Available online at www.sciencedirect.com

ScienceDirect

Agriculture and Agricultural Science Procedia 9 (2016) 134 – 145

Agriculture and Agricultural Science
Procedia

International Conference on Food, Agriculture and Natural Resources, IC-FANRes 2015

On The Development of Statistical Modeling in Plant Breeding: An Approach of Row-Column Interaction Models (RCIM) For Generalized AMMI Models with Deviance Analysis

Alfian Futuhul Hadi[a]*, Halimatus Sa'diyah[b]†

[a]Statistical Laboratory, Department of Mathematics, The University of Jember, Jl. Kalimantan 37, Jember, 68121, Indomesia
[a]Biometrics Laboratory, Department of Agronomy, The University of Jember, Jl. Kalimantan 37, Jember, 68121, Indomesia

Abstract

Generalized AMMI (GAMMI) model has been widely used to model the Genotype × Environment Interaction (GEI) with categorical (or in general, non-normal) response variables. It was developed by introduce the concept of Generalized Linear Model (GLM) into Additive Main Effect & Multiplicative Interaction (AMMI) model. GAMMI model will provide two major results (i) the stability analysis of some genotypes across environments and (ii) determine some others that have site specific for particular environment through Biplot of Singular Value Decomposition (SVD) of the interaction terms. This research will focus on major studies on counting data that is to round up the previous work of first author's on the Row Column Interaction Models (RCIMs) for the GEI by VGAM package of an R implementation with an addition on the deviance analysis. A simple illustrative comparison of both approaches (RCIM vs. GAMMI) was conducted on Poisson counting data of 4 rows × 5 columns. The defiance analysis was provided by log-likelihood of the model and ones of the residual. Deviance analysis will provide a way to determine the complexity of interaction component in the model, named by "rank" of model. The Biplot of both approaches seem not quite different. Finally, we did show that RCIMs be relied upon to fit well the GAMMI model and then applied it in an illustrative example to a real dataset. In addition, a simple scheme of simulation, adding some outlier on Poisson count data, will show an easy way handling the over dispersion problems, but firstly, we will talk about some statistical framework of Reduce Rank Regression (RR-VGLMs), the RCIMs, and then the approach of RCIMs for GAMMI models.

* Corresponding author. Tel.: ++6281234561932.
E-mail address: afhadi@unej.ac.id

doi:10.1016/j.aaspro.2016.02.108

Keywords: Reduce Rank Regression, RCIM, GAMMI, SVD, GEI, plant breeding, statistical modeling

1. Introduction

AMMI (Additive Main Effect &Multiplicative Interaction) model for interactions in two-way table provide the major mean for studying stability and adaptability through Genotype × Environment Interaction (GEI), which modeled by full interaction model. AMMI model represents observations into a systematic component that consists of main effect and interaction effect through multiplication of interactions components, apart from random errors component. AMMI model said to be most powerful one to analyze the GEI, by main effects plus multiplicative interaction terms. AMMI model basically presents interaction through dimension reduction techniques by using singular value decomposition (SVD) of its interaction matrix (Hadi et al., 2010; Hadi et al., 2007; Hadi & Sadiyah, 2004).

Multiplicative models (or bilinear models) bridge the gap between the main effect models (in ANOVA or GLM) and complete interaction models with interaction parameters for each cell in two way table. This model also gives a pattern of the main interaction visually through Biplot. Therefore, developing of GLM theory by accommodate the multiplicative component of interaction is very necessary. Van Eeuwijk (1995) proposed the multiplicative model in term of GLM as an extension of AMMI model called as Generalized AMMI (GAMMI) model. It is also possible to visualize the interaction terms trough Biplot. Nowaday, GAMMI model has been developed to be more general including handling Poisson count (Hadi et al., 2010).

Yee & Hadi (2014) introduce the Row Column Interaction Models (RCIM) for many interaction models including Poisson GAMMI Models. In case of Poisson count model, the formula of RCIM looks identical to GAMMI log-link. RCIM is an approach of Reduce Rank Regression (RRR) in the GLM class called Reduce Rank Vector GLM or RR-VGLM for short (Yee & Hastie, 2003), while GAMMI using a criss-cross regression (Eeuwijk, 1995). In term of parameter estimation and statistical modeling algorithms, both RR-VGLM and criss-cross GLM uses a similar algorithm called alternating regression (Falgurolous, 1996). Generally, the equivalence of both had been mentioned in Turner & Firth (2007) and also Yee & Hadi (2014) as well. Although these two approaches was using similar methods of parameter estimation and computational algorithms, but it differ in model parameters setting and the constraints were used. These remains possible to give rise to the difference results both.

