Data Engineer Position
Multivariate Statistics applied to Metabolomics

Crossover methodological approaches for analyzing metabolomics data with time component

1. Context and objectives

Biochemical processes are intrinsically dynamic and taking the **temporal behaviour** of the system under study into account constitutes a pivotal element for its understanding in many situations in life sciences. As it provides a phenotypic description of biological phenomena, the characterisation of time-resolved metabolic states is expected to reveal important biochemical information. In this perspective, collection of metabolomics data over time is especially relevant in clinical research. Statistical methods able to grasp the **dynamic nature** of the studied phenomenon are therefore needed in this context. The ANR project GDM MILK **(ANR-22-CE17-0039)** aims to identify specific metabolic signatures in the milk of GDM mothers (and their modifications depending on GDM treatments). To do this, the project aims to select among constituents of breast milk measured in various lactation stage (colostrum, transition and mature milk), those with the capacity to interact with longitudinal biomarkers identified associated to function of pancreatic and intestinal endocrine cells and to metabolic trajectory of the offspring, taking into account sexual dimorphism.

Several batteries of data sets with time component were acquired as part of GDM project: *i.e.*, in the experimental rat model of GDM, maternal breast milk composition, offspring plasma metabolomic and lipidomic signatures, blood and tissues biochemical parameters and liver, pancreatic or intestinal gene expression; and in human ongoing breast milk-blood-feces biocollection, maternal breast milk composition, infant fecal microbiota. For the same individual, these data have only a small number of points in time (less than eight) and a large number of variables (metabolites) as well as until four treatment groups of observations. These two characteristics make the use of functional approaches [1] developed for the analysis of longitudinal data ineffective. Nowadays, there is very little methodological work dedicated to analyze metabolomic data with time components. Among the methods that have attracted interest in this context is ANOVA–Simultaneous Component Analysis (ASCA) and related methods [2].

ASCA and related methods consider the temporal dimension as a factor of variation among the others in the experimental design, offering an easy interpretation of the variations induced by the time factor. At least two other alternative approaches are possible in the context of analyzing metabolomic data with time component. The first considers the time dimension as forming the third way of a third order tensor. This approach allows the implementation of a tensor approach using the PARAFAC model for the analysis of metabolomic data with time component [3]. The second approach takes into account time dimension through a path (past→present→future). This makes it possible to consider a specific model of directed relationships between several data blocks, each one representing a time point. The implementation of a structural equation model with latent variables, typically PLSPM [4], will be considered for this second approach because it allows flexibility and richness in terms of interpretation.

The objective is twofold: (i) to investigate the relative merits of these three approaches for the analysis of GDM-MILK data that have a time component and (ii) to extend these approaches to integrate several data blocks, all or part of them having a time component.

2. Tasks and expected results

The work carried out will serve as a basis for the development of methodological and operational solutions, applicable to the GDM MILK data and disseminated to the community of statisticians and biologists. The activities take place in two steps as summarized in the following table.

Step 1 (9 months) : Evaluate the relevance of the three methodological approaches
Main tasks : • Summarize a list of selected publications and bibliographic research. • Implement PLSPM, ASCA and PARAFAC on GDM MILK data. Main expected results : • Executable script under R. • Write a first publication for submission in scientific journals.
step 2 (9 months) : GDM MILK Data fusion
Main tasks : • Develop an algorithm for GDM MILK data fusion. • Implement executable script under R. Expected results : • Executable script under R. • Write a second publication for submission in scientific journals.

3. Candidate profile

The ideal candidate is motivated by methodological problems and related applications in metabolomics. He (she) has first experience with one of the three methodological approaches mentioned in the description (see section 1) or has solid background at machine learning, applied statistics or chemometrics. The ideal candidate has good scientific writing and oral communication skills in English language. A proven experience and taste in computer programming and data analysis is required. Experience in one of the following areas is required: Applied statistics, chemometrics, Machine learning, or related fields and proven ability to solve research problems, with demonstrable research experience in at least one of them.

4. Funding and location

The contract will be in effect from May 1, 2025, to September 30, 2026, with a closing date for applications set for April 25, 2025. The net salary is around 1500 euros/month. The successful postdoctoral candidate will be affiliated to StatSC (Oniris VetAgroBio) at Nantes in France. He (she) will collaborate with members of this laboratory and two partners of the GDM MILK project. He (she) will be required to participate in consortium TIMeID "Interdisciplinary Work on Temporal Data Integration Methodologies in Biology".

5. Application

Application files must be sent to jean-michel.galharret@oniris-nantes.fr as soon as possible and must include:

- Cover letter.
- CV, including contact information for at least one referee.
- Research outcome (Ph.D's thesis or paper) written by the candidate.
- Letters of recommendation will be appreciated.

It should be noted that incomplete application will not be considered.