

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
March 4th, 2016



Statistic Variables

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.

Quantitative Variables

They are mathematical variables measured in terms of numerical quantities.

Qualitative Variables

These variables express the qualities or characteristics of a sample or population.

Principal Component Analysis

It considers a set of variables x_1, x_2, \dots, x_p upon a group of objects or individuals and based on them a new set of variables y_1, y_2, \dots, y_p is calculated, but these new variables are uncorrelated with each other and their variances should decrease gradually [1].

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

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

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

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

Correspondence Analysis

- The analysis of simple correspondence is often used for showing data that can be represented as contingency tables [2].
- 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.
- 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.

Mixed 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.

```
graph TD; Q[Quantitative] --> PCAMIX[PCAMIX]; Qual[Qualitative] --> PCAMIX;
```

PCAMIX

Proposed methods for analysis of mixed variables

- De Leeuw, 1973 [3].
- Escofier, 1979 [3].
- De Leeuw and Van Rijckevorsel, 1980 [4].
- Nishisato, 1980 [5].

Proposals for PCAMIX method

- **PCAMIX** method was proposed by De Leeuw and Van Rijckevorsel (1980) [4].
- The most recent proposal for an alternative of **PCAMIX** was proposed by Kiers (1988, 1989) [6] denoted **INDOOR** or also known as **INDOMIX**.

Importance of this Research

- Today it is quite common to have mixed data.
- Most developed methods focus on analysis of pure quantitative or qualitative data.
- Is very important to conduct a research to get a better understanding, deepening and characterization of **PCAMIX** method and 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.
- So far it has not been developed an indicator.

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.

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



Proposed Methodology

- Review of technical literature to identify different types of quantifying matrices.
- Better understanding, deepening and characterization of **PCAMIX** procedure to analyze properly the information from quantitative and qualitative variables.
- Build an indicator that shows the relationship between the variables.
- Developed an application through a suitable programming language.

References I

-  A. C. Rencher, *Methods of Multivariate Analysis*, Wiley Series in Probability and Statistics, 1934.
-  S. de la Fuente Fernández, *Análisis correspondencias simples y múltiples*, Universidad Autónoma de Madrid (2011) 1–9.
-  H. Kiers, Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables, *Psychometrika* 56 (2) (1991) 197–212.
-  J. de Leeuw, J. van Rijckevorsel, HOMALS and PRINCALS, some generalizations of principal components analysis, *Data analysis and informatics II* (1980) 231–242.

References II

-  S. Nishisato, W. Sheu, Piecewise method of reciprocal averages for dual scaling of multiple-choice data, *Psychometrika* 45 (1980) 467–478.
-  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* (1988) 67–81.

THANKS FOR YOUR
ATTENTION