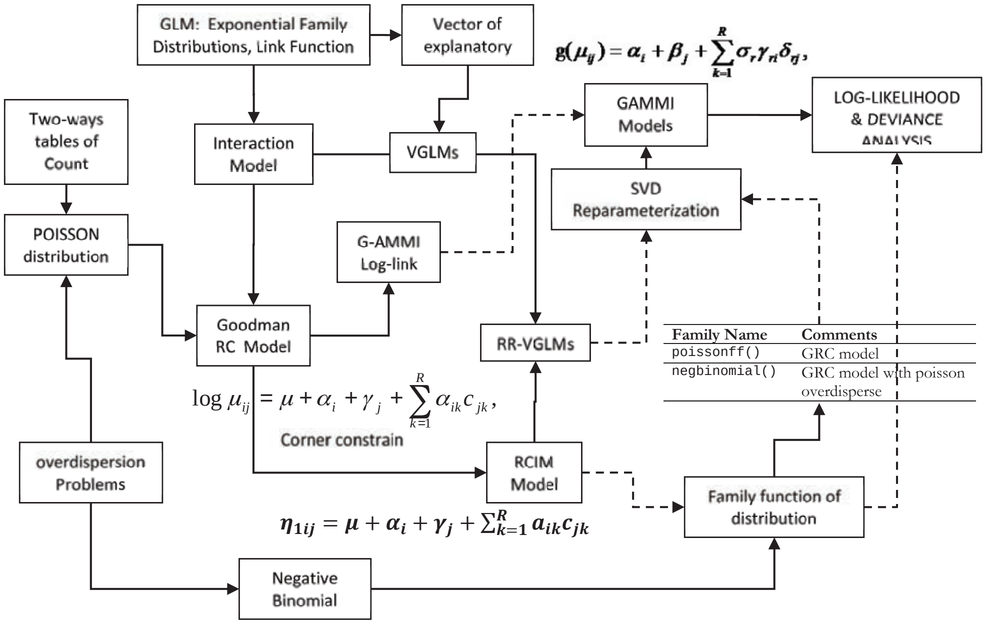
GLM: Exponential Family Distributions, Link Function
Vector of explanatory
$g(\mu_{ij}) = \alpha_i + \beta_j + \sum_{k=1}^{R} \sigma_r \gamma_{ri} \delta_{rj}$,
GAMMI Models
LOG-LIKELIHOOD & DEVIANCE ANALYSIS
Two-ways tables of Count
Interaction Model
VGLMs
SVD Reparameterization
POISSON distribution
G-AMMI Log-link
Goodman RC Model
RR-VGLMs
Family Name
Comments
poissonff()
GRC model
negbinomial()
GRC model with poisson overdisperse
$\log \mu_{ij} = \mu + \alpha_i + \gamma_j + \sum_{k=1}^{R} \alpha_{ik} c_{jk}$,
Corner constrain
overdispersion Problems
RCIM Model
Family function of distribution
$\eta_{1ij} = \mu + \alpha_i + \gamma_j + \sum_{k=1}^{R} a_{ik} c_{jk}$
Negative Binomial

Fig. 1. Statistical framework of The RCIM with others relevant to these articles

This paper describes a statistical framework and software for fitting RCIM to two-way table of Poisson count then provides the deviance analysis to determine the complexity of rank of multiplicative interaction term of the model. RCIMs apply some link function to the mean of a cell equaling a row effect plus a column effect plus an interaction term is modeled as a reduced-rank regression with rank of 2, and then will be visualized by Biplot. These comprise to the GAMMI models with the row and column effects plus one or more multiplicative interaction which is factored out by singular value decompositions and also often visualized by Biplot for the first two singular vectors. Both approaches provide very similar result, only different in numerical computation but not statistically essential. In the next early section we will talk about some statistical framework of RR-VGLM, the RCIM, and then the approach of RCIM for GAMMI models.

2. RR-VGLM and RCIM

Yee and Hastie (2003) firstly introduces a class of Reduced - Rank Vector Generalized Linear Models (RR-VGLMs) which implement a concept of reduced rank regression into multivariate class of VGLM (Yee, 2015). Let say our data was in the form of $(\mathrm{x_i}, \mathrm{y_i})$ for $\mathrm{i} = 1, \cdots, \mathrm{n}$ where $\mathrm{x_i}$ is a vector of exploratory variable for i-th observation and $\mathrm{y_i}$ is a response variable (or can be a vector). Generally, VGLM similar to GLM, but VGLM provide multiple linear predictors. VGLM can handle a number of M linear predictors (M depends on the models to be fitted) where the$-j$ one is:

$$\eta_j = \eta_j{}^*(x) = \beta_j^T x = \sum_{k=1}^{p} \beta_{(j)k} x_k \,, \; j = 1, \dots, M. \tag{1}$$

The η_jf VGLM may be applied directly to the distribution parameter $\theta_j$rather than just to the mean $\mu = E(Y)$like in GLMs. Generally,

$$\eta_j = g_j(\theta_j) \tag{2}$$

For some parameters of link function$g_j$and parameter θ_j . A collection of linear predictors here is:

$$\eta = \eta(x) = \begin{pmatrix} \eta_j(x) \\ \vdots \\ \eta_M(x) \end{pmatrix} = B^T(x) = \begin{pmatrix} \beta_j^T x \\ \vdots \\ \beta_M^T x \end{pmatrix} \tag{3}$$

where $B$is a $p{\times}M$ matrix ofregression coefficients. In many cases, regression coefficients were related to each other. For example, some of $\beta_{(j)k}$maybe are equal, or set to be zero, or add up to a certain quantity. These situations may be dealt with by use of constraint matrices. In general, VGLMs have

$$\eta_j(x) = \sum_{k=1}^{p} H_k \, \beta_{(k)}^* \,, \; j = 1, \dots, M. \tag{4}$$

where $H_1, \; H_2, \dots \; H_p$ are known constraint matrices of full column-rank and$\beta_{(k)}^*$ are vector of unknown coefficients, possibly containing reduced set of regression coefficients. With none of constraint at all, $H_1, \; H_2, \dots \; H_p = I_M$. Then for VGLM

$$B^T = (H_1\beta_{(1)}^* H_2\beta_{(2)}^* \; \dots \; H_p\beta_{(p)}^*) \tag{5}$$

Equation (5) is an expression of (4) concentrating on columns rather than rows. We need both (3) and (5) because sometimes we focus on the η_j and at other times on the variables, x_k.

