JOURNAL OF APPLIED MATHEMATICS AND DECISION SCIENCES, 5(1), 35–45


# CONFIDENCE CIRCLES FOR CORRESPONDENCE ANALYSIS USING ORTHOGONAL POLYNOMIALS

ERIC J. BEH†
*School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, NSW, 2522, Australia*

**Abstract.** An alternative approach to classical correspondence analysis was developed in [3] and involves decomposing the matrix of Pearson contingencies of a contingency table using orthogonal polynomials rather than via singular value decomposition. It is especially useful in analysing contingency tables which are of an ordinal nature. This short paper demonstrates that the confidence circles of Lebart, Morineau and Warwick (1984) for the classical approach can be applied to ordinal correspondence analysis. The advantage of the circles in analysing a contingency table is that the researcher can graphically identify the row and column categories that contribute or not to the hypothesis of independence.

## 1. Introduction

The correspondence analysis technique of [3] was shown to be mathematically similar to the classical correspondence analysis approach discussed by several authors, including Lebart, Morineau and Warwick (1984), [8] and [9]. However there is a major difference between the approaches, and this is concerned with the method of decomposing the Pearson chi-squared statistic. The classical approach decomposes the statistic into singular values by partitioning the matrix of Pearson contingencies using singular value decomposition. The approach of [3] decomposes the Pearson chi-squared statistic into bivariate moments, such as linear–by–linear, linear–by–quadratic, etc, by partitioning the matrix of Pearson contingencies using the orthogonal polynomials defined in [4]. Therefore the interpretation of the correspondence plots is very different. The ordinal correspondence plots of [3] graphically show how categories within a variable are similar or not by their proximity from each other along the first (linear), second (dispersion) and higher axes. The interpretation of the correspondence plots from the classical correspondence analysis technique is unclear. Points significantly

† Requests for reprints should be sent to E. J. Beh, School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, NSW, 2522, Australia.

far from the origin indicate that they contribute to the dependency between the row and column variables, while points close to the origin indicate they do not make such a contribution. While the interpretation of the ordinal plots allows us to reach the same conclusions, the classical correspondence plot will not explain how two points far from each other are different; the classical approach will only make the conclusion that they are different.

With ordinal correspondence plots, we can determine which row and column categories, if any, contribute to the dependency between the two variables using confidence circles. Lebart et al. (1984) defined such circles for classical correspondence analysis. This paper shows that similar confidence circles can be calculated for each row and column profile co-ordinate by using ordinal correspondence analysis. The derivations presented here are for the row categories, while those for the column categories can be made in a similar manner.

Section 2 defines the notation to be used in this presentation as well as defining the radii length of the confidence circle for the $i'th$ row profile co-ordinate in a plot using classical correspondence analysis. Section 3 shows that for the correspondence analysis approach of [3] the radius of the confidence circles can be derived in exactly the same way as those from classical correspondence analysis. Section 4 shows the relationship between the marginal frequencies of a set of categories and the radii length of the confidence circles. Section 5 consists of two examples which show the application of the confidence circle using doubly ordered correspondence analysis.

## 2. Confidence Circles for Classical Correspondence Analysis

Consider an $I \times J$ two-way contingency table, N, where the $(i,j)'th$ cell entry is denoted as $n_{ij}$ for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$. Let the grand total of N be $n$ and the probability matrix be P so that the $(i,j)'th$ cell entry is $p_{ij} = n_{ij}/n$ for which $\sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij} = 1$. Define the $i'th$ row marginal proportion as $p_{i\bullet} = \sum_{j=1}^{J} p_{ij}$ and the $j'th$ column marginal probability as $p_{\bullet j} = \sum_{i=1}^{I} p_{ij}$ so that $\sum_{i=1}^{I} p_{i\bullet} = \sum_{j=1}^{J} p_{\bullet j} = 1$.

The confidence circle of Lebart et al. (1984) is a method of observing the importance of a profile's position in a correspondence plot. Generally, if the origin lies outside the confidence circle for a particular category, then that category contributes to the dependency between the row and

column categories of the contingency table. If the origin lies within the circle for a particular category, then that category does not contribute to the dependency between the variables.

