Sélection d'effets aléatoires en grande dimension en modèle de survie

Que ce soit dans les domaines de la santé, de l'agronomie ou de la génétique, de nombreuses études de survie ont pour objectif d'étudier le temps d'apparition d'un évènement d'intérêt tels que la mort ou l'apparition d'une maladie. Ces études permettent d'étudier les liens entre les variables explicatives et les temps de survie. En génétique, l'un des enjeux majeurs est de sélectionner les régions du génome impliquées dans l'apparition de l'évènement. Or avec les outils modernes de la génomique, l'information génétique disponible au travers de marqueurs moléculaires, est devenue si massive que des méthodes appropriées ont été développées.

Depuis quelques années de nombreux travaux ont été menés pour la sélection de variables en grande dimension dans le cadre des modèles de survie (Bonato et al., 2010 ; Hastie et al., 2015; Tang et al., 2017). Ces approches s'appuient notamment sur les approches de pénalisation (Lasso, elastic net,..) développées dans le cadre des modèles linéaires. Elles permettent la sélection d'effets fixes comme les marqueurs moléculaires. Récemment de nouvelles approches combinant l'information moléculaire et pedigree ont fournis une information génétique le long du génome non plus au travers d'effets fixes mais d'effets aléatoires. Ces derniers sont associés à des positions génomiques avec des structures connues de matrices de variance-covariance. La sélection porte désormais sur les effets aléatoires afin d'identifier les régions du génome impliquées.

Dans le cadre des modèles linéaires mixtes (LMM) des approches bayésiennes de sélection d'effets aléatoires de type Spike and Slab ont été développées au cours d'un stage du master 2 biostatistique de Montpellier réalisé l'année dernière par Benjamin Heuclin (Heuclin 2017). L'objectif de ce stage est d'adapter les approches bayésiennes de sélection développées dans le cadre des LMM au modèle de survie avec l'exploration de différents modèles tels que le modèle de Cox ou « the accelerated failure time » modèle. Puis en fonction de l'avancée et des difficultés rencontrées, l'extension de ces méthodes à la sélection de groupes de variables sera envisagée.

Les différentes approches développées seront appliquées dans un contexte d'amélioration génétique du palmier à huile sur des données issues d'essais génétiques mis en place par le CIRAD (Centre de coopération internationale en recherche agronomique pour le développement) et PalmElit, sa filiale pour la sélection et la production de semences.

Contacts encadrants : Marie Denis, Frédéric Mortier (marie.denis@cirad.fr, frederic.mortier@cirad.fr)

Durée : 4-6 mois

Indemnité mensuelle : 550 euros

Thématiques : Sélection de variables, modèles de survie, méthodes Bayésiennes, génétique quantitative, étude de génétique

Lieu du stage : Montpellier, CIRAD Lavalette, UMR AGAP

Bonato, V., Baladandayuthapani, V., Broom, B. M., Sulman, E. P., Aldape, K. D., & Do, K. A. (2010). Bayesian ensemble methods for survival prediction in gene expression data. *Bioinformatics*, 27(3), 359-367.

Hastie, T. et al. (2015) *Statistical Learning with Sparsity—the Lasso and Generalization*. CRC Press, New York.

Heuclin (2017). Sélection d'effets fixe et aléatoires en grande dimension dans les modèles linéaires mixtes

Tang, Z., Shen, Y., Zhang, X., & Yi, N. (2017). The spike-and-slab lasso Cox model for survival prediction and associated genes detection. *Bioinformatics*, btx300.