Confidence Regions and Approximate P-values for Classical and Non-Symmetric Correspondence Analysis

ERIC J. BEH[1]

ROSARIA LOMBARDO[2]

[1]School of Mathematical and Physical Sciences, University of Newcastle, University Drive, Callaghan, NSW 2308, Australia.

[2]Economics Faculty, Second University of Naples, Capua, CE 81043, Italy

Recently a procedure was developed for constructing 100(1 – α)% confidence ellipses for points in a low-dimensional plot obtained from a classical correspondence analysis. This paper will review the construction of confidence regions for classical and non-symmetric correspondence analysis and propose a simple procedure for determining p-values of each of the points in this space. Such features enable the researcher to determine the statistical significance of a category to the association structure between the categorical variables being analysed. They also reflect the information contained in dimensions higher than those that typically allow for a visual inspection of the association structure.

Keywords: Confidence circle; Confidence ellipse; Correspondence plot; P-value

[1] Address correspondence to Eric J. Beh, School of Mathematical & Physical Sciences, University of Newcastle, Callaghan, NSW, 2308, Australia; Email address: eric.beh@newcastle.edu.au

1. Introduction

The most utilised feature when performing classical (or symmetric) or non-symmetric correspondence analysis on a two-way contingency table is the low dimensional space (referred to as a correspondence plot). Such a plot allows the researcher to visualise the association structure of the categorical variables and is often constructed using two dimensions (sometimes three dimensions are considered). An important issue when visualising the association between categorical variables using any of the correspondence analysis techniques is the need to identify the statistical significance of the proximity of a point from the origin of the plot. Since the origin of a correspondence plot can be interpreted as the position of all the points when there is complete independence between the categorical variables, determining the statistical significance of the distance of a point from the origin is appealing. While Lebart, Morineau and Warwick (1984, pg 182 – 186) proposed the construction of a $100(1-\alpha)\%$ confidence circle for each category to overcome this issue, such circles have only been considered more recently for Beh's (1997) ordered correspondence analysis technique (Beh, 2011) and non-symmetric correspondence analysis (Beh and D'Ambra, 2009). Gower, Lubbe and le Roux (2011) also consider confidence circles (and confidence ellipses) from a biplot perspective. However, since the axes of a correspondence plot are generally weighted differently, elliptical regions offer a more intuitive and appealing alternative. In many practical situations, more than three dimensions are required to reflect all the association that exists between the categorical variables. Recently, Beh (2010) proposed the construction of elliptical regions which allows the researcher to visually represent the association structure using only two dimensions and reflecting all of the information contained in the optimal correspondence plot.

Until now, no such regions have been considered for non-symmetric correspondence analysis. Therefore this paper advances the Lebart et al. (1984) circles and Beh (2010) ellipses in two major ways. Firstly, we shall be adapting the elliptical regions of Beh (2010) for performing non-symmetrical correspondence analysis. Secondly, we shall use the foundations of these circular and elliptical regions to propose ways of calculating approximations of the p-values designed to reflect the statistical significance of the distance of a point from the origin. Such p-values allow one to determine the statistical significance of a category to the association structure between the two categorical variables.

To address these two issues, this paper will be organised as follows. Section 2 will briefly review the distinctions and mathematical issues of classical and non-symmetric

correspondence analysis. Section 3 will provide a review of the issues surrounding the identification of the statistical significance of a point from the origin in a low-dimensional correspondence plot. The development of confidence regions for classical correspondence analysis will be discussed in Section 4 while Section 5 is concerned with the development of these regions for non-symmetric correspondence analysis. Expressions identifying the p-values that reflect the statistical significance of a point from the origin in a classical correspondence plot will be derived in Section 6. Section 7 considers this issue for non-symmetric correspondence analysis. In Section 8, these advances will be applied using data from the study of mother–child attachment of van IJzendoorn (1995). Some final remarks will be left for the discussion (Section 9).

2. The Two Correspondence Analysis Techniques

2.1. Classical Correspondence Analysis

Consider an $I \times J$ two-way contingency table, **N**, where the (i, j)'th cell entry is given by n_{ij} for i = 1, 2, . . . , I and j = 1, 2, . . . , J. Denote the grand total of **N** by n and the (i, j)'th relative frequency by $p_{ij} = n_{ij}/n$. Define the i'th row relative marginal frequency by $p_{i\bullet} = \sum_{j=1}^{J} p_{ij}$ and the j'th column relative marginal frequency by $p_{\bullet j} = \sum_{i=1}^{I} p_{ij}$.

To obtain a visual summary of the structure of the association between the variables one may perform classical correspondence analysis. To do so, consider decomposing the matrix of Pearson residuals using generalised singular value decomposition (Beh, 2004) such that

$$\frac{p_{ij}}{p_{i\bullet}p_{\bullet j}} = 1 + \sum_{m=1}^{M} a_{im}\lambda_m b_{jm}$$

where $M = \min(I, J) - 1$. Here a_{im} is the m'th element of the row singular vector associated with the i'th row profile. Similarly b_{jm} is the m'th element of the column singular vector associated with the j'th column profile. These elements are constrained such that

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

The value λ_m is the m'th singular value of the standardised residuals. As usual, the singular values are arranged in descending order such that $1 > \lambda_1 \geq \lambda_2 \geq \ldots \geq 0$.

To graphically view the association between the row and column profile coordinates in a low $(< M)$ dimensional space the i'th row profile and j'th column profile may be simultaneously represented by the principal coordinates

$$f_{im} = a_{im}\lambda_m \qquad \text{and} \qquad g_{jm} = b_{jm}\lambda_m \tag{2}$$

respectively. In the case where a plot consists of all M dimensions, it is referred to as the optimal correspondence plot. Based on these results, the Pearson chi-squared statistic of **N** can be expressed in terms of the coordinates by

$$X^2 = n\phi^2 = n\sum_{m=1}^{M}\lambda_m^2 = n\sum_{m=1}^{M}\sum_{i=1}^{I} p_{i\bullet} f_{im}^2 = n\sum_{m=1}^{M}\sum_{j=1}^{J} p_{\bullet j} g_{jm}^2 \, . \tag{3}$$

where the weight, or principal inertia, associated with the m'th axis is λ_m^2. When performing classical correspondence analysis the underlying structure of the association between the row and column variables is assumed to be symmetric. That is, they are both considered to be associated such that neither of them is regarded as a response variable of another variable. The reader is directed to, for example, Greenacre (1984) and Beh (2004) for a more comprehensive mathematical description of the issues underlying classical correspondence analysis. Clausen (1988) and Greenacre and Blasius (1994) may be considered for a more applied focus.

2.2. Non-symmetric Correspondence Analysis

Suppose now we treat the column variable as a predictor variable and the row variable as its response variable. For such an asymmetrically associated variable structure, non-symmetrical correspondence analysis can be used to provide a graphical summary of the row and column points; see, for example, D'Ambra and Lauro (1989). For such a variable structure we may consider the generalised singular value decomposition of the following

$$\tilde{r}_{ij} = \frac{p_{ij}}{p_{\bullet j}} - p_{i\bullet} = \sum_{m=1}^{M} \tilde{a}_{im}\tilde{\lambda}_m\tilde{b}_{jm} \, .$$

The values $\tilde{a}_{im}$ and $\tilde{b}_{jm}$ are akin to the a_{im} and b_{jm} values, respectively, of classical correspondence analysis. These quantities have the property

$$\sum_{i=1}^{I} \tilde{a}_{im}\tilde{a}_{im'} = \begin{cases} 1, & m = m' \\ 0, & m \neq m' \end{cases} \quad \text{and} \quad \sum_{j=1}^{J} p_{\bullet j}\tilde{b}_{jm}\tilde{b}_{jm'} = \begin{cases} 1, & m = m' \\ 0, & m \neq m' \end{cases}.$$

