NOVA

University of Newcastle Research Online

nova.newcastle.edu.au

Beh, Eric J. "Simple correspondence analysis using adjusted residuals" *Journal of Statistical Planning and Inference* Vol. 142, Issue 4, p. 965-973 (2012)

Available from: <http://dx.doi.org/10.1016/j.jspi.2011.11.004>

Accessed from: <http://hdl.handle.net/1959.13/1054339>

Simple Correspondence Analysis using Adjusted Residuals

Eric J. Beh

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

Abstract

Correspondence analysis is a versatile statistical technique that allows the user to graphically identify the association that may exist between variables of a contingency table. For two categorical variables, the classical approach involves applying singular value decomposition to the Pearson residuals of the table. These residuals allow for one to use a simple test to determine those cells that deviate from what is expected under independence. However, the assumptions concerning these residuals are not always satisfied and so such results can lead to questionable conclusions.

One may consider instead, an adjustment of the Pearson residual, which is known to have properties associated with the standard normal distribution. This paper explores the application of these adjusted residuals to correspondence analysis and determines how they impact upon the configuration of points in the graphical display.

Key words: Association; Pearson residuals; Standardised Residuals; Adjusted Pearson Residuals; Weighted Chi-Squared Statistic.

1 Introduction

The detection of associations between two or more variables of a contingency table is the primary aim of categorical data analysis. There are many techniques available that provide numerical summaries of association but it is correspondence analysis that has long been used to provide a graphical perspective to this issue. Classically, the procedure involves the decomposition of Pearson residuals using singular value decomposition thereby allowing the user to view the “correspondence” between categories in low-dimensional space.

Email address: `eric.beh@newcastle.edu.au` (Eric J. Beh).

There is much in the literature that describes the theoretical aspects and application of correspondence analysis. For example, one may refer, but not feel restricted, to consulting Greenacre (1984), Lebart, Morineau and Warwick (1984), Weller and Romney (1990), Benzecri (1992), Beh (2004a), Le Roux and Rouanet (2005), Murtagh (2005) and Greenacre and Blasius (2006) for further details on the theory of correspondence analysis.

Pearson residuals have the advantage that their sum of squares leads to the Pearson chi-squared statistic - the most common measure of association for cross-classifications. They are also known, in some cases, to be asymptotically standard normally distributed. However, studies in the past have found that the residuals do not exhibit unit variance and so using them to identify those cells that are not consistent with the hypothesis of independence can lead to spurious conclusions. As a result, Haberman (1973) and Agresti (2002) suggest that one may consider instead an adjustment of the Pearson residuals taking into account their variance. As a result, adjusted Pearson residuals are generally better suited to determine those cells that deviate from what is expected under the hypothesis of independence. For the sake of brevity, the adjusted Pearson residuals will be simply referred to as adjusted residuals throughout this paper.

This paper will demonstrate how adjusted residuals may be used in the correspondence analysis of a two-way contingency table. Such an approach to correspondence analysis will allow for additional variation in the categories to be graphically detected that may exist in the cells; variation that may otherwise be missed if focusing on the Pearson residuals.

The adjusted residuals considered in this paper are those proposed by Haberman (1973) and have been used for determining the adequacy of a class of log-linear and logit models (Lang and Agresti, 1994). Upton (2000) also used the adjusted residuals for the construction of “cobweb” diagrams. Generally though, consideration of adjusted residuals for the analysis of association in contingency tables has been rare. This paper explores the application of such residuals to correspondence analysis.

This paper is divided into three further sections. Section 2 will define the Pearson residual and its adjusted version for the two-way contingency table. The correspondence analysis of a table using both these residuals will be described with an example. It is shown that, for such an analysis, going from Pearson residuals to their adjusted versions results in a shift in the position of the profile points in the low-dimensional correspondence plot. Such a shift, referred to here as stretching, is further studied in Section 3. The paper is concluded (Section 4) with a brief discussion of the procedure and some possible avenues for further research.

2 Correspondence Analysis

2.1 Pearson Residuals

Consider an $I \times J$ two-way contingency table N , which has a grand total of n , and consists of I row categories and J column categories. Denote the matrix of relative proportions by $P = N/n$ where the (i, j) 'th cell entry of P is p_{ij} so that $\sum_{i=1}^I \sum_{j=1}^J p_{ij} = 1$. Let D_I be a diagonal matrix of size $I \times I$ where the (i, i) th element is the i th row marginal proportion by $p_{i\bullet} = \sum_{j=1}^J p_{ij}$. Similarly, let D_J be the $J \times J$ diagonal matrix where the (j, j) th element is the j th column marginal proportion by $p_{\bullet j} = \sum_{i=1}^I p_{ij}$.

When identifying those row (or column) categories that are similarly, or differently, distributed correspondence analysis considers the difference between the profiles of the row categories, denote in matrix form by $D_I^{-1}P$, and the column profiles, $D_J^{-1}P^T$. This allows for one to compare, not the joint proportion of a cell, but its proportion conditional the marginal frequency of that category.

