MS-E2112 Multivariate Statistical Analysis (5cr)
Lecture 7: Multiple Correspondence Analysis

Lecturer: Pauliina Ilmonen
Slides: Ilmonen/Kantala

Contents

Multiple Correspondence Analysis

Multiple Correspondence Analysis

Multiple correspondence analysis (MCA) is an extension of bivariate correspondence analysis to more than 2 variables.

Lecturer: Pauliina Ilmonen
Slides: Ilmonen/Kantala

Frequency Tables

Contingency Tables

We consider a sample of size n described by P qualitative variables $Y_1, ..., Y_P$. The variable Y_p has K_p modalities (categories), and $\sum_{p=1}^{P} K_p$ is the total number of the categories. The number of individuals having the modality l of the variable Y_p is denoted by n_{pl}. We set a variable $x_{ipl} = 1$ if individual i has modality l of Y_p, and we set $x_{ipl} = 0$ otherwise. Now

$$\sum_{l=1}^{K_p} n_{pl} = n,$$

and

$$\sum_{p=1}^{P} \sum_{l=1}^{K_p} n_{pl} = nP.$$

Table of Dummy Variables

The table of K_p dummy variables associated with variable Y_p.

	1	2	$\cdots$	K_p	
1	x_{1p1}	x_{1p2}	$\cdots$	x_{1pK_p}	1
2	x_{2p1}	x_{2p2}	$\cdots$	x_{2pK_p}	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{np1}	x_{np2}	$\cdots$	x_{npK_p}	1
	n_{p1}	n_{p2}	$\cdots$	n_{pK_p}	n

Table: Table of dummy variables

Complete Disjunctive Table

Now we introduce the $n \times K$ table/matrix $X = [X_1, \ldots, X_P]$, called the complete disjunctive table.

	X_1			$\cdots$	X_P			
	X_{11}	$\cdots$	X_{1K_1}	$\cdots$	X_{P1}	$\cdots$	X_{PK_P}	$\sum_{p=1}^{P} \sum_{l=1}^{K_p} x_{ipl}$
1	x_{111}	$\cdots$	x_{11K_1}	$\cdots$	x_{1P1}	$\cdots$	x_{1PK_P}	P
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	x_{i11}	$\cdots$	x_{i1K_1}	$\cdots$	x_{iP1}	$\cdots$	x_{iPK_P}	P
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{n11}	$\cdots$	x_{n1K_1}	$\cdots$	x_{nP1}	$\cdots$	x_{nPK_P}	P
$\sum_{i=1}^{n} x_{ipl}$	n_{11}	$\cdots$	n_{1K_1}	$\cdots$	n_{P1}	$\cdots$	n_{PK_P}	nP

Table: Complete disjunctive table

Example: Party Snacks

A group of kids were asked to select party snacks. Each kid chose one cookie, one milk shake and one salty snack. Here we have a sample of 4 individuals and 3 variables — $n = 4, P = 3$.

- Variable X_1 cookie has two options (modalities/categories) — chocolate chip cookie (1) and oat cookie (2).
- Variable X_2 milk shake has three options — vanilla (1), strawberry (2), and chocolate (3).
- Variable X_3 salty snack has two options — pop corn (1), and potato chips (2).

Now $K = K_1 + K_2 + K_3 = 2 + 3 + 2 = 7$.

Example: Party Snacks

We display the party snack data as a complete disjunctive table.

	X_{11}	X_{12}	X_{21}	X_{22}	X_{23}	X_{31}	X_{32}	$\sum_{p=1}^{7}\sum_{l=1}^{K_p} x_{ipl}$
1	0	1	1	0	0	1	0	3
2	0	1	1	0	0	0	1	3
3	1	0	0	0	1	1	0	3
4	0	1	0	1	0	0	1	3
$\sum_{i=1}^{n} x_{ipl}$	1	3	2	1	1	2	2	12

Table: Complete disjunctive table

- The first kid chose an oat cookie, vanilla milk shake and pop corn.
- The third kid chose a chocolate chip cookie, chocolate milk shake and pop corn.

Multiple Correspondence Analysis

Bivariate correspondence analysis is now applied to the complete disjunctive table!

