A Few Moments for Simple Correspondence Analysis

Eric J. Beh

The University of Newcastle, Callaghan, NSW, 2308, AUSTRALIA
eric.beh@newcastle.edu.au

Abstract

This paper provides a simple interpretation of the coefficients of skewness and kurtosis of points in a correspondence plot. It helps to provide further information on the configuration of the coordinates and adds to that already provided by the first two moments which have helped to form the foundations of the mathematical development of correspondence analysis.

Keywords: Correspondence analysis, total inertia, skewness, kurtosis

1. Introduction

Suppose we consider an $I \times J$ two-way contingency table, N, where the (i, j)th cell entry is given by n_{ij} for $i = 1, \ldots, I$ and $j = 1, \ldots, J$. Let the grand total of N be n and the matrix of joint relative frequencies be P, so that the (i, j)th cell entry is $p_{ij} = n_{ij} / n$ and $\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = 1$. Define the ith row marginal proportion by $p_{i\bullet} = \sum_{j=1}^{J} p_{ij}$ and define the jth column marginal proportion as $p_{\bullet j} = \sum_{i=1}^{I} p_{ij}$. Let

$$\pi_{ij} = \frac{p_{ij} - p_{i\bullet}p_{\bullet j}}{\sqrt{p_{i\bullet}p_{\bullet j}}}$$

be the standardised residual of the (i, j)th cell. Classical correspondence analysis can be performed by applying a singular value decomposition to the standardised residuals so that

$$\pi_{ij} = \sum_{m=1}^{M} a_{im}\lambda_m b_{jm}$$

where $M = \min(I, J) - 1$ is the maximum number of dimensions required to graphically depict the association between the categories of the two variables. Here, a_{im} is the mth singular vector element associated with the ith row category and b_{jm} is the mth singular vector element associated with the jth row category. These quantities are constrained so that

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

The value λ_m is the mth largest singular value of the standardised residuals such that the Pearson chi-squared statistic can be partitioned by $X^2 = n\sum_{m=1}^{M}\lambda_m^2$.

In order to obtain a graphical summary of the association between the row and column variables, we may jointly represent the following row and column coordinates onto a low-dimensional plot :

$$f_{im} = \frac{1}{\sqrt{p_{i\bullet}}} a_{im}\lambda_m \qquad g_{jm} = \frac{1}{\sqrt{p_{\bullet j}}} b_{jm}\lambda_m .$$

Here, f_{im} is the coordinate of the ith row category along the mth axis. Similarly, g_{jm} is the coordinate of the jth column category along the same axis. Our concern in this paper is to consider the moments of these coordinates to quantify characteristics of the general configuration of the correspondence plot. In section 2 we briefly describe these moments and do so focusing on the row coordinates. Section 3 provides a simple illustration of how these moments may help to identify features of the configuration on the low-D plot. More information on the mathematical aspects of

correspondence analysis can be found in [1], [2], [3] and [5].

2. Moments

When identifying the characteristics of a configuration of points in a correspondence plot, the correspondence analysis literature exclusively focuses on the first two moments. However there has been no attention paid to the quantification or interpretation of third and fourth moments. Here we shall briefly describe the first four moments by adopting the notation in [4, section 3.11]. We start denoting the rth moment of the row coordinates about its mean (μ):

$$\mu'_r = E\left(f_{im}^r\right) = \sum_{i=1}^{I}\sum_{m=1}^{M} p_{i\bullet}\left(f_{im} - \mu\right)^r$$

2.1. The First Moment - the Mean of f_{im}

Define μ to be the weighted mean of the row profile coordinates such that $\mu = \sum_{i=1}^{I}\sum_{m=1}^{M} p_{i\bullet} f_{im}$. Therefore, the first moment, or mean, of these coordinates can be shown to be zero since,

$$\begin{aligned}\mu'_1 &= E\left(f_{im}\right) \\ &= \sum_{i=1}^{I}\sum_{m=1}^{M} p_{i\bullet}\left(f_{im} - \mu\right) \\ &= \sum_{i=1}^{I}\sum_{m=1}^{M} p_{i\bullet} f_{im} - \mu = 0\end{aligned}$$

Therefore, the configuration of points is centred at the origin of the correspondence plot. This is a well known characteristic of correspondence analysis.

