The Cressie-Read Divergence Statistic and Digital Data Visualisation

Eric J. Beh[1] and Rosaria Lombardo[2]

[1] School of Mathematical & Physical Sciences, University of Newcastle, Australia
[2] Department of Economics, University of Campania «L. Vanvitelli», Italy

Abstract:
The foundations of correspondence analysis rest with Pearson's chi-squared statistic as the core measure used to assess the association structure between categorical variables. Not only is this statistic extremely popular and versatile, using it for the purposes of performing correspondence analysis ensures that the squared distance between two intra-variable points are Euclidean. More recently, it was shown how the Freeman-Tukey statistic plays a role in correspondence analysis and how such distances can be assessed using the Hellinger distance. Both Pearson's and the Freeman-Tukey statistics are special cases of the Cressie-Read divergence statistic. Therefore, we shall be exploring the features of correspondence analysis where the association, and the resulting low-dimensional visual display, have at its foundations this divergence statistic. By doing so, we shall describe the properties of correspondence analysis when special cases of the divergence statistic (including the log-likelihood ratio statistic and the Cressie-Read statistic) are considered. The applicability of this will be shown by analysing digital data. Such an approach means that a "better" quality two-dimensional visual display can be found that exceeds that of the "best" possible display obtained using the classical approach to correspondence analysis.

Keywords:
Correspondence Analysis; Freeman-Tukey statistic; Likelihood-ratio statistic; Pearson's chi-squared statistic; Singular value decomposition

1. Introduction:

Correspondence analysis (CA) is a popular method used for visualising the association between two or more categorical variables. Such a visualisation assesses the nature of the association by depicting the position of a categories *profile* in a low-dimensional space, typically consisting of two-dimensions. Traditionally, CA relies on using Pearson's chi-squared statistic for assessing the association between the categorical variables. However, the last 20 years or so have seen various amendments made to CA that, while not explicitly saying so, are very much linked to alternative chi-squared measures. For example, Greenacre (2009) proposed his log-ratio analysis which involves the modified log-likelihood ratio statistic as the underlying measure of association. While Greenacre (2009) did not examine this link, he did study the differences in the configuration of points when considering the logarithm transformation of a profile with the un-transformed profile. Cuadras & Cuadras (2006) proposed a "parametric correspondence analysis approach" using a Helllinger Distance Decomposition (HDD) and, while the authors did not do so, it can be shown that this technique uses the Freeman-Tukey statistic.

To demonstrate the link that CA has to the various chi-squared statistics, Beh & Lombardo (2021) demonstrated that the Cressie-Read divergence statistic (Cressie & Read, 1984) can be used as the underlying measure of association. Such a statistic provides a family of chi-

squared statistics which includes as special cases Pearson's statistic, the Freeman-Tukey statistic, the modified log-likelihood ratio statistic amongst others. Therefore, this paper briefly outlines how CA can be performed using the divergence statistic and discusses some of the features that come from such an analysis. A more comprehensive description of this method can be found in Beh and Lombardo (2021).

Our discussion of this approach to CA is made in the following three sections. Section 2 outlines the method and some of its features (which includes defining the principal coordinates and their properties, and some measures of distance) while Section 3 provides a demonstration of this method using data from the 2018 European Social Survey (ESS 2018). Some final comments are left for Section 4.

2. Methodology:

2.1 The Cressie-Read Statistic

Suppose we have an $I \times J$ contingency table, **N**, where the (i, j)th cell entry is denoted by n_{ij}, for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$. Let the grand total of **N** be n and the (i, j)th relative frequency be $p_{ij} = n_{ij}/n$ so that $\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = 1$. Denote the ith row and jth column marginal proportion by $p_{i\bullet} = \sum_{j=1}^{J} p_{ij}$ and $p_{\bullet j} = \sum_{i=1}^{I} p_{ij}$ respectively.

