International Statistical Review (2004), **72**, 2, 257–284, Printed in The Netherlands

Simple Correspondence Analysis: A Bibliographic Review

Eric J. Beh

School of Quantitative Methods and Mathematical Sciences, University of Western Sydney, Australia

Summary

Over the past few decades correspondence analysis has gained an international reputation as a powerful statistical tool for the graphical analysis of contingency tables. This popularity stems from its development and application in many European countries, especially France, and its use has spread to English speaking nations such as the United States and the United Kingdom. Its growing popularity amongst statistical practitioners, and more recently those disciplines where the role of statistics is less dominant, demonstrates the importance of the continuing research and development of the methodology.

The aim of this paper is to highlight the theoretical, practical and computational issues of simple correspondence analysis and discuss its relationship with recent advances that can be used to graphically display the association in two-way categorical data.

Key words: Computing issues; Correspondence analysis family; Graphical displays; Inertia; Ordered categories; Orthogonal polynomials; Pearson ratio; Profile; Reconstitution model; Singular value decomposition; Transition formula; Two-way contingency table.

1 Introduction

The analysis of the contingency table, is a very important component of multivariate statistics with many different types of analysis dedicated solely to this type of data set. Fienberg (1982) points out that the term *contingency* seems to have originated with Karl Pearson (1904) who used it to describe the measure of the deviation from complete independence between the rows and columns of such a data structure. More recently, the term has come to refer to the counts and the marginal frequencies in the contingency table. As a result, a contingency table contains information which is of a discrete or categorical nature.

The development of techniques to handle problems involving contingency tables are due most importantly to Karl Pearson, G. Udny Yule and R.A. Fisher (see Goodman, 1996). One of the most influential techniques developed to measure the association between two categorical variables is the Pearson chi-squared statistic. Pearson (1900) developed the ground work for the chi-squared statistic which is used to compare the observed counts with what is expected under the hypothesis of independence between the two variables.

One of the first examples used to investigate the application of measuring association in contingency tables was that of Fisher (1940) and is reproduced here as Table 1. It is the cross-classification of 5387 children from Caithness, Scotland according to their hair and eye colour. Fisher was interested in determining how the two variables were associated. Goodman (1981) also considered this example in his investigation of association for contingency tables where the variables consist of ordered responses.

Table 1

Two-way contingency table classifying 5387 children in Caithness, Scotland, according to hair colour and eye colour.

Eye Colour	Hair Colour				
	Fair	Red	Medium	Dark	Black
BLUE	326	38	241	110	3
LIGHT	688	116	584	188	4
MEDIUM	343	84	909	412	26
DARK	98	48	403	681	85

A test of the departure from independence between hair colour and eye colour produces a Pearson chi-squared statistic of 1240.039 which is highly significant.

This simple example demonstrates that, while the Pearson chi-squared statistic has long been used to determine the level of association between two variables, it does not divulge how the association is constructed, nor does the statistic allow for an investigation of similar, or different, categories. Many methodologies have been discussed in the literature which do permit an investigation of these issues, however it is not the purpose of this paper to review these. Instead we will focus our discussion on the development and application of correspondence analysis to two-way contingency tables; such an analysis is commonly referred to as simple or, more recently, as classical, correspondence analysis. The term "simple" is not meant to reflect the ease of execution, or interpretation, of the analysis. Instead it refers to its application to the most basic, or simple, data set—a two-way contingency table, as apposed to "multiple" correspondence analysis which applies to more than two categorical variables. The term "classical" has also been used to describe the original graphical methodology developed since there are adjustments to the classical approach that can be implemented.

Correspondence analysis is a technique that represents graphically the row and column categories and allows for a comparison of their "correspondences", or associations, at a category level. For example, Figure 1 shows a two-dimensional space where the association between the row and column categories of Table 1 can be visualised. It shows that, in general, the fair haired children in Caithness, Scotland, tend to have blue or light coloured eyes, while dark haired children tend to be those with dark coloured eyes. This figure is referred to as a correspondence plot, and is an important component of the output generated from the classical correspondence analysis of Table 1. Also from the output, we find that the first dimension visualises 86.56% of the association between the row and column categories, while the second axis visualises 13.07%. Thus Figure 1 visually shows 99.63% of the association that exists between the hair colour and eye colour of the children classified in Caithness.

In parts of Section 3 we will discuss particular points relevant to the construction of Figure 1 and the analysis of Table 1.

Part of this review involves discussing adaptations of correspondence analysis made over the years for its application to complex cross-classifications such as ordinal data, ranked data, cohort data etc. It should be evident from this paper that correspondence analysis is applicable in a diverse range of situations. For example, correspondence analysis has been beneficial in the areas of social science, engineering, health science, medicine, archeology, ecology, software development and market research.

The paper consists of eight further Sections.

Section 2 discusses some of the important developments of correspondence analysis made between the early 20th century to the present. An excellent review of the history of the methodology, from the French perspective, can be found in van Meter, Schiltz, Cibois & Mounier (1994).

Section 3 briefly describes some technical aspects of correspondence analysis, including methods of decomposing the Pearson ratios of two-way contingency tables, the definition of profile co-ordinates and its relationship with the Pearson chi-squared statistic. While multivariate statistics

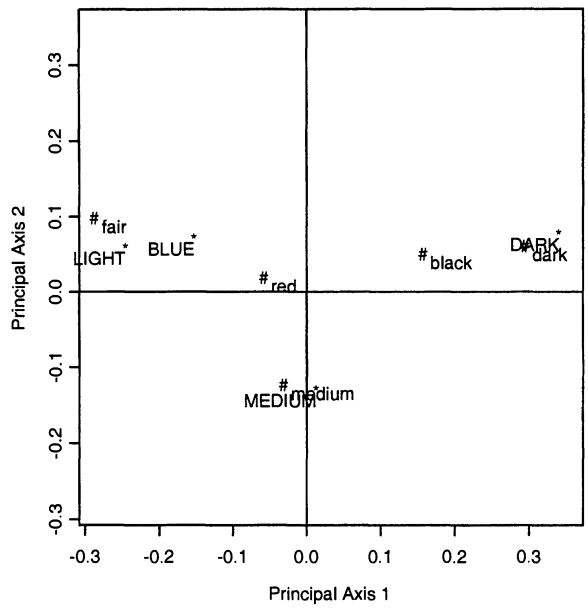

Figure 1. *Correspondence plot of Table 1.*

is generally described using matrix notation, we steer away from it in this paper to help simplify the discussion.

Section 4 describes some ways to model two-way contingency tables using correspondence analysis and descriptions of the members of the correspondence analysis family are made in Section 5.

Section 6 outlines many of the text books that discuss in detail the advancement of research into correspondence analysis. Accompanying these texts is a very general review of applications made using the analysis. There are many applications of correspondence analysis in nearly all disciplines and those discussed in this paper are not intended to fully summarise such applications. Instead, an aim of this paper is to demonstrate the method's diversity in applied situations.

In Section 7, we discuss the development of various computing issues that have evolved for the application correspondence analysis. We outline some of the early programs used to make such an analysis, and include a discussion of some of the commercially available software that performs correspondence analysis.

Some other issues to do with correspondence analysis are discussed in Section 8 and some final remarks are made in Section 9.

2 Development of Correspondence Analysis

The theoretical issues associated with correspondence analysis date back to the early 20th century and its foundation is algebraic rather than geometric.

The foundation of the technique was nearly laid with the 1904 and 1906 papers of Karl Pearson, as argued by de Leeuw (1983), when he developed the correlation coefficient of a two-way contingency table using linear regression. As Pearson (1906) states:

> The conception of linear regression line as giving this arrangement with the maximum

> degree of correlation appears of considerable philosophical interest. It amounts primarily to much the same thing as saying that if we have a fine classification, we shall get the maximum correlation by arranging the arrays so that the means of the arrays fall as closely as possible on a line.

de Leeuw (1983) then notes:

> this is exactly what correspondence analysis does. Pearson just ... was not familiar with singular value decomposition, although this had been discovered much earlier by Beltrami, Sylvester and Jordan.

See also Gifi (1990).

However, the original algebraic derivation of correspondence analysis is often accredited to Hirschfeld (1935) who developed a formulation of the correlation between the rows and columns of a two-way contingency table.

Others to contribute to such developments include Richardson & Kuder (1933) and Horst (1935). In fact, Horst, who discussed his findings in early 1934 before the Psychology Section of the Ohio Academy of Science, was the first to coin the term "method of reciprocal averaging", an alternative derivation of correspondence analysis.

The simplest derivation of correspondence analysis was made by the biometrician R.A. Fisher in 1940 when he considered data relating to hair and eye colour in a sample of children from Caithness, Scotland, see Table 1.

While the original development of the problem aimed at dealing with two-way contingency tables, a more complex approach dealing with multi-way contingency tables was not discussed until 1941 when psychometrician Louis Guttman discussed his method, called dual (or optimal) scaling, which is now referred to as the foundation of multiple correspondence analysis. Later applications of multiple correspondence analysis were considered using the Burt matrix of Burt (1950). In fact Guttman (1953) writes of Burt:

> ... it is gratifying to see how Professor Burt has independently arrived at much the same formulation. This convergence of thinking lends credence to the similarity of the approach.

Fisher and Guttman presented essentially the same theory in the biometric and psychometric literature. Thus biometricians regard Fisher as the inventor of correspondence analysis, while psychometricians regard it as being Guttman.

In the 1940's and 1950's further advances were made to the mathematical development of correspondence, particularly in the field of psychometrics, by Guttman and his researchers. In Japan, a group of data analysts led by Chikio Hayashi also further developed Guttman's ideas, which they referred to as the quantification of qualitative data.

The 1960's saw the biggest leap in the development of correspondence analysis when it was given a geometric form by linguist Jean-Paul Benzécri and his team of researchers at the Mathematical Statistics Laboratory, Faculty of Science in Paris, France. This work culminated in two volumes on data analysis; Benzécri (1973b, 1973a). As a result the method of *l'analyse des correspondances*, as coined by Benzécri, is very popular in France not just among statisticians, but among researchers from most disciplines in the country. The popularity of correspondence analysis in France resulted in a journal dedicated to the development and application of the technique as well as methods of classification, *Cahiers de l'Analyse des Données*, founded by Benzécri.

In 1974, this new method was widely exposed to English speaking researchers with the popular paper by M.O. Hill (Hill, 1974). He was the first to coin the method's name correspondence analysis which is the English translation of Benzécri's l'analyse des correspondances. Hill showed that the method is mathematically similar to already popular methods of data analysis such as principal

components analysis, canonical correlation analysis and reciprocal averaging (which he discussed the previous year).

Since Hill's (1974) contribution, the theory of correspondence analysis, especially its application to multivariate data, has been reinvented many times and given different names, such as homogeneity analysis (Gifi, 1990) and dual scaling (Nishisato, 1980, 1994).

