Multivariate outliers detection on GGE Biplot

Cite as: AIP Conference Proceedings **2471**, 020001 (2022); https://doi.org/10.1063/5.0083440
Published Online: 16 June 2022

A. F. Hadi, D. Anggraeni, H. Sadiyah, et al.

Multivariate Outliers Detection on GGE Biplot

A. F. Hadi[1,a)], D. Anggraeni[1], H. Sadiyah[2] and DBC. Wicaksono[3]

[1]*Data Science Research Graoup, Department of Mathematics, University of Jember, Jl. Kalimantan 37 Jember 68121, Indonesia.*
[2]*Biometrics Laboratory, Department of Agronomy, University of Jember, Jl. Kalimantan 37 Jember 68121, Indonesia.*
[3]*Department of Biostatistics & Epidemiology, University of Jember, Jl. Kalimantan 37 Jember 68121, Indonesia.*

[a)]Corresponding author: afhadi@unej.ac.id

Abstract. Outliers are a very interesting problem, statisticians are well aware of the potential effects of outliers on data analysis, especially in multivariate data. As part of our research in modeling the two-ways table data with the Row Column Interaction Model (RCIM), we had discussed the applicable RCIM model in the Genotypes × Environments Interaction (GEI) analysis which is frequently used the Genotype and Genotype × Environments Interaction (GGE) Biplot for displaying interaction in low-dimensional space. Previously, we have studied the influenced of the outlying observations on the visualization of the interaction effects in the GGE and Genotype × Environments (GE) by RCIM modeling. Now we focused on how to detect the presence of any outliers in data of two ways table and make some suggestions for practitioners by conducted simple scheme outlying observation scenario. We also proposed the use of Robust Biplot GGE as graphical techniques for detecting outliers in our data by visualizing them on two-dimensional space.

INTRODUCTION

Outlier or unusual observation is one of the main tasks in the statistical analysis of GEI data. Especially in a wide archipelagic agricultural country area like Indonesia, not every region has a similar condition. Therefore, some varieties of cultivar cannot be grown well in any particular region. The variation of the environment may lead to observations having different characteristics to the other observations, known as outlier. Such outliers often excluded from analytical data processing. But, in some cases of plant breeding research, the outliers have very useful information [1]. The common multivariate analysis techniques (e.g. principal components, discriminant analysis, and multivariate regression) are typically based on arithmetic means, covariance and correlation matrices, and least squares fitting. All of these can be strongly affected by even a few outliers [2].

Traditionally, despite the fact that GEI data sets are always multivariate, outliers are most often identified for every single variable in a particular data set. Multivariate outlier detection is the important task of statistical analysis of multivariate data. Extreme values can naturally provide environmental measures that can be interpreted specifically and if the value is not only extreme, but 'shocking' extreme or unrepresentative, that value may once again show that some unexpected influences are present in the data source. Many methods have been proposed for univariate outlier detection. The identification effort for outliers is usually based on location and spread of the data.

The higher (lower) the analytical result of a sample, the greater the distance of the observation from the central location of all observations; thus outliers, typically, have large distances. The basis for multivariate outlier detection is the Mahalanobis distance. The standard method for multivariate outlier detection is a robust estimation of the parameters in the Mahalanobis distance and the comparison with a critical value of the χ^2 [3]. However, although the estimated values larger than this critical value are not necessarily outliers, they could still belong to the data distribution.

Plant breeding effort plays an important rule in two ways (i) the Multi-environment Trial (MET) and (ii) the GEI. The MET is an experiment frequently used before breeders release the new genotype(s) become launched varieties.

