

# A Unified Approach for the Multivariate Analysis of Contingency Tables

Carles M. Cuadras<sup>1</sup>, Daniel Cuadras<sup>2</sup>

<sup>1</sup>Department of Statistics, University of Barcelona, Barcelona, Spain

<sup>2</sup>Statistical Service, Sant Joan de Deu Research Foundation, Barcelona, Spain

Email: [cmcuadras@gmail.com](mailto:cmcuadras@gmail.com), [daniccuadras@gmail.com](mailto:daniccuadras@gmail.com)

Received 21 January 2015; accepted 22 April 2015; published 28 April 2015

Copyright © 2015 by authors and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

---

## Abstract

We present a unified approach to describing and linking several methods for representing categorical data in a contingency table. These methods include: correspondence analysis, Hellinger distance analysis, the log-ratio alternative, which is appropriate for compositional data, and the non-symmetrical correspondence analysis. We also present two solutions working with cumulative frequencies.

## Keywords

Correspondence Analysis, Hellinger Distance, Log-Ratio Analysis, Generalized Pearson Contingency Coefficient, Correspondence Analysis with Cumulative Frequencies

---

## 1. Introduction

In multivariate analysis, it is usual to link several methods in a closed expression, which depends on a set of parameters. Thus, in cluster analysis, some criteria (single linkage, complete linkage, median), can be unified by using parametric coefficients. The biplot analysis on a centered matrix  $X$ , is based on the singular value decomposition (SVD)  $X = U\Lambda V'$ . The general solution is  $X = U\Lambda^\alpha \Lambda^{1-\alpha} V'$  with  $0 \leq \alpha \leq 1$ , providing the GH, JK, SQ and other biplot types depending on  $\alpha$ . Also, some orthogonal rotations in factor analysis (varimax, quartimax) are particular cases of an expression depending on one or two parameters.

There are several methods for visualizing the rows and columns of a contingency table. These methods can be linked by using parameters and some well-known matrices. This parametric approach shows that correspondence analysis (CA), Hellinger distance analysis (HD), non-symmetric correspondence analysis (NSCA) and log-ratio analysis (LR), are particular cases of a general expression. In these methods, the decomposition of the inertia is used as well as a generalized version of Pearson contingency coefficient. With the help of triangular

matrices, it is also possible to perform two analyses, Taguchi's analysis (TA) and double accumulative analysis (DA), both based on cumulative frequencies. This paper unifies and extends some results by Cuadras and Greenacre [1]-[4].

## 2. Weighted Metric Scaling

A common problem in data analysis consists in displaying several objects as points in Euclidean space of low dimension.

Let  $\Omega = \{\omega_1, \dots, \omega_k\}$  be a set with  $k$  objects,  $\delta$  a distance function on  $\Omega$  providing the  $k \times k$  Euclidean distance matrix  $\Delta_k = (\delta_{ij})$ , where  $\delta_{ij} = \delta(\omega_i, \omega_j)$ . Let  $\mathbf{w} = (w_1, \dots, w_k)'$  a weight vector such that  $\mathbf{w}'\mathbf{1} = \sum_{i=1}^k w_i = 1$  with  $w_i > 0$  and  $\mathbf{1}$  the column vector of ones.

The weighted metric scaling (WMS) solution using  $\Delta_k$  finds the spectral decomposition

$$D_w^{1/2} (I - \mathbf{1}\mathbf{w}') \left( -\frac{1}{2} \Delta_k^{(2)} \right) (I - \mathbf{w}\mathbf{1}') D_w^{1/2} = U \Lambda^2 U', \quad (1)$$

where  $I$  is the identity matrix,  $\Delta_k^{(2)} = (\delta_{ij}^2)$ ,  $\Lambda^2$  is  $p \times p$  diagonal with  $p$  positive eigenvalues arranged in descending order,  $U$  is  $k \times p$  such that  $U'U = I$ , and  $D_w = \text{diag}(\mathbf{w})$  [5].

The  $k \times p$  matrix  $X = D_w^{-1/2} U \Lambda$  contains the principal coordinates of  $\Omega$ , which can be represented as a configuration of  $k$  points  $P_1, \dots, P_k$  in Euclidean space. This means that the Euclidean distance between the points  $P_i, P_j$  with coordinates the rows  $\mathbf{x}_i, \mathbf{x}_j$  of  $X$ , equals  $\delta_{ij}$ .

The geometric variability of  $\Omega$  with respect to  $\delta$  is defined by

$$V_\delta = \frac{1}{2} \sum_{i,j=1}^k w_i \delta_{ij}^2 w_j = \frac{1}{2} \mathbf{w}' \Delta_k^{(2)} \mathbf{w}.$$

The geometric variability (also called inertia) can be interpreted as a generalized variance [6].