As was the case when considering a symmetrically associated variable structure between the rows and columns, the singular values, $\left(\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_M\right)$ are arranged in descending order. Therefore, the i'th row (response) category and the j'th column (predictor) category along the m'th axis of a correspondence plot is defined in terms of principal coordinates

$$\tilde{f}_{im} = \tilde{a}_{im}\tilde{\lambda}_m \quad \text{and} \quad \tilde{g}_{jm} = \tilde{b}_{jm}\tilde{\lambda}_m$$

respectively. For our asymmetric variable structure, the aim is to depict the prediction of the rows given the columns in a low $(< M)$ dimensional space where the Goodman-Kruskal tau index (Goodman and Kruskal, 1954)

$$\tau_{GK} = \frac{\tau_{num}}{1 - \sum_{j=1}^{J} p_{\bullet j}^2}$$

where

$$\tau_{num} = \sum_{i=1}^{I}\sum_{j=1}^{J} p_{\bullet j}\left(\frac{p_{ij}}{p_{\bullet j}} - p_{i\bullet}\right)^2$$

is used as the asymmetric measure of association. For non-symmetric correspondence analysis, the variation of the row and column categories can be measured using the numerator of the index such that

$$\tau_{num} = \sum_{i=1}^{I}\sum_{j=1}^{J} p_{\bullet j}\left(\frac{p_{ij}}{p_{\bullet j}} - p_{i\bullet}\right)^2 = \sum_{m=1}^{M}\tilde{\lambda}_m^2 = \sum_{m=1}^{M}\sum_{i=1}^{I}\tilde{f}_{im}^2 = \sum_{m=1}^{M}\sum_{j=1}^{J} p_{\bullet j}\tilde{g}_{jm}^2 .$$

A test the statistical significance of the association structure in this case can be made by considering the C-statistic of Light and Margolin (1962)

$$C = \frac{(n-1)(I-1)\tau_{num}}{1-\sum_{i=1}^{I} p_{i\bullet}^2} \sim \chi^2_{\alpha,(I-1)(J-1)} \,.$$

where $\chi^2_{\alpha,(I-1)(J-1)}$ is the $1 - \alpha$ percentile of a chi-squared distribution with $(I - 1)(J - 1)$ degrees of freedom. Note that the C-statistic may be expressed as the (weighted) sum of squares of the (column and) row coordinates such that

$$C = \frac{(n-1)(I-1)}{\left(1-\sum_{i=1}^{I} p_{i\bullet}^2\right)} \sum_{m=1}^{M}\sum_{i=1}^{M} f_{im}^2 = \frac{(n-1)(I-1)}{\left(1-\sum_{i=1}^{I} p_{i\bullet}^2\right)} \sum_{m=1}^{M}\sum_{j=1}^{J} p_{\bullet j}\tilde{g}_{jm}^2 \,.$$

Further information on the mathematical and practical issues concerned with non-symmetric correspondence analysis may be found by referring to, for example, Kroonenberg and Lombardo (1998, 1999), Lombardo, Beh and D'Ambra (2007) and Beh, Lombardo and Simonetti (2011).

3. On the Statistical Significance of a Categorical Point

When performing a classical or a non-symmetric correspondence analysis one needs to consider the proximity of a categorical point from other points in the same space (there are various issues on this point that we will not consider here) and the proximity of a point from the origin of the correspondence plot. It is this second issue that we shall direct our focus and thus the interpretation of the origin is of fundamental importance in this paper; the origin coincides with where all the points would be if there was complete independence in the contingency table. Therefore, points located close to the origin indicate that those categories do not play a major role in describing the association structure of the variables. On the other hand, the further a point lies from the origin, generally, the more important this category is for describing the association structure between the variables. An obvious question then is

> "*how close (or far) from the origin does a point need to be, with any amount of confidence, before the category becomes a statistically significant contributor to the association structure?*"

The following sections demonstrate ways in which this question may be answered.

4. Confidence Regions for Classical Correspondence Analysis

4.1. Confidence Circles

For a two-way contingency table Lebart et al. (1984, pg 182 – 186) proposed a simple answer to the question raised in Section 3. They showed that the radii length of the 95% confidence circle for the i'th row category in a two-dimensional correspondence plot is

$$r_{i(0.05)} = \sqrt{\frac{5.99}{n_{i\bullet}}}$$

where 5.99 represents that 95'th percentile of the chi-squared distribution with two degrees of freedom; this value reflects the two dimensions used to graphically depict the association between the row and column variables. More generally, the $100(1 - \alpha)\%$ confidence circle in a two-dimensional correspondence plot for the i'th row category using classical correspondence analysis is

$$r_{i(\alpha)} = \sqrt{\frac{\chi^2_{\alpha,2}}{n_{i\bullet}}}\,. \tag{4}$$

Confidence circles constructed in this manner allow the user to identify the statistical significance of those points in a two-dimensional correspondence plot that contribute to the association structure of the categorical variables being considered. In practice, if the origin is included within the circle then that particular category does not contribute to the association structure between the variables. Conversely, a confidence circle that does not include the origin means that, at the specified level of significance, the category to which it is related, contributes to the association structure.

4.2. Confidence Ellipses

A disadvantage of Lebart et al's (1984) confidence circles is that they do not take into consideration the unequal weighting of the axes of a correspondence plot; these weights being the principal inertia values. A second limitation is that they do not take into consideration the information contained in the higher dimensions of a plot. To overcome these two problems, the simple approach of constructing Beh's (2010) confidence ellipses can be considered.

Suppose we consider a two-dimensional (D = 2) correspondence plot. Beh (2010) proposed as an alternative to the circular regions of Lebart et al. (1984). For the i'th row category, a 100(1 – α)% confidence ellipse in a two-dimensional plot can be constructed with a semi-axis length along the m'th principal axis of

$$x_{im(\alpha)} = \lambda_m \sqrt{\frac{\chi_\alpha^2}{X^2}\left(\frac{1}{p_{i\bullet}} - \sum_{m=3}^{M} a_{im}^2\right)} \tag{5}$$

for m = 1, 2 where χ_α^2 is the chi-squared statistic with (I – 1)(J – 1) degrees of freedom at the α level of significance. Here, $x_{i1(\alpha)}$ is the semi-major axis length of the confidence ellipse while $x_{i2(\alpha)}$ is the semi-minor axis length of the ellipse. Ellipsoids can be constructed for three- or higher- dimensional correspondence plots by considering m > 2. Constructing confidence ellipses using $x_{im(\alpha)}$ takes into account the information of the i'th row coordinate in dimensions higher than the second. If the information contained in the third and higher dimensions is minimal, or (for some reason) is ignored, then the semi axis length along the m'th dimension is

$$\tilde{x}_{im(\alpha)} = \lambda_m \sqrt{\frac{\chi_\alpha^2}{X^2 p_{i\bullet}}}\,. \tag{6}$$

For more information on the link between (4) and (5) refer to Beh (2010). The following section briefly describes the comparison of confidence ellipses for correspondence plots of varying dimension and of varying levels of significance (α). We will then progress on to the calculation of p-values for each category using the foundations of the regions proposed by Lebart et al. (1984) and Beh (2010).

It must be noted that the construction of confidence regions has also been a topic of discussion in the past. Ringrose (1992, 1996) explored the use of bootstrapping for the construction of convex hulls defining a confidence region. Recently Ringrose (2011) provided a comparison of Beh's (2010) confidence ellipses with ellipses generated through bootstrapping for a two-dimensional display and showed via example that, while their construction is based on rather different approaches, the regions are equivalently interpretable. Greenacre (2007, p. 196–197) also describes the implementation of

bootstrapping to construct confidence regions by implementing a peeling step to remove the impact of 5% of the more extreme outlying replicates.

