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 today's 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:

\[
y_j = a_{j1}x_1 + a_{j2}x_2 + ... + a_{jp}x_p = a_j^\prime x
\]

where \( a_j^\prime = (a_{1j}, a_{2j}, ..., a_{pj}) \) is a vector of constants, and
The goal is to maximize the variance, so \( a_{ij} \) coefficients are increased. Besides that, to keep transformation orthogonality it is required that vector \( a_j \) module be 1, that is,

\[
a_j^T a_j = \sum_{k=1}^{p} a_{kj}^2 = 1
\]

The first component is calculated by choosing \( a_1 \) in such way that \( y_1 \) gets the greatest variance subjected to the constraint that \( a_1^T a_1 = 1 \). The second principal component is calculated by means of an \( a_2 \) 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** \( G_j G_j^T \): 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 \( G_j G_j^T \) are given by:

\[
S_{ij} = \begin{cases} 
1 & \text{if object } i \text{ and object } i' \text{ belong to the same category} \\
0 & \text{if object } i \text{ and object } i' \text{ belong to different category}
\end{cases}
\]

where \( S_{ij} \) 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** \( G_j (G_j^T G_j)^{-1} G_j^T \): In this case the \( G \) matrix is like the one in the previous case but here it is multiplied by itself transposed \( G_j^T G_j \) giving the Burt matrix inverted.

Here the value of the diagonal elements of the matrix \( G_j^T G_j \) 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 \( G_j (G_j^T G_j)^{-1} G_j \) we obtain a matrix where we find the values of the diagonal of \( G_j^T G_j \) 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 \( G_j (G_j^T G_j)^{-1} G_j^T \) are given by:

\[
S_{ij} = \begin{cases} 
f_{g}^{-1} & \text{if object } i \text{ and object } i' \text{ belong to the same category} \\
0 & \text{if object } i \text{ and object } i' \text{ belong to different category}
\end{cases}
\]

where \( f_{g}^{-1} \) is the \( g-th \) diagonal element of \( (G_j^T G_j)^{-1} \) (Kiers (1989)).

3) **Quantification Matrix** \( J G_j (G_j^T G_j)^{-1} G_j^T J \): 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 \( \chi^2 \) measure. Where \( \chi^2 = 0 \) if variables are statistically independent.

The elements of the quantification matrix \( J G_j (G_j^T G_j)^{-1} G_j^T J \) are given by:

\[
S_{ij} = \begin{cases} 
f_{g}^{-1} - n^{-1} & \text{if object } i \text{ and object } i' \text{ belong to the same category} \\
-n^{-1} & \text{if object } i \text{ and object } i' \text{ belong to different category}
\end{cases}
\]
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:

$$\text{Max } X'AX$$
$$\text{s.t } X'X = 1$$

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

$$X'AX = X'K\Lambda K'X$$
$$= Y'\Lambda Y$$

$$= \begin{bmatrix} y_1 & y_2 & \cdots & y_m \end{bmatrix}
\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\
0 & \lambda_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_n \end{bmatrix}
\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

$$= \sum_{i=1}^{n} \lambda_i y_i^2$$

$$\leq \sum_{i=1}^{n} \lambda_i y_i^2, \lambda_i \leq \lambda_1 \forall i$$

$$= \lambda_1 \sum_{i=1}^{n} y_i^2$$

where $\lambda_1$ 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.

$$g(K_1) = K_1'AK_1 = K_1'1K_1 = \lambda_1 K_1'K_1 = \lambda_1$$

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 $G_j$, of order $nk_j$, 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.

$$G_j = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\
2 & 1 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
n & 0 & 0 & \cdots & 1 \end{bmatrix}$$

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.

$$G = [G_1 \ G_2 \ \cdots \ G_m]$$

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

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

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

$$\text{s.t } X'X = n$$

$$\sum_{i=1}^{n} X_i = 0$$

Restrictions have been introduced to avoid getting trivial solutions.

