NOVA

University of Newcastle Research Online

nova.newcastle.edu.au

Beh, Eric J. "Elliptical confidence regions for simple correspondence analysis" Journal of Statistical Planning and Inference Vol. 140, Issue 9, p. 2582-2588 (2010)

Available from: <http://dx.doi.org/10.1016/j.jspi.2010.03.018>

Accessed from: <http://hdl.handle.net/1959.13/921672>

Elliptical confidence regions for simple correspondence analysis

Eric J. Beh

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

Abstract

Confidence regions for simple correspondence analysis allow for the identification of categories that are consistent with independence, and those that are not. This paper describes a procedure for constructing elliptical regions which takes into account the unequal weighting of each of the axes of the plot.

Key words: Confidence circle; Confidence ellipse; Correspondence Plot; Eccentricity.

1. Introduction

Simple correspondence analysis is a popular graphical technique that is used to identify how the categories of a two-way contingency table are associated with one another. The most utilised output from performing such an analysis is its ability to graphically show how categories are associated with one another by constructing a correspondence plot. For such a plot the origin indicates where the points will lie if there is complete independence between the two categorical variables. The distance of a point from the origin therefore provides an indication of the strength to which that particular category contributes to the null hypothesis of complete independence between the two variables.

However, if it is found that a statistically significant association exists between the two variables, it is also important to understand how each row/column category contributes to this dependent structure. In the past confidence circles, or circular regions, have been used as a means of identifying whether a category is statistically consistent with what is expected under the hypothesis of independence. Such regions are important since, given a level of significance, α , if the origin falls within the region then that point is considered to be a poor contributor to the dependence of the variables. If the origin lies outside of the circular region, then the category for which it was constructed helps to contribute to the association between the two regions. For simple correspondence analysis Lebart, Morineau and Warwick (1984) and Beh (2001) considered circular regions for nominal and ordinal categorical variables respectively.

The problem with circular regions is that they are constructed by assuming that the axes of the correspondence plot are equally weighted. However this, in general, is not the case. Based on the principal inertia's obtained from performing a correspondence analysis on a contingency table, it is are such that the first axis represents more of the association than the second axis, and the second axis represents more of the association than the third axis and so on. Therefore it is more appropriate to construct confidence regions to reflect the unequal weighting assigned to the axes – this can be achieved by considering an elliptical region rather than a one that is

circular. Non-circular regions for correspondence analysis have been a topic of discussion by Ringrose (1992, 1996) who explored the use of bootstrapping for the construction of convex hulls defining a confidence region. Greenacre (2007, p. 196 – 197) also describes the implementation of bootstrapping to construct confidence regions but proposed a peeling step to remove the impact of 5% of the more extreme outlying replicates. Greenacre (2007, p. 197) also considers using the delta method for the construction of elliptical regions but notes some problems with its use.

This paper will algebraically show a relatively simple procedure for constructing confidence regions of an elliptical shape for a two-dimensional simple correspondence plot. Unlike those techniques mentioned above, it will be shown that such regions are of benefit since they also incorporate the information contained in the third, and higher dimensions, of such a plot and therefore provide a comprehensive description of the association between two categorical variables.

2 Circular Confidence Regions

2.1 Correspondence Analysis & Confidence Circles

Consider an $r \times c$ two-way contingency table, N , where the (i, j) th cell entry is given by n_{ij} for $i = 1, 2, \dots, r$ and $j = 1, 2, \dots, c$. 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 $p_{i\cdot} = \sum_{j=1}^c p_{ij}$ and the j 'th column relative marginal frequency by

$$p_{\cdot j} = \sum_{i=1}^r p_{ij}.$$

To determine the nature of the association between the variables one may perform correspondence analysis to obtain a visual summary of how each category compares with other categories. Such comparisons are made keeping in mind the relative distribution of the classifications between the categories and focus on visually comparing the I row profiles $\{p_{ij} / p_{i\cdot} : j = 1, 2, \dots, c\}$ and the J column profiles $\{p_{ij} / p_{\cdot j} : i = 1, 2, \dots, r\}$ in a joint low-dimensional space. There are many ways in which correspondence analysis can be performed and Nishisato (2007, chap 2) provides a good account of the various ways that it can be done. Consider decomposing the matrix of standardised residuals using singular value decomposition (Beh, 2004) such that

$$r_{ij} = \frac{p_{ij} - p_{i\cdot} p_{\cdot j}}{\sqrt{p_{i\cdot} p_{\cdot j}}} = 1 + \sum_{m=1}^{M^*} a_{im} \lambda_m b_{jm}$$

where $M^* = \min(r, c) - 1$. 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 have the property

