SÉLECTION DE VARIABLES DANS UN MODÈLE PLN
APPLICATION À L'ÉTUDE DES COMMUNAUTÉS MICROBIENNES DANS LE PROCESSUS DE PRODUCTION DU LAIT

Grollemund PM[1], Chassard C[2], Chauvet J[3], et Theil S[2].
[1]LMPB, Université Clermont Auvergne; [2]UMRF, INRAE; [3]ISPED, Institut Catholique de Vendée

EN RECHERCHE D'UNE SECONDE DEMI-BOURSE DE THÈSE

CONTEXTE ET MOTIVATIONS

Comprendre ce qui sous-tend la qualité du lait
- Qualité sensorielle et composition biochimique
- Biodiversité prairiale et pratiques d'élevage
- Lien avec les différentes **communautés microbiennes**

Développer des approches à l'échelle du système agri/agroalimentaire :
- Impact des pratiques d'élevage
- Les **flux** microbiens d'amont en aval

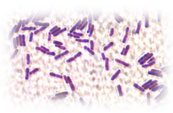
PROCESSUS DE PRODUCTION DU LAIT

Plusieurs écosystèmes concernés
- L'environnement : sol, herbe, air
- La ferme : grange, litière, nourriture
- La vache : les trayons , les fécès, le rumen, le lait
- Stockage du lait, fromage

Informations relevées
- **Composition** microbienne de chaque écosystème
- Composition physico-chimique
- Présence de **pathogènes**
- Typologie du système d'élevage

COMMUNAUTÉS MICROBIENNES

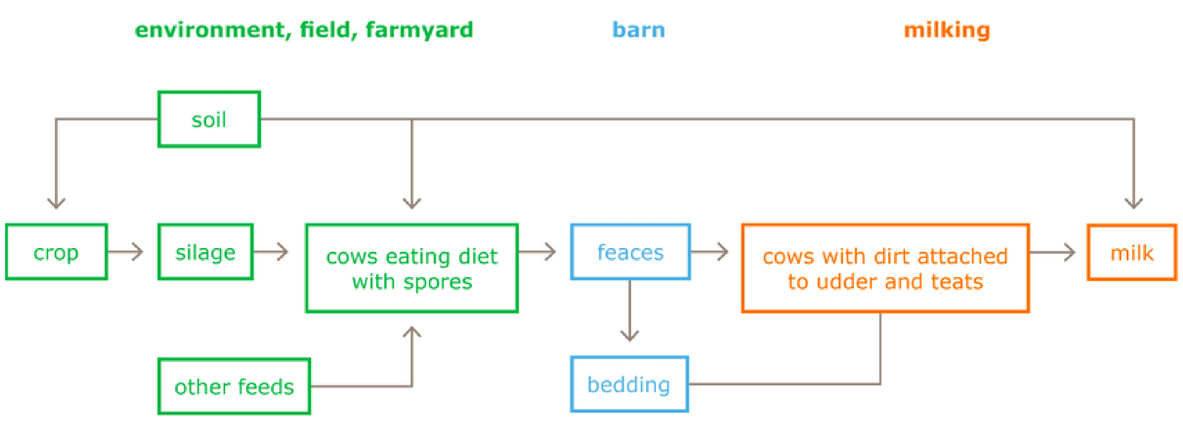
Mesure de la présence de microbes dans le lait [1, 2]
- Techniques modernes de génomiques : abondance de chaque espèce
- Suivi des espèces ou des souches de bactéries sur plusieurs écosystèmes
- Donne de **gros jeux de données** (besoin en **bioinformatique**)



Strain 1	ATCGCTAGTCAGCTAGCTAGCT
Strain 2	ATCGCTACTCAGCTAGCTATCT
Strain 3	ATCGCTAGTCAGCTAGCTATCT

Bases de données : projets en cours
- Projet Amont Saint Nectaire : environnement des animaux
- Projet MINDS : différence de microbiote en fonction de la diversité botanique
- Projet TANDEM : différence de microbiote, agroécologie / agriculture intensive, résilience aux perturbations

Données recouvrant les différents écosystèmes susceptibles de contenir des communautés microbiennes d'intérêt



Source : Galama et al. (2015) Sustainability aspects of ten bedded pack dairy barns in the Netherlands.

