ORIGINAL ARTICLE
DOI: 10.5433/1679-0375.2022v43n2p135

Heterogeneity among contingency tables diagnosed by hierarchical log-linear models and their effect on Biplots

Heterogeneidade entre tabelas de contingência diagnosticadas por modelos log-lineares hierárquicos e seu efeito em Biplots

Carla Regina Guimarães Brighenti[1]; Daniela Aparecida Mafra[2];
Marcelo Angelo Cirillo[3]

Abstract

The theory of singular value decomposition of matched matrices is used to verify the heterogeneity of rows, columns and between matched two-way tables. An exploratory analysis that can be visualized in biplots and through simulations studies with the hierarchical log-linear model using ordinary residuals and the components of residual deviance. The effect of heterogeneity was studied generating different sample sizes and their behavior was checked by adjusting Poisson's model. We concluded that the model of ordinary residuals is the one that best reflects the degree of heterogeneity among the matched tables. Finally, an illustrative example is presented in order to guide the researcher to interpret the relationship between the results of the log-linear models with the biplots considering the effects between the sum and difference between the tables.

Keywords: simulate; biplots; model; heterogeneity; tables.

Resumo

A teoria da decomposição de valores singulares é utilizada para verificar a heterogeneidade de linhas, colunas entre tabelas de dupla entrada. Em uma análise exploratória, essa relação pode ser visualizada em biplots e através de estudos de simulações com o modelo log-linear hierárquico, por meio dos resíduos ordinários e da componente da deviance residual. O efeito da heterogeneidade foi estudado, gerando diferentes tamanhos de amostra e seu comportamento foi verificado ajustando o modelo Poisson. Concluiu-se que dado o resíduo ordinário, a heterogeneidade entre as tabelas é melhor explicada pelos biplots. Por fim, apresenta-se um exemplo ilustrativo para orientar o pesquisador na interpretação da relação entre os resultados dos modelos log-lineares com os biplots considerando os efeitos entre soma e diferença entre as tabelas.

Palavras-chave: simulação; biplots; modelo; heterogeneidade; tabelas.

[1] Prof. Dr., Department of Zootechnics, UFSJ, São João Del-Rei, MG, Brasil, E-mail: carlabrighenti@ufsj.edu.br
[2] PhD student in Agri Stat and Experimentation, UFLA, Lavras, MG, Brazil, E-mail: daniela_profmatematica@outlook.com
[3] Prof. Dr., Department of Statistics, UFLA, Lavras, MG, Brazil. E-mail: macufla@ufla.br

Introduction

One of the main purposes of categorical data analysis is studying the association between two or more variables, but this identification may be complicated if tables with more complex structures are considered. The application of multivariate techniques, such as correspondence analysis (CARLIER; KROONENBERG, 1998; GREENACRE, 2003) and multiple factor analysis (ABDI; WILLIANS; VALENTIN, 2013; BÉCUE-BERTAUT; PAGÈS, 2008), may be an alternative to such problem. On the other hand, these methodologies use techniques of residual adjustment (BET, 2012) and, depending on the sample, the percentage of variability to be restored by the components may be low.

Another way to look into the association of categorical variables is the use of biplots, whose coordinates may be obtained by decomposing singular values (AITCHISON; GREENACRE, 2001). The contribution biplot can be applied to a wide variety of analyses, such as correspondence analysis, principal component analysis, log-ratio analysis, and various forms of discriminant analysis, and, in fact, to any method based on dimension reduction through the singular value decomposition (GREENACRE, 2013; VAN DER HEIJDEN; MOOIJAART, 1995). Dossou-Gbété and Grorud (2002), however, mention that high dispersions among the frequency of categorical variables result in an asymmetric biplot, which may lead to a wrong association of the variables under analysis.

Biplots may be applied in situations that involve "matched" two-way tables. According to Dossou-Gbété and Grorud (2002), the term "matched" appears as a simple extension of the concept of two-way tables, which are cross-classified according to two homologous factors (with levels in one-to-one relationship). This context shows an improvement to the technique of singular value decomposition that allows simultaneous estimations of coordinates to be used in the construction of biplots for the sum and difference of two or more contingency tables (FALGUEROLLES, 2000; GREENACRE, 2003). Nevertheless, it is worth mentioning that due to high contingency dispersion, these biplots tend to be asymmetric.

Given the relation of high dispersion with the resulting asymmetric biplots, the use of generalized methods has been common as a research tool for the relation of biplots with the structure or hierarchy of categorical variables (BRZEZINSKA, 2012). Because of such problems, Van Der Heijden, Falguerolles and Leeuw (1989), have contributed by using log-linear and as, more specifically, association models (POWERS; XIE, 2000) and bilinear models (FALGUEROLLES, 2000). Problem is that, regardless of the model applied into the analysis of categorical data, the results obtained with the biplot technique are exploratory and, consequently, the use of models applied to categorical data will bring inferential results, such as parameter interpretation and residue analysis.

The hierarchical log-linear model connection and an exploratory analysis given by the biplots technique, for combined double-entry tables, making it possible to infer about the degree of heterogeneity of the tables. The importance of this result is verified not only in the centroid of the biplots, but also in the behavior of the components of the sums of squares, given the residual information of the model, which influence the estimates of the effects $A+B$ (the sum of the tables) and $A-B$ (difference between the tables), considering different degrees of heterogeneity between them, thus, the importance of having a simulation study is justified.

Considering that the main purpose is to carry out a study using a simulation that points out the performance of a methodology involving the relation between hierarchical log-linear methods and the biplots, this work aims at proposing a Monte Carlo simulation study on hierarchical log-linear models, thus allowing the identification of the most adequate residue to contemplate the heterogeneity between matched two-way tables at the construction of Biplots. The importance to carry out an inferential procedure is highlighted among the several advantages provided by this study, relating to model adjustment and its contribution to the interpretation of a Biplot, which heterogeneity effect results in more asymmetric biplots. In such context, the researcher is expected to have more confidence while interpreting biplots, because such interpretation is mostly subjective.