If  $G = XX'$  and  $\mathbf{g}$  is the column vector with the diagonal entries in  $G$ , then  $\Delta_k^{(2)} = \mathbf{g}\mathbf{1}' + \mathbf{1}\mathbf{g}' - 2G$ . Since  $\mathbf{w}'X = \mathbf{0}$  and  $\mathbf{w}'\mathbf{1} = 1$ , we have  $\mathbf{g}'\mathbf{w} = \text{tr}(D_w^{1/2} G D_w^{1/2}) = \text{tr}(U \Lambda^2 U') = \text{tr}(\Lambda^2)$ . Thus, if  $k' = \text{rank}(\Lambda^2)$ , the geometric variability is

$$V_\delta = \sum_{i=1}^{k'} \lambda_i^2.$$

We should use the first  $m$  columns of  $X$  to represent the  $k$  objects in low dimension  $m$ , usually  $m = 2$ . This provides an optimal representation, in the sense that the geometric variability taking  $m \leq k$  first dimensions is  $V_\delta(m) = \sum_{i=1}^m \lambda_i^2$  and this quantity is maximum.

## 3. Parametric Analysis of Contingency Tables

Let  $N = (n_{ij})$  be an  $I \times J$  contingency table and  $P = n^{-1}N$  the correspondence matrix, where  $n = \sum_{ij} n_{ij}$ . Let  $K = \min\{I, J\}$  and  $\mathbf{r} = P\mathbf{1}$ ,  $D_r = \text{diag}(\mathbf{r})$ ,  $\mathbf{c} = P'\mathbf{1}$ ,  $D_c = \text{diag}(\mathbf{c})$ , the vectors and diagonal matrices with the marginal frequencies of  $P$ . In order to represent the rows and columns of  $N$ , Goodman [7] introduces the generalized non-independence analysis (GNA) by means of the SVD:

$$D_r^{1/2} (I - \mathbf{1}\mathbf{r}') \cdot R(D_r^{-1} P D_c^{-1}) \cdot D_c^{1/2} = U \Lambda V',$$

where  $\Lambda$  is diagonal with the singular values in descending order, and  $U, V$  are matrices of appropriate order with  $U'U = I$ , and  $V$  orthogonal.  $R(x)$ , with  $x > 0$ , is any monotonically increasing function. Here  $R(M)$  with  $M = (m_{ij})$ , means  $(R(m_{ij}))$ . The principal coordinates for rows and columns are given by  $A = D_r^{-1/2} U \Lambda$ ,  $B = D_c^{-1/2} V \Lambda$ . Clearly GNA reduces to CA when  $R(x) = 1$ .

A suitable choice of  $R(x)$  is the Box-Cox transformation

$$R(x) = \begin{cases} (x^\alpha - 1)/\alpha, & \text{if } \alpha > 0; \\ \ln(x), & \text{if } \alpha = 0. \end{cases}$$

With this transformation, let us consider the following SVD depending on three parameters:

$$D_r^{1/2} (I - \gamma \mathbf{1r}') \left\{ \frac{1}{\alpha} \left[ (D_r^{-1} P D_c^{-1})^{(\alpha)} - \mathbf{11}' \right] \right\} D_c^\beta = U \Lambda V', \tag{2}$$

where  $M^{(\alpha)} = (m_{ij}^\alpha)$  and  $0 \leq \alpha, \beta \leq 1$ . Then the principal coordinates for the  $I$  rows and the standard coordinates for the  $J$  columns of  $N$  are given by  $A = D_r^{-1/2} U \Lambda$  and  $B_* = D_c^{-\beta} V$ , respectively, in the sense that these coordinates reconstitute the model:

$$(I - \gamma \mathbf{1r}') \left\{ \frac{1}{\alpha} \left[ (D_r^{-1} P D_c^{-1})^{(\alpha)} - \mathbf{11}' \right] \right\} = AB_*'.$$

However, different weights are used for the column representation, e.g.,  $B = D_c^\beta V \Lambda$ . Implicit with this (row) representation is the squared distance between rows

$$\delta_{ii'}^2 = \sum_{j=1}^J \left[ \left( \frac{p_{ij}}{r_i c_j} \right)^\alpha - \left( \frac{p_{i'j}}{r_{i'} c_j} \right)^\alpha \right]^2 c_j^{2\beta}. \tag{3}$$

The first principal coordinates account for a relative high percentage of inertia, see Section 2. This parametric approach satisfies the principle of distributional equivalence and has been explored by Cuadras and Cuadras [2] and Greenacre [4]. Here we use Greenacre’s parametrization.

The geometric variability for displaying rows, is the average of the distances weighted by the row marginal frequencies:

$$V_\delta = \frac{1}{2} \mathbf{r}' \Delta^{(2)} \mathbf{r},$$

where  $\Delta^{(2)} = (\delta_{ii'}^2)$  is the  $I \times I$  matrix of squared parametric distances (3).