4.3. Comparison of Confidence Regions

It should be apparent that if one changes the level of significance then this will impact upon the radii length of a confidence circle and the semi-axis length for a confidence ellipse. Suppose we consider a comparison of the semi-axis length of a 100(1 – α)% confidence ellipse and a 100(1 – α')% confidence ellipse. The ratio between these two quantities when considering the association structure in the optimal correspondence plot is

$$r(\alpha, \alpha') = \frac{x_{im(\alpha)}}{x_{im(\alpha')}} = \frac{\lambda_m \sqrt{\frac{\chi^2_\alpha}{X^2}\left(\frac{1}{p_{i\bullet}} - \sum_{m=3}^{M} a^2_{im}\right)}}{\lambda_m \sqrt{\frac{\chi^2_{\alpha'}}{X^2}\left(\frac{1}{p_{i\bullet}} - \sum_{m=3}^{M} a^2_{im}\right)}} = \sqrt{\frac{\chi^2_\alpha}{\chi^2_{\alpha'}}} \ .$$

Therefore, when $\alpha < \alpha'$ then $r(\alpha, \alpha') > 1$ since $\chi^2_\alpha > \chi^2_{\alpha'}$. For example, consider a two-way contingency table consisting of four row categories and four column categories, so that the chi-squared statistic considered has nine degrees of freedom. Then, the semi-axis length for the m'th dimension with a 0.01 level of significance, when compared with the semi-axis length with a 0.05 level of significance, changes by a factor of

$$r(0.01, 0.05) = \sqrt{\frac{\chi^2_{0.01}}{\chi^2_{0.05}}} = \sqrt{\frac{21.666}{16.919}} = 1.132 \ .$$

That is, the semi-axis length for a 99% confidence ellipse along all dimensions will be 1.132 times longer than its corresponding semi-axis length for a 95% confidence ellipse. Similarly, for such a contingency table, the semi-axis length of a 99% confidence ellipse will be 1.215 times longer than the semi-axis length of a 90% confidence ellipse. Depending on the magnitude of the semi-axis length such differences may appear minimal or even quite large. In particular, when a category has a dominant role in the structure of the association (so that the area of the confidence ellipse is small) the impact of changing the level of significance may be quite small. However, if a category is not a statistically significant contributor to the association structure (so that the area of the confidence ellipse is quite large), the impact of changing the level of significance may be quite large. Such a ratio that is given by $r(\alpha, \alpha')$

may also be shown to be applicable to the confidence circles of Lebart et al. (1984) and also to the confidence ellipse (if the association structure reflected in the third and higher dimensions is ignored).

One may also compare the semi-axis length when considering a 100(1 – α)% confidence ellipse using only the first two dimensions with the semi-axis length of the ellipse that reflects the association in the optimal correspondence plot. Such a ratio may be defined as

$$q(2, M) = \frac{x_{im(\alpha)}}{\tilde{x}_{im(\alpha)}} = \frac{\lambda_m \sqrt{\frac{\chi^2_\alpha}{X^2}\left(\frac{1}{p_{i\bullet}} - \sum_{m=3}^{M} a^2_{im}\right)}}{\lambda_m \sqrt{\frac{\chi^2_\alpha}{X^2 p_{i\bullet}}}} = \sqrt{1 - p_{i\bullet} \sum_{m=3}^{M} a^2_{im}} \; .$$

Therefore, if the optimal correspondence plot consists of M = 2 dimensions then such a ratio becomes $q(2, 2) = 1$ and there is no change in the size of the confidence ellipse. However, if in the more general case when $2 < D \leq M$, $q(2, D) < 1$. Therefore, when considering the association structure reflected in the optimal correspondence plot, the semi-axis length along the m'th principal axis is $\sqrt{1 - p_{i\bullet} \sum_{m=3}^{M} a^2_{im}} = \sqrt{1 - p_{i\bullet} \sum_{m=3}^{M} f^2_{im} / \lambda^2_m}$ times shorter than the semi-axis length when only considering the information in the first two dimensions. This suggests that if the coordinate of the i'th row category lies close to the origin (relative to the magnitude of the principal inertia along each of the higher dimensions) for all high (> 2) dimensions then there is very little change in the construction of the confidence ellipse. However, if the category has a dominate high (> 2) dimension then the area of the confidence ellipse in the two-dimensional correspondence plot will decrease.

5. Confidence Regions for Non-Symmetric Correspondence Analysis

In the previous section, we have described the construction of confidence regions for classical correspondence analysis. When it is know (or assumed) that an asymmetric association structure exists between the two categorical variables one may instead consider non-symmetric correspondence analysis. While such a method of correspondence analysis has received very little attention when compared to its "symmetric" kin, confidence regions for non-symmetric correspondence analysis have recently been discussed in the literature. When the column variable is treated as the predictor variable and the row variable is treated as a

response variable, Beh and D'Ambra (2009) showed that the radii length of the 95% confidence circle for the i'th row (response) category in the two-dimensional plot will be

$$r_j^J = \sqrt{\frac{5.99\left(1-\sum_{i=1}^{I} p_{i\bullet}^2\right)}{p_{\bullet j}(n-1)(I-1)}} \ . \tag{7}$$

Lombardo, Beh and D'Ambra (2007) also derived such a radii length for the non-symmetric correspondence analysis of a contingency table consisting of ordered categorical variables. The radii length (7) is akin to (4) and so only considers the asymmetric association structure between the two variables that is reflected in a two-dimensional non-symmetric correspondence plot. The interpretation of the confidence circles derived by considering this radii length is analogous to the radii length (4) for classical correspondence analysis.

To reflect the association structure in dimensions higher than the second, and to reflect the different weighting of each of the principal axes, the 100(1–α)% confidence ellipse for the i'th row (response) category is constructed using a semi-axis length along the m'th principal axis of

$$x_{im(\alpha)} = \lambda_m \sqrt{\frac{\chi_\alpha^2}{\tau_{num}} \frac{\left(1-\sum_{i=1}^{I} p_{i\bullet}^2\right)}{(n-1)(I-1)} \left(1-\sum_{m=3}^{M} a_{im}^2\right)} \ . \tag{8}$$

One may also consider the change in the size of the ellipse when using the level of significance α and α', or comparing the ellipses in a two-dimensional and optimal correspondence plot. The comments made in Section 3.3 are pertinent here. Crisci and D'Ambra (2011) derive analogous semi-axis lengths for the purposes of conducting a multiple non-symmetric correspondence analysis of manufacturing enterprises in the Campania region of Italy.

6. Approximate P-values and Classical Correspondence Analysis

6.1 Approximate P-values and Confidence Circles

The theory concerned with the construction of 100(1 – α)% confidence circles for a row and column coordinate in a correspondence plot may be amended for deriving an approximation of the p-value for this point. By doing so, one may determine the statistical significance of a

row (or column) category to the association structure between the variables using such a value. P-values may also be achieved by considering Ringrose's (2011) bootstrapping approach, but we shall leave this for future consideration.

To derive an approximate p-value, we first consider the null and alternative hypotheses under which it will be generated. As described above, the relative distance of a row, or column, profile coordinate from the origin of the correspondence plot reflects the variation of the category associated with that coordinate from the hypothesis of complete independence. Therefore, the contribution to the chi-squared statistic, or alternatively the total inertia, of the i'th row profile point can be made by considering its proximity from the origin. This suggests that the null and alternative hypothesis of the i'th row profile is

$$\begin{aligned} H_0 &: f_{im} = 0 \\ H_A &: f_{im} \neq 0 \end{aligned}'$$

for m = 1, 2, . . . , M, may be considered for identifying the statistical significance of the row profile coordinate to the association structure of the two variables forming the contingency table. For such hypotheses, we may consider

$$X^2_{i,\,M} = np_{i\bullet}\sum_{m=1}^{M} f^2_{im}$$

as the test statistic that is tested against the $1 - \alpha$ percentile of the chi-squared distribution with M degrees of freedom. Such a test statistic reflects the position of the coordinate in the optimal correspondence plot which consists of M dimensions. Often, a subset of D < M dimensions are used to visually represent the association between the row and column variables; typically D = 2 or D = 3. Therefore,

$$(p-\text{value})_{i,\,D} = P\left\{\chi^2 > X^2_{i,\,D}\right\} \approx P\left\{\chi^2 > np_{i\bullet}\sum_{m=1}^{D} f^2_{im}\right\} \qquad (9)$$

and is the p-value of i'th row profile coordinate in a D-dimensional correspondence plot. Therefore a p-value that is less than the specified level of significance provides evidence that the category does play a statistically significant role in describing the association structure since it is deemed that the particular point in the configuration is not consistent with zero.