3 Theoretical Development

3.1 Notation

Consider an $I \times J$ two-way contingency table, N, where the (i, j)-th cell entry is given by n_{ij} for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$. Let the grand total of N be n and the correspondence matrix, or matrix of 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 i-th row marginal proportion by $p_{i\bullet} = \sum_{j=1}^{J} p_{ij}$ and define the j-th column marginal proportion as $p_{\bullet j} = \sum_{i=1}^{I} p_{ij}$. The row marginal proportions are called row masses and the column probability marginals are called column masses.

3.2 Pearson Ratio's

The aim of correspondence analysis, like many multivariate data analytic techniques, is to determine scores which describe how similar or different responses from two or more variables are. For a two-way contingency table, upon which our discussion is focused, the strength of association between the rows scores and column scores should also be measured.

For our discussion of simple correspondence analysis, first consider the model of complete independence between the rows and columns:

$$p_{ij} = p_{i\bullet} p_{\bullet j}. \tag{1}$$

Of course, complete independence will hardly ever be satisfied, and so a multiplicative measure of the departure from the model of complete independence can be considered such that

$$p_{ij} = \alpha_{ij} p_{i\bullet} p_{\bullet j}. \tag{2}$$

For the model of complete independence given by (1), $\alpha_{ij} = 1$ for all $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$. As complete independence will seldom be observed, one can determine which elements $\alpha_{ij} \neq 1$. These elements can easily be observed by calculating

$$\alpha_{ij} = \frac{p_{ij}}{p_{i\bullet} p_{\bullet j}} \tag{3}$$

which Goodman (1996) refers to as Pearson ratio's.

By considering the Pearson ratio's, the Pearson chi-squared statistic can be expressed by

$$X^2 = n \sum_{i=1}^{I} \sum_{j=1}^{J} p_{i\bullet} p_{\bullet j} \left(\alpha_{ij} - 1\right)^2 \tag{4}$$

and has a Pearson chi-squared distribution with $(I - 1)(J - 1)$ degrees of freedom; $\chi^2_{(I-1)(J-1)}$. Therefore, a small Pearson chi-squared statistic which is consistent with the hypothesis of independence (depending on the degrees of freedom), will be achieved when each $\alpha_{ij} \approx 1$.

A property of the Pearson chi-squared statistic is that as n increases, so too does the statistic. This

can hinder tests of association in contingency tables. To overcome this problem simple correspondence analysis considers X^2/n—referred to as the total inertia of the contingency table—to describe the level of association, or dependence, between two categorical variables.

By decomposing the total inertia the researcher can identify important sources of information that help describe this association. Using different decompositions will yield different interpretations of the association, and lead to different graphical outputs. The most common type of decomposition used, with a few exceptions, in correspondence analysis is singular value decomposition (SVD). The next subsection describes the use of SVD to perform simple correspondence analysis. Other types of decompositions can be used, and two others are described at the end of this section.

3.3 Singular Value Decomposition

Classically, simple correspondence analysis is conducted by performing a singular value decomposition (SVD) on the Pearson ratio's. The method of SVD, also referred to as the "Eckart–Young" decomposition, is the most common tool used to decompose the Pearson ratio's. For the application to the analysis of contingency tables, Eckart & Young (1936) conjectured that the Pearson ratio may be decomposed into components by

$$\alpha_{ij} = \sum_{m=0}^{M^*} a_{im}\lambda_m b_{jm} \tag{5}$$

where $M^* = \max(I, J) - 1$ is the maximum number of dimensions required to graphically depict the association between the row and column responses. For example, for Table 1, only $\max(4, 6) - 1 = 3$ dimensions are required to graphically depict all of the association between the hair and eye colour of the children classified in Caithness. However, for a simple interpretation of this association, generally only the first two dimensions are used to construct such a graphical summary. The result, (5), was formally proven for any rectangular matrix by Johnson (1963).

Consider the RHS of equation (5). The vector $\underset{\sim}{\mathbf{a}}_m = (a_{1m}, a_{2m}, \ldots, a_{Im})$ is the m-th row singular vector and is associated with the I row categories. Similarly, the vector $\underset{\sim}{\mathbf{b}}_m = (b_{1m}, b_{2m}, \ldots, b_{Jm})$ is the m-th column singular vector and is associated with the J column categories. The elements of vector $\underset{\sim}{\lambda} = (\lambda_0, \lambda_1, \ldots, \lambda_{M^*})$ are real and positive and are the first M^* singular values and are arranged in descending order so that

$$\lambda_0 = 1 \geq \lambda_1 \geq \ldots \geq \lambda_{M^*} \geq 0. \tag{6}$$

These singular values can also be calculated by

$$\lambda_m = \sum_{i=1}^{I}\sum_{j=1}^{J} a_{im} b_{jm} p_{ij} \tag{7}$$

while the singular vectors have the property

$$\sum_{i=1}^{I} p_{i\bullet} a_{im} a_{im'} = \begin{cases} 1 & m = m' \\ 0 & m \neq m' \end{cases} \qquad \sum_{j=1}^{J} p_{\bullet j} b_{jm} b_{jm'} = \begin{cases} 1 & m = m' \\ 0 & m \neq m' \end{cases}. \tag{8}$$

To remove the trivial values of $\lambda_0 = 1$, $a_{i0} = 1$ and $b_{j0} = 1$, consider again equation (5). It becomes

$$\alpha_{ij} = 1 + \sum_{m=1}^{M^*} a_{im}\lambda_m b_{jm}. \tag{9}$$

By considering this expression, the Pearson chi-squared statistic can be expressed as follows

$$\begin{aligned} X^2 &= n\sum_{i=1}^{I}\sum_{j=1}^{J} p_{i\bullet}p_{\bullet j}\left(\alpha_{ij}-1\right)^2 \\ &= n\sum_{i=1}^{I}\sum_{j=1}^{J} p_{i\bullet}p_{\bullet j}\left(\sum_{m=1}^{M^*} a_{im}\lambda_m b_{jm}\right)^2 \\ &= n\sum_{m=1}^{M^*}\lambda_m^2\left(\sum_{i=1}^{I} p_{i\bullet}a_{im}^2\right)^2\left(\sum_{j=1}^{J} p_{\bullet j}b_{jm}^2\right)^2. \end{aligned}$$

By using the orthogonality properties of a_{im} and b_{jm}, the total inertia can be written in terms of the singular values such that

$$\frac{X^2}{n} = \sum_{m=1}^{M^*}\lambda_m^2. \tag{10}$$

For example, $\lambda_1^2 = 0.1992$, $\lambda_2^2 = 0.0301$ and $\lambda_3^2 = 0.0009$ so that $X^2/n = 0.2302$ for Table 1. So the first axis explains $0.1992/0.2302 = 0.8656$ of the total variation that exists in the table, while the second axis explains $0.0301/0.2302 = 0.1307$ of this variation. Thus Figure 1 accounts for 99.63% of the total variation in Table 1.

Hence, the total variation in the contingency table (or the Pearson chi-squared statistic) can be partitioned into M^* components, which are referred to as the principal inertia values. Each principal inertia can be partitioned further into sub-components to identify how a particular row or column category contributes to the principal axis.

The first principal axis, with an inertia value of λ_1^2 is the axis that describes most of the variation. Generally, the m-th principal axis is the m-th most important axis and a correspondence plot containing the first two axes will be more descriptive than if any other axes were included.

3.4 Co-ordinate Systems

3.4.1 Standard profile co-ordinates

In order to visualise the associations between row categories or column categories, the set of singular vectors, $\{a_{im}\}$ and $\{b_{jm}\}$ for $i = 1, 2, \ldots, I$ and $j = 1, 2, \ldots, J$ may be plotted as coordinates onto the m-th dimension of a correspondence plot. For such a plot each axis is referred to as a principal axis. For example, the first axis is called the first principal axis, while the second axis is called the second principal axis. However these vectors in such a plotting system do not take into consideration the strength of the relationship between the rows and columns along each axis. In fact the axes are equally weighted as can be seen from (8). Therefore, these axes have associated with them unit inertias and Greenacre (1984, p.93) refers to the singular vectors as a system of co-ordinates as standard co-ordinates.

3.4.2 Principal profile co-ordinates

Instead of defining the row and column co-ordinates using a_{im} and b_{jm}, let the row and column profile co-ordinates be defined by

$$f_{im} = a_{im}\lambda_m \tag{11}$$

$$g_{jm} = b_{jm}\lambda_m \tag{12}$$

respectively. Then (8) becomes

$$\sum_{i=1}^{I} p_{i\bullet} f_{im} f_{im'} = \begin{cases} \lambda_m^2 & m = m' \\ 0 & m \neq m' \end{cases} \qquad \sum_{j=1}^{J} p_{\bullet j} g_{jm} g_{jm'} = \begin{cases} \lambda_m^2 & m = m' \\ 0 & m \neq m' \end{cases}. \tag{13}$$

The system of co-ordinates defined using (11) and (12) involve the measure of association λ_m and are plotted in Figure 1 for $m = 1, 2$. Instead of each axis having unit inertias, the m-th principal axis has an inertia value equal to λ_m^2.

Note, from (10) and (13) that

$$\frac{X^2}{n} = \sum_{i=1}^{I} \sum_{m=1}^{M^*} p_{i\bullet} f_{im}^2. \tag{14}$$

Similarly

$$\frac{X^2}{n} = \sum_{j=1}^{J} \sum_{m=1}^{M^*} p_{\bullet j} g_{jm}^2. \tag{15}$$

Equations (14) and (15) show that profile co-ordinates close to the origin do not contribute much to the variation of the data since they only make a small contribution to the total inertia. Profile co-ordinates far from the origin do make such a contribution.

An alternative expression for the row profile co-ordinates obtained by multiplying (3) by $p_{\bullet j} b_{jm}$ and using the column orthogonality property of (8) is

$$f_{im} = \sum_{j=1}^{J} \frac{p_{ij}}{p_{i\bullet}} b_{jm}. \tag{16}$$

Similarly, the column profile co-ordinates can be alternatively expressed by

$$g_{jm} = \sum_{i=1}^{I} \frac{p_{ij}}{p_{\bullet j}} a_{im}. \tag{17}$$

Equation (16) is the weighted sum of the i-th row profile, while (17) is the weighted sum of the j-th column profile. These equations also show the link between the profile co-ordinates and standard co-ordinates.

3.4.3 Goodman's profile co-ordinates

Consider again the Pearson ratios. Then

$$\alpha_{ij} - 1 = \sum_{m=1}^{M^*} a_{im}\lambda_m b_{jm} = \sum_{m=1}^{M^*} f_{im} b_{jm} = \sum_{m=1}^{M^*} a_{im} g_{jm}. \tag{18}$$

This shows, and as was noted before, that profile co-ordinates near the origin contribute to the hypothesis of independence, while those far from the origin do not. Equation (18) provides an adequate reason if one wants to include on the same display both row and column profile co-ordinates.

This suggests that for the comparison of a row profile co-ordinate with a column profile co-ordinate, instead of using the row profile co-ordinates as previously defined, there is some advantage in using a re-scaled version.

