Handling Missing Rows in Multi-Omics Data Integration: Multiple Imputation in Multiple Factor Analysis Framework

V. Voillet[a], **P. Besse**[b], **L. Liaubet**[a] **M. San Cristobal**[a] and **I. González**[c]

[a]Génétique, Physiologie et Systèmes d'Élevage
INRA, Castanet-Tolosan
valentin.voillet@toulouse.inra.fr
laurence.liaubet@toulouse.inra.fr
magali.san-cristobal@toulouse.inra.fr

[b]Institut de Mathématiques
Université de Toulouse III, Toulouse
philippe.besse@math.univ-toulouse.fr

[c]Mathématiques et Informatiques Appliquées
INRA, Castanet-Tolosan
ignacio.gonzalez@toulouse.inra.fr

Mots clefs : Multiple omics data integration, Multivariate factor analysis, Missing individuals, Multiple imputation.

In omics data integration studies, it is common, for a variety of reasons, that some individuals are not present in all data tables. Missing row values are challenging to deal with because most statistical methods cannot be directly applied to incomplete datasets. To overcome this issue, we propose a multiple imputation (MI) approach [1] in a multivariate framework. In this study, we focus on multiple factor analysis (MFA) [2] as a tool to compare and integrate multiple layers of information. MI involves filling the missing rows with plausible values, resulting in m completed datasets. MFA is then applied to each completed dataset leading to m different component configurations. Finally, the m configurations are combined to yield one consensus solution.

We assessed the performance of our method, named MI-MFA, on two real omics datasets. Incomplete artificial datasets were created from these data with different patterns of missingness. The MI-MFA results were compared to two other approaches, regularized iterative MFA (RI-MFA) [3] and mean variable imputation (MVI-MFA). For each component configuration resulting from these three strategies, we determined the suitability of the component solution against the true MFA configuration obtained from the original data. The overall results showed that MI-MFA outperformed the RI-MFA and MVI-MFA approaches in nearly all settings of missingness.

Two approaches, confidence ellipses and convex hulls, to visualize and assess the uncertainly due to missing values were also described. We showed how the areas of ellipses and convex hulls increased as variability was added to the data. These graphical representations provide scientists with considerable guidance in order to evaluate the reliability of the results.

Références

[1] Rubin, D.B. (2004). *Multiple Imputation for Non-Response in Surveys.* Wiley-Interscience, Hoboken, New Jersey, USA

[2] Escofier, B., Pagès, J. (1994). Multiple factor analysis (afmult package). *Computational*

Statistics & Data Analysis, **18**(1), 121-140
[3] Josse, J., Husson, F. (2012) Missing values in exploratory multivariate data analysis methods. *Journal de la SFdS*, **153**(2), 79-99