For measuring the dispersion in model (2), let us introduce the generalized Pearson contingency coefficient

$$\phi^2(\alpha, \beta) = \sum_{i=1}^I \sum_{j=1}^J \left[ \left( \frac{p_{ij}}{r_i c_j} \right)^\alpha - 1 \right]^2 r_i c_j^{2\beta}.$$

Note that  $V_\delta = \phi^2(\alpha, \beta) = 0$  if  $P = \mathbf{rc}'$ , i.e., under “statistical independence” between row and column variables. In general  $V_\delta \neq \phi^2(\alpha, \beta)$ .

The unified approach for all methods (centered and uncentered) discussed below, are given in Table 1. It is worth noting that, from

$$(I - \mathbf{1r}') (D_r^{-1} P D_c^{-1} - \mathbf{11}') = D_r^{-1} P D_c^{-1} - \mathbf{11}', \tag{4}$$

the centered ( $\gamma = 1$ ) and uncentered ( $\gamma = 0$ ) solutions coincide in CA, NSCA and TA (Taguchi’s analysis, see below).

To give a WMS approach compatible with (1), we mainly consider generalized versions without right-centering, i.e., post-multiplying  $\left[ (D_r^{-1} P D_c^{-1})^{(\alpha)} - \mathbf{11}' \right]$  by  $(I - \mathbf{c1}')$ . In fact, we can display columns in the same

**Table 1.** Four methods for representing rows and columns in a contingency table.

Method	Uncentered		Centered	
	$\gamma = 0$		$\gamma = 1$	
	$\alpha$	$\beta$	$\alpha$	$\beta$
CA correspondence analysis	1	1/2	1	1/2
HD Hellinger distance analysis	1/2	1/2	1/2	1/2
NSCA non-symmetric CA	1	1	1	1
LR Log-ratio analysis	0	1/2	0	1/2

graph of rows without applying this post-multiplication. To do this compute the SVD  $(H_I Q)'(H_I A) = RDS'$  with  $D$  diagonal and  $H_I$  the unweighted  $I \times I$  centering matrix. Then  $(H_I Q) = (H_I A)RS$  and if we take principal coordinates  $H_I A$  for the rows, and identify each column as the dummy row profile  $(0, \dots, 0, 1, 0, \dots, 0)$ , then the centered projection  $B = H_I RS'$  provides standard coordinates for the columns, see [2] [3].

#### 4. Testing Independence

Suppose that the rows and columns of  $N = (n_{ij})$  are two sets of categorical variables with  $I$  and  $J$  states, and that  $n_{ij}$  is the observed frequencies of the corresponding combination, according to a multinomial model. Assuming  $\beta = 1/2$ , the test for independence between row and column variables can be performed with  $\phi^2(\alpha, 1/2)$ . Under independence we have, as  $n \rightarrow \infty$ ,  $(n/\alpha^2)\phi^2(\alpha, 1/2) \rightarrow \chi^2_{(I-1)(J-1)}$  if  $\alpha > 0$ , and  $n\phi^2(0, 1) \rightarrow \chi^2_{(I-1)(J-1)}$  if  $\alpha = 0$ , where  $\chi^2_{(I-1)(J-1)}$  is the chi-square distribution with  $(I-1)(J-1)$  d.f. The convergence is in law.

To prove this asymptotic result, suppose  $\alpha > 0$  a fix value. Let  $x = p_{ij}/(r_i c_j)$ . From  $[p_{ij}/r_i c_j - 1]^2 r_i c_i = (p_{ij} - r_i c_j)^2 / (r_i c_j)$  we get

$$(x^\alpha - 1)^2 r_i c_j = \left(\frac{x^\alpha - 1}{x - 1}\right)^2 (x - 1)^2 r_i c_j.$$

But  $\lim_{x \rightarrow 1} [(x^\alpha - 1)/(x - 1)]^2 = \alpha^2$ . Hence, under independence,  $x \rightarrow 1$  as  $n \rightarrow \infty$ . Thus

$$\begin{aligned} \lim_{n \rightarrow \infty} n \sum_{i=1}^I \sum_{j=1}^J \left[ \left( \frac{p_{ij}}{r_i c_j} \right)^\alpha - 1 \right]^2 r_i c_i &= \alpha^2 \lim_{n \rightarrow \infty} n \sum_{i=1}^I \sum_{j=1}^J \left( \frac{p_{ij} - r_i c_j}{r_i c_j} \right)^2 r_i c_j \\ &= \alpha^2 \chi^2_{(I-1)(J-1)}. \end{aligned}$$

If  $\alpha \rightarrow 0$  then  $\lim_{x \rightarrow 1, \alpha \rightarrow 0} \frac{1}{\alpha^2} \left( \frac{x^\alpha - 1}{x - 1} \right)^2 = 1$  and the above limit reduces to  $n\phi^2(0, 1) \rightarrow \chi^2_{(I-1)(J-1)}$ .

#### 5. Correspondence Analysis

In this and the following sections, we present several methods of representation, distinguishing, when it is necessary, the centered from the uncentered solution. The inertia is given by the geometric variability and the generalized Pearson coefficient, respectively.