2.2 The Second Moment – the Variance of f_{im}

The second moment, or variance, of the row coordinates is

$$\mu'_2 = \sigma^2 = \sum_{i=1}^{I}\sum_{m=1}^{M} p_{i\bullet} f_{im}^2$$

In the correspondence analysis literature, the right hand side of this expression is referred to as the *total inertia* of the contingency table and is mathematically equivalent to X^2/n. Also, $\sum_{i=1}^{I} p_{i\bullet} f_{im}^2$ is the principal inertia of the mth axis and quantifies a proportion of the total inertia that the mth principal axis makes in graphically depicting the association.

2.3. The Third Moment – the Skewness of f_{im}

The coefficient of skewness for the row coordinates can be calculated to be

$$\begin{aligned}\gamma_1 &= \frac{\mu'_3}{\sigma^3} \\ &= \frac{\sum_{i=1}^{I}\sum_{m=1}^{M} p_{i\bullet} f_{im}^3}{\left(\sqrt{\sum_{i=1}^{I}\sum_{m=1}^{M} p_{i\bullet} f_{im}^2}\right)^3} \\ &= \left(\frac{n}{X^2}\right)^{3/2} \sum_{i=1}^{I}\sum_{m=1}^{M} p_{i\bullet} f_{im}^3.\end{aligned}$$

Consideration of this quantity has not been made in the correspondence analysis literature and may be used to quantify how evenly spread, or symmetrically structured, around a particular axis, or plot, a configuration of points is. For example, if, for the mth axis, the skewness is negative this indicates that most of the row coordinates are positive. In a two-dimensional plot, a negative skewness coefficient will indicate that the configuration of points is dominated in the first quadrant of the display. A zero skewness coefficient will indicate that the points are evenly spread around the origin of the correspondence plot.

2.4 The Fourth Moment – the Kurtosis of f_{im}

The coefficient of kurtosis for the row coordinates can be calculated to be

$$\begin{aligned}\gamma_2 &= \frac{\mu'_4}{\sigma^4} - 3 \\ &= \frac{\sum_{i=1}^{I}\sum_{m=1}^{M} p_{i\bullet} f_{im}^4}{\left(\sum_{i=1}^{I}\sum_{m=1}^{M} p_{i\bullet} f_{im}^2\right)^2} - 3 \\ &= \left(\frac{n}{X^2}\right)^2 \sum_{i=1}^{I}\sum_{m=1}^{M} p_{i\bullet} f_{im}^4 - 3.\end{aligned}$$

A positive kurtosis coefficient suggests the row coordinates are clustered towards the origin of the display, indicating that there may be a number of row categories that do not contribute to the association structure of the variables. Similarly, a negative kurtosis coefficient will suggest that the configuration is not clustered near the origin, thus indicating that many of the categories contribute to the association.

3 Example and Interpretation

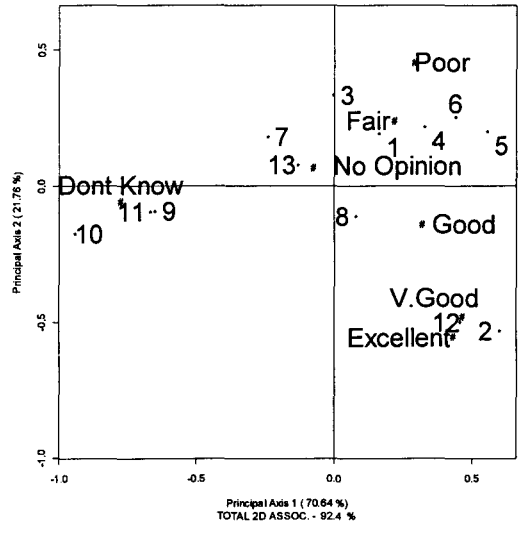
Consider Table 1 which was studied by [2, chapter 11] and is based on a 1969 study conducted by the French Radio & Television Organization (ORTF). A sample of 400 people were asked to evaluate each of 13 variety shows on a seven five point scale, with the option of indicating that they have "No Opinion" of the show, or "Don't Know". Unfortunately not all of the people who participated in the study provided their evaluation – for each show between 389 and 391 people responded.