2021 Asia-Pacific Conference on Applied Mathematics and Statistics
AIP Conf. Proc. 2471, 020001-1–020001-7; https://doi.org/10.1063/5.0083440
Published by AIP Publishing. 978-0-7354-4339-6/$30.00

The analysis of interaction would be difficult when the GEI was appeared [1]. The interaction term was modeled by a statistical technique of reduction dimension called Singular Value Decomposition (SVD). SVD will visualize the interaction terms graphically by Biplot and makes the GEI analysis become easier. With this feature of Biplot, The additive main effects and multiplicative interaction (AMMI) model said to be the most powerful model for the GEI [4]. In MET, the genotype (G) is applied to the different environments (E) to evaluate the interaction between G by E. AMMI models is commonly used to analyze stability and adaptability on the Genotype × Environments interaction (GEI) studies. Since G and GE must be considered simultaneously when making decisions on cultivar selection, Yan *et al.* [5] conducted evaluation of GEI and stability performance by deleting the main effect of environment (E), while the main effect (G) and the interaction effect of genotype by environment (GE) is kept and combined as GGE.

Alternatively, one can use the Row Column Interaction Model (RCIM) [6]. In the RCIM perspectives, AMMI or Generalized AMMI model was a model with row and column main effects plus one or more components of the multiplicative interaction. The singular value corresponding to each multiplicative component is often factored out, as a measure of the strength of association between the row and column scores, indicating the importance of the component, or axis. The Biplots were than produced by this SVD, which they were vulnerable to outliers in theory, as reported by [4] but the GGE has potentially robustness by itself [1].

Despite handling outliers in the modeling process is the most important thing, but for practitioners, the "first thing first" is to detect the presence of outliers in their data sets. With those reasons for the RCIM and GGE Biplot robustness, this paper wants to deliver a graphical tool for detecting the multivariate outliers in the GEI data sets using Biplot with elliptical confidence region. To address this situation an adaptive outlier identification method has been developed before as in [3, 7] was applied, here. We proposed the use of GGE Biplot of RCIM [1] featuring a confidence region for detecting outlier in the GEI data set, with comparison to the used of multivariate outlier with adaptive outlier identification of [3] as an alternative. Finally, with these two biplots, we introduced a helpful graphical tool for outlier identification in GEI data set with informative interpretation of it in multivariate data.

THE METHODOLOGY

Technically, this research was conducted by (1) simulating outliers in the GEI data set, (2) extracting the GEI effect according to the GGE model of RCIM, (3) identify the outlying observation (Genotype) by robust multivariate outlier detection using robust distance and confidence region, (4) evaluating the use of the GGE Biplot [1] in detection outlier, comparing to the multivariate outliers Biplot by the adaptive outlier identification [3].

Simulating The Multivariate Outliers in The GEI Datasets

An outlier is defined as the part of observation which has different characteristics from most corresponding observation data set. An observation is considered as an outlier when its value of the *k*-multiplied of standard deviation is greater than its original mean, where the k is greater than 3. According to [8], the outlier was mathematically expressed by $y_i^* \geq \mu_j + k \times stdev(y_i)$.

We generated the outliers randomly following the normal distribution $N(\mu_j + k\mu_j, \sigma_j^2)$, as suggested by [8]. The μ_j is the average value of the data for the j-th column, the σ_j^2 is the variance of the error term (or the variance of certain environment), and k is a constant value of the magnitude of the outliers, k= 10. We then conducted a simple scheme of simulation for adding outliers to the data matrix. The outliers were added to the generated data, placed randomly as we conducted before on [1] for (i) no outlier at all, (ii) small number of outlier (1%, 2%, 3%), (iii) a few outliers (5%, 6%), and (iv) a lot of outliers (8%, 10%).

Those outliers would be placed on the data table of RCIM2 in two kinds of placement (i) Scattered Outlier and (ii) Single Environment Outlier. The scattered outlier was a simple random placement at the whole data matrix, while the single environment outlier placement was done by systematic column-wise placement. For both placements detail, one can see [1].

