

UDOPIA

Universit  Paris-Saclay's Doctoral Program in Artificial Intelligence



Projet doctoral

Titre du projet, en fran ais et en anglais

R duction de dimension par m thodes d'apprentissage   facteurs latents structur s et r gularis s pour la mod lisation de la r ponse d'un syst me biologique   la modification de son environnement.

Dimension reduction by structured and regularised latent factor learning methods for modelling the response of a biological system to changes in its environment.

1. Résumé du projet (15 lignes au plus en français et en anglais)

La thèse vise à proposer plusieurs modèles à variables latentes avec structuration de dépendances existantes au sein des variables réponses (données omiques de différentes nature) et les méthodes d'apprentissage associées.

L'objectif est d'identifier ou de mesurer l'effet d'un changement de l'environnement (présence de contaminants par exemple) sur un système biologique à partir de données environnementales, de données sur l'état du système (par exemple son métabolome) et de son écosystème (son métagénome). Le modèle et son inférence doivent pouvoir prendre en compte la grande dimension des données (vecteur réponse $p=10^3$ - 10^4 pour un nombre d'échantillons de l'ordre de $n=500$) et leurs spécificités telle que la non-normalité des données de métagénomique et une dépendance éventuellement non linéaire entre les variables latentes et la variable réponse. Dans un premier temps, le doctorant s'intéressera aux modèles du type analyse en composantes principales (ACP) probabiliste de façon à réduire la dimension des variables réponses. Des versions régularisées seront introduites afin d'obtenir une interprétabilité des résultats. Par la suite, des versions permettant de prendre en compte des dépendances entre variables latentes seront considérées. La non-gaussianité des données et la non-linéarité des relations seront abordées par des modèles du type Poisson avec excès de zéros ou par des approches du type auto-encodeur variationnel. Les méthodes seront appliquées dans le cadre de deux collaborations portées par l'unité.

The thesis aims at proposing latent variable models with a structure of existing dependencies within the response variables (omic data of different nature) and associated learning methods.

The objective is to identify or measure the effect of an environmental change (e.g. presence of contaminants) on a biological system from environmental data, data on the state of the system (e.g. its metabolome) and its ecosystem (its metagenome). The model and its inference must be able to take into account the large size of the data (response vector $p=10^3$ - 10^4 for a number of samples of the order of $n=500$) and their specificities such as the non-gaussianity of the metagenomic data and a possibly non-linear dependence between the latent variables and the response variable. Initially, the doctoral student will focus on probabilistic principal component analysis (PCA) models in order to reduce the dimension of the response variables. Regularized versions will be introduced in order to obtain an interpretability of the results. Subsequently, versions allowing to take into account dependencies between latent variables will be considered. The non-gaussianity of the data and the non-linearity of the relationships will be addressed by zero inflated Poisson models or by variational auto-encoder type approaches. The methods will be applied in the framework of two collaborations carried out by the unit.

2. Mots-clefs (en français et en anglais)

Mots Clefs: apprentissage statistique • réduction de dimension • intégration de données • données omiques • données dépendantes • écologie

Keys words: statistical learning • dimension reduction • data integration • omics data • dependent data • ecology

3. Thèmes UDOPIA (liste)

Statistical learning • Energy, climate and environment

4. Type de financement demandé à UDOPIA (à 100% ou co-financement)

100%

5. La, ou les unité(s) de recherche soutenant le projet

Unité MIA-Paris-Saclay, UMR INRAE AgroParisTech,

6. Le(s) directrices ou directeur(s) de thèse,

Julien Chiquet (DR, MIA Paris-Saclay) et Sophie Donnet (CR-HDR, MIA Paris-Saclay). Julie Aubert (IR INRAE) viendra renforcer l'encadrement notamment pour l'application en métagénomique.

7. Ecole Doctorale de rattachement et l'Etablissement opérateur d'inscription

EDMH. Inscription Université Paris Saclay

8. Numéro du dépôt ADUM, si existant

39957

Description détaillée du projet doctoral

1. Contexte scientifique, social et sociétal du projet

Dans un contexte de réchauffement climatique ou plus largement d'accroissement de la pression de l'activité humaine sur les milieux naturels, il est nécessaire d'être capable de surveiller la réponse d'un système biologique (plante par exemple) aux changements dûs la modification de son environnement. Il est maintenant assez courant de mesurer l'état d'un système biologique par son métabolome (i.e. la liste de métabolites présentes dans l'échantillon biologique).

D'autre part, l'écosystème dans lequel est plongé le système biologique d'intérêt est décrit par l'analyse du microbiome, c'est-à-dire les abondances des espèces microbiennes entourant le système biologique (aussi appelées données métagénomiques). Enfin, l'environnement du système biologique peut être décrit par un certain nombre de variables environnementales qui peuvent être géographiques, climatiques, ou chimiques (présence d'engrais, de contaminants par exemple ...). L'hypothèse biologique est que la modification de l'environnement du système (présence de contaminants, présence d'engrais...) entraîne une modification non seulement du système biologique mais aussi de son environnement écologique et des interactions entre système et environnement.