To begin our discussion of correspondence analysis let us consider the hypothesis of independence between the row categories and column categories. For complete independence to exist the relative cell proportions can be expressed as

$$P = rc^T$$

where $r = \text{diag}(D_I)$ and $c = \text{diag}(D_J)$. To formally measure the departure from independence, the Pearson chi-squared statistic

$$X^2 = n \times \text{trace} \left[D_I^{-1/2} (P - rc^T) D_J^{-1} (P - rc^T)^T D_I^{-1/2} \right]$$

can be calculated and compared with theoretical chi-squared value χ_α^2 with $(I - 1)(J - 1)$ degrees of freedom at the α level of significance. If $X^2 > \chi_\alpha^2$ then there is enough information in the sample to suggest that there is an association between the two categorical variables. Since the Pearson chi-squared statistic is proportional to the size of the sample, correspondence analysis uses as a measure of association for the contingency table X^2/n . This is referred to as the total inertia of the table.

One may consider the matrix of Pearson residuals

$$R = D_I^{-1/2} (P - rc^T) D_J^{-1/2} \tag{1}$$

for determining those cell's of the table that are not consistent with the hypothesis of independence. Of course the sum of squares of these residuals gives the total inertia $X^2/n = \text{trace}(R^T R) = \text{trace}(R R^T)$. For large sample sizes, one may consider $\sqrt{n}r_{ij}$ (where r_{ij} is the (i, j) th element of R and correspondences to the Pearson residual for that cell) to be asymptotically standard normally distributed. This is a very useful property since identifying those cells that are not consistent with independence is easily done. The normality property of (1) holds under the multivariate central limit theorem for simple random sampling (Rayner and Best, 2001, pg 33-34).

Haberman (1973) refers to $\sqrt{n}r_{ij}$ as the *standardised residual* of the (i, j) th cell of the contingency table. Agresti (2002, pg 81) calls them *Pearson residuals*. This is because of their obvious link with the Pearson chi-squared statistic. For this reason, this paper will also refer to the residuals (1) as Pearson residuals.

The aim of correspondence analysis is to provide the user with a low-dimensional graphical display of the association between categorical variables. Such a display is referred to as a correspondence plot. The classical approach to performing such an analysis involves applying singular value decomposition to the Pearson residuals of the contingency table so that

$$R = AD_\lambda B^T \quad (2)$$

Here, the m th column vector of A , $a_m = (a_{1m}, a_{2m}, \dots, a_{Im})$, is the m th row singular vector and is associated with the I row categories. Similarly, the m th column vector of B , $b_m = (b_{1m}, b_{2m}, \dots, b_{Jm})$, is the m th column singular vector and is associated with the J column categories. In both cases, $m = 1, 2, \dots, M^* = \max(I, J) - 1$.

The (m, m) th element of the diagonal matrix D_λ is λ_m and is real and positive and is the m th largest singular value of the matrix of R . The singular values are arranged in descending order so that

$$1 > \lambda_1 \geq \dots \geq \lambda_{M^*} \geq 0.$$

and the matrix D_λ can be calculated by

$$D_\lambda = A^T P B$$

such that

$$A^T A = B^T B = I$$

where I here is the identity matrix.

To obtain a graphical representation of the association between the row and column variables one way to define the points in a low-dimensional space is to consider the following row and column profile coordinates

$$F = D_I^{-1/2} A D_\lambda \quad (3)$$

$$G = D_J^{-1/2} B D_\lambda. \quad (4)$$

Alternatively the coordinates

$$F' = D_I^{-1/2} A D_\lambda^\gamma \quad G' = D_J^{-1/2} B D_\lambda^\delta$$

with $\gamma + \delta = 1$ may also be used. Refer to Goodman (1986), Aitchison and Greenacre (2002) and Beh (2004a) for further details on these coordinates.

Using the coordinates (3) and (4), and the singular vectors λ_m , the total inertia of the contingency table can be expressed as

$$\begin{aligned} \frac{X^2}{n} &= \text{trace} \left(F^T D_I F \right) \\ &= \text{trace} \left(G^T D_J G \right) \\ &= \text{trace} \left(D_\lambda^2 \right). \end{aligned} \quad (5)$$

For a proof of these expressions refer to, for example, Greenacre (1984) or Beh (2004a). The first two of these expressions show that coordinates close to the origin reflect a small association between the variables. Therefore, the origin of the display coincides with where the coordinates would be if there was complete independence in each of the cells of the table.

2.2 Adjusted Residuals

Consideration of residuals other than the Pearson residual is not new to correspondence analysis. There are a number papers that have appeared in the literature that use residuals other than (1), although they are primarily concerned with modelling alternatives to independence. For example one may refer to Greenacre (1988) who looked at approximating N by another matrix but of lower rank. van der Heijden, de Falguerolles and de Leeuw (1989) demonstrate the usefulness of correspondence analysis when measuring the departure from what is expected under the hypothesis of symmetry, quasi-symmetry and quasi-independence.

