Handling Outlier in Two-Ways Table Data: The Robustness of Row-Column Interaction Model

IOP Conf. Series: Journal of Physics: Conf. Series **1028** (2018) 012222 doi:10.1088/1742-6596/1028/1/012222

Handling Outlier in Two-Ways Table Data: The Robustness of Row-Column Interaction Model

Alfian Futuhul Hadi[1*] , Moh. Hasan[2] , Halimatus Sadiyah[3]

[1]Statistical Laboratory, Department of Mathematics, University of Jember, Jember 68121, Indonesia
[2]Department of Mathematics, University of Jember, Jember 68121, Indonesia
[3]Biometrical Laboratory, Department of Agronomy, University of Jember, Jember 68121, Indonesia

*afhadi@unej.ac.id

Abstract. As part of our recent statistical research on modelling of the two-ways table data, here we will to investigate of the robustness of Row Column Interaction Model (RCIM). Row Column Interaction Model is a Reduced-Rank Vector Generalized Linear Models (RR-VGLM) class of modelling with the first linear predictor is modelled by the sum of the column effect, row effect, and interaction effect. The interaction effect was shown as a reduced-rank regression. We focused on outlying observations in the two ways data table. Outliers known as sample points that have unique characteristics, they differ from the majority of the whole sample. But there are some outliers that are difficult to identify due to the location and size of the data. Our previous proposed of handling outlier in Additive Main Effect and Multiplicative (AMMI) modelling by applying Robust Alternating Regression in Factor Analytics model. The two models will be compared in analysing two-ways table data that containing some outliers. In this research, two-ways table data are generated randomly follows normal distribution on Additive Main Effect and Multiplicative Interaction model by first two principal components (or AMMI2 for short), with two different types of outlier's placement. The RCIM model seem provide a better result in fitting the data than Robust factor model, the RCIM model have smaller error, even for Normal distribution or Asymmetric Laplace Distribution (ALD).

1. Introduction

Outliers observations often draw some attentions in statistical analysis. The frequently used to measure the central tendency and spreading size of data, which are the average and the variance, naturally vulnerable to the presence of outlier. Due to one observation was very much different from each other the average tends to be larger. Likewise, the variance, since the variety is measured by the distance of the observed to its average, it is also susceptible to have the same vulnerability. Almost all statistical procedures that are based on statistical averages and variance, will theoretically mimic the same kind of vulnerability.

Hawskin [1] defines an outlier as a numerically distorted observation of other observations suspected by different mechanisms. Johnson [2] defines an outlier as an observation on the data set which raises inconsistencies with the rest of the data set. In general, outliers may occur due to human error, instrument error, fraudulent behavior, changes in system behavior or system error, or constitute a natural deviation in the population. The presence of outliers often adversely affects to skew the

statistical tests based on two classic estimators is the sample mean and sample covariance. In the normally distributed data, the skewness tends to be detected as abnormalities. In the two-ways table data analysis, the Additive Main Effect and Multiplicative (AMMI) model and other, such as Principal Component Analysis (PCA) and Factor Analysis (FA), the interaction matrix was decomposed using Singular Value Decomposition (SVD). Although it is not many reported before, since theoretically SVD is least square based, the AMMI model, FA model and other kind of interaction matrix modeling mimic some problem due to the outlying observations. Please see Hadi [3] for more motivation. Hadi [3] has been show the application of Robust Alternating Regression (RAR) in Factor Analysis of Variance (FANOVA) of Croux [4] for developing a Robust AMMI models. Ardian [5] did the R implementation by Robust Principle Component Analysis (robPCA) approach to get evaluation result of handling outlier on two-ways table data. The robPCA based on PCA method to overcome the data with outlier.

In other hand, Yee & Hadi [6] introduce Row Column Interaction Model (RCIM) which is an extension from Reduced-Rank Vector Generalized Linear Models (RR-VGLM). RCIM is a model that can be applied like a Generalized-AMMI (GAMMI) model to analyze the two-ways table data for the response data matrix with more widely kind of family distribution than AMMI model. RCIM was the model used for interaction of Genotype × Environment on two ways table as an alternative approach of AMMI and GAMMI model [6,7,8].