Therefore, this paper is structured as follows: Introduction; Materials and methods; Random generation of matched two-way tables by log-linear models; Singular value decomposition of matched two-way tables implemented to biplot visualizations; Results and discussion; Influence of the residuals to identify the level of heterogeneity according to sum of matched tables; Influence of the residuals on the identification of levels of heterogeneity, according to the difference between matched tables; Numerical example; Conclusions.

Methodology

The methodology used in this work was divided into the following structure: In section *Random generation of matched two-way tables by log-linear models*, the structure of the design is given, as well as the model specification, predictor and linking function were mentioned. After adjusting the model, the frequency in each cell was obtained, and then replaced with ordinary residuals or with deviance residue components, creating three situations to be assessed in each scenario.

In section *Singular value decomposition of matched two-way tables implemented to the biplot visualizations*, the decomposition of singular values was implemented by obtaining coordinates in each of these situations and scenarios. As for heterogeneity of matched tables, assessing "sum" and "difference" between tables is important. Thus, the empirical distributions of sum of squares of such components were collected after 2000 simulations and their averages were computed so as to be used to build plots and biplots.

Random generation of matched two-way tables by log-linear models

There are several types of log-linear models for two-way contingency tables (FALGUEROLLES; FRANCIS, 1994). A saturated model that includes all the possible effects to explain every single expected cell frequency given by equation (1)

$$\eta_{jk} = \beta_0 + \beta_{1j} + \beta_{2k} + \beta_{12k}, \tag{1}$$

where β_0 represents an overall effect or a constant, β_{1j} represents the main or marginal effect of the j^{th} row, β_{2k}, represents the main or marginal effect of the k^{th} column and β_{12k} represents the interaction.

Considering that y_{ijk} represents the observation that belongs to the i^{th} group, i =1, 2, to the j^{th} row j =1,$\cdots$,5, and k^{th} column, k =1,$\cdots$,3, the hierarchical structure was given according to the layout of Table 1.

Following such specifications, the log link function is

$$g(\mu_{ijk}) = \log(\mu_{ijk)} = \eta_{ijk}, \tag{2}$$

where $\mu_{ijk} = g^{-1}(\eta_{ijk})$.

Using the structure as reference, see Table 1, each y_{ijk} was simulation of Poisson model considered a linear predictor, defined by

$$\eta_{ijk} = \beta_0 + \beta_{1i} + \beta_{2j} + \beta_{3k}, \tag{3}$$

Table 1 – Contingency table with structured groups.

		Column (k)			
Group (i)	**Row (j)**	**1**	**2**	**3**	**Total**
1	1	y_{111}	y_{112}	y_{113}	y_{11+}
	2	y_{121}	y_{122}	y_{123}	y_{12+}
	3	y_{131}	y_{132}	y_{133}	y_{13+}
	4	y_{141}	y_{142}	y_{143}	y_{14+}
	5	y_{151}	y_{152}	y_{153}	y_{15+}
2	1	y_{211}	y_{212}	y_{213}	y_{21+}
	2	y_{221}	y_{222}	y_{223}	y_{22+}
	3	y_{231}	y_{232}	y_{233}	y_{23+}
	4	y_{241}	y_{242}	y_{243}	y_{24+}
	5	y_{251}	y_{252}	y_{253}	y_{25+}

Source: The authors.

that included the effect of the table, but did not consider interaction parameters.

Assuming log link function equation (2), and η_{ijk} defined in equation (3), the as a hierarchical log-linear model written as, equation (4)

$$\log(\mu_{ijk}) = \beta_0 + \beta_{1i} + \beta_{2j} + \beta_{3k}, \tag{4}$$

where $\beta_{11} = 0$, β_{1i}, $i = 2$, is the effect of the i^{th} table (group), $\beta_{21} = 0$, β_{2j}, $j = 2,\cdots,5$, the effect of the j^{th} row and $\beta_{31} = 0$, β_{1k}, $k = 2,\cdots,4$, the effect of the k^{th} column. Thus, $y_{ijk} = E(e^{\eta_{ijk}})$.

Therefore, upon the independence of tables, rows and columns, our model has $I+J+K$ number of independent parameters. The marginal distribution was defined by y_{ij+} and y_{i+k}, i =1 or 2, so the sample size was given by $\eta_i = y_{ij+} + y_{i+k}$, $i = 1,2$.

The specification as to what parametric values should be used in the Monte Carlo simulation to generate y_{ijk} was determined considering the increase in the level of heterogeneity among the tables, as well as the discrepancies related to the effect of rows and columns, as described on Table 2. The combination of these parametric values resulted in 32 scenarios, which were assessed in 2000 Monte Carlo simulations, implemented to software R (R CORE TEAM, 2015).

Singular value decomposition of matched two-way tables implemented to the biplot visualizations

As Greenacre (2003) recommended, Table 1 was partitioned into two tables G_i, named G_1 and G_2, according to rows, in order to extract the components that relate the influence of the variables represented in the columns and the differences between tables, simultaneously represented by $G_1 + G_2$ and $G_1 - G_2$.

Table 2 – Parametric values used in the simulation to obtain frequencies on the contingency tables, according to the crescent order of heterogeneity among tables.

Parameter			Description
Group	Row	Column	
β_1	β_2	β_3	
0 1	0.2	0.2	Low and equal row and column effects
1.5 2	1.0	1.0	High and equal row and column effects
2.5 3	0.2	1.0	Low row effect and high column effect
3.5 4	1.0	0.2	High row effect and low column effect

Source: The authors.

To do so, the singular value decomposition was applied to the matrix N given by equation (5)

$$N = \begin{bmatrix} G_1 & G_2 \\ G_2 & G_1 \end{bmatrix}. \tag{5}$$

At this stage, it is worth mentioning that elements y_{ijk} that constitute matrices G_1 and G_2 were also replaced by ordinary residuals expressed by r_{ijk} and given by

$$r_{ijk} = \hat{y_{ijk}} - y_{ijk}, \tag{6}$$

that is, the difference between predicted and observed values, and the components of deviance residuals r^D_{ijk}, given by equation (7)

$$r^D_{ijk} = \text{sign}\left(y_{ijk} - \hat{y_{ijk}}\right)\left\{2y_{ijk}\log\left(\frac{y_{ijk}}{\hat{y_{ijk}}}\right) + 2\left(\eta_{ijk} - y_{ijk}\right)\log\left(\frac{\eta_{ijk} - y_{ijk}}{\eta_{ijk} - \hat{y_{ijk}}}\right)\right\}^{\frac{1}{2}}. \tag{7}$$

Assessing the residuals in singular value decomposition applied to matched tables is important, because the log-linear model was generated upon independence of rows and columns; however, no restriction was made as to the homogeneity or heterogeneity of the residual variance. Thus, considering ordinary residuals and deviance residue components, respectively, both residue heterogeneity and homogeneity are shown in the biplot visualizations.