Rather than using the co-ordinates defined above, Goodman (1986) suggested using co-ordinates of the form

$$\tilde{f}_{im} = \lambda_m^{\gamma} a_{im} \tag{19}$$

$$\tilde{g}_{jm} = \lambda_m^{\delta} b_{jm} \tag{20}$$

with $\delta + \gamma = 1$. These co-ordinates were also described in Aitchison & Greenacre (2002).

Different values of γ (and therefore δ) will produce different co-ordinates. For example, Gabriel (1971, p.458) describes three sets of co-ordinates used to construct a biplot that coincide with $\gamma = 1/2$, $\gamma = 1$ and $\gamma = 0$ so that $\delta = 1/2$, $\delta = 0$ and $\delta = 1$ respectively. The assignment of these values have been described as "symmetric", "row isometric" and "column isometric" factorisations respectively (Lombardo, Carlier & D'Ambra, 1996). Gabriel & Odoroff (1990) also describes the column metric preserving (CMP) biplot co-ordinates (similar to those for correspondence analysis).

The relationship between Goodman's co-ordinates and those of simple correspondence analysis is evident when $\gamma = 1$ and $\delta = 0$. In this case, the row profile co-ordinates are those defined by (11), while the columns are plotted using their standard co-ordinates. Similarly, when $\gamma = 0$ and $\delta = 1$ the column profile co-ordinates are just those defined by (12), while the rows are plotted using their standard co-ordinates.

Since the classical correspondence analysis approach focuses on partitioning the total inertia into singular values, the profile co-ordinates of (11) and (12) obtain similar results. For example,

$$\sum_{i=1}^{I} p_{i\bullet} \tilde{f}_{im}^2 = \sum_{i=1}^{I} p_{i\bullet} \frac{f_{im}^2}{\lambda_m^{2\delta}} = \lambda_m^{2-2\gamma} = \lambda_m^{2\delta} \tag{21}$$

$$\sum_{j=1}^{J} p_{\bullet j} \tilde{g}_{jm}^2 = \sum_{j=1}^{J} p_{\bullet j} \frac{g_{jm}^2}{\lambda_m^{2\gamma}} = \lambda_m^{2-2\delta} = \lambda_m^{2\gamma}. \tag{22}$$

Therefore, if γ and δ are chosen so that $\gamma = \delta = 1/2$, then

$$\sum_{i=1}^{I} p_{i\bullet} \tilde{f}_{im}^2 = \sum_{j=1}^{J} p_{\bullet j} \tilde{g}_{jm}^2 = \lambda_m$$

while the relationship between the total inertia and the profile co-ordinates of (19) and (20) is

$$\frac{X^2}{n} = \sum_{m=1}^{M^*} \lambda_m.$$

A simple generalisation of some graphical procedures used for the analysis of categorical data can be found in Beh (2003b) and includes the Goodman co-ordinates as a special case.

3.5 *Transition Formulae*

The transition formulae (Hill, 1973), or barycentric formulae (Benzécri, 1992, p.111), are equations for obtaining profile co-ordinates for one variable from the co-ordinates of the other variable.

Suppose we consider the application of classical correspondence analysis. Using (16) and (12), equation (11) can be alternatively expressed by

$$f_{im}\lambda_m = \sum_{j=1}^{J} \frac{p_{ij}}{p_{i\bullet}} b_{jm}\lambda_m = \sum_{j=1}^{J} \frac{p_{ij}}{p_{i\bullet}} g_{jm}.$$

Therefore, we can obtain the row profile co-ordinates when the column profile co-ordinates are known by

$$f_{im} = \frac{1}{\lambda_m} \sum_{j=1}^{J} \frac{p_{ij}}{p_{i\bullet}} g_{jm}. \tag{23}$$

Similarly, we can obtain the column profile co-ordinates when the row profile co-ordinates are known by

$$g_{jm} = \frac{1}{\lambda_m} \sum_{i=1}^{I} \frac{p_{ij}}{p_{\bullet j}} f_{im}. \tag{24}$$

Therefore the profile co-ordinate, f_{im}, is a scaled combination of the column profile co-ordinate, g_{jm}, as j varies. Alternatively (23) can be viewed as the weighted sum of the i-th row profile across the columns, where the weight is g_{jm}/λ_m. Thus, if p_{ij} is relatively large, g_{jm} will be heavily weighted and so will influence f_{im}. However, a direct comparison between a row and column profile is not possible using the scaling approach of (11) and (12). Refer also to discussions made in Section 3.6.4.

3.6 *Distances*

3.6.1 *Centring of profile co-ordinates*

The row and column profile co-ordinates are centred about the origin of the correspondence plot, called a centroid. As it will be shown, the origin is where the expected cell values $\{p_{i\bullet}p_{\bullet j}\}$ lie.

It can be shown that the row profile co-ordinates are centred about the centroid of the correspondence plot. That is

$$\sum_{i=1}^{I} p_{i\bullet} f_{im} = 0$$

for $m = 1, 2, \ldots, M^*$.

To show that this is true, recall the definition of the row profile co-ordinate of (11). Then

$$\sum_{i=1}^{I} p_{i\bullet} f_{im} = \sum_{i=1}^{I} \sum_{m=1}^{M^*} p_{i\bullet} a_{im} \lambda_m = \sum_{m=1}^{M^*} \left(\sum_{i=1}^{I} p_{i\bullet} a_{im} \right) \lambda_m = 0.$$

Similarly, it can be shown that the column profile co-ordinates are centred about the centroid. That is

$$\sum_{j=1}^{J} p_{\bullet j} g_{jm} = 0$$

for $m = 1, 2, \ldots, M^*$.

3.6.2 *Distance from the origin*

The squared distance of the i-th row profile from the origin is

$$d_I^2(i, 0) = \sum_{j=1}^{J} \frac{1}{p_{\bullet j}} \left(\frac{p_{ij}}{p_{i\bullet}} - p_{\bullet j} \right)^2 = \sum_{m=1}^{M^*} \left(\sum_{j=1}^{J} p_{\bullet j} b_{jm}^2 \right) a_{im}^2 \lambda_m^2$$

which simplifies to

$$d_I^2(i, 0) = \sum_{m=1}^{M^*} f_{im}^2 \tag{25}$$

and is the Euclidean distance of the i-th row profile co-ordinate from the origin. Therefore, equation (14) becomes

$$\frac{X^2}{n} = \sum_{i=1}^{I} p_{i\bullet} d_I^2(i, 0). \tag{26}$$

Hence the larger the distance of the i-th row profile in the M^*-dimensional correspondence plot from the origin, the larger the weighted discrepancy between the profile of category i to the average profile of the column categories. It follows that points far from the origin indicate a clear deviation from what we would expect under complete independence, while a point near the origin indicates that the frequencies in row i of the contingency table fits the independence hypothesis well. In fact, Lebart *et al.* (1984) showed by using confidence circles that the researcher is able to test graphically whether the position of a particular row or column category contributes to the hypothesis of independence for the contingency table. Beh (2001a) demonstrated that these circles can be usefully applied for the correspondence analysis of ordinal two-way contingency tables. Generally, if the origin lies outside of the confidence circle, then that category can be said to contribute to the dependency between the row and column categories of the contingency table. If the origin lies within the circle, then that category does not make such a contribution.

The same conclusion can be made for the Euclidean distance of the column profile co-ordinates to the origin.

3.6.3 Chi-squared distances

One of the advantages of using a correspondence plot is that the researcher is able to graphically establish similar and/or different profiles from the same variable.

The squared distance between two row profiles i and i' in an optimal correspondence plot is given by

$$d_I^2(i, i') = \sum_{j=1}^{J} \frac{1}{p_{\bullet j}} \left(\frac{p_{ij}}{p_{i\bullet}} - \frac{p_{i'j}}{p_{i'\bullet}} \right)^2 \tag{27}$$

and is the weighted Euclidean distance between these profiles.

Equation (27) can be written in terms of f_{im} and $f_{i'm}$, the profile co-ordinates of rows i and i' along the m-th principal axis.

However, when $m = 0$, the first term of the sum is zero since $a_{im} = a_{i'm} = 1$. Using the row profile co-ordinate definition of (11), this distance is measured in the Euclidean space, and for M^*-dimensional correspondence plot is defined by

$$d_I^2(i, i') = \sum_{m=1}^{M^*} (f_{im} - f_{i'm})^2 \tag{28}$$

or equivalently

$$d_I^2(i, i') = d_I^2(i, 0) + d_I^2(i', 0) - 2 \sum_{m=1}^{M^*} f_{im} f_{i'm}$$

so that $d_I^2(i, i) = d_I^2(i, 0)$.

Similarly, the Euclidean distance between columns j and j' can be measured by

$$d_J^2(j, j') = \sum_{i=1}^{I} \frac{1}{p_{i\bullet}} \left(\frac{p_{ij}}{p_{\bullet j}} - \frac{p_{ij'}}{p_{\bullet j'}} \right)^2 = \sum_{m=1}^{M^*} (g_{jm} - g_{j'm})^2. \tag{29}$$

These results lead to the conclusion that when two row profiles, or two column profiles, are similar, then they will be positioned closely to one another in the correspondence plot. If two profiles are different, then they will be positioned at a distance from one another. Therefore correspondence analysis can determine how profiles within a variable correspond to one another—thus the etymology of the technique.

The relationship between (27) and (28) also verifies the property of distributional equivalence as stated by Lebart, Morineau & Warwick (1984).

1. If two row profiles having identical profiles are aggregated, then the distance between them remains unchanged;
2. If two row profiles having identical distribution profiles are aggregated, then the distance between them remains unchanged.

The distributional equivalence results also apply to the column profile co-ordinates.

More recently, Yamakawa, Ichihashi & Miyoshi (1998, 1999) and Yamakawa, Kanaumi, Ichihashi & Miyoshi (1999) considered the development and application of a technique that involves, for the analysis of two-way tables, minimising the distance between two row profile co-ordinates

$$d_I^2(i, i') = \sum_{m=1}^{M^*} (f_{im} - f_{i'm})^S$$

for any positive value of S, where S is determined using the interior point method for the neural solution algorithm. This is used as an alternative to (28) and shifts correspondence analysis from the L_2-norm planar space to the more general L_S-norm space.

3.6.4 Inter-point distances

As the row and column co-ordinates can be simultaneously represented on the same correspondence plot, it seems reasonable to assume that one is able to measure the distance between a row and column profile. Such distances have been referred to as "inter-point" distances

Carroll, Green & Schaffer (1986, 1987, 1989) proposed a way of measuring these distances by recoding the two-way contingency table to be of the form of an indicator matrix. However Greenacre (1989) demonstrated that the claims made by these authors are flawed. According to Hoffman, de Leeuw & Arjunji (1995) the difference in opinion between Carroll, Green and Schaffer (CGS) and Greenacre can be described as follows:

> Greenacre was trained in the 'French' school, which appears to correspond nicely with the fact that he takes simple correspondence analysis to be a more fundamental and satisfactory technique than MCA. It also means that he tends to emphasise the so called 'chi-squared distance' interpretation of within-set distances. CGS have their starting point in multidimensional scaling and unfolding theory, which naturally leads them to emphasise between-set distance relations.

