Sélection de variables par approximation de la norme L0 dans un modèle de Poisson log-normal

Togo Jean Yves Kioye*[1], Paul-Marie Grollemund[2,3], Jocelyn Chauvet[4,5], and Christophe Chassard[3]

[1]Unité Mixte de Recherche sur le Fromage – Université Clermont Auvergne, INRAE, VetAgro Sup, UMRF0545 Unité Mixte de Recherche sur le Fromage, 20 Côte de Reyne, 15000 Aurillac, France – France

[2]Laboratoire de Mathématiques Blaise Pascal – Centre National de la Recherche Scientifique, Université Clermont Auvergne – France

[3]Unité Mixte de Recherche sur le Fromage – VetAgro Sup - Institut national d'enseignement supérieur et de recherche en alimentation, santé animale, sciences agronomiques et de l'environnement, Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement, Université Clermont Auvergne – France

[4]Laboratoire Angevin de Recherche en Ingénierie des Systèmes – Université d'Angers, Université d'Angers : EA7315 – France

[5]Centre de recherche de l'ICES – Institut Catholique de Vendée – France

Résumé

Les questions sur les communautés microbiennes déterminant la qualité du lait se sont développées ces dernières années grâce à des techniques modernes en génomique qui fournissent des données sur l'abondance des espèces. Dans ce contexte, il est possible d'utiliser le modèle de Poisson log-normal multivariée pour ajuster les données de comptage multivariées relatives aux abondances des espèces. Cette modélisation offre la possibilité d'intégrer une couche de régression permettant de modéliser la relation entre des covariables et les données de comptage. Toutefois, dans un contexte caractérisé par une multitude de covariables, le modèle actuel et son implémentation ne sont pas en mesure d'identifier les covariables qui ont une pertinence majeure pour expliquer les variations d'abondance au sein des communautés microbiennes, ou autrement dit : faire de la sélection de variables. Pour résoudre ce problème, des méthodes telles que le lasso sont couramment utilisées, mais elles nécessitent l'ajustement d'un paramètre de régularisation. Ce paramètre est généralement choisi en minimisant l'erreur de validation croisée ou en optimisant un critère d'information. Une alternative récente est le critère d'information lisse appelé SIC (Smooth Information Criterion). Cette méthode est hybride, car elle optimise simultanément un critère d'information et une approximation de la norme L0 des coefficients de régression. Nous proposons d'insérer le critère SIC dans l'implémentation de l'algorithme d'ajustement du modèle PLN sans augmenter considérablement le temps de calcul. Contrairement au lasso et à ses extensions, l'application de l'approche SIC ne se fait pas par le biais d'un algorithme coûteux comme la procédure de validation croisée. Les performances de cette méthode de sélection de variables seront évaluées au travers d'une étude de simulation et seront illustrées dans le cadre d'une étude cherchant à identifier les facteurs importants qui contribuent à la diversité des communautés microbiennes intervenant dans le processus de production du lait.

*Intervenant

Mots-Clés: Sélection de variables, modèle Poisson Log, Normal, régularisation, critère d'information lissé, communautés microbiennes