After obtaining different residuals in different effect levels in tables, rows and columns, the next step was singular value decomposition, based on partitioned matrices. Considering matrix N, equation (5), constituted by G_1 and G_2, there was correction by average, resulting in block matrix M, constituted by matrices A_1 and A_2 so that, equation (8):

$$M = \begin{bmatrix} G_1 - 1\bar{c}^T & G_2 - 1\bar{c}^T \\ G_2 - 1\bar{c}^T & G_1 - 1\bar{c}^T \end{bmatrix} = \begin{bmatrix} A_1 & A_2 \\ A_2 & A_1 \end{bmatrix}, \tag{8}$$

where $\bar{c} = 0.5(\bar{g}_1 + \bar{g}_2)$ being $\bar{g}_i$ the average row of G_i.

The sum of the squared elements in matrix M, equation (8), may be decomposed into two components: one due to the matrix sum $A_1 + A_2$ and one due to matrix difference $A_1 - A_2$ written as equation (9)

$$2\sum_i\sum_j(a_{1jk} - \bar{c_k})^2 + 2\sum_i\sum_j(a_{2jk} - \bar{c_k})^2 = \sum_i\sum_j(a_{1jk} - \bar{c_k} + a_{2jk} - \bar{c_k})^2 + \sum_i\sum_j(a_{1jk} + a_{2jk})^2. \tag{9}$$

If an evaluation of averages A_1 and A_2 is necessary, rather than summing, the sum component $A_1 + A_2$ must be divided by the number of columns.

However, according to Greenacre (2003), a single singular value decomposition applied to matrix M, equation (8), may provide both components $A_1 + A_2$ and $A_1 - A_2$, considering the expressions defined as

$$A_1 + A_2 = UD_\alpha V^T \tag{10}$$

and

$$A_1 - A_2 = XD_\beta Y^T, \tag{11}$$

where U and X are singular vector matrices to the left and V and Y are singular vector matrices to the right, each with k orthonormal columns. D_α and D_β represent diagonal matrices with positive singular values γ at a decreasing order of magnitude. Therefore, the decomposition of Matrix M, equation (8), is described by equation (12).

$$\begin{bmatrix} A_1 & A_2 \\ A_2 & A_1 \end{bmatrix}_{2j\times 2k} = \frac{1}{\sqrt{2}}\begin{bmatrix} U & X \\ U & -X \end{bmatrix} \times \begin{bmatrix} D_\alpha & 0 \\ 0 & D_\beta \end{bmatrix}\frac{1}{\sqrt{2}}\begin{bmatrix} V & Y \\ V & -Y \end{bmatrix}^T. \tag{12}$$

Still according to Greenacre (2003), altering the signal of singular values in X and Y, corresponding to the matrix difference, and the presence of factor $\frac{1}{\sqrt{2}}$ multiplying the singular values to the left and to the right ensures a correct normalization of the solution and the fact that singular values to the left and to the right of the equation are orthonormal. Note that components $A_1 + A_2$ and $A_1 - A_2$, equations (10) and (11), are not separated, but interspersed according to the magnitude of the corresponding singular values, which are disposed at a decreasing order.

Discriminating the vectors associated to components $A_1 + A_2$ and $A_1 - A_2$ implies that the singular vectors to the left (U) and to the right (V) refer to the component $A_1 + A_2$ and have two identical copies to each vector, which are arranged in the same column, one on top of the other, resulting in a vector twice as big as expected, according to equation (8). On the other hand, for singular vectors X and Y, which correspond to component $A_1 - A_2$, the initial vector and the arranged vector have opposite signs, which allow us to separate the k related to component $A_1 + A_2$ and the$k$ related to component $A_1 - A_2$ from the resulting vectors $(2k)$.

Once these vectors associated to components were identified, we obtained the coordinates for $A_1 + A_2$ and $A_1 - A_2$. The biplots were created considering the averages of both main coordinates, gathered from 2000 Monte Carlo simulations.

Thus, in order to build the biplots, the contribution of each autovalue should not be calculated in total of n autovalues, but as $n/2$, since half of these autovalues are related to the sum $A_1 + A_2$ and the other half, to the difference $A_1 - A_2$.

According to Greenacre (2003), in order to report these aspects $A_1 + A_2$ and $A_1 - A_2$, it is possible to calculate two matrices, one for the sum, or average, and another one for the difference, and carry out separate analyses, but it is also possible to reach both results by making a single SVD analysis of a bigger matrix, formatted as a special block matrix, though the result will be a decentralized difference matrix. In such case, the resulting biplots show, simultaneously, the differences between columns and rows, as well as differences between both matrices (AITCHISON; GREENACRE, 2001; GREENACRE, 1993).

Results

When it comes to the applicability of Monte Carlo simulations, as used herein, we find it proper to obtain the empirical distribution of biplot coordinates by adjusting log-linear models linked to decomposition into singular values. Therefore, an analytical study is inviable.

In order to validate the methodology, a great number of experiments had been necessary, which would only be possible by implementing computational system. It is possible to assess heterogeneity through the magnitude of average coordinates, obtained from 2000 simulations. For further details on the computational implementation we included, please refer to the script attached in appendix, where we describe all functions used to obtain the results. The heterogeneity between contingency tables is not only reported numerically, but also graphically.

We obtained the results shown in Table 3 using the contingency table as reference, structured according to the layout of Table 1 and, aiming at comparing the heterogeneity among the tables, characterized by the parametric values, as defined in β_1. We also kept the effect of categorical variables constant, as represented in the rows and columns, with the respective parametric values, as specified in $\beta_{2j} = 0.2$ and $\beta_{3k} = 0.2$. Such results correspond to the sum of squared components $A_1 + A_2$ and $A_1 - A_2$, considering the numbers gathered from the hierarchical log-linear models, adjusted by maximum likelihood method.