L'analyse intégrative des données (métabolomiques et métagénomiques) sur une collection d'échantillons biologiques doit permettre de comprendre et quantifier les réactions du système biologique à un type de perturbation donné de son environnement, une fois pris en compte l'effet des autres facteurs environnementaux.

2. Objectif scientifique

La difficulté du problème vient de la dimension du vecteur réponse comparée au nombre d'échantillons (réalistement 500). En effet, un système biologique est classiquement décrit par 10^2 à 10^4 métabolites et l'analyse d'un microbiome aboutit facilement à l'identification de 10^3 à 10^4 espèces microbiennes. La structure de dépendance de ces données réponses est intéressante d'un point de vue biologique.

En raison de la taille des données réponses, il est inenvisageable d'estimer tous les effets des covariables, ni la structure de dépendance entre toutes les réponses sans hypothèses ou contraintes additionnelles. Une approche par **réduction de dimension**, couplée à un contrôle du nombre de paramètres par **régularisation** est donc nécessaire. Ceci peut intervenir à plusieurs niveaux, i.e. à la fois sur les variables réponses et sur les variables explicatives.

Une approche fondée sur une analyse en composantes principales (ACP) probabiliste pour réduire la dimension des réponses paraît une première piste à envisager. L'idée est de considérer qu'un petit nombre de variables latentes décrit une grande partie de la variabilité des réponses. L'effet des covariables environnementales serait alors mesuré sur cet ensemble de variables latentes (résumées).

Cependant, le risque est d'aboutir à un modèle ininterprétable biologiquement. Une première solution consisterait alors à introduire de la régularisation par pénalisation forçant l'identification d'un petit nombre de facteurs environnementaux dans l'espace latent (sur les effets des covariables, et non sur les axes les composantes de l'ACP comme cela est fait dans l'ACP sparse). Par ailleurs, la structure de dépendance des variables latentes peut aussi avoir un sens biologique. Traditionnellement supposée sphérique dans l'ACP probabiliste, nous pouvons relâcher cette hypothèse dans le cadre des modèles à facteurs tout en contrôlant le nombre de paramètres en supposant cette structure de dépendance creuse ou structurée, afin d'aboutir à plus d'interprétabilité biologique.

Par la suite, des solutions adaptées à la spécificité des données multi-omiques doivent être envisagées. En particulier, l'ACP probabiliste suppose une dépendance linéaire entre les variables latentes et les variables réponses, dans un cadre gaussien. Cependant, la gaussianité des données est une hypothèse très forte qui est loin d'être respectée dans le cas des données de métagénomique. Les méthodes proposées devront d'adapter aux données de comptage avec excès de zéros (qui correspondent à des absences de métabolites). On pense à des versions inflatées en zéro du modèle Poisson lognormal par exemple, modèle largement étudié par plusieurs des encadrants. De plus, les approches d'apprentissage profond type auto-encodeurs variationnels fournissent des approches de réduction de dimension alternatives et un cadre génératif qui permettent d'échapper à la linéarité, et dont la pertinence devra être testée.

L'objectif de la thèse est donc l'écriture d'un ou plusieurs modèles statistiques répondant à ces attentes et le développement de la méthode d'inférence statistique correspondante. Cette tâche d'apprentissage statistique nécessite l'utilisation de méthodes d'optimisation adaptées à la grande dimension.

3. Impact attendu

Les méthodes développées seront implémentées dans des packages facilement diffusables (R par exemple) et appliquées sur des données fournies par nos partenaires.

En particulier, le MNHN s'intéresse à la mesure de l'effet des contaminants métalliques sur l'état de santé des mousses. Une collaboration est déjà en cours entre notre unité et l'unité Mixte de Service Patrimoine Naturel du Muséum d'Histoire Naturelle (MNHN) qui est le coordinateur pour la France d'un programme européen de surveillance des dépôts métalliques atmosphériques par les mousses terrestres, le dispositif BRAMM (Biosurveillance des Retombées Atmosphériques Métalliques par les Mousses). Les données de métabolites ont déjà été collectées. Les données de métagénomique seront collectées dans l'année à venir.

Par ailleurs, l'unité a obtenu le financement d'un projet sur un fond d'amorçage de la Direction de la Recherche et de la Valorisation d'AgroParisTech pour le projet intitulé SYMBIOMODEL (*Modélisation des interactions trophiques entre racines et microorganismes bénéfiques du sol*) impliquant 2 autres UMR (IJPG -Institut Jean-Pierre Bourgin- et EcoSys -Écologie fonctionnelle et écotoxicologie des agroécosystèmes-). Ce projet s'intéresse à la modélisation des interactions trophiques entre les

plantes (ici les céréales) et leur microbiote tellurique (ici champignons mycorhiziens) et à la détection de différences entre 2 lignées de plants.

Ces deux contextes d'application seront sollicités pour appliquer les méthodologies proposées.