Table 1: Evaluation of 13 variety shows in 1969

	Evaluation						
Variety Show	*Excellent*	*Very Good*	*Good*	*Fair*	*Poor*	*No Opinion*	*Don't Know*
1	9	28	89	124	51	19	71
2	31	87	165	63	24	4	17
3	7	21	65	103	83	8	103
4	3	26	121	142	45	11	43
5	17	40	117	111	83	16	7
6	8	35	115	119	78	6	28
7	4	22	73	56	77	12	147
8	15	44	102	83	32	25	90
9	5	18	63	61	15	9	219
10	8	15	40	37	8	12	271
11	5	16	64	54	15	17	220
12	29	87	140	62	24	9	40
13	12	18	89	95	41	9	127

The Pearson chi-squared statistic of Table 1 is 1675.52 and with a p-value < 0.0001 indicates that there is a statistically significant association between *Evaluation* and *Variety Show*. With n = 5079, the total inertia of the data is 1675.52/5079 = 0.3299 and is a measure of the strength of the association removing the impact of the sample size.

Figure 1: 2-D Correspondence Plot of Table 1



By observing Figure 1 it appears that, while the row and column coordinates are centred about the origin, the configuration has a horseshoe shape where the first and second quadrant contain most of the points. This suggests a certain amount of skewness and kurtosis in the plot. Table 2 provides, for the row points, a summary of the quantities of centre, spread, skewness and kurtosis for each axis of the two-dimensional plot of Figure 1 and of the optimal (three-dimensional) plot. As expected the row points are centred about the origin of the display and so the mean of the coordinates along each axis is zero. Similarly, the quantities that reflect the spread of the coordinates are equivalent to the principal inertia for each axis, or in the case of the optimal plot, the total inertia. Table 3 provides a similar summary for the column points.

Table 2: Moments for the Row Coordinates in Figure 1

	Axis 1	*Axis 2*	*2-D Plot*	*Optimal Plot*
Centre	0	0	0	0
Spread	0.2330	0.0718	0.3048	0.3299
Skewness	-0.5477	-0.7511	-0.1309	-0.4042
Kurtosis	-0.9161	-0.5498	-1.6461	-1.8358

Table 3: Moments for the Column Coordinates in Figure 1

	Axis 1	*Axis 2*	*2-D Plot*	*Optimal Plot*
Centre	0	0	0	0
Spread	0.2330	0.0718	0.3048	0.3299
Skewness	-0.9078	-0.0754	-0.6154	-0.5655
Kurtosis	-1.0327	-0.4273	-1.7074	-1.8802

Figure 1 provides a graphical perspective of the association among the variables and accounts for 92.4% of the total inertia. It indicates that the variety shows 2 and 12 were generally rated as *Excellent* while shows 1, 3, 4, 5 and 6 were rated *Fair* to *Poor*. There was some disinterest in shows 7 and 13 where people had no opinion of them, while people were generally unsure on how to evaluate shows 9, 10 and 11.

Figure 2a: Row coordinates along the first principal axis

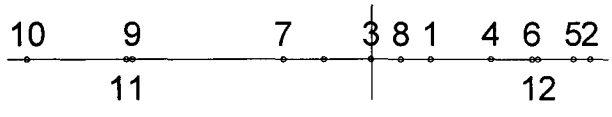
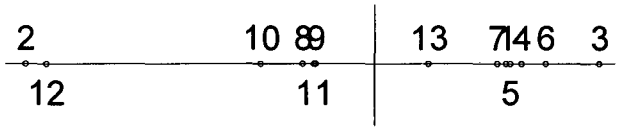


