Profile Transformations for Reciprocal Averaging & Singular Value Decomposition

Ting-Wu Wang, Eric J. Beh

School of Mathematical & Physical Sciences,
University of Newcastle, Australia

1 Abstract

The traditional approach to correspondence analysis (CA) and the method of reciprocal averaging (RA) has close links in the analysis of the association between categorical variables. The need for power transformation of cell frequencies have been of interest in the past decade in the CA literature. This issue has been discussed based on the power transformation of the proportional frequencies which involves taking the power of the sum of untransformed cells. This paper has demonstrated the connections between CA and RA under the proportional frequencies of power-transformed cells. A brief application was used to demonstrate the algorithms between how the two methods acquire solutions that are a weighted average of one another. In the calculation of the scores and the association between them, two alternative methods of choosing an appropriate power parameter are considered. One approach calculates the power parameter that explains the optimal association as a percentage of the phi-squared statistic from the first dimension of the solution. The second method involves determining a power transformation, such that the data becomes non-dispersed.

2 Introduction

The scoring of rows and columns for a two-way contingency table have been a topic of much discussion in the statistics literature for nearly a century. Two inter-related techniques that have become widespread include reciprocal averaging (RA) and correspondence analysis (CA) (Hill, 1974; Beh and Lombardo, 2014). RA typically determines a one-dimensional solution to the set of scores that reflects the maximum association possible between them but it can also be generalised to determine scores along multiple orthogonal dimensions. These scores also highlight those rows (and columns) that are similar or different in terms of their relative distribution which, in CA, is referred to as the categories *profile*. The mathematical derivation of the method that calculates these scores while maximising the correlation between them was established by Hirschfeld (1935) and its links to RA and CA were highlighted by Hill (1974).

The multi-dimensional solution to RA can also be determined by applying the singular value decomposition (SVD) to the matrix of Pearson's residuals. In this case the singular values are akin to the maximum correlations found from RA and their sum-of-squares gives Pearson's phi-squared statistic. Therefore, finding the scores and their correlation via SVD is directly linked to more formal tests of association used in categorical data analysis and typically done in the CA literature.

An issue that has arisen over the past decade is the need to consider a power transformation of the cell frequencies. Greenacre (2009) studied the problem when linking his log-ratio analysis (LRA) to the traditional approach to CA. Such a strategy is related to the CA approach discussed by Cuadras and Cuadras (2006) which uses, at its heart (although not explicitly stated in their paper) the Freeman-Tukey statistic. More recently, Beh and Lombardo (2021) expanded upon their methods and showed how a more general parametric approach to CA could be achieved where the Cressie-Read divergence statistic

can be used as the core measure of association, thereby including as special cases the log-likelihood ratio statistic, the Freeman-Tukey statistic and others.

Despite the linkages of these techniques they are based on the power transformation of the profile elements. A limitation of this is that, while they involve the power transformation of the cells of the contingency table, they involve the power transformation of the marginal row frequencies of the untransformed cells. Therefore, while Section 3 provides an overview of the link between RA and the SVD of the matrix of Pearson residuals, this paper will outline how the power transformation applied to each cell of the contingency table applies to the method RA (Section 4), thereby giving marginal information that involves the summation of these power-transformed cells. We shall also show how such a solution can be achieved using SVD (Section 4). A brief practical example will be described that shows the equivalency of these two approaches for three values of the power applied to the cells of the contingency table (Section 5).

3 Classical Approach to Reciprocal Averaging

Suppose that the (i, j)th element of an $I \times J$ contingency table is denoted by n_{ij} and that $p_{ij} = n_{ij}/n$, where $n = \sum_{i=1}^{I} \sum_{j=1}^{j} n_{ij}$ is the sample size. Denote $p_{i\bullet} = \sum_{j=1}^{J} p_{ij}$ and $p_{\bullet j} = \sum_{i=1}^{I} p_{ij}$ to be the ith row and jth column, respectively, marginal proportion of the contingency table.

