Dr. Strangelove or: How I Learned to Stop Worrying and Love the Correspondence Analysis

Jörg Breitung*
University of Cologne

January 30, 2023

Abstract

This paper analyses the properties of alternative distance metrics for analysing contingency tables. It is argued that the usual χ^2-metric possesses some undesirable features and it is therefore interesting to consider alternative metrics. Furthermore it is shown that the "standardized residuals" used as a starting point for the correspondence analysis are not properly standardized as their variances depend on cell probabilities. As an alternative, the deviation from independence is measured by the underlying correlation coefficient. It is argued that this measure is similar to the (signed square-root) of the log-likelihood difference, suggesting that this metric shares some more appealing statistical properties.

*University of Cologne, Institute of Econometrics, Albertus Magnus Platz, 50923 Köln, Germany, email: breitung@statistik.uni-koeln.de.

1 Introduction

What do you do, if you don't know anything about something that you do not really care about? Right, you ask Google. So I typed: "What is correspondence analysis"? As usual, one of the first entries of the search results is Wikipedia. I read: "*Correspondence Analysis is conceptually similar to principal component analysis, but applies to categorical rather than continuous data.*" Oh that's great because I have some experience with principal component analysis (see e.g. Breitung and Eickmeier 2011, Breitung and Tenhofen 2011, Breitung 2013). I therefore started to study the relevant textbooks like Blasius (2001) and Greenacre (2007) and try to understand the relationship between correspondence analysis (*henceforth: CA*) and principal component analysis (PCA). Obviously the literature on CA is huge and for this small note I was not able to carefully study the relevant literature. So I stick to the position of an alien that is visiting the planet *"correspondence analysis"*. This alien is grown up on the planet called "Econometrics", where data matrices are typically pretty large and observed on metric scales. Utmost importance is attached to define and identify what you are estimating, and no one will take you seriously unless you can demonstrate that your results are statistically significant and that the assumptions of your model are valid.

This said I carefully approached the CA. If you do so, you first have to survive an exhausting trip through a (non)Euclidian geometric space, which is somewhat unusual in statistical analysis. To be honest, this geometric journey has left me a bit perplexed. To understand better the geometric approach, I first studied the geometry of CA and found that the χ^2-distance has some undesirable features and it may therefore be appealing to consider alternative metrics as well. My favorite is to transform the entries of the contingency table into correlation coefficients. Interestingly, correlations are constructed similarly as the χ^2-distance but imply a slightly different denominator. Another measure is the likelihood ratio (LR) that seems most natural from a statistical point of view. Moreover, LR statistics allow us to find out whether some specific cell frequency (or row/column) is significantly different from the expected cell frequency under the assumption of independent outcomes.

To illustrate the issues involved, I borrowed the empirical example of Blasius and Greenacre (2006) and compare different distance measures and visualization

techniques. In particular I present the correlation matrix in form of a "balloon plot" that highlights the distances from the expected cell frequencies. I argue that the information presented by the balloon plot is similar to the biplot in CA but it does not involve a loss of information from dimensionality reduction. Furthermore I visualize the correlation matrix by a respective biplot that reveals interesting differences to the standard correspondence plot in CA.

The rest of the paper is organized as follows. In Section 2 I compare PCA and CA thereby highlighting some conceptual differences between these approaches. The standard χ^2-distance measure is considered in Section 3. Following Rao (1995) I argue that this distance measure possesses some undesirable properties. Another drawback is that under the assumptions of independent outcomes the variances of the χ^2-distances depend on the cell probabilities and, therefore, the entries of the contingency matrix are not properly standardized. To overcome this disadvantages, Section 4 proposes a distance measure constructed from the underlying correlation coefficient that also provides a proper standardization. In Section 5 I consider the standard statistical distance measure namely the difference of the log-likelihood function (resp. the logarithm of the likelihood ratio, LR). It turns out that the LR distance is related but not identical to the χ^2-distance. By construction the LR distances shares some optimality properties for statistical inference. The theoretical results are illustrated in Section 6 by using the empirical example of Blasius and Greenacre (2006). Section 7 offers some concluding remarks.

2 The relationship between CA and PCA

PCA is typically applied to large correlation matrices. Let X denote an $N \times J$ matrix where N indicates the number of observations on each of the J variables. The data is standardized such that the diagonal elements of the matrix $N^{-1}X'X$ are equal to one. The PCA is a dimension reduction technique that maps the J variable in a subspace spanned by $J \gg r$ linear combinations of the original variables (the so-called principal components).

Whereas PCA considers (standardized) observations from a large number of variables, the CA starts from a matrix constructed from a particular distance metric (see Section 3). Furthermore, in most textbook examples and empirical

applications the size of the contingency table is rather small, with K and L typically less than 10. In such situations it is not obvious that it is necessary to reduce the dimensionality. As far as I can see the main reason for reducing the dimension of the correspondence matrix is the possibility to represent the data in form of a two-dimensional diagram (biplot). Obviously, such a dimensionality reduction implies some loss of information but helps to recover some unknown structure behind the contingency matrix.