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

The value λ_m is the m 'th singular value and are arranged in descending order such that

$$1 > \lambda_1 \geq \lambda_2 \geq \dots \geq 0.$$

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 coordinates

$$f_{im} = a_{im} \lambda_m \quad \text{and} \quad g_{jm} = b_{jm} \lambda_m \quad (2)$$

respectively. Thus the Pearson chi-squared statistic of N can be expressed in terms of the coordinates by

$$X^2 = n \sum_{m=1}^{M^*} \lambda_m^2 = n \sum_{m=1}^{M^*} \sum_{i=1}^r p_{i\bullet} f_{im}^2 = n \sum_{m=1}^{M^*} \sum_{j=1}^c p_{\bullet j} g_{jm}^2.$$

For the two-way contingency table, N , Lebart, Morineau and Warwick (1984) showed that the radii length of the $100(1-\alpha)\%$ confidence circle in a two-dimensional correspondence plot for the i 'th row category using correspondence analysis is

$$r_{i(\alpha)} = \sqrt{\frac{\chi_{\alpha}^2}{n_{i\bullet}}}. \quad (3)$$

Here χ_{α}^2 is the $1-\alpha$ percentile of a chi-squared distribution with 2 degree of freedom. Confidence circles constructed in this manner are similar to those considered by Mardia, Kent and Bibby (1982, p. 345 – 348) derived for canonical analysis. In situations where a categorical variable consists of ordinal responses, Beh (2001) showed that the radii length of these circles using the ordinal correspondence analysis approach of Beh (1997) are identical to those of Lebart *et al.* (1984).

2.2 Example

Consider the two-way contingency table of Table 1 that cross-classifies a mothers attachment to her child, and the child's response to their mother's level of attachment. The data is 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, Kaplan and Main, 1985) while the four row categories are observed from the Ainsworth Strange Situation (Ainsworth, Blehar, Waters and Wall, 1978).

A chi-squared test of independence concludes that, with a Pearson chi-squared statistic of 252.3982 and 9 degrees of freedom (p -value = 0), there is ample evidence to conclude that the two categorical variables are associated.

Table 1

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

<i>Infant Response</i>	<i>Mother's Attachment Classification</i>			
	<i>Dismissing</i>	<i>Autonomous</i>	<i>Preoccupied</i>	<i>Unresolved</i>
<i>Avoidant</i>	62	29	14	11
<i>Secure</i>	24	210	14	39
<i>Resistant</i>	3	9	10	6
<i>Disorganised</i>	19	26	10	62

To investigate further the nature of the association, a simple correspondence analysis can be performed where the two-dimensional correspondence plot is given by Fig. 1. With a total inertia of $252.3982/548 = 0.4606$, the two-dimensional plot of Fig. 1 graphically accounts for 89.89% of the association between the variables. The first axis, with a principal inertia of $\lambda_1^2 = 0.24897$ accounts for 54.06% of the inertia, while the second axis accounts for 35.84% (principal inertia is $\lambda_2^2 = 0.16505$).

Fig. 1. Correspondence plot of Table 1 with 95% confidence circles

Superimposed on Fig. 1 are the 95% confidence circles for each row and column category using the radii length of Lebart *et al.* (1984), (3). It shows that infant response *Resistant* is consistent with what is expected if the two variables are independent since its circle overlaps the origin. Conversely the confidence circles of

the remaining categories indicate that those infant response and mother attachment classifications contribute to the dependent nature of the two variables. However, the confidence circle of *Preoccupied* is close to overlapping the origin.

Such conclusions are not as apparent when viewing the three-dimensional correspondence plot of Table 1. Fig. 2 provides an optimal view of the association between the two sets of categories and so accounts for 100% of the total inertia of Table 1. Contrary to what Fig. 1 shows, *Preoccupied* and *Resistant* lie at some distance away from the origin when considering the third dimension. While this dimension only contributes to 10.11% of the association that exists between the variables, much of this is accountable by the “correspondence” between these two categories. This suggests that *Preoccupied* and *Resistant* perhaps are not as consistent with what is expected when the two categorical variables are independent as what Fig. 1 indicates. Note that the coordinate for the third dimension of the remaining categories is very close to zero.