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Relative Frequency Tables

From the complete disjunctive table, one can compute the associated relative frequency table (F), where the elements of the complete disjunctive table are divided by the total sum nP leading to

$$f_{ipl} = \frac{x_{ipl}}{nP} \quad (i = 1, \ldots, n; p = 1, \ldots, P; l = 1, \ldots, K_p).$$

We have P variables and n individuals and $f_{ipl} = \frac{1}{nP}$ or $f_{ipl} = 0$. Thus the marginal relative frequencies are computed as

$$f_{i..} = \sum_{p=1}^{P} \sum_{l=1}^{K_p} \frac{x_{ipl}}{nP} = P \frac{1}{nP} = \frac{1}{n}$$

and

$$f_{.pl} = \sum_{i=1}^{n} \frac{x_{ipl}}{nP} = \frac{n_{pl}}{nP}.$$

Example: Party Snacks

We display the party snacks data as a relative frequency table.

	X_{11}	X_{12}	X_{21}	X_{22}	X_{23}	X_{31}	X_{32}	$f_{i..}$
1	0	$\frac{1}{12}$	$\frac{1}{12}$	0	0	$\frac{1}{12}$	0	$\frac{1}{4}$
2	0	$\frac{1}{12}$	$\frac{1}{12}$	0	0	0	$\frac{1}{12}$	$\frac{1}{4}$
3	$\frac{1}{12}$	0	0	0	$\frac{1}{12}$	$\frac{1}{12}$	0	$\frac{1}{4}$
4	0	$\frac{1}{12}$	0	$\frac{1}{12}$	0	0	$\frac{1}{12}$	$\frac{1}{4}$
$f_{.pl}$	$\frac{1}{12}$	$\frac{3}{12}$	$\frac{2}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{2}{12}$	$\frac{2}{12}$	1

Table: Relative frequency table

Row Profiles

Row Profiles

The idea behind MCA, like in bivariate correspondence analysis, is to apply a PCA type approach on one hand to the row profiles, and on the other hand to the column profiles of the relative frequencies table F. The coordinate pl of the row profile $l_i(1 \times K)$ associated with individual i is given as

$$(l_i)_{pl} = \frac{f_{ipl}}{f_{i..}} = \frac{\frac{x_{ipl}}{nP}}{\frac{1}{n}} = \frac{x_{ipl}}{nP}\frac{n}{1} = \frac{x_{ipl}}{P}, \qquad i = 1, \ldots, n.$$

As

$$\sum_{i=1}^{n} \frac{1}{n}(l_i)_{pl} = \sum_{i=1}^{n} \frac{1}{n}\frac{x_{ipl}}{P} = \frac{n_{pl}}{nP},$$

the n row profiles weighted by the marginal relative frequencies $(1/n)$ compose a point cloud in $\mathbb{R}^K$ with a center given by the relative marginal profile

$$G_l = (\frac{n_{11}}{nP}, \ldots, \frac{n_{1K_1}}{nP}, \ldots, \frac{n_{P1}}{nP}, \ldots, \frac{n_{PK_P}}{nP}).$$

Example: Party Snacks

The row profiles of the party snacks data is given below.

	X_{11}	X_{12}	X_{21}	X_{22}	X_{23}	X_{31}	X_{32}	
1	0	$\frac{1}{3}$	$\frac{1}{3}$	0	0	$\frac{1}{3}$	0	1
2	0	$\frac{1}{3}$	$\frac{1}{3}$	0	0	0	$\frac{1}{3}$	1
3	$\frac{1}{3}$	0	0	0	$\frac{1}{3}$	$\frac{1}{3}$	0	1
4	0	$\frac{1}{3}$	0	$\frac{1}{3}$	0	0	$\frac{1}{3}$	1

Table: Row profiles

Row Profiles

Intuitively, the distance between two individuals is small if they have many modalities in common, and the distance between the individual i and the center increases as the modalities taking by the individual i becomes rare ($x_{ipl} = 1$ for n_{pl} small).