In searching of the genotype with some superior inherited traits, outliers become something valuable, therefore ignoring the outlier really is not wise thing. Now we need to investigate the robustness of RCIM model to the presence of outlier. This paper will discuss the handling strategy to overcome the outlier on two-ways table data matrix using the RCIM model.

2. Experimental Details: A Simulation Study

The data of two-ways table was generated randomly with rows (genotypes) by columns (environments) under normal distribution. The interaction of row and column modeled by two multiplicative terms, called RCIM of rank=2 (RCIM2). The step of generating RCIM2 data follows Rodrigues, et. al[9], as seen below:

1. Create a design matrix X of $n = 50$ rows (genotypes) by $p = 8$ columns (environments) drawn from a $U(-0.5, 0.5)$
2. Do the SVD of matrix X and obtained the U, V, and D matrices.
3. Fixed the grand mean $\mu \sim N(15,3)$, row/genotype effect $\alpha \sim N(5,1)$, and column/environment effect $\beta \sim N(5,1)$
4. Generate two-ways table data with AMMI2 or RCIM2 structur
   $$Y = 1_I 1_J^T \mu + \alpha_I 1_J^T + 1_I \beta_J^T + 28 \times U[,1] D[1,1] V[,1]^T + 15 \times U[,2] D[2,2] V[,2]^T + \varepsilon$$
   with $A[,i]$ is the i^{th} column on matrix X and $A[i,i]$ is a the i^{th} row and the i^{th} column of element of matrix X where $i = 1,2$, $\varepsilon \sim N(0, \sigma^2)$. This scheme was following AMMI model with two components of interaction (AMMI2) according to Rodrigues et. al [9], but has an error term and less of term $1_I \beta_J^T$ in the model for adjusting RCIM2 or AMMI2 model.

Here in our two-ways table data was then added some outliers with a type of Pure Shift Outlier [9, 10] has the form of $N(\mu + k\mu, \sigma^2)$, where the σ^2 is the variance of the error term (or the variance of certain environment), and the $k = 4, 10$. The outliers are placed on the data table of RCIM2 in two kinds of placement as follow [9]:

- Scattered outliers
  Every single outlier was placed at random for representation the random position. This placement was conducted by choosing one of the row (genotype) at random then choosing one of the column/environment at random, then the first outlier was placed at the row and column chosen. And then with the same ways we placed the next outlier. Thus all of the outliers were placed on row and column which have been randomly selected.

- Single Environment outliers
  Here, we firstly begin with choosing the column (or environment) randomly. The outlier then will be placed on this chosen column for certain row (or genotype) which chosen also at random. After that we turn to other row to place the next outlier until fulfil the number of outlier required or equal to the number of rows at most.

In this scheme of simulation, we generate some number of outliers of 2%, 5%, and 10% of data matrix. So, for the 10% single-environment outliers we will have a specific column with high column mean otherwise we will have 10% outlier placed scattered randomly at the whole matrix data elements.

3. Data Analysis

These simulated data were analyzed by two methods:

1. The RCIM. The RCIM with normal distribution will run use the **`VGAM`** package of R with `uninormal` family function:
   `rcim(data, uninormal, Svd.arg=TRUE, Alpha=0.5, Rank=0, trace=TRUE.`

The RCIM modelling steps as follows:

- Conduct the RCIM modelling with **`rcim`** function to determine the number of multiplicative term or the rank of 1, 2, 3, 4, or 5 using deviance analysis.
- Evaluating the Goodness of Fit (GoF) by log-Likelihood, in deviance analysis and the Mean Square Error (MSE) affected by the outliers.

2. The Robust Factor Model. There are three algorithms used here with RAR-FANOVA of Croux [4] which had been implemented in R by Ardian [5]:
   - The `weight.wl1`
     It will provide the weight from every single row and column with weighted L1 regression from two-ways matrix data.
   - The `RobPCA`
     This is a robust principle component analysis by a Projection-Pursuit (PP) method.
   - The `twoway.rob`
     It conducts the two-ways table data modelling by least absolute median criterion.