RA involves calculating row (and column) scores that maximise the correlation between the variables. This is achieved by identifying similarities and differences in a row (and column) profile. Profiles can be defined in the following manner: the jth element of the ith centred row profile, and the ith element of the jth centred column profile are defined, respectively, by

$$\frac{p_{ij}}{p_{i\bullet}} - p_{\bullet j} \qquad \text{and} \qquad \frac{p_{ij}}{p_{\bullet j}} - p_{i\bullet}\,.$$

Hirschfeld (1935) and Hill (1974) describe RA as follows. The row scores $\boldsymbol{a} = (a_1, a_2, \cdots, a_I)^T$, and the column scores, $\boldsymbol{b} = (b_1, b_2, \cdots, b_J)^T$, are calculated such that their correlation, λ, is maximised. These quantities can be determined by solving (reciprocally) the following two equations which, in matrix form, are

$$\lambda \boldsymbol{a} = \left(\mathbf{R}^{-1}\mathbf{P} - \mathbf{1}_I \mathbf{c}^T\right) \boldsymbol{b} \tag{1}$$

and

$$\lambda \boldsymbol{b} = \left(\mathbf{C}^{-1}\mathbf{P}^T - \mathbf{1}_J \mathbf{r}^T\right) \boldsymbol{a}. \tag{2}$$

Here, $\mathbf{1}$ denotes the vector of 1's of the length specified by its subscript. The matrix $\mathbf{P}$ is consists of the joint relative frequencies whose (i, j)th element is p_{ij}, while $\mathbf{R}$ denotes the diagonal matrix of marginal row proportions diag $(p_{i\bullet})$ for $i = 1, 2, \ldots, I$ and $\mathbf{C}$ is the diagonal matrix of marginal column proportions diag $(p_{\bullet j})$ for $j = 1, 2, \ldots, J$. After a number of iterations, the scores will converge to give an optimal solution which maximises the correlation between them. The solution of the row scores, $\boldsymbol{a}$, is constrained by the Gram-Schmidt orthogonalisation, $\boldsymbol{a}^T\mathbf{R}\boldsymbol{a} = 1$ while column score, $\boldsymbol{b}$, is constrained by $\boldsymbol{b}^T\mathbf{C}\boldsymbol{b} = 1$.

Beh and Lombardo (2014, Section 3.5.2) showed that by pre-multiplying both side of Equation (1) by $\mathbf{R}^{1/2}$ and Equation (2) by $\mathbf{C}^{1/2}$, the solution to $\boldsymbol{a}$, $\boldsymbol{b}$ and λ found via RA can also be determined from the one-dimensional solution to the singular value decomposition (SVD) of

$$\mathbf{Z} = \mathbf{R}^{-1/2}(\mathbf{P} - \mathbf{r}\mathbf{c}^T)\mathbf{C}^{-1/2}\,.$$

Here, $\mathbf{Z}$ is the matrix of Pearson's residuals and finding the scores in this manner is typically undertaken in the CA literature.

4 Power Transformation and Reciprocal Averaging

There are times when one may wish to perform a power transformation to the contingency table. Greenacre (2009) and Beh and Lombardo (2021) describe when it may be appropriate to do so and discusses the technical merits of the approach. Their approach involves, for some value of δ, performing CA where a comparison of the centred profiles are of the form

$$\left(\frac{p_{ij}}{p_{i\bullet}}\right)^{\delta} - p_{\bullet j}^{\delta} \qquad \text{and} \qquad \left(\frac{p_{ij}}{p_{\bullet j}}\right)^{\delta} - p_{i\bullet}^{\delta}\,.$$