Detecting Outlying Observation with Robust Distance

In the multivariate case, not only the distance of an observation from the centroid of the data have to be considered, but also the shape of the data. Recently achieved by direct estimation of the percentiles and visual

inspection of the data. When computers were not widely available an approximation of the 97.5th percentile was obtained by estimating the mean and standard deviation (SD) for each variate and computing the value of mean + 2SD. If candidates for outliers are defined to be observations falling in the extreme 2% fractions of the univariate data for each variable, the rectangle visualized with bold dots separates potential outliers from non-outliers. This procedure ignores the elliptical shape of the bivariate data and therefore it is not effective.

Multivariate outliers can now simply be defined as observations having a large (squared) Mahalanobis Distance (MD). As noted above for the univariate case, when no prior threshold is available a certain proportion of the data or quantile of the normal distribution is selected for identifying extreme samples for further study. Similarly, in the multivariate case a quantile of the chi-squared distribution (e.g., the 98% quantile $\chi^2_{p;0.98}$) could be considered for this purpose. The Mahalanobis distances need to be estimated by a robust procedure in order to provide reliable measures for the recognition of outliers.

Our simulated data here was assumed to be normally multivariate distributed. Some outliers were difficult to be detected by MD, in this case, the selection of t and C in the following equation can be a solution:

$$MD := \sqrt[2]{(x_i - t)^T C^{-1} (x_i - t)} \quad (1)$$

where the t is estimated multivariate location and C was the estimated covariance matrix. Usually, t was the average (centroid) and C is a sample of the covariance matrix. But we were not directly used it, we then used the adaptive outlier detection as in [3]. We were used the FastMCD estimator of [9] to get the robust estimate for the covariance matrix and the centroid.

The basic idea we used here is using the robust estimators of the centroid and scatter in the formula Eq. (1) for the Mahalanobis distance leads to the so-called robust distances (RDs). As used in [10] for multivariate outlier detection. If the squared RD for an observation is larger than, say, $\chi^2_{2;0.98}$ it can be declared a candidate outlier.

Ellipticals Confidence Region

Ellipses are formed by using square roots of some χ^2_2 quantiles of 0.25, 0.50, 0.75 and 0.98 by these following algorithm steps below. Note that, the quantile of 0.98 then be used if the steps in the Adaptive Outlier detection produce an infinity.

1. Calculate the covariance matrix and centroid vector using the FastMCD algorithm [9]
2. Decomposing the covariance matrix, $\mathbf{C}$, using SVD, $\mathbf{C} = U\lambda V$
3. Determine the circle coordinates, for each χ^2_2 quantile of 0.25, 0.50, 0.75 and/or 0.98:

$$a_{ipj} = U_{i1}\sqrt{\lambda_1} \times \cos\left(\frac{i}{m} \times 2\pi\right) \times \alpha_j + t_p$$

$$b_{ipj} = U_{i2}\sqrt{\lambda_1} \times \sin\left(\frac{i}{m} \times 2\pi\right) \times \alpha_j + t_p$$

where:

i = 1,2, ..., m; with m = 1000

U_{i1} = the 1st column of the matrix U resulted by SVD in step 2

$\sqrt{\lambda_1}$ = the 1st singular value or the square root of the 1st eigenvalue of covariance matrix $\mathbf{C}$ resulted by SVD in step 2

α_j = the χ^2_2 quantile with j = 0.25, 0.50, 0.75 and/or 0.98 so the α_j are in $\{\chi^2_{(2,0.25)}, \chi^2_{(2,0.50)}, \chi^2_{(2,0.75)}, \chi^2_{(2,0.98)}\}$

t_p = the MCD estimated multivariate centroid in the 2-dimensional coordinate, with p=1,2

4. Find the elliptical coordinates:
   a. $x_{ij} = a_{ipj} + b_{ipj}$, for p=1
   b. $y_{ij} = a_{ipj} + b_{ipj}$, for p=2
5. Plot all point of the first elliptical points coordinates
6. Repeat the 1st until the 5th step for the next quantile elliptical points coordinates

Then we conducted the algorithm above by R script shown in table 1.