Figure 2b: Row coordinates along the second principal axis



Suppose we consider the row coordinates that are depicted in Figure 1. The negative skewness coefficients along each of the two axes indicates that most of the points have a positive first and second coordinate (hence lying in the first quadrant of Figure 1). This is evident by observing Figures 2a and 2b which focus on their relative position along the first and second axis, respectively. Since the configuration is centred about the origin, the points that lie on the negative side of each axis may potentially be outlying categories, or categories that contribute to the association in a very different manner than the remaining categories. The negative kurtosis coefficient indicates, for each axis, that the configuration is not dominated by points that are clustered near the origin. Therefore the negative skewness (-0.1309) and kurtosis (-1.6461) coefficients associated with the two-dimensional plot of Figure 1 (or of Figures 2a and 2b) suggest that the row points are not heavily clustered near the origin (as it shows). Therefore, there may be some potential for nearly all of the categories to contribute to the association structure between the two variables. Incorporating the confidence circles of Lebart, Morineau & Warwick (1984) can aid in verifying which points make such a contribution.

Figure 3a: Column coordinates along the first principal axis

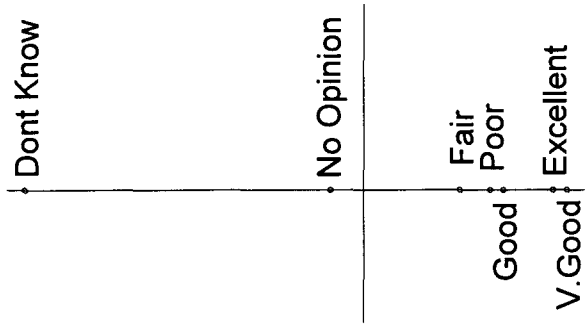
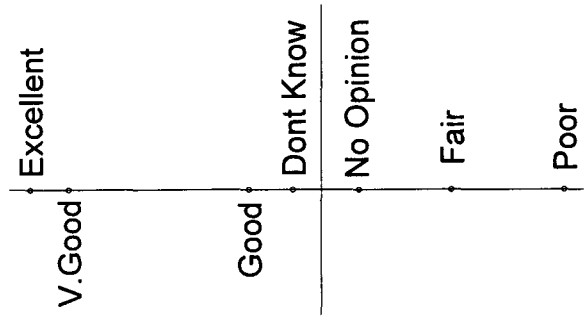


Figure 3b: Column coordinates along the second principal axis



Figures 3a and 3b depict the column coordinates along the first and second principal axes, respectively. Their large negative skewness coefficient along the first axis (-0.9708) is apparent when observing the relative position of points in Figure 3a, while this coefficient is close to zero along the second axis (-0.0754). We can see that, for Figure 1, their is a much larger negative skewness coefficient for the column points than for the row points (-0.6154). This is largely because of the dominance of "Dont Know" on the left hand side of Figure 1 (and Figure 3a). The kurtosis coefficient of the column points (-1.7074) in the two-dimensional display is similar to that of the row points.

4 Discussion

While the moments of centre and spread have become widely adopted and interpretable in the correspondence analysis literature, this paper provides a means of interpreting the coefficients of skewness and kurtosis of a configuration of points in a correspondence plot. We have briefly shown that skewness reflects how evenly distributed around the origin the configuration of points is, leading to the potential to identify categories that impact upon the association structure differently than the other categories. Similarly, kurtosis provides some guide as to whether there is a dominant cluster of points near the origin. Of course, such quantitative measures can be used in the context of analysing the association structure between multiple categorical variables. This issue will be left for future consideration.

References

[1] E. J. Beh, "Simple correspondence analysis: A bibliographic review", *International Statistical Review*, 72, 257-284. 2004.

[2] J. P. Benzecri, (1992), *Correspondence Analysis Handbook*, Marcel Dekker, Inc: New York. 1992.

[3] M. J. Greenacre, *Theory and Applications of Correspondence Analysis*, Academic Press: London. 1984.

[4] R. J. Larsen, M. L. Marx, *An Introduction to Mathematical Statistics and its Applications (Second Edition)*, Prentice-Hall: New Jersey. 1986.

[5] L. Lebart, A. Morineau, K. M. Warwick, *Multivariate Descriptive Statistical Analysis*, Wiley: New York. 1984.