3-year PhD position in Statistics - Machine Learning for Biological processes in Lyon, France

Vivian Viallon (International Agency for Research on Cancer, Lyon)
Yohann De Castro (Ecole Centrale de Lyon, Institut Camille Jordan)

March 2024

Scientific and funding environment

Applications are invited for a 3-year PhD position in statistics and machine learning at the International Agency for Research on Cancer (IARC, World Health Organization, Lyon, France) and Ecole Centrale de Lyon (ECL, Lyon, France). The PhD candidate will join the MOBiL project (Multi-omics data integration to investigate biological mechanisms underlying the link between lifestyle behaviors and gastro-intestinal cancers), which is one of the 11 "Projets Structurants" funded within the Shape-Med Lyon initiative.

Under the supervision of Vivian Viallon (leader of the Biostatistics and Data Integration team at IARC) and Yohann De Castro (Professor at ICJ, junior member of Institut Universitaire de France), the PhD candidate will develop and study high-dimensional nonlinear latent variable models for the integration of multi-omics data. This methodological project is motivated by applications in cancer epidemiology: our models will eventually be applied in the European Prospective Investigation into Cancer and nutrition (EPIC) and UK Biobank studies to characterize the human metabolic responses to specific lifestyle exposures (e.g., habitual alcohol intake), while accounting for the inter-individual variability of that metabolic response. The epidemiological results will be key to enhance our understanding of the biological processes through which lifestyle might influence human health, in particular cancer development.

Proposal and team

Previous metabolic signatures of lifestyle exposures were constructed using simple statistical tools. This project aims at inferring the inter-individual variability across subgroups of the population. A key challenge is that the subgroups to which the different subjects belong are unknown. Using statistical jargon, the categorical variable G that represents the different subgroups is latent. From a computational perspective, it will have to be "predicted" from the data, before estimating model parameters.

In MOBiL, we will assume that G is determined by genome-wide genotyping data, blood proteomics levels and standard epidemiological data, such as age, sex, and auxiliary lifestyle exposures. Models developed in MOBiL will perform multi-omics data integration and dimension reduction through the latent categorical variable G. Previous studies of metabolic signatures implicitly worked under a much simpler setting ignoring the possible influence of the other variables. The main novelty of our approach lies in the consideration of variable G to model the modulation of lifestyle's metabolic impact by genetic determinants, blood protein levels and standard epidemiological variables. Another novelty of our approach is that it will enable the modeling of multivariate exposures, to study the impact on metabolism of several exposures or several mixtures of exposures at a time.

We will pay a particular attention on the trade-off between complexity (to ensure appropriate goodness-of-fit on real data) and tractability (to ensure our model can be fitted on real data). To reach this trade-off, which constitutes another key novelty of our approach, different versions of our model will be tested successively. When needed, we will enlarge our inference by considering non-linearities in the definition of our signatures.

On the other hand, the complexity of our model could be reduced by discarding irrelevant variables. For example, variables could focus on a handful of selected SNPs and/or polygenic risk scores, such as those found to be associated with metabolic response or all-cause mortality in the literature. The final MOBiL model will be selected based on regular exchanges within the MOBiL team, especially using IARC experts feedback on the scalability and goodness-of-fit of the successive models.

MOBiL will be led by a multi-disciplinary team, combining expertise in statistics and machine learning, but also in biochemistry, molecular epidemiology and cancer epidemiology. MOBiL will benefit from the expertise the IARC team acquired over the years through the coordination of research projects on the link between metabolic biomarkers with cancer risk, and on the identification of metabolic biomarkers and metabolic signature for specific lifestyle exposures.

Profile of the candidates

No background in Biology or Medicine is required, though experience in biochemistry, molecular epidemiology and cancer epidemiology can be appreciated. The ideal applicant hold a Master in Applied Mathematics, or is about to graduate (a 4-6 months Master 2 internship is possible before the Ph.D.). Background in High-Dimensional Statistics, or Variational Inference, or Bayesian computing, or Statistical Learning, or Bio-statistics is required.

Timeframe of the call

The proposal is open and the position can be taken within the next months. Applications for a starting date in 2025 will be also considered. Non-French speakers applicants are welcome, full English job is possible.
Phase 1 (open): Candidates are required to send a full CV, list of publications and motivation letter to viallonv@iarc.who.int and yohann.de-castro@ec-lyon.fr
Phase 2 (invited): Invited candidates will be interviewed online. Up to 2 reference letters can be sent to M. Viallon and M. De Castro before the interview.

Salary and support

Computer, access to large computing facilities, travel expanses and conference fees will be fully covered.

The salary is based on the french academic system of 2,100 euros gross salary per month (2,200 euros in 2025 and 2,300 euros in 2026, around 3000 euros total payroll cost to the employer) plus work/home transport reimbursements and participation in mutual insurance. Depending on your situation, the tax can be around 3%.

A teaching duty (64hours per year, possibily in English) in a French Grande École (Centrale de Lyon) is possible (around 45 euros gross per hour).