Recall from Section 1 that the asymptotic normality property of the Pearson

residual, $\sqrt{nr_{ij}}$, allows for a simple exploration of those cell's that are not consistent with what is expected under independence. While this is a very useful property studies have shown that it is rarely observed. Agresti (2002, pg 81) comments that “their asymptotic variances are less than 1.0, averaging $[(I - 1)(J - 1)] / (\text{number of cells})$ ”. Haberman (1973) points out that, under independence, the maximum likelihood estimate (using a Poisson or multinomial model) of the variance of $\sqrt{nr_{ij}}$ is $(1 - p_{i\bullet})(1 - p_{\bullet j})$. A recent proof of this result can be found by referring to Agresti (2002, pg 588-589).

To modify the Pearson residual in the manner considered by Haberman (1973) and Agresti (2002) a correspondence analysis can be performed by considering

$$\begin{aligned}\tilde{R} &= (I - D_I^{-1/2}) R (I - D_J^{-1/2}) \\ &= [D_I^{-1/2} (I - D_I^{-1/2})] (P - rc^T) [D_J^{-1/2} (I - D_J^{-1/2})].\end{aligned}\quad (6)$$

By considering the (i, j) th element of $\tilde{R}$, $\tilde{r}_{ij}$, Haberman (1973) refers to $\sqrt{n}\tilde{r}_{ij}$ as the *adjusted residual* of the (i, j) cell of the contingency table while Agresti (2002, pg 81) calls it the *standardised Pearson residual*. By modifying the Pearson residuals so that its variance is taken into account, $\sqrt{n}\tilde{r}_{ij}$ is asymptotically standard normally distributed.

By considering the matrix of adjusted residuals (6), the adjusted, or weighted, Pearson chi-squared statistic becomes

$$\begin{aligned}\tilde{X}^2 &= n \times \text{trace} \left[D_I^{-1/2} (I - D_I)^{-1/2} \right. \\ &\quad \left. (P - rc^T) D_J^{-1} (I - D_J)^{-1} (P - rc^T)^T D_I^{-1/2} (I - D_I)^{-1/2} \right]\end{aligned}\quad (7)$$

Golden (2000, pg 38) presents a means of calculating p-values based on the computational procedure outlined in Sheil and O’Muircheartaigh (1977). Further information on the weighted chi-squared statistic can be found in Davies (1980), Wood and Cady (1981, pg 6) and Davis (1982). Alternatively, formal tests of association between two cross-classified categorical variables can be made by considering the M-test of Fuchs and Kenett (1980). In this paper we shall consider the Monte Carlo p-value based on 10,000 simulations. We shall also identify what impact using adjusted Pearson residuals has on the correspondence plot of a two-way contingency table.

For the correspondence analysis of N using adjusted residuals the singular value decomposition of $\tilde{R}$ can be performed so that

$$\tilde{R} = \tilde{A} D_{\tilde{\lambda}} \tilde{B}^T$$

where

$$\tilde{A}^T \tilde{A} = I \quad \tilde{B}^T \tilde{B} = I. \quad (8)$$

In this case, the row and column profile coordinates can be expressed as

$$\begin{aligned} \tilde{F} &= D_I^{-1/2} \tilde{A} D_{\tilde{\lambda}} \\ \tilde{G} &= D_J^{-1/2} \tilde{B} D_{\tilde{\lambda}} \end{aligned}$$

respectively. These coordinates can be derived in an analogous manner to those defined for classical correspondence analysis (using Pearson residuals).

By considering the use of standardised residuals, the adjusted total inertia may be expressed in terms of $\tilde{F}$, $\tilde{G}$ and $D_{\tilde{\lambda}}$ by

$$\begin{aligned} \frac{X^2}{n} &= \text{trace} \left(\tilde{F}^T D_I \tilde{F} \right) \\ &= \text{trace} \left(\tilde{G}^T D_J \tilde{G} \right) \\ &= \text{trace} \left(D_{\tilde{\lambda}}^2 \right). \end{aligned} \quad (9)$$

Therefore, the properties associated with these results when using Pearson residuals also applies when adjusted residuals are considered as part of the correspondence analysis.

2.3 Profile Distances

One important feature of correspondence analysis is that the distance between the intra-variable profiles calculated from the contingency table are preserved in the correspondence plot. Here we discuss the impact on these profiles when adjusted residuals are used to perform correspondence analysis.

Suppose the distance between the i th and i' th row profile in the M^* -dimensional correspondence plot is of interest. When classical correspondence analysis is performed (so that the Pearson residuals are decomposed) the squared Euclidean distance between these profiles is

$$d_I^2(i, i') = \sum_{m=1}^{M^*} (f_{im} - f_{i'm})^2 = \sum_{j=1}^I \frac{1}{p_{\bullet j}} \left(\frac{p_{ij}}{p_{i\bullet}} - \frac{p_{i'j}}{p_{i'\bullet}} \right)^2.$$

Alternatively, this squared Euclidean distance can be expressed as

$$d_I^2(i, i') = \sum_{m=1}^{M^*} (f_{im} - f_{i'm})^2 = \sum_{j=1}^I \frac{1}{p_{\bullet j}} \left[\left(\frac{p_{ij}}{p_{i\bullet}} - p_{\bullet j} \right) - \left(\frac{p_{i'j}}{p_{i'\bullet}} - p_{\bullet j} \right) \right]^2$$

where the profiles are centred about zero. When the two categorical variables are completely independent, $f_{im} = f_{i'm} = 0$, for all i and m .