Typically, the data for a $K \times L$ contingency table come in form of n bivariate vectors $(a_1, b_1), (a_2, b_2), \ldots, (a_n, b_n)$ of independent multinomially distributed random variables, where $a_i \in \{1, 2, \ldots, K\}$, $b_i \in \{1, 2, \ldots, L\}$ and $i = 1, \ldots, n$. Let us define $K + L$ corresponding dummy variables as

$$d_i^a(k) = \begin{cases} 1 & \text{if } a_i = k \\ 0 & \text{otherwise} \end{cases}$$

$$d_i^b(\ell) = \begin{cases} 1 & \text{if } b_i = \ell \\ 0 & \text{otherwise} \end{cases}$$

Then, the relative frequency results as

$$p_{k\ell} = \frac{n_{k\ell}}{n} = \frac{1}{n} \sum_{i=1}^{n} d_i^a(k) d_i^b(\ell),$$

where $n_{k\ell}$ indicates the number of observations in the (k, ℓ)-cell of the contingency table. The difference between actual and expected (by assuming independence) relative cell frequencies result as

$$q_{k\ell} := p_{k\ell} - p_{k\cdot} p_{\cdot\ell} \;=\; \frac{1}{n} \sum_{i=1}^{n} d_i^a(k) d_i^b(\ell) - \left(\frac{1}{n} \sum_{i=1}^{n} d_i^a(k) \right) \left(\frac{1}{n} \sum_{i=1}^{n} d_i^b(\ell) \right),$$

where $p_{k\cdot} = \sum_{\ell=1}^{L} p_{k\ell}$ and $p_{\cdot\ell} = \sum_{k=1}^{K} p_{k\ell}$ are called the row and column masses, respectively. This shows that the "residual" $q_{k\ell}$ can be interpreted as the sample covariance between the dummy variables $d_i^a(k)$ and $d_i^b(\ell)$. In Section 4 this interpretation is employed to construct a correlation matrix between the outcomes of the contingency table.

For the CA we divide the residuals $q_{k\ell}$ by the factor $\sqrt{p_{k\cdot}p_{\cdot\ell}}$ yielding what is often called the "standardised residuals" (e.g. Blasius 2001: 89, and Greenacre 2007: 202). In Section 3 I argue that dividing the residuals by $\sqrt{p_{k\cdot}p_{\cdot\ell}}$ does not result in a proper standardization, as the variances of the χ^2-distances depend on the cell probabilities. Let S denote the $K \times L$ matrix with typical element

$$s_{k\ell} = \frac{q_{k\ell}}{\sqrt{p_{k\cdot}p_{\cdot\ell}}} = \frac{p_{k\ell} - p_{k\cdot}p_{\cdot\ell}}{\sqrt{p_{k\cdot}p_{\cdot\ell}}} \,. \tag{1}$$

The reduction of dimensionality is obtained by applying the singular value decomposition (SVD) with $S = UDV'$. Let U_2 (V_2) denote the matrix of the first two left (right) singular vectors and D_2 is the upper-left (2×2) submatrix of D. In PCA the biplot coordinates are given by $U_2 D_2^{\alpha}$ and $V_2 D_2^{1-\alpha}$, where an asymmetric version is most popular by choosing either $\alpha = 0$ or $\alpha = 1$. The set of points depicting the variables is typically drawn as arrows from the origin to reinforce the idea that they represent biplot axes onto which the observations can be projected when approximating the original data.

The SVD results in a least-squares minimal approximation of the element in S given by

$$s_{k\ell} = u'_{2k} D_2 v_{2\ell} + \tilde{e}_{k\ell}$$

where u_{2k} ($v_{2\ell}$) represents the k-th (ℓ-th) row of U_2 (V_2) and $\tilde{e}_{k\ell}$ is the approximation error. So far so comprehensible. But now correspondence analysis introduces another transformation yielding

$$\begin{aligned} \frac{1}{\sqrt{p_{k\cdot}p_{\cdot\ell}}}\, s_{k\ell} &= \frac{p_{k\ell}}{p_{k\cdot}p_{\cdot\ell}} - 1 = \left(\frac{u'_{2k}}{\sqrt{p_{k\cdot}}}\right) D_2 \left(\frac{v_{2\ell}}{\sqrt{p_{\cdot\ell}}}\right) + \frac{1}{\sqrt{p_{k\cdot}p_{\cdot\ell}}}\tilde{e}_{k\ell} \\ &:= \phi'_k D_2 \gamma_\ell + e_{k\ell} \end{aligned}$$

where ϕ_k and γ_ℓ are called "standard coordinates" (cf. Greenacre 2007: 202). It is important to note, however, that the error $e_{k\ell}$ is no longer a least-squares minimal approximation error. Instead, the least-squares minimal approximation is obtained from the SVD applied to the matrix $D_r^{-1/2} S D_c^{-1/2}$, where D_r and D_c are diagonal matrices with the row and column masses ($p_{k\cdot}$ resp. $p_{\cdot\ell}$) on the

main diagonals. The latter SVD yields a different representation with smaller approximation error. As an example consider the Asbestos data set (see Selikoff 1981).[1] The two alternative approaches for computing the standard coordinates yield the following matrices of standard row coordinates:

$$\Phi = D_r^{-1/2} U_2 \;=\; \begin{pmatrix} -0.847 & 0.472 \\ 0.416 & -1.334 \\ 1.800 & 0.879 \\ 2.161 & 2.167 \end{pmatrix}, \quad \widetilde{\Phi} \;=\; \begin{pmatrix} -0.224 & 0.476 \\ 0.045 & -0.689 \\ 0.582 & -0.338 \\ 0.780 & 0.429 \end{pmatrix}$$

where $\widetilde{\Phi}$ is obtained as the two left eigenvectors of the SVD for $D_r^{-1/2} S D_c^{-1/2}$.[2] This alternative ("direct") way of computing the standard coordinates is also mentioned by Greenacre (2007, eq. (A.13)). What is obvious from the above results is that the two approaches may render quite different coordinates.

