

MS-E2112 Multivariate Statistical Analysis (5cr)

Lecture 6: Bivariate Correspondence Analysis - part II

Lecturer: Pauliina Ilmonen
Slides: Ilmonen/Kantala

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Contents

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Correspondence Analysis

Chi-square Test Statistic

Chi-square Distances

Correspondence Analysis, the Row Profiles

Correspondence Analysis, the Column Profiles

Association Between the Profiles

References

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Correspondence Analysis

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Correspondence Analysis (CA)

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Correspondence analysis is a PCA-type method appropriate for analyzing categorical variables. The aim in bivariate correspondence analysis is to describe dependencies between the variables and to visualize approximate attraction repulsion indices in lower dimensions. We consider a sample of size n described by two qualitative variables, x with categories A_1, \dots, A_J and y with categories B_1, \dots, B_K . We use the same notations as last week and start by independence testing and by looking at chi-square distances between the row (or column) profiles of the variables.

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Chi-square Test Statistic

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

The independence between variables x and y can be tested using chi-square statistic. The null hypothesis of the test is

$$H_0 : p_{jk} = p_{j.} \cdot p_{.k}, \text{ for all } j, k$$

and the test statistic is given by

$$\chi^2 = \sum_{j=1}^J \sum_{k=1}^K \frac{(n_{jk} - n_{jk}^*)^2}{n_{jk}^*}.$$

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Independence

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Under random sampling, the n_{jk} follow multinomial distribution with parameters n, p_{11}, \dots, p_{JK} and $E[n_{jk}] = np_{jk}$. In the test statistics above, the np_{jk} , under the null, are estimated by n_{jk}^* . When the sample size n is large, the test statistic has, under the null hypothesis, approximately chi-square distribution with $(K - 1)(J - 1)$ degrees of freedom. Thus the null hypothesis (independence between variables x and y) is rejected at the level α if

$$\chi^2 > \chi_{(K-1)(J-1), 1-\alpha}^2.$$

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Chi-square distribution

Multinomial distribution

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Decomposition of the Chi-square Statistic

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Let $Z \in \mathbb{R}^{J \times K}$, where

$$Z_{jk} = \frac{f_{jk} - f_{j.} \cdot f_{.k}}{\sqrt{f_{j.} \cdot f_{.k}}}.$$

Thus, the matrix Z gives shifted and scaled relative frequencies of the variables. Moreover, the variables are scaled such that the elements

$$Z_{jk} = \frac{f_{jk} - f_{j.} \cdot f_{.k}}{\sqrt{f_{j.} \cdot f_{.k}}} = \frac{f_{jk} - f_{jk}^*}{\sqrt{f_{jk}^*}} = \frac{n_{jk} - n_{jk}^*}{\sqrt{n_{jk}^*}}$$

are the terms that are squared and summed in the chi-square statistic that is used for testing the independence of the variables.

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Decomposition of the Chi-square Statistic

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

A large positive value Z_{jk} indicates a large contribution to the chi-square statistic. This indicates a positive association between row j and column k . (More observations than expected under independence.) A large negative value Z_{jk} also indicates a large contribution to the chi-square statistic, but this indicates a negative association between row j and column k . (Less observations than expected under independence.) Values near zero indicate no contribution to the test statistic. (The number of observations is equal to the expected number under independence.)

Let

$$V = Z^T Z$$

and let

$$W = Z Z^T.$$

Now the chi-square statistic

$$\chi^2 = n(\text{trace}(V)) = n(\text{trace}(W)).$$

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distance

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Chi-square Distances

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Chi-square Distance

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

When the data is in the form of frequency distribution, the distance between the rows (or columns) can be measured using weighted euclidian distances. The so called chi-square distance between two rows j_1 and j_2 is given by

$$d(j_1, j_2) = \sum_{k=1}^K \frac{1}{f_{\cdot k}} \left(\frac{f_{j_1 k}}{f_{j_1 \cdot}} - \frac{f_{j_2 k}}{f_{j_2 \cdot}} \right)^2.$$

The euclidian distance gives the same weight to each column. The chi-square distance gives the same relative importance to each column proportionally to the average frequency. The division of each squared term by the expected frequency is variance standardizing and compensates for the larger variance in high frequencies and the smaller variance in low frequencies. If no such standardization were performed, the differences between larger proportions would tend to be large and thus dominate the distance calculation, while the differences between the smaller proportions would tend to be swamped. The weighting factors are used to equalize these differences.