TABLE 1. Script to build the elliptical confidence region detecting outlying observation in the GEI data set on R software

```
library(robustbase)
rob <- covMcd(xx, alpha =1/2)
covr <- rob$cov
mer <- rob$center
covr.svd <- svd(covr, nv = 0)
rr <- covr.svd[["u"]] %*%
diag(sqrt(covr.svd[["d"]]))
m <- 1000
alpha <- sqrt(qchisq(c(0.975, 0.75, 0.5, 0.25),
ncol(xx)))
rd <- sqrt(mahalanobis(xx, mer, covr))
lpch <- c(3, 3, 16, 1, 1)
lcex <- c(1.5, 1, 0.5, 1, 1.5)
lalpha = length(alpha)
for (j in 1:lalpha) {
 e1 <- cos(c(0:m)/m * 2 * pi) * alpha[j]
 e2 <- sin(c(0:m)/m * 2 * pi) * alpha[j]
 e <- cbind(e1, e2)
 ttr <- t(rr %*% t(e)) + rep(1, m + 1) %o% mer
 if (j == 1) {
  xmax <- max(c(xx[, 1], ttr[, 1]))
  xmin <- min(c(xx[, 1], ttr[, 1]))
  ymax <- max(c(xx[, 2], ttr[, 2]))
  ymin <- min(c(xx[, 2], ttr[, 2]))
plot(xx, xlab = "PC1", ylab = "PC2",
  xlim = c(xmin, xmax), ylim = c(ymin,
  ymax), type = "n", main = "Color
  according to Euclidean distance")
points(xx[rd >= alpha[j], ], pch=lpch[j],
  cex = lcex[j])
}
if (j > 1 & j < lalpha)
  points(xx[rd < alpha[j - 1] & rd >=
    alpha[j], ],cex = lcex[j],
    pch = lpch[j])
if (j == lalpha) {
    points(xx[rd < alpha[j - 1] & rd >=
      alpha[j], ], cex = lcex[j],
      pch = lpch[j])
    points(xx[rd < alpha[j], ],
      pch = lpch[j+1], cex = lcex[j+1])

 }
 lines(ttr[, 1], ttr[, 2], lty = 3)
}
```

Visualizing the GEI with Outliers Marking Using Biplot

Here we proposed visualization methods of the GEI into a 2-dimensional biplot, with features of outlier detection by these following steps:

1. Calculate the covariance matrix and location vector using FastMCD algorithm [9],
2. Calculate the square of the RD from the estimated covariance matrix and its location vector of the FastMCD estimator.
3. Each Genotype and Environment will be marked as an outlier when it had an RD greater than $\chi^2_{(2;0.975)}$

This visualization also be superimposed the elliptical confidence region to get a more informative figure. We have done these visualizations using (1) the GGEBiplots packages [11] of RCIM [1, 6], compared to the multivariate outliers Biplot by adaptive outlier identification of the mvoutlier packages [7].

RESULTS AND DISCUSSION

Detecting Outlying Observations

To identify unusual observation which tends to be outliers, we use script as in table 1 in the simulation data that has been made as section 2.1. How does the elliptical confidence region conduct identification outlier will be shown in figure 1. The GGEBiplots elliptical confidence region detecting Genotype of G09 and G03 (left), where the mvoutlier (right) detects the Genotype of G17 and G01 as outliers in the GEI Data with Scattered outlier.

To see the different identification between the two methods we show all the results in table 2. We see that in the first two rows of table 2 the GGE-RCIM identified an outlying observation of G17 as scattered outlier, the mvoutlier does not detect any outlier. It means that the GGE more sensitive than the mvoutlier. We also see that mvoutlier has higher consistency that it detects the G17 as outlier than follow by G01 in the higher percent outliers. While the GGEBiplots seem to be more sensitive but less consistent detecting the true outlier. G17 had been detected before as an outlier in the low percent of outlier, but they lose to detect G17 in the higher percentage of outliers.

