RIS
Réseau Interdisciplinaire autour de la Statistique

Journée thématique « Statistique sous contraintes »

Muséum national d'Histoire naturelle, Paris

Mardi 27 mars 2018

Le RIS, réseau thématique pluridisciplinaire CNRS de l'INEE, organise une journée dédiée à la présentation de méthodes pour le traitement de données sous contraintes. Six présentations s'articuleront autour de deux problématiques : les petits effectifs et les données manquantes. Chaque communication durera 30 minutes et sera suivie de 15 minutes de discussion avec la salle.

Programme

9 h 45 — 10 h 00	Accueil des participants
10 h 00 — 10 h 15	Présentation du RIS par Sandrine Cabut, membre du comité de pilotage du réseau.
10 h 15 — 11 h 00	**Marie-Anne Vibet**, Laboratoire de Mathématiques Jean Leray — Université de Nantes. *Statistique bayésienne : une solution au traitement des petits échantillons ?*
11 h 00 — 11 h 45	**Paul-Marie Grollemund**, Laboratoire IMAG — Université de Montpellier. *Régression linéaire fonctionnelle : une modélisation bayésienne parcimonieuse*
11 h 45 — 12 h 30	**Camelia Protopopescu**, SESSTIM, UMR 1252 — Inserm, IRD, Aix-Marseille Université. *Modèle d'Heckman pour correction du biais de sélection dans le contexte des données longitudinales*
12 h 30 — 14 h 00	Pause repas.
14 h 00 — 14 h 45	**François Husson**, Department Statistics & Computer science — Agrocampus Ouest, Rennes. *Imputation multiple par des méthodes d'analyse factorielle*
14 h 45 — 15 h 30	**Jean Dumoncel**, UMR 5288, Laboratoire d'Anthropologie Moléculaire et Imagerie de Synthèse — Université de Toulouse. *Les contraintes en paléoanthropologie : comment établir des modèles statistiques ?*
15 h 30 — 16 h 15	**Dominique Soudant**, ODE/VIGIES — Ifremer, Nantes. *Application de modèles dynamiques bayésiens aux séries temporelles de surveillance de l'environnement marin*
16 h 15 — 16 h 45	Discussion générale et clôture de la journée.

Inscription

L'inscription à cette journée est gratuite mais obligatoire, avec une date limite fixée au **mardi 27 février 2018**. Le formulaire d'inscription est disponible à l'adresse suivante :

`https://frama.link/RIS2018`

Organisateurs

La journée est organisée, avec le soutien financier de l'INEE, par le groupe de travail « Statistique sous contraintes » du RIS :

Claire Berticat,
CNRS, UMR 5554 (ISEM), Montpellier

Amandine Blin,
CNRS, UMS 2700 (OMSI), Paris

Sandrine Cabut,
CNRS, UMR 7269 (LAMPEA), Aix en Provence

Isabelle Clerc-Urmès,
CHRU, Plateforme d'Aide à la Recherche Clinique, Nancy

Lionel Granjon,
CNRS, UMR 8242 (LPP), Paris

Sandrine Loubière,
AP-HM (DRCI), EA 3279 (CEReSS), Marseille

Bérengère Saliba-Serre,
CNRS, UMR 7268 (ADES), Marseille

Frédéric Santos,
CNRS, UMR 5199 (PACEA), Pessac

Accès

Muséum national d'Histoire naturelle, Amphi Rouelle, 43–47 rue Cuvier, 75005 Paris.
Bus : lignes 24 et 63 (arrêt Cuvier), lignes 67 et 89 (arrêt Cuvier – Jardin des Plantes).
Métro : lignes 7 et 10 (arrêt Jussieu).

Plan de quartier

Résumés des conférences

Session « Petits effectifs »

10 h 15 — 11 h 00 : **Marie-Anne Vibet & Anne Philippe**, Laboratoire de Mathématiques Jean Leray — Université de Nantes.

Statistique bayésienne : une solution au traitement des petits échantillons ?

L'une des particularités de la statistique bayésienne est l'interprétation du paramètre d'intérêt à estimer, appelé θ, en quantité aléatoire et inconnue. Afin de conduire une modélisation bayésienne, une loi sur le paramètre d'intérêt doit être donnée. Cette loi, dite *a priori*, traduit toute l'information disponible sur le paramètre et sur son incertitude. Le choix de cette loi *a priori* a une grande importance dans le cas de petits échantillons de données et doit donc être choisie avec précaution et justifications.