Centered and Uncentered ( $\alpha = 1, \beta = 1/2$ )

$$D_r^{1/2} (D_r^{-1} P D_c^{-1} - \mathbf{1}\mathbf{1}') D_c^{1/2} = U \Lambda V'.$$

- 1) Chi-square distance between rows:  $\delta_{ii'}^2 = \sum_{j=1}^J \left( \frac{p_{ij}}{r_i} - \frac{p_{i'j}}{r_{i'}} \right)^2 \frac{1}{c_j}$ .
- 2) Rows and columns coordinates:  $A = D_r^{-1/2} U \Lambda$ ,  $B = D_c^{-1/2} V \Lambda$ .
- 3) Inertia:  $\phi^2(1, 1/2) = V_\delta = \sum_{i=1}^I \sum_{j=1}^J \left( \frac{p_{ij}}{r_i c_j} - 1 \right)^2 r_i c_j$ .

Some authors considered CA the most rational method for analyzing contingency tables, because its ability to display in a meaningful way the relationships between the categories of two variable [8]-[10]. For the history of CA, see [11], and for a continuous extension, see [12] [13]. CA can be understood as the first order approximation to the alternatives HD and LR given below [3]. Besides, LR would be a limiting case of parametric CA [14].

#### 6. Hellinger Distance Analysis

Centered ( $\alpha = 1/2, \beta = 1/2, \gamma = 1$ ), Uncentered ( $(\alpha = 1/2, \beta = 1/2, \gamma = 0)$ )

$$\text{Centered} \quad D_r^{1/2} (I - \mathbf{1r}') (D_r^{-1/2} P^{(1/2)} D_c^{-1/2} - \mathbf{11}') D_c^{1/2} = U \Lambda V'.$$

$$\text{Uncentered} \quad D_r^{1/2} (D_r^{-1/2} P^{(1/2)} D_c^{-1/2} - \mathbf{11}') D_c^{1/2} = U \Lambda V'.$$

$$1) \text{ Hellinger distance between rows: } \delta_{ii'}^2 = \sum_{j=1}^J \left( \sqrt{p_{ij}/r_i} - \sqrt{p_{i'j}/r_{i'}} \right)^2.$$

$$2) \text{ Rows and columns coordinates: } A = D_r^{-1/2} U \Lambda, \quad B_* = D_c^{-1/2} V.$$

$$3) \text{ Inertia: } V_\delta = 1 - r' D_r^{-1/2} P^{1/2} P^{1/2} D_r^{-1/2} r = 1 - \sum_{j=1}^J \left( \sum_{i=1}^I \sqrt{p_{ij} r_i} \right)^2,$$

$$\phi^2(1/2, 1/2) = 2 \left( 1 - \sum_{i=1}^I \sum_{j=1}^J \sqrt{p_{ij} r_i c_j} \right).$$

Although the distances between rows are the same, the principal coordinates in the centered and uncentered solutions are distinct. Note that  $\left( \sum_{i,j} \sqrt{p_{ij} r_i c_j} \right)$  is the so-called affinity coefficient and that  $V_\delta < \phi^2(1/2, 1/2)$ . HD is suitable when we are comparing several multinomial populations and the column profiles should not have influence on the distance. See [15] [16].

## 7. Non-Symmetric Correspondence Analysis

Centered and Uncentered ( $\alpha = 1, \beta = 1$ )

$$D_r^{1/2} (D_r^{-1} P D_c^{-1} - \mathbf{11}') D_c = U \Lambda V'.$$

$$1) \text{ Distance between rows: } \delta_{ii'}^2 = \sum_{j=1}^J \left( \frac{p_{ij}}{r_i} - \frac{p_{i'j}}{r_{i'}} \right)^2.$$

$$2) \text{ Rows and columns coordinates: } A = D_r^{-1/2} U \Lambda, \quad B = V \Lambda.$$

$$3) \text{ Inertia: } \phi^2(1, 1) = V_\delta = \sum_{i=1}^I \sum_{j=1}^J \left( \frac{p_{ij}}{r_i} - c_j \right)^2 r_i.$$

Note that  $V_\delta$  is related to the Goodman-Kruskal coefficient  $\tau$  in a contingency table. This measure is

$$\tau = \frac{\sum_{i=1}^I \sum_{j=1}^J \left( \frac{p_{ij}}{r_i} - c_j \right)^2 r_i}{1 - \sum_{i=1}^I r_i^2}.$$

The numerator of  $\tau$  represents the overall predictability of the columns given the rows. Thus NSCA may be useful when a categorical variable plays the role of response depending on a predictor variable, see [17]-[19].

