PrAIrie Postdoctoral Position

Topic: High-dimensional inference in genomic data

Supervision: Chloé-Agathe Azencott, Mines ParisTech & Institut Curie (PSL Research Institute, Paris) http://cazencott.info

Funding: through PrAIrie https://prairie-institute.fr/

Location: CBIO Mines ParisTech & Institut Curie (Paris) https://cbio.ensmp.fr/

Duration: 12 months. Start date is flexible, starting from Fall 2021.

Qualifications: A PhD in statistics of machine learning, preferably with experience in post-selection inference and/or high-dimensional data. Prior experience with genomics data is not required, but interest for the application is necessary.

Scientific context: The goal of this project is to **propose new efficient procedures for high-dimensional inference, motivated by applications to high-dimensional genomic data.** More specifically, we are interested in identifying regions of the genome associated with a phenotype, through procedures that provide p-values.

In the simplest setup, regions of the genome can be identified by a single binary or ternary feature (a single nucleotide polymorphism, or SNP), and phenotypes are continuous or binary responses. The number of features is orders of magnitudes larger than that of samples (typically, 105-107 features, for 104-105 samples), which greatly hinders statistical power.

To alleviate this, we can use the following additional information:

- SNPs that are nearby on the genome are correlated, and it's possible to consider blocks of SNPs rather than individual SNPs and search for association at the block level (see for example [1, 2]).
- SNPs can be mapped to genes, based on position on the genomic sequence, regulatory information (the SNP is known to regulate the expression of genes) or 3-dimensional information (the SNPs and the genes are physically nearby). This allows us to combine SNP information with gene-level information. In particular:
    - we can also have access to levels of gene expression, which are continuous features, following a normal distribution or a Poisson distribution, depending on the acquisition technique, which leads us to take a multiomics view of the problem;

    - we can use biological networks in which nodes are genes and edges represent known relationships between genes, as we can make the assumption that relevant features tend to be connected on such a network (see for example [3]).
- We may have access to multiple related phenotypes, which allows us to use multitask approaches. In addition, we may have information about task relatedness, in the form of similarities between tasks; vectorial representation of tasks (see for example [4]); or networks in which tasks are nodes and edges model relationship between them.

While many methods already exist to perform feature selection accounting for such information, in particular through variants of the lasso, few techniques allow one to obtain p-values efficiently in this context. A possible starting point would be existing literature on post-selection inference, including our work on performing post-selection inference with kernels (kernelPSI) [1].

[1] L. Slim, C. Chatelain, C.-A. Azencott (2020). Nonlinear post-selection inference for genome-wide association studies, BioRxiv. https://doi.org/10.1101/2020.09.30.320515
[2] A. Nouira, C.-A. Azencott (2021). Multitask group Lasso for Genome-Wide Association Studies in admixed populations, MLCSB abstract
https://www.iscb.org/cms_addon/conferences/ismbeccb2021/tracks/mlcsb
[3] H. Climente-González et al. (2021). Boosting GWAS using biological networks: A study on susceptibility to familial breast cancer, PLOS Comp Bio.
https://doi.org/10.1371/journal.pcbi.1008819
[4] V. Bellón, V. Stoven, C.-A. Azencott (2016). Multitask feature selection with task descriptors, PSB. http://psb.stanford.edu/psb-online/proceedings/psb16/bellon.pdf

Applying: Please email chloe-agathe.azencott@mines-paristech.fr with your CV and a motivation letter (in either English or French).