The MSE value of the RCIM and Robust Factor model will be compared to evaluate the GoF and determine their robustness.

4. Experimental Result

4.1. The Reference Model: The Rank of 2 RCIM model

Since the data was randomly generated with normality distributed error term, so it does not have any outlier. It was shown that the RCIM of rank 2 model was significant with p-value less than 0.002 (see Table 1). Since the deviance analysis for other higher rank model, let say for rank=3, 4 and even for rank=5 they all have the p-value larger than 0.1 this RCIM model rank of 2 was determined as the best fit model following the data that was generated under model of RCIM rank of 2.

Table 1. The Deviance Analysis of The Rank of 2 Rcim Model

Model	df	Deviance	Mean Deviance	Ratio	p value
Row Eff.	9	20.562	2.285	0.7823103	0.634
Column Eff.	7	37.390	5.341	1.8289703	0.113
Rank 1	**15**	**139.146**	**9.276**	**3.1763397**	**0.002**
Rank 2	**13**	**410.130**	**31.548**	**10.802554**	**0.000***
Residual	35	102.216	2.920		

*) The term of rank=2 is significant

4.2. The influence of the Scattered Outlier to the RCIM

Fitting the data with scattered outliers by `rcim()`, we then provide the analysis of deviance in Table 2. It shows there was a principal change affected by the 2% scattered outlier. The scattered outlier had affect the complexity of the model. The deviance analysis of the model with 2% and k=4 scattered outliers show that the most fit model is RCIM with rank=3, more complex than before for data with no outlier. The term of rank = 3 has significant p-value (Table 2).

Table 2. The Effect of 2% Scattered Outlier With *K*=4 On Rcim's Deviance Analysis

Model	Df	Deviance	Mean Dev.	Ratio	p value
Row Eff	9	20.396	2.266	0.247	0.983
Column Eff	7	42.976	6.139	0.670	0.695
Rank 1	15	191.795	12.786	1.396	0.226
Rank 2	13	127.808	9.831	1.073	0.424
Rank 3	**11**	**227.802**	**20.709**	**2.261**	**0.046***
Residual	24	219.838	9.160		

*) The term of rank=3 is significant

But for more number of scattered outlier (5%), it seem do not have any changing of deviance analysis. It is because of the randomness of relative-position of the outliers in the two-ways table.

We can say here that since scattered outliers were placed at random, the more percentage number of outliers the less effect on RCIM. This is also happened for higher value of outlier (*k*=10), only for 2% number of outliers affect the more complexity of RCIM (Table 3). For the more percentage number of the scattered outliers do not affect any change complexity of the Rcim.

Table 3. The Effect of 2% Scattered Outlier With *K*=10 On RCIM's Deviance Analysis

Model	Df	Deviance	Mean Dev.	Ratio	p value
Row Eff	9	20.474	2.275	0.290	0.971
Column Eff	7	45.095	6.442	0.822	0.578
Rank 1	15	159.419	10.628	1.357	0.245
Rank 2	13	164.317	12.640	1.614	0.150
Rank 3	**11**	**273.200**	**24.836**	**3.171**	**0.009***
Residual	24	187.994	7.833		

*) The term of rank=3 is significant

4.3. Influence of the Single Environment outlier to the RCIM model

Here our data of two-ways table now contain a number of outliers were placed at only a certain specific column or environment. Since we generated the 2%, 5%, and 10% of an 8 × 10 data matrices, so we had only a particular column containing 2, 4 or 8 number of outliers. With this kind of placement of the outlier, we thought that it would be some changing in the interaction since specific environment will be have larger column mean differ from the other column. The question is how large the single environment outliers will make any interaction changing? Table 4 shows us that there was no change in complexity of RCIM due to the *k*=4 of the 2% single-environment outlier. With *k*=4 no one of 2%, 5%, and 10% of outliers there make any change in the complexity of interaction model (Table 4 and 5). Now we turn to see whether the larger value of single-environment outliers will affect the complexity of the interaction.

