Programme des Journées d'automne du GT "Entropie mots stat"

Université de Caen, Campus II salle S3-247.

Jeudi 18 novembre 2021 14h30

Pablo Rotondo Université Gustave Eiffel Champs-sur-Marne

Entropie renormalisée : partitions balancées et changement de base

Dans le cas de deux sources de deux sources d'entropie positive, Dajani et Fieldsteel ont démontré en 2001 que la quantité de symboles d'une source S_2 qui sont produits par chaque symbole d'une source S_1 est essentiellement le quotient $h(S_1) / h(S_2)$ des entropies respectives. Plus précisément, si, grâce à n symboles de S_1 sur l'entrée, on peut produire m symboles dans la source S_2 , alors le quotient m/n tend vers $h(S_1)/h(S_2)$ lorsque n tend vers l'infini.

Nous obtenons un cadre plus général qui s'applique aussi aux sources d'entropie nulle. Notre généralisation du théorème de Dajani-Fieldsteel fait intervenir une fonction f dite de poids ou entropie renormalisée, qui joue un rôle analogue à l'entropie. Quand cette fonction de poids f existe (presque partout ou en mesure), disons f_1 pour S_1 et f_2 pour S_2 , nous montrons que la limite qui intervient dans le théorème de Dajani-Fieldsteel est $\lim f_2(m)/f_1(n) = 1$.

Quand l'entropie h est positive, la fonction de poids f est tout simplement linéaire $f(t) = h \cdot t$. Pour certains cas d'entropie nulle, et en particulier pour plusieurs sources importantes de théorie des nombres, nous démontrons que la fonction de poids existe toujours et est sous-linéaire.

Collaboration avec Valérie Berthé, Eda Cesaratto et Martín D. Safe.

15h15

Blanche Saint Bêat Laboratoire d'Ecologie Pélagique (PDG-ODE-DYNECO-PELAGOS) Ifremer Brest
Estimation des flux du réseau trophiques par modélisation inverse (LIM) : choix d'une fonction objective.

En écologie marine, les propriétés des réseaux trophiques révèlent le fonctionnement des écosystèmes. Ces propriétés nécessitent un réseau quantifié. Alors que les mesures de flux in situ sont souvent parcellaires, la quantification des flux devient un problème sous-déterminé pour lequel il existe une infinité de solutions qui s'ajustent au jeu de données in situ. Une manière rigoureuse de résoudre un tel problème est la MODELISATION LINEAIRE INVERSE (LIM). La version la plus récente du LIM permet de déterminer un set de solution pour chaque flux grâce à un algorithme d'échantillonnage aléatoire. La description de l'incertitude associée à chaque flux permet de développer des tests statistiques indispensables dans un contexte de comparaison d'écosystèmes ou d'analyse d'impact d'une perturbation. Cependant, pour certaines analyses plus précises une seule solution par flux est nécessaire. Une question se pose alors : comment choisir une seule solution parmi toutes celles proposées ? Déterminées selon des critères mathématiques ou écologiques, différentes fonctions objectives, permettant la sélection d'une seule valeur pour chaque flux, ont été testées. Par une stratégie de dégradation de modèles LIM très bien documentés et la réestimation des flux des réseaux trophiques associés, 10 fonctions objectives ont été testées. Si certaines fonctions écologiques donnent de bons résultats, la fonction de la moyenne apparaît comme celle qui estime le mieux les flux du réseau trophique quelle que soit la quantité d'information intégrée au modèle LIM. La reprise récente de cette étude, avec une approche mathématique, a mis à jour de nouvelles conclusions, qui vous seront présentées par Valérie Girardin.

16h

Valérie Girardin UNICAEN LMNO Caen

Quelques questions sur l'utilisation de la théorie de l'information en analyse inverse de réseaux trophiques par des méthodes MCMC

L'algorithme LIM-MCMC de Van Den Meersche (2009) est couramment utilisé par les écologues pour décrire le polygone multidimensionnel déterminé par les contraintes linéaires biologiques sur un écosystème marin. Saint Bêat et al. (2013) ouvre la voie sur une méthodologie nouvelle permettant de

déterminer une solution particulière à partir de l'ensemble des solutions ainsi simulées, par optimisation de fonctions de coût.

Dans son stage, Yann Burkhardt a envisagé différentes fonctions provenant de la théorie de l'information, en les comparant de manière statistique sur un modèle synthétique qui supporte une dégradation systématique. Des questions sur la méthode LIM-MCMC en comparaison de méthodes algorithmiques plus récentes utilisées dans d'autres domaines sont apparues en cours d'étude.