De nouvelles expériences
- Phase suivante du projet TANDEM, en condition réelle et non-contrôlée
- Besoin de développer des procédures d'analyse pour ce type de données et ce type de problématiques
- Besoin grandissant de suivi des pathogènes (biosécurité), identifier les facteurs dominants expliquant la présence de pathogènes

OBJECTIFS EN AGRONOMIE

Les liens entre écosystèmes microbiens et qualité du lait
- Impact de la diversité botanique
- Lien avec les différentes communautés microbiennes
- Identifier ce qui assure la **biosécurité**
- Impact des pratiques d'élevage
- Réaction des écosystèmes microbiens à une **transition agroécologique**
- Communautés microbiennes spécifiques à certains écosystèmes ?
- Connaissane et prise en compte des entités microbiennes selon le système

Besoin d'outils pour l'étude des flux microbiens
- **Modélisation (biostatistique)**, bioinformatique
- Les différentes questions peuvent se reformuler comme : modéliser un lien entre les abondances microbiennes des écosystèmes (et leurs flux) et des **facteurs exogènes**

MODÉLISATION

Analyse standard
- Analyse différentielle
- Prétraitement et normalisation ?
- Surabondance de communautés dans des conditions différentes

Complément méthodologique proposé
- **Modèle Poisson Log-Normal**
- Modéliser les comptages multivariées [3, 4, 5]
  $y_i = (y_{i,1}, \ldots, y_{i,q})$
- Inclut des régresseurs dans une sous-couche
  $x_{i,1}, \ldots, x_{i,p}$

$$y_{ij}|\lambda_{ij} \sim P\left(\exp(o_{ij} + \lambda_{ij})\right)$$
$$\lambda_i \sim N_q(\mu_i, \Sigma)$$
$$\mu_{ij} = x_i^T \theta_j$$

Informations à en retirer
Différences dues à certains facteurs entres communautés et écosystèmes (qualité du lait)
Contribution de chaque régresseur concernant l'abondance microbienne (biosécurité)

Inférence
- **Approximation variationnelle** : ELBO
- Approximation de la log vraisemblance au sens de KL
- Optimisation via (V)EM

SUJET DE THÈSE

Axe 1 : Modélisation, sélection de variables
- Reformuler problématique comme une sélection de variables [6, 7]
- A mettre en place dans un modèle PLN (régularisation ou spike-and-slab)
- Approches bayésiennes [8, 9, 10] ou fréquentistes [11] ?

Axe 2 : Implémentation et estimation
- Problèmes poser par une modélisation parcimonieuse sur l'implémentation
- Autres approches possibles ?

Axe 3 : Travail fondamental
- Modèle parcimonieux et procédure d'optimisation **VEM** de la **ELBO** (Evidence Lower BOund)

Axe 4 : Application
- Appliquer la méthodologie proposé sur des données des différents projets
- Mettre en place un **processus d'analyse** pour les prochaines études
- Implémentation efficace : `Rcpp` et/ou sous python

VERROUS IDENTIFIÉS

Implémentation :
- Méthode d'optimisation complexe
- Implémentation efficace déjà existante : package `PLNmodels`
- Opportunité possible : utilisation d'algorithme d'optimisation en machine learning

Application :
- Big data, et données complexes
- Gestion de bases de données (bioinformatique)

Travail fondamental :
- Quantification de l'approximation ELBO dans un contexte de sur-paramétrisation
- Garanti de converge vers un optimum en norme L_0

RÉFÉRENCES

[1] **Chassard, C.** et al. "Lactic Starter Dose Shapes S. aureus and STEC O26: H11 Growth, and Bacterial Community Patterns in Raw Milk Uncooked Pressed Cheeses". In: *Microorganisms* 9.5 (2021).
[2] P. L. Ruegg. "The bovine milk microbiome–an evolving science". In: *Domestic Animal Endocrinology* 79 (2022), p. 106708.
[3] Y. Choi et al. "A Poisson log-normal model for constructing gene covariation network using rna-seq data". In: *Journal of Computational Biology* 24.7 (2017), pp. 721–731.
[4] J. Chiquet, S. Robin, and M. Mariadassou. "Variational inference for sparse network reconstruction from count data". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 1162–1171.
[5] J. Chiquet, M. Mariadassou, and S. Robin. "The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances". In: *Frontiers in Ecology and Evolution* 9 (2021).
[6] G. Heinze, C. Wallisch, and D. Dunkler. "Variable selection–a review and recommendations for the practicing statistician". In: *Biometrical journal* 60.3 (2018), pp. 431–449.
[7] J. T Ormerod, C. You, and S. Müller. "A variational Bayes approach to variable selection". In: *Electronic Journal of Statistics* 11.2 (2017), pp. 3549–3594.
[8] P. Carbonetto and M. Stephens. "Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies". In: *Bayesian analysis* 7.1 (2012).
[9] **Grollemund, P.M.** et al. "Bayesian functional linear regression with sparse step functions". In: *Bayesian Analysis* 14.1 (2019), pp. 111–135.
[10] C. Ma et al. "Bayesian EDDI: Sequential Variable Selection with Bayesian Partial VAE". In: (2019).
[11] **Chauvet, J.**, C. Trottier, and X. Bry. "Component-Based Regularization of Multivariate Generalized Linear Mixed Models". In: *Journal of Computational and Graphical Statistics* 28.4 (2019).