L'inférence bayésienne est une méthode d'estimation probabiliste, l'information obtenue sur le paramètre d'intérêt θ étant donnée par une densité de probabilité appelée la densité *a posteriori*. Cette loi est en fait la mise à jour des connaissances sur θ au vu des données collectées. Généralement, la forme analytique d'une telle densité de probabilité n'est pas facile à exprimer. Cependant, il est possible de simuler une chaine de Markov dont la loi stationnaire est la densité *a posteriori* désirée. En effet, l'algorithme de Monte Carlo par chaines de Markov (MCMC) permet de simuler un échantillon de la densité *a posteriori* pour des modèles bayésiens. À partir d'un tel échantillon, il est possible d'estimer avec une bonne précision la moyenne *a posteriori*, un intervalle de confiance bayésien.

Dans cet exposé, nous présenterons la démarche de la modélisation bayésienne et nous verrons comment mener une inférence bayésienne en traitant un exemple concret. Dans un deuxième temps, nous aborderons la question spécifique du choix de la loi *a priori* sur le paramètre d'intérêt en comparant l'approche non-informative à l'approche subjective. Nous verrons l'importance de ce choix dans le cas de petits échantillons. Nous envisagerons également les avantages et inconvénients de la statistique bayésienne dans le cas de petits échantillons.

Nous illustrerons cet exposé à l'aide d'exemples traités avec le logiciel R et le package `rjags`.

—

11 h 00 — 11 h 45 : **Paul-Marie Grollemund**, Laboratoire IMAG — Université de Montpellier.

Régression linéaire fonctionnelle : une modélisation bayésienne parcimonieuse

Lorsque l'objectif est de comprendre le lien entre un ensemble de courbes (covariable fonctionnelle x) et un ensemble de réels (variable réponse y), on peut utiliser le modèle de régression linéaire appliqué aux données fonctionnelles. La principale difficulté de l'inférence est alors d'estimer la fonction coefficient, un paramètre de grande dimension. Pour cela, une approche largement utilisée consiste à réduire la dimension du problème en écrivant cette fonction à estimer dans une base finie de fonctions et d'estimer ces coefficients dans la base. Si on ne dispose que de peu de données, on fait face à des problèmes d'estimations rendant inaccessible la compréhension du lien entre x et y. De plus, pour comprendre ce lien, une quantité importante est le support de la fonction coefficient. L'estimation de cette quantité est un problème statistique en soi, qui devient d'autant plus complexe lorsqu'on ne dispose que de peu de données.

Nous proposons d'écrire la fonction coefficient de la manière la plus parcimonieuse possible, comme une fonction en escalier, avec très peu d'escaliers. Notre contribution est 1° d'établir un modèle bayésien dont la loi a priori charge seulement l'ensemble de ces fonctions, et 2° de définir des estimateurs de la fonction coefficient et de son support. Nous présenterons une application de la méthode proposée sur des données réelles avec un petit effectif afin de comprendre l'impact des précipitations sur la production de truffes noires du Périgord.

Session « Données manquantes »

11 h 45 — 12 h 30 : **Camelia Protopopescu**, SESSTIM, UMR 1252 — Inserm, IRD, Aix-Marseille Université.

Modèle d'Heckman pour correction du biais de sélection dans le contexte des données longitudinales

Premièrement, la typologie des données manquantes sera présentée et les méthodes statistiques applicables en fonction du type de données manquantes. Ensuite sera présentée une des méthodes permettant la correction du biais de sélection dans le cas général des données manquantes non aléatoires : le modèle d'Heckman, avec la version standard dans le cas des données i.i.d. et la généralisation dans le cas des données longitudinales (mesures répétées). La méthode sera illustrée par un exemple sur données réelles.

—

14 h — 14 h 45 : **François Husson**, Department Statistics & Computer science — Agrocampus Ouest, Rennes.