Then we partition x_i into

$$\left(x_{(1)}^T, x_{(2)}^T\right) \; (of\ dimension\ of\ p_1 + \; p_2 = p)$$

and $B = \left(B_{(1)}^T, B_{(2)}^T\right)$.

In general, B is a full rank matrix, i.e., min (M, p).The $p\times M$regression coefficients tobe estimated, means for some models and data sets, this is too large and is susceptibleto over fitting. A dimension reduction method is warranted. In case of B_2.consist of multiple regression coefficients, then we can reduce it number by reduced-rank regression. A simple solution is to replace B_2 by a RRR

$$\eta = B_1^T x_1 + B_2^T x_2 \qquad (6)$$

where we approximate$B_2$by its reduced-ranked

$$B_2 = CA^T. \qquad (7)$$

C and A are $p_2\times R$ and $M \times R$ matrix respectively, and they are "thin" because the rank R is low, i.e. R = 1 or 2 hence the number of columns is much less than the number of rows. If R is low then the number of regression coefficients can be reduced enormously. When R = 2 the estimated $\hat{A}$ and$\hat{C}$may be biplotted (Yee and Hastie, 2003). The RRR is applied toB_2,because we want to make provision the variables in $x_1$to be left alone, e.g., the intercepts. Animportant fact is that RR-VGLMs are VGLMs where some of the constraint matrices areestimated. Thus,

$$\eta = B_1^T x_1 + AC^T x_{2i} = B_1^T x_1 + Av_i \qquad (8)$$

where v = v = $C^T x_2$is a vector of R latent variables. To make the parameter unique, it is common to enforce corner constraint to A. By default the top of R × R sub matrix is fixed to be I_R and the remaider of A is estimated.

We now will use Goodman's RC model, GRC for short (Goodman, 1981)to explain what a RCIM is. GRC model will fit the data by a framework of VGLM. Suppose $Y = [(y_{ij})]$ is the$n \times M$ matrix of count. GRC model fits a Reduced-rank type model to Y by assume that Y_{ij} has Poisson distributed, firstly. And the log scale of its means has the form of:

$$\eta_{ij} = log(\mu_{ij}) = \mu + \alpha_i + \gamma_j + \sum_{k=1}^{R} a_{ik} c_{jk} \qquad (9)$$

Where $\mu_{ij} = E(Y_{ij})$ is the mean of the$-ij$ cell Model in (9) need identifiable constraint, for the row and column effects,α_i andγ_j a corner constraint $\alpha_i = \gamma_j = 0$ were used. The parameter a_{ik} and c_{jk}need constraint as well, we usea_{ik} =$c_{jk} = 0$ with $k = 1, \dots, R$. Then we can write (9) as:

$$log(\mu_{ij}) = \mu + \alpha_i + \gamma_j + \delta_{ij} \qquad (10)$$

where the $n\times M$ matrix of interaction terms, $\Delta = [(\delta_{ij})]$ isapproximated by the reduced rank quantity of$\sum_{k=1}^{R} a_{ik} c_{jk}$. GRC model fits within the VGLMs framework by letting

$$\eta_i = log\,\mu_i \qquad (11)$$

where μ_i = E(Y_i)is the mean of the i^{th} row of Y.

RCIM is part of RR-VGLM where the first linear predictor modeled by the sum of the effects of rows, columns, and the effect of the interaction, in which the interaction effect is shown as reduced-rank regression. So that the model of RCIM is generally defined as RR-VLGM applied on the response variable Y as follows:

$$g_1(\theta_1) \equiv \eta_{1ij} = \mu + \alpha_i + \gamma_j + \sum_{r=1}^{R} c_{ir} a_{jr} \qquad (12)$$

where$g$is a link function, $\alpha_i$and$\gamma_j$are rows and columns effects,$\delta_{ij=} \sum_{r=1}^{R} c_{ir} a_{jr}$ are the interaction terms and the rank $R \leq \min(M, p-2)$.This means that the first parameter of a statistical model relating to a response matrix, after a suitable transformation, is equal to the sum of the intercept, a row and column effect plus optional interaction term of the formA^TC. Note that (12) applies to the first linear/additive predictor; for models with M > 1 one can leave $\eta_2. \dots, \eta_M$. unchanged because these are functions of nuisance parameters (e.g.,scale and shape parameters) that are best left alone (probably as intercept-only: $\eta_j = \beta_{(j)1}$ .Of course, choosing η_j for (12) is only for convenience.

3. The RCIM approach to fit GAMMI Model

Turner and Firth (2007) defines Generalized Additive Main Effects and Multiplicative Interaction (GAMMI) in a row-column model, by adopting the two-way table consisting of the effect of rows, columns and one or more components on a multiplicative interaction. Associate with the singular value decomposition of its multiplicative components, such as a measure of the strength of the relationship or interaction between the row and column, which indicates the importance of the component. Model GAMMI-K is defined as follows:

$$\eta_{1ij} = \mu + \alpha_i + \beta_j + \sum_{k=1}^{K} \sqrt{\lambda_k} \gamma_{ki} \delta_{kj} \tag{13}$$

where$\alpha_i$and$\beta_j$are rows and columns effects, as well as its γ_{ki} and$\delta_{kj}$are row and column values for the k^{th} multiplicative component of the interaction terms and$\lambda_k$is the singular value for the k-th component. K is the rank. Based on (13) GAMMI model appear to be identical to RCIMs. Here they apply a SVD to A^TC. While our interaction term uses corner constraints, their SVD parameterization is quite interpretable and is related to some of the other parameterizations described in Yee and Hastie (2003). The advantage of RCIMs is that it should work for any VGAM family function, thus the family size is much bigger. It is easy to perform some post-transformations such as applying `svd()` to the VGAM output to obtain the SVD parameterization for GAMMI models (Yee, 2010;Yee & Hadi, 2014).