In the case where correspondence analysis is performed by decomposing the adjusted residuals, the squared Euclidean distance between the row profile coordinates $\tilde{f}_{im}$ and $\tilde{f}_{i'm}$ of a M^* -dimensional correspondence plot is

$$\begin{aligned} \tilde{d}_I^2(i, i') &= \sum_{m=1}^{M^*} (\tilde{f}_{im} - \tilde{f}_{i'm})^2 \\ &= \sum_{j=1}^I \frac{1}{p_{\bullet j}} \left[\frac{1}{\sqrt{(1 - p_{i\bullet})(1 - p_{\bullet j})}} \left(\frac{p_{ij}}{p_{i\bullet}} - p_{\bullet j} \right) - \right. \\ &\quad \left. \frac{1}{\sqrt{(1 - p_{i'\bullet})(1 - p_{\bullet j})}} \left(\frac{p_{i'j}}{p_{i'\bullet}} - p_{\bullet j} \right) \right]^2. \end{aligned}$$

The proof of this result can be verified by substituting the profile coordinate of (3) into the first line of this statement and simplifying using (8). Since the adjusted residuals are calculated to take into account the variance of the Pearson residuals by dividing them by $\sqrt{(1 - p_{i\bullet})(1 - p_{\bullet j})}$ the centred row profiles are standardised in the same manner. Therefore, the squared Euclidean distance between the row profile coordinates using the adjusted residuals reflects the weighted squared Euclidean distance between the standardised row profiles, where the i th of these profiles is

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

Similarly it can be shown that the squared Euclidean distance between the j th and j' th column profile coordinates, g_{jm} and $g_{j'm}$ respectively, can be expressed in terms of the standardised column profiles such that

$$\begin{aligned} \tilde{d}_J^2(j, j') &= \sum_{m=1}^{M^*} (\tilde{g}_{jm} - \tilde{g}_{j'm})^2 \\ &= \sum_{i=1}^J \frac{1}{p_{i\bullet}} \left[\frac{1}{\sqrt{(1 - p_{i\bullet})(1 - p_{\bullet j})}} \left(\frac{p_{ij}}{p_{\bullet j}} - p_{i\bullet} \right) - \right. \\ &\quad \left. \frac{1}{\sqrt{(1 - p_{i\bullet})(1 - p_{\bullet j'})}} \left(\frac{p_{ij'}}{p_{\bullet j'}} - p_{i\bullet} \right) \right]^2. \end{aligned}$$

$$\frac{1}{\sqrt{(1 - p_{i\bullet})(1 - p_{\bullet j'})}} \left(\frac{p_{ij'}}{p_{\bullet j'}} - p_{i\bullet} \right) \Bigg]^2.$$

The distance measures of $\tilde{d}_I^2(i, i')$ and $\tilde{d}_J^2(j, j')$ also ensures that the property of distributional equivalence holds (Lebart, Morineau and Warwick, 1984, pg 35). This states that two standardised profiles that are identical will have identical positions in the correspondence plot. Similarly, two standardised profiles having identical positions in the correspondence plot are also identical.

2.4 Example

To demonstrate the application of correspondence analysis using Pearson residuals and their adjusted counterparts, consider Table 1. The table is a cross-classification of the daily consumption of wine and the highest level of education attained for liver patients. The data are based on the findings of a 2003 survey of 826 patients suffering from liver sickness which was conducted by the Department of Clinic Medicine and Infectious Disease, Second University of Naples, Italy. The variable summarising the highest level of education that a patient completed includes three categories: *Primary*, *Secondary* and *University*. The row categories summarise the daily consumption (number of glasses) of wine. Lombardo, Beh and D'Ambra (2007) consider the non-symmetric correspondence analysis of Table 1.

Table 1

Cross-classification of 826 liver patients according to attained education and daily wine consumption

<i>Daily Wine Consumption</i>	<i>Highest Attained Education</i>		
	Primary	Secondary	University
0 - 1 Glasses	179	192	172
2 - 3 Glasses	59	87	50
4 - 10 Glasses	29	36	5
11+ Glasses	9	7	2

The Pearson chi-squared statistic of Table 1 is 25.52 so that, with a p-value of 0.0003, there is ample evidence to suggest that daily wine consumption of a liver patient is associated with their highest level of education attained.

Differences in each cell with what is expected under the hypothesis of independence can be identified by considering the quantities $\sqrt{nr_{ij}}$ that are summarised in Table 2.

Table 2 shows that, for a standard normal random variable at the 5% level of significance, there are less people with a university education in the sample who drink between 4 and 10 than what is expected under the hypothesis of independence. Despite there being a statistically significant association between the two variables (when considering the Pearson chi-squared statistic) this is the only cell to indicate any significant departure from independence. To explore further the nature of the association between *Daily Wine Consumption* and *Highest Attained Education* a classical correspondence analysis is performed.