## 8. Log-Ratio Analysis

Centered ( $\alpha = 0, \beta = 1/2, \gamma = 1$ ), Uncentered ( $\alpha = 0, \beta = 1/2, \gamma = 0$ )

$$\text{Centered} \quad D_r^{1/2} (I - \mathbf{1r}') \ln(D_r^{-1} P D_c^{-1}) D_c^{1/2} = U \Lambda V'.$$

$$\text{Uncentered} \quad D_r^{1/2} \ln(D_r^{-1} P D_c^{-1}) D_c^{1/2} = U \Lambda V'.$$

$$1) \text{ Log-ratio distance between rows: } \delta_{ii'}^2 = \sum_{j=1}^J c_j \left( \ln \frac{p_{ij}}{r_i} - \ln \frac{p_{i'j}}{r_{i'}} \right)^2.$$

$$2) \text{ Rows and columns coordinates: } A = D_r^{-1/2} U \Lambda, \quad B_* = D_c^{-1/2} V \Lambda.$$

$$3) \text{ Inertia: } V_\delta = \sum_{j=1}^J c_j \left[ \sum_{i=1}^I r_i \left( \ln \frac{p_{ij}}{r_i} \right)^2 - \left( \sum_{i=1}^I r_i \ln \frac{p_{ij}}{r_i} \right)^2 \right],$$

$$\phi^2(0,1/2) = \sum_{i=1}^I \sum_{j=1}^J \left( \ln \frac{p_{ij}}{r_i c_j} \right)^2 r_i c_j.$$

In spite of having the same distances, the principal coordinates (centered and uncentered) are different. Note that  $V_\delta < \phi^2(0,1/2)$ . This method satisfies the principle of subcompositional coherence and is appropriate for positive compositional data [20].

The inertia and the geometric variability in these four methods, as well as Taguchi’s method given in Section 2, are summarized in **Table 2**. For a comparison between CA, HD, and LR see [3] [21]. Besides, by varying the parameters there is the possibility of a dynamic presentation linking these methods [22].

### 9. Double-Centered Log-Ratio Analysis

In LR analysis Lewi [23] and Greenacre [4] considered the weighted double-centered solution

$$D_r^{1/2} (I - \mathbf{1r}') \ln(D_r^{-1} P D_c^{-1}) (I - \mathbf{1c}')' D_c^{1/2} = U \Lambda V'$$

called “spectral map”. The unweighted double-centered solution, called “variation diagram”, was considered by Aitchison and Greenacre [20]. They show that log-ratio and centered log-ratio biplots are equivalent. In this solution the role of rows and columns is symmetric.

### 10. Analysis Based on Cumulative Frequencies

Let  $N = (n_{ij})$  be the  $I \times J$  contingency table,  $n_{i\cdot}$  and  $n_{\cdot j}$  the row and column marginals. Given a row  $i$  let us consider the cumulative frequencies

$$z_{i1} = n_{i1}, \quad z_{i2} = n_{i1} + n_{i2}, \quad \dots, \quad z_{iJ} = n_{i1} + \dots + n_{iJ},$$

and cumulative column proportions

$$d_1 = \frac{n_{\cdot 1}}{n}, \quad d_2 = \frac{n_{\cdot 1} + n_{\cdot 2}}{n}, \quad \dots, \quad d_J = \frac{n_{\cdot 1} + \dots + n_{\cdot J}}{n}.$$

The Taguchi’s statistic [24], is given by

$$T = \sum_{j=1}^{J-1} w_j \left( \sum_{i=1}^I n_i \left( \frac{z_{ij}}{n_i} - d_j \right)^2 \right),$$

**Table 2.** Inertia expressions for five methods for representing rows in contingency tables. In CA and NSCA the geometric variability coincides with the contingency coefficient. This coefficient does not apply in TA.

Method	Inertia (centered) $V_\delta = \sum \lambda_i^2$	Inertia (uncentered) $\phi^2(\alpha, \beta) = \sum \lambda_i^2$
CA Benzécri-Greenacre-Lebart	$\sum_{i,j} \left( \frac{p_{ij}}{r_i c_j} - 1 \right)^2 r_i c_j$	$\phi^2(1,1/2) = V_\delta$
HD Domenge-Volle-Rao	$1 - \sum_i \left( \sum_j \sqrt{p_{ij} r_i} \right)^2$	$\phi^2(1/2,1/2) = 2 \left( 1 - \sum_{i,j} \sqrt{p_{ij} r_i c_j} \right)$
NSCA Lauro-D’Ambra	$\sum_{i,j} \left( \frac{p_{ij}}{r_i} - c_j \right)^2 r_i$	$\phi^2(1,1) = V_\delta$
LR Aitchison-Greenacre	$\sum_{j=1}^J c_j \left[ \sum_{i=1}^I r_i \left( \ln \frac{p_{ij}}{r_i} \right)^2 - \left( \sum_{i=1}^I r_i \ln \frac{p_{ij}}{r_i} \right)^2 \right]$	$\phi^2(0,1/2) = \sum_{i,j} \left( \ln \frac{p_{ij}}{r_i c_j} \right)^2 r_i c_j$
TA Beh-D’Ambra-Simonetti	$\sum_{i,j} \frac{w_j (p_{ij} - r_i c_j)^2}{r_i}$	Same $V_\delta$

where  $w_1, \dots, w_{J-1}$  are weights. Two choices are possible:  $w_j = [d_j(1-d_j)]^{-1}$  and  $w_j = 1/J$ . The test based on  $T$  is better than Pearson chi-square when there is an order in the categories of the rows or columns of the contingency table [25].