One may note that, while (9) takes into consideration the proximity of a point from the origin, it ignores the magnitude of the principal inertia's for each of the D axes in the sub-optimal correspondence plot. As we shall see in the following section, we can amend (9) so that this problem is overcome.

6.2. Approximate P-values and Elliptical Regions

Consider now the elliptical regions generated using the semi-axis length defined by (5). Suppose that a D (< M) – dimensional correspondence plot is used to simultaneously represent the association between the row and column points. Equation (5) suggests that

$$\chi^2 \sim n\phi^2\left(\frac{1}{p_{i\bullet}} - \sum_{m=D+1}^{M} a_{im}^2\right)^{-1} \sum_{m=1}^{D}\left(\frac{f_{im}}{\lambda_m}\right)^2 . \tag{10}$$

Thus, if the information contained in dimensions D + 1, D + 2, . . . , M is reflected in the construction of a confidence interval (as it is when considering (5)), the p-value of the i'th row point in a D-dimensional correspondence plot may be approximated by

$$(p-\text{value})_{i,D} = P\left\{\chi^2 > n\phi^2\left(\frac{1}{p_{i\bullet}} - \sum_{m=D+1}^{M} a_{im}^2\right)^{-1} \sum_{m=1}^{D}\left(\frac{f_{im}}{\lambda_m}\right)^2\right\} . \tag{11}$$

where the subscript (i, D) refers to the i'th row category for which the p-value is approximated using a D-dimensional correspondence plot. Since, from (2), $a_{im} = f_{im}/\lambda_m$, the p-value (11) may be expressed in terms of the principal coordinates in the optimal correspondence plot by

$$(p-\text{value})_{i,D} = P\left\{\chi^2 > n\phi^2 p_{i\bullet}\left(1 - p_{i\bullet}\sum_{m=D+1}^{M}(f_{im}/\lambda_m)^2\right)^{-1} \sum_{m=1}^{D}(f_{im}/\lambda_m)^2\right\} .$$

Due to the inclusion $\sum_{m=D+1}^{M}(f_{im}/\lambda_m)^2$, the proximity of a point from the origin in dimensions higher than the D'th is reflected in this p-value. Thus, by considering the elliptical regions and p-values described here a two- (or even three-) dimensional correspondence plot can be constructed to reflect the significance of a category that would otherwise require an optimal correspondence plot to view the association structure. One may also note that if the coordinate of the i'th row category in the optimal correspondence plot lies

at the origin, the p-value is unity since $P\{\chi^2 > 0\} = 1$. Conversely, if the position of a point lies at a distance from the origin (in the optimal correspondence plot) then the p-value is, approximately, $P\{\chi^2 > n\phi^2\} = P\{\chi^2 > X^2\}$.

In many practical situations, the first D dimensions may be sufficient to adequately describe the association structure between the categorical variables. In this case, one may ignore the higher dimensions without any significant loss of information of this structure. In this case, the p-value for the i'th row category as given by (11) may be amended to give

$$(p-\text{value})_{i,D} = P\left\{\chi^2 > n\phi^2 p_{i\bullet} \sum_{m=1}^{D} (f_{im} / \lambda_m)^2\right\} . \qquad (12)$$

It is apparent from considering the p-values of (11) and (12) that, unlike (9), the inequality of the principal inertia values can be taken into consideration, thereby reflecting the relative weighting of each of the axes in a D – dimensional correspondence plot.

7. Approximate P-values and Non-Symmetric Correspondence Analysis

We may follow the same arguments made above to derive approximations of the p-value of row (response) and column (predictor) points in a low, or optimal, correspondence plot obtained from performing non-symmetric correspondence plot.

Suppose we consider this time the derivation of the approximate p-value for the j'th column (predictor) category that reflects the association structure contained in the optimal correspondence plot. In a manner that is analogous to that described in Section 6, for a D-dimensional non-symmetric correspondence plot, this value is

$$(p-\text{value})_{i,D} = P\left\{\chi^2 > \frac{(n-1)(I-1)\tau_{num}}{1-\sum_{i=1}^{I} p_{i\bullet}^2}\left(1-\sum_{m=D+1}^{M}\left(\frac{\tilde{g}_{jm}}{\tilde{\lambda}_m}\right)^2\right)^{-1}\sum_{m=1}^{D}\left(\frac{\tilde{g}_{jm}}{\tilde{\lambda}_m}\right)^2\right\} . \qquad (13)$$

Therefore, by considering the elliptical regions these p-values, a two-dimensional correspondence plot can be constructed to reflect the statistical significance of a category that would otherwise require an optimal correspondence plot to view the asymmetric association structure. If only the first two dimensions are considered when constructing the elliptical regions for non-symmetric correspondence plot (so that D = 2) then, for the j'th column predictor category, the approximate p-value of its point is

$$(\text{p}-\text{value})_{\text{i},2} = \text{P}\left\{\chi^2 > \frac{(\text{n}-1)(\text{I}-1)\tau_{\text{num}}}{1-\sum_{\text{i}=1}^{\text{I}} \text{p}_{\text{i}\bullet}^2}\left[\left(\frac{\tilde{\text{g}}_{\text{j}1}}{\tilde{\lambda}_1}\right)^2 + \left(\frac{\tilde{\text{g}}_{\text{j}2}}{\tilde{\lambda}_2}\right)^2\right]\right\}. \qquad (14)$$

Expressions for the calculation of these p-values for the row (response) categories may be analogously derived.

8. Application

8.1. The Data

Consider the two-way contingency table of Table 1 that cross-classifies a mother's attachment to her child, and the child's response to their mother's level of attachment. The column variable is therefore defined as *Mothers Attachment Classification* and the row variable is defined as *Infant Response*. The data are based on an extensive study of mother–child attachment conducted by van IJzendoorn (1995). The four column categories are a result of the adult attachment interview (George et al., 1985) while the four row categories are observed from the Ainsworth strange situation (Ainsworth et al., 1978). Table 1 was analysed using non-symmetric correspondence analysis by Kroonenberg and Lombardo (1999) and was considered by Beh (2010) and Ringrose (2011) in their derivation of elliptical confidence regions.

Table 1 about here

While it is apparent that the association structure between *Mothers Attachment Level* and *Infant Response* may be treated asymmetrically, we shall first consider that the association structure between the variables is symmetric. In doing so, we shall highlight the confidence regions and p-values of Table 1 by performing a classical correspondence analysis.

8.2. Classical (Symmetric) Correspondence Analysis

By considering the row and column variables to be symmetrically associated, the Pearson chi-squared statistic is 252.3982 and has a p-value <0.0001. Therefore, there is ample evidence to conclude that an association exists between the two variables. By performing a correspondence analysis on Table 1, the first principal inertia is $\lambda_1^2 = 0.249$ and the second principal inertia is $\lambda_2^2 = 0.165$. Together these two values account for 89.9% of the association that exists between the two variables, and this association structure is reflected in the two-dimensional correspondence plot of Figure 1. Superimposed on this figure are the 95% confidence circles of Lebart et al. (1984) for each row and column point.

Figure 1 suggests that, with the exception of *Resistant* (whose circular region overlaps the origin), all of the row and column categories contribute to the association structure between the two variables. It is also evident that the confidence circle of *Preoccupied* includes the origin within the region. This suggests that the p-value for *Resistant* and *Preoccupied* is more than 0.05. The row category p-values based on the confidence circles of Lebart et al. (1984), obtained using (9) when D = 2, are summarised in the second column of Table 2. Similarly, the column category p-values of the contingency table are summarised in the second column of Table 3. It can be seen that the p-value of *Resistant* is 0.287. However, the p-value of *Preoccupied* is 0.0203, *less* than 0.05. This suggests that perhaps the confidence circles of Lebart *et al.* (1984) are not effective for monitoring the statistical significance of a category from the hypothesis of no association. This may be due to the equal weighting that is assumed of each of the axis when constructing Lebart *et al's* (1984) confidence circles.