Table 4. The Effect of 2% Single Environment Outlier With *k*=4 on RCIM's Deviance Analysis

Model	df	Deviance	Mean Dev.	Ratio	p value
Row Eff	9	20.251	2.250	0.297	0.969
Column Eff	7	37.551	5.364	0.708	0.665
Rank 1	15	138.186	9.212	1.216	0.325
Rank 2	**13**	**409.840**	**31.526**	**4.162**	**0.001**
Rank 3	11	97.625	8.875	1.172	0.356[ns]
Residual	24	181.794	7.575		

[ns]) The term of rank=3 is non-significant, but rank=2 significant

Table 5. The Effect of 5% and 10% Single Environment Outlier With *k*=4 on RCIM's Deviance Analysis

The 5% Single Environment-Outlier with *k*=4

Model	Df	Deviance	Mean Dev.	Ratio	p value
Row Eff	9	22.706	2.523	0.305	0.966
Column Eff	7	36.243	5.178	0.626	0.729
Rank 1	15	163.735	10.916	1.321	0.264
Rank 2	**13**	**365.415**	**28.109**	**3.401**	**0.005**
Rank 3	11	130.591	11.872	1.436	0.220 [ns]
Residual	24	198.355	8.265		

The 10% Single Environment - Outlier with *k*=4

Model	Df	Deviance	Mean Dev.	Ratio	p value
Row Eff	9	6.884	0.765	0.100	0.999
Column Eff	7	78.330	11.190	1.462	0.228
Rank 1	15	192.931	12.862	1.680	0.125
Rank 2	**13**	**215.145**	**16.550**	**2.162**	**0.049**
Rank 3	11	128.495	11.681	1.526	0.186 [ns]
Residual	24	183.707	7.654		

[ns]) The term of rank=3 is non-significant, but rank=2 significant

Table 6. The Effect of the 2% Single Environment Outlier With *k*=10 On RCIM's Deviance Analysis

Model	**df**	**Deviance**	**Mean Dev.**	**Ratio**	**p value**
Row Eff	9	13.934	1.548	0.792	0.62672
Column Eff	7	30.624	4.375	2.237	0.06683
Rank 1	**15**	**217.113**	**14.474**	**7.403**	**0.00001**
Rank 2	**13**	**265.600**	**20.431**	**10.449**	**0.00000**
Rank 3	**11**	**237.046**	**21.550**	**11.021**	**0.00000***
Residual	24	46.926	1.955		

*) The term of rank=3 is significant

Table 6 shows us that for larger value of outliers with k=10, the interaction become more complex, and it need the 3rd term of interaction in the model. In fact, it only happens for the few number of outliers, but does not occur when the number of single environment outliers increase. We can see that there is no complexity change at the 5% and 10% of single-environment outliers.

4.4. The MSE of RCIM: Compare to Robust Factor Model

On the same complexity of the model, we still can evaluate the GoF of the model using the MSE. The MSE measure how close the fitted value to the original observation value in the data. Here we provide a comparison to the previous method of handling outlier in additive and multiplicative modeling by applying Robust Alternating Regression in FANOVA. We take a look at the effect of outlier to the MSE of RCIM firstly (Table 7).

According to Table 7, we can say that scattered outliers do affect the model less fit, the MSE become larger than before. It happens for small value of outlier (k=4) and also for the larger value of outlier (k=10). Since scattered outliers were placed at random on matrix data, we can see here that for 10% outliers has largest MSE. But it does not indicate "the more outliers the bigger the error" in general. If we do increase the number of outlier in matrix data until the value of almost every single element of matrix data shifted, then we will not investigate a matrix data containing any outliers as contaminant, but we turn to have a new matrix data with new column mean and different structure of interaction and distribution.

The single-environment outliers affect the MSE differently. A larger value of outlier with k=10 change the MSE become larger when we put only 2% in a certain column. The MSE was increase from 0.030 to 0.033. But when we put 5% outliers, we have a better fitting than original data without outlier. The MSE was 0.029, less then 0.030 of the original data. We can explain here that when we put 2% outliers and use the RCIM rank of 2, we did not meet the most fitted model. Since the most fitted one is, RCIM rank of 3.

