missMDA: a package to handle missing values in and with multivariate exploratory data analysis methods

Julie Josse[1], **Francois Husson**[1]

1. Applied Mathematics Department, Agrocampus, Rennes, France
*Contact author: julie.josse@agrocampus-ouest.fr

Keywords: Missing values, Principal component methods, Multiple imputation, Confidence ellipses

In this presentation, we describe the **missMDA** package which is dedicated to handle missing values in exploratory data analysis methods such as principal component analysis (PCA) and multiple correspondence analysis. This package provides the classical outputs (scores, loadings, graphical representations, etc.) of principal component methods despite the missing values. It also gives confidence areas around the position of the points (individuals and variables) representing the uncertainty due to missing values. The package can also be used to perform single or multiple imputation for continuous and categorical variables in a general framework. In this presentation, we describe the underlying method through PCA.

A common approach to handle missing values in PCA consists in minimizing the loss function (the reconstruction error) over all nonmissing elements. This can be achieved by the iterative PCA algorithm (also named expectation maximization PCA, EM-PCA) described in Kiers (1997). It consists in setting the missing elements at initial values, performing the analysis (the PCA) on the completed data set, filling-in the missing values with the reconstruction formula (the PCA model, Caussinus (1986)) and iterate these two steps until convergence. The parameters (axes and components) and the missing values are then simultaneously estimated. Consequently, this algorithm can be seen as a single imputation method. To avoid overfitting problems, regularized version of the EM-PCA algorithm have been proposed (Josse and Husson, 2011; Ilin and Raiko, 2010).

After the point estimate, it is natural to focus on the variability of the parameters. However, the variance of the axes and components estimated from the completed data set (obtained with the EM-PCA algorithm) is underestimated. Indeed, the imputed values are considered as observed values and consequently the uncertainty of the prediction is not taken into account in the subsequent analysis. It is possible to resort to multiple imputation (Rubin, 1987) to avoid this problem. Multiple imputation consists first in generating different plausible values for each missing values. Then it consists in performing the statistical analysis on each imputed data set and combining the results. Josse and Husson (2011) have proposed a new method to generate multiple imputed data sets from the PCA model. They have also proposed two ways to visualize the influence of the different predictions of the missing values onto the PCA results. It leads to confidence areas around the position of the individuals and of the variables on the PCA maps.

References

Caussinus, H. (1986). Models and uses of principal component analysis. In J. de Leeuw, W. Heiser, J. Meulman, and F. Critchley (Eds.), *Multidimensional Data Analysis*, pp. 149–178. DSWO Press.

Ilin, A. and T. Raiko (2010). Practical approaches to principal component analysis in the presence of missing values. *Journal of Machine Learning Research 11*, 1957–2000.

Josse, J. and F. Husson (2011). Multiple imputation in principal component analysis. *Advances in Data Analysis and Classification.*

Kiers, H. A. L. (1997). Weighted least squares fitting using ordinary least squares algorithms. *Psychometrica 62*, 251–266.

Rubin, D. B. (1987). *Multiple imputation for non-response in survey.* Wiley.