4. Material and Methods

This research uses two datasets, originally obtained from experimental trial conducted by Indonesian Legumes and Tuber Crops Research Institute (ILETRI), Malang, Indonesia. The first dataset contain a population count of 5 types of leaves pest(*Bemissia tabacci*, *Emproasca sp*., *Agromyza phaseoli*, *Lamprosema indicata* and also *Longitarsus suturellinus*) on four soybean genotypes (Wilis, IAC-100, IAC-80-596-2 and W/80-2-4-20). The second data was also obtained from a multi-environment trial of ILETRI Malang. It involves 15 genotypes of soybean, grown at 8 locations. The observation interest is the count of filled pods.

A simple scheme of simulation was conducted by adding some outlier to make an ill condition of over dispersion on the 2nd dataset. Technically, the 2nd data set of Poisson distributed will be modified by imposing an outlier observation in it, one by one. Outlier observations we redefined as a value observation that at least equal to maximum value of each row (column)plus four times its standard deviation of row (or column). So that mathematically we can write as outlier >= max(yi) + 4*stdev.

Data will be modeled into the GAMMI model through RCIM approach on a VGAM package with Rank = 0, 1, 2, ..., n. The function of `rcim(data, poissonff, Svd.arg = TRUE, Alpha = 0.5, Rank = k, trace = TRUE)` (Yee, 2010).RCIM approach will be carried out by using rank = 0 to the maximum rank of 7. Rank = 0 stand for additive model, rank = 1 stand for one component interaction, rank = 2 for the Biplot model or 2 components of interactions and so on. Maximum rank or we can say it as the complexity of interaction terms will be determined from ratio of the log-likelihood of Poisson regression model and it of residuals. An analysis of deviance then obtained from its ratio of log-likelihood. An illness condition of over dispersion on the data will be handled by Negative Binomial Regression (NB model for short) in RCIM approach. We will then evaluate the effectiveness of the Negative Binomial model to Poisson model in handling over dispersed data by its log-likelihood ratio or/and by its MSE.

5. Result and Discussion

5.1. The Deviance Analysis & Biplot Model of RCIM

The first data was analyzed by Hadi et al. (2010), it contain a population of five types of pests leaf on four soybean genotypes (Wilis, IAC-100, IAC-80-596-2 dan W/80-2-4-20). The count data of four replicates shown in Table 1

Table 1. Population of Leaf Pests on some Soybean Genotype: a study of its endurance

	Leaf Pests				
Genotype	Bemisia tabacci	Empooascasp.	Agromyza phaseoli	Lamprosema indicate	Longitarsus suturellinus
AC-100	2	7	9	2	7
IAC-80	12	11	4	7	13
W/80	14	12	5	8	8
Wilis	16	12	4	7	16

The deviance for a model of μ is defined as the Likelihood ratio of saturated model with the full model as follows (Pawitan, 2001):

$$D = 2\, log \frac{L(y;y)}{L(\mu;y)} \tag{14}$$

where $L(y;y)$ is the likelihood of the full model, and $L(\mu;y)$ is the Likelihood of the saturated models. Equation (14) means that from subtraction of the two log-likelihood value can be obtained deviance value of the model. Log-Likelihood value on the null model would be subtracted by each model according to equation 14. Thus, we can obtain the deviance value of each model as follows:

Table 2 Residual deviance of RCIM model

RCIMs Model	Residual Deviance	Residual Degree of Freedom
Null Model	46.66258	19
Leaf Pests (column)	29.92456	15
Genotype (row)	35.31915	16
Row and Column : Rank = 0	18.58113	12
Interaction Effects : Rank = 1	3.89752	6
Rank = 2	0.1067687	2
Rank = 3	-7.33E-09	0

The deviance of the model without interaction (leaf pests and genes) can be obtained by subtracting the residual deviance of Null Model by residual deviance in each model. Model's deviance of GAMMI1 obtained from a subtracting the residual deviance of Rank = 0 model by it's of Rank = 1 model, that is 18.58113 - 3.89752 = 14.68361 with 12-6 = 6 degree of freedom. We can obtain it for GAMMI2model as well. Finally, from the whole RCIM model, either for GAMMI2 and GAMMI1 models, the deviance analysis presented in Table 3 or4 show that the GAMMI2 model meets the eligibility due to the ratio of the average deviance on GAMMI2 is significant at p-value less than 0.06. This means that the GAMMI2 model is the best model, so the GAMMI2 model with the Poisson distribution can be fitting the data well. This result identical to that obtained by Hadi et al. (2010) using

GAMMI Van Eeuwijk, which was run by GENSTAT. The Biplot models was carried out based on rank = 2 RCIM model, it performed by SVD reparameterization (fig. 2). It was statistically verified that there is clearly no difference to Hadi et al. (2010).

