Principal Component Analysis for Mixed Quantitative and Qualitative Data

Susana Agudelo-Jaramillo
sagudel9@eafit.edu.co
Manuela Ochoa-Muñoz
mochoam@eafit.edu.co
Francisco Iván Zuluaga-Díaz
fzuluag2@eafit.edu.co

Research Practise
Mathematical Engineering
Research Group in Mathematical Modelling
Department of Mathematical Sciences
EAFIT University
Medellin

***Abstract*—The purpose of this paper is to present the research results on the PCAMIX method showing how useful it is in todays real world. In most current databases we commonly find mixed variables, that is, quantitative and qualitative variables. PCAMIX method can deal with these mixed databases and make possible to obtain significant statistical elements over a population under study. Besides that, we construct a useful indicator for result analysis and deeper study of certain selected population characteristics. An application case used for the understanding of the method shows its evident effectiveness and an indicator is obtained giving important information about the quality of life of relevant places.**

***Index Terms*—Indicator, Mixed data, Principal Components, Quantification.**

I. Introduction

Currently most of the databases information is of mixed nature, which means a mixture of qualitative and quantitative variables like statistical variables that are a representation of the characteristics associated to a surveyed population to perform diverse analysis.

When we talk about quantitative variables, we refer to mathematical variables measured in numerical quantities such as age, weight, area, volume, etc. These variables can be either of discrete nature where intermediate values are not allowed in an established scale, or of continuous nature, which are those variables that allow any value within a specific range; and when we refer to qualitative variables, we discuss the variables that express qualities or characteristics of a sample population and among them ordinals and nominal variables can be distinguished, ordinals because they can take different sorted values according to an established scale values and nominal variables because they can not be subjected to a sorting criteria.

Methods for quantifying qualitative data and especially methods for the analysis of mixed data are relatively recent, because most data analysis methods have focused over the years in the treatment of only pure quantitative data or pure qualitative data. Within these categories are two important methods, the Principal Component Analysis (PCA) for quantitative variables and Correspondence Analysis (CA) for qualitative variables.

To deal with mixed data several methods have been proposed by different authors, those are the cases of PCAMIX proposed by de Leeuw & van Rijckevorsel (1980), and later on during the years 1988 and 1989 of an alternative for the PCAMIX method called INDOOR proposed by Kiers (1988).

The appearance of these methods for mixed databases has meant great and specific support to decision-makers by providing important information and indicators over population problems requiring solutions.

In this article we can find a description of the methods used throughout the research and an application case with results obtained by applying the PCAMIX method to a mixed selected database. Also an important indicator is obtained giving important population related information.

II. Methodology

A. Principal Component Analysis (PCA)

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

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

The goal is to maximize the variance, so a_{ij} coefficients are increased. Besides that, to keep transformation orthogonality it is required that vector $\mathbf{a}'_j$ module be 1, that is,

$$\mathbf{a}_j^{'}\mathbf{a}_j = \sum_{k=1}^{p} a_{kj}^2 = 1$$

The first component is calculated by choosing $\mathbf{a}'_1$ in such way that y_1 gets the greatest variance subjected to the constraint that $\mathbf{a}'_1\mathbf{a}_1 = 1$. The second principal component is calculated by means of an $\mathbf{a}'_2$ that makes y_2 uncorrelated to y_1. The same procedure is applied to select y_3 through y_p.

B. Principal Component Analysis for Mixed data (PCAMIX)

This method considers the mixture of qualitative and quantitative variables. For qualitative variables it is necessary to calculate the quantification matrix associated (this procedure is explained in more detail in the next subsection). Then this matrix is concatenated with the matrix of quantitative variables to perform the Principal Component Analysis, (Kiers (1991)). It is important to highlight that this process is done with the purpose of reducing the number of variables that describe the problem, therefore the components that explain 80% of the variance of the data is selected.

C. Quantification Matrices

The idea of using quantification matrices is to define correlation coefficients between the variables.

