Multiway high-dimensional lasso-penalized analysis with imputation of missing data applied to postgenomic data in an Ebola vaccine trial

Hadrien Lorenzo* , Rodolphe Thiébaut[1], and Jérôme Saracco[2]

[1]Institut de Santé Publique d'Epidémiologie et de Developpement (ISPED) – Université de Bordeaux – France

[2]CQFD (INRIA Bordeaux - Sud-Ouest) – CNRS : UMR5251, INRIA – France

Abstract

Several sets of variables can be analyzed simultaneously by canonical correlation in a multi-way analysis. These sets of variables are often high-dimensional and repeated over time. For instance, full-transcriptome measured by RNA-Seq used to be performed in longitudinal studies as well as other measures such as peptides or cells. Hence, canonical correlation analysis has been extended with regularized approaches to deal with several high dimensional data. However, some measurements can be missing for technical reasons and therefore introduce undesired structures due to the huge dimension of the datasets.
Our objective is to find an efficient method allowing to impute the missing values taking into account the three-way structure, participant-transcriptome-time, and also the missing path structure.
We proposed an EM-like covariance-maximization lasso-penalized high-dimensional completion matrix algorithm to reach that goal.
We compared our approach on simulated data-sets with the mean imputation per gene per time step, the missMDA-imputeMFA algorithm which takes structure into account and the softImpute solution initially designed to solve the Netflix competition a high-dimensional problem. We used two criterions: the L2-error between estimated and simulated values and the L2-error between estimated and simulated covariance matrices. The numerical results exhibited the superiority of the proposed method in most of the scenarii. We also illustrated our approach on a real data-set from a phase I Ebola vaccine trial measuring RNA-Seq data after vaccination (richtien, cell report 2017) in 20 participants at 4 different times on whole-blood samples, representing 74 sequenced-samples, among which 24 samples were missing because of technological issues.

*Speaker