Lebart et al. (1984) showed that for classical correspondence analysis the radii length of the confidence circle fot the $i'th$ row profile co-ordinate can be calculated by

$$r_i = \sqrt{\frac{\chi^2_{(J-1)}}{np_{i\bullet}}} \tag{1}$$

where $\chi^2_{(J-1)}$ is the theoretical chi-squared value with $J-1$ degrees of freedom at the $\alpha$ level of significance. Generally a correspondence plot consists of only two dimensions, but can include three or more. However, visually representing multiple dimensions is conceptually difficult; [2], [7] and [11]([11], [12]) presented some novel approaches to visualising multiple dimensions. If a correspondence plot consists of two dimensions, then with 2 degrees of freedom and at the 5% level of significance, $\chi^2_{(2)} = 5.99$. Therefore, the radius of the confidence circle for the $i'th$ row profile co-ordinate can be approximated by

$$r_i = \sqrt{\frac{5.99}{np_{i\bullet}}} \tag{2}$$

## 3. Confidence Circles for Ordinal Correspondence Analysis

The radii length of the confidence circle for the $i'th$ row profile using the correspondence analysis of [3] is mathematically identical to the radii length using classical correspondence analysis. [6] calculated confidence circles for their analysis using the same orthogonal polynomial definitions as we do here but the plotting system they considered is different.

Suppose that a doubly ordered correspondence analysis is applied to a two-way contingency table. Then denote the row profile co-ordinate of the $i'th$ row category along the $k'th$ axis as $f^*_{ik}$ for $k = 1, 2, \ldots, J-1$ which is defined by

$$f^*_{ik} = \sum_{j=1}^{J} \frac{p_{ij}}{p_{i\bullet}} b_k(j)$$

This row profile co-ordinate is the weighted sum of the column orthogonal polynomials or order k, $\{b_k(j)\}$, where the weights used are from the profile of the $i'th$ row category, $\{p_{ij}/p_{i\bullet}\}$; see [3] for a derivation of $f^*_{ik}$.

By using equations (3.1.10) and (3.1.11) of [3], the relationship between the chi-squared statistic and the row profile co-ordinates is

$$X^2 = n \sum_{k=1}^{J-1} \sum_{i=1}^{I} p_{i\bullet} \left(f_{ik}^*\right)^2 \tag{3}$$

For the $i'th$ row profile co-ordinate, the contribution to $X^2$ is $X_i^2$ where

$$X_i^2 = n \sum_{k=1}^{J-1} p_{i\bullet} \left(f_{ik}^*\right)^2 \tag{4}$$

for all $i = 1, 2, \ldots, I$ and where $X_i^2$ has a Pearson chi-squared distribution with $J-1$ degrees of freedom; $\chi^2_{(J-1)}$.

From (4)

$$\sum_{k=1}^{J-1} \left(f_{ik}^*\right)^2 = \frac{X_i^2}{np_{i\bullet}} \tag{5}$$

By comparison with (2) the radii length for the confidence circle of the $i'th$ row profile co-ordinate can be taken to be the square root of the right hand side of (5) with $X_i^2$ replaced by the $100\,(1-\alpha)\,\%$ point of its approximate distribution; $\chi^2_{J-1}$. When the ordinal correspondence plot consists of two dimensions, the square root of (5) with this replacement is identical to (2).

Confidence circles can also be calculated with the centre at the origin. Those points not contained within the circle all contribute to the dependency of the row and column variables that form the table. Those points lying within the circle, do not make such a contribution. [6] considered confidence circles with the centre at the origin, as well as circles with the origin at the position of the profile co-ordinate. However, Lebart et al. (1984, p183) state that

> In practice, instead of drawing concentric circles around the origin, it is clearer and easier to draw them around each point concerned, and look at the position of the origin.

The disadvantage of drawing a circle with the centre at the origin is that it assumes that points close to the origin will never significantly contribute to the dependency of the row and column variables, while those far from the origin will always make such a contribution. While this may occur in many situations, it will not always occur.

## 4. Relationship Between a Marginal Frequency and its Radii Length

Observing (2), the radii length will depend on the proportion of observations classified into a category of the contingency table.

A large proportion of observations classified will have a relatively small radii length, while a small proportion of classified observations will have a relatively large radii length. These observations can be seen in the application of confidence circles in Lebart et al. (1984, p51, Table 5).