Among the most commonly quantification matrices used, due to its effectiveness and characteristics in the results, we can find:

1) Quantification Matrix $G_jG'_j$: The G matrix is an indicator matrix with a binary structure, where 1 denotes that the individual meets certain characteristic and 0 tells us that the object does not meet certain feature of any category.

After obtaining this indicator matrix the product of this matrix is by itself but transposed, getting a square matrix, also known as Burt matrix which also give information about the frequency and the relationship between individuals and variables, through contingency tables.

In the diagonal blocks containing diagonal matrices appear marginal frequencies of each of the variables analyzed. Outside the diagonal cross frequency tables appear, corresponding to all combinations 2 to 2 of the variables analyzed.

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

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

where $S_{ii'j}$ 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)).

2) Quantification Matrix $G_j(G'_jG_j)^{-1}G'_j$: In this case the G matrix is like the one in the previous case but here it is multiplied by itself transposed G'_jG_j giving the Burt matrix inverted.
Here the value of the diagonal elements of the matrix G'_jG_j tell us when i and i' objects belong to the same category, that is, when individuals are similar, because when making the product of the matrices $Gj(G'_jG_j)^{-1}Gj$ we obtain a matrix where we find the values of the diagonal of G'_jG_j for individuals having similarity and zero for those not having.

It can be seen again as a measure of similarity, because their values are higher when objects are in the same category. The similarity between objects that are in the same category now depend on the number of objects in this category. The higher the frequency of a category, the greater the probability of two objects being in the same categories.

The elements of the quantification matrix $G_j(G'_jG_j)^{-1}G'_j$ 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 $(G'_jG_j)^{-1}$ (Kiers (1989)).

3) Quantification Matrix $JG_j(G'_jG_j)^{-1}G'_jJ$: Saporta proposed J matrix with the objective of centering the observations.

The J matrix is the subtraction between the identity matrix of dimension n and a vector product between a column of ones and a row of ones, divided by the sample size.

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

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

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

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

These similarities differ from the previous matrix in which these are reduced by n^{-1}, which leads to negative similarities between objects belonging to different categories and slightly reduced, but always positive between objects that fall into the same category.

Next, the procedure of PCAMIX method explained in detail below:

$$\begin{aligned} &\text{Max} \quad X'AX \\ &\text{s.t} \quad X'X = 1 \end{aligned}$$

where X is a nx1 vector and A is a nxn matrix.

$$\begin{aligned} X'AX &= X'K\Lambda K'X \\ &= Y'\Lambda Y \\ &= \begin{bmatrix} y_1 & y_2 & ... & y_n \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \lambda_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \\ &= \sum_{i=1}^{n} \lambda_i y_i^2 \\ &\leq \sum_{i=1}^{n} \lambda_1 y_i^2 \quad \lambda_i \leq \lambda_1 \forall i \\ &= \lambda_i \frac{\sum_{i=1}^{n} y_i^2}{1} \end{aligned}$$

where λ_1 is an upper bound for $g(x) = X'AX$. This upper bound can be reached by taking $X = K_1$, the first column of K, and this results.

$$g(K_1) = K_1'AK_1 = K_1'\lambda_1 K_1 = \lambda_1 K_1'K_1 = \lambda_1$$

due to,

$$[A - \lambda_1 I_n]K_1 = 0$$

and it found the maximum of the quadratic form $X'AX$ subject to the restriction $X'X = 1$.

The maximum is the largest eigenvalue of A, and is reached when X is the eigenvector associated.

Homogeneity analysis also known as Multiple Correspondence Analysis, is a generalization of PCA for qualitative variables.

A qualitative variable can, in this context, be conveniently represented by a matrix G_i, of order nxk_i, where k_j is the number of categories of the variable j. Each of the n individuals has a score of 1 for the category to which he/she belongs, and a score of zero in any other case.

$$G_i = \begin{matrix} 1 \\ 2 \\ \vdots \\ n \end{matrix} \begin{bmatrix} 1 & 0 & 0 & ... & 0 \\ 0 & 1 & 0 & ... & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & ... & 1 \end{bmatrix}$$

A matrix of this kind is called indicator matrix. In the case of m qualitative variables, m indicator matrices will be involved, they can be collected in a supermatrix.

$$G = \begin{bmatrix} G_1 & G_2 & ... & G_m \end{bmatrix}$$

of order nxk, where $k = k_1 + k_2 + ... + k_m$, the total number of categories of the m variables.

