Étude de la répartition des mutations le long du génome à l'aide de chaînes de Markov cachées : exemple de la bactérie Borrelia sp

Sylvain Coly, Myriam Garrido, David Abrial

▶ **To cite this version:**

Sylvain Coly, Myriam Garrido, David Abrial. Étude de la répartition des mutations le long du génome à l'aide de chaînes de Markov cachées : exemple de la bactérie Borrelia sp. 45. Journées de Statistiques, May 2013, Toulouse, France. , pp.1-5, 2013, Actes des JdS 2013. hal-02745377

HAL Id: hal-02745377
https://hal.inrae.fr/hal-02745377v1

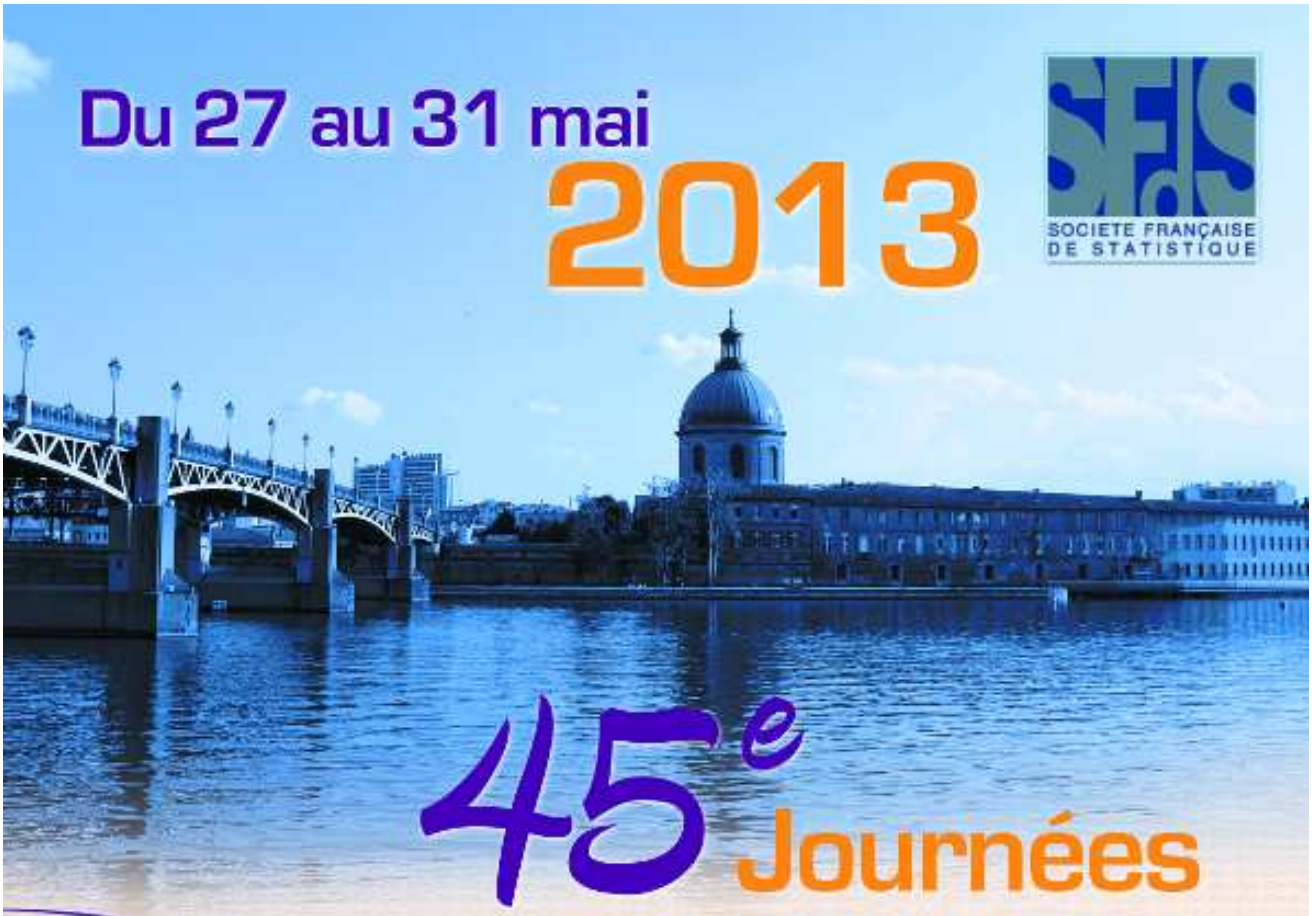
Submitted on 3 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Du 27 au 31 mai 2013

SOCIETE FRANÇAISE DE STATISTIQUE



45e Journées de Statistique de la SFdS

JdS 2013

ESC Toulouse,
20, bd Lascrosses

INRA
INP ENSEEIHT
INSA
ESC TOULOUSE
UNIVERSITÉ TOULOUSE 1 CAPITOLE
Université Paul Sabatier
cnrs
Toulouse School of Economics
Université de Toulouse
Inserm
UNIVERSITE DE TOULOUSE LE MIRAIL
INP ENSAT

Contact : jds2013@sfds.asso.fr

Site : http://jds2013.sfds.asso.fr

Comité de Programme

Président :

