Comparaison de méthodes de régularisation bayésienne pour la prédiction de la survie en grande dimension

Encadrant/Contact : Damien Drubay, PhD - Service de Biostatistique et d'Epidémiologie de Gustave Roussy / INSERM U1018 - Equipe Oncostat (damien.drubay@gustaveroussy.fr)
Lieu : Gustave Roussy, 114 Rue Edouard Vaillant 94800, Villejuif
Début du stage : Entre début Février et début Avril
Durée du stage : 5 à 6 mois
Date limite de candidature : 31/01/2020

Contexte

A l'ère du "big data", la grande dimension est présente dans tous les domaines impliquant l'analyse de données (analyse d'image, actuariat, finance, marketing, écologie,...). Plus spécifiquement dans le cadre de la santé (et particulièrement de la cancérologie), la démocratisation des techniques de recueil d'information "omiques" permet de générer quotidiennement de grandes quantités de données. Leur analyse est un problème complexe du fait du faible nombre de patients (quelques dizaines/centaines) par rapport au très grand nombre d'acteurs impliqués dans les mécanismes biologiques complexes (ex : dizaines de milliers de gènes). Dans ce contexte de très grande dimension ($N << P$) où les méthodes d'analyse classiques présentent leurs limites, les pénalisations type norme L1 (ex : LASSO) et L2 (ex : Ridge) sont devenues des références pour l'analyse de ce type de données grâce à leur simplicité et leur efficacité. Dans le contexte bayésien, ces estimateurs correspondent aux modes de posteriors obtenus à partir de priors spécifiques (ex : LASSO $\leftrightarrow$ prior Laplacien). Aussi, de nombreux autres priors ont été développés, ayant chacun leurs spécificités : normal-exponentiel-gamma, generalized double Pareto, generalized beta, Dirichlet-Laplace, horseshoe, R2-D2, Inverse-gamma-gamma... Leurs propriétés en termes oracle et de minimaxité ont été établies dans le cadre du modèle linéaire, mais peu de travaux se sont intéressés à leurs performances dans le cadre d'autres modèles. Au cours de ce stage, nous nous proposons d'évaluer et comparer leur performance en terme de prédiction dans le cadre de la modélisation de la survie par simulations.

Objectif du travail

L'objectif principal de ce stage sera de comparer une liste exhaustive de priors de régularisation à l'aide d'une série de simulations. En se basant sur le modèle de Weibull, ces simulations prendront en compte différents ratios P/N pour évaluer leur extensibilité à des jeux de données réelles qui peuvent être de très haute dimension pour de faibles effectifs. De plus, pour étudier leur comportement dans le cadre de l'analyse de données génomiques, l'influence de la structure de corrélation des prédicteurs sera évaluée. Ces priors seront également utilisés pour l'analyse d'un ou plusieurs jeux de données réelles disponibles au sein de l'institut Gustave Roussy.

Profil du candidat

Ce stage d'une durée de 5 à 6 mois s'adresse à des étudiants d'école d'ingénieur, Master 2 ou équivalent en biostatistique/statistique/mathématique, ayant un attrait pour la modélisation prédictive. Le candidat devra maitriser les bases de l'inférence bayésienne (prior, posterior, algorithmes de Gibbs/Metropolis-Hastings). Des connaissances sur des méthodes plus avancées (prior de rétrécissement, Hamiltonian Monte Carlo) seront un plus. Le candidat devra également avoir

des connaissances en R (et/ou Python) qui sera utilisé pour réaliser les simulations. Les posteriors seront obtenus en utilisant l'algorithme de Monte Carlo hamiltonien, implémenté dans le logiciel Stan (interfaçage avec R : package rstan; avec Python: PyStan). Un niveau d'anglais écrit convenable sera également demandé, dans la perspective de la rédaction d'un article dans une revue internationale.

Matériel mis à disposition de l'étudiant

Pour atteindre ses objectifs, l'étudiant aura en plus de son ordinateur fixe personnel accès à un serveur de calcul, afin de pallier la potentielle demande de calcul intensif.