Homogeneity analysis is a technique in which weights are obtained and collected in the vectors $Y_1, Y_2, ..., Y_m$ of order (k_jx1) to quantify the categories of variables. As a result the quantified variables are constructed of the form $G_1Y_1, G_2Y_2, ..., G_mY_m$. G_jY_j is of order (nx1), (Gifi (1990)).

The weights are chosen to make quantified variables as homogeneous as possible.

This means that the quantified variables deviate a least as possible of a certain vector X, in the sense of minimum squares specifically, the weights and X are chosen to minimize, (ten Berge (1993)).

$$l(Y_1, ..., Y_m, X) = \sum_{j=1}^{m} (G_jY_j - X)'(G_jY_j - X)$$

$$\begin{aligned} &\text{s.t} \quad X'X = n \\ &\qquad \sum_{i=1}^{n} X_i = 0 \end{aligned}$$

Restrictions have been introduced to avoid getting trivial solutions.

$X = 0$ and $Y_j = 0$ or $X = j$ and $Y_j = j_{kj}$, respectively, $j = 1, ..., m$

$$G_jJ_{kj} = J$$

A way to find the minimum of $l(\cdot)$ subject to the restrictions is as follows.
Independently of X the associated $Y_j = (j = 1, ..., m)$ must satisfy:

$$\begin{aligned} Y_j &= (G_j'G_j)^{-1}G_j'X \\ &= D_j^{-1}G_j'X \end{aligned}$$

where D_j is defined as the diagonal matrix of order k_jxk_j containing the diagonal frequencies of different categories of $j - th$ variable. Using the above expression the problem can be simplified or minimize.

$$\begin{aligned}\tilde{l}(x) =& \sum_{j=1}^{m}(G_jD_j^{-1}G_jX - X)'(G_jD_j^{-1}G_j'X - X)\\ =& \sum_{j=i}^{m}(X'G_jD_j^{-1}G_j' - X')(G_jD_j^{-1}G_jX - X)\\ =& X'\left[\sum_{j=1}^{m}G_jD_j^{-1}G_j'G_jD_j^{-1}G_j'\right]X - X'\left[\sum_{j=1}^{m}G_jD_j^{-1}G_j'\right]X\\ & - X'\left[\sum_{j=1}^{m}G_jD_j^{-1}G_j'\right]X + \sum_{i=1}^{m}X'X\\ =& X'\left[\sum_{j=1}^{m}G_jD_j^{-1}G_j'\right]X - 2X'\left[\sum_{j=1}^{m}G_jD_j^{-1}G_j'\right]X + nm\\ =& nm - X'\left[\sum_{j=1}^{m}G_jD_j^{-1}G_j'\right]X\end{aligned}$$

The remaining problem is to maximize the quadratic form:

$$g(x) = X'\left[\sum_{j=1}^{m}G_jD_j^{-1}G_j'\right]X$$
$$\text{s.t} \quad X'X = n \quad \text{and} \quad 1'X = 0$$

The restriction $j'X = 0$ is equivalent to the restriction $JX = X$ where:

$$\begin{aligned}J &= (I_n - \frac{11'}{n})\\ JX - X &= 0\\ [J - I_n]X &= 0\\ [I_n - \frac{11'}{n} - I_n]X &= 0\\ \frac{11'}{n}X &= 0\\ 11'X &= 0\\ 1'X &= 0\end{aligned}$$

whereupon we have

$$\begin{aligned}g(x) &= g(JX)\\ &= X'J\left[\sum_{j=1}^{m}G_jD_j^{-1}G_j'\right]JX\\ &= X'\left[\sum_{j=1}^{m}JG_jD_j^{-1}G_j'J\right]X\\ &= X'WX\\ &= n\left(\frac{x}{\sqrt{n}}\right)W\left(\frac{X}{\sqrt{n}}\right)\\ &\leq n\lambda_1(W)\end{aligned}$$

where $\lambda_1(W)$ is the largest eigenvalue of W.