The so-called Taguchi's inertia  $T_a = T/n$  is

$$\begin{aligned} T_a &= \sum_{j=1}^{J-1} w_j \left( \sum_{i=1}^I r_i \left( \frac{z_{ik}/n}{r_i} - d_j \right)^2 \right) \\ &= \sum_{j=1}^{J-1} w_j \left( \sum_{i=1}^I \left( \frac{z_{ik}}{n} - r_i d_j \right)^2 \frac{1}{r_i} \right). \end{aligned}$$

By using  $\mathbf{d} = (d_1, d_2, \dots)'$  and the  $J \times J$  triangular matrix

$$M = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 1 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 1 & 1 & 1 & 1 \end{pmatrix},$$

then  $\mathbf{d} = M\mathbf{c}$  and  $(z_{ik}/n) = PM'$ . Thus  $T_a$  depends on  $(PM' - \mathbf{rd}') = (P - \mathbf{rc}')M'$  and can be expressed as

$$T_a = \text{tr} \left( D_r^{-1/2} (P - \mathbf{rc}') M' W M (P - \mathbf{rc}')' D_r^{-1/2} \right).$$

As it occurs in CA, where the inertia is the trace  $\text{tr}(QQ')$  with  $Q = D_r^{-1/2} (P - \mathbf{rc}') D_r^{-1/2}$ , Beh *et al.* [26] considered the decomposition of Taguchi's inertia. In our matrix notation, using the above  $M$ , we have

$$D_r^{1/2} (D_r^{-1} P D_c^{-1} - \mathbf{11}') D_c M' W^{1/2} = U \Lambda V'.$$

From (4), centering is not necessary here. This SVD provides an alternative for visualizing the rows and columns of  $N$ . The main aspects of this solution, where  $P_{ij} = p_{i1} + \dots + p_{ij}$  is the cumulative sum for row  $i$  and  $C_j = c_1 + \dots + c_j$ , are:

- 1) Distance between rows:  $\delta_{ii'}^2 = \sum_{j=1}^J w_j \left( \frac{P_{ij}}{r_i} - \frac{P_{i'j}}{r_{i'}} \right)^2$ .
- 2) Rows and columns coordinates:  $A = D_r^{-1/2} U \Lambda$ ,  $B = W^{-1/2} V \Lambda$ .
- 3) Inertia:

$$T_a = \sum_{i=1}^I \sum_{j=1}^J \frac{w_j (P_{ij} - r_i C_j)^2}{r_i} = \sum_{i=1}^K \lambda_i^2,$$

where  $K = \min\{I, J\}$ .

There is a formal analogy between  $T_a$  and the Goodman-Kruskal coefficient  $\tau$ . Also note that the last column in  $PM'$  and  $\mathbf{rc}'$  are equal, so in  $T_a$  the index  $j$  can run from 1 to  $J-1$ .

## 11. Double Acumulative Frequencies

More generally, the analysis of a contingency table  $N$  may also be approached by using cumulative frequencies for rows and columns. Thus an approach based on double accumulative (DA) frequencies is

$$D_r^{-1/2} L (P - \mathbf{rc}') M' W^{1/2} = D_r^{-1/2} (H - RC') W^{1/2} = U \Lambda V',$$

where  $L$  is a suitable triangular matrix with ones. Clearly matrices  $H = LPM'$ ,  $R = L\mathbf{r}$ ,  $C = M\mathbf{c}$  contain the cumulative frequencies [1]. However, both cumulative approaches TA and DA may not provide a clear display of the contingency table.

Finally, from

$$D_r \left[ (D_r^{-1} P D_c^{-1})^{(\alpha)} - \mathbf{11}' \right] D_c = D_r^{1-\alpha} P^{(\alpha)} D_c^{1-\alpha} - \mathbf{rc}',$$

all (uncentered) methods CA, HD, NSCA, LR, TA and DA can be unified by means of the SVD

$$D_r^{-1/2} L \left\{ \frac{1}{\alpha} \left[ D_r^{1-\alpha} P^{(\alpha)} D_c^{1-\alpha} - \mathbf{rc}' \right] \right\} M W^{1/2} = U \Lambda V'$$

as it is reported in **Table 3**. If  $\alpha = 1$ , we suppose  $0^{1-\alpha} = 0$  in the null entries of  $D_r^{1-\alpha}$  and  $D_c^{1-\alpha}$ .