Table 7. The MSE of RCIM Affected by Scattered and Single-Environment Outlier

RCIM (rank =2)	Number of Outlier			
	0%	2%	5%	10%
Scatter (k = 4)	0.030	**0.049**	0.048	0.070
Scatter (k = 10)		0.052	0.044	0.063
Single Environment (k = 4)		**0.033**	0.037	0.039
Single Environment (k = 10)		0.031	**0.029**	0.038

We now turn to compare the RCIM to the previous method of Robust Factor model. Robust Factor does not fit the data well. The data generated randomly with AMMI scheme rank of 2.

Table 8. shows us that the scattered outliers was affecting the robust factor model greatly. Scattered outliers were breaking down the GoF of Robust Factor model (Fig. 1d.). As the percentage of the scattered outliers increase, the MSE becomes larger.

Table 8. The MSE of Robust Factor Model Affected by Effect Of Scattered and Single-Environment Outliers

Robust Factor	Number of Outlier			
	0%	2%	5%	10%
Scatter (k = 4)	0.387	24.644	47.079	316.754
Scatter (k = 10)		10.581	22.799	297.614
Single Environment (k = 4)		58.342	47.723	1.516
Single Environment (k = 10)		35.426	4.185	9.132

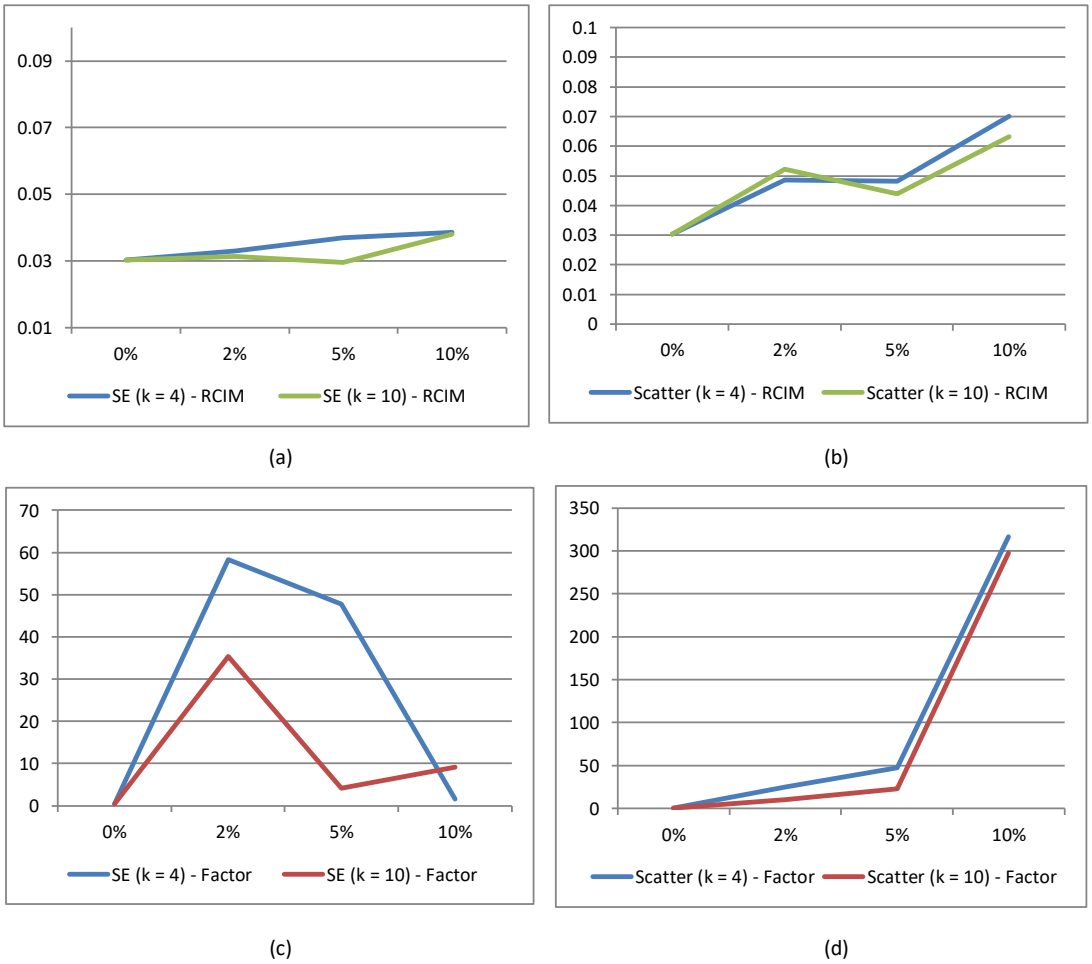


Figure 1. The plot of the MSE of RCIM rank of 2 affected by (a) Single-environment (SE) and (b) by Scattered Outlier; (c) The plot of the MSE of Robust Factor model affected by Single-environment (SE) and (d) by Scattered Outlier

In case of single-environment outliers, Robust Factor model turn to have larger MSE for low percentage (2%) of outliers in the matrix data. If we put more outlier the MSE getting better from the worst but not better than the original data. See Figure 1c for more detail. Scattered outliers inflict greater damage to the GoF of than the single-environment. Although the scattered and the single-environment outliers was affect both models differently, generally we can conclude here that for overall k and the percentage of outliers the RCIM fit the data better than Robust Factor model, since it has smaller MSE. The RCIM itself has potential robustness to the single-environment and as well to scattered outlier.