$$X = 0 \text{ and } Y_j = 0 \text{ or } X = j \text{ and } Y_j = jk_j, \text{ respectively, } j = 1, \ldots, m$$

$$G_jJ_{k_j} = J$$

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

$$Y_j = (G_j'G_j)^{-1}G_j'X$$

$$= D_j^{-1}G_j'X$$

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.
\[ \tilde{l}(x) = \sum_{j=1}^{m} (G_j D_j^{-1} G_j' X - X') (G_j D_j^{-1} G_j' X - X) \]
\[ = \sum_{j=1}^{m} (X' G_j D_j^{-1} G_j' - X') (G_j D_j^{-1} G_j' X - X) \]
\[ = X' \left[ \sum_{j=1}^{m} G_j D_j^{-1} G_j' G_j' \right] X - X' \left[ \sum_{j=1}^{m} G_j D_j^{-1} G_j' \right] X \]
\[ = X' \left[ \sum_{j=1}^{m} G_j D_j^{-1} G_j' \right] X + \sum_{i=1}^{m} X' X \]
\[ = X' \left[ \sum_{j=1}^{m} G_j D_j^{-1} G_j' \right] X - 2X' \left[ \sum_{j=1}^{m} G_j D_j^{-1} G_j' \right] X + nm \]
\[ = nm - X' \left[ \sum_{j=1}^{m} G_j D_j^{-1} G_j' \right] X \]

The remaining problem is to maximize the quadratic form:

\[ g(x) = X' \left[ \sum_{j=1}^{m} G_j D_j^{-1} G_j' \right] X \]

\[ \text{s.t. } X' X = n \text{ and } 1' X = 0 \]

The restriction \( j' X = 0 \) is equivalent to the restriction \( JX = X \) where:

\[ J = (I_n - \frac{11'}{n}) \]
\[ JX - X = 0 \]
\[ [J - I_n] X = 0 \]
\[ [I_n - \frac{11'}{n} - I_n] X = 0 \]
\[ \frac{11'}{n} X = 0 \]
\[ \frac{11'}{n} X = 0 \]
\[ 1' X = 0 \]

whereupon we have

\[ g(x) = g(JX) \]
\[ = X' J \left[ \sum_{j=1}^{m} G_j D_j^{-1} G_j' \right] JX \]
\[ = X' \left[ \sum_{j=1}^{m} JG_j D_j^{-1} G_j' \right] X \]
\[ = X' WX \]
\[ = n \left( \frac{x}{\sqrt{n}} \right) W \left( \frac{X}{\sqrt{n}} \right) \]
\[ \leq n\lambda_1(W) \]

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 application case.

### III. 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. 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>
<thead>
<tr>
<th>DATA SET</th>
<th>VARIABLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Housing</td>
<td>Percentage of households</td>
</tr>
<tr>
<td></td>
<td>Percentage of social housing</td>
</tr>
<tr>
<td>Services</td>
<td>Number of butcheries</td>
</tr>
<tr>
<td></td>
<td>Number of bakeries</td>
</tr>
<tr>
<td></td>
<td>Number of post offices</td>
</tr>
<tr>
<td></td>
<td>Number of dental offices</td>
</tr>
<tr>
<td></td>
<td>Number of supermarkets</td>
</tr>
<tr>
<td></td>
<td>Number of nurseries</td>
</tr>
<tr>
<td></td>
<td>Number of doctor’s offices</td>
</tr>
<tr>
<td></td>
<td>Number of chemical locations</td>
</tr>
<tr>
<td></td>
<td>Number of restaurants</td>
</tr>
</tbody>
</table>

- **Quantitative Variables:**

<table>
<thead>
<tr>
<th>DATA SET</th>
<th>VARIABLES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employment</td>
<td>Percentage of managers</td>
</tr>
<tr>
<td></td>
<td>Average income</td>
</tr>
<tr>
<td>Natural environment</td>
<td>Percentage of buildings</td>
</tr>
<tr>
<td></td>
<td>Percentage of water</td>
</tr>
<tr>
<td></td>
<td>Percentage of vegetation</td>
</tr>
</tbody>
</table>

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:
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:

\[
Z_1 = 0.278Y_1 + 0.262Y_2 + 0.298Y_3 + 0.325Y_4 + 0.301Y_5 + \\
+ 0.336Y_6 + 0.156Y_7 + 0.193Y_8 + 0.340Y_9 + \\
+ 0.350Y_{10} + 0.309Y_{11} + 0.112Y_{12} + 0.198Y_{14}
\]

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:

\[
Indicator = \frac{Z_i - \text{min}(Z_i)}{\text{max}(Z_i) - \text{min}(Z_i)} \times 100
\]