### 12. An Example

The data in **Table 4** is well known. This table combines the hair and eye colour of 5383 individuals. We present the first two principal coordinates (centered solution) of the five hair colour categories for CA, HD, LR and NSCA. We multiply the NSCA solution (denoted by  $NS$ ) by 2 for comparison purposes.

$$\begin{aligned}
 CA &= \begin{bmatrix} -0.5437 & -0.1722 \\ -0.2324 & -0.0477 \\ -0.0402 & 0.2079 \\ 0.5899 & -0.1070 \\ 1.0784 & -0.2743 \end{bmatrix}, & HD &= \begin{bmatrix} -0.5776 & -0.1368 \\ -0.2145 & -0.0416 \\ -0.0139 & 0.1791 \\ 0.5818 & -0.1057 \\ 1.0711 & -0.2182 \end{bmatrix}, \\
 LR &= \begin{bmatrix} -0.6501 & -0.1367 \\ -0.1971 & 0.0282 \\ 0.0073 & 0.1654 \\ 0.6039 & -0.0830 \\ 1.2866 & -0.4127 \end{bmatrix}, & NS &= \begin{bmatrix} -0.5356 & -0.1841 \\ -0.2517 & -0.0726 \\ -0.0413 & 0.2246 \\ 0.5881 & -0.1128 \\ 1.0649 & -0.3018 \end{bmatrix}.
 \end{aligned}$$

These four solutions are similar.

Finally, we show the first two coordinates for Taguchi's and double accumulative solutions ( $\alpha = 1$ ), but multiplying by 3 for comparison purposes.

**Table 3.** Correspondence analysis, Hellinger analysis, non-symmetric correspondence analysis, log-ratio analysis and two solutions based on cumulative frequencies. The right column suggests the type of categorical data.

SVD $D_r^{-1/2} L \left\{ \left[ D_r^{1-\alpha} P^{(\alpha)} D_c^{1-\alpha} - \mathbf{rc}' \right] / \alpha \right\} M W^{1/2} = U \Lambda V'$					
Method	$\alpha$	$L$	$M$	$W$	Suitable in case of
CA	1	Identity	Identity	$D_c^{-1}$	Relating two variables
HD	1/2	Identity	Identity	$D_c^{-1}$	Multinomial populations
NSCA	1	Identity	Identity	Identity	Responses/predictors
LR	0	Identity	Identity	$D_c^{-1}$	Compositional data
TA	1	Identity	Triangular	Weight	One ordinal variable
DA	1	Triangular	Triangular	Weight	Two ordinal variables

**Table 4.** Classification of a large sample of people combining the hair colour and the eye colour.

Eye colour	Fair	Red	Hair medium	Colour dark	Black	Total
Light	688	116	584	188	4	1580
Blue	326	38	241	110	3	718
Medium	343	84	909	412	26	1774
Dark	98	48	403	681	81	1311
Total	1455	286	2137	1391	114	5383



$$TA = \begin{bmatrix} -0.5481 & -0.0760 \\ -0.2555 & -0.0424 \\ 0.0056 & 0.1070 \\ 0.5389 & -0.0625 \\ 0.9559 & -0.1658 \end{bmatrix}, \quad DC = \begin{bmatrix} -0.5532 & -0.0134 \\ -3.0731 & -0.0812 \\ -0.3936 & 0.0948 \\ -0.0763 & 0.0224 \\ 0.0000 & 0.0000 \end{bmatrix}.$$

Both solutions are quite distinct from the previous ones.

