Engineer offer: development of an R package to handle missing values

1 Introduction

missMDA is an R package, created by Julie Josse and François Husson which was designed to perform principal components (PC) methods, such as PCA, correspondence analysis, with missing values, *i.e* estimating loadings and scores from an incomplete dataset. These functionalities are available in the package, but as it turns out, missMDA has been most used as a package to do single and multiple imputation for quantitative and categorical data.

Indeed, the main algorithms to perform PC methods despite missing values also directly output an imputed data set which can be used for further analyses. The quality of imputation is usually very high, which can be explained by the fact that imputation takes into account similarities between observations as well as relationships between variables, using a rather small number of parameters due to the dimensionality reduction property of the methods.

Following our works in research, we incrementally added new functionalities in the package such as multiple imputation for categorical data (Audigier, Husson, and Josse, 2015). In comparison to state of the art methods and implementations of single and multiple imputation (Little and Rubin, 2002, van Buuren, 2012), missMDA fills a gap because it handles larger datasets, with many categorical variables and rare categories, as well as small sample sizes in comparison to the number of variables. In addition, it is equipped with all the visualization tools of principal components methods (such as biplots) and outputs graphical displays, in order to better assess the quality of imputation and provide new diagnostic plots for the users.

2 Profil

We are looking for an excellent and highly motivated engineer able to implement methodologies to handle missing values. The engineer will be hired for one year with the aim to further develop the missMDA packages. The candidate should hold a degree in Statistics/ CS/ data-sciences and have excellent coding capabilities in R and in C++. Knowledge of the missing values literature or of matrix completion problems could be valuable. Interested graduates (undergraduates) should apply as early as possible since the position will be filled when a suitable candidate is found. The candidate will also have excellent opportunities to interact with researchers in public health

3 Laboratory

The position takse place in the applied mathematics department of Ecole Polytechnique CMAP (http://www.cmap.polytechnique.fr/spip.php?rubrique141). The department is a dynamic environment of international renown with many students, PhD students and researchers. The student will be integrated into the statistical team and the data-sciences initiative. https://portail.polytechnique.edu/datascience/fr

4 Missions

The engineer will have to further develop missMDA as follows.

- Recode some parts of the package to increase the speed of execution: sparse coding, option for fast SVD, function to combine the results of multiple imputation from different machines

- Work on visualization tools and users interface (Shiny application)

- Add new functions in the R package to implement the current research works on (distributed) imputation for multi-level (hierarchical) data. This work is part of a project with collaborators in public health. Ask Julie Josse for more details on the project on distributed imputation.

- Work on the documentation/webpage of the package.

5 Contact

Julie Josse whose research focuses on handling missing values. She organized the first MissData conference, gives many lectures/tutorials on the topic and is preparing a Statistical Science special issue to have a snapshot of the state of the art on the topic. **julie.josse@polytechnique.edu**