Considering the hierarchical log-linear model, the ratio between sample sizes from the contingency tables A_1 and A_2 follow the relation:

$$\frac{n_{A_1}}{n_{A_2}} = e^{\beta_{1i}} \tag{13}$$

where β_{1i} corresponds to the heterogeneity between tables. Thus, the higher β_{1i} is, the larger the heterogeneity between sample sizes from the tables.

According to Falguerolles (2000), in case of matched matrices that represent the analysis of categorical data from two contingency tables A_1 and A_2, the resulting matrix for the sum $A_1 + A_2$ results in the accumulation, cell to cell, of frequencies from both tables, and $A_1 - A_2$ is the difference between frequencies. If such frequencies between tables differ substantially from the total or marginal frequencies, such fact will supersede the analysis of the difference matrix, i.e., $A_1 - A_2$. According to Greenacre (2003), when sample frequencies n_{A_1} and n_{A_2} have similar sizes, there will not be such problem, especially when the data are not collected according to a given design that exposes the difference matrix.

Note that in the results shown on Table 3, there is statistical evidence that there is effect based on type of residuals to the log-linear model, of expressive differences as to the components.

Table 3 – Sample size and sum of squares of components $A_1 + A_2$ and $A_1 - A_2$, as a function of group and distribution parameters.

β_1	Sample size		Simulated data			
			Ordinary residuals		Residual deviance components	
	n_{A_1}	n_{A_2}	A_1+A_2	A_1-A_2	A_1+A_2	A_1-A_2
0.0	43	43	40	74	17.06	29.61
1.0	318	117	212	346	15.25	26.67
1.5	865	193	496	755	15.82	25.50
2.0	2347	318	1235	1701	16.03	24.72
2.5	6380	523	3207	4023	16.41	23.79
3.0	17349	864	8534	9847	17.21	23.43
3.5	47149	1424	23103	25124	17.35	22.38
4.0	128157	2349	61371	65276	17.71	22.29

Source: The authors.

When we used the residual deviance components, there is proximity of results obtained in relation to the sum of squares to the components. So it is possible to state that such waste does not establish the heterogeneity between the tables. Thus, this discussion emphasizes the results related to the adjustment of log-linear model, using ordinary residuals.

Influence of the residuals to identify the level of heterogeneity according to sum of matched tables $A_1 + A_2$.

In order to assess the performance of ordinary residuals, Figure 1(a), and the residual deviance components, Figure 1(b), we considered the parametric values specified for β_{1i}, which characterize the heterogeneity among tables, as a function of the sum of squares of component $A_1 + A_2$.

Considering the residual deviance components, Figure 1(b), we observed that the low effect of variables organized in "rows" caused the component $A_1 + A_2$ to provide higher numbers in all situations. Considering ordinary results, the total synthesized variation in $A_1 + A_2$ did not cause differences to the effects among rows and columns. Such result is noticeable when verifying the similarity between curves, Figure 1(a). Similarly, as a comparison, the occurrence of this result was similar to situations in which the observed frequencies, Figure 1(c), were used in the study of table heterogeneity for component $A_1 + A_2$.

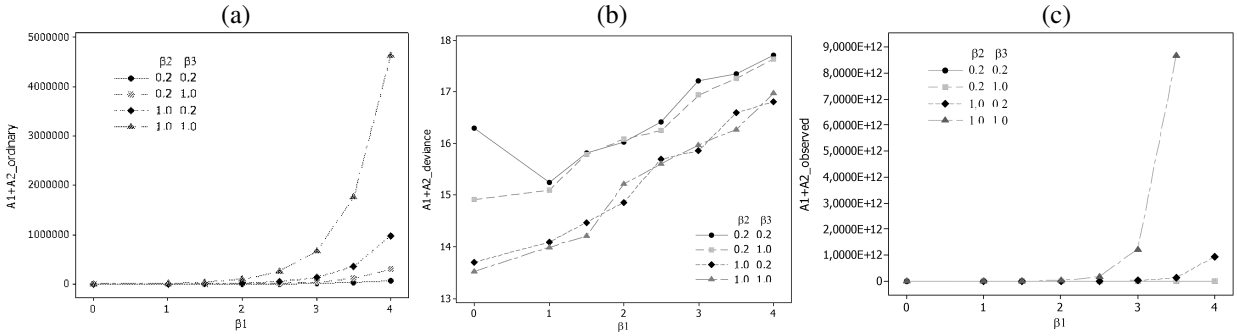
In this way, we understand that the contribution of parameters β_2 to the categorical levels described in the table rows and β_3 the levels arranged in the columns has a greater impact on the effect of groups β_1, when biplots are constructed with residual deviance, Figure 1(b). Considering the ordinary residuals, Figure 1(a), this effect is more stable, therefore, the use of ordinary residuals is recommended because it presents a more homogeneous behavior in relation to the magnitude of the parameters associated with the rows and columns of the tables.

Thus, it can be said that use of ordinary residuals allows to verify, in a manner similar to the use of the observed frequencies, the behavior of the sum of squares of components $A_1 + A_2$ due to the heterogeneity of the combined table and, moreover, has the advantage add information of the adjusted model. Thus, it was found the performance of using ordinary residuals in the building biplots from the eigenvectors and eigenvalues of components $A_1 + A_2$ considering the parametric values β_{1i} increasing and with fixed $\beta_{2j} = 0.2$ and $\beta_{3k} = 0.2$.

The biplots built to $\beta_{1i} = 0$ are illustrated in Figure 2(a). In this case, the generated tables are homogeneous and it is noticed a symmetric biplot and very close to zero residual, as expected. For the parameter $\beta_{1i} = 2.0$, Figure 2(b), the heterogeneity is identified by the asymmetry of Biplot and increasing the magnitude of both the first coordinate on the second axis. This occurs similarly when used $\beta_{1i} = 4.0$, Figure 2(c), but there is not a proportional increase in the magnitude of the coordinates. This may be associated with the fact that the sum component, having as main characteristic the overall average of the column, not being able to distinguish the heterogeneity levels generated by the log-linear model, but only the existence or not of this.

