

MS-E2112 Multivariate Statistical Analysis (5cr)

Lecture 6: Bivariate Correspondence Analysis - part II

Lecturer: Pauliina Ilmonen
Slides: Ilmonen/Kantala

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

Contents

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Correspondence Analysis

Chi-square Distances

Correspondence Analysis, Row Profiles

Correspondence Analysis, Column Profiles

Association Between the Profiles

References

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

Correspondence Analysis

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, 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 looking at chi-square distances between the row (or column) profiles of the variables.

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

Chi-square Distances

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

Chi-square Distance

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

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

Euclidean distance gives the same weight to each column. The chi-square distance gives the same relative importance to each column proportionally to the marginal relative row frequency. The division of each squared term by the marginal relative column 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.

Chi-square Distance

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

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 Euclidean metric can be used on the new scaled data.

Example, Chi-square Distance

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

	B_1	B_2	
A_1	0.10	0.20	0.30
A_2	0.20	0.40	0.60
A_3	0.01	0.09	0.10
	0.31	0.69	1

Table: Relative frequencies

- The chi-square distances between the first and the second row profile is $\frac{1}{0.31}(\frac{0.1}{0.3} - \frac{0.2}{0.6})^2 + \frac{1}{0.69}(\frac{0.2}{0.3} - \frac{0.4}{0.6})^2 = 0$.
- The chi-square distances between the second and the third row profile is $\frac{1}{0.31}(\frac{0.2}{0.6} - \frac{0.01}{0.1})^2 + \frac{1}{0.69}(\frac{0.4}{0.6} - \frac{0.09}{0.1})^2$
- Note that the chi-square distances between the second and the third row profile is equal to the chi-square distances between the first and the third row profile.

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

Chi-square Distance

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

The distance between two columns k_1 and k_2 is given by

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

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

Correspondence Analysis, Row Profiles

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

Correspondence Analysis, Row Profiles

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Recall that traditional principal component analysis is based on maximizing Euclidean distances. As discussed, in the context of frequency distributions, the proper distance between the variables is the chi-square distance. Thus, in correspondence analysis, a PCA type approach is applied to modified data. Instead of the original relative frequencies f_{jk} , we work on scaled relative frequencies

$$\frac{f_{jk}}{f_{j.} \sqrt{f_{.k}}}.$$

The scaling here is the scaling used in calculating the chi-square distances between the rows. **Correspondence analysis is based on maximizing chi-square distances.**

Note that the relative row frequency weighted sum

$$\sum_{j=1}^J f_{j.} \frac{f_{jk}}{f_{j.} \sqrt{f_{.k}}} = \sqrt{f_{.k}}.$$

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

Correspondence Analysis, Row Profiles

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

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

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

Let R_j denote the j th row of R and let

$$V = \sum_{j=1}^J f_{j.} R_j^T R_j.$$

The matrix R now contains the scaled and centered relative frequencies and the matrix V is a relative row frequency weighted covariance matrix of the rows of R . The data is centered using the relative row frequency weighted mean and the observations are scaled by relative row frequencies. (In traditional covariance matrix the scale is $\frac{1}{n}$.)

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

The maximization problem in correspondence analysis 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 .

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

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,

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

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

The vectors u_i are the eigenvectors of the matrix V . In constructing the matrices R and V , 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 Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

Remark

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Recall the matrix $Z \in \mathbb{R}^{J \times K}$, where

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

that is connected to the chi-square independence test. One can show that the matrix

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

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

Score Vectors

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

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

The score vector ϕ_i is centered such that

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

and the variance of ϕ_i is λ_i .

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

Contribution of the Modalities

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

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

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

Quality of the Representation

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

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

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

Correspondence Analysis, Column Profiles

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

Correspondence Analysis, Column Profiles

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Correspondence analysis on the column profiles is conducted exactly as correspondence analysis on the row profiles.

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

Correspondence Analysis, 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 and let

$$W = \sum_{k=1}^K f_{.k} C_k C_k^T.$$

The matrix C now contains the scaled and centered relative frequencies and the matrix W is a relative column frequency weighted covariance matrix of the rows of C .

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

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 Euclidean 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 Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

Score Vectors

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

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

The score vector ψ_h is centered such that

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

and the variance of ψ_h is λ_h .

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

Contribution of the Modalities

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

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

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

Quality of the Representation

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

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.

Association Between the Profiles

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, 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 Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

Association Between the Profiles

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

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

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

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 Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, 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 Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, 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 Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

Example of Correspondence Analysis

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

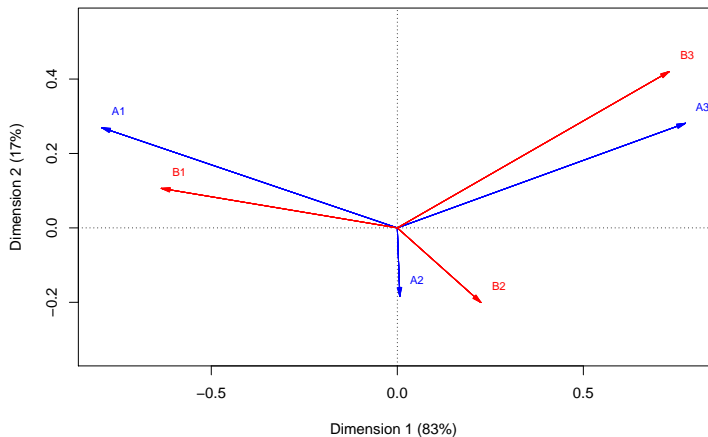


Figure: Salary and education (A1=Primary School education, A2=High School education, A3=University level education, B1=low salary, B2=average salary, B3=high salary)

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

Next Week

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

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

References

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

References I

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Correspondence
Analysis


Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles


Association Between
the Profiles

References

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

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 Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

References III

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Correspondence
Analysis

Chi-square Distances

Correspondence
Analysis, Row Profiles

Correspondence
Analysis, Column
Profiles

Association Between
the Profiles

References

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