Table 2

$\sqrt{n} \times$ Pearson residuals of Table 1

<i>Daily Wine Consumption</i>	<i>Highest Attained Education</i>		
	Primary	Secondary	University
0 - 1 Glasses	-0.18	-1.33	1.77
2 - 3 Glasses	-0.77	1.21	-0.59
4 - 10 Glasses	1.18	1.67	-3.27
11+ Glasses	1.23	-0.01	-1.34

For such an analysis, the total inertia, X^2/n , can be partitioned into the sum of λ_1^2 and λ_2^2 - refer to equation (9). Here $\lambda_1 = 0.1645$ and $\lambda_2 = 0.06197$. Since Table 1 is a 4×3 contingency table, $M^* = 2$ so that the two-dimensional correspondence plot of Figure 1 graphically depicts all of the association between the two variables.

The points “*” in Figure 1 indicate the position of the profile coordinates for each of the categories of Table 1 calculated from a classical correspondence analysis of the data. This figure shows that liver patients who drink 0-1 glasses of wine a day are associated with those in the study whose highest attained level of education is a university qualification. Similarly *2-3 Glasses* is associated with those who have a secondary qualification while liver patients who have a history of drinking at least 10 glasses of wine a day are associated with those who have attained only a primary level of education.

Given the small number of cells in Table 1, the average variance of the Pearson residuals is $(I - 1)(J - 1)/IJ = 0.5 < 1$ (Agresti, 2002, pg 81). Therefore it seems appropriate that adjusted residuals be considered for the correspondence analysis of Table 1. These residuals are summarised in Table 3 and provide an indication of how significant each cell frequency deviates from what is expected under the null hypothesis.

Table 3 shows that, compared with the values of Table 2, using adjusted

Fig. 1. Classical and adjusted correspondence plot of Table 1

Table 3
 $\sqrt{n} \times$ Adjusted residuals of Table 1

<i>Daily Wine Consumption</i>	<i>Highest Attained Education</i>		
	Primary	Secondary	University
0 - 1 Glasses	-0.38	-2.90	3.56
2 - 3 Glasses	-1.09	1.78	-0.79
4 - 10 Glasses	1.51	2.23	-4.02
11+ Glasses	1.52	-0.01	-1.59

residuals has identified a further three cells that are not consistent with what is expected under independence. These cells are those people who have attained a secondary qualification and have 0-1 and 4-10 drinks a day, and university educated liver patients who have drunk only 0-1 drinks daily.

Suppose we consider again the correspondence plot of Figure 1. The “#” points in the figure are associated with position of the row and column profile coordinates when the adjusted residuals are decomposed. Figure 1 shows that

going from analysing the Pearson residuals to their adjusted version has lead to the position of each coordinate being moved further away from the origin. By eye-balling the plot, it appears that the two categories that have the biggest relative shift in position are *0-1 Glasses* and *University*, followed then by *4-10 Glasses*. This is not surprising since the cell corresponding to the categories *0-1 Glasses* and *University* has a statistically insignificant Pearson residual (of 1.77) but a highly significant adjusted Pearson residual of 3.56 - a 100% increase.

As indicated earlier, the Pearson chi-squared statistic of Table 1 is 25.52. When using adjusted residuals as part of the analysis the adjusted chi-squared statistic is 54.42 with a Monte Carlo p-value $p < 0.0001$. Therefore, by adjusting for the variance of the Pearson residuals, there is more than twice as much association between the variables being accounted for.

The following section quantifies the shift, or “stretching”, in the position of the profiles when moving from the analysis of the Pearson residual to the analysis of the adjusted residual.

3 “Stretching” in the Plot

The example in Section 2 demonstrated that there is a stretching of the profile coordinates in a correspondence plot when going from decomposing the Pearson residuals (1) to the adjusted residuals (6). In order to determine the shift in position for each of the coordinates, consider that the standardised residuals may be expressed as

$$R = (I_r - D_r)^{1/2} \tilde{A} D_{\tilde{\lambda}} \tilde{B}^T (I_c - D_c)^{1/2}$$

$$r_{ij} = \sqrt{(1 - p_{i\bullet})(1 - p_{\bullet j})} \sum_{m=1}^{M^*} \tilde{a}_{im} \tilde{b}_{jm} \tilde{\lambda}_m.$$

Let $D_{\tilde{\lambda}} = W^{1/2} D_{\lambda} W^{1/2}$ $\tilde{\lambda}_m = w_m \lambda_m$ where W is a $M^* \times M^*$ diagonal matrix where the (m, m) th element is $w_{mm} > 1$. This is the amount by which the configuration has stretched along the m th axis of the correspondence plot. Therefore

$$R = (I_r - D_r)^{1/2} \tilde{A} \left(W^{1/2} D_{\lambda} W^{1/2} \right) \tilde{B}^T (I_c - D_c)^{1/2} \quad (10)$$

$$\begin{aligned}
r_{ij} &= \sqrt{(1 - p_{i\bullet})(1 - p_{\bullet j})} \sum_{m=1}^{M^*} \tilde{a}_{im} \tilde{b}_{jm} w_m \lambda_m \\
&= \sum_{m=1}^{M^*} \left[\tilde{a}_{im} \sqrt{w_m (1 - p_{i\bullet})} \right] \lambda_m \left[\tilde{b}_{jm} \sqrt{w_m (1 - p_{\bullet j})} \right]. \tag{11}
\end{aligned}$$