An important issue in correspondence analysis is the proper scaling of the biplot axes. The `R` package "ca" (cf. Nenadić and Greenacre 2007) offers 8 different scaling options for the coordinates. Obviously, the issue of the "best scaling" of the coordinates is not settled completely. The standard scaling employs *principal coordinates* given by $F = D_r^{-1/2} U_2 D_2$ and $G = D_c^{-1/2} V_2 D_2$. This particular scaling is chosen such that the weighted sum-of-squares of the principal coordinates (i.e. their inertia in the direction of this dimension) is equal to the square of the singular value (the principal inertia).

Summing up, PCA and CA share the idea of representing some matrix by a lower dimensional approximation. But the details of the analysis are quite different. Statistical inference using PCA typically assumes an i.i.d. sample of J correlated variables, where $r \ll J$ linear combinations of the variables (principal components) are constructed that best represent the linear dependence among the variables.[3] On the other hand, the CA analyses the variability of the rows/columns of the contingency table. Notwithstanding these conceptual differences the CA benefits from adapting useful tools like the biplot in order to visualize the data.

[1] The R code for this and the other computations are provided on the homepage of the author.

[2] The coordinates are available from the output ($rowcoord) of the R prackage "ca".

[3] The i.i.d. assumption may be dropped by allowing for some "weak correlation" (e.g. Bai 2003) but the data matrix of the CA may be strongly correlated in both dimensions.

3 The χ^2-distance

Textbooks on CA (e.g. Blasius 2001 and Greenacre 2007) typically start with some geometric reasoning in Euklidian space in order to explain how to measure the distance between the row resp. column profiles. Then Pearson's X^2-statistic for independence is introduced and it is argued that this test statistic gives rise to the χ^2-distance measure for two row profile vectors r_k and $r_{k\prime}$ defined as (cf. Greenacre 2007: 31)

$$||r_k - r_{k^*}||_\chi = \sqrt{\sum_{\ell=1}^{L} \frac{[(p_{k\ell}/p_{k\cdot}) - (p_{k^*\ell}/p_{k^*\cdot}]^2}{p_{\cdot\ell}}}$$

where $r_k = (p_{k1}/p_{k\cdot}, \ldots, p_{kL}/p_{k\cdot})'$ and r_{k^*} denote two $L \times 1$ vectors of row profiles. The X^2-statistic results as a weighted average of the row (resp. column) distances

$$X^2 = n \sum_{k=1}^{K} p_{k\cdot} ||r_k - c||_\chi^2$$

where $c = (p_{\cdot 1}, \ldots, p_{\cdot L})'$ is the average row profile under the assumption of independence (i.e. the vector of column masses). A similar representation can be derived for the column distances.

As far as I can see, the main reason for introducing such a distance metric is the desire to develop a geometric interpretation for Pearson's X^2 statistic. This is achieved by defining the χ^2 distance as a *weighted* Euclidian distance. For me as an alien it is difficult to see why the squared distances should be weighted by the (inverted) column masses. The only reason seems to be that applying this particular weighting scheme gives rise to the X^2 statistic. But why should the distance between the entries of two rows profiles depend on the respective column masses? This implies that when we add or drop rows, then the corresponding distances may get smaller or larger. Another problem is that if the column masses gets small, then undue emphasis is given to the corresponding row distances. This let the famous (and now 102 years old) C.R. Rao (1995) to advocate the Hellinger

distance defined as

$$d_H(r_k, r_{k*})^2 = \sum_{\ell=1}^{L} (\sqrt{p_{k\ell}/p_{k\cdot}} - \sqrt{p_{k^*\ell}/p_{k^*\cdot}})^2.$$

Beside the fact that this distance measure only depends on the row profiles themselves and do not imply any weighting related to information from outside, this distance measure satisfies the principle of *distributional equivalence* and the distances do not get arbitrarily large if the column masses tend to zero. Cuadras and Cuadras (2006) proposed a generalized distance measure that entails the CA distances and the Hellinger distance as a special case. Other alternatives are the L_1-type distance of Benzécri (1982) and the log-ratios considered in Cuadras and Cuadras (2006). This list of proposed distance measures is not complete. Consider, for example a distance measure that is popular in machine learning when it comes to analyzing the similarity of discrete distributions[4] (e.g. Yang et al. 2015), which is defined as

$$||r_k - r_{k*}||_S = \sqrt{\sum_{\ell=1}^{L} \frac{(p_{k\ell} - p_{k*\ell})^2}{2(p_{k\ell} + p_{k*\ell})}} \tag{2}$$

Note that under independence we have $p_{k\ell} \approx p_{k\ell}$ and, therefore, for independent outcomes this we have $4n \sum_{k=1}^{K} ||r_k - c||_S^2 \approx X^2$.