Imputation multiple par des méthodes d'analyse factorielle

Les valeurs manquantes sont problématiques car la plupart des méthodes statistiques ne peuvent pas être directement appliquées à un ensemble de données incomplet. L'une des approches les plus populaires pour traiter les valeurs manquantes consiste à utiliser des méthodes d 'imputation simple. Ceci est fait en complétant les valeurs manquantes avec des valeurs plausibles qui conduisent à un jeu de données complet, qui peut être analysé par n'importe quelle méthode statistique.

Cependant, l'imputation simple ne permet pas de refléter l'incertitude de la valeur imputée et si une méthode statistique est utilisée, la variabilité des estimateurs est sous-estimée. C'est pour cette raison que l'imputation multiple (Rubin 1987 ; Little et Rubin 1987, 2002) est de plus en plus utilisée.

Dans cet exposé, nous verrons que de nouvelles méthodes basées sur des méthodes d'analyse factorielle permettent de faire de l'imputation multiple, quel que soit la nature des variables du jeu de données (quantitatives et/ou qualitatives). Une visualisation des tableaux imputés permet de plus d'appréhender l'incertitude autour des valeurs imputées, ce qui est un outil précieux pour décider ou non de poursuivre l'analyse avec n'importe quelle autre méthode statistique.

14 h 45 — 15 h 30 : **Jean Dumoncel**, UMR 5288, Laboratoire d'Anthropologie Moléculaire et Imagerie de Synthèse — Université de Toulouse.

Les contraintes en paléoanthropologie : comment établir des modèles statistiques ?

La paléoanthropologie se consacre à l'étude de la lignée humaine à travers les représentants fossiles de nos ancêtres. Dans ce cadre, la morphométrie est largement utilisée pour l'étude des formes anatomiques. Cependant, la paléoanthropologie souffre du nombre très réduit de spécimens, qui sont par ailleurs souvent fragmentaires. Afin de pallier ces contraintes, nous proposons des nouvelles méthodes d'analyse surfacique afin d'établir des modèles statistiques.

Durant cette présentation, nous nous intéresserons notamment à deux problématiques : comment réaliser des analyses sur des échantillons réduits et comment établir des modèles statistiques lorsque certains échantillons sont incomplets.

Dans la chaine de traitement des données tridimensionnelles employées en paléoanthropologie, la méthode la plus souvent utilisée est basée sur des points de repère (en général, anatomiques) dont les coordonnées sont analysées à l'aide d'outils mathématiques tels que la « morphométrie géométrique ». Plus récemment, une autre classe de méthodes a été proposée et permet des comparaisons globales entre les surfaces complètes de structures anatomiques sans avoir besoin de définir des points de repère. On obtient ainsi une analyse statistique de la forme moyenne et de sa variabilité en tout point. Nous utilisons des outils de déformations pour des recalages surfaciques basés sur des difféomorphismes. Les déformations obtenues sont ainsi utilisées dans les modèles statistiques pour proposer des analyses taxinomiques. Cette méthode permet également de répondre efficacement à des problématiques de données manquantes, comme par exemple pour estimer la moyenne à partir d'échantillons incomplets.

Nous verrons des exemples d'applications sur différentes structures anatomiques comme des dents et des endocrânes.

—

15 h 30 — 16 h 15 : **Dominique Soudant**, ODE/VIGIES — Ifremer, Nantes.

Application de modèles dynamiques bayésiens aux séries temporelles de surveillance de l'environnement marin

Les données disponibles à travers les réseaux d'observation et de surveillance de l'environnement marin se présentent sous la forme de séries temporelles. Elles ne sont généralement pas gaussiennes, pas stationnaires, ni en moyenne ni en variance. Cette non-stationnarité peut être le fait de phénomènes écologiques (e.g. changement globaux, pressions anthropiques) ou artéfactuels (e.g. changement de méthodes, d'agents, de laboratoire). Elles présentent des données manquantes, des données exceptionnelles voire des données fausses. L'ensemble de ces caractéristiques rendent ces séries temporelles particulièrement délicates à traiter. Cependant, les modèles linéaires dynamiques (i.e. Dynamic Linear Models, DLM) permettent de traiter les séries temporelles non-stationnaires comportant des données manquantes et peuvent prendre en compte les changements intervenant dans les séries via des « interventions » (West & Harrison, 1997). De ce fait, ils constituent une approche particulièrement bien adaptée à l'analyse des séries temporelles environnementales marines.