By comparing (2) with (11), a_{im} , b_{jm} and λ_m may be expressed in terms of their adjusted counterparts by

$$a_{im} = \tilde{a}_{im} \sqrt{w_m (1 - p_{i\bullet})}, \tag{12}$$

$$b_{jm} = \tilde{b}_{jm} \sqrt{w_m (1 - p_{\bullet j})}, \tag{13}$$

$$\lambda_m = \frac{\tilde{\lambda}_m}{w_m}. \tag{14}$$

These results show that b_{jm} is proportional to $\tilde{b}_{jm}$ by a factor of $\sqrt{w_m (1 - p_{\bullet j})}$. Similarly the i th value of the m th singular vector obtained from the Pearson residuals, a_{im} , is also proportional to the $\tilde{a}_{im}$.

Consideration of these results leads to a set of transition formulae for the row coordinates f_{im} and $\tilde{f}_{im}$ (and similarly g_{jm} and $\tilde{g}_{jm}$). By recalling the profile coordinates for Pearson residuals, given by (3) and (4), and using (12) - (14),

$$\begin{aligned}
f_{im} &= \tilde{a}_{im} \tilde{\lambda}_m \sqrt{\frac{1 - p_{i\bullet}}{w_m}} = \tilde{f}_{im} \sqrt{\frac{p_{i\bullet} (1 - p_{i\bullet})}{w_m}}, \\
g_{jm} &= \tilde{b}_{jm} \tilde{\lambda}_m \sqrt{\frac{1 - p_{\bullet j}}{w_m}} = \tilde{g}_{jm} \sqrt{\frac{p_{\bullet j} (1 - p_{\bullet j})}{w_m}}.
\end{aligned}$$

Since $w_m > 1$, $\lambda_m < \tilde{\lambda}_m$, $f_{im} < \tilde{f}_{im}$ and $g_{jm} < \tilde{g}_{jm}$. Hence $\tilde{r}_{ij} > r_{ij}$ and

$$\frac{\tilde{X}^2}{n} = \sum_{m=1}^{M^*} \tilde{\lambda}_m^2 > \sum_{m=1}^{M^*} \left(\frac{\tilde{\lambda}_m}{w_m} \right)^2 = \sum_{m=1}^{M^*} \lambda_m^2 = \frac{X^2}{n}.$$

The amount by which the i th row profile coordinate is stretched along the m th axis can be determined from the ratio $\tilde{f}_{im}/f_{im}$ or equivalently $\psi_{im} = \sqrt{w_m/[p_{i\bullet}(1 - p_{i\bullet})]}$. It is also easy to verify this result if one considers the squared Euclidean distance between the row coordinates $\tilde{f}_{im}$ and f_{im} ,

$$d_i^2(\tilde{f}_{im}, f_{im}) = \sum_{m=1}^{M^*} (\tilde{f}_{im} - f_{im})^2 = \sum_{m=1}^{M^*} (\psi_{im} - 1)^2 f_{im}^2.$$

Similarly, the squared Euclidean distance between the column coordinates $\tilde{g}_{jm}$ and g_{jm} is

$$d_j^2(\tilde{g}_{jm}, g_{jm}) = \sum_{m=1}^{M^*} (\tilde{g}_{jm} - g_{jm})^2 = \sum_{m=1}^{M^*} (\varphi_{jm} - 1)^2 f_{im}^2$$

where $\varphi_{jm} = \sqrt{w_m / [p_{\bullet j} (1 - p_{\bullet j})]}$ is the amount by which the j th column profile coordinate is stretched along the m th dimension of the correspondence plot.

These distance measures indicate that generally the further away a point is from the origin of a classical correspondence plot, the more stretching that is done on that point when using adjusted Pearson residuals. Of course such a scaling will depend on the marginal proportion of that point and the weight w_m associated with the m th singular value.

For a correspondence plot consisting of $M < M^*$ dimensions, the amount by which the i th row profile is stretched can be calculated by

$$\psi_i(M) = \frac{\sum_{m=1}^M p_{i\bullet} \tilde{f}_{im}^2}{\sum_{m=1}^M p_{i\bullet} f_{im}^2} = \frac{\sum_{m=1}^M \tilde{f}_{im}^2}{\sum_{m=1}^M f_{im}^2} = \frac{1}{p_{i\bullet} (1 - p_{i\bullet})} \frac{\sum_{m=1}^M w_m f_{im}^2}{\sum_{m=1}^M f_{im}^2}.$$

Similarly, the amount by which a correspondence analysis of the adjusted residuals will stretch the j th column profile coordinate in an M dimensional correspondence plot is

$$\varphi_j(M) = \frac{1}{p_{\bullet j} (1 - p_{\bullet j})} \frac{\sum_{m=1}^M w_m g_{jm}^2}{\sum_{m=1}^M g_{jm}^2}.$$

The stretching of the configuration of points in an M dimensional correspondence plot may be determined by calculating

$$\vartheta(M) = \sqrt{\frac{\sum_{m=1}^M (w_m \lambda_m)^2}{\sum_{m=1}^M \lambda_m^2}}$$

so that $\vartheta(1) = w_1$ and $\vartheta(M^*) = \sqrt{\tilde{X}^2 / X^2}$.

