Principal Component Analysis for Mixed Quantitative and Qualitative Data

Susana Agudelo-Jaramillo   Manuela Ochoa-Muñoz

Tutor: Francisco Iván Zuluaga-Díaz

EAFIT University
Medellín-Colombia

Research Practise
April 12th, 2016

Mixed Quantitative and Qualitative Data

Quantitative

There are many methods to analyze pure quantitative data.
→ Principal Component Analysis.

Qualitative

There exist also several techniques to deal with pure qualitative data.
→ Correspondence Analysis.

PCAMIX

Correspondence Analysis

→ It is a graphical technique to represent information of a contingency table with two inputs, which contains the count of elements for cross-classification of two categorical variables.

→ These tables are based on two qualitative nominal or ordinal variables where categories of one variable appear in rows and other variable categories are represented in columns [de la Fuente Fernández, 2011].

→ Correspondence analysis can be useful to identify categories that are similar, which therefore can be combined.

Example

The following example illustrated how a quantification matrix works for a sample of 12 people and 4 categorical variables.

Figure 1: Categories for the four variables taken from [Rencher, 1934]

Variable	Levels
Gender	Male, female
Age	Young, middle-aged, old
Marital status	Single, married
Hair color	Blond, brown, black, red

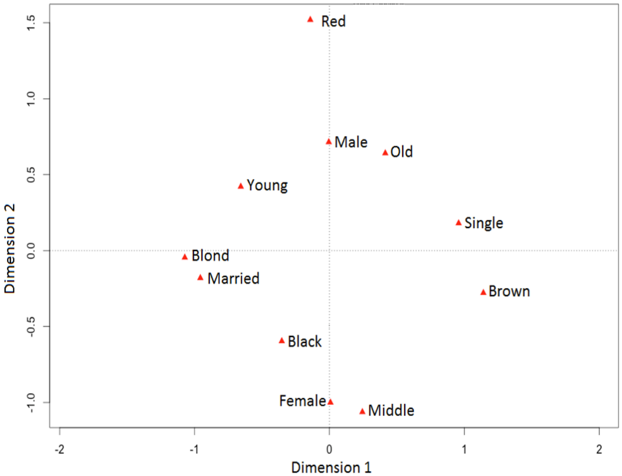
Example

Figure 2: List of 12 people and their categories on four variables taken from [Rencher, 1934]

Person	Gender	Age	Marital Status	Hair Color
1	Male	Young	Single	Brown
2	Male	Old	Single	Red
3	Female	Middle	Married	Blond
4	Male	Old	Single	Black
5	Female	Middle	Married	Black
6	Female	Middle	Single	Brown
7	Male	Young	Married	Red
8	Male	Old	Married	Blond
9	Male	Middle	Single	Blond
10	Female	Young	Married	Black
11	Female	Old	Single	Brown
12	Male	Young	Married	Blond

Example

Figure 3: Correspondence analysis of the four variables



Indicator Matrix

$$S_{ij} = \begin{cases} 1 & \text{if object } i \text{ belongs to the category of the variable } j \\ 0 & \text{if object } i \text{ does not belong to the category of the variable } j \end{cases}$$

Figure 4: Indicator matrix G for the data taken from [Rencher, 1934]

Person	Gender	Age	Marital Status	Hair Color
1	1 0	1 0 0	1 0	0 1 0 0
2	1 0	0 0 1	1 0	0 0 0 1
3	0 1	0 1 0	0 1	1 0 0 0
4	1 0	0 0 1	1 0	0 0 1 0
5	0 1	0 1 0	0 1	0 0 1 0
6	0 1	0 1 0	1 0	0 1 0 0
7	1 0	1 0 0	0 1	0 0 0 1
8	1 0	0 0 1	0 1	1 0 0 0
9	1 0	0 1 0	1 0	1 0 0 0
10	0 1	1 0 0	0 1	0 0 1 0
11	0 1	0 0 1	1 0	0 1 0 0
12	1 0	1 0 0	0 1	1 0 0 0

From the indicator matrix G we can get the G'G matrix known as the Burt matrix.

Figure 5: Burt Matrix G'G for the matrix G taken from [Rencher, 1934]

Gender		Age			Marital Status		Hair Color			
7	0	3	1	3	4	3	3	1	1	2
0	5	1	3	1	2	3	1	2	2	0
3	1	4	0	0	1	3	1	1	1	1
1	3	0	4	0	2	2	2	1	1	0
3	1	0	0	4	3	1	1	1	1	1
4	2	1	2	3	6	0	1	3	1	1
3	3	3	2	1	0	6	3	0	2	1
3	1	1	2	1	1	3	4	0	0	0
1	2	1	1	1	3	0	0	3	0	0
1	2	1	1	1	1	2	0	0	3	0
2	0	1	0	1	1	1	0	0	0	2

Burt Matrix

In the diagonal blocks appear matrices containing the marginal frequencies of each of the variables analyzed.