Suppose we define the *Cressie-Read residual* of the (i, j)th cell of **N** to be

$$r_{ij}(\delta) = \frac{1}{\delta}\left(\left(\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}}\right)^{\delta} - 1\right)$$

for some value of δ. Then, the Cressie-Read divergence statistic, $\phi(\delta)$, can be obtained from the weighted sum-of-squares of these residuals so that

$$\phi(\delta) = n\sum_{i=1}^{I}\sum_{j=1}^{J} p_{i\bullet}p_{\bullet j}r_{ij}(\delta)\,.$$

Specific values of δ give us some of the best-known chi-squared statistics. In particular, the modified log-likelihood ratio statistic, the Freeman-Tukey statistic, Cressie-Read statistic and Pearson's statistic can be obtained when $\delta = 0$, $\delta = 1/2$, $\delta = 2/3$ and $\delta = 1$, respectively. While a CA of **N** is typically performed using Pearson's chi-squared statistic ($\delta = 1$) generalisations of the analysis can be performed by considering $\phi(\delta)$. To do so, we consider the SVD of the matrix of *Cressie-Read residuals* such that, for the (i, j)th entry,

$$r_{ij}(\delta) = \sum_{m=1}^{M} a_{im}(\delta)\,\lambda_m(\delta)\,b_{jm}(\delta)\,.$$

Here, $a_{im}(\delta)$ is the ith element of the mth left singular vector while $b_{jm}(\delta)$ is the jth element of the mth right singular vector. These elements are constrained so that

$$\sum_{i=1}^{I} p_{i\bullet}a_{im}(\delta)\,a_{im'}(\delta) = \begin{cases}1 & m = m'\\ 0 & m \neq m'\end{cases} \qquad \sum_{j=1}^{J} p_{\bullet j}b_{jm}(\delta)\,b_{jm'}(\delta) = \begin{cases}1 & m = m'\\ 0 & m \neq m'\end{cases}.$$

The value $\lambda_m(\delta)$ is the mth largest singular value of the matrix of Cressie-Read residuals for a given value of δ. Such values are arranged in descending order so that $1 > \lambda_1(\delta) > \lambda_2(\delta) > \cdots > \lambda_M(\delta) > 0$. The value of M depends on the choice of δ; when $\delta = 0$ or $\delta = 1$ then $M = \min(I, J) - 1$ while $M = \min(I, J)$ for all other values of δ.

The choice of value that δ can take is important. While certain values of δ lead to some of the most popular chi-squared statistics (as we described above), Beh & Lombardo (2021) point out that Greenacre's (2009) log-ratio analysis is a special case of this approach, as is the "parametric correspondence analysis" of Cuadras and Cuadras (2006). For log-ratio analysis, Greenacre (2009) confined δ to lie within the interval [0, 1] so that a comparison could be made between the classical approach to CA (δ = 1) and his log-ratio analysis (when δ = 0). Cuadra & Cuadras (2006) studied the link between classical CA and their HDD method ($\delta = 1/2$) and so confined δ to lie within $[1/2, 1]$. However, by using $\phi(\delta)$ for performing CA virtually any value of δ can be considered. This is because Cressie & Read (2006) were not explicit about what values of δ should be used, citing that even large values (including $\delta = 5$) could be used for detecting departures from independence. However, they did suggest that $0 \leq \delta \leq 3/2$ is appropriate for goodness-of-fit purposes. They also argued that $1/3 \leq \delta \leq 2/3$ provides good coverage when assessing deviations from independence.

