Tutorial: Handling missing values with a special focus on the use of principal components methods

Julie Josse, Applied mathematics department, Agrocampus Ouest, Rennes France, josse@agrocampus-ouest.fr

François Husson, Applied mathematics department , Agrocampus Ouest, Rennes France, husson@agrocampus-ouest.fr

Overview

Any statistician or user of statistics has been confronted with the problem of missing values for a variety of reasons: measuring devices that fail, data that has been accidentally destroyed or lost, individuals who have failed to respond to certain questions in a questionnaire, dead animals, damaged plants, etc.

Missing values are problematic since most statistical methods cannot be applied directly to an incomplete dataset.

One of the most popular approaches to deal with missing values consists in using “single imputation” methods. This is done by filling in the missing values with plausible values which leads to a completed dataset, that can be analyzed by any statistical method. There is a huge literature on imputation methods for continuous data or categorical data which is well summarized in Schaefer (1997) and Little & Rubin (2002) and limited proposals for mixed data. Recently, methods based on principal component models such as principal component analysis (PCA) have shown good performances to complete data (Mazumder, R., Hastie, T. & Tibshirani, R., 2010; Josse, J. & Husson, F., 2013).

However, single imputation is limited because it does not take into account the uncertainty associated with the prediction of missing values based on observed values. Thus, if a statistical method is applied on a completed data table, the variability of the estimators is underestimated.

To avoid this problem, an Expectation-Maximisation (EM) algorithm (Dempster et al., 1977) combined for instance with a Supplemented Expectation-Maximisation algorithm (Meng and Rubin,1991) could be used to get the maximum likelihood estimate as well as their variance from an incomplete data. Another solution is to perform multiple imputation (Rubin, 1987; Little and Rubin, 1987, 2002).

This tutorial presents a small overview of the literature on the missing data topic. We first present the classical methods of single and multiple imputation as well as the main R packages dedicated to this issue. Then, we focus on recent methods based on principal component methods.

Goal

Handling missing values for continuous, categorical and mixed data

As an aside perform principal component methods, PCA and multiple correspondence analysis (MCA) with missing values.

Outline

1) Introduction to the missing values issues (aims - imputation or estimation of the parameters? - EM algorithm)

2) Presentation of single imputation methods for continuous, categorical and mixed data. Comparison of a method based on PCA versus one based on random forests (Stekhoven, D.J. & Buhlmann, P., 2011)

3) Presentation of multiple imputation methods for continuous, categorical and mixed data. Comparison of methods based on joint or conditional modelling (Multivariate Imputation by Chained Equations - MICE) assuming a Gaussian distribution for the variables versus methods based on PCA.

4) Examples of analyzing and visualization of continuous incomplete data with PCA and with MCA for categorical data such as survey data. Visualization of the variability due to the missing values.

The different methods will be illustrated with numerous examples from different fields such as genomics (human tumor data), sensometrics (wine data) and survey (questionnaire data) and we will use the R packages Amelia, MICE and missMDA.

Intended audience

Users of statistics who would like to analyze data with missing values.

Background

Basic knowledge in PCA

Related link

More information will be available (notes, scripts, and data sets) at our website http://factominer.free.fr.

Some videos are available on http://www.youtube.com/user/HussonFrancois

References

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., Rubin, D. B., 2006. Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation, 76, 1049–1064.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* B, 39, 1-38.

Honaker J, King G, Blackwell M (2011). Amelia II: A Program for Missing Data. *Journal of Statistical Software*, 45(7), 1-47.

Josse J, Husson F (2012). Handling missing values in exploratory multivariate data analysis methods. *Journal de la Société Française de Statistique*, 153 (2), 1-21.

Little RJA, Rubin DB (1987, 2002). Statistical Analysis with Missing Data. Wiley series in probability and statistics, New-York.

Mazumder R, Hastie T, Tibshirani R (2009) Spectral regularization algorithms for learning large incomplete matrices. *Journal machine learning research*, 11, 2287- 2322.

Schafer J (1997). Analysis of Incomplete Multivariate Data. Chapman & Hall/CRC.

Stekhoven D, B uhlmann P (2011) Missforest - nonparametric missing value imputation for mixed-type data. *Bioinformatics*, 28, 113-118.