3.1 Table 1 Revisited

Consider again Table 1. Since $X^2 = 25.52$ and $\tilde{X}^2 = 54.43$, $\vartheta(2) = 1.46$. Therefore by using adjusted residuals, the overall configuration of profile co-

ordinates in the classical correspondence plot in Figure 1 has stretched by 46%. For each of the axes, $w_1 = 1.468$ and $w_m = 1.404$, so that the coordinates in the correspondence plot have been stretched further along the first axis (by 47%) than the second axis. Thus, the additional variation accounted for by considering the adjusted residuals, rather than the Pearson residuals, is reflected more by the relatively large shift along the first axis than the second axis.

The shift in position for each of the categories in Table 1 can also be measured. Tables 4 and 5 summarise the shift (as measured by ψ_{im} and φ_{jm}) of each of the row categories and column categories, respectively. The tables also provide an indication of the impact of analysing adjusted residuals in Figure 1, by showing the amount by which the configuration of points in the plot has shifted, as summarised by $\psi_i(2)$ and $\varphi_j(2)$.

Table 4

Stretching of the row profile coordinates of Table 1

Row Category	ψ_{i1}	ψ_{i2}	$\psi_i(2)$
0 - 1 Glasses	2.11	1.33	1.46
2 - 3 Glasses	1.65	1.33	1.42
4 - 10 Glasses	1.24	3.89	1.47
11+ Glasses	1.09	1.40	1.45

Table 4 shows that of the four row categories, *4-10 Glasses* has experienced the greatest shift in position. That category's position in Figure 1 stretched by a further 47%. For each of the axes of the plot, its point shifted by another 25% along the first axis and nearly four times along the second axis. Similarly, the position of the point for the row category *0-1 Glasses* was stretched by a further 46% while for *2-3 Glasses* this shift was 42%. Row category *0-1 Glasses* stretched further along the first axis than all any of the other six categories of Table 1, more than doubling its original position when classical correspondence analysis is employed.

Table 5

Stretching of the column profile coordinates of Table 1

Column Category	φ_{j1}	φ_{j2}	$\varphi_j(2)$
Primary	1.11	1.44	1.43
Secondary	1.74	1.30	1.45
University	1.41	2.83	1.47

Table 5 shows that the point associated with the column category *University* was stretched further than the position of the remaining two education levels - by a further 47%. This column category also stretched further along the second axis of Figure 1, while the *Secondary* category was stretched the most along the first axis (by a further 74%).

Importantly, these results provide a graphical insight into those cells, and categories, that deviate from the hypothesis of independence. For example, Section 2.3 showed that by using adjusted residuals, university qualified liver patients who drink 0-1 glasses of wine per day deviated significantly from what was expected under independence. However by analysing Pearson residuals this conclusion could not be made. By observing the quantities summarised in Table 5 it can be seen that, of the three educational levels, *University* experienced the biggest shift. Similarly, Table 4 shows that 0 – 1 and 4 – 10 had the biggest shift in position, certainly - along the first and second axis respectively.

4 Discussion

This paper has described the application of simple correspondence analysis using adjusted residuals. The asymptotic properties of these residuals allow for one to take into account the variability in the categories better than what the Pearson residuals can do. The variability is reflected in the shifting, or stretching, of the profile coordinates in the correspondence plot. The simple procedure outlined in this paper will always lead to a stretching effect when comparing the plots obtained from decomposing the Pearson residuals and adjusted residuals. Whether this stretching is statistically significant for a particular profile coordinate can be made by determining if $\phi_{im} = 0$ for the i th row category, or $\varphi_{jm} = 0$ for the j th column category. However, this issue will be left for further investigation. Although, by observing plots of the type of Figure 1, correspondence analysis using adjusted residuals can be used to easily determine those cells and profiles that exhibit variability.

Interestingly, Haberman (1973, pg 206) comments that with the use of these adjusted residuals “one may apply various graphical and analytical techniques to detect deviations from the proposed model”. The cobweb diagrams of Upton (2000) is one such graphical technique. The correspondence analysis described in this paper is another. Adjusted residuals could be further incorporated into the suite of simple correspondence analysis procedures. For example, the ordinal correspondence analysis technique of Beh (1997) could be adapted such that bivariate moment decomposition is applied to the residuals (6) as an alternative to singular value decomposition. Non-symmetric correspondence analysis of two cross-classified nominal categorical variables (D’Ambra and

Lauro, 1989) or ordinal variables (Lombardo, Beh and D'Ambra, 2007) could also be performed keeping in mind a variation of the adjusted residuals. Computationally, the SPLUS code of Beh (2004b, 2005) - which can also be incorporated into R - can be modified to incorporate the decomposition of adjusted residuals when using them to perform correspondence analysis.