For a description of unfolding theory for categorical data refer to Heiser (1981).

This is where CGS made their mistake. They tried to enforce characteristics of multidimensional scaling (MDS) to be applicable to correspondence analysis, when MDS can be used to measure such distances. Distance issues in simple correspondence analysis should be confined to categories within a chosen variable (chi-squared distances). However, conclusions about inter-point distances should only be used as an informal guide to the correspondence between categories from different variables. For a more formal description of inter-point correspondences, some description of the association between the two variables needs to be taken into account. This is a goal of the simple correspondence analysis procedure of Beh (1997); to obtain a graphical description of the linear (and non-linear) association of two-way contingency tables by isolating generalised correlations (Davy, Rayner & Beh, 2003). Also refer to Greenacre (1989, p.363) for more comments on this issue.

3.7 Other Pearson Ratio Decompositions

3.7.1 The general decomposition

The decomposition of Pearson ratio's can be generalised by considering any approach that is of the form

$$\alpha_{ij} = D(\underset{\sim}{\lambda}, \underset{\sim}{\mathbf{a}}_i, \underset{\sim}{\mathbf{b}}_j) = \sum_{u=0}^{M_a^*} \sum_{v=0}^{M_b^*} a_{iu} \lambda_{uv} b_{jv} \tag{30}$$

where $D(\bullet)$ denotes the type of decomposition used to identify the scores and the measures of association between the two variables. For example, SVD is a special case where $M_a^* = M_b^* = M^* = \min(I, J) - 1$ and

$$\lambda_{uv} = \begin{cases} \lambda_m & u = v = m \\ 0 & u \neq v \end{cases}.$$

The term λ_{uv} quantifies a level of association between the two variables, for $u = 1, 2, \ldots, M_a^*$ and $v = 1, 2, \ldots, M_b^*$, such that

$$\lambda_{uv} = \sum_{i=1}^{I} \sum_{j=1}^{J} a_{iu} b_{jv} p_{ij}$$

where $\underset{\sim}{\mathbf{a}}_i = (a_{i0}, a_{i1}, \ldots, a_{iM_a^*})$ is the vector of scores for the i-th row category and $\underset{\sim}{\mathbf{b}}_j = (b_{j0}, b_{j1}, \ldots, b_{jM_b^*})$, is the vector of scores for the j-th column category. Here, M_a^* is the optimal number of dimensions required to represent the row scores, while M_b^* is the optimal number of dimensions required to represent the column scores. The elements of these vectors are constrained by the orthogonality property (8).

Therefore, the decomposition of Pearson ratios may also be expressed as

$$\alpha_{ij} - 1 = D^*\left(\underset{\sim}{\lambda}\,,\, \underset{\sim}{\mathbf{a}}_i\,,\, \underset{\sim}{\mathbf{b}}_j\right) = \sum_{u=1}^{M_a^*} \sum_{v=1}^{M_b^*} a_{iu} \lambda_{uv} b_{jv}.$$

Other than SVD, there are many ways that the decompositions of the Pearson ratio that can be made. We will be briefly considering another two—bivariate moment decomposition and hybrid decomposition.

As long as the decomposition of the Pearson ratio is of the form (30), correspondence analysis will maintain the same mathematical structure as the classical approach. However, the interpretation of the output depends on the decomposition used.

3.7.2 Moment decomposition

The method of correspondence analysis developed in Beh (1997) decomposes the Pearson ratio's such that

$$D^*\left(\underset{\sim}{\lambda}\,,\, \underset{\sim}{\mathbf{a}}_i\,,\, \underset{\sim}{\mathbf{b}}_j\right) = \sum_{u=1}^{I-1} \sum_{v=1}^{J-1} a_{iu} \lambda_{uv} b_{jv}$$

where

$$\lambda_{uv} = \sum_{i=1}^{I} \sum_{j=1}^{J} a_{iu} b_{jv} p_{ij}$$

are asymptotically standard normally distributed random variables. The values a_{iu} and b_{jv} are row and column orthogonal polynomials, respectively. Those polynomials that are considered are those presented by Best (1994) and Beh (1998). For each variable, the polynomials require a set of scores to reflect the structure of the categories. For example, ordinal scores should be used for ordinal categories. Therefore performing this type of correspondence analysis is especially informative for graphically displaying, and taking into consideration, the structure of ordered categories. Refer to Best & Rayner (1996) and Rayner & Best (1996) for a full interpretation of this decomposition.

An important theoretical implication of using the bivariate moment decomposition is that the λ_{uv}

terms have a clear and simple interpretation; they define the (u, v)-th bivariate moment between the row and column categories of N. As a result, Davy *et al.* (2003) refer to them, and their multivariate extensions, as generalised correlations. Consider as an example

$$\lambda_{11} = \sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij}\left(\frac{s_I(i)-\mu_I}{\sigma_I}\right)\left(\frac{s_J(j)-\mu_J}{\sigma_J}\right)$$

where $s_I(i)$ and $s_J(j)$ are the set or row and column scores used to construct the orthogonal polynomials, and $\mu_I = \sum_{i=1}^{I} s_I(i)p_{i\bullet}$ and $\sigma_I^2 = \sum_{i=1}^{I} s_I(i)^2 p_{i\bullet} - \mu_I^2$. The quantities μ_J and σ_J^2 are similarly defined.

When integer valued scores, $1, 2, \ldots$, are used as the scores to specify the underlying structure of both sets of categories, then λ_{11} is Pearson's product moment correlation (Rayner & Best, 1996), and when midrank scores are used, λ_{11} is Spearman's rank correlation (Best & Rayner, 1996).

3.7.3 Hybrid decomposition

Another method of decomposing α_{ij} is to consider a hybrid decomposition that consists of a mixture of elements from singular value decomposition and bivariate moment decomposition. It is of the form

$$D^*\left(\underset{\sim}{\lambda}\,,\ \underset{\sim}{\mathbf{a}}_i\,,\ \underset{\sim}{\mathbf{b}}_j\right) = \sum_{m=0}^{M^*}\sum_{v=0}^{J-1} a_{im}\lambda_{(m)v}b_{jv}$$

where

$$\lambda_{(m)v} = \sum_{i=1}^{I}\sum_{j=1}^{J} a_{im}b_{jv}p_{ij}$$

and uses singular vectors for the nominal (row) variable and orthogonal polynomials to take into account the complex structure of the (columns) variable. It is based on the decomposition of the Pearson chi-squared statistic presented in Beh (2001b) and has been used to perform ordinal correspondence analysis on singly ordered two-way contingency tables; see Beh (2003a).

An important implication of using this decomposition is that the square of the singular values obtained from singular value decomposition can be expressed as the sum of squares of the $\lambda_{(u)v}$. That is, for the m-th largest singular value

$$\lambda_m^2 = \lambda_{(m)1}^2 + \lambda_{(m)2}^2 + \cdots + \lambda_{(m)J-1}^2.$$

Therefore, using such a decomposition means that singular values can be partitioned into linear $(\lambda_{(m)1})$, quadratic $(\lambda_{(m)2})$ and higher order terms.

4 Modeling Simple Correspondence Analysis

4.1 RC Correlation Model

In Section 2.3, the departure from the independence hypothesis can be made by testing (2) against (1).

Such a measure can be quantified by comparing the definition of the Pearson ratios (3) and the

decomposition (30)

$$p_{ij} = p_{i\bullet}p_{\bullet j}\left(1 + \sum_{u=1}^{M_a^*}\sum_{v=1}^{M_b^*} a_{iu}\lambda_{uv}b_{jv}\right).$$

For example when singular value decomposition is applied to the Pearson ratio's

$$p_{ij} = p_{i\bullet}p_{\bullet j}\left(1 + \sum_{m=1}^{M^*} a_{im}\lambda_m b_{jm}\right). \tag{31}$$

Model (31) is termed the Fisher's identity by Lebart *et al.* (1984) and Lancaster (1969) and is named so to reflect the work of Fisher (1940). However, it is also more commonly referred to as the saturated RC canonical correlation model.

Model (31) is saturated when $M^* = \min(I, J) - 1$. The work of de Leeuw & van der Heijden (1991) showed that model (31) is equivalent to the canonical correlation model when M^* of the canonical correlations between the row and column variables are non-zero.

The unsaturated model is

$$p_{ij} \approx p_{i\bullet}p_{\bullet j}\left(1 + \sum_{m=1}^{M} a_{im}\lambda_m b_{jm}\right) \tag{32}$$

where $1 \leq M < M^*$.

Grassi & Visentin (1994) note that Escoufier (1983, 1984) proposed a generalisation of a correspondence analysis, where the generalised RC correlation model is

$$p_{ij} = q_{ij} + q_{i1}q_{j2}\sum_{m=1}^{M^*} a_{im}\lambda_m b_{jm} \tag{33}$$

where q_{ij} are joint proportions obtained from some specified hypothesis, and q_{i1} and q_{j2} are any weights with the restrictions $q_{i1} > 0$, $q_{j2} > 0$, $\sum_{i=1}^{I} q_{i1} = \sum_{j=1}^{J} q_{j2} = 1$. For example, under the hypothesis of independence, $q_{ij} = p_{i\bullet}p_{\bullet j}$, $q_{i1} = p_{i\bullet}$ and $q_{j1} = p_{\bullet j}$.

Also refer to de Leeuw & van der Heijden (1991) for a discussion of another generalised version of (31).

4.2 *Basic Correspondence Model*

When N is a contingency table of size $I \times 2$, or $2 \times J$, then $M^* = 1$. For such a table, the RC canonical correlation model (31) becomes

$$p_{ij} = p_{i\bullet}p_{\bullet j}\left(1 + a_{i1}\lambda_1 b_{j1}\right). \tag{34}$$

Model (34) has been studied by Goodman (1985b), Gilula, Krieger & Ritov (1988), Rom & Sarkar (1992), Ritov & Gilula (1993), Williams (1952) and Gilula & Haberman (1986) and is called the rank-2 canonical correlation model, or basic correspondence model. It can be seen from this model that the measure of correlation λ_1 between the row and column scores, $\{a_{i1}\}$ and $\{b_{j1}\}$, respectively, can be calculated by

$$\lambda_1 = \sum_{i=1}^{I}\sum_{j=1}^{J} a_{i1}b_{j1}p_{ij}. \tag{35}$$

When $\lambda_1 = 0$, model (34) becomes the model of independence (1). By rearranging (34) we see that

$$\frac{p_{ij} - p_{i\bullet}p_{\bullet j}}{p_{i\bullet}p_{\bullet j}} = \lambda_1 a_{i1}b_{j1}. \tag{36}$$

Squaring both sides of (35) and multiplying by $p_{i\bullet}p_{\bullet j}$ yields

$$\frac{\left(p_{ij} - p_{i\bullet}p_{\bullet j}\right)^2}{p_{i\bullet}p_{\bullet j}} = \lambda_1^2 a_{i1}^2 b_{j1}^2 p_{i\bullet}p_{\bullet j}. \tag{37}$$

So summing over the I rows and J columns of (37) and using (8) gives

$$\frac{X^2}{n} = \lambda_1^2 \tag{38}$$

which is just equation (10) when $M^* = 1$.