2.2 Principal Coordinates

The benefit of using the divergence statistic, $\phi(\delta)$, in CA is that the technique involves a power transformation of the elements of the centred row and column profiles since

$$r_{ij}(\delta) = \frac{1}{\delta p_{\bullet j}^{\delta}}\left(\left(\frac{p_{ij}}{p_{i\bullet}}\right)^{\delta} - p_{\bullet j}^{\delta}\right) = \frac{1}{\delta p_{i\bullet}^{\delta}}\left(\left(\frac{p_{ij}}{p_{\bullet j}}\right)^{\delta} - p_{i\bullet}^{\delta}\right).$$

A visual summary of the association between the row and column variables of **N** can then be made by constructing a two-dimensional (say) correspondence plot. Such a visual summary plots the principal coordinate for each category. The ith row and jth column principal coordinate along the mth dimension of the plot is defined as $f_{im}(\delta) = a_{im}(\delta)\lambda_m(\delta)$ and $g_{im}(\delta) = b_{jm}(\delta)\lambda_m(\delta)$, respectively. Therefore, the ith row and jth column categories can be jointly represented in a two-dimensional plot by plotting $\left(f_{i1}(\delta), f_{i2}(\delta)\right)$ and $\left(g_{j1}(\delta), g_{j2}(\delta)\right)$.

The properties concerning these principal coordinates are as follows. Firstly, it can be shown that

$$\sum_{i=1}^{I} p_{i\bullet} f_{im}(\delta)\, f_{im'}(\delta) = \begin{cases} \lambda_m^2(\delta) & m = m' \\ 0 & m \neq m' \end{cases} \qquad \sum_{j=1}^{J} p_{\bullet j} g_{jm}(\delta)\, g_{jm'}(\delta) = \begin{cases} \lambda_m^2(\delta) & m = m' \\ 0 & m \neq m' \end{cases}$$

so that, irrespective of the choice of δ, the contribution that each dimension makes to the association decreases as $m \to M$. It can also be shown that the total inertia of **N** can be expressed as

$$\frac{\phi(\delta)}{n} = \sum_{i=1}^{I} p_{i\bullet} f_{im}^2(\delta) = \sum_{j=1}^{J} p_{\bullet j} g_{jm}^2(\delta) = \sum_{m=1}^{M} \lambda_m^2(\delta).$$

Therefore, principal coordinates located far from the origin highlight those categories that play an important role in defining the nature of the association between the variables. Points that are close to the origin show those categories that are not deemed to be as dominant. More on the interpretation of distances of points from the origin will be now be discussed.

2.3 Distance Measures

A feature of this approach to CA is that it involves the power transformation of the elements of the row and column profiles; the jth element of the ith centred row profile is $\left(p_{ij}/p_{i\bullet}\right)^{\delta} - p_{\bullet j}^{\delta}$ and the ith element of the jth centred column profile is $\left(p_{ij}/p_{\bullet j}\right)^{\delta} - p_{i\bullet}^{\delta}$. Therefore, the Cressie-Read residual can be expressed in terms of these elements such that

$$r_{ij}(\delta) = \frac{1}{\delta p_{\bullet j}^{\delta}}\left(\left(\frac{p_{ij}}{p_{i\bullet}}\right)^{\delta} - p_{\bullet j}^{\delta}\right) = \frac{1}{\delta p_{i\bullet}^{\delta}}\left(\left(\frac{p_{ij}}{p_{\bullet j}}\right)^{\delta} - p_{i\bullet}^{\delta}\right).$$

So, this approach is consistent with Greenacre's (2009) "power family 2", although he did not establish the link between this "family" and $\phi(\delta)$ and considered $0 \leq \delta \leq 1$. Therefore, the weighted squared Euclidean distance of the ith row profile, say, from the origin is

$$d_I^2(i, 0;\ \delta) = \sum_{j=1}^{J} p_{\bullet j} r_{ij}^2(\delta) = \frac{1}{\delta^2}\sum_{j=1}^{J} \frac{1}{p_{\bullet j}^{2\delta-1}}\left[\left(\frac{p_{ij}}{p_{i\bullet}}\right)^{\delta} - p_{\bullet j}^{\delta}\right]^2 = \sum_{m=1}^{M} f_{im}^2(\delta)\,.$$

Thus, the Cressie-Read divergence statistic can be expressed in terms of this distance by

$$\phi(\delta) = n\sum_{i=1}^{I} p_{i\bullet} d_I^2(i, 0;\ \delta) = n\sum_{i=1}^{I}\sum_{m=1}^{M} p_{i\bullet} f_{im}^2(\delta)$$

This result shows that, irrespective of the value of δ, if all principal coordinates lie at the origin of the correspondence plot, then the total inertia, $\phi(\delta)/n$ and hence the Cressie-Read divergence statistic will be zero. The further away from the origin that a point lies then the more dominant that its category is in defining the nature of the association in **N**.