Figure 1 about here

As described above, the confidence circles of Lebart et al. (1984) are constructed by assuming that the principal inertias of the first two dimensions are equal. Since the two values are quite different, one may instead consider the elliptical regions proposed by Beh (2010).

These regions appear in Figure 2 but reflect only the information contained in the first two dimensions. Note that the relative size of these regions is consistent with those that appear in Figure 1. By taking into consideration the unequal weighting of the two principal axes, the region for *Preoccupied* again includes the origin which suggests that this column category does not play a statistically significant role in the association structure between the two variables. Indeed, the p-values associated with elliptically generated regions reflect the importance (or not) of these categories. The third column of Table 2 summarises the p-values of the row categories and are calculated using (12) where D = 2, while the third column of Table 3 provides those p-values of the column categories. Table 2 shows that, when considering the first two dimensions only, the p-value for *Resistant* is 0.802 indicating that this particular row category does not play a significant part in the association. However, the remaining three row categories, which have a p-value less than 0.001 do play a significant role in the association structure; in the following Tables a zero p-value represents those categories with a p-value less than 0.001. Similarly, Figure 3 shows that, at the 0.05 level of significance, *Preoccupied* (which has a p-value of 0.108 when considering only the first two dimensions) does not play a significant role in the association structure between the variables of Table 1.

Figure 2 about here

These p-values ignore the association reflected by the third (and, in general, higher) axis. However, such information may be reflected by considering the p-values derived from (11) where D = 2. In such a case, all of the row and column categories have a p-value that is less than 0.001 thus concluding that all of these categories play a statistically significant role in the association between the row and column variables of Table 1. The elliptical regions of Figure 3 for both sets of categories provide a graphical representation of this result.

Figure 3 about here

The key difference between Figure 2 and Figure 3 is the size of the elliptical region for *Resistant* and *Preoccupied*. This suggests that these two categories have a non-zero coordinate in the third dimension of the optimal correspondence plot; Figure 2 of Beh (2010) confirms that this is the case. As a result, the p-value of these categories dramatically changes due to the inclusion of the additional information on the association structure contained in the higher dimension – in both cases reducing the p-value from a relatively large quantity to less than 0.001. Such a dramatic change in the conclusions yielded from the regions and p-values shows that it is important to reflect the information contained in higher dimensions rather than relying on findings of the association structure on just the first two dimensions as it is typically done when performing correspondence analysis.

Table 2 about here

Table 3 about here

To validate the statistical significance of each of the categories described above, we may consider

$$r_{ij} = \sqrt{n}\,\frac{p_{ij} - p_{i\bullet}p_{\bullet j}}{\sqrt{p_{i\bullet}p_{\bullet j}\left(1-p_{i\bullet}\right)\left(1-p_{\bullet j}\right)}}.$$

Haberman (1973) refers to r_{ij} as the adjusted standardised residual for the (i, j)th cell and is a random variable from the standard normal distribution. The significance of a cell to the association structure between the two categorical variables may be assessed by comparing these residuals with 1.96 (assuming that a 0.05 level of significance is considered). Table 4 gives these residuals and those values in bold are significant at the 0.05 level of significance.

We can see that each of the rows and columns provide a very strong contribution to the association structure between the two categorical variables of Table 1. Therefore there is evidence to suggest that each of the categories plays a vital role in describing the significance of the association between *Infant Response* and *Mothers Attachment Classification*. If we had considered the regions, or p-values, associated with Lebart et al.'s (1984) confidence circles we would have neglected the significance of the (3, 3)'th, or (*Resistant, Preoccupied*)'th, cell in Table 1. Such important association features would have also been missed by considering the elliptical regions of Figure 2.

Table 4 about here

8.3. Non-Symmetric Correspondence Analysis

One may consider that the association between *Mother's Attachment Classification* and *Infant Response* to be more naturally asymmetrically structured such that the column variable (*Mother's Attachment Classification*) is treated as a predictor variable and the row variable (*Infant Response*) is treated as a response variable. Indeed Beh and D'Ambra (2009) considered such an association structure and so performed a non-symmetric correspondence analysis. In doing so, the Goodman-Kruskal tau index is $\tau = 0.1990$ and has a numerator of $\tau_{\text{num}} = 0.1259$. Therefore, with a C – statistic of 326.51, there is a highly statistically significant (p-value < 0.0001) asymmetric association between the column (predictor) and row (response) variables. Thus, the attachment classifications given for the mothers do impact upon how the infant responds. Investigating more precisely the nature of the asymmetric association can be undertaken by considering non-symmetric correspondence analysis. A two-dimensional visual summary of the asymmetric association can be found by considering Figure 4. Superimposed on this figure are the 95% confidence circles of Beh and D'Ambra (2009) for the predictor categories.

Figure 4 about here

By performing a non-symmetric correspondence analysis on Table 1, the first principal inertia is $\lambda_1^2 = 0.0874$ and the second principal inertia is $\lambda_2^2 = 0.0355$. Together these two values account for 97.63% of the asymmetric association that exists between the two variables. This Figure 4 shows that, if we confine our attention to just the first two dimensions, and ignore the differently weighted axes, then all of the four levels of *Mother's Attachment Classification* play a significant role in determining their *Infants Response*. This suggests that the p-values associated with the four column categories are less than 0.05. In fact, the second column of Table 6 shows that the p-value for *Preoccupied* is 0.002 and for the remaining categories it is less than 0.0001. Similarly, Figure 4 also shows that all of the infant responses (with the exception of *Resistant*) are impacted by the mother's attachment levels. Thus, *Resistant* has a p-value greater than 0.05 (the second column of Table 5 shows that it is 0.760) and the remaining infant responses are all less than 0.001.

Figure 5 about here

Such confidence circles for non-symmetric correspondence analysis do not reflect the information contained in dimensions higher than the second, nor do they reflect the different principal inertia values attached to the axes. For this reason, confidence ellipses may be considered. Figure 5 shows these regions when considering the information contained in the optimal non-symmetric correspondence plot, while Figure 6 shows those ellipses which reflect the association in the first two dimensions only. In both cases all mother attachment levels play a significant role in the asymmetric structure of the two categorical variables. The differences in these two figures are consistent when comparing the configuration of points

using classical correspondence analysis (see Figure 2 and Figure 3). The key difference between Figure 5 and Figure 6 is the size of the confidence ellipse for *Preoccupied*. If we consider only the information contained in the first two dimensions, Figure 6 suggests that the p-value for this row category is larger than the p-value when considering its confidence circle (the third column of Table 6 shows that this p-value is 0.027). If we take into account the information contained in the optimal plot, the size of the ellipse in Figure 5 is a lot smaller, indicating that this particular category has at least one non-zero coordinate in one of the higher dimensions. Therefore its p-value reduces to less than 0.001 (see the fourth column of Table 6).

Figure 6 about here

Table 5 about here

Table 6 about here

---

9. Discussion

This paper has examined procedures for quantifying and interpreting the p-value of points in a classical (symmetric) and non-symmetric correspondence plot. We have shown by way of example that, for classical correspondence analysis, while the confidence circles of Lebart et al. (1984) and the p-values that may be derived from them are the most simplest of those considered, they do not take into consideration the inequality of the principal inertia values. Nor do such regions reflect the information contained in higher dimensions. On the other

hand, the p-values derived from considering the elliptical regions of Beh (2010) allow the user to take into consideration these two important aspects of the correspondence plot. We have also extended these features to be applicable when performing non-symmetric correspondence analysis and allow for a clearer understanding of the asymmetric structure of the categorical variables.

Our application has been focused on a simple two-way contingency table, but the same technique may be applied to consider multiple correspondence analysis using the indicator matrix, Burt matrix, or any other equivalent approach. Similarly, this paper has focused on the classical correspondence plots, but other plotting mechanisms could have been considered (such as biplots). However, we shall leave these issues for future investigation.

