Multiway SIR for biological data integration

Valérie Sautron, Marie Chavent, Nathalie Viguerie, Nathalie N. Villa-Vialaneix

Multiway SIR for biological data integration

Valérie Sautron [1], Marie Chavent [2], Nathalie Viguerie[3], Nathalie Villa-Vialaneix [4]

[1] *INRA, GenPhySE (Génétique Physiologie et Systèmes d'Elevage), F-31326 Castanet-Tolosan cedex, valerie.sautron@toulouse.inra.fr*

[2] *IMB - Institut de Mathématiques de Bordeaux, marie.chavent@u-bordeaux.fr*

[3] *Inserm - UMR1048, Obesity Research Laboratory, Institut of Metabolic and Cardiovascular Diseases (I2MD), Toulouse, nathalie.viguerie@inserm.fr*

[4] *INRA UR0875 MIAT Mathématiques et Informatique Appliquées Toulouse, 24 chemin de Borde-Rouge - Auzeville CS 52627, 31326 Castanet-Tolosan cedex, nathalie.villa@toulouse.inra.fr*

1 Abstract

High-throughput sequencing techniques are increasingly used and allows to measure a large quantity of data at different levels of the living organism under study. As the cost of such techniques becomes more affordable, the complexity of the design of the experiments also increases and a biological process can now be observed dynamically. In this case, repeated measurements of one or several omics are acquired in order to monitor the evolution of a given biological process. When wanting to integrate such data to have a global overview of the time-course evolution, the biologist has to face several difficulties: i) the longitudinal nature of the data has to be taken into account, ii) data coming from different sources have to be related.

In the present paper, we address the issue of predicting a given numerical target variable from time-course multivariate data. The proposed method is an extension of the STATIS approach [L'Hermier des Plantes, 1976] to a framework which is similar to SIR (Sliced Inverse Regression [Li, 1991]).

Given Y a real variable of interest and $(X_t)_{t=1,\ldots,T}$ the collection of T tables of p variables measured on the same n individuals, the proposed method can then be described in two steps: first, we analyze the between table similarity structure and derive a set of optimal weights to construct a compromise matrix designed so as to capture the stable structure of the differences over time between the means of X_t conditional to Y. Second, we perform a generalized PCA on the compromise matrix which estimates an EDR space for longitudinal data.

We will illustrate our approach on data collected in the DIOGenes project (www.diogenes-eu.org). These data contained thirteen clinical variables which were collected on 135 obese female patients at 3 stages of a specific diet. The goal of this exploratory analysis is to investigate the relationship between these clinical data and weight evolution rate of the patients between the beginning and the end of the study.

References

[L'Hermier des Plantes, 1976] L'Hermier des Plantes, H. (1976). *Structuration des tableaux à trois indices de la statistique*. PhD thesis, Université de Montpellier. Thèse de troisième cycle.

[Li, 1991] Li, K. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–342.