Table 3 Analysis Deviance of The RCIM Models for testing Rank = 2 (or GAMMI 2)

Source	df	Deviance	Mean Deviance	Ratio of Mean Deviance	p – value
Leaf Pests (column)	4	16.7380	4.1845	78.38	0.01283
Genotype (row)	3	11.3434	3.7812	70.83	0.01423
GAMMI1	6	14.6836	2.4473	45.84	0.02172
GAMMI2	4	3.7908	0.9477	17.75	0.05482
Residual	2	0.1068	0.0534		
Total	19	46.6626	2.4560		

Table 4 Analysis Deviance of The RCIM Models for testing Rank = 1 (or GAMMI 1)

Source	df	Deviance	Mean Deviance	Ratio of Mean Deviance	p – value
Leaf Pests (column)	4	16.7380	4.1845	6.44	0.02395
Genotype (row)	3	11.3434	3.7812	5.82	0.03430
GAMMI1	6	14.6836	2.4473	3.77	0.06696
Residual	6	3.8976	0. 6496		
Total	19	46.6626	2.4560		

The Biplot shows that genotypes W / 80 and IAC-80 were more likely to be susceptible to *Bemissia* rather than to *Agromyza*, and also with genotype IAC-100. Genotype W/80 seems had endurance to all types of leaf pests unless to *Emproasca*, it is contrary to IAC-100 that specifically susceptible to *Agromyza*. Biplot of interaction in log-bilinear model may lead us to find a pair(s) of soybean genotypes and a pair(s) of pest type's population which has odds ratio equal to one or log odds ratio equal to zero. On our data, we find that these pairs are genotypes W/80 and IAC-80, and pest *Bemissia* and *Agromyza*. If ones delivery straight-line these pairs of genotypes and pests, it seem "almost" perpendicular one to another. It means that log odds ratio "tends" to zero. Original data in Table 1 can verify that odds ratio between both of them tends to unity. It means that W/80 and IAC-80 has the same tendency, both are more susceptible to be attacked by *Bemissia* than the *Agromyza* in the same scale.

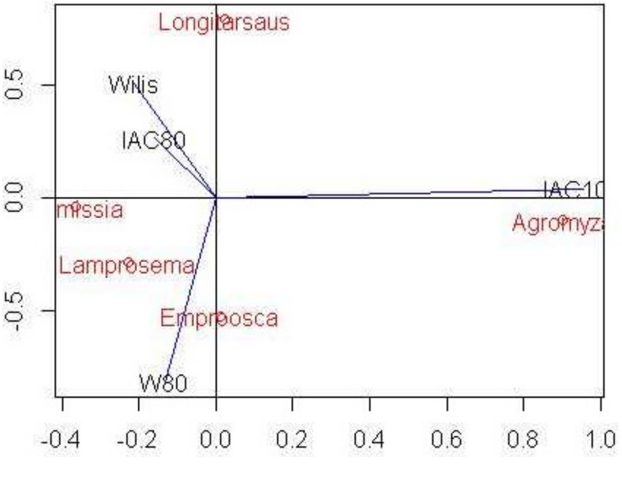

Fig. 2. The Biplot of interaction models for dataset of Table 2

5.2. Over dispersion, Outliers and Its Influence on RCIM Models

There are two strong assumptions for Poisson model to be checked: (i) events occur independently over of time or exposure period, (ii) the conditional mean and variance are equal. In practice, the Poisson with a large numbers of count, usually have greater variance than the mean are described as over dispersion. Over dispersion and the offense homoscedasticity assumptions have the same impact in linear regression model. The model's deviance become overestimated, it is greater than it should be.

Over dispersion the Poisson regression can be detected by checking the Pearson Chi-Square ($\chi2$) value or the deviance value divided by degrees of freedom. According to Hilbe (2011), a model may be over dispersed if the value of the Pearson (or $\chi2$) statistic divided by the degrees of freedom (df) is greater than 1.0. The ratio of either is called the estimated dispersion parameter (or dispersion, for short).Small amounts of over dispersion are of little concern; however, if the dispersion statistic is greater than 1.25 for moderate sized models, then a correction may be warranted. Models with large numbers of observations maybe over dispersed with a dispersion statistic of 1.05.The common causes that can lead to over dispersion is additional variation to the mean or heterogeneity (particularly may cause by outlier(s)), here a Negative Binomial model is often used.

Second observational data is the number of pods from 15 strains of soybeans grown in 8 different locations, which is available in the form of a matrix size of 15 × 8. Based on this data the simulation was conducted by adding the outliers in the data to show whether it will increase the estimated value of dispersion. The addition of outliers in the data was done by firstly calculating the standard deviation of each row and column then adds it up to the maximum value in each row and column. Mathematically, for the row it can be written as $\max(\text{row}) + 4\sqrt{stdev}$. We add 20 outliers observation in row and column of the data matrix, so we have a simulated dataset of 19.2% outliers in the matrix data. In the last part of this section we conducted another simple scheme of simulation by adding outliers with increment of0.8%, 1.7%, 2.5%, . . , 19.2% of the 15 × 8 matrix data.

Table 5 shows that there is a potential over dispersion problem either in the real data or in the simulated data which are used in this study. On both data, the deviance divided by its degrees of freedom is greater than 1. The simulated-data containing outlier has a value greater than the dispersion of real data. Table 5 indicates that outliers may cause over dispersion on RCIM modeling. RCIM Poisson with rank = 1 to rank = 4 fails to model the counting data containing outliers, since it's estimated value of dispersion parameter is larger than 1 or it was already occurred an over dispersion. However, it should be noted here that the outlier increase slightly estimated dispersion parameters on RCIM Poisson with rank = 6. It was only increasing from 0.487 on real data to 0.562 for simulated data with outliers in it. In other way, one can say that with increasing rank of model on RCIM influence of outliers to over dispersion getting less. The over dispersion problems at low rank of the RCIM can be overcome by using a negative binomial regression in RCIM scheme. Poisson regression was a special form of negative Binomial Regression with α = 0.