References

Ainsworth, M. D., Blehar, M. C., Waters, E., Wall, S. (1978). *Patterns of Attachment: A Psychological Study of the Strange Situation*. Erlbaum: Hillsdale, NJ.

Beh, E. J. (1997). Simple correspondence analysis of ordinal cross-classifications using orthogonal polynomials. *Biometrical Journal* 39: 589 – 613.

Beh, E. J. (2001). Confidence circles for correspondence analysis using orthogonal polynomials. *Journal of Applied Mathematics and Decision Sciences* 5: 35 – 45.

Beh, E. J. (2004). Simple correspondence analysis: A bibliographic review. *International Statistical Review* 72: 257 – 284.

Beh, E. J. (2010). Elliptical confidence regions for simple correspondence analysis. *Journal of Statistical Planning and Inference* 140: 2582 – 2588.

Beh, E. J., D'Ambra, L. (2009). Some interpretative tools for non-symmetrical correspondence analysis. *Journal of Classification* 26: 55 – 76.

Beh, E. J., Lombardo, R., Simonetti, B. (2011). A European perception of food using two methods of correspondence analysis. *Food Quality and Preference* 22: 226 – 231.

Clausen, S.-E. (1988). *Applied Correspondence Analysis: An Introduction.* Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-121. Thousand Oaks, CA: Sage.

Crisci, A., D'Ambra, L. (2011). The confidence ellipses in multiple non-symmetrical correspondence analysis for the evaluation of the innovative performance of the manufacturing enterprises in Campania. *Statistica & Applicazioni* 9: 175 – 187.

D'Ambra, L., Lauro, N. C. (1989). Non-symmetrical correspondence analysis for three-way contingency table. In: Coppi, R., Bolasco, S. eds. *Multiway Data Analysis.* Elsevier: Amsterdam, pp 301 - 315.

Goodman, L. A., Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association* 49: 732–764.

Gower, J., Lubbe, S., le Roux, N. (2011). *Understanding Biplots*. Wiley: Singapore.

George, C., Kaplan, N., Main, M. (1985). *Adult attachment interview*. University of California, Berkeley, unpublished manuscript.

Greenacre, M. (1984). *Theory and Application of Correspondence Analysis*. Academic Press: London.

Greenacre, M. (2007). *Correspondence Analysis in Practice*, 2nd ed. Chapman & Hall/CRC: London.

Greenacre, M., Blasius, J. (eds.) (1994). *Correspondence Analysis in the Social Sciences*. Academic Press: London.

Haberman, S. J. (1973). The analysis of residuals in cross-classified tables. *Biometrics* 29: 205 – 220.

Kroonenberg, P. M., Lombardo, R. (1998). Nonsymmetric correspondence analysis: A tutorial. *Kwantitatieve Methoden* 58: 57-83.

Kroonenberg, P. M., Lombardo, R. (1999). Non-symmetric correspondence analysis: A tool for analysing contingency tables with a dependence structure. *Multivariate Behavioral Research* 34: 367 – 396.

Lebart, L., Morineau, A., Warwick, K. M. (1984). *Multivariate Descriptive Statistical Analysis*. Wiley: New York.

Light, R. J., Margolin, B. H. (1971). An analysis of variance for categorical data. *Journal of the American Statistical Association* 66: 534–544.

Lombardo, R., Beh, E. J., D'Ambra, L. (2007). Non-symmetrical correspondence analysis with ordinal variables using orthogonal polynomials. *Computational Statistics & Data Analysis* 52: 566 – 577.

Ringrose, T. J. (1992). Bootstrapping and correspondence analysis in archaeology. *Journal of Archaeological Science* 19: 615–629.

Ringrose, T. J. (1996). Alternative confidence regions for canonical variate analysis. *Biometrika* 83: 575–587.

Ringrose, T. J. (2011). Bootstrap confidence regions for correspondence analysis. *Journal of Statistical Computation and Simulation*, (to appear), doi.org/10.1080/00949655.2011.579968.

van IJzendoorn, M. H. (1995). Adult attachment representations, parental responsiveness, and infant attachment. A meta-analysis on the predictive validity of the adult attachment interview. *Psychological Bulletin* 117: 387–403.

Table 1

Cross-classification of the attachment classification of a mother and her infant

Infant Response	*Mother's Attachment Classification*				Total
	Dismissing	*Autonomous*	*Preoccupied*	*Unresolved*	
AVOIDANT	62	29	14	11	116
SECURE	24	210	14	39	287
RESISTANT	3	9	10	6	28
DISORGANISED	19	26	10	62	117
Total	108	274	48	118	548

Table 2

Row category p-values from a classical correspondence analysis of Table 1 based on (i) Lebart et al.'s (1984) circular regions, (ii) 2-D elliptical regions and (iii) optimal elliptical regions

Row Category	*Confidence Circle*[(i)]	*Confidence Ellipse*[(ii)] *(D = 2)*	*Confidence Ellipse*[(iii)] *(D = M = 3)*
Avoidant	0	0	0
Secure	0	0	0
Resistant	0.287	0.802	0
Disorganised	0	0	0

Table 3

Column category p-values from a classical correspondence analysis of Table 1 based on (i) Lebart et al's (1984) circular regions, (ii) 2-D elliptical regions and (iii) optimal elliptical regions

Column Category	*Confidence Circle*[(i)]	*Confidence Ellipse*[(ii)] *(D = 2)*	*Confidence Ellipse*[(iii)] *(D = M = 3)*
Dismissing	0	0	0
Autonomous	0	0	0
Preoccupied	0.0203	0.108	0
Unresolved	0	0	0

Table 4

Adjusted standardised residuals of Table 1. Bolded values are significant at the 0.05 level of significance. The values in parentheses are the residuals' p-values

Infant Response	*Mother's Attachment Classification*			
	Dismissing	*Autonomous*	*Preoccupied*	*Unresolved*
Avoidant	**8.19** **(<0.001)**	**-3.81** **(<0.001)**	1.20 (0.114)	**-2.80** **(0.003)**
Secure	**-4.33** **(<0.001)**	**5.55** **(<0.001)**	**-2.22** **(0.013)**	**-2.90** **(0.002)**
Resistant	-1.07 (0.142)	-1.34 (0.091)	**4.82** **(<0.001)**	-0.01 (0.495)
Disorganised	-0.85 (0.199)	**-4.25** **(<0.001)**	-0.08 (0.469)	**7.33** (<0.001)

Table 5

Row category p-values from a non-symmetric correspondence analysis of Table 1 based on (i) Beh and D'Ambra's (2009) circular regions, (ii) 2-D elliptical regions and (iii) optimal elliptical regions

Row Category	*Confidence Circle*[(i)]	*Confidence Ellipse*[(ii)] *(D = 2)*	*Confidence Ellipse*[(iii)] *(D = M = 3)*
Avoidant	0	0	0
Secure	0	0	0
Resistant	0. 760	0.999	0.925
Disorganised	0	0	0

Table 6

Column category p-values from a non-symmetric correspondence analysis of Table 1 based on (i) Beh and D'Ambra's (2009) circular regions, (ii) 2-D elliptical regions and (iii) optimal elliptical regions

Column Category	*Confidence Circle*[(i)]	*Confidence Ellipse*[(ii)] *(D = 2)*	*Confidence Ellipse*[(iii)] *(D = M = 3)*
Dismissing	0	0	0
Autonomous	0	0	0
Preoccupied	0.002	0.027	0
Unresolved	0	0	0