TABLE 2. Detected outlying observation in the GEI by the mvoutlier and the GGEBiplots of RCIM for simulated Scattered Outliers

Percentage of Scattered Outliers	GGE-RCIM		Mvoutlier	
	Number of identified outlier(s)	Outlying Observation (Genotype)	Number of identified outlier(s)	Outlying Observation (Genotype)
0 %	1	G17	0	-
1 %	1	G17	0	-
2 %	3	G09, G11, G17	1	G17
3 %	3	G09, G11, G17	1	G17
5 %	**2**	**G08, G09**	**1**	**G17**
6 %	1	G09	2	G17, G01
8 %	1	G09	2	G17, G01
10 %	**2**	**G03, G09**	**2**	**G17, G01**

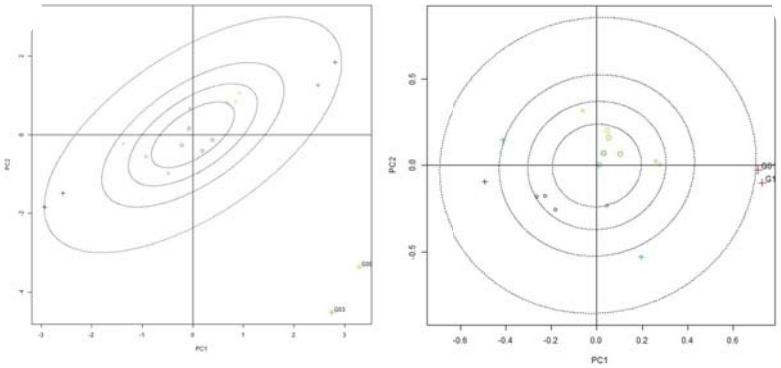

FIGURE 1. Elliptical Confidence Region detecting outlier observation of Genotype using the GGEBiplots (left) and mvoutlier (right) in the GEI Data with 10% of a scattered outlier

Table 3 shows that it seems more difficult to detect the outlying observation of Genotype when the outlier inputted in the GEI data set was Single Environment outlier. The number of outlier(s) identified was higher than in table 2 it means that both mvoutlier and GGEBiplots were more sensitive to the single environment outlier than to scattered outlier.

TABLE 3. Detected outlying observation in the GEI by mvoutlier and GGEBiplots of RCIM for simulated Single Environment Outliers

Percentage of Single Environment Outliers	GGE-RCIM		Mvoutlier	
	Number of identified outlier(s)	Outlying Observation (Genotype)	Number of identified outlier(s)	Outlying Observation (Genotype)
0 %	1	G17	1	G17
1 %	3	G16, G17, G18	1	G17
2 %	4	G10, G11, G16, G18	0	-
3 %	5	G04, G10, G11, G16, G18	0	-
5 %	**2**	**G01, G17**	**3**	**G01, G17, G18**
6 %	6	G03, G07, G09, G12, G13, G14	8	G03, G07, G09, G12, G13, G14, G15, G18
8 %	8	G01, G03, G06, G07, G12, G13, G14, G17	6	G03, G06, G07, G13, G13, G14, G18
10 %	1	G17	1	G17

