Improvement of the scalability of hypergraph inference based on infinite sparse independent component analysis in the context of high sample size and high dimensional context

Contact: Damien Drubay, PhD - Bureau de Biostatistique et d'Epidémiologie de Gustave Roussy / INSERM U1018 - Equipe Oncostat (damien.drubay@gustaveroussy.fr)
Place: Gustave Roussy, 114 Rue Edouard Vaillant 94800, Villejuif
Starting date: from the February 3rd to the April 1st, 2025
Duration: from 5 to 6 months
Application deadline: January 15th, 2025

Context of the project

Identifying individual mechanisms involved in complex diseases, such as cancer, is essential for the precision medicine. Their characterization is particularly challenging due to the unknown relationships of the high-dimensional omics data and their inter-patient heterogeneity.

Since the underlying number of biological mechanisms is unknown, but certainly large, we proposed to model them by deconvoluting the high-dimensional gene expression using infinite independent component analysis, where each component would represent a biological mechanism. The inter-patient heterogeneity is modeled allocating an individual combination of components to each patient imposing a strict sparsity to the infinite component weight matrix by its elementwise product with a random binary matrix drawn from a beta-bernoulli process [1]. We demonstrated that this infinite sparse independent component analysis was able to blindly identify some known molecular signatures in the breast cancer, some of them could impact the patient prognostic [2].

Given that the random binary matrix defining the hyperedges of an infinite hypergraph, this model could offer an interpretable structure for data generated by more recent technologies. For example, each hyperedge could represents cell communities from single-cell sequencing or cellular niches from spatial transcriptomics. However, the large size of datasets generated by these technologies (several thousands of cells/spots and genes per tumor sample) presents a significant challenge, primarily due to the large number of parameters required for non-parametric inference.

To address this, we aim to develop a more scalable algorithm for practical use. We propose two main improvements: first, the current algorithm's primary bottleneck lies in the discrete optimization of elements of the binary matrix. We introduced a continuous relaxation of these discrete variables, which should speed up gradient-based inference using differentiable functions and reduce the number of parameters. Second, we propose leveraging GPU computing for high-performance computing for the inference using modern probabilistic programming framework such as NumPyro, which support a wide range of MCMC sampling algorithm such as the Hamiltonian Monte Carlo algorithm, or approximative methods such as the (stochastic) variational inference approaches (ADVI, Stein,..).

Objectives

- Familiarize with the notions of non-parametric Bayesian approach and the underlying stochastic processes
- Implement the algorithm with GPU computing framework (ex: using Numpyro, a Python framework for probabilistic programming). The student will have access to the high-performance computing cluster of Gustave Roussy including several GPU modules and high memory capacity for the analysis of large datasets

- Perform numerical simulations in order to validate the new algorithm
- Apply the algorithm to real datasets
- This work will open the perspective of a PhD thesis for the development of several extensions for multiomics data

Expected profile

- Students in engineering schools, in the second year of a Master's program, or equivalent, in Statistics, Machine Learning, data-science, or a closely aligned discipline
- Strong mathematical background for computational science, and willingness to address the challenge of developing models of complex biological systems. Bayesian theory knowledge would be a plus
- Good programming skills, Python would be a plus
- In addition to your statistical/mathematical skills, you are known for your rigor, autonomy and interpersonal skills. You enjoy working in a multidisciplinary environment and are able to communicate complex results to a non-specialist audience.

How to apply?

As part of our proactive policy to promote the integration of people with disabilities, all applications received are considered on an equal basis.
To submit an application for this vacancy, please send your curriculum vitae and a cover letter by email to Damien Drubay (damien.drubay@gustaveroussy.fr), indicating the reference "InternCBio2025" in the object of the email.

References

[1] Sarah-Laure Rincourt, Stefan Michiels, and Damien Drubay. Complex Disease Individual Molecular Characterization Using Infinite Sparse Graphical Independent Component Analysis. *Cancer Informatics*, 21:11769351221105776, 2022.

[2] Sarah-Laure Rincourt, Stefan Michiels, and Damien Drubay. A non-parametric Bayesian joint model for latent individual molecular profiles and survival in oncology. *Journal of Bioinformatics and Computational Biology*, 20(5):2250022, October 2022.