The radii length defined by (2) shows that a variable with equi-probable responses will have equal length radii for each of the response. Therefore, when conducting an ordinal correspondence analysis on ranked data, as has been done by [5], the radius of the confidence circle for each of the row and column profile co-ordinates will be identical. For such an application, the length of the radii for all of the categories can be taken to be

$$r = \sqrt{t\,\frac{\chi^2_{(t-1)}}{n}} \tag{6}$$

where n is the number of judges (or consumers) who rank, according to their preference for, t products/treatments. The value of $\chi^2_{(t-1)}$ is the theoretical Pearson chi-squared value with t-1 degrees of freedom at the $\alpha$ level of significance. However, [1] noted that, as the rankings of a product/treatment are not independent, the Pearson chi-squared statistic does not have a chi-squared distribution, although $(t-1)\,X^2/t$ does. So in order to use (6), we use not the $X^2$ profile but the $(t-1)\,X^2/t$ profile which is $\{p_{ij}/p_{i\bullet}\}$ multiplied by $\{\sqrt{\frac{t-1}{t}}b_v\,(j)\}$.

## 5. Examples

### 5.1. Example 1 – Drug Data

Consider the contingency table given by Table 1 which was analysed in [4].

The study was aimed at testing four analgesic drugs (named A, B, C and D) and their effect on 121 hospital patients. The patients were given an ordered five point scale consisting of the categories *Poor*, *Fair*, *Good*, *Very Good* and *Excellent* on which to make their judgement.

It can be seen that Table 1 consists of ordered column categories and non-ordered categories. The natural scores 1, 2, 3, 4 and 5 are applied to the

*Table 1.* Cross-classification of 121 Hospital Patients According to Analgesic Drug and its Effect

| | Poor | Fair | Good | Very Good | Excellent |
|---|---|---|---|---|---|
| Drug A | 5 | 1 | 10 | 8 | 6 |
| Drug B | 5 | 3 | 3 | 8 | 12 |
| Drug C | 10 | 6 | 12 | 3 | 0 |
| Drug D | 7 | 12 | 8 | 1 | 1 |

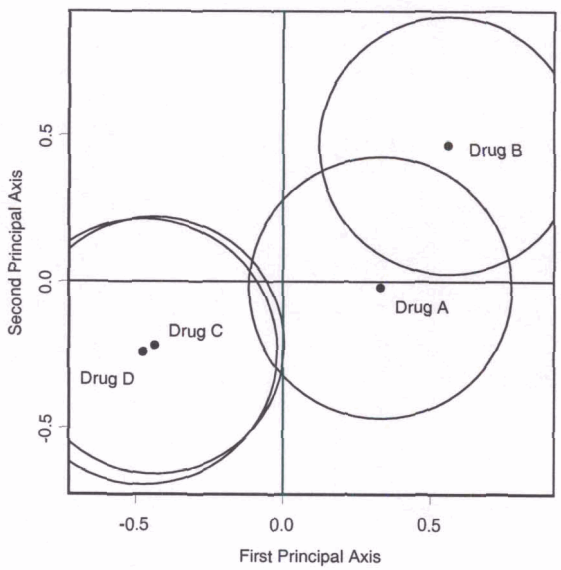


*Figure 1.* 95% Confidence Circles for the Drugs in Table 1

ordered column (judgement) categories, and the mean rank scores of 3.30, 3.23, 2.26 and 2.21 are applied to the non-ordered row (drug) categories.

The Pearson chi-squared statistic of the contingency table is 47.072, which at 12 degrees of freedom is highly significant. Therefore there is an asso-

ciation between the drug used and its effect on the patients. The ordinal correspondence plot for the row (drugs) profile co-ordinates is given by Figure 1. Similarly, the ordinal correspondence plot for the column (judgement) profile co-ordinates is given by Figure 2.