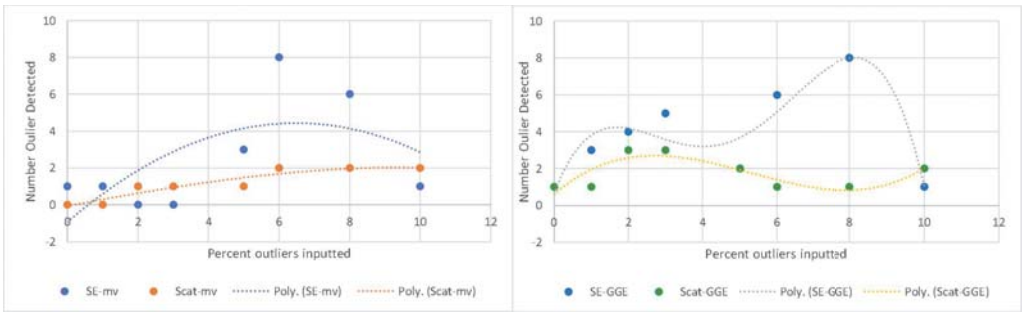

FIGURE 2. Plot number outlier detected vs percent outlier inputted with trend polynomial dotted line for mvoutlier (left) and GGEBiplots (right)

Figure 2 shows us that the mvoutlier always detects outliers less than the GGEBiplots. One can see also that due to increasing percentage of outliers, the mvoutlier shows a pattern that more systematic in number outliers detected by simple polynomial quadratic trend than the GGEBiplots for both Scattered and Single Environment outliers.

Visualizing Outliers in GGE Biplot

We than proposed a feature of outlier detection attached to the GGEBiplots of the GEI. This features will provide outlier marking in the interaction Biplot of the GEI. Interpretation of the interaction and stability analysis of the genotype or local specific adaptation will be easier to explore.

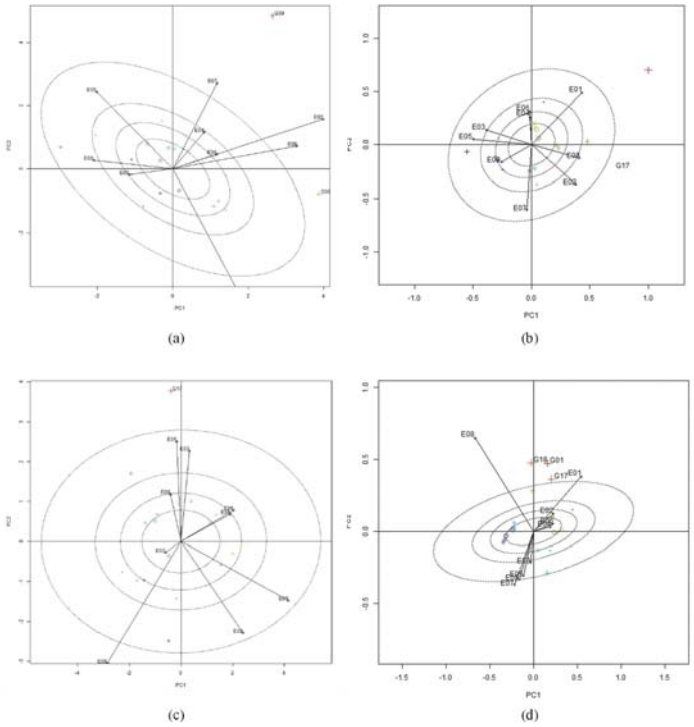

FIGURE 3. Biplot of GEI with feature if elliptical confidence region detecting outlier of Genotype and/or Environment using the GGEBiplots (left; a & c) and mvoutlier (right; b & d) in the GEI Data with 5% of scattered outlier (top; a & b) and single environment outlier (bottom; c & d)

Figure 3a was the GGE Biplot with marking outliers at G09 and G08 by red and blue colored. While the figure 3b was the mvoutlier Biplot with an outlier at G17 with red-colored. The figure 3a will be confirmed in table 2. The outliers identified in table 2 were G08 & G09 by the GGE Biplot and G17 by mvoutlier. Here we see that after we accommodate the environments in the Biplot, there were no changes in the identification of outlying point in the Biplot. This is also parallel with figure 3b which can be confirmed by table 2 that G17 was an outlier identified in the Biplot of GGE interaction was not differ from the previous identification. For the Single Environment outliers,

figure 3c was identified outlier as marking the G17 and figure 3d identified the G17, G18, G01. These two figures were confirmed parallel with table 3, except for the G01 in figure 3c did not appear as an outlier while in table 3 identified as an outlier.