Table 5 Estimated value of dispersion parameter

Model	Real Data			Simulated Data with Outliers		
	Deviance	Df	Dispersion	Deviance	df	Dispersion
Rank= 1	170.615	20	8.531	201.868	20	10.093
Rank= 2	99.284	18	5.516	119.123	18	6.618
Rank= 3	61.391	16	3.837	69.151	16	4.322
Rank= 4	30.348	14	2.168	35.177	14	2.513
Rank= 5	14.192	12	1.183	16.680	12	1.390
Rank= 6	4.869	10	0.487	5.617	10	0.562

In the Poisson regression model, or generally in additive model, over dispersion will reflect badly on the testing of the model parameter, in this case, the log-likelihood value. As shown in Table 6, the value of the log-likelihood models of all rank tends to decrease by the presence of outliers. But there is something interesting in Negative

Binomial (NB) regression. First, either in the real data which is over dispersed or in simulated data with outliers, NB regression improve RCIM model to fit the data better than Poisson. It always successful to increase the log-likelihood model, especially for model with low rank or for the additive model in general, i.e. (i) model null, (ii) additive models with rows or columns component only, or (iii) models with additive components rows and columns but without interaction (rank = 0). On RCIM models with rank = 1, it clearly appears that the negative binomial model can only improve the Poisson models of over dispersed data by increasing the log-likelihood values just slightly. While for the rank = 2 or more complex model, RCIM Poisson can fit the data properly, as good as done by RCIM NB models, regardless of estimate of dispersion parameter still larger than 5. This indicates that RCIM or GAMMI model has ability to fit Poisson over dispersed data by additional multiplicative terms. Certainly with that, the model becomes more complicated.

Table 6 Log-likelihood of RCIM models affected by outliers

	Real Data		Simulated Data	
Model	Poisson Regression	Negative Binomial Regression	Poisson Regression	Negative Binomial Regression
Null	-2198,594	-631,832	-2281,542	-634,415
Row	-691,869	-532,965	-739,6529	-540,196
Column	-2086,342	-628,163	-2149,931	-630,240
Rank = 0	-547,002	-496,746	-588,291	-508,626
Rank = 1	-454,704	-451,278	-471,524	-462,886
Rank = **2**	-419,038	-419,038	-430,151	-430,076
Rank = 3	-400,092	-400,092	-405,165	-405,165
Rank = 4	-384,570	-384,570	-388,178	-388,178
Rank = 5	-376,492	-376,492	-378,929	-378,930
Rank = 6	-371,831	-371,831	-373,398	-373,398
Rank = 7	-369,397	-369,397	-370,590	-370,590

Table 7 The MSE of RCIM model with poison and negative binomial distribution

Data	Model	Poisson	Negative Binomial
Real Data	RCIM 1	0,021562830	0,020838680
	RCIM 2	0,012077778	0,012077740
	RCIM 3	0,008437611	0,008434989
	RCIM 4	0,005427515	0,005427515
	RCIM 5	0,002552477	0,002552321
	RCIM 6	0,000600769	0,000600608
Simulated Data	RCIM 1	0,025835524	0,024468820
	RCIM 2	0,014287709	0,014146090
	RCIM 3	0,009145594	0,009145374
	RCIM 4	0,005893773	0,005893766
	RCIM 5	0,002428332	0,002428258
	RCIM 6	0,001028197	0,001028055

Second, on simulated data with outliers that has more severe illness condition of over dispersion, NB models play an important role to improve the goodness of fit by increasing the value of the log-likelihoods, significantly. While an additional complexity of the RCIM (or GAMMI) model in its interaction terms has demonstrated that it has sufficient capability to fit over dispersed counting data. Generally we got similar information by comparing MSE of Poisson models versus the NB's on simulated data with outliers. Table 7 shows that, in general, the NB always has slightly smaller MSE than Poisson models, this shown in fitting real data (over dispersed) either simulated data with outliers. Particularly, on the data that has more severe illness condition of over dispersion by outliers, the difference of the MSE between Poisson models and NB models seem larger than it on real data.