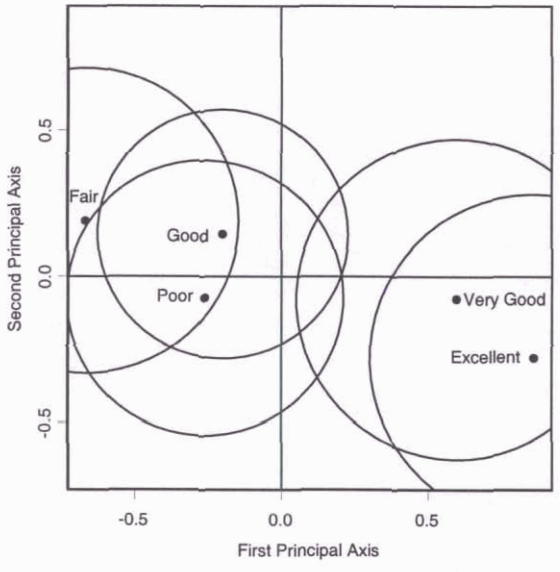


*Figure 2.* 95% Confidence Circles for the Judgements in Table 1

The row profile co-ordinates graphically depicted by Figure 1 are accompanied by the 95% confidence circle for each drug tested. Figure 2 also includes the 95% confidence circles for the judgement the patients gave for each drug. Both these figures consist of axes which reflect the variation in terms of the location (first principal axis) and dispersion (second principal axis) components. These two axis explain 75% of the variation in the drugs; location=54.07%, dispersion=20.93%. They also explain 80.21% of the variation in the judgements; location = 72.45%, dispersion = 7.76%. There are higher order moments which explain more of the variation in the categories than the dispersion, but we wish to highlight the variation only in terms of the location and dispersion components.

Figure 1 shows that drug A is the only drug to contribute to the independence hypothesis as the origin passes through its confidence circle. Therefore, drugs B, C and D have an effect on the results. We can see from Table 1 that drug B is rated as *Excellent*, drug C is rated *Poor* to *Good*, while drug D is considered to have a *Fair* effect on the patient.

Therefore, it would be advised that the drug associated with *Drug B* be used to treat patients suffering from analgesic illnesses.

Figure 2 shows that only *Poor* and *Good* contribute to the independence hypothesis. Therefore, *Excellent*, *Very Good* and *Fair* all can be used to characterise the drugs that were tested. In further studies we could possibly only consider a three-point scale rather than a five-point scale as was carried out in this experiment. By observing Figure 2, *Good* may be considered by some researchers as a descriptive response of the drugs as the origin barely falls within the confidence circle for this category.

### 5.2. Example 2 – Bean Data

Consider the bean data of [1] and analysed in [5]. A consumer study was conducted to determine which variety of snap bean was the most preferred. A lot of each of the three bean varieties were displayed in retail stores and 123 consumers were asked to rank the beans according to first, second and third choice. Table 2 lists the preferences of each variety of bean.

*Table 2.* Consumer Rankings of Three Varieties of Bean

| | Rank 1 | Rank 2 | Rank 3 | Total |
|---|---|---|---|---|
| Variety 1 | 42 | 64 | 17 | 123 |
| Variety 2 | 31 | 16 | 76 | 123 |
| Variety 3 | 50 | 43 | 30 | 123 |
| Total | 123 | 123 | 123 | 369 |

The Pearson chi-squared statistic for the Table 2 data is 79.561. Using the more appropriate Anderson chi-squared statistic, this value becomes 53.041, which at 4 degrees of freedom is highly significant. We use the Anderson chi-squared statistic rather than the Pearson value as Table 2 is a ranked data set where the rank assigned to a bean variety is not independent of the rank assigned to another bean. Therefore, there is a difference in ranking the three varieties of bean. However, by just observing

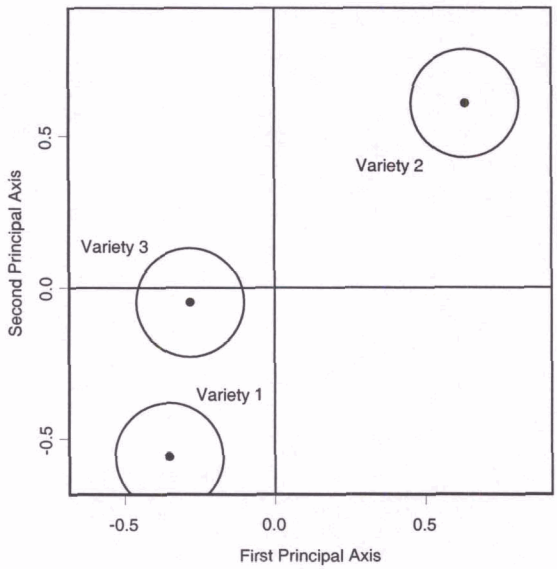


*Figure 3.* 95% Confidence Circles for the Bean Varieties in Table 2

Table 2 it is not evident which bean variety or rank contributes to this relationship. Confidence circles will determine which categories do so.