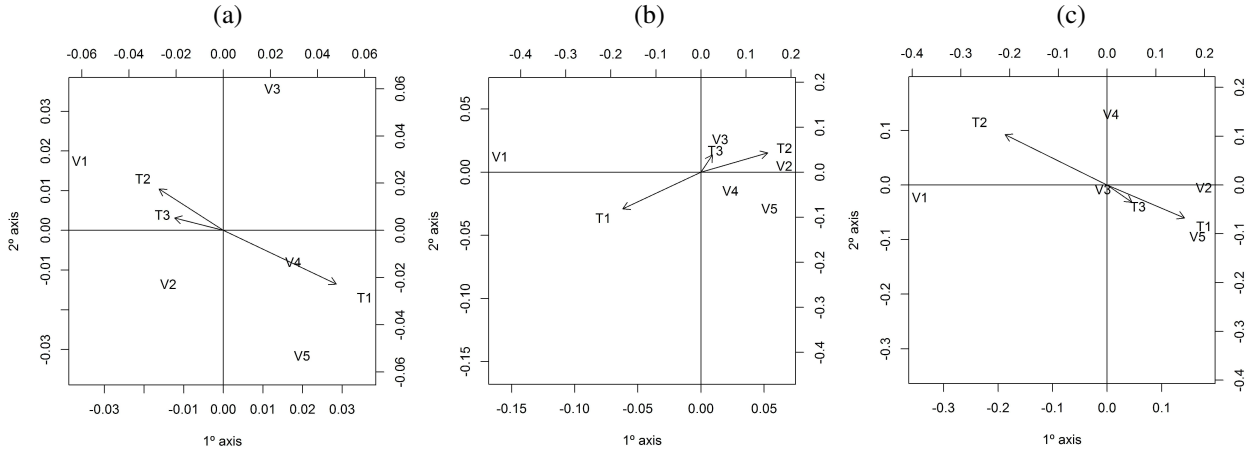
In case of the parameter $\beta_{1i} = 2.0$ and $\beta_{1i} = 4.0$, the biplot created is asymmetric, since the centroid is located next to the variables represented in the "row". Relating this result to the level of heterogeneity, Figure 1(b), there is statistical evidence to confirm that, in hierarchical log-linear models, the influence in component $A_1 + A_2$ is more pointed out towards categorical variables represented in the "row", which is different from what Greenacre (2003) mentions, when they say that adding two matrices, Tables 1 and 2, provides a general idea of column variation.

Figure 1 – Sum of squares of component $A_1 + A_2$ obtained as a function of different degree of heterogeneity β_1 for combined $\beta_2 = 0.1$ or 1.0 and $\beta_3 = 0.2$ or 1.0, using: (a) ordinary residuals; (b) residual deviance components and (c) observed frequencies.



Source: The authors.

Figure 2 – Biplots related to components $A_1 + A_2$ considering ordinary residuals for fixed $\beta_2 = 0.2$ and $\beta_3 = 0.2$ parameters and different degree of heterogeneity β_1: where in (a) $\beta_1 = 0.0$; (b) $\beta_1 = 2.0$ and (c) $\beta_1 = 4.0$. V_j the variables represented in lines and the T_k variables represented in the columns.



Source: The authors.

With such characteristics, the biplot obtained by ordinary residuals for matrix $A_1 + A_2$, used just to verify the presence or absence of total variation, is recommended.

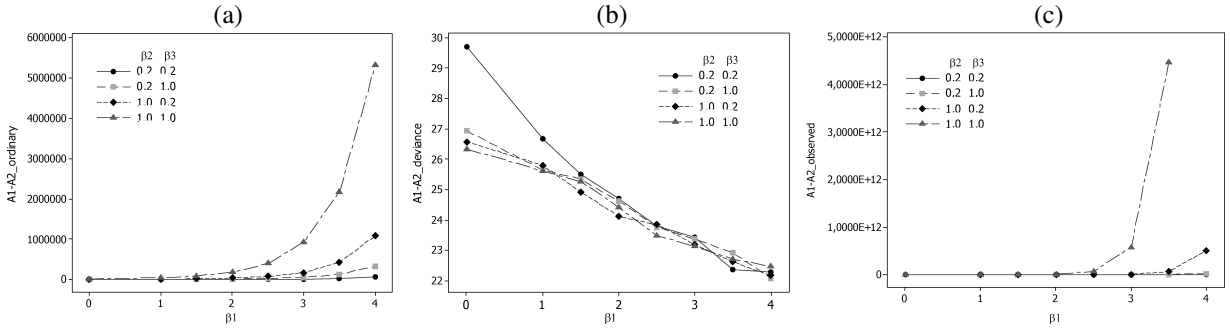
Influence of the residuals to identify the level of heterogeneity according to difference between matched tables $A_1 - A_2$

Following the same scenarios assessed in subsection *Influence of the residuals to identify the level of heterogeneity according to sum of matched tables* $A_1 + A_2$, the study of the influence of ordinary residuals, Figure 3(a). And the residual deviance components, Figure 3(b), as well as the observed frequencies, Figure 3(c), to estimate component $A_1 - A_2$ is illustrated below.

Considering the residual deviance components, Figure 3(b), there is an antagonistic effect in relation to the level of heterogeneity in the table, when compared to its influence upon component $A_1 + A_2$, Figure 1(b). In such contexts, the increase in heterogeneity among the tables caused a reduction in component $A_1 - A_2$, which is interpreted as the sum of squares of the difference among the tables. The results obtained when using residual deviance components, Figure 3(b), are not in accordance with Greenacre (2003), since they mention that the result between two tables, that is, $A_1 - A_2$, will fall upon the differences between the tables variables, matrices, within each row variable and upon the way they vary in between rows.