## References

- [1] Cuadras, C.M. (2002) Correspondence Analysis and Diagonal Expansions in Terms of Distribution Functions. *Journal of Statistical Planning and Inference*, **103**, 137-150. [http://dx.doi.org/10.1016/S0378-3758\(01\)00216-6](http://dx.doi.org/10.1016/S0378-3758(01)00216-6)
- [2] Cuadras, C.M. and Cuadras, D. (2006) A Parametric Approach to Correspondence Analysis. *Linear Algebra and its Applications*, **417**, 64-74. <http://dx.doi.org/10.1016/j.laa.2005.10.029>
- [3] Cuadras, C.M., Cuadras, D. and Greenacre, M. (2006) A Comparison of Different Methods for Representing Categorical Data. *Communications in Statistics-Simulation and Computation*, **35**, 447-459. <http://dx.doi.org/10.1080/03610910600591875>
- [4] Greenacre, M. (2009) Power Transformations in Correspondence Analysis. *Computational Statistics and Data Analysis*, **53**, 3107-3116. <http://dx.doi.org/10.1016/j.csda.2008.09.001>
- [5] Cuadras, C.M. and Fortiana, J. (1996) Weighted Continuous Metric Scaling. In: Gupta, A.K. and Girko, V.L., Eds., *Multidimensional Statistical Analysis and Theory of Random Matrices*, VSP, The Netherlands, 27-40.
- [6] Cuadras, C.M., Fortiana, J. and Oliva, F. (1997) The Proximity of an Individual to a Population with Applications in Discriminant Analysis. *Journal of Classification*, **14**, 117-136. <http://dx.doi.org/10.1007/s003579900006>
- [7] Goodman, L.A. (1993) Correspondence Analysis, Association Analysis, and Generalized Nonindependence Analysis of Contingency Tables: Saturated and Unsaturated Models, and Appropriate Graphical Displays. In: Cuadras, C.M. and Rao, C.R., Eds., *Multivariate Analysis: Future Directions 2*, Elsevier, Amsterdam, 265-294.
- [8] Beh, E.J. (2004) Simple Correspondence Analysis: A Bibliographic Review. *International Statistical Review*, **72**, 257-284.
- [9] Benzecri, J.-P. (1976) L'Analyse des Donnees. II. L'Analyse des Correspondances. Deuxieme Edition. Dunod, Paris.
- [10] Greenacre, M.J. (1984) Theory and Applications of Correspondence Analysis. Academic Press, London. <http://www.carme-n.org/?sec=books5>
- [11] Lebart, L. and Saporta, G. (2014) Historical Elements of Correspondence Analysis and Multiple Correspondence Analysis. In: Blasius, J. and Greenacre, M., Eds., *Visualization and Verbalization of Data*, CRC Press, Taylor & Francis Group, New York, 31-44.
- [12] Cuadras, C.M., Fortiana, J. and Greenacre, M. (2000) Continuous Extensions of Matrix Formulations in Correspondence Analysis, with Applications to the FGM Family of Distributions. In: Heijmans, R.D.H., Pollock, D.S.G. and Satorra, A., Eds., *Innovations in Multivariate Statistical Analysis*, Kluwer Academic Publishers, Dordrecht, 101-116. [http://dx.doi.org/10.1007/978-1-4615-4603-0\\_7](http://dx.doi.org/10.1007/978-1-4615-4603-0_7)
- [13] Cuadras, C.M. (2014) Nonlinear Principal and Canonical Directions from Continuous Extensions of Multidimensional Scaling. *Open Journal of Statistics*, **4**, 132-149. <http://dx.doi.org/10.4236/ojs.2014.42015>
- [14] Greenacre, M. (2010) Log-Ratio Analysis Is a Limiting Case of Correspondence Analysis. *Mathematical Geosciences*, **42**, 129-134. <http://dx.doi.org/10.1007/s11004-008-9212-2>
- [15] Domenges, D. and Volle, M. (1979) Analyse Factorielle Spherique: Une Exploration. *Annales de L'INSEE*, **35**, 3-84.
- [16] Rao, C.R. (1995) A Review of Canonical Coordinates and an Alternative to Correspondence Analysis Using Hellinger Distance. *Questio*, **19**, 23-63.
- [17] Beh, E.J. and D'Ambra, L. (2009) Some Interpretative Tools for Non-Symmetrical Correspondence Analysis. *Journal of Classification*, **26**, 55-76. <http://dx.doi.org/10.1007/s00357-009-9025-0>
- [18] Kroonenberg, P.M. and Lombardo, R. (1999) Nonsymmetric Correspondence Analysis: A Tool for Analyzing Contingency Tables with a Dependence Structure. *Multivariate Behavioral Research*, **34**, 367-396. [http://dx.doi.org/10.1207/S15327906MBR3403\\_4](http://dx.doi.org/10.1207/S15327906MBR3403_4)
- [19] Lauro, N. and D'Ambra, L. (1984) L'analyse non symetrique des correspondances. In: Diday, E., Jambu, M., Lebart, L., Pages, J. and Tomassone, R., Eds., *Data Analysis and Informatics III*, North Holland, Amsterdam, 433-446.
- [20] Aitchison, J. and Greenacre, M. (2002) Biplots of Compositional Data. *Applied Statistics*, **51**, 375-392.

- <http://dx.doi.org/10.1111/1467-9876.00275>
- [21] Greenacre, M. and Lewi, P. (2009) Distributional Equivalence and Subcompositional Coherence in the Analysis of Contingency Tables, Ratio-Scale Measurements and Compositional Data. *Journal of Classification*, **26**, 29-54. <http://dx.doi.org/10.1007/s00357-009-9027-y>
- [22] Greenacre, M. (2008) Dynamic Graphics of Parametrically Linked Multivariate Methods Used in Compositional Data Analysis. Universitat Pompeu Fabra, Barcelona. <http://www.econ.upf.edu/en/research/onepaper.php?id=1082>
- [23] Lewi, P.J. (1976) Spectral Mapping, a Technique for Classifying Biological Activity Profiles of Chemical Compounds. *Arzneimittel Forschung—Drug Research*, **26**, 1295-1300.
- [24] Taguchi, G. (1974) A New Statistical Analysis for Clinical Data, the Accumulating Analysis in Contrast with the Chi-Square Test. *Saishin Igaku (The New Medicine)*, **20**, 806-813.
- [25] Nair, V.N. (1987) Chi-Square Type Tests for Ordered Categories in Contingency Tables. *Journal of the American Statistical Association*, **82**, 283-291. <http://dx.doi.org/10.1080/01621459.1987.10478431>
- [26] Beh, E.J., D'Ambra, L. and Simonetti, B. (2011) Correspondence Analysis of Cumulative Frequencies Using a Decomposition of Taguchi's Statistic. *Communications in Statistics-Theory and Methods*, **40**, 1620-1632. <http://dx.doi.org/10.1080/03610921003615880>