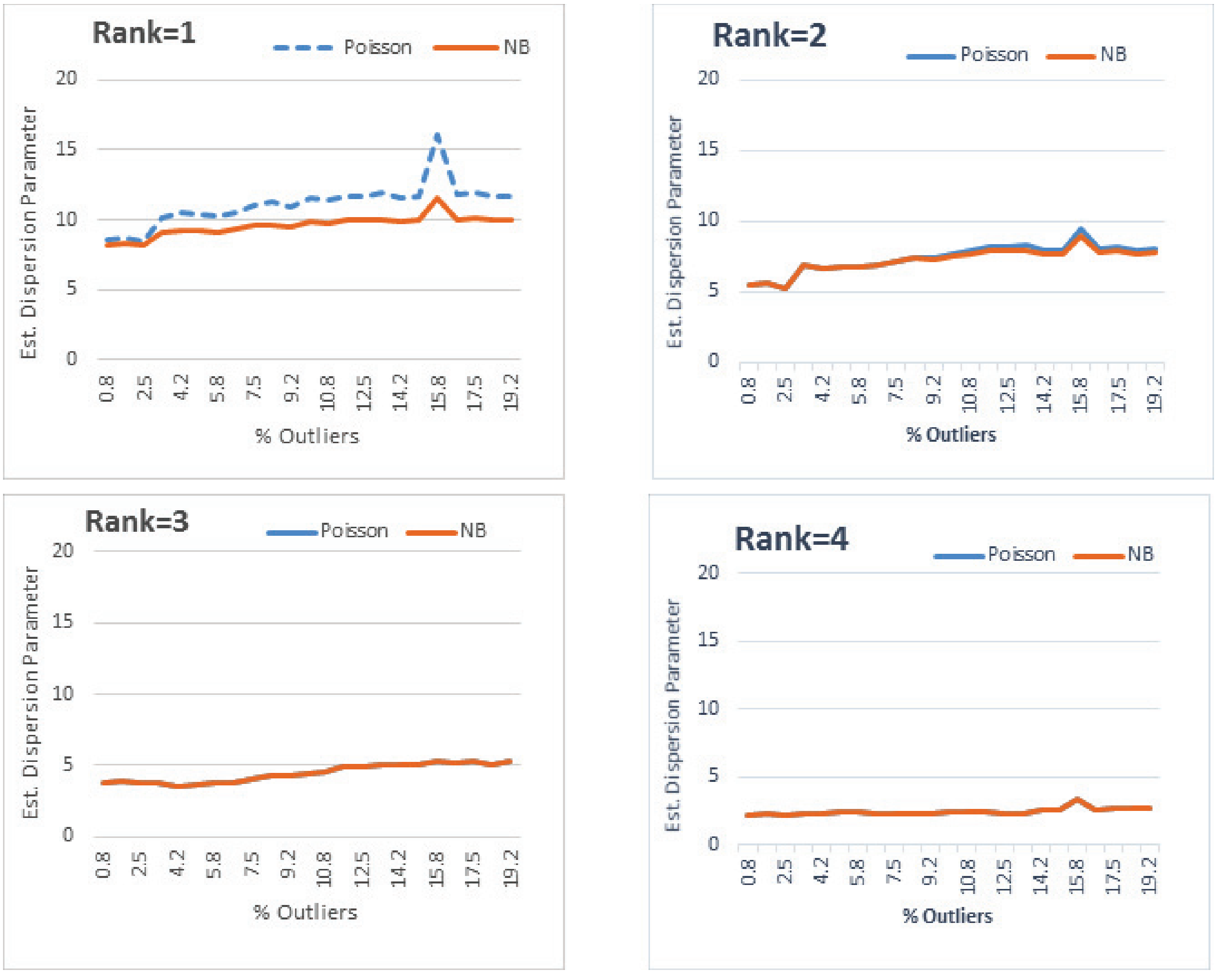

Fig. 3. The estimated dispersion parameter affected by outliers observation on both Poisson and NB model at rank = 1, 2, 3, 4.

Figure 3 shows the effect of outliers to the estimated dispersion parameter of Poisson and NB model at rank=1, 2, 3, 4. It can be noted here that at low rank (rank = 1 and 2), with some outliers, estimated dispersion parameter tend to be increase rapidly, and Poisson model seem get a worse consequence, since it has a greater estimated than NB model. Contrary to that, at higher rank it seems that the influence of outliers is felt equally by Poisson and NB.The influence of outliers to the MSE was shown in Figure 4. It shows the same phenomenon to figure 3. Both figure 3 and 4 indicate that over dispersion of count data can be handled by add an extra multiplicative term of GAMMI model (or use higher rank of RCIM).