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Chi-square Distance

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

The chi-square distances between two row profiles can be given as

$$\begin{aligned}d(j_1, j_2) &= \sum_{k=1}^K \frac{1}{f_{\cdot k}} \left(\frac{f_{j_1 k}}{f_{j_1 \cdot}} - \frac{f_{j_2 k}}{f_{j_2 \cdot}} \right)^2 \\ &= \sum_{k=1}^K \left(\frac{f_{j_1 k}}{f_{j_1 \cdot} \sqrt{f_{\cdot k}}} - \frac{f_{j_2 k}}{f_{j_2 \cdot} \sqrt{f_{\cdot k}}} \right)^2.\end{aligned}$$

Thus, if the row profiles are scaled, the usual euclidian metric can be used on the new scaled data.

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Chi-square Distance

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

The distance between two columns k_1 and k_2 is given by

$$d(k_1, k_2) = \sum_{j=1}^J \frac{1}{f_{j.}} \left(\frac{f_{jk_1}}{f_{.k_1}} - \frac{f_{jk_2}}{f_{.k_2}} \right)^2.$$

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Correspondence Analysis, the Row Profiles

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Correspondence Analysis (CA)

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Principal component analysis is based on maximizing euclidian distances. In the context of frequency distributions, the proper distance between the variables is the chi-square distance. In correspondence analysis, a PCA type approach is applied to modified data.

Whereas traditional PCA relies on euclidian distances, correspondence analysis is based on chi-square distances.

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Correspondence Analysis, the Row Profiles

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Let $R \in \mathbb{R}^{J \times K}$, where

$$R_{jk} = \frac{f_{jk}}{f_{j.} \sqrt{f_{.k}}} - \sqrt{f_{.k}}$$

The matrix R contains the scaled and shifted row profiles. Let R_j denote the j th row of R . In correspondence analysis on the row profiles, one finds orthonormal vectors (directions) u_i such that projection $P_i(\cdot)$ onto u_i maximizes the weighted sum of the euclidean distances,

$$\sum_{j=1}^J f_{j.} d^2(0, P_i(R_j)),$$

under the constraint that u_i is orthogonal to all u_l , $1 \leq l < i$.

Note that the row profiles are scaled and shifted to obtain a maximization problem that involves euclidean distances as optimization involving chi-square distances directly would be technically difficult.

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Correspondence Analysis, the Row Profiles

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

The problem is a problem of maximization under constraint, and similarly as in PCA, the solution is given by the eigenvalues and the eigenvectors of the matrix

$$V = \sum_{j=1}^J f_j \cdot R_j^T R_j$$

Some matrix algebra is needed to show that the matrix

$$V = \sum_{j=1}^J f_j \cdot R_j^T R_j = Z^T Z.$$

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Correspondence Analysis, the Row Profiles

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Let λ_i denote the i th largest eigenvalue of the matrix V and let u_i denote the corresponding unit length eigenvector. Let $u_{i,k}$ denote the k th element of u_i . The score of the row profile j (associated with modality A_j) on the i th CA component is given by

$$\phi_{i,j} = \sum_{k=1}^K u_{i,k} R_{jk}.$$

It can be proven that ϕ_i is centered such that

$$\sum_{j=1}^J f_j \cdot \phi_{i,j} = 0,$$

and that the variance of ϕ_i is λ_i .

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Contribution of the Modalities

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

The contribution of the modality A_j on construction of the axis u_i is given by

$$\frac{f_{j.}(\phi_{i,j})^2}{\lambda_i}.$$

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Quality of the Representation

The quality of the representation of the centered row profile R_j by the CA axis i is measured by the squared cosine of angle between the vector OR_j and u_i :

$$\cos^2(\alpha) = \left(\frac{\langle OR_j, u_i \rangle}{\|OR_j\| \cdot \|u_i\|} \right)^2 = \frac{(\phi_{i,j})^2}{\|OR_j\|^2}.$$

If the value is close to 1, the quality of the representation is good.