D'après le rapport de stage de magistère de Yann Burkhardt, Université Paris-Saclay Orsay

16h45 pause

17h15

Justine Lequesne Centre François Baclesse Caen

Construction d'un score de toxicités cumulées pour modéliser l'évolution de la qualité de vie pendant un traitement : application sur des données de patientes traitées pour un cancer de l'ovaire

Avec les progrès récents en matière de recherche médicale, la survie des personnes atteintes de cancer s'est nettement améliorée depuis ces dernières décennies, au prix de fortes contraintes pour le patient. Maintenir une bonne qualité de vie chez ces patients est devenue une préoccupation centrale dans leur prise en charge.

Dans cet exposé, nous analyserons les scores de toxicités des patientes de l'essai clinique AURELIA, traitées pour un cancer de l'ovaire. Les scores de toxicité seront calculés à partir de l'ensemble des événements indésirables, de ceux reliés au traitement ou seulement de ceux de grade élevé.

L'association entre chaque score de toxicité et la qualité de vie sera évaluée par des modèles mixtes ajustés sur plusieurs facteurs d'intérêt (traitement, age, etc.). Les paramètres de régression standardisés associés à chaque score seront ensuite comparés dans le but d'identifier le score de toxicité qui prédit au mieux l'évolution de la qualité de vie.

Collaboration avec Sophie Lefèvre-Arbogast, Elodie Coquan et Florence Joly

vendredi 19 novembre 2021

10h

Jean-Guy Caputo Laboratoire de Mathématiques INSA Rouen

Analyse de réseaux trophiques: une approche par l'optimisation

Nous introduisons une méthodologie pour étudier les possibles flux de matière dans un écosystème défini par des observations de biomasses et des contraintes biologiques réalistes. Les flux appartiennent à un polygone dans un espace multidimensionnel ce qui rend une exploration statistique difficile en pratique ; nous proposons alors de résoudre un problème d'optimisation convexe. Sept critères basés sur des index écologiques ont été choisis comme fonctions de coût convexes. Les résultats numériques montrent que la méthode est rapide et peut être utilisée pour de grands systèmes. Les solutions de flux minimal sont ensuite analysées par la décomposition chemins-cycles. Leur consistance est aussi vérifiée en introduisant un système différentiel vérifié par les biomasses sur lequel on examine la stabilité du point fixe. La méthode est illustrée sur un modèle réduit d'écosystème et appliquée à des systèmes réalistes.

Collaboration avec Valérie Girardin, Arnaud Knippel, Hieu Nguyen, Nathalie Niquil et Quentin Noguès.

10h45

Nathalie Niquil et Quentin Noguès UNICAEN, BOREA Caen

Quelques exemples d'applications du LIM MCMC à l'analyse des impacts humains sur le fonctionnement des écosystèmes marins.

Qu'est ce qu'un écosystème marin en « bon état écologique » ? Cette question est devenue capitale pour les états de l'Union Européenne lorsque cette notion de « bon état écologique » des réseaux trophiques est entrée dans le droit de l'UE par la Directive Cadre Stratégie pour le Milieu Marin (en 2008 à l'UE et 2010 dans le droit français). Nous proposons d'utiliser les indices ENA (Ecological Network Analysis), calculés à partir du LIM-MCMC, pour décrire le changement d'état du réseau trophique avant et après des impacts, observés ou simulés. Nous développerons en particulier ici le cas du futur parc éolien de Courseulles-sur-mer dont nous avons décrit l'effet prévisible sur le fonctionnement de l'écosystème en étudiant son cumul avec les effets du changement climatique sur la

répartition des espèces qui suivent vers le Nord leur optimum de température et de salinité. Les résultats démontrent la capacité des indices ENA à décrire et comprendre les effets cumulés à l'échelle de l'écosystème. En utilisant une approche d'analyse de sensibilité, notre étude montre que les indices ENA pourraient être des outils utiles pour aider les gestionnaires dans leurs décisions : par exemple, sur l'intérêt écologique d'ouvrir ou fermer la zone à la pêche entre les mats d'éoliennes.

11h30

Christine Kéribin Université Paris-Saclay Orsay

Classification non supervisée sous contrainte de mémoire de très grands jeux de données déséquilibrés à partir de comptages marginaux (Frugal Gaussian clustering of huge imbalanced datasets through a bin-marginal approach)

En classification non supervisée de grands volumes de données, il est intéressant de pouvoir détecter de petites classes qui peuvent avoir une grande valeur. Nous nous plaçons dans un cadre où la taille mémoire est limitée et le jeu de données ne peut être chargé en mémoire. Il est alors courant de recourir au sous-échantillonnage, mais sa capacité à détecter les petites classes est limitée. Considérant la classification non supervisée par modèle de mélange gaussien, nous proposons une approche de réduction de taille par comptages marginaux. Nous en établissons les propriétés théoriques, définissons un algorithme d'estimation et illustrons sa plus-value par rapport au sous-échantillonnage pour la détection de petits clusters.

Collaboration avec Filippo Antonazzo et Christophe Biernacki.