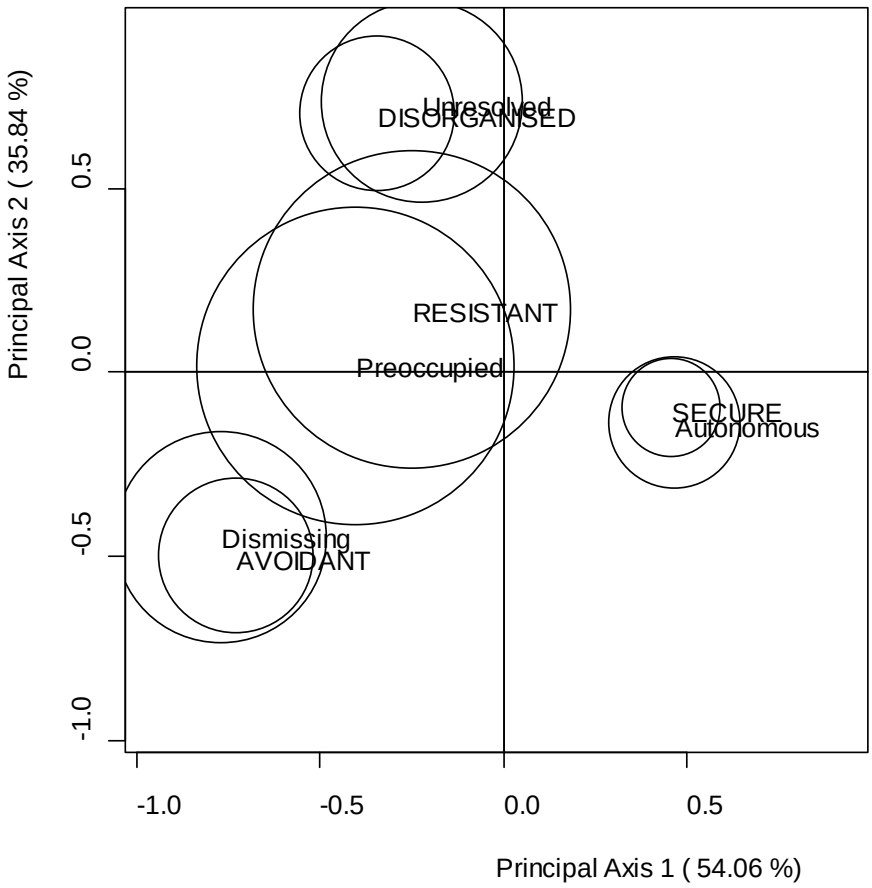

Figure 1. 95% Confidence circles from a classical correspondence analysis with a radii length of (4). These regions take into account the position of the row and column points in the first and second dimensions only.

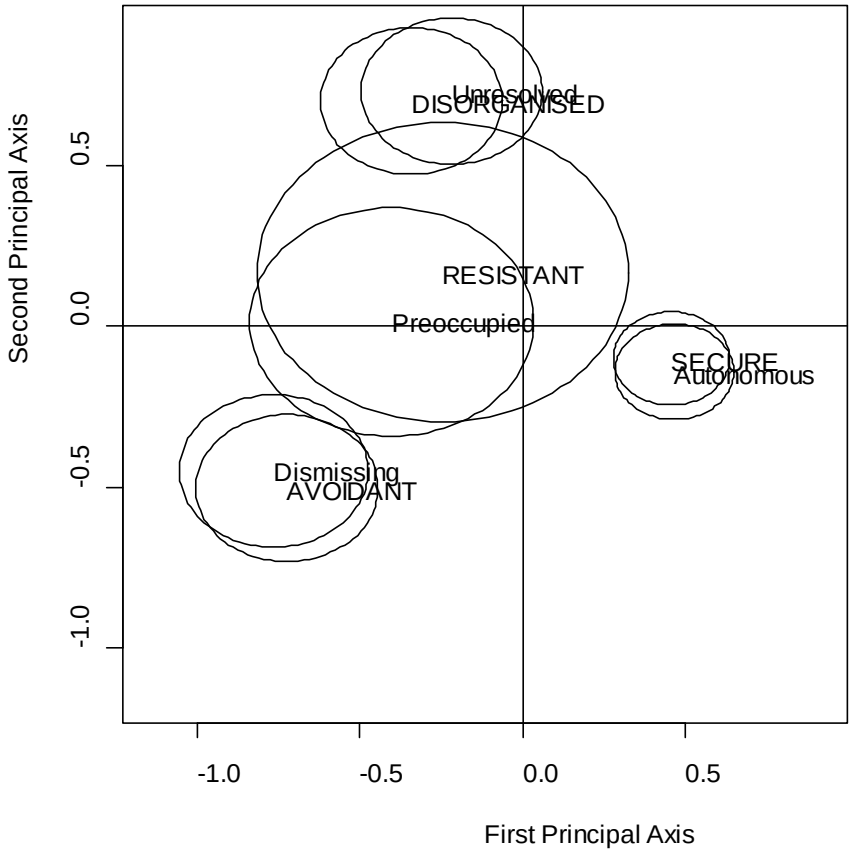

Figure 2. 95% Confidence ellipses from a classical correspondence analysis with a semi-axis length of (6). These regions take into account the row and column points in the first and second dimensions only.

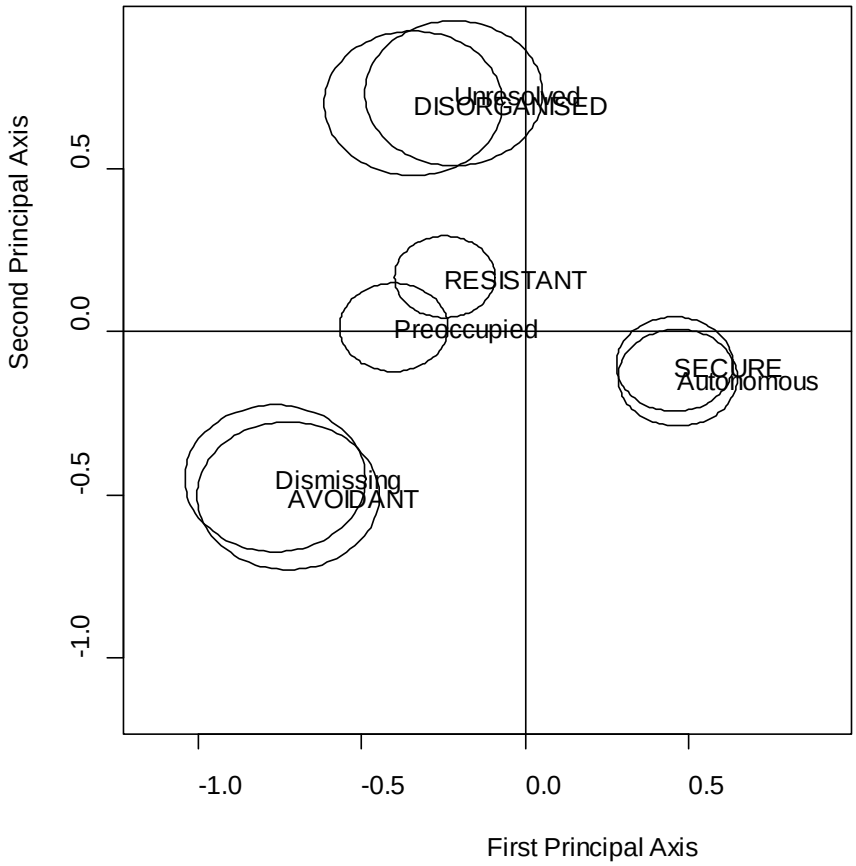

Figure 3. 95% Confidence ellipses from a classical correspondence analysis with a semi-axis length of (5). These regions take into account the position of the row and column points in the third, and higher, dimensions.