Therefore, the total inertia of a contingency table when $M^* = 1$ is just the square of the correlation between the rows and columns. Examples of where a $I \times 2$ or a $2 \times J$ contingency table will arise include those where there are two responses for one variable; "Yes" and "No", "Pass" and "Fail", "Male" and "Female", and so on.

Rom & Sarkar (1992) proposed the rank-2 model

$$p_{ij} = p_{i\bullet}p_{\bullet j}\left\{1 + \phi\left(\alpha_i + \beta_j + \lambda_1 a_{i1}b_{j1}\right)\right\}^{1/\phi} \tag{39}$$

where $\{\alpha_i\}$ and $\{\beta_j\}$ depend on the parameter ϕ. If $\phi = 1$ in (39) then

$$p_{ij} = p_{i\bullet}p_{\bullet j}\left(1 + \alpha_i + \beta_j + \lambda_1 a_{i1}b_{j1}\right) \tag{40}$$

so that

$$p_{ij} - p_{i\bullet}p_{\bullet j}\left(1 + \alpha_i + \beta_j\right) = \lambda_1 a_{i1}p_{i\bullet}b_{j1}p_{\bullet j}.$$

Multiplying this by $a_{i1}b_{j1}$ and summing over the I rows and J columns yields

$$\lambda_1 = \sum_{i=1}^{I}\sum_{j=1}^{J} a_{i1}b_{j1}p_{ij} - \sum_{i=1}^{I}\sum_{j=1}^{J}\left(a_{i1}p_{i\bullet}\right)\left(b_{j1}p_{\bullet j}\right)\left(1 + \alpha_i + \beta_j\right).$$

Since λ_1 is defined by (35), it can be seen that $\alpha_i = \beta_j = 0$, using (8).

Therefore, when $\phi = 1$, (39) reduces to be the rank-2 canonical correlation model of (34).

Also, if the correlation between the row and column scores is zero, the model of Rom & Sarkar (1992) reduces to the independence model (1).

4.3 Reconstitution Model

In the correspondence analysis context, the scores used to identify similarities and differences between categories within a particular variable are not the standard co-ordinates, instead they are the

profile co-ordinates defined by (11) and (12). Using this transformation, the RC canonical correlation model (31) becomes

$$p_{ij} = p_{i\bullet}p_{\bullet j}\left(1 + \sum_{m=1}^{M^*} \frac{f_{im}g_{jm}}{\lambda_m}\right). \tag{41}$$

Model (41) is referred to as the correspondence model by Goodman (1986), or the reconstitution model by Greenacre (1984) and many others.

The reconstitution model of (41), just like (31), can be used to determine the effect of reducing the problem from M^* dimensions to M dimensions. The better the approximation given by the model, the better the M-dimensions are in representing the row and column profiles of the contingency table.

The reconstitution formula verifies that if the row and column profile co-ordinates are close to zero then the hypothesis of complete independence is supported.

4.4 Other Models

There are many models that can be used as an alternative to (31) and (41). Here two more are briefly described.

One of the most popular alternatives to the RC correlation model of (31) is the Goodman RC model, as seen in Goodman (1979). This model is

$$p_{ij} = \alpha_i\beta_j \exp\left(\sum_{m=1}^{M^*} \varphi_m a_{im}b_{jm}\right) \tag{42}$$

where a_{im} and b_{jm} are as defined earlier. The parameter φ_m is termed the coefficient of intrinsic association, while α_i and β_j are positive parameters, or nuisance parameters as Gilula *et al.* (1988) call them. Escoufier & Juncar (1986) showed that these parameters can be calculated by

$$\alpha_i = \exp\left(\sum_{j=1}^{J} p_{\bullet j} \log p_{ij}\right) \tag{43}$$

$$\beta_j = \exp\left(\sum_{i=1}^{I} p_{i\bullet} \log p_{ij} - \sum_{i=1}^{I}\sum_{j=1}^{J} p_{i\bullet}p_{\bullet j} \log p_{ij}\right). \tag{44}$$

Model (42) has been extensively reviewed by many such as Goodman (1979, 1981, 1985b, 1985a, 1986, 1996), Escoufier (1988), Haberman (1981) and Becker & Clogg (1989). In fact, Goodman (1981) applied model (42) to two-way contingency tables with ordered variables.

Consider the rank-2 Goodman correlation model when $M^* = 1$

$$p_{ij} = \alpha_i\beta_j \exp\left(\varphi_1 a_{i1}b_{j1}\right). \tag{45}$$

This has been examined by Gilula, Krieger & Ritov (1988), Ritov & Gilula (1991, 1993), Gilula (1984) and Rom & Sarkar (1992).

Gilula, Krieger & Ritov (1988) show that the parameter λ_1 from equation (35) and ϕ_1 from (45) are different measurements. They point out that while λ_1 is a correlation value, as can be seen from (35), and thus lies within the interval $[-1, 1]$, ϕ_1 lies within $[0, \infty)$.

Goodman (1991) showed that the intrinsic association can be calculated by

$$\varphi_m = \sum_{i=1}^{I}\sum_{j=1}^{J} a_{im} b_{jm} \log p_{ij} \tag{46}$$

for $m = 1$.

For classical correspondence analysis, reparameterise the row and column scores so that

$$f_{im} = a_{im}\varphi_m \tag{47}$$

$$g_{jm} = b_{jm}\varphi_m \tag{48}$$

so that the reconstitution formula is

$$p_{ij} = \alpha_i \beta_j \exp\left(\sum_{m=1}^{M^*} \frac{f_{im} g_{jm}}{\varphi_m}\right). \tag{49}$$

Therefore, the constraints of (14) and (15) become

$$\sum_{i=1}^{I} p_{i\bullet} f_{im} f_{im'} = \begin{cases} \varphi_m^2, & m = m' \\ 0, & m \neq m' \end{cases} \tag{50}$$

$$\sum_{j=1}^{J} p_{\bullet j} g_{jm} g_{jm'} = \begin{cases} \varphi_m^2, & m = m' \\ 0, & m \neq m' \end{cases}. \tag{51}$$

In some situations, reconstituting the cell values using the unsaturated version of (41) can lead to negative estimates of p_{ij}. This problem was also pointed out for the unsaturated version of the reconstitution formula

$$p_{ij} = p_{i\bullet} p_{\bullet j}\left(1 + \sum_{u=1}^{I-1}\sum_{v=1}^{J-1} a_{iu}\lambda_{uv} b_{jv}\right)$$

and for similar models by Rayner & Best (1996) and Beh (2001b). Therefore, the exponential approximation should be considered as an alternative.

5 The Correspondence Analysis Family

Over the past few decades correspondence analysis has encountered many attempts to adjust the method so that it can cater for interdisciplinary problems that have arisen.

The first adjustment was made in the field of ecology and proposed by Hill & Gauch Jr. (1980). Their method, which is called detrended correspondence analysis, is a method of removing the technique's characteristic "arch effect" by cutting the first axis into segments during the calculation of the row and column scores and then resetting the average of each segment to zero. Palmer (1993) pointed out that such a refinement produces "inelegancies" that have been criticised by Minchin (1987), Oksanen (1987, 1988) and Wartenberg, Ferson & Rohlf (1987). The criticisms of Wartenberg *et al.* (1987) prompted Peet, Knox, Case & Allen (1988), although agreeing with some of their points, to "provide prospective users with a more balanced perspective on the advantages and disadvantages of DCA". Oksanen & Minchin (1997) investigated the instability of detrended correspondence analysis

using computer programs designed for the execution of correspondence analysis and its detrended form.

Another method of correspondence analysis, called canonical correspondence analysis, was developed by ter Braak (1986, 1987). It conducts correspondence analysis by including the additional step of selecting the linear combination of row variables that maximises the variation of the column scores. In fact Birks, Peglar & Austin (1994) provide a list of 378 references relating to the application and development of canonical correspondence analysis between 1986 to 1993. This 1994 list has since been updated and includes 402 references up to 1996. This bibliography can be found at the following web address;

`http://www.fas.umontreal.ca/BIOL/Casgrain/cca_bib/index.html`

Many more can now be found, for example, by utilising search engines available on the internet.

Another form of correspondence analysis that can be included in this family is joint correspondence analysis developed by Greenacre (1988, 1990, 1991) and improved upon by Boik (1996). This is a multiple correspondence analysis adjustment, but can naturally be used for the analysis of two-way contingency tables as well. For bivariate categorical data, the method involves transforming the table into a $(I + J) \times (I + J)$ Burt matrix (Burt, 1950) where the diagonal sub-matrices consist of row and column marginal frequencies and the off-diagonal matrices are N and its transpose. Applying classical correspondence analysis to the Burt matrix will lead to an inflated total inertia since the diagonal sub-matrices are included in the analysis. This leads to a problem since they are included in the analysis because the cell frequencies, and hence the marginal frequencies, are assumed to be fixed. Joint correspondence analysis is performed by omitting the diagonal sub-matrices and leads to an identical correspondence analysis when compared with the classical approach.

Thus, correspondence analysis can no longer be viewed as a single multivariate analytic technique. Instead there is, as Palmer (1993) calls it, a "correspondence analysis family" that incorporates all these correspondence analysis methodologies. Also included in this family are detrended canonical correspondence analysis, which is a combination of the two methods, non-symmetrical correspondence analysis (D'Ambra & Lauro, 1989, 1992; Kroonenberg & Lombardo, 1998, 1999; Lauro & Balbi, 1999; Lombardo *et al.*, 1996), and partial canonical correspondence analysis (ter Braak, 1988).

6 Texts and Applications

Despite the English introduction to correspondence analysis by Hill (1974), its initial development had been fairly slow to reach English speaking countries. Much of the earlier work, in the 1960's and early 1970's, was written in French. There are three possible reasons for this slow development:

1. The initial lag in the development outside France may be due to the problem non-English speaking researchers have with the language.
2. Correspondence analysis is often introduced without any reference to other methods of the statistical treatment of categorical data. Thus, initially, the application and development of correspondence analysis was rare.
3. Due to the difference in philosophies of data analysis between European and English speaking statisticians, correspondence analysis, in its early years, failed to mature outside of France.

There are only a few English written articles concerned with correspondence analysis, from the graphical point of view, between late 1960's and early 1970's. These include Benzécri (1969), Hill (1974), Maignan (1974) and Teil (1975). There are also only a few early English written articles discussing the application of correspondence analysis. Teil & Cheminee (1975) applied correspondence analysis to rock samples collected from an old volcanic region from the Erta Chain in Eastern

Ethiopia. The aim of this study was to determine important geological elements in the region. David, Campligio & Darling (1974) used correspondence analysis to study the major factors influencing geological processes in the volcanic belt along the Superior Province of the Canadian Shield. Another fairly early application of correspondence analysis can be seen in van Heel & Frank (1980). They applied the method to help identify images of biological macro-molecules using an electron microscope.