Let us now consider the Pearson's statistic given by

$$X^2 = n \sum_{k=1}^{K} \sum_{\ell=1}^{L} s_{k\ell}^2$$

where $s_{k\ell}$ as defined in (1) is often called the "standardized residuals". A proper standardization would imply that $s_{k\ell}$ has the same (unit) variance for all k and ℓ but it turns out that the distributional properties of $s_{k\ell}$ depend on the cell probabilities. The reason is that the denominator $\sqrt{p_{k\cdot}p_{\cdot\ell}}$ is different from the standard deviation of the numerator, even if the outcomes are independent. To illustrate this fact I performed a small Monte Carlo experiment. The data generating process resembles the Asbestos data set used in the previous section, where

[4]See also the function *chisqDistance(a,b)* in the R package *colordistance*.

Table 1: Variances of the residuals with different standardization

	$p_{\kappa\ell}$	$\mathrm{var}(s_{\kappa\ell})$	$\mathrm{var}(\varrho_{\kappa\ell})$	$p_{\kappa\ell}$	$\mathrm{var}(s_{\kappa\ell})$	$\mathrm{var}(\varrho_{\kappa\ell})$
row	first column			second column		
1	0.159	0.346	1.036	0.101	0.464	1.000
2	0.174	0.328	1.025	0.111	0.451	1.015
3	0.035	0.471	1.045	0.022	0.610	0.975
4	0.089	0.395	0.987	0.056	0.557	1.003
5	0.055	0.427	0.988	0.035	0.592	0.988
row	third column			fourth column		
1	0.034	0.597	0.975	0.013	0.629	0.954
2	0.038	0.616	1.051	0.015	0.631	1.001
3	0.007	0.853	1.033	0.003	0.904	1.017
4	0.019	0.764	1.042	0.007	0.772	0.978
5	0.012	0.826	1.044	0.004	0.805	0.945

Entries present the variances of the standardized residuals as defined in (1) and (3). Entries are based on 1000 replications of the Asbestos dataset generated under the assumption of independent outcomes.

the data is generated as independent multinomial distributed random variables with probabilities $p_{k\cdot}p_{\cdot\ell}$, that is, the data are generated under the assumption of independent outcomes. The results of the Monte Carlo simulation based on 1000 replications are presented in Table 1.

The results indicate that $s_{\kappa\ell}$ cannot be considered to be properly standardized as the variances tend to become larger for smaller cell probabilities $p_{\kappa\ell}$. The variances range from 0.328 for a cell probability of 0.174 up to 0.904 for a probability of 0.003. Accordingly, it is much more likely to observe large residuals when the corresponding probability is small. On the other hand, the standardisation proposed in the next section appear to work well (indicated by $\varrho_{\kappa\ell}$ in Table 1).

4 Correlation as a distance measure

There is a close relationship between distance and correlation. For example, consider the Euclidean distance

$$||x - y|| = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

between two vectors of standardized random variables $x = (x_1, \ldots, x_n)'$ and $y = (y_1, \ldots, y_n)'$. Since

$$||x - y||^2 = 2n - 2\sum_{i=1}^{n} x_i y_i$$

the squared Euclidian distance can be expressed as $||x - y||^2 = 2n(1 - \varrho_{xy})$, where ϱ_{xy} denotes the correlation coefficient between x_i and y_i. It therefore makes sense to consider the relationship between distances and correlations between the outcomes of the contingency table.

As noted in Section 2 the differences $p_{k\ell} - p_{i\cdot}p_{\cdot j}$ can be written as the covariance between the two dummy variables $d_i^a(k)$ and $d_i^b(\ell)$. This suggests to use the correlation between the dummy variables as measure of the deviation from independence. Since the dummy variables are binomially distributed with variances $var[d_i^a(k)] = p_{k\cdot}(1 - p_{k\cdot})$ and $var[d_i^b(\ell)] = p_{\cdot\ell}(1 - p_{\cdot\ell})$ and therefore, an estimator for the correlation between the two dummy variables results as

$$\varrho_{k\ell} = \frac{p_{k\ell} - p_{i\cdot}p_{\cdot j}}{\sqrt{(p_{k\cdot} - p_{k\cdot}^2)(p_{\cdot j} - p_{\cdot j}^2)}} \tag{3}$$

It is interesting to note the close correspondence between this correlation measure and the standardized residuals $s_{k\ell}$ used for the correspondence analysis. The only difference is the squared probabilities in the denominator. Since the cell probabilities are typically small, the differences between $s_{k\ell}$ and $\varrho_{k\ell}$ are usually moderate. An important property of the correlation coefficient is that

$$\sqrt{n}\,\varrho_{k\ell} \xrightarrow{d} \mathcal{N}(0, 1)$$

where $\xrightarrow{d}$ signifies convergence in distribution. Accordingly under the hypothesis of independent outcomes all correlations have the same asymptotic distribution and can therefore be considered to be properly standardized. In Section 6 I therefore use the correlations $\varrho_{k\ell}$ instead of $s_{k\ell}$ as an alternative starting point of the correspondence analysis.

5 The log-likelihood distance

It is well known that for statistical tests the difference between the log-likelihood functions under the null and alternative hypotheses results in most powerful test statistics (Neyman-Pearson lemma). It is therefore natural to consider the difference in the log-likelihood functions (that is the logarithm of the likelihood-ratio) as a distance measure between the actual and expected cell frequencies.

