*, **35**, 73–101.
Huber, P.J. (1981) *Robust Statistics*. Wiley, New York.
Hubert, M. *et al.* (2005) Robpca: a new approach to robust principal component analysis. *Technometrics*, **47**, 64–79.
Jung, K.-M. (2010) Robust singular value decomposition based on weighted least absolute deviation regression. *Commun. Kor. Stat. Soc.*, **17**, 803–810.
Krzanowski, W.J. (1979) Between-groups comparison of principal components. *J. Am. Stat. Assoc.*, **74**, 703–707.
Locantore, N. *et al.* (1999) Robust principal components for functional data. *Test*, **8**, 1–28.
Lourenço, V.M. *et al.* (2011) Robust linear regression methods in association studies. *Bioinformatics*, **27**, 815–821.
Maronna, R. (2005) Principal components and orthogonal regression based on robust scales. *Technometrics*, **47**, 264–273.
Mukherjee, B. *et al.* (2012) Principal interactions analysis for repeated measures data: application to gene-gene and gene-environment interactions. *Stat. Med.*, **31**, 2531–2551.
Paderewski, J. and Rodrigues, P.C. (2014) The usefulness of EM-AMMI to study the influence of missing data pattern and application to Polish post-registration winter wheat data. *Aust. J. Crop Sci.*, **8**, 640–645.
Piepho, H.P. (1997) Analyzing genotype-environment data by mixed models with multiplicative terms. *Biometrics*, **53**, 761–766.

- Rocke, D.M. and Woodruff, D.L. (1996) Identification of outliers in multivariate data. *J. Am. Stat. Assoc.*, **91**, 1047–1061.
- Rodrigues, P.C. et al. (2011) A comparison between JRA and AMMI: the robustness with increasing amounts of missing data. *Scientia Agricola*, **68**, 679–686.
- Rodrigues, P.C. et al. (2014) A weighted AMMI algorithm to study genotype-by-environment interaction and QTL-by-environment interaction. *Crop Sci.*, **54**, 1555–1570.
- Romagosa, I. et al. (1996) Use of the additive main effects and multiplicative interaction model in QTL mapping for adaptation in barley. *Theor. Appl. Genet.*, **93**, 30–37.
- Shin, S.M. et al. (2008) Comparison of shape variability in principal component biplot with missing values. *Kor. J. Appl. Stat.*, **21**, 1109–1116.
- Smith, A. et al. (2001) Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics*, **57**, 1138–1147.
- Todorov, V. and Filzmoser, P. (2009) An object oriented framework for robust multivariate analysis. *J. Stat. Softw.*, **32**, 1–47.
- Tsujitani, M. (1992) A note on the additive and multiplicative models in two-way contingency tables. *Biometrics*, **48**, 267–269.
- Yan, W. et al. (2000) Cultivar evaluation and mega-environment investigation based on the GGE biplot. *Crop Sci.*, **40**, 597–605.
- Zhang, L. et al. (2013) Robust regularized singular value decomposition with application to mortality data. *Ann. Appl. Stat.*, **7**, 1540–1561.