Outside the diagonal appear contingency tables of frequencies corresponding to all combinations 2 to 2 of the variables analyzed.

Figure 6: Part of the contingency tables for variables Gender and Age

Gender		Age		
7	0	3	1	3
0	5	1	3	1
3	1	4	0	0
1	3	0	4	0
3	1	0	0	4

Quantification Matrices

Quantification matrices transform qualitative data into components which facilitates the analysis of results.

- → The idea of using quantification matrices is to define correlation coefficients.
- → The quantification matrices are used to measure similarity and dissimilarity between the objects respect to a variable.

Quantification Matrix $G_j G_j'$

The elements of the quantification matrix $G_j G_j'$ are given by:

$$S_{ii'j} = \begin{cases} 1 & \text{if object } i \text{ and object } i' \text{ belong to the same category} \\ 0 & \text{if object } i \text{ and object } i' \text{ belong to different category} \end{cases}$$

$S_{ii'j}$ it is a measure of similarity between sample objects i and i' in terms of a particular variable j.

The frequency categories and the number of categories are not taken into account in this measure of similarity [Kiers, 1989].

Quantification Matrix $G_j G_j'$

Table 1: Quantification matrix GG' of hair color variable

Hair Color											
1	0	0	0	0	1	0	0	0	0	1	0
0	1	0	0	0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	1	1	0	0	1
0	0	0	1	1	0	0	0	0	1	0	0
0	0	0	1	1	0	0	0	0	1	0	0
1	0	0	0	0	1	0	0	0	0	1	0
0	1	0	0	0	0	1	0	0	0	0	0
0	0	1	0	0	0	0	1	1	0	0	1
0	0	1	0	0	0	0	1	1	0	0	1
0	0	0	1	1	0	0	0	0	1	0	0
1	0	0	0	0	1	0	0	0	0	1	0
0	0	1	0	0	0	0	1	1	0	0	1

Example

Figure 7: List of 12 people and their categories on four variables taken from [Rencher, 1934]

Person	Gender	Age	Marital Status	Hair Color
1	Male	Young	Single	Brown
2	Male	Old	Single	Red
3	Female	Middle	Married	Blond
4	Male	Old	Single	Black
5	Female	Middle	Married	Black
6	Female	Middle	Single	Brown
7	Male	Young	Married	Red
8	Male	Old	Married	Blond
9	Male	Middle	Single	Blond
10	Female	Young	Married	Black
11	Female	Old	Single	Brown
12	Male	Young	Married	Blond

Quantification Matrix $G_j(G_j'G_j)^{-1}G_j'$

In this case Burt matrix inverted is added:

Table 2: Burt matrix inverted of hair color variable

	Hair Color			
	Blond	Brown	Black	Red
Blond	0.25	0	0	0
Brown	0	0.33	0	0
Black	0	0	0.33	0
Red	0	0	0	0.5

Quantification Matrix $G_j(G_j'G_j)^{-1}G_j'$

The elements of the quantification matrix $G_j(G_j'G_j)^{-1}G_j'$ are given by:

$$S_{ii'j} = \begin{cases} f_g^{-1} & \text{if object } i \text{ and object } i' \text{ belong to the same category} \\ 0 & \text{if object } i \text{ and object } i' \text{ belong to different category} \end{cases}$$

where f_g^{-1} is the g^{th} diagonal element of $(G_j'G_j)^{-1}$ [Kiers, 1989].

Quantification Matrix $G_j(G_j'G_j)^{-1}G_j'$

Table 3: Quantification matrix $G(G'G)^{-1}G'$ of hair color variable

Hair Color											
0.33	0	0	0	0	0.33	0	0	0	0	0.33	0
0	0.5	0	0	0	0	0.5	0	0	0	0	0
0	0	0.25	0	0	0	0	0.25	0.25	0	0	0.25
0	0	0	0.33	0.33	0	0	0	0	0.33	0	0
0	0	0	0.33	0.33	0	0	0	0	0.33	0	0
0.33	0	0	0	0	0.33	0	0	0	0	0.33	0
0	0.5	0	0	0	0	0.5	0	0	0	0	0
0	0	0.25	0	0	0	0	0.25	0.25	0	0	0.25
0	0	0.25	0	0	0	0	0.25	0.25	0	0	0.25
0	0	0	0.33	0.33	0	0	0	0	0.33	0	0
0.33	0	0	0	0	0.33	0	0	0	0	0.33	0
0	0	0.25	0	0	0	0	0.25	0.25	0	0	0.25

Example

Figure 8: List of 12 people and their categories on four variables taken from [Rencher, 1934]

Person	Gender	Age	Marital Status	Hair Color
1	Male	Young	Single	Brown
2	Male	Old	Single	Red
3	Female	Middle	Married	Blond
4	Male	Old	Single	Black
5	Female	Middle	Married	Black
6	Female	Middle	Single	Brown
7	Male	Young	Married	Red
8	Male	Old	Married	Blond
9	Male	Middle	Single	Blond
10	Female	Young	Married	Black
11	Female	Old	Single	Brown
12	Male	Young	Married	Blond

Quantification Matrix $JG_j(G_j'G_j)^{-1}G_j'J$

Here the J matrix is added:

$$J{=}I_n - \frac{\mathbb{1}\mathbb{1}'}{n}$$

where I_n is the identity matrix, $\mathbb{1}$ is an ones vector and n is the sample size.