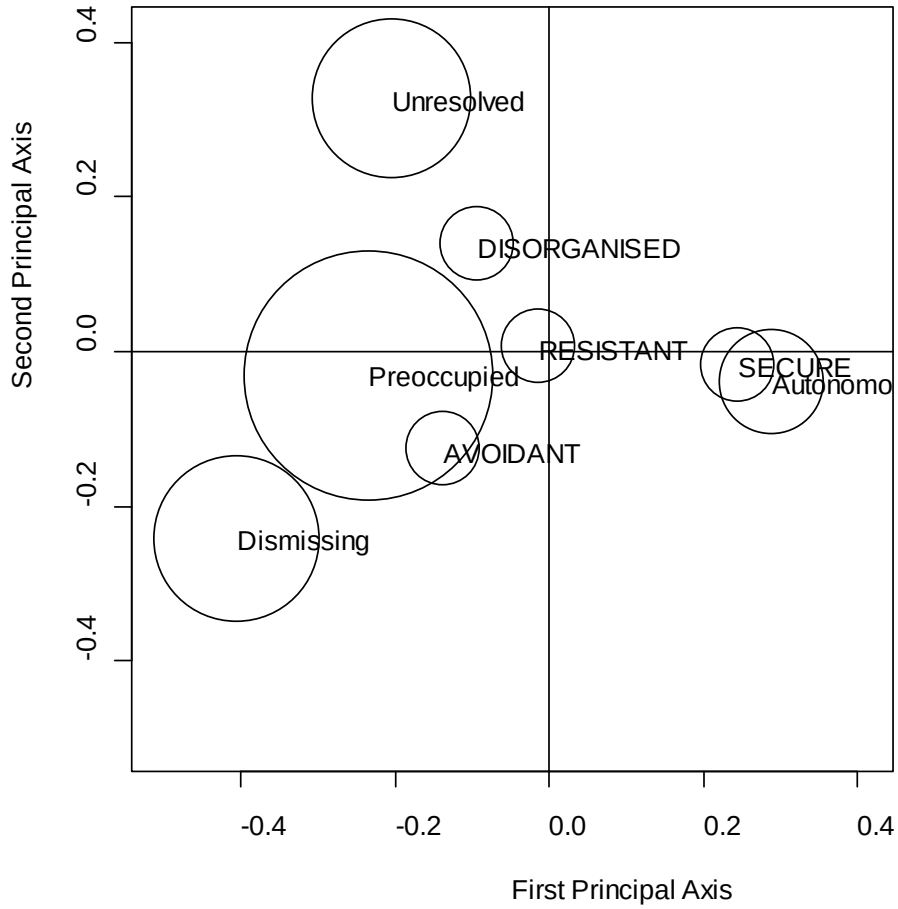

Figure 4. 95% Confidence circles for a non-symmetric correspondence analysis of Table 1. These regions take into account the position of the row and column points in the first and second dimensions only.

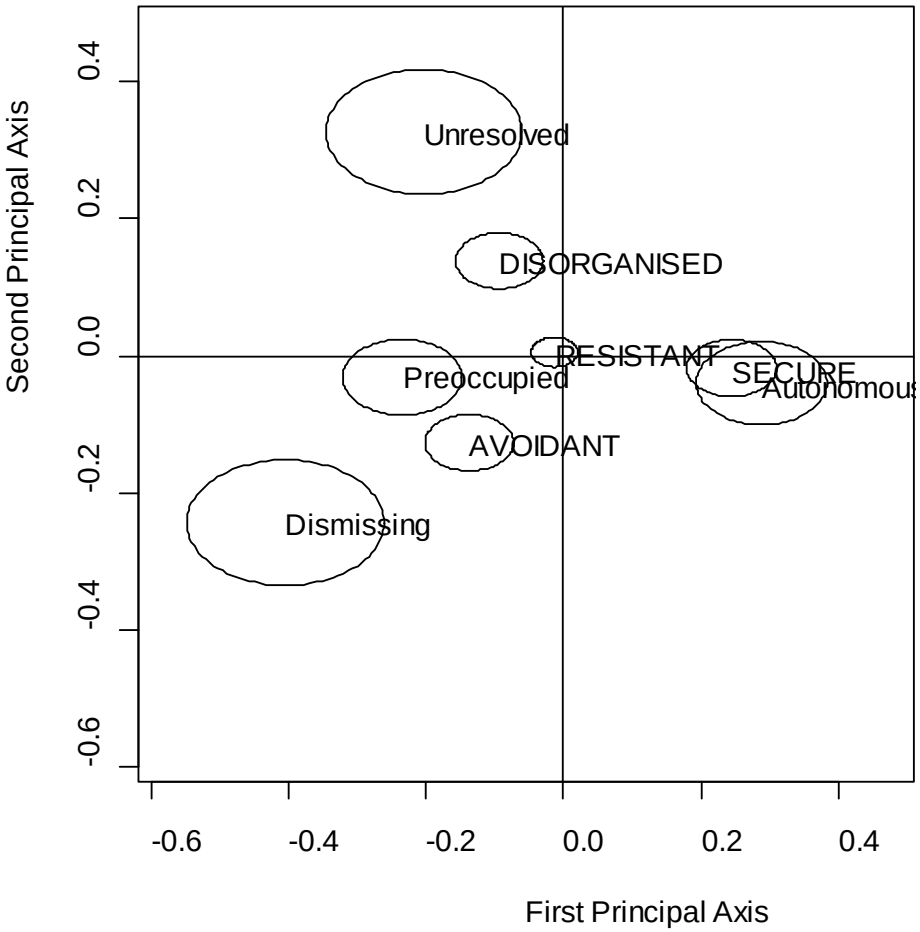

Figure 5. 95% Confidence ellipses for a non-symmetric correspondence analysis of Table 1. These regions take into account the position of the row and column points in the third, and higher, dimensions.

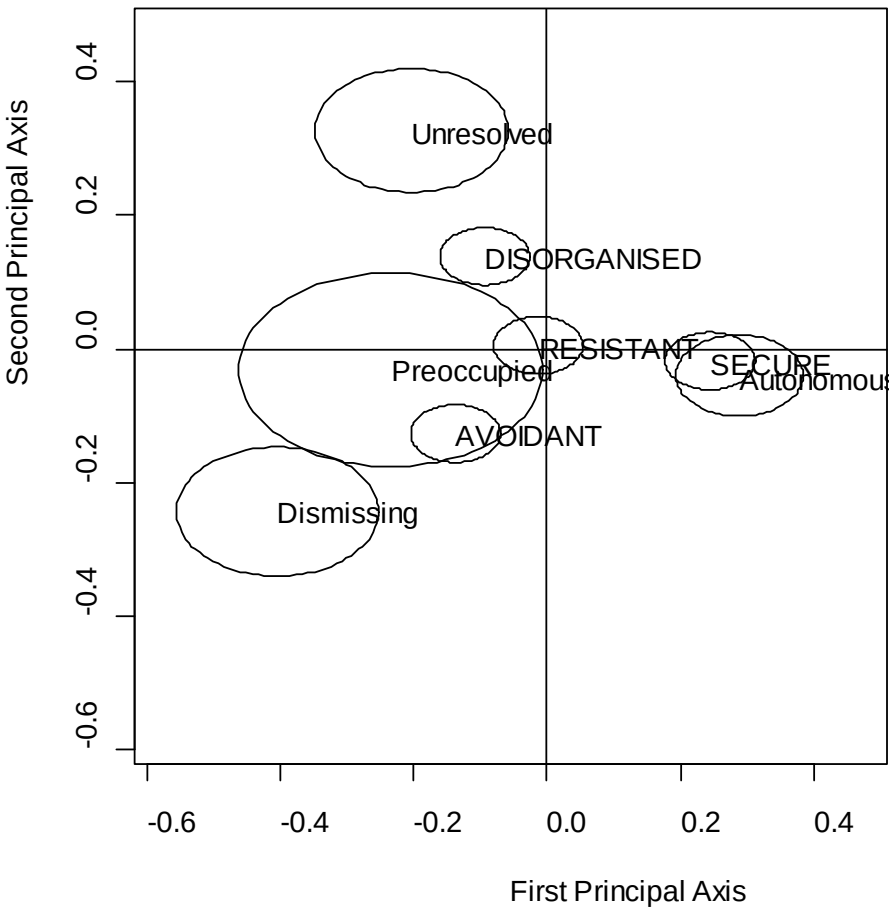

Figure 6. 95% Confidence ellipses for a non-symmetric correspondence analysis of Table 1. These regions take into account the position of the row and column points in the first and second dimensions only.