One can also measure the squared distance between the ith and i'th row (centred) profiles from by

$$d_I^2(i, i';\ \delta) = \frac{1}{\delta^2}\sum_{j=1}^{J} \frac{1}{p_{\bullet j}^{2\delta-1}}\left[\left(\frac{p_{ij}}{p_{i\bullet}}\right)^{\delta} - \left(\frac{p_{i'j}}{p_{i'\bullet}}\right)^{\delta}\right]^2 = \sum_{m=1}^{M}\left(f_{im}(\delta) - f_{i'm}(\delta)\right)^2.$$

For example

$$d_I^2(i, i';\ 0) = 2\sum_{j=1}^{J} p_{*j}\ln\left(\frac{p_{ij}}{p_{i\bullet}}\Big/\frac{p_{i'j}}{p_{i'\bullet}}\right), \qquad d_I^2\left(i, i';\ \frac{1}{2}\right) = 4\sum_{j=1}^{J}\left(\sqrt{\frac{p_{ij}}{p_{i\bullet}}} - \sqrt{\frac{p_{i'j}}{p_{i'\bullet}}}\right)^2,$$

$$d_I^2\left(i, i';\ \frac{2}{3}\right) = \frac{9}{4}\sum_{j=1}^{J} \frac{1}{p_{\bullet j}^{1/3}}\left[\left(\frac{p_{ij}}{p_{i\bullet}}\right)^{2/3} - \left(\frac{p_{i'j}}{p_{i'\bullet}}\right)^{2/3}\right]^2 \qquad d_I^2(i, i';\ 1) = \sum_{j=1}^{J}\frac{1}{p_{\bullet j}}\left(\frac{p_{ij}}{p_{i\bullet}} - \frac{p_{i'j}}{p_{i'\bullet}}\right)^2$$

are the logarithmic, Hellinger, "Cressie-Read" and chi-squared distances between the ith and i'th row profile, respectively. Thus, for these δ, the distance between two row principal coordinates in a correspondence plot can be assessed in terms of these measures of difference between their transformed profiles, thereby satisfying the "property of distributional equivalence". Such a property applies for any value of δ.

3. Result:

We analysed data from the 2018 European Social Survey (ESS2018). By focusing on the "Human Rights" and "socio-demographics" variables, we cross-classified the perceived level of "richness" (on a six point scale ranging from "Very much like me" to "Not like me at all") with the occupation of 19 different types of managers (as classified by ISCO88); see below. This data is available from the URL https://www.europeansocialsurvey.org/data.

We performed four variants of correspondence analysis that stem from the Cressie-Read divergence statistic; when $\delta = 0, 1/2, 2/3$ and 1. In doing so we obtained a two-dimensional correspondence plot for each; see Figure 1. All four plots provide a similar interpretation of the association between the variables, but their quality and interpretation of distances are all very different. Figure 1a) displays the association using the logarithmic distance ($\delta = 0$) between

the row (and column) profiles and visually shows about 72% of this association. By using $\delta = 1/2$ we obtain the same plot when using the technique of Beh, Lombardo & Alberti (2018); see Figure 1b). This plot displays the difference between the profiles using the Hellinger distance and describes about 66% of the association between the variables. Figures 1c) and 1d) are the correspondence plots obtained when $\delta = 2/3$ and $\delta = 1$, respectively, and visually describe 65% and 63% of the association. Of these δ values, the best quality plot arises for $\delta = 0$ so that the association is assessed using the modified log-likelihood ratio statistic. The worst quality display occurs using the classical approach to correspondence analysis ($\delta = 1$)!