4. Justification de l'intérêt du projet pour le programme UDOPIA

Le sujet proposé répond à un besoin identifié par les partenaires biologiques. La réponse nécessite le développement de méthodes d'apprentissage innovantes interprétables, s'appuyant sur une large littérature et une large expertise des encadrants dans le domaine.

5. Compétences spécifiques des encadrants et équipes

Julien Chiquet est spécialiste de l'apprentissage statistique pour les données de grande dimensions avec application pour les sciences du vivant, en particulier en .

Sophie Donnet travaille depuis de nombreuses années dans la modélisation et l'inférence d'interactions entre entités biologiques. En particulier, elle est spécialiste de l'analyse de réseaux impliquant des acteurs de diverses natures.

Enfin, Julie Aubert est spécialiste de l'analyse de données métagénomiques et métabolomiques et des méthodes qui s'y rapportent.

6. Programme de recherche envisagé

Le programme de recherche suivra les étapes de modélisations citées au paragraphe 2:

1- Année 1:

PCA probabiliste avec contraintes sur les facteurs et pénalité sur les co-variables.

Modélisation de la structure de dépendance des facteurs.

2- Année 2: Relaxation de la dépendance linéaire et utilisation d'auto-encodeurs.

3- Année 3: Relaxation de la gaussianité et adaptation au cas des données avec 0 sur-représentés.

7. Résultats attendus et perspectives en recherche et applications.

La thèse devra donner lieu à publication dans un journal statistique classique (*Ann. Applied. Stat., Stat & Comput*) mais aussi dans un journal méthodologique dédié à l'écologie (*Meth. Ecol. Evol*) ou dans un journal d'écologie classique (*Ecology Letters*). Le développement d'outils logiciels librement disponibles (package R par exemple) est également prévu.

8. Conditions de réalisation de la thèse

Le doctorant bénéficiera des mêmes conditions de travail que les autres chercheurs du laboratoire d'accueil (moyens de calcul, missions, etc). L'encadrement s'organisera autour au minimum d'un rendez-vous hebdomadaire formel et de discussions continues plus informelles. Les données seront fournies par les partenaires biologistes.

9. Références bibliographiques, ou autre produits de la recherche, des équipes d'accueil

Coupling ecological network analysis with high-throughput sequencing-based surveys: lessons from the Next-Generation Biomonitoring project, M. Dubart, P. Alonso, D. Barroso-Bergada, N. Becker, K. Bethune, D. A. Bohan, C. Boury, M. Cambon, E. Canard, E. Chancerel, **J. Chiquet**, P. David, N. de Manincor, **S. Donnet**, A. Duputié, B. Facon, E. Guichoux, T. Le Minh, S. Ortiz-Martínez, L. Piouceau, A. Sacco–Martret de Préville, M. Plantegenest, C. Poux, V. Ravigné, S. Robin, M. Trillat, C. Vacher, C. Vernière, François Massol. Accepted by Applied Ecology Resources

Block models for multipartite networks. Applications in ecology and ethnobiology. A. Bar-Hen, P. Barbillon & **S. Donnet**. Statistical Modelling, First Published December 18, 2020.

Accelerating Bayesian estimation for network Poisson models using frequentist variational estimates. **Sophie Donnet**, Stéphane Robin. JRSSC, 16 April 2021.

The Poisson-Lognormal Model as a Versatile Framework for the Joint Analysis of Species Abundances **Chiquet, Julien**, Mariadassou, Mahendra, and Robin, Stéphane, Frontiers in Ecology and Evolution, 2021

Joint species distributions reveal the combined effects of host plants, abiotic factors and species competition as drivers of species abundances in fruit flies. Facon, Benoit, Hafsi, Abir, Masselière, Maud Charlery, Robin, Stéphane, Massol, François, Dubart, Maxime, **Chiquet, Julien**, Frago, Enric, Chiroleu, Frédéric, Duyck, Pierre-François, and Ravigné, Virginie, Ecological Letters, 2021

Variational inference for probabilistic Poisson PCA; **Chiquet, Julien**, Mariadassou, Mahendra, and Robin, Stéphane Ann. Appl. Statist., 2018

A multivariate variable selection approach for analyzing LC-MS metabolomics data Perrot, M., Lévy-Leduc, C., **Chiquet, Julien**, Sansonnet, L., Brégère, M., Étienne, M.-P., Robin, Stéphane, and Genta-Gouve, G. SAGMB, 2018

Unraveling negative biotic interactions determining soil microbial community assembly and functioning. Romdhane, S., Spor, A., **Aubert, J.** Bru, D., Breuil, M.-C., Mounier, A., Ouadah, S., Tsiknia, M. and Philippot, L. ISME J 16, 296–306 (2022). <https://doi.org/10.1038/s41396-021-01076-9>

Model-based biclustering for overdispersed count data with application in microbial ecology. **Aubert, J.**, Schbath, S, Robin, S. Methods Ecol Evol. 2021; 12: 1050– 1061.
<https://doi.org/10.1111/2041-210X.13582>