5. Discussion

The RCIM itself has potential robustness to the single-environment and as well to scattered outlier in case of two-ways table data which is generated by AMMI2 model scheme. In the comparison to robust factor model above, the RCIM was used `uninormal` family distribution. Unfortunately, this family distribution has no mathematical background theory to have the properties of robustness. The potential robustness as shown before, merely arises by the characteristic of multiplicative modelling using RR-VGLM in the Vector Generalized Linear and Additive Model (VGAM) family function. There were some VGAM family function that potentially useful in conjunction with `rcim()` [10].

We were then interested in other VGAM family function here, that is an Asymmetric Laplace Distribution (ALD).

One are strongly recommended to pay attention in Yee [10] for mathematical properties and computational detail of ALD in `rcim()` using `alapace1()` and `alapcae2()` function. With the same simulation scheme before, we will show here that `rcim()` with `alapace1()` has interesting robustness to the scattered outliers for two-ways matrix generated data rank of 2 (data for AMMI2 model). From Figure 2, we see that although an ALD has larger MSE than `uninormal` with no outlier in the matrix data, it has higher potentially robustness than normal family function. It has smaller MSE than `uninormal` for scattered outliers matrix data at high or small value of outliers. We also got the similar result for single-environment outliers. For both scattered and single-environment outliers, here we use: `rcim(data, alapace1(tau=0.5), Svd.arg=TRUE, Alpha=0.5, Rank=0, trace=TRUE)` as suggestion of Yee [10].

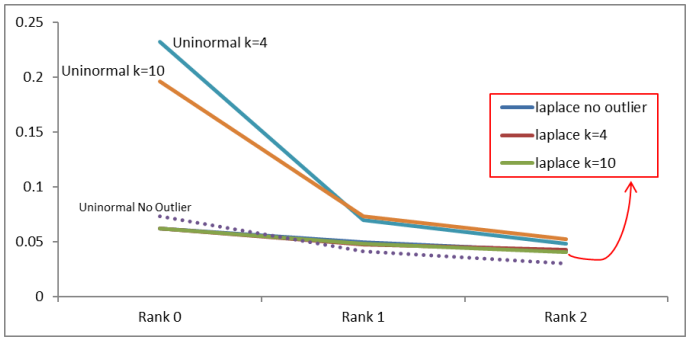
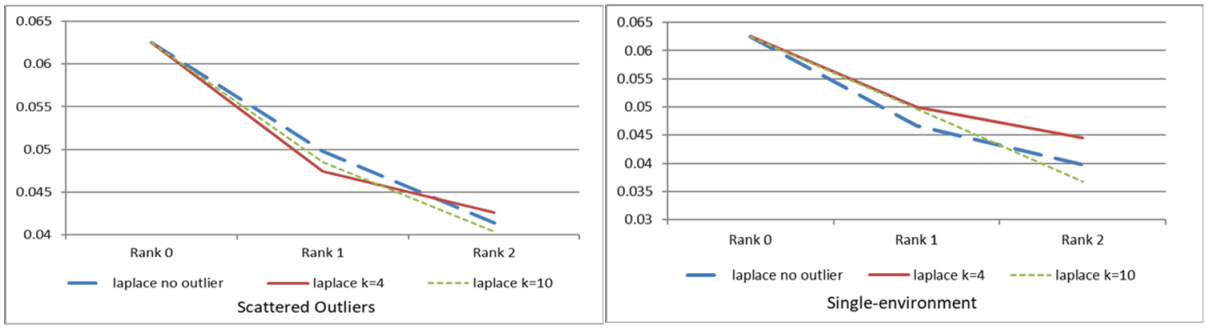


