Trabalho apresentado no XXXVIII CNMAC, Campinas - SP, 2018.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Principal Components Analysis in Mixed Epidemiological Data

Adrián Martínez Amarilla [1]
Facultad de Ciencias Exactas y Tecnológicas, UNC, Concepción, Paraguay
Francisco Medina Roa[2]
Facultad de Ciencias Exactas y Tecnológicas, UNC, Concepción, Paraguay
Jorge Daniel Mello Román[3]
Facultad de Ciencias Exactas y Tecnológicas, UNC, Concepción, Paraguay

The principal components analysis is a dimension reduction technique from multivariate data analysis, whose objective is to synthesize information from a dataset, reducing the number of variables to a lower number of components or factors that explain most of the variance found in the data. These components or factors constitute a linear combination of the original variables and are linearly independent [1].

The technique has a double purpose: to optimally represent in a space of small dimension, observations of a larger p-dimensional general space; and to identify possible "latent" variables that generate the variability of the data. The components are independent and therefore facilitate the interpretation of the data [2].

The principal component analysis can be applied to a mixed data set, that is, that contain both quantitative and qualitative variables. Let X_1 be a matrix of numerical data of $n \times p_1$, X_2 a matrix of categorical data of $n \times p_2$, and m the total number of categories. The algorithm is summarized in the following three steps [3].

1. Preprocess: Consists of building a matrix of numerical data $Z = (Z_1|Z_2)$ of dimension $n \times (p_1 + m)$ where Z_1 is the standardized version of the matrix X_1 and Z_2 the centered indicator matrix of the levels of X_2 .

   Next, build the diagonal matrix $\mathbf{N}$ of the weights of the rows, where n is weighted by $\frac{1}{n}$. Likewise, build the diagonal matrix $\mathbf{M}$ of the weights of the columns, where the p_1 first columns weighted by 1 and the last columns weighted by $\frac{n}{n_s}$, where n_s is the number of observations with level s. The total variance is:

$$p_1 + m - p_2 \tag{1}$$

[1]adrianmartinezamarilla@gmail.com
[2]franciscomedinaroa@gmail.com
[3]jorgedanielmello@hotmail.com

2. Decomposition of the Generalized Singular Value from the Z matrix is performed through the decomposition:

$$Z = UDV^t \tag{2}$$

where:

- $D = diag(p_1; \cdots ; p_r)$ is the $r \times r$ diagonal matrix of the singular values of ZMZ^tN and Z^tNZM, and r denotes the rank of Z;
- U is the $n \times r$ matrix of the first eigenvectors of ZMZ^tN such that $U^tNU = \mathbb{I}_r$.
- V is the $p \times r$ matrix of the first r eigenvectors of Z^tNZM such that $V^tMV = \mathbb{I}_r$.

3. Score processing: the set of factor scores for the rows, or scores of the main components, is calculated through formula $F = UD$, while the set of factor scores for columns is calculated by means of the expression:

$$A = MVD \tag{3}$$

This technique was applied to an epidemiological database corresponding to 78 suspected cases of Dengue Fever registered at the health centers of the Department of Concepción, Paraguay, in 2016. The database is made up of 28 qualitative variables and 9 quantitative variables that contain information about patients, including symptoms, medical evaluations and data about their recent events, such as mobility and the presence of similar cases in their environment.

The data was processed with the help of the R software and the PCAmixdata package [3]. Preliminary results show an adequate reduction in the dimension of the data set to a total of 14 components that explain 78.7% of the total variation of the data.

References

[1] O. R. Rojas. Análisis en componentes principales. *San José: Universidad de Costa Rica*, 2009.

[2] D. Peña. *Análisis de datos multivariantes.* McGraw-Hill España, 2013.

[3] M. Chavent, et al. Multivariate analysis of mixed data: The PCAmixdata R package. *arXiv preprint arXiv:1411.4911*, 2014.

[4] D. Beaton, C. R. C. Fatt, and A. Hervé. An ExPosition of multivariate analysis with the singular value decomposition in R. *Computational Statistics & Data Analysis*, 2014, vol. 72, p. 176-189.