Additional information here was about the environment vectors in the Biplot. Figure 3a marks that some of the environments as E07, E02, E08 identified as environment with large variance were plotted outside of the ellipse as the presence of 5% scattered outliers. While in the original data all of environments in the range of outer ellipse. The mvoutlier seems to be more robust, with marking the outlier of G17, all environments plotted inside the ellipse.

While for the single environment outlier, figure 3c and 3d show that the E08 being plotted outside the ellipse. It can be explained that in the scheme of a single environment outlier, the large value was simulated in the specific environment, let say here we put randomly in E08. So the variance of the E08 increase as the single environment outlier came in.

CONCLUDING REMARK

Developing a feature of outlier detection attached to the GGE Biplot of the GEI provides outlier marking in the interaction Biplot of the GEI. Interpretation of the interaction and stability analysis of the genotype or local specific adaptation will be easier to explore. The Robust Biplot of GGE with mvoutlier provides a good sensitivity in detection of the objects as outlier, but has to consider that if the outlier was coming in the single environment since the variance of the environment outlier will also increase directly. So the use of elliptical confidence region for detecting the outlying observation together with identification of environment with large variance observation can be attached at once in the Robust Biplot of GEI.

ACKNOWLEDGMENTS

This research was supported by Ministry of Research, Technology & Higher Education of Indonesia, Grant No. 1301/UN25.3.1/LT/2018. We thank to Salsabila (Statistical Laboratory, Department of Mathematics of UNEJ) for the data preparation in this research.

REFERENCES

1. Hadi, A. F., H. Sadiyah, & Moh. Hasan. 2018. On the development of the GE and the GGE interaction Biplot in the RCIM Model and the evaluation of its' robustness to the outlying observations. *J. Phys.: Conf. Ser.* **1132**, 012077 doi:10.1088/1742-6596/1132/1/012077
2. Hubert, M., P. J. Rousseeuw, & Van Aelst. 2005. Multivariate Outlier Detection and Robustness. Handbook of Statistics, Volume 23: *Data Mining and Computation in Statistics* (2005), Ed. C.R. Rao, E. Wegman, and J.L. Solka, Amsterdam: Elsevier North-Holland, pp. 263-302.
3. Filzmoser, P., R. G. Garrett, & C. Reimann, 2005. Multivariate outlier detection in exploration geochemistry. *Comp. & Geosciences*. 31(5) 579-587.
4. Hadi AF 2011 Handling outlier in the two-ways table by robust alternating regression of FANOVA models: towards robust AMMI models. *Jurnal Ilmu Dasar* 12(2) 123 -131
5. Yan W, Hunt LA, Sheng Q & Szlavnics Z. 2000 Cultivar evaluation and mega-environment investigation based on the GGE biplot. *Crop Sci*. 40 597-605.
6. Yee TW and Hadi AF 2014 Row-column interaction models, with an R implementation. *Computational Statistics*, **29(6)**1427-1445.
7. Filzmoser, P. and M. Gschwandtner. 2018. mvoutlier: Multivariate Outlier Detection Based on Robust Methods. https://cran.r-project.org/web/packages/mvoutlier/mvoutlier.pdf
8. Rocke, D. M. and Woodruff, D. L. 1996. Identification of Outliers in Multivariate Data, *Journal of the American Statistical Association*, 91:435, 1047-1061
9. Rousseeuw, P., & Van Driessen, K. 1999. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, 41(3), 212-223. doi:10.2307/1270566
10. Rousseeuw, P. J. and Van Zomeren. 1990. Unmasking Multivariate Outliers and Leverage Points. *Journal of the American Statistical Association* 85(411):633-639
11. Dumble, S., E. F. Bernal, & P. G. Villardon. GGE Biplots with 'ggplot2'. https://cran.r-project.org/web/packages/GGEBiplots/GGEBiplots.pdf