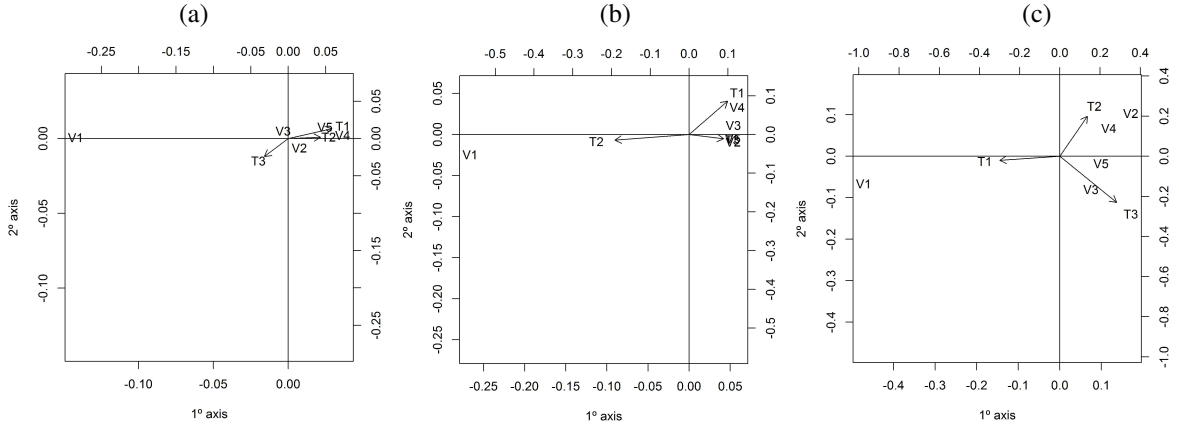
As to biplots, once the ordinary results were obtained, the same behavior happened in relation to the effect of component $A_1 + A_2$. Naturally, the vector sizes and the variable disposition differ from the components mainly, because they have different coordinates, since, for each case, the eigenvalues for $A_1 + A_2$ and $A_1 - A_2$ have been discussed. The difference occurs when using $\beta_{1i} = 0$, because the biplots obtained were asymmetric, Figure 4(a).

Figure 3 – Sum of squares of component $A_1 - A_2$ obtained as a function of different degree of heterogeneity β_1 for combined $\beta_2 = 0.1$ or 1.0 and $\beta_3 = 0.2$ or 1.0, using: (a) ordinary residuals; (b) residual deviance components and (c) observed frequencies.



Source: The authors.

Figure 4 – Biplots related to components $A_1 - A_2$ considering ordinary residuals for fixed $\beta_2 = 0.2$ and $\beta_3 = 0.2$ parameters and different degree of heterogeneity β_1: where in (a) $\beta_1 = 0.0$; (b) $\beta_1 = 2.0$ and (c) $\beta_1 = 4.0$. V_j the variables represented in lines and the T_k variables represented in the columns.



Source: The authors.

It is observed that, despite being asymmetrical, the growing magnitude of the coordinate values on the axes are associated with increased heterogeneity of the tables. Thus, to the degree of heterogeneity zero, the average coordinates of the adjusted ordinary residuals are very close to zero, allowing such tables to have similar frequencies, just like component $A_1 + A_2$, Figure 2(a).

On the other hand, in the case of the increase of the parameter considered for heterogeneity in the tables, there is a gradual increase in the average coordinate, and to $\beta_{1i} = 2.0$, Figure 4(b), approximately twice as much as obtained in $\beta_{1i} = 0$, and approximately four times greater when using $\beta_{1i} = 4.0$, Figure 4(c). It is possible to assess heterogeneity through the magnitude of average coordinates, obtained from 2000 simulations.

Numerical Example

In order to illustrate the methodology proposed, we present an application to sensory analysis, in which we observe the frequency of men and women who consume two kinds of coffee in Paris, New York and Tokyo, as a function of age, Table 4. Two tables with two-way are displayed, where matrix A_1 refers to coffee Bourbon and matrix A_2 to coffee Catuaí.

The adjusted hierarchical log-linear model is given as follows:

$$\begin{aligned}\log(\eta_{ijk}) &= 2.58 + 0.44\beta_{12} + 0.01\beta_{22} + \\ &+0.14\beta_{23} + 0.09\beta_{24} + 0.19\beta_{25} \\ &+0.22\beta_{26} + 0.08\beta_{32} + 0.01\beta_{33}, \quad (14)\end{aligned}$$

where β_{1i}, β_{2j} and β_{3k} are the parameters related to tables (2 types of coffee), rows (6 arrangement between sex and age) and columns (3 cities).

Table 4 – Data on the number of people who consume coffee.

Type of Coffee	Sex and Age	City		
		Paris	NY	Tokyo
Bourbon (A_1)	Men < 30	17	16	15
	Men 30-50	22	18	19
	Men > 50	24	19	27
	Women < 30	25	27	25
	Women 30-50	27	31	28
	Women > 50	28	32	25
Catuaí (A_2)	Men < 30	18	16	22
	Men 30-50	10	18	18
	Men > 50	13	19	18
	Women < 30	15	13	9
	Women 30-50	16	15	9
	Women > 50	14	20	11

Source: The authors.

By means of residual deviance, there is statistical evidence, p-value =0.268. The parameters for table β_1, representing the difference between coffees, was estimated in 0.44 and corresponds to the log of the ratio between sizes for each of the tables, that is, considering n as given by equation (13)

$$\frac{n_{A_1}}{n_{A_2}} = \frac{425}{274} = 1.55 = e^{0.44}. \tag{15}$$

When the general profile evaluation $A_1 + A_2$ for coffee consumption as a function of city, age and sex is gathered, considering ordinary residuals, the sum of squares is 222.16. The singular values corresponding to $A_1 + A_2$ are 13.95 and 5.25, and the third value is close to zero. The contribution to the biplot in two coordinates was 99.9%. When using observed frequencies, the sum of squares was 423.5, with singular values equal to 18.28, 8.16 and 4.77, 94.6% of contribution, Figure 5(a).

The individual behavior does not repeat the expected behavior, when it comes to observed frequencies, as observed when comparing Figure 2(c) and Figure 5(b). In relation to ordinary residuals, because data heterogeneity is considered, the expected behavior was preserved according to the simulated results, Figure 2(a) and Figure 5(a).