As there are three treatments, and therefore three rankings, the two-dimensional correspondence plot will describe all of the variation that exists for each category.

The row profile co-ordinates of Table 2 are presented as Figure 3 which also includes the 95% confidence circles for each bean variety. The column profile co-ordinates from the correspondence plot of Table 2 is included as Figure 4 and is constructed in the same manner as described in [5]. It also contains the 95% confidence circles for each rank.

As the row and column marginal frequencies are identical, the radius of each confidence circle will be equal; the radius length is 0.18018, using (2). This is verified by observing the radii length of each row and column profile co-ordinate in Figures 3 and 4 respectively.

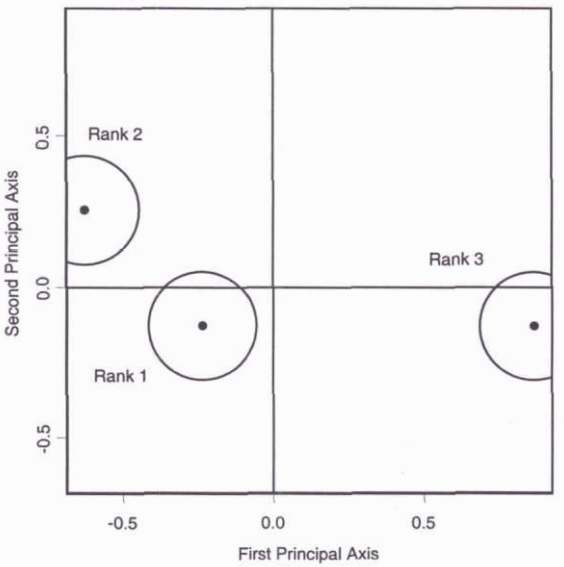


*Figure 4.* 95% Confidence Circles for the Bean Rankings in Table 2

Figures 3 and 4 show that all the bean varieties and the ranks contribute to the dependency in Table 2. These figures support the conclusion that there is a very strong association between the row and column categories.

## References

1. R.L. Anderson. Use of contingency tables in the analysis of consumer preference studies. *Biometrics*, 15:582–590, 1959.
2. D.F. Andrews. Plots of high dimensional data. *Biometrics*, 28:125–136, 1972.
3. E. J. Beh. Simple correspondence analysis of ordinal cross-classifications using orthogonal polynomials. *Biometrical Journal*, 39:589–613, 1997.
4. E. J. Beh. A comparative study of scores for correspondence analysis with ordered categories. *Biometrical Journal*, 40:413–429, 1998.
5. E. J. Beh. Correspondence analysis of ranked data. *Communication in Statistics (Theory and Methods)*, 28:1511–1533, 1999.

6. D. J. Best and J. C. W. Rayner. Product maps for ranked preference data. *The Statistician*, 46:347–354, 1997.
7. H. Chernoff. The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, 68:361–368, 1973.
8. M. J. Greenacre. *Theory and Application of Correspondence Analysis*. Academic Press, London, 1984.
9. D. L. Hoffman and G. R. Franke. Correspondence analysis : graphical representation of categorical data in marketing research. *The American Statistician*, 23:213–227, 1986.
10. L. Lebart, A. Morineau, and K.M. Warwick. *Multivariate Descriptive Statistical Analysis*. Wiley, New York, 1984.
11. P. A. Tukey and J. W. Tukey. Preparation; prechosen sequences of views. In V. Barnett, editor, *Interpreteting Multivariate Data*, pages 189–213. 1981.
12. P. A. Tukey and J. W. Tukey. Summarisation; smoothing; supplemented views. In V. Barnett, editor, *Interpreteting Multivariate Data*, pages 245–275. 1981.