For such a case, $p_{i\bullet}$, say, does not involve the summation (across the columns) of p_{ij}^{δ}. Nor does $p_{\bullet j}$ involve the summation (across the rows) of p_{ij}^{δ}. Instead, $p_{i\bullet}^{\delta}$ (say) is the power transformation of the marginal proportion of the ith row and does not involve the power transformation of the cells. Therefore, we shall consider the case where these two expressions are adjusted so that the marginal proportions reflect the summation (across both variables) of the power transformation of the cells of the contingency table. To do so, we start with the hypothesis of complete independence which, for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$, is $n_{ij} = n_{i\bullet}n_{\bullet j}/n$ where $n_{i\bullet} = \sum_{j=1}^{J} n_{ij}$ and $n_{\bullet j} = \sum_{i=1}^{I} n_{ij}$. This independence hypothesis may be equivalently expressed as $p_{ij} = p_{i\bullet}p_{\bullet j}$. Note that if a power transformation were applied to the cells then $p_{ij}^{\delta} \neq (p_{i\bullet}p_{\bullet j})^{\delta}$. Instead, we may consider a power transformation so that

$$p_{ij}(\delta) = p_{i\bullet}(\delta)p_{\bullet j}(\delta)$$

where $p_{ij}(\delta)$ is the proportional frequency after after every element of n_{ij} is raised to the power of δ. Namely, $p_{ij}(\delta) = n_{ij}^{\delta}/n_{\delta}$ where n_{δ} is the power-transformed sample size $n_{\delta} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}^{\delta}$. Similarly, $p_{i\bullet}(\delta)$ is the sum of the row marginal proportions and $p_{\bullet j}(\delta)$ is the column marginal proportions after every element of n_{ij} is raised to the power of δ. We point out that, unlike the power transformations of Greenacre (2009), Cuadras and Cuadras (2006) and Beh and Lombardo (2021) these three quantities involve dividing the elements (and sum of the elements) n_{ij}^{δ} by n_{δ} not n.

Under such a setting, the RA of the power-transformed centred profiles leads to the two formulae

$$\lambda_{\delta}\boldsymbol{a}_{\delta} = \left(\mathbf{R}_{\delta}^{-1}\mathbf{P}_{\delta} - \mathbf{1}_{I}\mathbf{c}_{\delta}^{T}\right)\boldsymbol{b}_{\delta} \tag{3}$$

and

$$\lambda_{\delta}\boldsymbol{b}_{\delta} = \left(\mathbf{C}_{\delta}^{-1}\mathbf{P}_{\delta}^{T} - \mathbf{1}_{J}\mathbf{r}_{\delta}^{T}\right)\boldsymbol{a}_{\delta} \tag{4}$$

where $\boldsymbol{a}_{\delta}$ is the set of row scores for a power transformation of the cells of the contingency table that is of the order δ and $\boldsymbol{b}_{\delta}$ is the set of column scores under such a transformation. Also, $\mathbf{R}_{\delta}^{-1}$ is the inverse of the diagonal matrix of elements of $p_{i\bullet}(\delta)$, $\mathbf{C}_{\delta}^{-1}$ is the inverse of the diagonal matrix of elements $p_{\bullet j}(\delta)$, P_{δ} is the matrix whose (i, j) element is $p_{ij}(\delta)$. Note that $\mathbf{1}$ is a vector of 1's with the length specified by its subscript, while $\mathbf{c}_{\delta} = \text{vec}\left(p_{\bullet j}(\delta)\right)$ and $\mathbf{r}_{\delta} = \text{vec}\left(p_{i\bullet}(\delta)\right)$. Where $\boldsymbol{a}_{\delta}$ is contrained by $\boldsymbol{a}_{\delta}^{T}\mathbf{R}_{\delta}\boldsymbol{a}_{\delta} = 1$ while $\boldsymbol{b}_{\delta}$ is constrained by $\boldsymbol{b}_{\delta}^{T}\mathbf{C}_{\delta}\boldsymbol{b}_{\delta} = 1$