Note that the formula above does not contain the weight f_j , and thus one modality can be:

- Close to the axis u_i and therefore be well represented (well explained).
- Due to a low weight f_j , it can have a low contribution to the axis.

Correspondence Analysis, the Column Profiles

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Correspondence Analysis, the Column Profiles

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Performing correspondence analysis on the column profiles does not differ from performing correspondence analysis on the row profiles. The solution is given by the eigenvalues and the eigenvectors of the matrix $W = ZZ^T$.

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Correspondence Analysis, the Column Profiles

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Let $C \in \mathbb{R}^{J \times K}$, where

$$C_{jk} = \frac{f_{jk}}{f_{.k} \sqrt{f_{.j}}} - \sqrt{f_{.j}}$$

The matrix C contains scaled and shifted column profiles. Let C_k denote the k th column of C . In correspondence analysis on the column profiles, one finds orthonormal vectors (directions) v_h such that projection $P_h(\cdot)$ onto v_h maximizes the weighted sum of the euclidian distances,

$$\sum_{k=1}^K f_{.k} d^2(0, P_h(C_k)),$$

under the constraint that v_h is orthogonal to all v_l , $1 \leq l < h$. The solution is given by the eigenvalues and the eigenvectors of the matrix $W = ZZ^T$.

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Correspondence Analysis, the Column Profiles

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Let λ_h denote the h th largest eigenvalue of the matrix W and let v_h denote the corresponding unit length eigenvector. Let $v_{h,k}$ denote the k th element of v_h . The score of the column profile k (associated with modality B_k) on the h th CA component is given by

$$\psi_{h,k} = \sum_{j=1}^J v_{h,j} G_{jk}.$$

It can be proven that ψ_h is centered such that

$$\sum_{k=1}^K f_{.k} \psi_{h,k} = 0,$$

and that the variance of ψ_h is λ_h .

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Contribution of the Modalities

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

The contribution of the modality B_k on construction of the axis v_h is given by

$$\frac{f_{.k}(\psi_{h,k})^2}{\lambda_h}.$$

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Quality of the Representation

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

The quality of the representation of the centered column profile C_k by the CA axis h is measured by the squared cosine of angle between the vector OC_k and v_h .

$$\cos^2(\beta) = \left(\frac{\langle OC_k, v_h \rangle}{\|OC_k\| \cdot \|v_h\|} \right)^2 = \frac{(\psi_{h,k})^2}{\|OC_k\|^2}.$$

If the value is close to 1, the quality of the representation is good.

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Association Between the Profiles

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Association Between the Profiles

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

It can be shown that the matrices V and W have the same nonzero eigenvalues. Moreover, the eigenvectors u_i can be given in terms of v_i and vice versa:

$$u_i = \frac{1}{\sqrt{\lambda_i}} Z^T v_i$$

and

$$v_i = \frac{1}{\sqrt{\lambda_i}} Z u_i.$$

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Association Between the Profiles

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Let $H = \text{rank}(V) = \text{rank}(W)$. The coolest thing in correspondence analysis is that the attraction-repulsion indices d_{jk} can be given in terms of ϕ and ψ as follows

$$d_{jk} = 1 + \sum_{h=1}^H \frac{1}{\sqrt{\lambda_h}} \phi_{h,j} \psi_{h,k}.$$

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Association Between the Profiles

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

The scores are often standardized defining

$$\hat{\psi}_{h,k} = \frac{1}{\sqrt{\lambda_h}} \psi_{h,k}$$

and

$$\hat{\phi}_{h,j} = \frac{1}{\sqrt{\lambda_1}} \phi_{h,j}.$$

Then

$$d_{jk} = 1 + \sqrt{\lambda_1} \sum_{h=1}^H \hat{\phi}_{h,j} \hat{\psi}_{h,k}.$$

The attraction-repulsion index d_{jk} is now larger than 1 if and only if the smallest angle between $(\hat{\phi}_{1,j}, \dots, \hat{\phi}_{H,j})$ and $(\hat{\psi}_{1,k}, \dots, \hat{\psi}_{H,k})$ is less than 90° .

