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
June 7th, 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

Important Concepts

→ **Indicator Matrix**

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

→ **Burt Matrix**

From the indicator matrix G we can get the $G'G$ matrix known as the 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 [Rencher, 1934].

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

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

J Matrix

$$J = I_n - \frac{11'}{n}$$

where I_n is the identity matrix, 1 is an ones vector and n is the sample size.

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

This Quantification Matrix is selected for the PCAMIX method, due to the frequency categories and the number of categories are taken into account in this measure of similarity.

PCAMIX Method

$$W = \sum_{j=1}^{m} JG_j(G_j'G_j)^{-1}G_j'J$$

$$X = \frac{\lambda_1(W)}{\sqrt{n}}$$

$$Y = (G'G)^{-1}G'X$$

Quantification

$$Q = GY$$

where
$\lambda_1(W)$ is the largest eigenvalue of W,
n is the sample size,
m is the number of qualitative variables.

After finding the Q matrix, it is concatenated with the quantitative variables matrix to apply Principal Component Analysis [Kiers, 1991].

Principal Component Analysis

It considers a set of variables $x_1, x_2, ..., x_p$ upon a group of objects or individuals and based on them a new set of variables $y_1, y_2, ..., y_p$ is calculated, but these new variables are uncorrelated with each other and their variances should decrease gradually [Rencher, 1934].

Each y_j (where $j = 1, ..., p$) is a linear combination of original $x_1, x_2, ..., x_p$ described as follows:

$$y_j = a_{j1}x_1 + a_{j2}x_2 + ... + a_{jp}x_p = \mathbf{a}_j'\mathbf{x}$$

where $\mathbf{a}_j' = (a_{1j}, a_{2j}, ..., a_{pj})$ is a vector of constants, and

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$$

Application Case

For the application case an R database taken from PCAmix-data library and named "Gironde" is used. This database consists of 4 data sets characterizing the living conditions of 540 cities in Gironde – France.

Quantitative Variables:

Table 1: Quantitative variables of Gironde database

DATA SET	VARIABLES
Employment	Percentage of managers
	Average income
Natural environment	Percentage of buildings
	Percentage of water
	Percentage of vegetation

Application Case

Qualitative Variables:

Table 2: Qualitative variables of Gironde database

DATA SET	VARIABLES
Housing	Percentage of households
	Percentage of social housing
Services	Number of butcheries
	Number of bakeries
	Number of post offices
	Number of dental offices
	Number of supermarkets
	Number of nurseries
	Number of doctor's offices
	Number of chemical locations
	Number of restaurants

Application Case

After applying the PCAMIX method to the selected database a reduction of 56.25% in the number of variables is obtained since 7 components account for 80% of the data variance.

This information can be seen in the following table:

Table 3: PCAMIX for Gironde database

	Standard deviation	Proportion of Variance	Cumulative Proportion
Comp 1	2,6692	0,4453	0,4453
Comp 2	1,2203	0,0931	0,5384
Comp 3	1,1749	0,0863	0,6247
Comp 4	1,0521	0,0692	0,6939
Comp 5	0,9351	0,0546	0,7485
Comp 6	0,8056	0,0405	0,7890
Comp 7	0,7279	0,0331	0,8221
Comp 8	0,7189	0,0323	0,8544
Comp 9	0,6771	0,0287	0,8831
Comp 10	0,6477	0,0262	0,9093
Comp 11	0,6204	0,0241	0,9334
Comp 12	0,5750	0,0207	0,9541
Comp 13	0,5248	0,0172	0,9723
Comp 14	0,4747	0,0141	0,9854
Comp 15	0,3744	0,0088	0,9942
Comp 16	0,3081	0,0058	1

Application Case

In the procedure of analyzing "Gironde" database the indicator of life quality is the first component chosen and it explains the 44.53% of data variance [Aguilar, 2004]. Therefore, the indicator gets established as follows:

$$
\begin{aligned}
Z_1 = 0,278\,Y_1 &+ 0,262\,Y_2 + 0,298\,Y_3 + 0,325\,Y_4 + 0,301\,Y_5 \\
&+ 0,336\,Y_6 + 0,156\,Y_7 + 0,193\,Y_8 + 0,340\,Y_9 \\
&+ 0,350\,Y_{10} + 0,309\,Y_{11} + 0,112\,Y_{12} + 0,198\,Y_{14}
\end{aligned}
$$

Application Case

Table 4: Variables that explain the indicator

VARIABLES	
Y_1	Percentage of households
Y_2	Percentage of social housing
Y_3	Number of butcheries
Y_4	Number of bakeries
Y_5	Number of post offices
Y_6	Number of dental offices
Y_7	Number of supermarkets
Y_8	Number of nurseries
Y_9	Number of doctor's offices
Y_{10}	Number of chemical locations
Y_{11}	Number of restaurants
Y_{12}	Percentage of managers
Y_{14}	Percentage of buildings

Based upon this indicator, a ranking of the 10 best and worst cities of Gironde is presented and for this, the scores obtained by means of Principal Components Method are unified in values ranging among 0 and 100, as follows:

$$Indicator = \frac{Z_i - min(Z_i)}{max(Z_i) - min(Z_i)} * 100$$

The resulting rank of cities is:

Table 5: Ranking of 10 best cities of Gironde

	Best cities of Gironde	**Score**
1	Bordeaux	100
2	Bouscat	98,4095
3	Talence	95,8205
4	Begles	92,9496
5	Sainte-Foy-La-Grande	92,0792
6	Arcachon	90,6155
7	Eysines	90,3977
8	Cenon	90,1268
9	Merignac	89,7749
10	Pessac	89,7638

Table 6: Ranking of 10 worst cities of Gironde

	Worst cities of Gironde	Score
531	Fosses-Et-Baleyssac	0,5042
532	Lartigue	0,4705
533	Saint-Exupery	0,3367
534	Saint-Hilaire-De-La-Noaille	0,2719
535	Roquebrune	0,2599
536	Lucmau	0,2540
537	Cauvignac	0,2305
538	Giscos	0,2262
539	Labescau	0,1128
540	Saint-Martin-Du-Puy	0

Application Case

Table 7: Comparison between Bordeaux and Saint-Martin-Du-Puy

VARIABLES		Bordeaux	Saint-Martin-Du-Puy
Y_1	Percentage of households	0,0027	0,0007
Y_2	Percentage of social housing	0,0025	0,0007
Y_3	Number of butcheries	0,0033	0,0009
Y_4	Number of bakeries	0,0028	0,0012
Y_5	Number of post offices	0,0020	0,0011
Y_6	Number of dental offices	0,0034	0,0010
Y_7	Number of supermarkets	0,0011	0,0005
Y_8	Number of nurseries	0,0044	0,0002
Y_9	Number of doctor's offices	0,0029	0,0012
Y_{10}	Number of chemical locations	0,0035	0,0012
Y_{11}	Number of restaurants	0,0027	0,0012
Y_{12}	Percentage of managers	13,9700	0
Y_{14}	Percentage of buildings	21,4418	0,5131

References

Aguilar, M. A. S. (2004).
"Construcción e Interpretación de indicadores estadísticos.
OTA, 1 edition.

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

Kiers, H. (1991).
Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables.
Psychometrika, 56(2):197–212.

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

THANKS FOR YOUR ATTENTION