Postdoc/Internship offers: handling missing values for heterogeneous data - developing models for traumatized patients

1 Summary

Fellowships (Postdocs, Master 2 internships) are focusing on missing data. Interested graduates (undergraduates) should apply as early as possible since the positions will be filled when suitable candidates are found. The Centre for Applied Mathematics (CMAP) is looking for excellent and highly motivated individuals able to develop methodologies to handle missing values (such as a general multiple imputation method for multivariate continuous and categorical variables) and their implementation in the free R software. The successful candidates will be part of research group in the statistical team on missing values. The candidates will also have excellent opportunities to collaborate with researcher in public health with partners on the analysis of a large register from the Paris Hospital (APHP) to model the decisions and events when severe trauma patients are handled by emergency doctors. Candidates should contact Julie Josse at polytechnique.edu

2 Laboratory

The postdoc will take 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

3 Project: multiple imputation for heterogeneous data

Key words: missing values, SVD, multiple imputation, latent variable models, matrix completion, heterogeneous data.

Multiple imputation [1] is a reference method for making inference in the presence of missing data. It operates in three steps. First, M plausible values are generated for each missing values leading to M imputed (completed) tables. Then, the quantity of interest θ and its variance are estimated on each table and the estimates are aggregated to obtain a point estimate and an estimator of the variance that incorporates the additional variability due to the missing data and thus coverage of confidence intervals at the nominal level.

This approach is very popular because once the data have been completed, it is possible to apply any statistical methods. However, multiple imputation methods still have many shortcoming that are active research fields. In particular, there are very few satisfactory solutions [2] to impute data with mixed variables (continuous, categorical, ordinal, count) in large dimensions.

The objective of this fellowship is to develop a multiple imputation method based on generalized PCA [3]. This work is motivated by the very good empirical results of imputation methods for categorial variables [4] based on decompositions in weighted singular values and recent developments of latent variable models for categorical variables [5]. The idea is to extend the latter model to the mixed variables in order to propose an imputation. To define a "proper" multiple imputation procedure in the sense of Rubin, ie, which ensures good properties for the resulting inference, we must reflect the uncertainty of the imputation model parameters from an imputation to the other. To do this, it is common to use a bootstrap approach [6] or a Bayesian approach. We shall endeavor to study finely the properties of these two alternatives and to discuss the interest of the "proper" multiple imputation, echoing the controversies between Nielson and Rubin [7, 8]. Finally, attention will be paid to the influence of Missing at Random and the simultaneous presence of the MAR and missing non-random (MNAR) types [9, 10] on the validity of the inference.

4 Collaboration

The candidates will be part of a collaboration with APHP (Public Assistance - Hospitals of Paris) on Analysis and modeling of the management of severely traumatized patients. When a patient arrives at the emergency room, there is a succession of important decisions which are taken, in particular, on the gravity of his condition and which lead to pay more or less immediate attention. Of course, diagnostic errors at these stages can be dramatic. Based on a very large, extremely incomplete database (consisting of the registers of several centers in Ile de France), which details the patient's path and their characteristics, the ultimate objective is to develop a model decision in order to

guide and support the emergency workers who must establish the actions to be carried out for the patient in a very short time. The methodological advances made to handle missing data are essential to be able to respond in an operational way to the problems raised by the Traumabase database. We will focus in particular on the development of logistic regression decision rules with missing data. Many interactions between theory and applications are envisaged.
Collaborators on this topic include: avec Jean-Pierre Nadal, research director CNRS and head of the analysis center of social mathematics (CAMS) at EHESS, Lab of Physic Statistics (CNRS-ENS-UPMC-Univ. Paris Diderot) and the group Traumabase with Pr Catherine Paugam-Burtz and Dr Tobias Gauss Anaesthetics, Emergency Medical Care, Traumatology - Hôpital Beaujon, APHP Hôpitaux Universitaires Paris Nord Val de Seine).

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**

Références

[1] Little, R.J.A & Rubin, D.B. (2002). *Statistical Analysis with Missing Data.* John Wiley & Sons.
[2] Murray, J. & Reiter, J. (2016). Multiple imputation of categorical and continuous via bayesian mixture models. *Journal of American Statistical Association.*
[3] Allen, G.I., Grosenick, G. & Taylor, J. (2014). A Generalized Least-Square Matrix Decomposition. *Journal of the American Statistical Association.*
[4] Audigier, V., Husson, F. & Josse, J. (2015). MIMCA: Multiple imputation for categorical variables with multiple correspondence analysis. *Statistics and Computing.*
[5] Fithian, W. and Josse, J. (2017) Multiple Correspondence Analysis & the Multilogit Bilinear Model. *Journal of Multivariate Analysis.* [6] Efron, B. (1994) Missing Data, Imputation, and the Bootstrap. *Journal of the American Statistical Association.*
with multiple correspondence analysis. *Statistics and Computing.*
[7] Nielsen, S.F. (2003). Proper and Improper Multiple Imputation. *International Statistical Review.*
[8] Rubin, D. B.. (2003). Discussion on Multiple Imputation. *International Statistical Review.*
[9] Seaman, S. Galati, J., Jackson, D. & Carlin, J (2013). What Is Meant by "Missing at Random"? *Statistical Sciences.*
[10] Franks, A.M., Airoldi, E. M. & Rubin, D.B. (2016). Non-standard conditionally specified models for non-ignorable missing data.