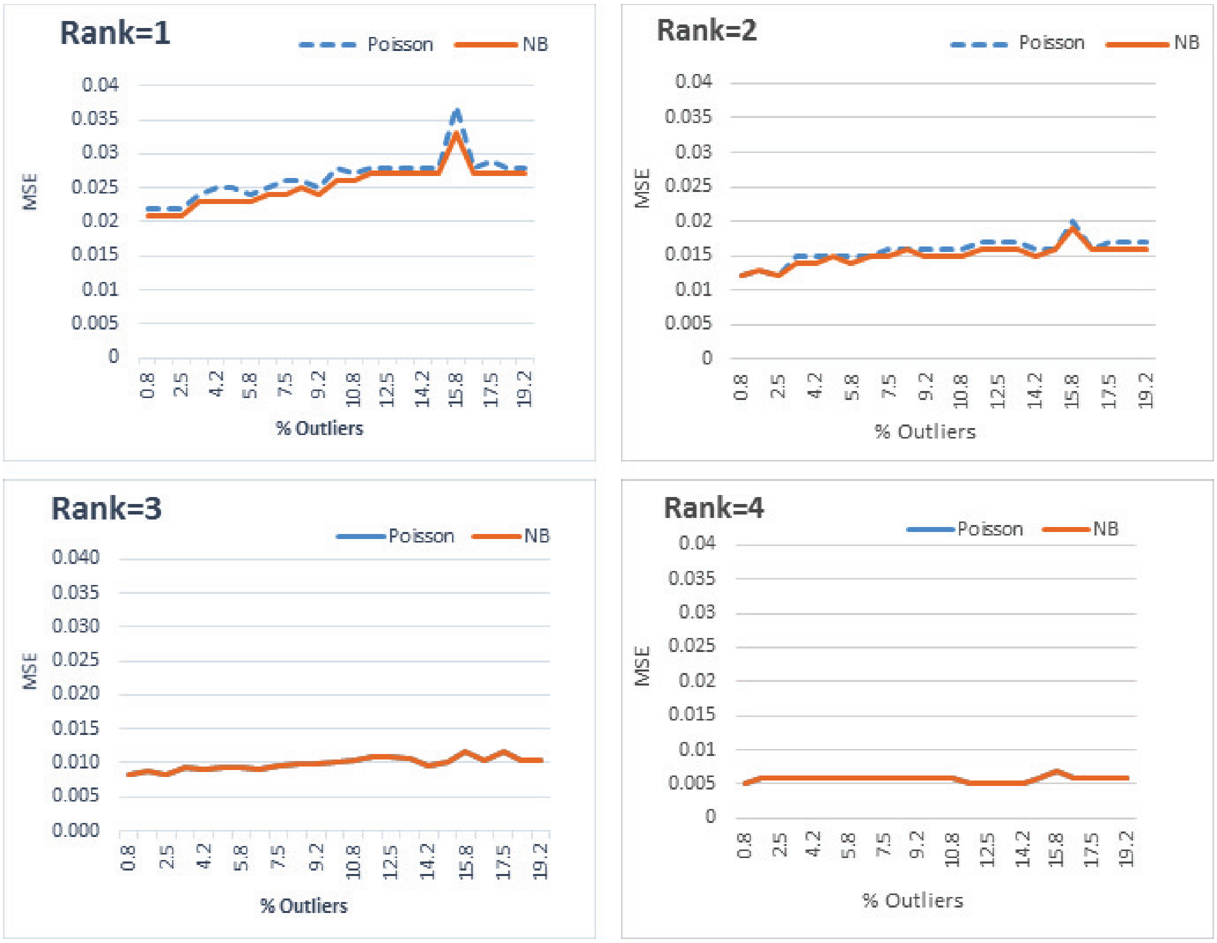

Fig. 4. The MSE of both Poisson and NB model at rank = 1, 2, 3, 4 affected by outliers observation.

6. Concluding Remark

We here have some remark concluding our discussion that is: (1) the RCIM by VGAM package of R can be model the GEI of count data, as well as GAMMI model of Van Eeuwijk. It also provides the deviance analysis as well. (2) RCIM for GEI of Poisson count data was also facing the problem of over dispersion but it was shown that one can use the NB to model over dispersed count data. (3) over dispersion in count data may be caused by outliers. Outliers tend to increase the estimated dispersion parameter, and also increase the MSE of the model, both Poisson and NB. (4) but this is not worrisome, since we can model over dispersed Poisson count data by adding an extra multiplicative term of GAMMI (or by using higher rank of RCIM) with Poisson log-link. With this Poisson log-link, we stay in the transformation scale of the log link function, so we can use the Biplot interpretation directly to its log odds ratio, as our illustrative example of Biplot above.

Acknowledgements

Hadi thanks to Thomas W. Yee, Department of Statistics, The University of Auckland for the VGAM package and for kindness during his previous visiting research. We thanks to Rudi Iswanto, Indonesian Legumes and Tuber Crops Research Institute, for the data, our discussion and collaborative scientific writing before. We also thanks to Fundamental Research Grant, Ministry of Research, Technology, and Higher Education, The Republic of Indonesia.

References

Falguerolles de A., 1996. Generalized Linear Bilinear Models. An Abstract.. The 2nd International Conference on Computing and Finance. Society of Computational Economics. Genewa, Switzerland, 26–28 June 1996.

Goodman, L.A., 1981. Association Models and Canonical Correlation in The Analysis of Cross-Classifications having Ordered Categories. J Am Stat Assoc 76, 320–334

Hadi, A.F., Sa'diyah, H., 2004. Model AMMI untuk Analisis Interaksi Genotipe x Lokasi. J. Ilmu Dasar 7(1), 33-41.

Hadi, A.F., Sa'diyah, H., Sumertajaya, I. M., 2007. Data Non-normality on AMMI Models: Box-Cox Transformations. J. Ilmu Dasar 8(2), 165-174.

Hadi, A.F, Mattjik, A.A., Sumertajaya, I M., 2010. Generalized AMMI Models for Assessing The Endurance of Soybean to Leaf Pest. J. Ilmu Dasar. 11 (2), 151-159.

Hilbe, J.M., 2011. Negative Binomial Regression. Second Edition. New York: Cambridge University Press.

Pawitan, Y., 2001. In All Likelihood: Statistical Modelling and Inference Using Likelihood. Ireland: Clarendon Press. Oxford.

Turner H., Firth D., 2007. GNM: a package for generalized nonlinear models. R News 7, 8–12.

Van Euwijk, F.A., 1995. Multiplicative Interaction in Generalized Linear Models. Biometrics, 51, 1017 – 1032

Yee T.W., 2008. The VGAM package. R News 8, 28–39

Yee, T.W., 2010. The VGAM Package for Categorical Data Analysis. Journal of Statistical Software, 32, 1--34.

Yee ,T.W., 2015. Vector generalized linear and additive models. Springer, NY.

Yee, T.W., Hadi, A.F., 2014. Row Column Interaction Models, with an R implementation. Computational Statistics. 29 (6), 1427-1445.

Yee T.W., Hastie T.J., 2003. Reduced-Rank Vector Generalized Linear Models.Stat Model 3, 15–41