Row Profiles

More formally, the chi-square distances between two row profiles l_{i_1} and l_{i_2} can be given as

$$d^2(l_{i_1}, l_{i_2}) = \sum_{p=1}^{P} \sum_{l=1}^{K_P} \frac{1}{f_{.pl}} ((l_{i_1})_{pl} - (l_{i_2})_{pl})^2$$

$$= \frac{n}{P} \sum_{p=1}^{P} \sum_{l=1}^{K_P} \frac{1}{n_{pl}} (x_{i_1 pl} - x_{i_2 pl})^2.$$

Example: Party Snacks

The distance between the first kid and the second kid in the party snacks data is

$$(\frac{4}{3}(1(0-0)^2+\frac{1}{3}(1-1)^2+\frac{1}{2}(1-1)^2+1(0-0)^2+1(0-0)^2+\frac{1}{2}(1-0)^2+\frac{1}{2}(0-1)^2))$$

$$=\frac{4}{3}\approx 1.33.$$

Column Profiles

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Column Profiles

The coordinate i of the column profile c_{pl} $(n \times 1)$ associated with the modality l of Y_p is given as

$$(c_{pl})_i = \frac{f_{ipl}}{f_{.pl}} = \frac{\frac{x_{ipl}}{nP}}{\frac{n_{pl}}{nP}} = \frac{x_{ipl}}{nP}\frac{nP}{n_{pl}} = \frac{x_{ipl}}{n_{pl}}, \qquad p = 1, \ldots, P; l = 1, \ldots, K_p.$$

As

$$\sum_{p=1}^{P}\sum_{l=1}^{K_p} f_{.pl}(c_{pl})_i = \sum_{p=1}^{P}\sum_{l=1}^{K_p} \frac{n_{pl}}{nP}\frac{x_{ipl}}{n_{pl}} = \sum_{p=1}^{P}\sum_{l=1}^{K_p} \frac{x_{ipl}}{nP} = \frac{P}{nP} = \frac{1}{n},$$

the K column profiles weighted by the marginal relative frequencies $(\frac{n_{pl}}{nP})$ compose a point cloud in $\mathbb{R}^K$ with the center given by the relative marginal profile $G_c = (\frac{1}{n}, \ldots, \frac{1}{n})$.

Example: Party Snacks

The column profiles of the party snacks data are given below.

	X_{11}	X_{12}	X_{21}	X_{22}	X_{23}	X_{31}	X_{32}
1	0	$\frac{1}{3}$	$\frac{1}{2}$	0	0	$\frac{1}{2}$	0
2	0	$\frac{1}{3}$	$\frac{1}{2}$	0	0	0	$\frac{1}{2}$
3	1	0	0	0	1	$\frac{1}{2}$	0
4	0	$\frac{1}{3}$	0	1	0	0	$\frac{1}{2}$
	1	1	1	1	1	1	1

Table: Column profiles

Column Profiles

Intuitively, the χ^2 distance between two modalities is small if the same individuals take these two modalities together, and the distance between the modality l of Y_p and the center increases as the modality becomes more rare (n_{pl} small).

Column Profiles

More formally, the chi-square distances between two column profiles $c_{p_1 l_1}$ and $c_{p_2 l_2}$ can be given as

$$d^2(c_{p_1 l_1}, c_{p_2 l_2}) = \sum_{i=1}^{n} \frac{1}{f_{i..}} ((c_{p_1 l_1})_i - (c_{p_2 l_2})_i)^2$$

$$= n \sum_{i=1}^{n} \left(\frac{x_{ip_1 l_1}}{n_{p_1 l_1}} - \frac{x_{ip_2 l_2}}{n_{p_2 l_2}}\right)^2.$$

Example: Party Snacks

The distance between modality 1 of Y_1 (chocolate chip cookie) and modality 2 of Y_2 (strawberry milk shake) is

$$4((0-0)^2 + (0-0)^2 + (1-0)^2 + (0-1)^2) = 8$$

Attraction Repulsion Indices

Attraction Repulsion Indices

Let $n_{p_1 l_1, p_2 l_2}$ be the number of individuals having the modality l_1 of the variable Y_{p_1} and the modality l_2 of the variable Y_{p_2}. Now the attraction repulsion index $d_{p_1 l_1, p_2 l_2}$ between the modality l_1 of the variable Y_{p_1} and the modality l_2 of the variable Y_{p_2} is given by

$$d_{p_1 l_1, p_2 l_2} = \frac{n_{p_1 l_1, p_2 l_2}/n}{n_{p_1 l_1}/n n_{p_2 l_2}/n} = \frac{n_{p_1 l_1, p_2 l_2}}{\frac{n_{p_1 l_1} n_{p_2 l_2}}{n}}.$$