We can show that under this framework, solving these two RA formulae gives an identical solution to the SVD of the power-transformed analog of the matrix of Pearson residuals. That is, by pre-multiplying both sides of Equation (3) by $\lambda\mathbf{R}_{\delta}^{1/2}$ and pre-multiplying both sides of Equation (4) by $\lambda\mathbf{C}_{\delta}^{1/2}$ leads to the SVD of the matrix

$$\mathbf{Z}_{\delta} = \mathbf{R}_{\delta}^{-1/2}[\mathbf{P}_{\delta} - \mathbf{r}_{\delta}\mathbf{c}_{\delta}^{T}]\mathbf{C}_{\delta}^{-1/2}\,.$$

5 Example: Selikoff's Asbestos Data

Consider an example from Selikoff's studies on the association of occupational exposure to asbestos to the grade of asbestosis diagnosed by chest X-ray (refer to Beh and Lombardo (2014, Table 1.3)). 1117 workers with different lengths of exposure to asbestos (in intervals of 9 years) were classified

into different grades (in severity) of asbestosis according to their diagnosis. Table 1 illustrates the comparison of the first dimension of the column scores and its association with the row scores between power transformation of RA and CA with $\delta = 1, 0.5$ and 0.001. This table demonstrates the row and column scores and their association are equivalent (to 5 d.p.) between RA and CA under the same power transformation.

Table 1: Comparison of scores

	$\delta = 1$		$\delta = 0.5$		$\delta \approx 0$	
	RA	CA	RA	CA	RA	CA
λ_δ	0.6994048	0.6994048	0.5507601	0.5507601	0.3455012	0.3455012
Column scores						
d_1	-0.6075334	-0.6075332	-0.65556607	-0.65556607	-0.4279660	-0.4279667
d_2	0.2378942	0.2378939	0.01693329	0.01693331	-0.4256346	-0.4256353
d_3	0.6043241	0.6043242	0.52262561	0.52262561	0.3240548	0.3240581
d_4	0.4572767	0.4572769	0.54480169	0.54480167	0.7284701	0.7284679

The results of RA discussed in this paper calculates a one-dimensional solution to the maximum association between the row scores and the column scores. A way to identify the most appropriate δ is by exploring the impact of change of δ on the percentage of association explained by the first dimension of λ_δ. The maximum percentage of association explained by the first dimension of λ_δ is 77% with a power transformation of $\delta = 0.36$. By applying a power transformation of RA and CA with $\delta = 0.36$ for Selikoff's data, we obtain the maximum association between the row scores and the column scores from the first dimension of the solution.

Another approach for finding the most appropriate power transformation aims to find a dispersion-free dataset. A common difficulty when analysing categorical variables of a contingency table is that cell frequencies are often prone to overdispersion as they are typically assumed to be Poisson random variables. One measure of dispersion is by taking the ratio of Pearson's chi-squared statistic to its degrees of freedom. The assumption of equal mean and variance is satisfied when this ratio is one, otherwise there is evidence of under/overdispersion. The degrees of freedom for Selikoff's data is (number of rows - 2)(number of columns - 2) = 6 degrees of freedom. The power transformation of $\delta = 0.18$ finds the chi-squared statistic of this data to be 6 (to 3 d.p.). Therefore, it is possible to analyse the contingency table dispersion-free by applying power transformation of profiles for RA or CA with $\delta = 0.18$.

References

Beh, E. J. and Lombardo, R. (2014). *Correspondence analysis: Theory, practice and new strategies.* John Wiley & Sons.

Beh, E. J. and Lombardo, R. (2021). Correspondence analysis and the cressie-read divergence statistic. *(In review).*

Cuadras, C. M. and Cuadras, D. (2006). A parametric approach to correspondence analysis. *Linear algebra and its applications*, 417(1):64–74.

Greenacre, M. (2009). Power transformations in correspondence analysis. *Computational Statistics & Data Analysis*, 53(8):3107–3116.

Hill, M. O. (1974). Correspondence analysis: a neglected multivariate method. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 23(3):340–354.

Hirschfeld, H. O. (1935). A connection between correlation and contingency. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 31, pages 520–524. Cambridge University Press.