Fig. 2. Optimal (3-dimensional) Correspondence plot of Table 1

Therefore, since the two principal inertia values associated with Fig. 2 are not equivalent, circular regions may not be considered an appropriate means of making such conclusions. Instead one may reflect the unequal weighting of the axes by generating elliptical regions. The following section describes how such elliptical regions may be generated.

3 The Elliptical Region

This section derives $100(1 - \alpha)\%$ confidence ellipses for the row and column categories of a two-way contingency table when analysing the association between the variables using simple correspondence analysis.

3.1 Constructing Confidence Ellipses

Suppose we consider the property of the elements of the i 'th row singular vector given by the left hand side of (1). For $M^* = \min(r, c) - 1$

$$\sum_{m=1}^{M^*} a_{im}^2 = \frac{1}{p_{i\bullet}}$$

so that

$$a_{i1}^2 + a_{i2}^2 = \left(\frac{1}{p_{i\bullet}} - \sum_{m=3}^{M^*} a_{im}^2 \right). \quad (4)$$

Therefore, by considering the definition of the profile coordinate for the i 'th row category, (2), (4) can be alternatively expressed as

$$\frac{f_{i1}^2}{\lambda_1^2} + \frac{f_{i2}^2}{\lambda_2^2} = \left(\frac{1}{p_{i\bullet}} - \sum_{m=3}^{M^*} a_{im}^2 \right)$$

and leads to the result

$$\frac{f_{i1}^2}{\lambda_1^2 \left(\frac{1}{p_{i\bullet}} - \sum_{m=3}^{M^*} a_{im}^2 \right)} + \frac{f_{i2}^2}{\lambda_2^2 \left(\frac{1}{p_{i\bullet}} - \sum_{m=3}^{M^*} a_{im}^2 \right)} = 1.$$

This is the equation for an ellipse centred at the point (f_{i1}, f_{i2}) in a two-dimensional correspondence plot with a semi-major and semi-minor axis length of

$$x_i = \lambda_1 \sqrt{\left(\frac{1}{p_{i\bullet}} - \sum_{m=3}^{M^*} a_{im}^2 \right)} \quad \text{and} \quad y_i = \lambda_2 \sqrt{\left(\frac{1}{p_{i\bullet}} - \sum_{m=3}^{M^*} a_{im}^2 \right)} \quad (5)$$

respectively. Semi-major and minor lengths for the j 'th column coordinate in a two-dimensional space can be obtained in a similar manner.

Note that, unless $a_{im} = 0$ for $m = 3, 4, \dots, M^*$, $x_i > y_i$ since $\lambda_1 > \lambda_2$. Therefore the elliptical nature of the confidence region will be such that the ellipse will be stretched more along the first principal axis than along the second principal axis. However, (5) can not be used, in their current state, for identifying the level of confidence with which a particular row (or column) category is consistent with what is expected under complete independence. To address this issue, we may consider the procedure outlined in the following section.

3.2 100(1- α)% Ellipse

To construct a 100(1- α)% ellipse, recall that λ_m^2 accounts for a certain proportion of the total inertia, X^2/n . If the 1- α percentile of the chi-squared statistic with $(r-1)(c-1)$ degrees of freedom is denoted by χ_α^2 and $n\tilde{\lambda}_{m(\alpha)}^2$ represents a percentage of χ_α^2 , such that

$$\frac{\lambda_m^2}{X^2/n} = \frac{\tilde{\lambda}_{m(\alpha)}^2}{\chi_\alpha^2/n}$$

then

$$\tilde{\lambda}_{m(\alpha)} = \lambda_m \sqrt{\frac{\chi_\alpha^2}{X^2}}$$

Therefore, the 100(1- α)% confidence ellipse for the i 'th row category has a semi-major length and a semi-minor length of

$$x_{i(\alpha)} = \lambda_1 \sqrt{\frac{\chi_\alpha^2}{X^2} \left(\frac{1}{p_{i\bullet}} - \sum_{m=3}^{M^*} a_{im}^2 \right)} \quad \text{and} \quad y_{i(\alpha)} = \lambda_2 \sqrt{\frac{\chi_\alpha^2}{X^2} \left(\frac{1}{p_{i\bullet}} - \sum_{m=3}^{M^*} a_{im}^2 \right)} \quad (6)$$