Let us therefore consider the log-likelihood difference for a test of the hypothesis of independent outcomes which given by

$$\mathrm{LR} = 2n \sum_{k=1}^{K} \sum_{\ell=1}^{L} p_{k\ell} \log \left(\frac{p_{k\ell}}{p_{k\cdot} p_{\cdot\ell}} \right). \tag{4}$$

How is this likelihood-ratio statistic related to Pearson's X^2-statistic? Let $p^0_{k\ell} = p_{k\cdot} p_{\cdot\ell}$. A second order Taylor expansion around $p^0_{k\ell}$ yields:

$$p_{k\ell} \left[\log(p_{k\ell}) - \log(p^0_{k\ell}) \right] \approx (p_{k\ell} - p^0_{k\ell}) + \frac{1}{2p^0_{k\ell}} (p_{k\ell} - p^0_{k\ell})^2.$$

Since $\sum_{k=1}^{K} \sum_{\ell=1}^{L} (p_{k\ell} - p^0_{k\ell}) = 0$ it follows that the likelihood ratio statistic LR can be approximated by Pearson's X^2 statistic whenever $p_{k\ell} - p^0_{k\ell}$ is small (that is, if the outcomes are nearly independent).

Let us now consider the LR test for the hypothesis that a particular row profile, say for $k = 1$, deviates from the other row profiles. It is important to notice that it is not possible to just pick the relevant summands for $k = 1$ from the LR statistic (4) as this would not result in a test statistic with the usual χ^2-distribution with $L - 1$ degrees of freedom. Instead we consolidate the contingency table such that

$k\downarrow$	$\ell:1$	2	$\cdots$	L
1	p_{11}	p_{12}	$\cdots$	p_{1K}
2	$\overline{p}_{21}$	$\overline{p}_{22}$	$\cdots$	$\overline{p}_{2K}$

where $\overline{p}_{2\ell} = \sum_{k=2}^{K} p_{k\ell}$. Accordingly, the remaining rows $k = 2, 3, \ldots, K$ are aggregated such that the contingency table is reduced to a $2 \times L$ table. The LR statistic for this row results as

$$\mathrm{LR}(k=1) = 2n \sum_{\ell=1}^{L} p_{1j} \log\left(\frac{p_{1j}}{(p_{1\ell}+\overline{p}_{2\ell})p_{1\cdot}}\right) + p_{1j}\log\left(\frac{\overline{p}_{2\ell}}{(p_{1\ell}+\overline{p}_{2\ell})\overline{p}_{2\cdot}}\right) \qquad (5)$$

where $\overline{p}_{2\cdot} = \overline{p}_{21}+\overline{p}_{22}+\cdots+\overline{p}_{2L}$. This test statistic is asymptotically χ^2-distributed with $L-1$ degrees of freedom. In what follows I refer to this LR statistic as the *"LR row distance"*. In a straightforward manner we can also define a log-likelihood distance for each cell. To this end we need to consolidate also the remaining columns such that for the upper right cell, for example, we obtain

$k\downarrow$	ℓ: 1	2
1	p_{11}	$\overline{p}_{12}$
2	$\overline{p}_{21}$	$\overline{p}_{22}$

where $\overline{p}_{12} = \sum_{k=2}^{L} p_{k\ell}$, $\overline{p}_{21} = \sum_{k=2}^{K} p_{k\ell}$ and $\overline{p}_{22} = \sum_{k=2}^{K}\sum_{\ell=2}^{L} p_{k\ell}$. The log-likelihood difference for testing independence in this 2×2 matrix is denoted by $\mathrm{LR}(k=1,\ell=1)$. Let us now consider the properties of this distance measure as an alternative to the χ^2-distance. First, it is obvious that under independence the log-likelihood difference is asymptotically χ^2-distributed whereas the distribution of the χ^2-distance depends on the cell probabilities. Furthermore it turns out that $\mathrm{LR}(k,\ell) \approx n\varrho_{k\ell}^2$ and, therefore, we may use the signed square-root of the $\mathrm{LR}(k,\ell)$ statistic as an alternative for constructing properly standardized residuals.

6 An empirical illustration

To illustrate the issues discussed in the previous sections, I borrow an example from Blasius and Greenacre (2006). The data are from the International Social

Table 2: Contingency table for "international sports × countries"

	U.K.	U.S.	Russia	Spain	France
agree strongly	230	400	1010	201	365
agree	329	471	530	639	478
neither nor	177	237	141	208	305
disagree	34	28	21	72	50
disagree strongly	6	12	11	14	97

Column profiles

	U.K.	U.S.	Russia	Spain	France	row masses
agree strongly	0.296	0.348	0.590	0.177	0.282	0.364
agree	0.424	0.410	0.309	0.563	0.369	0.403
neither nor	0.228	0.206	0.082	0.183	0.236	0.176
disagree	0.044	0.024	0.012	0.063	0.039	0.034
disagree strongly	0.009	0.010	0.006	0.012	0.075	0.023
column masses	0.128	0.189	0.282	0.187	0.213	
LR difference	37.41	21.98	568.00	265.68	260.89	

Souce: Blasius and Greenacre (2006: 7). "LR difference" indicates the LR statistic for the hypothesis that the respective row profile is different from the expected profile under independence.