As for the general profile evaluation $A_1 - A_2$ for coffee consumption in relation to cities, sex and age, the sum of squares for ordinary residuals was 847.44. The singular values corresponding to $A_1 + A_2$ are 26.19; 11.78 and 4.78. When observed frequencies were used, the sum of squares was 2222.89 and the singular values were 44.87; 12.919 and 6.55, Figure 6(a) and Figure 6(b).

For component $A_1 - A_2$ there was a similarity in behavior for both biplots, when using ordinary residuals, Figure 6(a), and observed frequencies, Figure 6(b). Both are asymmetric and have similar correspondences between variables "row" and "column", and only differed because of the rotation in relation to the first singular value (1st vs.-horizontal axis).

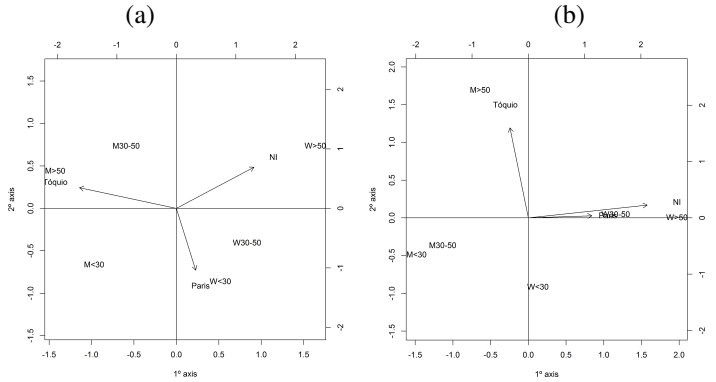
For purposes of comparison, we carried out a singular value decomposition, Figure 7 and Figure 8, having the data obtained from tables as reference, considering the total sum of coffee consumers per city and sex/age $A_1 + A_2$, Table 5 and the difference among consumers $A_1 - A_2$, Table 6.

For the sum of consumers, there are similar biplots, which confirms the efficacy of this methodology, both for observed frequencies and ordinary residuals obtained by adjusting the hierarchical log-linear model, Figure 5(a).

Note that the difference between the consumer's responses were more similar among the biplots, when using observed frequencies and ordinary residuals, Figure 7(a). In general, interpreting biplots is the same relation to the biplots illustrated in Figure 8(b).

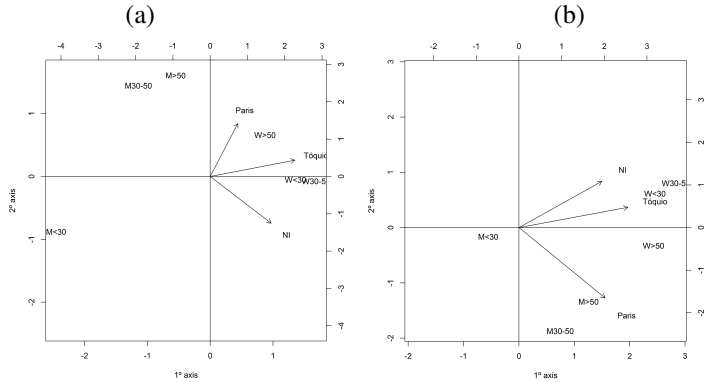
The percentage of explained variability and the singular values obtained in each case previously assessed are displayed in Table 7, where the percentage of sample variation, explained by the biplots, had adequate values, with special attention to component $A_1 + A_2$, where the ordinary residue is considered.

Figure 5 – Biplots related to component $A_1 + A_2$ (Bourbon + Catuaí), where considering: (a) ordinary residuals and (b) observed frequencies.



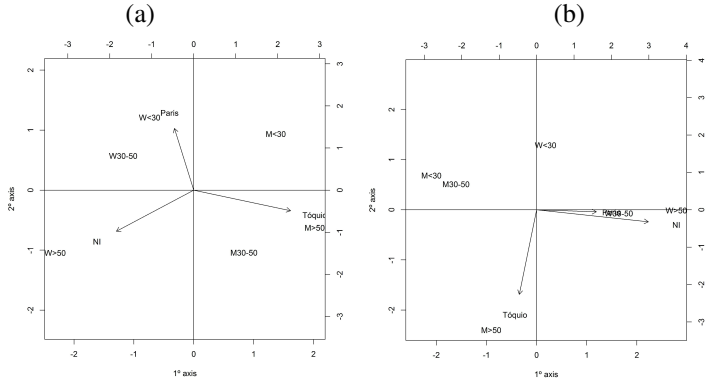
Source: The authors.

Figure 6 – Biplots related to component $A_1 - A_2$ (Bourbon + Catuaí), where considering: (a) ordinary residuals and (b) observed frequencies.



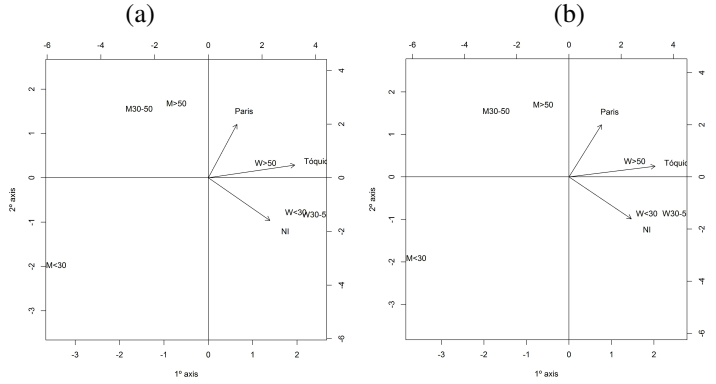
Source: The authors.

Figure 7 – Biplots related to component $A_1 + A_2$ using table data (Bourbon + Catuaí), where considering: (a) ordinary residuals and (b) observed frequencies.



Source: The authors.

Figure 8 – Biplots related to component $A_1 - A_2$ using table data (Bourbon + Catuaí), where considering: (a) ordinary residuals and (b) observed frequencies.



Source: The authors.

Table 5 – Total frequency of people who consume coffee, considering the types of coffee Bourbon and Catuaí.