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

### Table III – PCAMIX for Gironde database

<table>
<thead>
<tr>
<th>Comp</th>
<th>Standard deviation</th>
<th>Proportion of Variance</th>
<th>Cumulative Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.5692</td>
<td>0.4433</td>
<td>0.4453</td>
</tr>
<tr>
<td>2</td>
<td>1.2203</td>
<td>0.0931</td>
<td>0.5384</td>
</tr>
<tr>
<td>3</td>
<td>1.1749</td>
<td>0.0863</td>
<td>0.6247</td>
</tr>
<tr>
<td>4</td>
<td>1.0521</td>
<td>0.0692</td>
<td>0.6939</td>
</tr>
<tr>
<td>5</td>
<td>0.9351</td>
<td>0.0546</td>
<td>0.7485</td>
</tr>
<tr>
<td>6</td>
<td>0.8056</td>
<td>0.0405</td>
<td>0.7890</td>
</tr>
<tr>
<td>7</td>
<td>0.7279</td>
<td>0.0331</td>
<td>0.8221</td>
</tr>
<tr>
<td>8</td>
<td>0.7189</td>
<td>0.0323</td>
<td>0.8544</td>
</tr>
<tr>
<td>9</td>
<td>0.6771</td>
<td>0.0287</td>
<td>0.8831</td>
</tr>
<tr>
<td>10</td>
<td>0.6477</td>
<td>0.0262</td>
<td>0.9093</td>
</tr>
<tr>
<td>11</td>
<td>0.6204</td>
<td>0.0231</td>
<td>0.9334</td>
</tr>
<tr>
<td>12</td>
<td>0.5730</td>
<td>0.0207</td>
<td>0.9541</td>
</tr>
<tr>
<td>13</td>
<td>0.5248</td>
<td>0.0172</td>
<td>0.9723</td>
</tr>
<tr>
<td>14</td>
<td>0.4747</td>
<td>0.0141</td>
<td>0.9854</td>
</tr>
<tr>
<td>15</td>
<td>0.3744</td>
<td>0.0088</td>
<td>0.9942</td>
</tr>
<tr>
<td>16</td>
<td>0.3081</td>
<td>0.0058</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

### Table IV – Ranking of 10 best cities of Gironde

<table>
<thead>
<tr>
<th>Best cities of Gironde</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Bordeaux</td>
<td>100</td>
</tr>
<tr>
<td>2 Bousca</td>
<td>98.4095</td>
</tr>
<tr>
<td>3 Talence</td>
<td>95.8025</td>
</tr>
<tr>
<td>4 Begles</td>
<td>92.9596</td>
</tr>
<tr>
<td>5 Sainte-Foy-La-Grande</td>
<td>92.0792</td>
</tr>
<tr>
<td>6 Arcachon</td>
<td>90.6155</td>
</tr>
<tr>
<td>7 Eysines</td>
<td>90.3977</td>
</tr>
<tr>
<td>8 Cenon</td>
<td>90.1268</td>
</tr>
<tr>
<td>9 Merignac</td>
<td>89.7499</td>
</tr>
<tr>
<td>10 Pessac</td>
<td>89.7638</td>
</tr>
</tbody>
</table>

### Table V – Ranking of 10 worst cities of Gironde

<table>
<thead>
<tr>
<th>Worst cities of Gironde</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>531 Fosses-Ét-Baleyssac</td>
<td>0.5042</td>
</tr>
<tr>
<td>532 Laritg</td>
<td>0.4705</td>
</tr>
<tr>
<td>533 Saint-Exupery</td>
<td>0.3367</td>
</tr>
<tr>
<td>534 Saint-Hilaire-De-La-Noaille</td>
<td>0.2719</td>
</tr>
<tr>
<td>535 Roquebrune</td>
<td>0.2599</td>
</tr>
<tr>
<td>536 Lucma</td>
<td>0.2340</td>
</tr>
<tr>
<td>537 Canvignac</td>
<td>0.2305</td>
</tr>
<tr>
<td>538 Giscos</td>
<td>0.2262</td>
</tr>
<tr>
<td>539 Labesca</td>
<td>0.1128</td>
</tr>
<tr>
<td>540 Saint-Martin-Du-Pay</td>
<td>0</td>
</tr>
</tbody>
</table>

### IV. Concluding Remarks

It is found that the quantification matrices were 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