Survey Program (ISSP). Table 2 reports responses from five selected countries to the question: *"When my country does well in international sports, it makes me proud to be [Country Nationality]"*. The contingency table is provided in Table 2.

Let us first address the question: how special are the responses from the different countries? To assess the deviation of each column (country) to all other countries, I computed the log-likelihood differences as in (4) after transposing the matrix to obtain the LR column distances. The results are presented in the last row of Table 2. These results suggest that the responses of Russia are most different from the responses of the other countries, whereas the responses of Spain and France are less different but still quite far away from the "average column" (row masses). The columns of the U.K. and U.S. are much more similar to the average (resp. independent) pattern. We can also compare the outcomes of two countries with each other. For example the log-likelihood difference between the U.K. and U.S. is only 10.834, suggesting that the response pattern of these countries are pretty similar. On the other hand, the log-likelihood difference

between the columns of France and Spain is 145.06. Although the distance of these two columns to the mean column is similar, this does not mean that the response pattern of France and Spain is similar. This becomes also clear from the correspondence analysis (see below).

Table 3 compares the standardized residuals $s_{k\ell}$ to the alternative measures considered in Sections 3 – 5. The first line for each row presents the respective cell entries of the matrix S as defined in (1), whereas the second line reports the symmetric distances measure $||\cdot||_S$ as defined in (2). In most cases the differences between these two distance measures are pretty small. The correlation measure reveals more important differences to the χ^2-distance measures in particular if the distance gets large. For example for the $(1,3)$-cell the χ^2-distance is 0.199, whereas the correlation is 0.294. This suggests that larger distances are more accentuated by using correlations instead of χ^2-distances. By multiplying correlations with $\sqrt{n}$ we obtain a test statistic for the null hypothesis that the two features affecting the cell outcome are independent. This hypothesis is rejected at the 0.05 significance level for 15 out of 25 cells.

In order to visualize the correlation pattern, Figure 1 presents a balloon plot for the correlations. The larger the positive correlation the larger is the green (resp. black) balloon, whereas negative correlations are indicated by red (grey) balloons. As the more popular alternative Figure 2 presents the correspondence plot as provided by the `R` package "ca" (Nenadić and Greenacre 2007).[5] Let me summarize the main findings of Blasius and Greenacre (2006: 10f) from analysing this contingency table in their own words followed by the corresponding pattern in the balloon plot:

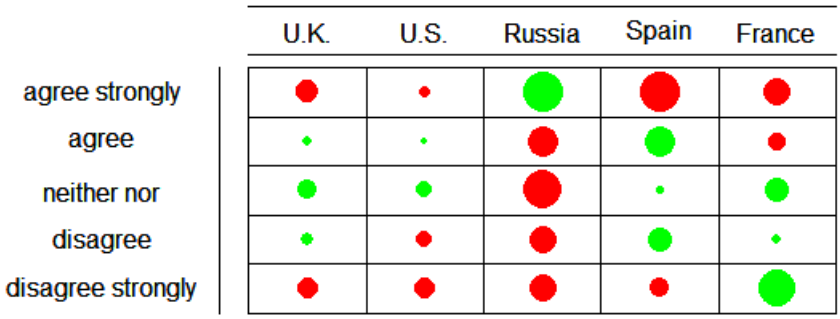
- *"The first dimension can be interpreted as 'level of pride towards achievement in international sport. ... As for the countries we see Russia on the left opposing the other countries of the right; thus of these five nations the Russians feel most proud when Russia is doing well in international sports. At the opposite right-hand side of this axis we see that the French and the Spanish are the least proud of the five nations in this respect..."*. This conclusion is confirmed by the balloon plot (Figure 1). The nationality Russian is most correlated with strongly agree, suggesting that Russians feel most

[5]The plot in figure 2 is slightly different from the CA plot in Blasius and Greenacre (2006), as the fist axis seems to be multiplied by -1. Note that the sign of the axes is not identified.

