# Principal Component Analysis for Mixed Quantitative and Qualitative Data

Proposal Report Research Practise

Susana Agudelo-Jaramillo[1]
Manuela Ochoa-Muñoz[2]

Tutor: Francisco Iván Zuluaga-Díaz[3]

Research Group in Mathematical Modelling
Department of Mathematical Sciences
EAFIT University
Medellín

February 12th 2016

[1] Email: sagudel9@eafit.edu.co
[2] Email: mochoam2@eafit.edu.co
[3] Email: fzuluag2@eafit.edu.co

# 1 Problem Formulation

The statistic variables of a sample or population surveyed represent the various features associated to their elements that are important to be analyzed and studied. By the nature of the data the statistic variables can be classified as:

- **Quantitative Variables:** They are mathematical variables measured in terms of numerical quantities such as age, weight, area, volume, etc. These variables can be whether of discrete nature where intermediate values are not allowed on an established scale, or of continuous nature, which are those variables that admit any value within a specific range.
- **Qualitative Variables:** These variables express the qualities or characteristics of a sample or 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.

A great amount of procedures developed for multivariate data analysis are supported only for quantitative variables, and one of the most used method is [1]:

- **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.
  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 a 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 $y_p$.

There exist also techniques to deal with qualitative data like the pure method used to perform the analysis [2]:

Simple Correspondence Analysis is often used in the representation of data that can be presented as contingency tables of two nominal or ordinal variables.

- **Correspondence Analysis:**
  This is a descriptive or exploratory technique with the main objective of summarizing a large amount of data into a small number of dimensions with the least possible loss of information.

  The analysis of simple correspondence is often used for showing data that can be represented as contingency tables of two nominal variables or of two ordinals.

  A contingency table based on two qualitative nominal or ordinal variables where categories of one variable appear in rows and other variable categories are represented in columns, the Correspondence Analysis consists of summarizing the rows and columns information in such a way that it can be projected on a reduced subspace of row points and column points and finally conclusions of two variable relationships can be drawn.

  The extension of Simple Correspondence Analysis to the case of several nominal variables (multidimensional contingency tables) is named as Multiple Correspondence Analysis and uses the same general principles of the described technique. In general, this application is oriented to cases where a variable represents items or individuals and the rest are qualitative or ordinal variables representing qualities.

What happens in practice is that a great part of existent information is of a mixed nature, meaning a mixture of quantitative and qualitative variables. Therefore, it is required to go deep in Principal Component Analysis for Mixed Data (**PCAMIX**) to construct an appropriated indicator.

## 2 Goals

*Main Goal*
Deepen and understand the fundamental characteristics of the **PCAMIX** procedure, to achieve a better analysis of mixed data problems.

*Specific Goals*

- Characterize different types of matrices used to quantify qualitative data.
- Validate statistically the results obtained under **PCAMIX** procedure.
- Report an application case where it is possible to study the usefulness of the implemented procedure.

## 3 Preceding Work

Although quantification of qualitative data and especially mixed data analysis problems are of relatively recent appearance, researchers are currently developing methods for analysis of mixed variables, including the method **PCAMIX** that will be implemented in this work. **PCAMIX** method was proposed by De Leeuw and Van Rijckevorsel (1980) [3], and includes Principal Component Analysis and Multiple Correspondence Analysis.

Over the past 40 years several authors have independently proposed methods for the analysis of mixed variables, such as De Leeuw, 1973; De Leeuw and Van Rijckevorsel, 1980 [3]; Escofier, 1979 [4] and Nishisato, 1980 [5]. These methods differ in the way in which the quantitative variables are transformed but for qualitative variables the same approach is used.

The most recent proposal for an alternative of **PCAMIX** was proposed by Kiers (1988, 1989) [6] denoted **INDOOR** or also known as **INDOMIX**. Kiers suggested this method as an alternative for the analysis of certain variables set through optimization and a different presentation of the coordinates of the objects.

## 4 Justification

Today it is quite common to have mixed data, that is, a mixture of data of quantitative and qualitative nature. Most developed methods focus on analysis of pure quantitative or qualitative data but there exist just few methods dealing with analysis of mixed data and besides that, some of them have been recently formulated. So that is why it becomes very important to conduct a research to get a better understanding, deepening and characterization of **PCAMIX** method and then to apply this technique in the analysis of mixed data. As a further analysis result a suitable indicator should be built up in order to evaluate, estimate or demonstrate the relationships and/or the importance of variables in the problem studied. According to a systematic literature review no such indicator has been developed so far.

## 5 Scope

In this research project will be developed a methodology for building an indicator using **PCAMIX** procedure.

## 6 Proposed Methodology

Through this research a review of technical literature will be conducted to identify different types of quantifying matrices and thus determining which of them would be the most suitable technique for the analysis of main components of mixed data.

Then, better understanding, deepening and characterization of **PCAMIX** procedure are required to analyze properly the information from quantitative and qualitative variables. Besides that, it is required statistically validate the results obtained under this procedure and build an indicator that shows the relationship between the variables.

Finally, an application case will be developed through a suitable programming language such as MATLAB or R, which will be selected according to the requirements of the method, in order to demonstrate developed procedure usefulness.

As a special note, there will be several tutoring meetings during the academic semester to evaluate project contents, progress, teamwork, results and advances.

## 7 Activity Schedule

The following tables sketch activities, semester weeks and dates that must be fulfilled in order to accomplish this project.

Table 1: Activity schedule during research.

| **Activity** | **Semester/Week** | **Dates** |
|---|---|---|
| Literature review. | 1 - 4 | January 25 - February 21 |
| Understanding, deepening and characterization of **PCAMIX** procedure. | 4 - 7 | February 15 - March 13 |
| Construction of the indicator using **PCAMIX**. | 8 - 10 | March 14 - April 10 |
| Application case using a suitable programming language. | 11 - 14 | April 11 - May 8 |
| Project report. | 15 - 16 | May 9 - May 22 |
| Project presentation. | 17 - 19 | May 23 - June 7 |

Table 2: Key dates during Research Practise.

| **Activity** | **Week** | **Dates** |
|---|---|---|
| Proposal report. | 3 | February 12 |
| Proposal presentation. | 4 | February 19 |
| Oral progress report. | 10 | April 8 |
| Project report. | 16 | May 20 |
| Project presentation. | 19 | June 7 |

Data for Table 2 was taken from the following web page related to this research course: http://www1.eafit.edu.co/asr/courses/research-practises-me/2016-1/index.html

## 8 Budget

This research does not require financing.

## 9 Intellectual Property

Susana Agudelo Jaramillo, Manuela Ochoa Muñoz and Francisco Iván Zuluaga Díaz share intellectual property in this research equally.

# References

[1] A. C. Rencher. *Methods of Multivariate Analysis.* Wiley Series in Probability and Statistics, 1934.

[2] S. de la Fuente Fernández. Análisis correspondencias simples y múltiples. *Universidad Autónoma de Madrid*, pages 1–9, 2011.

[3] J. de Leeuw and J. van Rijckevorsel. HOMALS and PRINCALS, some generalizations of principal components analysis. *Data analysis and informatics II*, pages 231–242, 1980.

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

[5] S. Nishisato and W. Sheu. Piecewise method of reciprocal averages for dual scaling of multiple-choice data. *Psychometrika*, 45:467–478, 1980.

[6] H. Kiers. 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, 1988.