Quantification Matrix $JG_j(G_j'G_j)^{-1}G_j'J$

Table 4: J matrix

J Matrix											
0.9166	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833
-0.0833	0.9166	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833
-0.0833	-0.0833	0.9166	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833
-0.0833	-0.0833	-0.0833	0.9166	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833
-0.0833	-0.0833	-0.0833	-0.0833	0.9166	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833
-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	0.9166	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833
-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	0.9166	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833
-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	0.9166	-0.0833	-0.0833	-0.0833	-0.0833
-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	0.9166	-0.0833	-0.0833	-0.0833
-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	0.9166	-0.0833	-0.0833
-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	0.9166	-0.0833
-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833	0.9166

Quantification Matrix $JG_j(G_j'G_j)^{-1}G_j'J$

This quantification matrix is a normalized version of the χ^2 measure. Where $\chi^2 = 0$ if variables are statistically independent [Kiers, 1989].

The elements of the quantification matrix $JG_j(G_j'G_j)^{-1}G_j'J$ are given by:

$$S_{ii'j} = \begin{cases} f_g^{-1} - n^{-1} & \text{if object } i \text{ and object } i' \text{ belong to the same category} \\ -n^{-1} & \text{if object } i \text{ and object } i' \text{ belong to different category} \end{cases}$$

Quantification Matrix $JG_j(G_j'G_j)^{-1}G_j'J$

Table 5: Quantification matrix $JG(G'G)^{-1}G'J$ of hair color variable

Hair Color											
0.25	-0.0833	-0.0833	-0.0833	-0.0833	0.25	-0.0833	-0.0833	-0.0833	-0.0833	0.25	-0.0833
-0.0833	0.4166	-0.0833	-0.0833	-0.0833	-0.0833	0.4166	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833
-0.0833	-0.0833	0.1666	-0.0833	-0.0833	-0.0833	-0.0833	0.1666	0.1666	-0.0833	-0.0833	0.1666
-0.0833	-0.0833	-0.0833	0.25	0.25	-0.0833	-0.0833	-0.0833	-0.0833	0.25	-0.0833	-0.0833
-0.0833	-0.0833	-0.0833	0.25	0.25	-0.0833	-0.0833	-0.0833	-0.0833	0.25	-0.0833	-0.0833
0.25	-0.0833	-0.0833	-0.0833	-0.0833	0.25	-0.0833	-0.0833	-0.0833	-0.0833	0.25	-0.0833
-0.0833	0.4166	-0.0833	-0.0833	-0.0833	-0.0833	0.4166	-0.0833	-0.0833	-0.0833	-0.0833	-0.0833
-0.0833	-0.0833	0.1666	-0.0833	-0.0833	-0.0833	-0.0833	0.1666	0.1666	-0.0833	-0.0833	0.1666
-0.0833	-0.0833	0.1666	-0.0833	-0.0833	-0.0833	-0.0833	0.1666	0.1666	-0.0833	-0.0833	0.1666
-0.0833	-0.0833	-0.0833	0.25	0.25	-0.0833	-0.0833	-0.0833	-0.0833	0.25	-0.0833	-0.0833
0.25	-0.0833	-0.0833	-0.0833	-0.0833	0.25	-0.0833	-0.0833	-0.0833	-0.0833	0.25	-0.0833
-0.0833	-0.0833	0.1666	-0.0833	-0.0833	-0.0833	-0.0833	0.1666	0.1666	-0.0833	-0.0833	0.1666

Example

Figure 9: List of 12 people and their categories on four variables taken from [Rencher, 1934]

Person	Gender	Age	Marital Status	Hair Color
1	Male	Young	Single	Brown
2	Male	Old	Single	Red
3	Female	Middle	Married	Blond
4	Male	Old	Single	Black
5	Female	Middle	Married	Black
6	Female	Middle	Single	Brown
7	Male	Young	Married	Red
8	Male	Old	Married	Blond
9	Male	Middle	Single	Blond
10	Female	Young	Married	Black
11	Female	Old	Single	Brown
12	Male	Young	Married	Blond

References

de la Fuente Fernández, S. (2011).
Análisis correspondencias simples y múltiples.
Universidad Autónoma de Madrid, pages 1–9.

Kiers, H. (1989).
Three-way methods for the analysis of qualitative and quantitative two-way data.
PhD thesis.

Rencher, A. C. (1934).
Methods of Multivariate Analysis.
Wiley Series in Probability and Statistics.

THANKS FOR YOUR ATTENTION