Attraction Repulsion Indices

It is clear that if the attraction repulsion index is larger than one, the individuals are more inclined to take both modalities simultaneously than under the hypothesis of independence. And vice-versa, if the attraction repulsion index is smaller than one, the individuals are less inclined to take both modalities simultaneously than under the hypothesis of independence.

Attraction Repulsion Indices

The attraction repulsion index $d_{i,pl}$ between the individual i and the modality l of the variable Y_p is defined as follows.

$$d_{i,pl} = \frac{f_{ipl}}{f_{i..}f_{.pl}} = \frac{x_{ipl}}{n_{pl}/n}.$$

Now, clearly

$$d_{i,pl} = 0,$$

if $x_{ipl} = 0$ and

$$d_{i,pl} = \frac{n}{n_{pl}},$$

if $x_{ipl} = 1$. Thus, if the individual i does not have the modality l of the variable Y_p, then the attraction repulsion index $d_{i,pl}$ is equal to 0, and if the individual i does have the modality l of Y_p, then the attraction repulsion index $d_{i,pl}$ increases as the l of Y_p becomes rare.

Multiple Correspondence Analysis

Multiple Correspondence Analysis

To maximize chi-square distances and to obtain a representation in lower dimension, PCA type transformation is applied on the two data clouds: the row profiles and the column profiles. A transformation of the profiles is necessary to center the variables, and to be able to base the maximization problem on euclidian distances instead of χ^2 distances directly:

$$(l_i^\circ)_{pl} = \frac{(l_i)_{pl}}{\sqrt{f_{.pl}}} - \sqrt{f_{.pl}} \text{ and } (c_{pl}^\circ)_i = \frac{(c_{pl})_i}{\sqrt{f_{i..}}} - \sqrt{f_{i..}}.$$

Solution

The solution of the problem of maximization associated with the transformed row and column profiles is given respectively by the eigenvalues and the eigenvectors of the matrices $V(K \times K)$ and $W(n \times n)$ where

$V = T^T T$ and $W = TT^T$ where the elements of T are given by $\dfrac{x_{ipl} - n_{pl}/n}{\sqrt{Pn_{pl}}}$.

Note that here also, the matrix V is a relative row frequency weighted covariance matrix of the scaled and shifted row profiles and the matrix W is a relative column frequency weighted covariance matrix of the scaled and shifted column profiles.

Multiple Correspondence Analysis

The MCA components for the individuals are derived from the eigenvectors of the matrix V, and the MCA components for the modalities from the eigenvectors of the matrix W.

Let $H = rank(V) = rank(W)$. The scores of the individuals are given as

$$\phi_{h,i} = \sum_{k=1}^{K} u_{h,k} (l_i^{\circ})_k \qquad h = 1, \ldots, H,$$

where $u_{h,k}$ is the kth element of the eigenvector associated with the hth largest eigenvalues of V.

The scores for the modalities are given as

$$\psi_{h,pl} = \sum_{i=1}^{n} v_{h,i} (c_{pl}^{\circ})_i \qquad h = 1, \ldots, H.$$

Contribution of the Modalities

Contribution of the modality l of Y_p on the variance of the new variable ψ_h is given by

$$C(pl, h) = \frac{f_{.pl}\psi^2_{h,pl}}{\lambda_h} = \frac{n_{pl}\psi^2_{h,pl}}{nP\lambda_h}.$$

