Ecole Doctorale Carnot-Pasteur

Proposition de sujet de thèse

Intitulé français du sujet de thèse proposé : Régularisation de problèmes inverses pour l'estimation des composants immunitaires cellulaires

Intitulé en anglais : Inverse problems regularization for estimation of cellular immune components

Unité de recherche : Institut de Mathématiques de Bourgogne

Nom, prénom et courriel du directeur (et co-encadrant) de thèse :
Hervé Cardot (PR. UB, herve.cardot@u-bourgogne.fr)
Samuel Vaiter (CR CNRS, samuel.vaiter@u-bourgogne.fr)
Caroline Truntzer (IR CGFL, caroline.truntzer@cgfl.fr)

Domaine scientifique principal de la thèse :
Statistique, Optimisation

Domaine scientifique secondaire de la thèse :
Cancérologie, Santé

Sujet à considérer dans le cadre de l'appel de I-SITE Bourgogne Franche-Comté :Oui

Description du projet scientifique

This PhD proposal is at the interface between mathematics and biology. The PhD student will be supervised by Hervé Cardot (PR) and Samuel Vaiter (CR), as well as Caroline Truntzer (IR). Hervé Cardot and Samuel Vaiter are two mathematicians working at Institut de Mathématiques de Bourgogne, with complementary skills, H. Cardot being an expert in statistics, in particular for functional data analysis, and S. Vaiter known for his work in signal processing and variational method analysis Caroline Truntzer is recognised as an expert in biomedical statistics for her work at CGFL. At the end of the PhD, the student is expected to have both mastered a large spectrum of statistics and optimization methods, and to have a deep knowledge relative to the application in biostatistics.

Context

The most recent and probably most important change in the field of oncology is the development of immuno-oncology. However, the complexity of tumor microenvironment is still not fully addressed. Knowing the global immune composition of tumors is thus of major importance. Recently, the biology literature proposed to use deconvolution methods for transcriptomic data in order to estimate this cellular composition. The first one was presented by Abbas et al. in 2009 (Abbas *et al.* 2009). This method was based on linear regression with a positivity constraint on the coefficients to be estimated. Since then, several other methods, like MCP-counter (Becht *et al.* 2016) or Cibersort (Newman *et al.* 2015) have been suggested. One of the main drawbacks of Cibersort is that the estimated quantities of the cell types are relative to the number of cell subtypes considered, and not absolute quantifications.

Objectives

In this PhD proposal we will investigate new regularization methods of inverse problems that provide an absolute quantification of immune cell subpopulations. The mathematical aspect of this PhD proposal is two-fold. The first goal is to enhance the underlying linear model through a more refined construction of the expression matrix. The second goal is, given this linear model, to derive the best possible estimator. These two issues can be treated in a decoupled way, which is the standard for existing methods such as Cibersort, or as a coupled optimization problem (which is known as blind deconvolution in signal processing).

We advocate in this PhD proposal the use of variational methods which could include some hard constraints such as positivity, imposing that the regressor is a probability vector, or a regularity prior, including low-rank or parsimony of the expression matrix. The existing methods used to estimate cellular immune components are modifications of the ordinary least squares estimator which does not take into account the structure of the data from a mathematical point of view. Our objective is to propose an underlying model of the procedure, in order to be able to provide better guarantees.

We believe that a formulation which includes the group structure of the genomic expression, and the parsimony could help both estimation speed and precision. More precisely, following the seminal paper (Tibshirani, 1996) on sparse regression through a convex program (LASSO) which is connected to Support Vector Regression for a particular choice of kernel, we aim to deal with a large collection of cell populations. Note that every regularization methods lead to biased estimates, and a modern strategy of debiasing (Deledalle *et al.* 2017) will be used in practice to enhance the results (again in a one-step procedure). To help us achieve a better FDR rate, we will also consider modification (such as group-based) of the SLOPE estimator proposed in Bogdan *et al* (2015) which shares similar properties of LASSO but with better guarantees regarding false positives. All the mentioned regularization methods can be solved efficiently, thanks to proximal splitting (Combettes *et al.* 2011), even for very large populations.

These models will be validated on public RNAseq datasets for whom clinical information is available and on CGFL private databases. The achievements of the PhD student will be applied on different types of cancer to enable further explorations of the disease-specific clinical impact of the tumor microenvironment and thus a disease-specific therapy strategy.

Bibliography

Abbas A *et al.,* Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *Plos One,* 2009. 4: 1-16.

Becht E. *et al*. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression,» *Genome Biology,* 2016, 17:218- 2016

Bogdan M, *et al*. SLOPE-Adaptative Variable Selection via Convex Optimization. *Ann. Appl. Stat.* ,2015. 9 (3): 1103-1140

Combettes, P and Pesquet, J-C. Proximal Splitting Methods in Signal Processing. In: Fixed-Point Algorithms for Inverse Problems in Science and Engineering. 2011. 185-212. Springer, New York, NY

Deledalle C, Papadakis N, Salmon, J., Vaiter S. (2017) CLEAR: Covariant LEAst Square Refitting with applications to image restoration. *SIAM Journal on Imaging Sciences (*SIAM J. Imaging Sci).10 (1), 243-284

Newman A *et al.*, Robust enumeration of cell subsets from tissue expression profiles, *Nature Methods,* 2015. 12: 453-457.

Tibshirani R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B,* 1996. 58(1): 267-288.

Yuan M, Lin, Y. Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society, Series B*, 2006. 68: 1, 49-67.

Connaissances et compétences requises :

This PhD proposal is at the interface between biology and mathematics. We are looking for candidates with very good skills in statistics and optimization, as well as a real interest for biological problems.