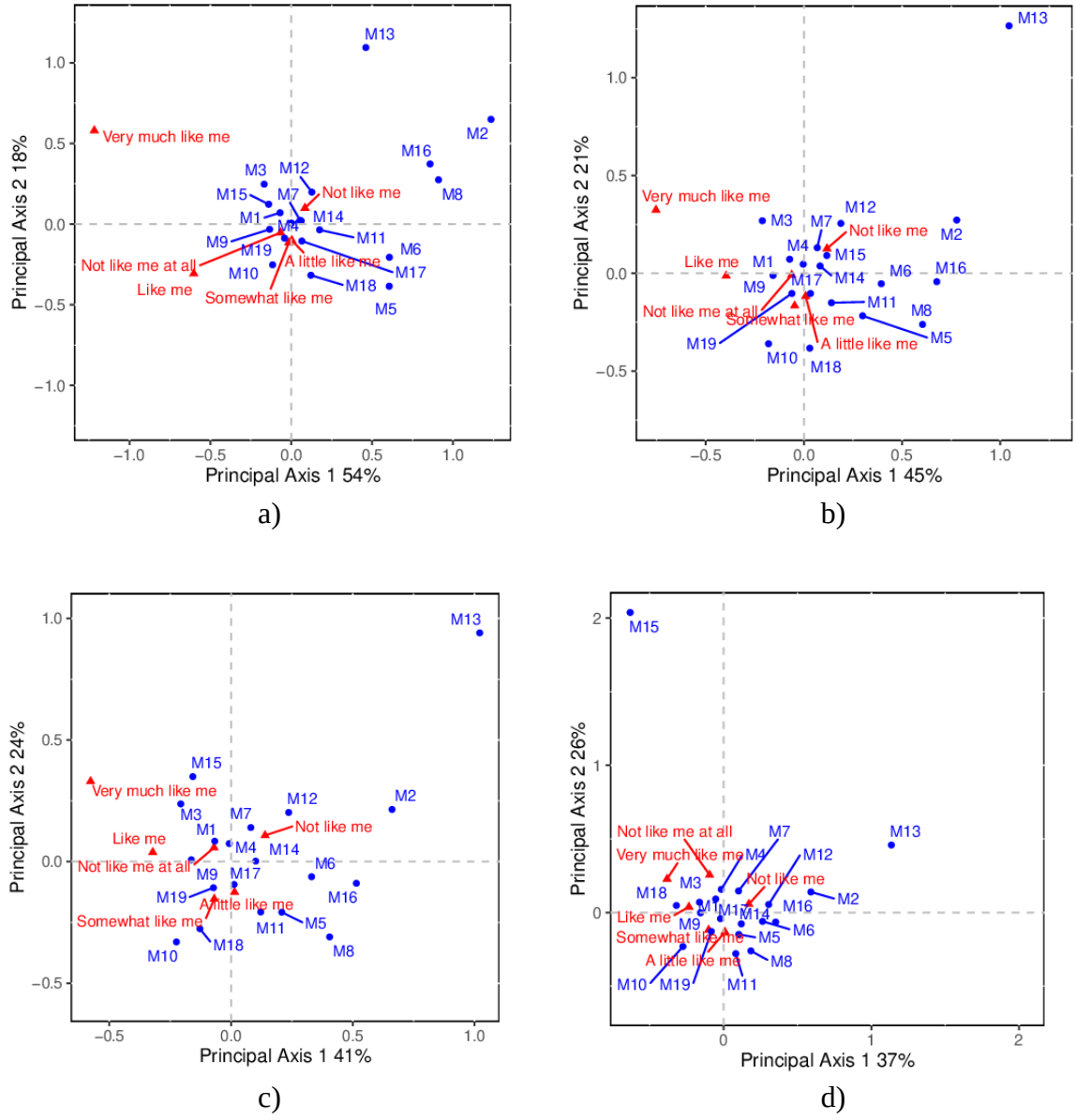

Figure 1: Correspondence plot of the ESS2018 data using the Cressie-Read divergence statistic with a) $\delta = 0$, b) $\delta = 1/2$, c) $\delta = 2/3$ and d) $\delta = 1$.