respectively. Unless $M^* = 2$, the radii length of a confidence ellipse using (6) takes into consideration the information contained in dimensions higher than the first and second principal axes. Such information is reflected by the presence of a_{im} and b_{jm} in (6). In fact, the semi-major and semi-minor axis lengths of (6) may be expressed in terms of a coordinate of a profile in higher dimensions by

$$x_{i(\alpha)} = \lambda_1 \sqrt{\frac{\chi_\alpha^2}{X^2} \left(\frac{1}{p_{i\bullet}} - \sum_{m=3}^{M^*} \left(\frac{f_{im}}{\lambda_m} \right)^2 \right)} \quad \text{and} \quad y_{i(\alpha)} = \lambda_2 \sqrt{\frac{\chi_\alpha^2}{X^2} \left(\frac{1}{p_{i\bullet}} - \sum_{m=3}^{M^*} \left(\frac{g_{jm}}{\lambda_m} \right)^2 \right)}.$$

Thus, unlike the circular regions of Lebart *et al.* (1984), conclusions obtained by constructing elliptical confidence regions using (6) take into account the information that exists in the third and higher dimensions of a simple correspondence plot.

When the information in the third, and subsequent dimensions, is ignored, the semi-axis lengths of (6) may be alternatively expressed as

$$x_{i(\alpha)} = \lambda_1 \sqrt{\frac{\chi_\alpha^2}{X^2 p_{i\bullet}}} \quad \text{and} \quad y_{i(\alpha)} = \lambda_2 \sqrt{\frac{\chi_\alpha^2}{X^2 p_{i\bullet}}}. \quad (7)$$

Therefore, the sum of squares of these semi-axis lengths is

$$x_{i(\alpha)}^2 + y_{i(\alpha)}^2 = \phi_2 \frac{X_\alpha^2}{n_{i\bullet}}$$

where ϕ_2 is the proportion of the total inertia accounted for by a two-dimensional correspondence plot constructed using the first two principal axes. When $M^* = 2$, so that the optimal correspondence plot consists of only two-dimensions, the sum of squares of these half-axis lengths is such that $x_{i(\alpha)}^2 + y_{i(\alpha)}^2 = r_{i(\alpha)}^2$.

By considering the semi-axis lengths described by (6) as the level of confidence increases (so that α decreases), the elliptical region for each of the r rows and c columns expands. For example, the area of the region for the i 'th row category is

$$A_{i\alpha}^E = \pi x_i y_i = \chi_\alpha^2 \frac{\pi \lambda_1 \lambda_2}{X^2} \left(\frac{1}{p_{i\bullet}} - \sum_{m=3}^{M^*} a_{im}^2 \right)$$

and is determined by the magnitude of the singular values, and the size of the contingency table through M . Thus, for a contingency table with 4 degrees of freedom, the area of the 95% and 99% confidence ellipse for a particular category will increase by a factor of $13.27670/9.487729 = 1.4$.

3.3 Eccentricity of Elliptical Regions

Another important characteristic of any ellipse is to measure its eccentricity which is bounded by the interval $[0, 1]$ – note that a circle has zero eccentricity. Such a measure can prove beneficial when performing correspondence analysis since the eccentricity of an ellipse reflects the relative magnitude of the first principal inertia compared with the second principal inertia. By considering the definition of eccentricity of Weir, Hass and Giordano (2005, p. 698), for example, the eccentricity of the $100(1-\alpha)\%$ elliptical region for the i 'th row category can be shown to be equivalent to

$$E_{i(\alpha)} = \sqrt{1 - \frac{y_{i(\alpha)}^2}{x_{i(\alpha)}^2}}.$$

By considering the semi-axis lengths of (6), the eccentricity of a $100(1-\alpha)\%$ confidence ellipse in a two-dimensional correspondence plot is

$$E_{i(\alpha)} = \sqrt{1 - \frac{\lambda_2^2}{\lambda_1^2}} \quad (8)$$

and is therefore fixed for all values of α and for each and every row category. It can be similarly shown that the eccentricity of a $100(1-\alpha)\%$ confidence ellipse for each column category, denoted by $E_{j(\alpha)}$, is equivalent to the right hand side of (8). Therefore the relative weight of each axis of a correspondence plot is reflected in the construction of a confidence region by taking into account the relative size of the first two singular values. If $\lambda_1 = \lambda_2$ so that the principal inertia of the first two axes are equivalent, then the confidence region will be circular rather than elliptical. In this case the region of each of the rc points in the correspondence plot will have an eccentricity of zero. Therefore the radii length of such a circular region is

$$x_{i(\alpha)} = y_{i(\alpha)} = r_{i(\alpha)} \sqrt{\frac{\lambda_1^2}{X^2/n} - p_i \cdot \frac{\lambda_1^2}{X^2} \sum_{m=3}^M a_{im}^2}.$$