A decade of further development brought about the English publication of two independent texts, Greenacre (1984) and Lebart *et al.* (1984). They are both now used as standard books and can be used as an introduction to correspondence analysis. Michael Greenacre was a student of Jean-Paul Benzécri for two years from 1973, and as a result, much of today's literature on correspondence analysis uses Benzécri's style. The text of Ludovic Lebart, Alain Morineau and Kenneth M. Warwick is based on an earlier French work by Lebart, Morineau & Tabard (1977) and discusses other multivariate techniques such as principal component analysis, canonical correlation analysis and cluster analysis.

Since these books have become standard texts on correspondence analysis, the technique has experienced an explosion of its application and development in most fields of research. Of particular note is the method's growth in popularity from the early 1990's.

Correspondence analysis has been applied in areas such as sensory evaluation, ecology, psychometry, health care, economics, medicine, literature and engineering. Refer to Beh (2002), which can be found at the web address;

`http://www.uws.edu.au/about/acadorg/clb/sqmms/research/reports`

for a description of articles relevant to these disciplines.

Archaeology is a discipline that has experienced a strong interest in the application of correspondence analysis. Recently, Clouse (1999) investigated the application of correspondence analysis to aiding archaeological studies. Baxter (1994) provides some excellent discussions of correspondence analysis in this area. Of special interest are the reference lists that he has put together. They provide an extensive overview of the many ways correspondence analysis can be used in archaeological studies. These lists can be viewed on the internet at the following addresses;

`http://science.ntu.ac.uk/msor/mjb/oldcabib.htm`—references prior to 1993
`http://science.ntu.ac.uk/msor/mjb/corranbib.html`—references from 1993.

Over the years, there are many more articles that have appeared in the literature that span a wide range of disciplines. The disciplines discussed in this paper, and the articles noted by Beh (2002), are included to give a flavour of the diverse applications and popularity that correspondence analysis has maintained.

7 Computing

There are several computer packages and review journal articles (both relatively old and new reviews) relating to many of the computational issues associated with correspondence analysis. For example, published review articles include those by Lebart & Morineau (1982), Greenacre (1986), Carr (1990), Hoffman (1991), Gorman & Primavera (1993), Tian, Sorooshian & Meyers (1993), Thompson (1995) and Bond & Michailidis (1997).

There are also many commercially available programs, and computer programs published in journal form, that perform classical correspondence analysis. For example, Carr (1990) details FORTRAN code, called CORSPOND, that executes simple correspondence analysis. It is limited to the

analysis of a two-way contingency table with 80 rows and 20 columns with information associated with a maximum of the 10 dimensions calculated. However, Carr (1990) points out that these limitations are easily extended by modifying program dimensions. While these three values are used as input values, the program's output consists of all the input data, a summary of eigenvalues and the percentage contribution they make with the total inertia, the profile co-ordinates and two-dimensional plots. For example, if M-dimensions are specified, there are $M\ (M-1)/2$ two-dimensional plots that can be displayed. The Carr (1990) program produces each of these plots.

Tian *et al.* (1993) outlines a program called MATCORS written using the language `Matlab`. Their program allows for supplementary row and column categories to be included. The program of Tian and his co-authors allow for the analysis of a contingency table with a large number of rows and columns. The output is thorough with graphs for relative and absolute contributions of profiles to each principal axis, error profiles and supplementary projections, including an optimal correspondence plot.

Everitt (1994) offers a very simple program for classical correspondence analysis using S-PLUS. The correspondence plot using S-PLUS can easily incorporate colour and contain all the relevant information, such as profile names, and partial inertia values. Shi & Carr (2001) provides details for R code, CORSPONDA, which also performs simple correspondence analysis.

There are also many commercially available packages that perform correspondence analysis. A brief search of the internet will also reveal many packages available from commercial vendors. However, we will only briefly discuss those that have appeared in the research literature. Hoffman (1991) and Thompson (1995) give a good review of, between them, eight packages. Hoffman (1991) discussed Dual3 (version 3.2), MAPWISE (version 2.01), PC-MDS (version 5.0) and SimCA (version 1.5). SimCA (version 1.0) was discussed by Greenacre (1986). Dual3 deals with the dual scaling approach to correspondence analysis of Nishisato (1980) and was written by Shizuhiko and Ira Nishisato. It is written in BASIC and so is command orientated rather than menu driven. Hoffman (1991) suggests that Dual3 is easy to use only if the researcher has prior knowledge of dual scaling (or simple correspondence analysis). However, the program will only calculate solutions up to three dimensions. Hoffman also suggests that MAPWISE is not the best package to use. The documentation is wrought with "incorrect assertions and misleading statements" (Hoffman, 1991, p. 308). However, comparing MAPWISE with other correspondence analysis programs, it is menu orientated, easy to use and is colourful. MAPWISE can handle contingency tables containing up to 100 rows and 100 columns, and is largely written for industrial applications.

SimCA is a computer package largely written for the correspondence analysis of two-way contingency tables, but can perform a multiple correspondence analysis only if the data is presented in the form of an indicator matrix. It is written in `Turbo` BASIC and unlike Nishisato's program (Dual3) can calculate solutions up to the tenth dimension. As it is written to be able to analyse indicator matrices, SimCA can analyse a contingency table with up to 175 columns and virtually an unlimited number of rows (however, the larger the data set, the slower the computational power), and thus is a very good program to use.

Thompson (1995) reviewed four other commercially available packages; BMDP (version 7.0), NCSS (version 5.3), SAS (version 6.07) and SPSS (version 6.0).

As far as worked examples are concerned, Thompson felt at the time that BMDP is a good start for those learning correspondence analysis, while SAS is also very good in this respect (but with fewer examples). However, the SAS documentation is more technical than the other three packages reviewed, while the SPSS documentation is good. All of the programs will conduct a simple correspondence analysis. However, all except NCSS, will calculate a multiple correspondence analysis. The three programs that perform this analysis do so via the Burt matrix. SAS, NCSS and BMDP will permit an analysis of supplementary categories, yet SPSS will not.

Lebart (1982) reviewed their program which performs a correspondence analysis of two-way

and multi-way contingency tables. Their program, written in FORTRAN, also performs a principal component analysis and cluster analysis on the data, and can analyse contingency tables consisting of hundreds of rows and thousands of columns.

Bond & Michailidis (1997) have also written a program capable of performing correspondence analysis on two-way and multi-way contingency tables. Their program, called ANACOR, is written in `Lisp-Stat` and its performance is claimed (by the authors) to be as good as the commercially available products, and in some respects, better. Their package has one advantage that many of the others do not have, and that is mouse-driven zooming capabilities for the correspondence plot. When analysing large contingency tables, the correspondence plot can often look very cluttered, especially toward the centroid, where many profile categories may be positioned. ANACOR allows the user to zoom-in or zoom-out of the plot by drawing a square around the region that is to be investigated. ANACOR is also menu driven and consists of colour graphics.

Gorman & Primavera (1993) described their program MCA.EXE, which performs a multiple correspondence analysis via the indicator or Burt matrices (see Greenacre, 1984, for more details on this type of analysis). The program is written in QUICKBASIC 4.5 and can be executed on an MS/DOS or DC/DOS machine. The program can handle any number of observations/people classified into a contingency table, however, the total number of categories must not exceed 70.

Correspondence analysis modules, or add-ins, also appear in other popular packages. STATA includes the module CORANAL, as described by van Kerm (1998). XLSTAT (Version 5) is available as an "add-in" to Microsoft Excel to perform correspondence analysis as well as multiple correspondence analysis, multidimensional scaling, principal component analysis and discriminant analysis. MVSP (Multivariate Statistical Package) is a Windows package that performs correspondence analysis and its detrended and canonical forms.

The statistical package MINITAB (Version 13) also contains a module CORRES.MTB that performs simple, and multiple, correspondence analysis. STATISTICA also contains a module to perform canonical and detrended correspondence analysis.

For the simple correspondence analysis associated with ordinal two-way contingency tables, Beh (2004) discusses in some detail programs written using S-PLUS.

Refer to Beh (2002) for a more comprehensive list of contributors to the computation of correspondence analysis.

8 Other Issues

The aim of this paper has been to provide a discussion of the development and the application of correspondence analysis. However, there are many other issues associated with correspondence analysis that have not been discussed in any detail due to space restrictions. We will therefore provide references relating to several of these issues.

Recent theoretical developments in Japan have been made into the correspondence analysis of artificial shaped data, especially cylindrical and binary cylindrical shapes, disk, torus, symmetric polyhedron, multiple circular, spherical and other typical geometric figures. See Okamoto (1994a, 1994b, 1994c, 1995a, 1995b, 2000), Endo (1995, 1996) and Okamoto & Endo (1995) for details.

The issue of influence of cells, and responses, in correspondence analysis has been investigated. Kim (1992, 1994), and Pack & Jolliffe (1992) looked at the impact on the analysis by including and excluding, or deleting, categories, while Pack & Jolliffe (1992) also looked at the topic of influence. An issue similar to that of influence was discussed by Krzanowski (1993). For an incidence matrix, the author examines methods for identifying columns (attributes) which highlight important row (incidence) differences.

Some of the most popular discussions concern the theoretical similarities between correspondence analysis and log-linear models, despite their differences in philosophy of data analysis. Goodman

(1985a, 1986), van der Heijden & de Leeuw (1985), Choulakian (1988), van der Heijden & Worsley (1988) and van der Heijden, de Falguerolles & de Leeuw (1989) showed the link between correspondence analysis and non-ordinal log-linear models. Gower (1989) comments on the link between correspondence analysis and log-linear models:

> Correspondence analysis (CA) has been enthusiastically developed in France and widely adopted in other continental countries but has had a more cautious reception in Britain. In part this has been a consequence of claims that CA is a descriptive method and not model based. Links between CA and log-linear analysis (LLA) have helped to gain more acceptance in Britain, and perhaps for LLA to gain more acceptance abroad.

We refer to the above mentioned articles for more details. For the link between correspondence analysis and log-linear models of ordinal categorical data refer to Beh (2001b). Of particular interest is their non-iterative estimation procedure of parameters from an ordinal log-linear model.

9 Discussion

The development of correspondence analysis is a long and interesting one, and one that has not being exclusively confined to statisticians. Its diversity of development and application range the fields of biometry, psychometry, linguistics to health care and vegetation science. Therefore, correspondence analysis makes a very versatile method of data analysis in all situations where an exploratory or more in-depth analysis of categorical data is required. In a sense this is a reflection of all statistical techniques, and is nicely summed up by Kendall (1972, p.194):

> It is hard to think of any subject which has not made some kind of contribution to statistical theory—agriculture, astronomy, biology, chemistry and so on through the alphabet. The remarkable thing, perhaps, is that these lines of development remained relatively independent for so long and only in the present century have been seen to have a common conceptual content.