Sex and age	Sum of observed frequencies			Sum of adjusted ordinary residuals		
	City			City		
	Paris	NY	Tokyo	Paris	NY	Tokyo
Men < 30	35	32	37	0.934	-4.30	3.38
Men 30 - 50	32	36	37	-2.395	-0.66	3.05
Men > 50	37	38	45	-2.310	-3.89	6.21
Women < 30	40	40	34	2.650	0.22	-2.86
Women 30-50	43	46	37	1.730	2.02	-3.74
Women > 50	42	52	36	-0.580	6.62	-6.03

Source: The authors.

Table 6 – Difference among frequencies of number of people who consume coffees Bourbon and Catuaí.

Sex and age	Differences between observed frequencies			Differences between adjusted ordinary residuals		
	City			City		
	Paris	NY	Tokyo	Paris	NY	Tokyo
Men < 30	-1	0	-7	-8.37	-7.84	-14.26
Men 30-50	12	0	1	4.57	-7.92	-6.33
Men > 50	11	0	9	2.51	-9.05	0.63
Women < 30	10	14	16	1.93	5.40	8.04
Women 30-50	11	16	19	2.09	6.50	10.20
Women > 50	14	12	14	4.80	2.20	4.93

Source: The authors.

Table 7 – Difference among frequencies of number of people who consume coffees Bourbon and Catuaí.

Operation carried out in the original tables	Type of data used	Singular values (sv)	Contribution to the first two sv's		
			1st	2nd	3rd
$A_1 + A_2$	Ordinary residue	13.95	5.24	0.01	99.99%
	Observed frequency	18.28	8.16	4.77	94.60%
$A_1 - A_2$	Ordinary residue	26.18	10.51	4.61	97.40%
	Observed frequency	28.37	10.58	4.47	97.87%

Source: The authors.

Conclusion

The use of hierarchical log-linear models is viable to verify the level of heterogeneity among matched two-way tables. The use of ordinary residuals, based in the results described in Figure 1(a). The adjustment of hierarchical log-linear models was also found to be a promising alternative to build and interpret biplots, when compared to biplots with observed frequencies.

Acknowledgments

The authors would like to thank the funding for the realization of this study provided by the Brazilian agencies CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and Fapemig (Fundação de Amparo a Pesquisa do Estado de Minas Gerais - Brasil), Finance Code: APQ 0242-16.

References

ABDI, H.; WILLIANS, L.; VALENTIN, D. Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *Wiley Interdisciplinary reviews*: computational statistics, Hoboken, v. 5, n. 2, p. 149-179, 2013. DOI: https://doi.org/10.1002/wics.1246.

AITCHISON, J.; GREENACRE, M. Biplots of compositional data. *Journal of the Royal Statistical Society*: Series C (Applied Statistics), Oxford, v. 51, n. 4, p. 375-392, 2001.

BÉCUE-BERTAUT, M.; PAGÈS, J. Multiple factor analy sis and clustering of a mixture of quantitative, categorical and frequency data. *Computational Statistics & Data Analysis*, Amsterdam, v. 52, n. 6, p. 3255-3268, 2008. DOI: https://doi.org/10.1016/j.csda.2007.09.023.

BEH, E. J. Simple correspondence analysis using adjusted residuals. *Journal of Statistical Planning and Inference*, Amsterdam, v. 142, n. 4, p. 965-973, 2012. DOI: https://doi.org/10.1016/j.jspi.2011.11.004.

BRZEZINSKA, J. Hierarquical log-linear models for contingency tables. *Acta Universitatis Lodziensis Folia Oeconomica*, Łódź, v. 269, p. 123-129, 2012.

CARLIER, A.; KROONENBERG, P. M. The case of the French cantons: an application of three-way correspondence analysis. *In*: BLASIUS, J.; GREENACRE, M. *Visualization of categorical data*. Cambridge: Academic Press, 1998. p. 253-275.

DOSSOU-GBÉTÉ, S.; GRORUD, A. Biplots for matched two-way tables. *Annales de la Faculté des sciences de Toulouse*: Mathématiques, Toulouse, v. 11, n. 4, p. 469-483, 2002.

FALGUEROLLES, A. GBMs: GLMs with bilinear terms. *In*: BETHLEHEM, J. G.; VAN DER HEIJDEN; P.G.M. *Compstat 2000*. Heidelberg: Physica, 2000. p. 53–64.

FALGUEROLLES, A.; FRANCIS, B. An algorithmic approach to bilinear models for two-way contingency tables. In: DIDAY, E.; LECHEVALLIER, Y.; SCHADER, M.; BERTRAND, P.; BURTSCHY, B. *New approaches in classification and data analysis.* Berlim: Springer Berlin Heidelberg, 1994. p. 518-524.

GREENACRE, M. Biplots in correspondence analysis. *Journal of Applied Statistics*, London, v. 20, n. 2, p. 251 – 269, 1993. DOI: https://doi.org/10.1080/02664769300000021.

GREENACRE, M. Contribution biplots. *Journal of Computational and Graphical Statistics*, Alexandria, v. 22, n. 1, p. 107-122, 2013.

GREENACRE, M. Singular value decomposition of matched matrices. *Journal of Applied Statistics*, London, v. 30, n. 10, p. 1101-1113, 2003. DOI: https://doi.org/10.1080/0266476032000107132.

POWERS, D. A; XIE, Y. *Statistical methods for categorical data analysis*. San Diego: Academic Press, 2000.

R CORE TEAM. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2015. Available from: http://www.R-project.org. Access in: Nov. 10, 2022.

VAN DER HEIJDEN, P. G. M.; DE FALGUEROLLES, A.; DE LEEUW, J. A combined approach to contingency table analysis using correspondence analysis (with Discussion). *Journal of the Royal Statistical Society*: Series C (Applied Statistics), Oxford, v. 38, n. 249–292, 1989.

VAN DER HEIJDEN, P. G. M.; MOOIJAART, A. Some new-bilinear models for the analysis of asymmetry in a square contingency table. *Sociological Methods and Research*, Beverly Hills, v. 24, n. 1, p. 7–29, 1995. DOI: https://doi.org/10.1177/0049124195024001002.

Received: July 6, 2022
Accepted: Nov. 25, 2022
Published: Dec. 17, 2022