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

If the row profile j and the column profile k are well represented by the first two CA components, then the attraction-repulsion index

$$d_{jk} \approx 1 + \sqrt{\lambda_1} \sum_{h=1}^2 \hat{\phi}_{h,j} \hat{\psi}_{h,k}.$$

We can therefore say that the modalities A_j and B_k are attracted to each if the angle between $(\hat{\phi}_{1,j}, \hat{\phi}_{2,j})$ and $(\hat{\psi}_{1,k}, \hat{\psi}_{2,k})$ is less than 90° and they repulse each other if the angle between $(\hat{\phi}_{1,j}, \hat{\phi}_{2,j})$ and $(\hat{\psi}_{1,k}, \hat{\psi}_{2,k})$ is larger than 90° . In this case, one can simply observe the angle from the (double) biplot of the first two components of $\hat{\phi}$ and $\hat{\psi}$.

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Example of Correspondence Analysis

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Correspondence analysis using the data presented in lecture five. Variable x Education is divided to categories A_1 Primary School, A_2 High School, and A_3 University, and variable y Salary is divided to categories B_1 low, B_2 average, and B_3 high.

	B_1	B_2	B_3	
A_1	150	40	10	200
A_2	190	350	60	600
A_3	10	110	80	200
	350	500	150	1000

Table: Contingency table

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

Example of Correspondence Analysis

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

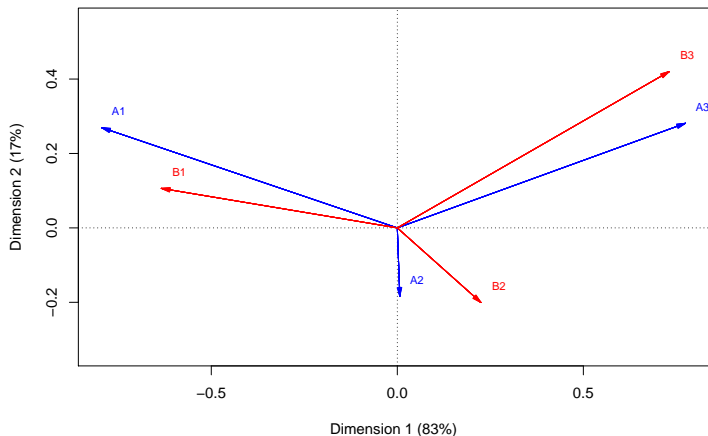


Figure: Salary and education (A1=Primary School education, A2=High School education, A3=University level education, B1=low salary, B2=average salary, B3=high salary)

Correspondence Analysis
Chi-square Test Statistic
Chi-square Distances
Correspondence Analysis, the Row Profiles
Correspondence Analysis, the Column Profiles
Association Between the Profiles
References

Next Week

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Next week we will talk about multiple correspondence analysis (MCA).

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

References

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles


Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

References I

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

 K. V. Mardia, J. T. Kent, J. M. Bibby, *Multivariate Analysis*, Academic Press, London, 2003 (reprint of 1979).

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

References II

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

-  R. V. Hogg, J. W. McKean, A. T. Craig, Introduction to Mathematical Statistics, Pearson Education, Upper Sadle River, 2005.
-  R. A. Horn, C. R. Johnson, Matrix Analysis, Cambridge University Press, New York, 1985.
-  R. A. Horn, C. R. Johnson, Topics in Matrix Analysis, Cambridge University Press, New York, 1991.

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References

References III

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

 L. Simar, An Introduction to Multivariate Data Analysis,
Université Catholique de Louvain Press, 2008.

Correspondence
Analysis

Chi-square Test
Statistic

Chi-square Distances

Correspondence
Analysis, the Row
Profiles

Correspondence
Analysis, the Column
Profiles

Association Between
the Profiles

References