The upper bound is reached when X is chosen as the first eigenvector of W, scaling a sum of squares n. This eigenvector also satisfies both restrictions and clearly it has been found the minimum of $l(Y_1, ..., Y_m, X)$.

This procedure was implemented in R programming language for a specific aplication case.

III. Application case

For the application case an R database taken from PCAmixdata library and named "Gironde" is used. This database consists of 4 data sets characterizing the living conditions of 540 cities in Gironde – France. The aim of using this database is to construct an indicator that provides additional and complementary information about the life quality in cities.

The data set related to employment, housing and services come from the 2009 Census conducted by INSEE (Institut National de la Statistique et des Etudes Economiques) and the data set related to natural environment comes from IGN (Institut National de l'Information Geographique et forestiere).

This application case works with 16 variables from which 11 are qualitative and 5 quantitative, as shown in Table I and Table II:

- **Qualitative Variables:**

Table I – *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

- **Quantitative Variables:**

Table II – *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

After applying the PCAMIX method to the selected database a reduction of 56.25% in the number of variables is obtained since seven components account for 80% of the data variance. This information can be seen in Table III:

Table III – *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

An indicator is a numeric data as the result of a process that scientifically quantifies a characteristic of a sample. It gives information about the status of a particular situation or any particular characteristic at any given time and space, (Aguilar (2004)).
An indicator is generally a statistical data that synthesizes information of certain variables or parameters that affect the situation analyzed. Indicators can be qualitative and quantitative. In this case, as all variables carry quantitative terms the resulting indicator will carry them too.

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. Therefore, the indicator gets established as follows:

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

where Y_1 is the percentage of households, Y_2 is the percentage of social housing, Y_3 is the number of butcheries, Y_4 is the number of bakeries, Y_5 is the number of post offices, Y_6 is the number of dental offices, Y_7 is the number of supermarkets, Y_8 is the number of nurseries, Y_9 is the number of doctor's offices, Y_{10} is the number of chemical locations, Y_{11} is the number of restaurants, Y_{12} is the percentage of managers and Y_{14} is the percentage of buildings.

This way it is quite evident that the higher the value in each of the abovementioned variables, the higher the city life indicator.
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:

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

and the resulting rank of cities is shown in Table IV and Table V:

Table IV – *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 V – *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

IV. Concluding Remarks

It is found that the quantification matrices are quite useful to work with mixed data bases, since the qualitative variables do not contain numerical information needed to implement methods that require qualitative variables; in addition, the utility of the matrices to measure similarity and dissimilarity between individuals respect to a variable is verified.

The process has the purpose of reducing the number of variables that describe the problem selecting the components that explain 80% of the variance of the data. Besides that, weights are chosen to make quantified variables as homogeneous as possible, and restrictions are introduced to avoid trivial solutions.

The PCAMIX method was very useful when it was applied in a real life case, since it was found that it is possible to extract relevant information from mixed variables and it is not necessary to study the pure quantitative or qualitative variables separately when the data base is mixed.

V. Future work

Develop indicators that give important information from mixed data using PCAMIX method and implement this method on a database related to EAFIT University.

References

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

de Leeuw, J. & van Rijckevorsel, J. (1980), 'HOMALS and PRINCALS, some generalizations of principal components analysis', *Data Analysis and Informatics II* pp. 231–242.

Gifi, A. (1990), *Nonlinear Multivariate Analysis*, Wily and Sons.

Kiers, H. (1988), 'Principal components analysis on a mixture of quantitative and qualitative data based on generalized correlation coefficients', *In M. G. H. Jansen and W. H. van Schuur (Eds.), The many faces of multivariate analysis* **1**, 67–81.

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.

ten Berge, J. M. (1993), *Least Squares Optimization in Multivariate Analysis*, DSWO Press.