Table 3: Alternative distance measures

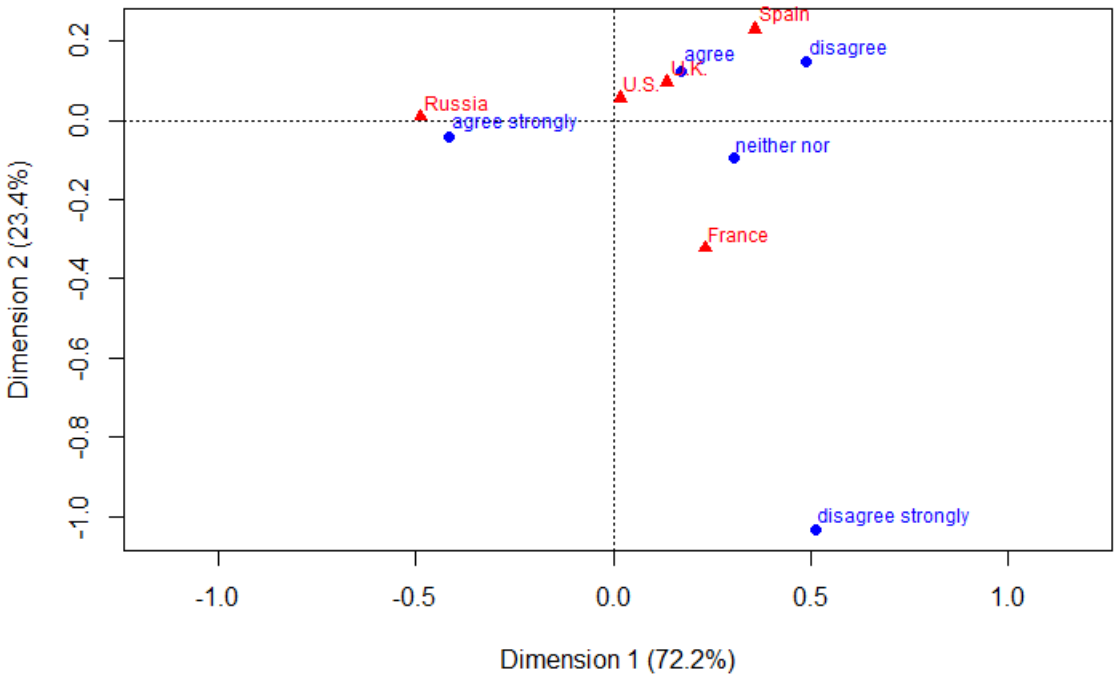
(row,column):	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)
$s_{k\ell}$ (residual)	−0.039	−0.010	0.199	−0.133	−0.062
symmetric	−0.041	−0.011	0.174	−0.155	−0.066
correlation	−0.053	−0.015	0.294	−0.185	−0.088
$\sqrt{n}$·corr.	−4.171	−1.191	22.94	−14.47	−6.900
LR distance	−4.226	−1.194	22.68	−15.13	−7.000
(row,column):	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)
$s_{k\ell}$ (residual)	0.011	0.004	−0.078	0.108	−0.024
symmetric	0.011	0.004	−0.083	0.100	−0.025
correlation	0.016	0.006	−0.120	0.156	−0.036
$\sqrt{n}$·corr.	1.250	0.527	−9.361	12.18	−2.835
LR distance	1.248	0.527	−9.459	12.07	−2.846
(row,column):	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)
$s_{k\ell}$ (residual)	0.044	0.031	−0.118	0.007	0.065
symmetric	0.041	0.030	−0.138	0.007	0.060
correlation	0.052	0.038	−0.154	0.009	0.081
$\sqrt{n}$·corr.	4.074	3.001	−12.02	0.721	6.334
LR distance	3.957	2.953	−12.77	0.718	6.161
(row,column):	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)
$s_{k\ell}$ (residual)	0.019	−0.022	−0.062	0.069	0.012
symmetric	0.018	−0.023	−0.075	0.058	0.012
correlation	0.021	−0.025	−0.074	0.078	0.013
$\sqrt{n}$·corr.	1.654	−1.958	−5.822	6.137	1.081
LR distance	1.594	−2.036	−6.394	5.661	1.064
(row,column):	(5,1)	(1,2)	(5,3)	(5,4)	(5,5)
$s_{k\ell}$ (residual)	−0.036	−0.036	−0.058	−0.030	0.157
symmetric	−0.044	−0.042	−0.073	−0.035	0.108
correlation	−0.039	−0.040	−0.069	−0.034	0.179
$\sqrt{n}$·corr.	−3.048	−3.164	−5.420	−2.669	14.00
LR distance	−3.469	−3.462	−6.085	−2.871	12.33

Figure 1: Balloon plot of correlations



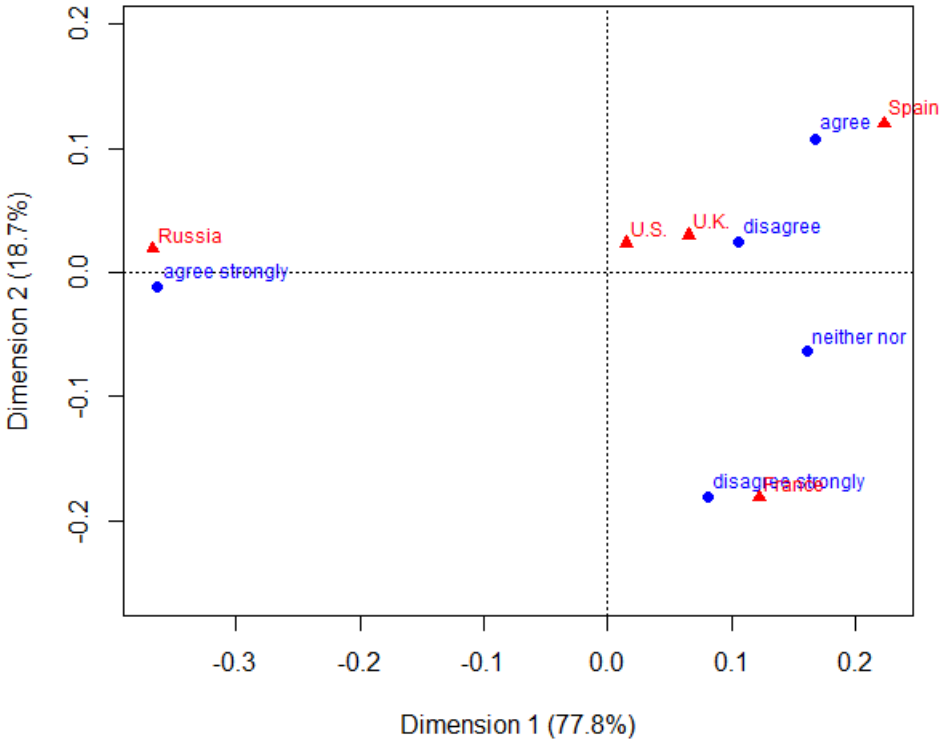
Note: Green (resp. black) balloon: positive correlation, red (grey) balloon: negative correlation. The size of the balloon corresponds to the absolute correlation between the dummy variables indicating the (i, j) cell of the contingency table.