Figure 2. The MSE of RCIM model using Normal Distribution and ALD with 2% scattered for small and larger value of outliers



(a) (b)

Figure 3. Robustness of RCIM model using ALD family function to 2% scattered and single-environment for small and larger value outliers

Figure 3 show us that for matrix data with scattered outliers, the RCIM with `alapace1(tau=0.5)` provide a robust fitted value. The MSE of both contaminated data were still similar to its fitting for no outlier data. In case we focus on RCIM of rank=2, the larger scattered outliers value (k=0, 4, and 10), the smaller MSE we got. For single-environment outliers, RCIM of rank=2 with `alapace1(tau=0.5)` have more variety of MSE than scattered outliers. But it has similar property that is the larger single-environment outliers value (k=10), the smaller MSE we got. Finally, we got an evidence to conclude that RCIM with ALD has robustness to the scattered and single-environment outliers.

Acknowledgments
This work was supported in part by PDUPT 2017 under Grant No. 0446/UN25.31/LT/2017. Hadi thanks to Paulo C. Rodrigues for his paper and the discussion.

References
[1] Hawkins, D. M., Liu, L., & Young, S. S. 2001. *Robust Singular Value Decomposition. National Institute of Statistical Science.Technical Report Number 122.*
[2] Johnson, R. 1992. *Applied Multivariate Statistical Analysis*. Prentice Hall.
[3] Hadi, A. F. 2011. Handling Outlier in Two-Ways Table by Robust Alternating Regression of FANOVA Models: Towards Robust AMMI Models. *Jurnal Ilmu Dasar*, 12(2), 123-131.
[4] Croux, C., Filzmoser, P., Pison, G., & Rousseeuw, P. J. 2003. Fitting Multiplicative Models by Robust Alternating Regressions. *Statistics & Coomputin*, 13, 23-36.
[5] Ardian, J. 2016. *Implementasi Algoritma Model Robust Faktor untuk Tabel Dua Arah pada Program R.* http://repository.unej.ac.id/handle/123456789/77980.
[6] Yee, T. W., & Hadi, A. F. 2014. Row-column interaction models, with an R implementation. *Computational Statistics*, 29 (6), 1427-1445.
[7] Hadi, A. F., & Sadiyah, H. 2016. On the development of statistical modeling in plant breeding: An approach of RCIM for Generalized AMMI model with deviance analysis. *Agriculture and Agricultural Science Procedia,* 9, 134 - 145.
[8] Hadi, A. F., Sadiyah, H., & Iswanto, R. 2017. On Generalization of Additive Main Effect and Multiplicative Interaction (AMMI) Models: An Approach of Row Column Interaction Models for Counting Data. *Malaysian Journal of Mathematical Sciences,* 11(S), 115 - 141.
[9] Rodrigues, P. C., Monteiro, A., & Lourenco, V. M. 2015. A Robust AMMI model for the Analysis of Genotype by Environment Data. *Bioinformatics Advance Access*.
[10] Rocke, D. M., & Woodruff, D. L. 1996. Identification of Outliers in multivariate data. *Journal of the American Statistical Association,* 91, 1047-1061.
[11] Yee, T. W. 2015. *Vector Generalized Linear and Additive Models, With an Implementation in R*. Springer. New York.