Suppose we now confine our attention to observing the distance between the row points in the four plots of Figure 1. As we described in Section 2.3, these distances can all be measured by a power transformation of the profiles thereby remaining consistent with Greenacre's (2009) "power family 2" approach. However, each (row) distance measure emphasises the importance of the columns differently. The Hellinger distance is calculated without any influence from the column marginal information; this is one advantage of using this distance and was advocated by many including and Cuadras & Cuadras (2009). The distances between

the row points for the remaining three plots all are all influenced by the columns in some way. The column information is weighted most heavily when calculating the chi-squared distances ($\delta = 1$) while the logarithmic distance ($\delta = 0$) gives the least weight to the columns.

The managers occupation labels are defined as follows:

M1: Managing director & chief executives
M2: Administration and commercial
M3: Business services & administration
M4: Finance
M5: Human resources
M6: Policy & planning
M7: Other business services & admin
M8: Sales, marketing & development
M9: Sales & marketing
M10: Advertising & public relations
M11: Research & development
M12: Production & specialised services
M13: Production managers in agriculture, Forestry & fisheries
M14: Agricultural & forestry production
M15: Aquaculture & fisheries production
M16: Manufacturing, mining, construction & distribution
M17: Manufacturing
M18: Mining
M19: Construction

4. Discussion and Conclusion:

This technique can be used to determine the "best" possible and "worst" possible visual display of the association using correspondence analysis by determining the value of δ that gives the "best" and "worst" quality correspondence plot. The advantage of this technique is that, irrespective of the value of δ, the interpretation of such a visual display can be made in terms of distances that are quantifiable and meaningful and involve power transformations of the centred profiles. One may refer to Beh & Lombardo (2021) for more details on this issue.

While the technique outlined above is confined to the analysis of two cross-classified categorical variables, there is scope for it to be extended for the correspondence analysis of multiple categorical variables. This can be achieved using multivariate extensions of the Tucker3 method of decomposition; see Kroonenberg (2008) for a comprehensive discussion of this method. This could also be achieved by incorporating the extensions made to the divergence statistic outlined in Pardo & Pardo (2003) and Pardo (2010) into the framework described in this paper. Such work is yet to be undertaken and so is an exciting avenue to pursue as the development of correspondence analysis continues to grow.

References:

1. Beh, E.J. & Lombardo, R. (2021). Correspondence analysis and the Cressie-Read divergence statistic. (in review)
2. Cressie, N.A.C. & Read, T.R.C. (1984). Multinomial goodness-of-fit tests. Journal of the Royal Statistical Society, Series B, 46, 440 – 464.
3. Cuadras, C.M. & Cuadras, D. (2006). A parametric approach to correspondence analysis. Linear Algebra and its Applications, 417, 64 – 74.
4. Greenacre, M. (2009). Power transformations in correspondence analysis. Computational Statistics and Data Analysis, 53, 3107 – 3116.
5. Kroonenberg (2008)
6. Pardo, J.A. (2010). An approach to multiway contingency tables based on ϕ-divergence test statistics. Journal of Multivariate Analysis, 101, 2305 – 2319.
7. Pardo, L. & Pardo, M.C. (2003). Minimum power-divergence estimator in three-way contingency tables. Journal of Statistical Computation and Simulation. 73, 819 – 831.
8. Rao, C.R. (1995). A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. Qüestiió, 19, 23 – 63.