References

Aitchison, J. & Greenacre, M. (2002). Biplots in compositional data. *Applied Statistics*, **51**, 375–392.

Baxter, M.J. (1994). *Exploratory Multivariate Analysis in Archaeology*. Edinburgh: Edinburgh University Press.

Becker, M.P. & Clogg, C.C. (1989). Analysis of sets of two-way contingency tables using association models. *Journal of the American Statististical Association*, **84**, 142–151.

Beh, E.J. (1997). Simple correspondence analysis of ordinal cross-classifications using orthogonal polynomials. *Biometrical Journal*, **39**, 589–613.

Beh, E.J. (1998). A comparative study of scores for correspondence analysis with ordered categories. *Biometrical Journal*, **40**, 413–429.

Beh, E.J. (2001a). Confidence circles for correspondence analysis using orthogonal polynomials. *Journal of Applied Mathematics and Decision Sciences*, **5**, 1–11.

Beh, E.J. (2001b). Partitioning Pearson's chi-squared statistic for singly ordered two-way contingency tables. *The Australian and New Zealand Journal of Statistics*, **43**, 327–333.

Beh, E.J. (2002). Simple correspondence analysis: A bibliographic review. Research Report No. QMMS2002.9. School of Quantitative Methods and Mathematical Sciences, University of Western Sydney, Australia.

Beh, E.J. (2003a). Singly ordered simple correspondence analysis. Research Report No. QMMS2003.2. School of Quantitative Methods and Mathematical Sciences, University of Western Sydney, Australia.

Beh, E.J. (2003b). A simple generalisation of approaches for the graphical analysis of cross-classified data. Research Report QMMS2003.1, School of Quantitative Methods and Mathematical Sciences, University of Western Sydney, Australia.

Beh, E.J. (2004). S-PLUS code for ordinal correspondence analysis. *Computational Statistics*, (to appear).

Benzécri, J.P. (1969). Statistical analysis as a tool to make patterns emerge from data. In *Methodologies of Pattern Recognition*, Ed. S. Watanabe, pp. 35–74.

Benzécri, J.P. (1973a). *L'Analyse des donnees: II. La Taxonomie*. Paris: Dunod.

Benzécri, J.P. (1973b). *L'Analyse des donnees: I. La Taxonomie*. Paris: Dunod.

Benzécri, J.P. (1992). *Correspondence Analysis Handbook*. New York: Marcel Dekker.

Best, D.J. (1994). Nonparametric comparison of two histograms. *Biometrics*, **50**, 538–541.

Best, D.J. & Rayner, J.C.W. (1996). Nonparametric analysis for doubly ordered two-way contingency tables. *Biometrics*, **52**, 1153–1156.

Birks, H., Peglar, S. & Austin, H. (1994). *An Annotated Bibliography of Canonical Correspondence Analysis and Related Constrained Ordination Methods 1986–1993*. Tech. Rep. Botanical Institute, University of Bergen, Allegaten, Bergen, Norway.

Boik, R.J. (1996). An efficient algorithm for joint correspondence analysis. *Psychometrika*, **61**, 255–269.

Bond, J. & Michailidis, G. (1997). Interactive correspondence analysis in a dynamic object—oriented environment. *Journal of Statistical Software*, **2**.

Burt, C. (1950). The factorial analysis of qualitative data. *J. Statist. Psychology*, **3**, 166–185.

Carr, J.R. (1990). CORSPOND: A portable FORTRAN-77 program for correspondence analysis. *Computers and Geosciences*, **16**, 289–307.

Carroll, J.D., Green, P.E. & Schaffer, C.M. (1986). Interpoint distance comparisons in correspondence analysis. *Journal of Marketing Research*, **23**, 271–280.

Carroll, J.D., Green, P.E. & Schaffer, C.M. (1987). Comparing interpoint distances in correspondence analysis. *J. Marketing Research*, **24**, 445–450.

Carroll, J.D., Green, P. & Schaffer, C.M. (1989). Reply to Greenacre's commentary on the Carroll–Green–Schaffer scaling of two-way correspondence analysis solutions. *Journal of Marketing Research*, **26**, 366–368.

Choulakian, V. (1988). Exploratory analysis of contingency tables by log-linear formulation and generalisations of correspondence analysis. *Psychometrika*, **53**, 235–250.

Clouse, R.A. (1999). Interpreting archaeological data through correspondence analysis. *Historical Archaeology*, **33**, 99–107.

D'Ambra, L. & Lauro, N. (1989). Non symmetrical analysis of three-way contingency tables. In *Multiway Data Analysis*, Eds. R. Coppi and S. Bolasco, pp. 301–315. Amsterdam: North-Holland.

D'Ambra, L. & Lauro, N.C. (1992). Non symmetrical exploratory data analysis. *Statistica Applicata*, **4**, 511–529.

David, M., Campligio, C. & Darling, R. (1974). Progresses in R- and C-Mode analysis: correspondence analysis and applications to the study of geological processes. *Canadian Journal of Earth Sciences*, **11**, 131–146.

Davy, P.J., Rayner, J.C.W. & Beh, E.J. (2003). Generalised correlations. Preprint, No.4/03, School of Mathematics and Applied Statistics, University of Wollongong, Australia.

de Leeuw, J. (1983). On the prehistory of correspondence analysis. *Statistica Neerlandica*, **37**, 161–164.

de Leeuw, J. & van der Heijden, P. (1991). Reduced rank models for contingency tables. *Biometrika*, **78**, 229–232.

Eckart, C. & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, **1**, 211–218.

Endo, H. (1995). Correspondence analysis of an artificial binary cylinder data. *Statistics & Probability Letters*, **25**, 231–240.

Endo, H. (1996). Correspondence analysis of artificial binary data with circular structure. *Mathematica Japonica*, **43**, 339–355.

Escoufier, B. (1983). Analyse de la difference entre deux medures sur le produit de deux memes ensembles. *Les Cahiers de l'Analyse des Donnees*, **8**, 325–329.

Escoufier, B. (1984). Analyse factorielle en reference a un modele: application a l'analyse de tableaux d'échange. *Revue de Statistique Appliquee*, **32**, 25–36.

Escoufier, Y. (1988). Beyond correspondence analysis. In *Classification and Related Methods of Data Analysis*, Ed. H.H. Bock, pp. 505–514. Amsterdam: North-Holland.

Escoufier, Y. & Juncar, S. (1986). Least-squares approximation of frequencies or their logarithms. *International Statistical Review*, **54**, 279–283.

Everitt, B.S. (1994). *A Handbook of Statistical Analysis using S-plus*. Chapman and Hall.

Fienberg, S.E. (1982). Contingency tables. *Encyclopedia of Statistical Sciences*, **2**, 161–171.

Fisher, R.A. (1940). The precision of discriminant functions. *Annals of Eugenics*, **10**, 422–429.

Gabriel, K.R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, **58**, 453–467.

Gabriel, K.R. & Odoroff, C.L. (1990). Biplots in biomedical research. *Statistics in Medicine*, **9**, 469–485.

Gifi, A. (1990). *Non-linear Multivariate Analysis*. Chichester: Wiley.

Gilula, Z. (1984). On some similarities between canonical correlation models and latent class models for two-way contingency tables. *Biometrika*, **71**, 523–529.

Gilula, Z. & Haberman, S.J. (1986). Canonical analysis of contingency tables by maximum likelihood. *Journal of the American Statistical Association*, **81**, 780–788.

Gilula, Z., Krieger, M. & Ritov, Y. (1988). Ordinal Association in contingency tables: Some interpretative aspects. *Journal of the American Statistical Association*, **83**, 540–545.

Goodman, L.A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association*, **74**, 537–552.

Goodman, L.A. (1981). Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *Journal of the American Statistical Association*, **76**, 320–334.

Goodman, L.A. (1985a). Correspondence analysis models, log-linear models and log-bilinear models for the analysis of contingency tables. *Bulletin of the International Statistical Institute*, **51**, 28.1-1–28.1-14.

Goodman, L.A. (1985b). The analysis of cross-classified data having ordered and/or unordered categories: association models, correlation models and asymmetry models for contingency tables with or without missing entries. *The Annals of Statistics*, **13**, 10–69.

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.

Goodman, L.A. (1991). Measures, models and graphical displays in the analysis of cross-classified data. *Journal of the American Statististical Association*, **86**, 1085–1111.

Goodman, L.A. (1996). A single general method for the analysis of cross-classified data: Reconciliation and synthesis of some methods of Pearson, Yule, and Fisher, and also some methods of correspondence analysis and association analysis. *Journal of the American Statistical Association*, **91**, 408–428.

Gorman, B.S. & Primavera, L.H. (1993). MCA—a simple program for multiple correspondence analysis. *Educational Psychological Measurement*, **53**, 685–687.

Gower, J.C. (1989). Discussion of "a combined approach to contingency table analysis using correspondence analysis and log-linear analysis". *Applied Statistics*, **38**, 249–292.

Grassi, M. & Visentin, S. (1994). Correspondence analysis applied to grouped cohort data. *Statistics in Medicine*, **13**, 2407–2425.

Greenacre, M.J. (1984). *Theory and Application of Correspondence Analysis*. London: Academic Press.

Greenacre, M.J. (1986). SIMCA: A program to perform simple correspondence analysis. *The American Statistician*, **40**, 230–231.

Greenacre, M.J. (1988). Correspondence analysis of multivariate categorical data by weighted least-squares. *Biometrika*, **75**, 457–467.

Greenacre, M.J. (1989). The Carroll–Green–Schaffer scaling in correspondence analysis: a theoretical and empirical appraisal. *Journal of Marketing Research*, **26**, 358–365.

Greenacre, M.J. (1990). Some limitations of multiple correspondence analysis. *Computational Statistics Quarterly*, **3**, 249–256.

Greenacre, M.J. (1991). Interpreting multiple correspondence analysis. *Applied Stochastic Models and Data Analysis*, **7**, 195–210.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In *The Prediction of Personal Adjustment*, Ed. P. Horst, pp. 319–348. Social Science Research Council, New York.

Guttman, L. (1953). A note on Sir Cyril Burt's "Factorial analysis of qualitative data". *The British Journal of Statistical Psychology*, **6**, 1–4.

Haberman, S.J. (1981). Tests for independence in two-way contingency tables based on canonical correlation and on linear-by-linear interaction. *The Annals of Statistics*, **9**, 1178–1186.

Heiser, W.J. (1981). *Unfolding Analysis of Proximity Data*, Doctor of Social Sciences Thesis. Department of Data Theory, University of Leiden, The Netherlands.

Hill, M.O. (1973). Reciprocal averaging: an eigenvector method of ordination. *Journal of Ecology*, **61**, 237–251.

Hill, M.O. (1974). Correspondence analysis: a neglected multivariate method. *Applied Statistics*, **23**, 340–354.