For any correspondence analysis using Pearson residuals, the correspondence plot remains dependent on inferential tests of association - the plot itself is not enough to indicate any significant association between two categorical variables. This is also true for the analysis of adjusted residuals. Therefore future work integrating the link between the weighted Pearson chi-squared statistic, (7), and correspondence analysis will help advance the technique so that it will adequately describe the variation, and association, between profiles of two-way contingency tables and the cells that occupy them.

References

- [1] Agresti, A., 2002. Categorical Data Analysis (Second Edition). Wiley, New Jersey.
- [2] Aitchison, J., Greenacre, M. 2002. Biplots in compositional data. *Applied Statistics* 51, 375-392.
- [3] Beh, E.J., 1997. Simple correspondence analysis of ordinal cross-classifications using orthogonal polynomials. *Biometrical Journal* 39, 589-613.
- [4] Beh, E.J., 2004a. Simple correspondence analysis: A bibliographic review. *International Statistical Review* 72, 257-284.
- [5] Beh, E.J., 2004b. S-PLUS code for ordinal correspondence analysis. *Computational Statistics* 19, 593-512.
- [6] Beh, E.J., 2005. S-PLUS code for simple and multiple correspondence analysis. *Computational Statistics* 20, 415-438.
- [7] Benzecri, J.P., 1992. Correspondence Analysis Handbook. Marcel Dekker, New York.
- [8] D'Ambra, L., Lauro, N., 1989. Non symmetrical analysis of three-way contingency tables. In: Coppi, R. Bolasco, S. (Eds.), *Multiway Data Analysis*, 301-315.
- [9] Davies, R.B., 1980. Algorithm AS 155: The distribution of a linear combination of χ^2 random variables. *Applied Statistics* 29, 323-333.
- [10] Davis, C.S., 1982. The distribution of a linear combination of chi-squared variables. *Biometrics* 38, 279.

- [11] Fuchs, C., Kenett, R., 1980. A test for detecting outlying cells in the multinomial distribution and two-way contingency tables. *Journal of the American Statistical Association* 75, 395-398.
- [12] Golden, R.M., 2000. Examples of complex probability models in the social and behavioural sciences (Presentation). Retrieved April 3, 2007 from www.utdallas.edu/~golden/ANNTALKS/Feb21talk.
- [13] Goodman, L.A., 1986. Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. *International Statistical Review* 54, 243-309.
- [14] Greenacre, M.J., 1984. *Theory and Application of Correspondence Analysis*. Academic Press, London.
- [15] Greenacre, M.J., 1988. Correspondence analysis of multivariate categorical data by weighted least-squares. *Biometrika* 75, 457-467.
- [16] Greenacre, M.J., Blasius, J., 2006. *Multiple Correspondence Analysis and Related Methods*. Chapman & Hall, London.
- [17] Haberman, S.J., 1973. The analysis of residuals in cross-classified tables. *Biometrics* 29, 205-220.
- [18] Lang, J.B., Agresti, A., 1994. Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association* 89, 625-632.
- [19] Le Roux, B., Rouanet, H., 2005. *Geometric Data Analysis: From Correspondence Analysis to Structured Data Analysis*. Springer, London.
- [20] Lebart, L., Morineau, A., Warwick, K., 1984. *Multivariate Descriptive Statistical Analysis*. Wiley, New York.
- [21] Lombardo, R., Beh, E.J., D'Ambra, L., 2007. Non-symmetric correspondence analysis with ordinal variables using orthogonal polynomials. *Computational Statistics and Data Analysis*, 52, 566-577.
- [22] Murtagh, F., 2005. *Correspondence Analysis and Data Coding with Java and R*. Chapman & Hall, Florida.
- [23] Rayner, J.C.W., Best, D.J., 2001. *A Contingency Table Approach to Nonparametric Testing*. Chapman & Hall, London.
- [24] Sheil, J., O'Muircheartaigh, I., 1977. Algorithm AS 106: The distribution of non-negative quadratic forms in normal variables. *Applied Statistics* 26, 92-98.
- [25] Upton, G.J.G., 2000. Cobweb diagrams for multiway contingency tables. *The Statistician* 49, 79-85.
- [26] van der Heijden, P.G.M., de Falguerolles, A., de Leeuw, J., 1989. A combined approach to contingency table analysis using correspondence analysis and log-linear analysis. *Applied Statistics* 38, 249-292.

- [27] Weller, S.C., Romney, A.K., 1990. Metric Scaling: Correspondence Analysis. Sage University Paper Series of Quantification Applications in Social Sciences, 07-075, Newbury Park, CA.
- [28] Wood, C.L., Cady, F.B., 1981. Intersite transfer of estimated response surfaces. Biometrics 37, 1-10.