This result shows that the radii length of a confidence circle constructed using the approach of Lebart *et al.* (1994) and those derived from the elliptical semi-axis lengths are proportional to one another.

3.5 Example Revisited

Suppose we consider again Table 1. Recall that Fig. 1 is two-dimensional correspondence plot where the 95% confidence circles provide a way of identifying those categories that are consistent with the hypothesis of independence and those that are not. Since the principal axes are not equally weighted ($\lambda_1^2 = 0.24897$ and $\lambda_2^2 = 0.16505$) elliptical confidence regions can be constructed.

Fig. 3. Correspondence plot of Table 1 with 95% confidence ellipses taking into consideration the information in the third dimension; equation (6)

Fig. 3 is the correspondence plot with the 95% confidence ellipses included as defined by (6). The elliptical regions were constructed using R code based on that of MacDonald (2002). It shows that, while the figure is a two-dimensional representation

of the association, the ellipses conclude that all of the categories are not consistent with what would be expected under complete independence. If we just considered the first two singular vectors, Fig. 4 concludes that, just like the confidence circles of Fig. 1, *Resistant* is consistent with what is expected under independence. The “stretching” of the ellipse along the first principal axis yields the same conclusion for *Preoccupied*. The confidence ellipse for the other responses is virtually unchanged when comparing Fig. 3 with Fig. 4 since, as shown by Fig. 2, these categories have a near zero coordinate along the third principal axis. Note that the eccentricity of each ellipse in Fig. 3 and Fig. 4 is

$$E_{0.05} = \sqrt{1 - \frac{\lambda_2^2}{\lambda_1^2}} = \sqrt{1 - \frac{0.16505}{0.24897}} = 0.58058$$

and indicates that the regions in Fig. 3 and Fig. 4 are far from circular in shape.

Fig. 4. Correspondence plot of Table 1 with 95% confidence ellipses ignoring the information in the third dimension; equation (7)

4 Discussion

This paper has described a simple means of constructing elliptical regions for a configuration of points on a correspondence plot by performing classical and non-

symmetrical correspondence analysis. As Section 3 has shown these two-dimensional regions are advantageous to consider since they reflect information contained in the third and higher dimensions. However, there are certainly avenues that can be explored in the near future. For example, incorporating this approach into the construction of uncertainty circles (Gabriel, 1995) has the potential to identify other key sources of association that may exist between intra-variable categories. The construction of such elliptical regions for multiple correspondence analysis can also be pursued.

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
- Gabriel, K. R., 1995. MANOVA biplots for two-way contingency tables, in Krzanowski, W. J. ed, *Recent Advances in Descriptive Multivariate Analysis*. Oxford University Press, 227 – 268.
- George, C., Kaplan, N., Main, M., 1985. *Adult Attachment Interview*. Unpublished manuscript, University of California, Berkeley.
- Greenacre, M., 2007. *Correspondence Analysis in Practice*. Chapman & Hall/CRC, London, 2nd ed.
- Lebart, L., Morineau, A., Warwick, K. M., 1984. *Multivariate Descriptive Statistical Analysis*. John Wiley & Sons, New York.
- Macdonald, P. D. M., 2002. Drawing an Ellipse in Splus or R. Last accessed from URL http://www.math.mcmaster.ca/peter/s4m03/s4m03_0102/ellipse.html on the 11th August, 2009.
- Mardia, K. V., Kent, J. T., Bibby, J. M., 1982. *Multivariate Analysis*. Academic Press, New York.
- Nishisato, S., 2007. *Multidimensional Nonlinear Descriptive Analysis*. Chapman & Hall/CRC, London.
- 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.
- 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.
- Weir, M. D., Hass, J., Giordano, F. R., 2005. *Thomas' Calculus*. Pearson, Boston, 11th ed.