Hill, M.O. & Gauch Jr., H.G. (1980). Detrended correspondence analysis: an improved ordination technique. *Vegetatio*, **42**, 47–58.

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

Hoffman, D., de Leeuw, J. & Arjunji, R. (1995). Multiple correspondence analysis. In *Advanced Methods of Marketing Research*, Ed. R.P. Bagozzi, pp. 260–294.

Hoffman, D.L. (1991). Review of four correspondence analysis programs for the IBM PC. *The American Statistician*, **45**, 305–311.

Horst, P. (1935). Measuring complex attitudes. *The Journal of Social Psychology*, **6**, 369–375.

Johnson, R.M. (1963). On a theorem stated by Eckart and Young. *Psychometrika*, **28**, 259–263.

Kendall, M.G. (1972). The history and future of statistics. In *Statistical Papers in Honor of George W. Snedecor*, Ed. T.A. Bancroft, pp. 193–210. The Iowa State University Press, Ames, Iowa.

Kim, H. (1992). Measures of influence in correspondence analysis. *Journal of Statistical Computation and Simulation*, **40**, 201–217.

Kim, H. (1994). Influence functions in multiple correspondence analysis. *Korean Journal of Applied Statistics*, **7**, 69–74.

Kroonenberg, P.M. & Lombardo, R. (1998). Nonsymmetric correspondence analysis: A tutorial. *Kwantitatieve Methoden*, **58**, 57–83.

Kroonenberg, P.M. & Lombardo, R. (1999). Nonsymmetric correspondence analysis: A tool for analysing contingency tables with a dependence structure. *Multivariate Behavioral Research*, **34**, 367–396.

Krzanowski, W.J. (1993). Attribute selection in correspondence analysis of incidence matrices. *Applied Statistics*, **42**, 529–541.

Lancaster, H.O. (1969). *The Chi-squared Distribution*. New York: Wiley.

Lauro, C. & Balbi, S. (1999). The analysis of structured qualitative data. *Applied Stochastic Models and Data Analysis*, **15**, 1–27.

Lebart, L. (1982). Exploratory analysis of large matrices with applications to textual data. In *COMSTAT 1982*, Eds. P.E.H. Caussinus and R. Tomassone, pp. 67–75. Wien: Physica-Verlag.

Lebart, L. & Morineau, A. (1982). SPAD—A system of FORTRAN programs for correspondence analysis. *Journal of Marketing Research*, **19**, 608–609.

Lebart, L., Morineau, A. & Tabard, N. (1977). *Techniques de la Description Statistique: methodes et logiciels pour l'analyse des grands tableaux*. Paris: Dunod.

Lebart, L., Morineau, A. & Warwick, K.M. (1984). *Multivariate Descriptive Statistical Analysis*. New York: Wiley.

Lombardo, R., Carlier, A. & D'Ambra, L. (1996). Nonsymmetric correspondence analysis for three-way contingency tables. *Methodologica*, **4**, 59–80.

Maignan, M.F. (1974). Correspondence factorial analysis. In *COMSTAT 1974*, Eds. G. Bruckmann, F. Ferschl and L. Schmetterer, pp. 234–243. Wien: Physica-Verlag.

Minchin, P.R. (1987). An evaluation of the relative robustness of techniques for ecological ordination. *Vegetatio*, **69**, 89–107.

Nishisato, S. (1980). *Analysis of Categorical Data: Dual Scaling and its Applications*. Toronto: University of Toronto Press.

Nishisato, S. (1994). *Elements of Dual Scaling: An Introduction to Practical Data Analysis*. Hillsdale, N.J.: L. Erlbaum Associates.

Okamoto, M. (1994a). Correspondence analysis of an artificial cylinder data. *Statistics & Probability Letters*, **20**, 101–112.

Okamoto, M. (1994b). Correspondence analysis of an artificial disk data. *Journal of the Japanese Statistical Society*, **24**, 157–168.

Okamoto, M. (1994c). Correspondence analysis of an artificial torus data. *Behaviormetrika*, **21**, 149–161.

Okamoto, M. (1995a). Correspondence analysis of artificial data based on non-regular symmetric polyhedron. *Mathematica Japonica*, **44**, 61–66.

Okamoto, M. (1995b). Correspondence analysis of some artificial data with multiple circular structure. *Mathematica Japonica*, **42**, 201–212.

Okamoto, M. (2000). Correspondence analysis of typical geometric figures. *Mathematica Japonica*, **51**, 145–152.

Okamoto, M. & Endo, H. (1995). Spherical trait in correspondence analysis of artificial data. *Journal of the Japan Statistical Society*, **25**, 181–191.

Oksanen, J. (1987). Problems of joint display of species and site scores in correspondence analysis. *Vegetatio*, **72**, 51–57.

Oksanen, J. (1988). A note on the occasional instability of detrending in correspondence analysis. *Vegetatio*, **74**, 29–32.

Oksanen, J. & Minchin, P.R. (1997). Instability of ordination results under changes in input data order: explanations and remedies. *Journal of Vegetation Science*, **8**, 447–454.

Pack, P. & Jolliffe, I.T. (1992). Influence in correspondence analysis. *Applied Statistics*, **41**, 365–380.

Palmer, M.N. (1993). Putting things in even better order: the advantage of canonical correspondence analysis. *Ecology*, **74**, 2215–2230.

Pearson, K. (1900). On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine (Series 5)*, **50**, 157–175.

Pearson, K. (1904). On the theory of contingency and its relation to association and normal correlations. *Drapers Company Research Memoirs (Biometric Series)*, **1**.

Pearson, K. (1906). On certain points connected with scale order in the case of a correlation of two characters for some arrangement give a linear regression line. *Biometrika*, **5**, 176–178.

Peet, R.K., Knox, R.G., Case, J.S. & Allen, R.B. (1988). Putting things in order: the advantages of detrended correspondence analysis. *American Naturalist*, **131**, 924–934.

Rayner, J.C.W. & Best, D.J. (1996). Smooth extensions of Pearsons product moment correlation and Spearmans Rho. *Statistics and Probability Letters*, **30**, 171–177.

Richardson, M. & Kuder, G.F. (1933). Making a rating scale that measures. *Personnel Journal*, **12**, 36–40.

Ritov, Y. & Gilula, Z. (1991). The order restricted RC model for ordered contingency tables: estimation and testing for fit. *Annals of Statistics*, **19**, 2090–2101.

Ritov, Y. & Gilula, Z. (1993). Analysis of contingency tables by correspondence models subject to ordered constraints. *Journal of the American Statistical Association*, **88**, 1380–1387.

Rom, D. & Sarkar, S.K. (1992). A generalised model for the analysis of association in ordinal contingency tables. *Journal of Statistical Planning and Inference*, **33**, 205–212.

Shi, M. & Carr, J.R. (2001). A modified code for R-mode correspondence analysis of large-scale problems. *Computers & Geosciences*, **27**, 139–146.

Teil, H. (1975). Correspondence factor analysis: An outline of its method. *Mathematical Geology*, **7**, 3–12.

Teil, H. & Cheminee, J.L. (1975). Application of correspondence factor analysis to the study of major and trace elements in the Erta Ale Chain (Afar, Ethiopia). *Mathematical Geology*, **7**, 13–30.

ter Braak, C.J.F. (1986). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, **67**, 1167–1179.

ter Braak, C.J.F. (1987). Ordination. In *Data Analysis in Community and Landscape Ecology*, Eds. R.H. Jongman, C.F.J. ter Braak and O.F.R. van Tongeren, pp. 91–173. Wageningen: Pudoc.

ter Braak, C.J.F. (1988). Partial canonical correspondence analysis. In *Classification and Related Methods of Data Analysis*, Ed. H.H. Bock, pp. 551–558. Amsterdam: North-Holland.

Thompson, P.A. (1995). Correspondence analysis in statistical package programs. *The American Statistician*, **49**, 310–316.

Tian, D.Q., Sorooshian, S. & Myers, D.E. (1993). Correspondence analysis with Matlab. *Computers and Geosciences*, **19**, 1007–1022.

van der Heijden, P. & de Leeuw, J. (1985). Correspondence analysis used complimentary to log-linear analysis. *Psychometrika*, **50**, 429–447.

van der Heijden, P. & Worsley, K.J. (1988). Comment on "Correspondence analysis used complimentary to log-linear analysis". *Psychometrika*, **53**, 287–291.

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.

van Heel, M. & Frank, J. (1980). Classification of particles in noisy electron micrographs using correspondence analysis. In *Pattern Recognition in Practice*, Eds. E.S. Gelsema and L.N. Kanal, pp. 235–243. Amsterdam: North-Holland.

van Kerm, P. (1998). Simple and multiple correspondence analysis in STATA. *STATA Technical Bulletin Reprints*, **10**, 210–217.

van Meter, K.M., Schiltz, M.-A., Cibois, P. & Mounier, L. (1994). Correspondence analysis: A history and French sociological perspective. In *Correspondence Analysis in the Social Sciences: Recent Developments and Applications*, Eds. M. Greenacre and J. Blasius, pp. 128–137. San Diego: Academic Press.

Wartenberg, D., Ferson, S. & Rohlf, F.J. (1987). Putting things in order: a critique of detrended correspondence analysis. *American Naturalist*, **129**, 434–448.

Williams, E.J. (1952). Use of scores for the analysis of association in contingency tables. *Biometrika*, **39**, 274–289.
Yamakawa, A., Ichihashi, H. & Miyoshi, T. (1998). Multiple correspondence analysis based on L_S-Norm and its application to an analysis of senior simulation. In *Proceedings of the 2nd Japan–Australia Joint Workshop on Intelligent and Evolutionary Systems*, pp. 99–106.
Yamakawa, A., Ichihashi, H. & Miyoshi, T. (1999). Multiple correspondence analysis of mimetic experiences of advanced aged persons. In *Proceedings of International Conference on Production Research*, Vol. **2**, pp. 1065–1068.
Yamakawa, A., Kanaumi, Y., Ichihashi, H. & Miyoshi, T. (1999). Simultaneous application of clustering and correspondence analysis. In *Proceedings of International Conference on Neural Networks*, No. **625**.

Résumé

Au cours des dernières décennies, l'analyse des correspondances a gagné une réputation internationale d'outil statistique puissant pour l'analyse graphique des tableaux de contingence. Cette popularité provient de son développement et de son application dans de nombreux pays Européens, particulièrement la France, et son utilisation s'est étendue à des pays anglophones tels que les Etats Unis et le Royaume Uni. Sa popularité croissante parmi les praticiens de la statistique, et plus récemment dans des disciplines où le rôle de la statistique est moins dominant, démontre l'importance de la recherche et du développement continuels sur la méthodologie. Le but de cet article est de souligner les aspects théoriques, pratiques et informatiques de l'analyse des correspondances simple et de discuter sa relation avec des avancées récentes qui peuvent tre utilisées pour représenter graphiquement l'association en données catégorielles à deux dimensions.

[*Received September 2002, accepted December 2003*]