Robust modeling in statistical genetics

Marcelo B. Fonsêca[1], Vanda M. Lourenço[2], Paulo Canas Rodrigues[1]
[1] Federal University of Bahia, Brasil
[2] NOVA University Lisbon, Portugal

Abbreviated abstract: The main objective in plant breeding programs is the study and understanding of the interaction between genotypes and environment (GEI), which allows generating important results for the selection of genotypes with high productivity, production stability and adaptability in different environments. Simpler models such as the AMMI model does not perform well when the data is contaminated. In this work, we show that new weightings schemes further improve robust models.

Related publications:

– Gauch H. *et al*, Elsevier Science Publishers (1992)
– Rodrigues, P. C. et al, Bioinformatics 32 (1), 58-66 (2016)

Departamento de Estatística
IME · UFBA

4th Conference on
Statistics and
Data Science
Salvador, Brazil (online)
December 1-3, 2022

Problem and Previous Works

When the phenotypic data are contaminated with outliers or when the error variance of the environments is heterogeneous, simpler models such as the AMMI model do not produce reliable results for the analysis of the interaction because it takes into account that all the data have the same weights which the sometimes it is not an appropriate assumption to make.

• The weighted AMMI model was proposed for when the error variances are heterogeneous, so that the model takes into account that each set of genotype and environment can have a different weight, improving the interaction analysis.

• To work around problems with data that are contaminated by outliers, the robust AMMI model was proposed, which uses Huber's M estimation to give weights to each cell of the dataset, but depending on the dataset, the estimation can take a lot of computational time.

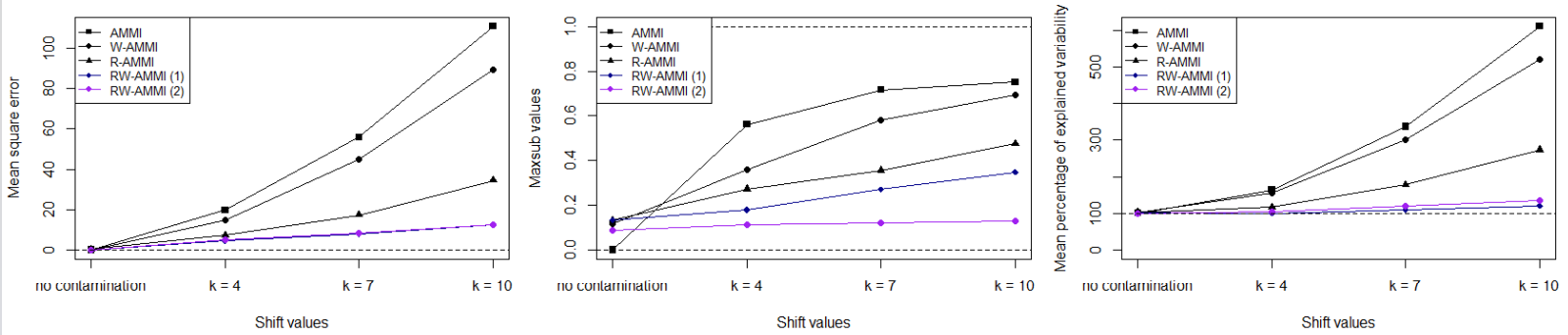
Methods

The robust weighted AMMI model is developed in this work because it combines parts of the robust AMMI model and the weighted AMMI with the objective of improving the estimations and accelerating the computational time to obtain the results.

In order to improve the estimates, new weighting schemes based on the error variance of the environments and on the error variance of the genotypes using the linear mixed effects model were proposed. And to accelerate the computational time, the robust linear model was used in the first step and the weighted singular value decomposition in the second step.

To compare the models, Markov chain Monte Carlo method were used to simulate data and some metrics such as the mean square error (MSE), maxsub and mean percentage of explained variability (MPEV) were used to identify the model that had the best performance.

Results and Conclusions



The figures above show three types of metrics (mean square error, maxsub and mean percentage of explained variability) for five models (AMMI, weighted AMMI, robust AMMI, robust weighted AMMI (1) and robust weighted AMMI (2)), for data with 10% contamination and shift type (4, 7 and 10).

(1): Weight scheme using a combination of robust linear model, genotype error variance and environment error variance;
(2): Weight scheme using a combination of robust linear model and genotype error variance.

With these results, using the RW-AMMI model with a combination of weightings and robust statistical techniques, the model outperforms the AMMI, W-AMMI and R-AMMI models according to the metrics used.