Global contribution of the variable Y_p is given by

$$C(p, h) = \sum_{l=1}^{K_p} C(pl, h).$$

Graphical Presentation

Comparison of the Modalities

The attraction repulsion index

$$d_{p_1 l_1, p_2 l_2} = 1 + \sum_{h=1}^{H} \psi_{h,p_1 l_1} \psi_{h,p_2 l_2}.$$

The graphical output of MCA is the approximation of the previous formula using few dimensions. Suppose that the modalities are well represented in two dimensions. Then we can plot the two first MCA components and interpret the proximity between the points on the first principal plan with the following approximation

$$d_{p_1 l_1, p_2 l_2} \approx 1 + \sum_{h=1}^{2} \psi_{h,p_1 l_1} \psi_{h,p_2 l_2}.$$

Comparison of the Individuals

The proximity between two individuals i_1 and i_2 is defined as

$$d_{i_1,i_2} = 1 + \sum_{h=1}^{H} \phi_{h,i_1}\phi_{h,i_2}.$$

Two individuals are close if they have in general the same modalities.

Now d_{i_1,i_2} can be approximated by

$$d_{i_1,i_2} \approx 1 + \sum_{h=1}^{2} \phi_{h,i_1}\phi_{h,i_2}.$$

Simultaneous Comparison

The attraction repulsion index

$$d_{i,pl} = 1 + \sum_{h=1}^{H} \frac{1}{\sqrt{\lambda_h}} \phi_{h,i} \psi_{h,pl},$$

and thus again

$$d_{i,pl} \approx 1 + \sum_{h=1}^{2} \frac{1}{\sqrt{\lambda_h}} \phi_{h,i} \psi_{h,pl}.$$

Simultaneous Comparison

The scores are often standardized defining

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

and

$$\hat{\phi}_{2,j} = \frac{1}{\sqrt{\lambda_2}} \phi_{2,j}.$$

Then

$$d_{i,pl} \approx 1 + \sum_{h=1}^{2} \hat{\phi}_{h,i} \psi_{h,pl},$$

and the final graphical representation can be given simultaneously as a double biplot.

Example

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Example of MCA: Extended Party Snack Data

Disclaimer: This example data set is randomly generated. Please do not draw real life conclusions from it.

	X_{11}	X_{12}	X_{21}	X_{22}	X_{23}	X_{31}	X_{32}	$\sum_{p=1}^{7}\sum_{l=1}^{K_p} x_{ipl}$
1	0	1	1	0	0	1	0	3
2	0	1	1	0	0	0	1	3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
25	1	0	0	0	1	0	1	3
$\sum_{i=1}^{n} x_{ipl}$	16	9	9	6	10	14	11	

Table: Complete disjunctive table

Example of MCA

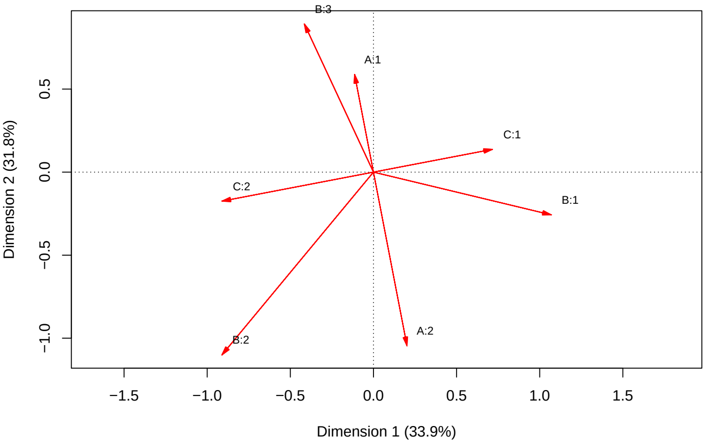

Figure: Result of MCA (A1=chocolate chip cookie, A2=oat cookie, B1=vanilla milk shake, B2=strawberry milk shake, B3=chocolate milk shake, C1=pop corn, C2=potato chips.) It seems that kids that like chocolate chip cookies like chocolate milk shake as well.

Some Remarks

Some Remarks

When performing MCA, it is better to take into account variables that have more or less the same number of modalities. (The number of modalities has an effect on the analysis.) It is also advised to avoid having very rare modalities. (Rare modalities have a big impact on analysis, and that makes MCA quite nonrobust method.) One can preprocess the data by grouping modalities if necessary.

Next Week

Next week we will talk about canonical correlation analysis.

References

References I

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

References II

- 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.

References III

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