Figure 2: Symmetric correspondence plot



Note: Correspondence plot as obtained from the `R` package "ca".

Figure 3: Symmetric biplot of correlations



Note: Modified correspondence plot based on correlations.

proud for the achievements in international sports. By contrast, the category "strongly agree" is highly negatively correlated with "France" and "Spain" which corresponds with the findings of the correspondence plot, where these countries are far away from the category "strongly agree".

- *"The second dimension mainly reflects the outlying position "disagree strongly" as well as France compared with the other categories and countries."* Indeed this "outlier' is represented by the high positive correlation (green balloon) between "disagree strongly" and France.

- *"...the U.S. and U.K. have very similar response pattern, which are not much different from the overall, or average, pattern. Geometrically this is depicted by these two countries lying close to each other, towards the origin of the map."* In the balloon plot this fact can be identified by the rather small correlations for all categories of these two countries.

It appears that many features of the correspondence plot can also be obtained from studying the correlations presented by the balloon plot. Categories that are close together in the correspondence plot are highly correlated, whereas observations far away from each other are negatively correlated. Outcomes close to the origin of the correspondence plot correspond to low correlations. A somewhat puzzling phenomenon is the isolated location of "disagree strongly" in the correspondence plot. The balloon plot indicates that this outlying position results from the fact that for France this category is highly over-represented, while for all other countries this response category behaves quite similar.

It may be interesting to see how a biplot for the correlation looks like. To this end I apply an SVD to the correlation matrix $R = (\varrho_{k\ell})$ resulting in $R = UDV'$. Let U_2 (V_2) denote the first two columns of the matrix of the left (right) singular vectors and D_2 is the upper-left (2×2) submatrix of D. Figure 3 presents the (symmetric) biplot based on $U_2 D_2^{1/2}$ (as row coordinates) and $V_2 D_2^{1/2}$ (as column coordinates). Overall the biplot resembles the original correspondence plot in Figure 2 but some interesting differences emerge. In Figure 3 the category "disagree strongly" is no longer as isolated as in the original correspondence plot. Now this category is located much closer to "France" (such that their labels overlap) representing the high correlation between "France" and "disagree strongly". Furthermore, the biplot also reveals the outlying position of the combination

"strongly agree" and "Russia". This corresponds to the largest correlation of the whole table with nearly 0.295, whereas the correlation between "France" and "disagree strongly" is substantially lower (0.18).

7 Conclusion

As it's time to leave the planet "correspondence analysis" I am going to leave a message in the bottle at the Schwarzrheindorfer waterfront: "Whoever finds this bottle, let it be said that 65 years after the fateful 1957 an alien from a distant planet has strayed into this inhospitable territory. He left some cryptical notes on alternative distance measures, although everyone on this planet is happy about this geometry, which is not encountered anywhere else." So in this bottle you will find a collection of red and green balloons. If you want to escape the limitations of a two-dimensional plane, you may fill these balloons with helium and they will carry you away, right in the direction of my home planet...".

References

Bai, J. (2003): Inferential Theory for Factor Models of Large Dimensions. In: Econometrica 71, S. 135–172.

Blasius, Jörg (2001): Korrespondenzanalyse. München: Oldenbourg.

Blasius, Jörg/Greenacre, Michael (2006): Correspondence Analysis and Related Methods in Practice. In: Greenacre, Michael and Blasius, Jörg (eds.): Multiple Correspondence Analysis and Related Methods. London: Chapman & Hall/CRC. S. 3–40.

Breitung, Jörg (2013):] Factor Models. In: Hashimzade, Nigar and A. Thornton, Michael A. (eds.): Empirical Macroeconomics. Edward Elgar, 249–265.

Breitung, Jörg/Eickmeier, Sandra (2011): Testing for structural breaks in dynamic factor models. In: Journal of Econometrics, 163, S. 71–84.

Breitung, Jörg/Tenhofen, Jörn (2011): GLS estimation of dynamic factor

models. In: Journal of the American Statistical Association, 106, S. 1150–1166.

Greenacre, Michael (2007): Correspondence Analysis in Praxis. 2nd ed. London: Chapman & Hall/CRC.

Nenadic, Olec/Michael Greenacre (2007): Correspondence Analysis in R, with Two- and Three-dimensional Graphics: The ca Package. In: Journal of Statistical Software, Volume 20, Issue 3.

Rao, C.R. (1995): A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. Qüestiió, 19, S. 23–63.

Yang, Wei/Xu, Luhui/Chen, Xiaopan/Zheng, Fengbin/Liu, Yang (2015): Chi-Squared Distance Metric Learning for Histogram Data. In: Mathematical Problems in Engineering, 2015, Volume 2015, Article ID 352849.