Low rank estimation with non-ignorable missing data

Claire Boyer[1], Julie Josse[2], and Aude Sportisse[1,2]

[1]Laboratoire de Probabilités Statistique et Modélisation, Sorbonne Université, France
[2]Centre de Mathématiques Appliquées, Ecole Polytechnique, France

The problem of missing data is ubiquitous in the practice of data analysis. A classic solution for dealing with missing values is to impute them to get a completed data. The matrix completion methods based on low-rank approximation of the data matrix caught the attention due to their ability to handle large matrices with large amount of missing entries. However, the theoretical guarantees of these completion methods ensuring the correct prediction of missing values are only valid under the restrictive assumptions of completely random missing data (MCAR) or random missing data (MAR) where the missing data are totally independent of the value of variables or only depend on the value of observable variables. In our talk, we will focus on the least restrictive assumption of not random missing data (MNAR), when the unavailability of the data depends of the values of other variables and its value itself. There is little literature on how to deal with missing MNAR data just focusing on cases with one variable missing which is not our case.

We suggest two approaches to take into account MNAR data in the low rank model to impute data and recover the low rank structure. First we maximize a penalized likelihood using an Expectation Maximization (EM) algorithm where the missing-data mechanism is modeled with a logistic regression model. More specifically, we use a Monte Carlo approximation by simulating the samples with the Sampling Importance Resampling algorithm. The second approach we suggest is to concatenate the data matrix and the missing data indicator matrix and use classical methods, such as the softimpute methods ([2]) using iterative thresholded SVD, without modeling the missing-data mechanism. Finally, we compare the approaches to the case where we ignore the MNAR mechanism and apply classical methods. Part of our work will be positioned within the framework developed by Pearl ([3]) on the graphical models.

This work is motivated by a public health application with the APHP Traumabase Group (Assistance Publique - Hôpitaux de Paris) on the management of polytraumatised patients who have suffered a major trauma (injuries that endanger the life or functional integrity of a person). The decision support models that we wish to establish will here focus on the prediction of hemorrhagic shock and head injury, which are the main causes of mortality in major trauma. The MNAR data are extremely frequent in this database where, for example, the patient's blood pressure cannot be measured because his or her health condition is such that the measurement cannot be made. Different imputation techniques will be compared on these real data.

In this talk, we will highlight that modeling the mechanism gives better imputation results. However, in this specific low rank structure, the gain of using a specific model for the mechanism is not enough significant in comparison to its computation burden. The second approach by concatenating the data matrix with the missing data indicator matrix can be empirically shown to be a good compromise: it makes few hypothesis, performs well and is computationally efficient.

References

[1] Roderick J.A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., 2002.

[2] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug):2287–2322, 2010.

[3] Karthika Mohan, Felix Thoemmes, and Judea Pearl. Estimation with incomplete data: The linear case. In *IJCAI*, pages 5082–5088, 2018.