- Gilles Celeux (INRIA Saclay Île-de-France)

Membres :

- Liliane Bel (INRA, AgroParisTech)
- Christophe Biernacki (INRIA Lille, Université Lille 1)
- Delphine Blanke (Université d'Avignon)
- Laurent Bordes (Université de Pau et des pays de l'Adour)
- Guillaume Chauvet (ENSAI, Bruz)
- Florence Forbes (INRIA Rhône-Alpes)
- Fabrice Gamboa (Institut de Mathématiques de Toulouse)
- Stéphane Girard (INRIA Rhône-Alpes)
- Sylvie Huet (INRA Jouy-en-Josas)
- Jean-Michel Marin (Université Montpellier 2)
- Alberto Pasanisi (EDF)
- Christine Thomas-Agnan (Toulouse School of Economics)
- Anne VanHems (ESC Toulouse)

Comité d'organisation

Présidente :

- Christine Thomas-Agnan (Toulouse School of Economics, UT1C)

Membres :

- Philippe Berthet (Institut de Mathématiques de Toulouse, UPS)
- Philippe Besse (Institut de Mathématiques de Toulouse, INSA)
- Sandrine Casanova (Toulouse School of Economics, UT1C)
- Sébastien Déjean (Institut de Mathématiques de Toulouse, UPS)
- Robert Faivre (INRA de Toulouse)
- Nadine Galy (ESC Toulouse)
- Bernard Garel (Institut de Mathématiques de Toulouse, INP-ENSEEIHT)
- Christophe Genolini (INSERM U1027, Hôpital Purpan)
- Agnès Lagnoux (Institut de Mathématiques de Toulouse, UT2M)
- Thibault Laurent (Toulouse School of Economics, UT1C)
- Eve Leconte (Toulouse School of Economics, UT1C)
- Cathy Maugis-Rabusseau (Institut de Mathématiques de Toulouse, INSA)
- Elie Maza (INP-ENSAT)
- Christèle Robert-Granié (INRA de Toulouse)
- Anne Ruiz-Gazen (Toulouse School of Economics, UT1C)
- Rémi Servien (INRA de Toulouse)
- Anne Vanhems (ESC Toulouse)
- Nathalie Villa-Vialaneix (UPVD-IUT STID, en délégation à l'INRA de Toulouse)

Lundi 27 mai 2013

8h30-9h45	Accueil des participants et café					
9h45-10h15	Ouverture des Journées					
10h15-11h15	Conférence Le Cam : **Peter J. Bickel**					
11h20-12h20	Finance 1	Apprentissage 1	Statistique médicale 1	Régression 1	Grande dimension	Statistique non paramétrique 1
12h20-14h00	Déjeuner					
14h00-15h00	**Janine Illian**			**Sylvia Frühwirth-Schnatter**		
15h05-16h25	Statistique spatiale 1	Statistique mathématique 1	Sélection de modèles	Études de cas 1	Données de survie 1	Segmentation
16h25-16h45	Pause café					
16h45-17h45	Graphes	Modèles de mélange	Données fonctionnelles 1	Études de cas 2	Analyse de données 1	Enseignement 1
19h00	Réception à la mairie de Toulouse					

Mardi 28 mai 2013

8h30-9h30	Prix du Docteur Norbert Marx : **Guy S. Mahiane**					
9h35-10h35	Données fonctionnelles 2	Statistique mathématique 2	Analyse de données 2	Statistique médicale 2	Ingénierie	
10h35-11h	Pause café					
11h-12h20	Modèles semi-paramétriques	Statistique spatiale 2	Statistique bayésienne	Séries temporelles	Analyse de données 3	Extrêmes 1
12h20-14h00	Déjeuner					
14h00-15h00	**Christian Robert**			**Juan Cuesta-Albertos**		
15h05-16h05	Etude de cas 3	Trafic routier	Apprentissage 2	Statistique médicale 3	Régression 2	
16h05-16h25	Pause café					
16h25-17h25	Statistique et sciences humaines	Etude de cas 4	Apprentissage 3	Génétique/ Génomique 1	ENBIS	
17h30-18h30	Assemblée générale de la SFdS					

Mercredi 29 mai 2013

8h30-9h35	**Christian Francq**			**Johan Segers**		
9h35-10h35	Finance 2	Statistique d'enquête 1	Enseignement 2	Processus 1	Statistique non-paramétrique 2	Analyse de données 4
10h35-11h	Pause café					
11h-12h00	Fiabilité et incertitudes 1	Processus 2	Enseignement 3	Classification non supervisée 1	Histoire	Société Française de Biométrie
12h00	Pique-Nique / Programme social et culturel					
19h	Repas de Gala					

Jeudi 30 mai 2013

9h00-10h00	Prix Pierre Simon de Laplace : **Christian Gouriéroux**					
10h00-10h15	Pause café					
10h15-11h15	**Michael Goldstein**			**David Hunter**		
11h20-12h40	Extrêmes 2	Processus ponctuels spaciaux	Computer experiments	Statistique non-paramétrique 3	Génétique/ Génomique 2	Finance/ Econométrie
12h40-14h10	Déjeuner					
14h10-15h10	**Christophe Ambroise**			**Jan Johannes**		
15h15-16h35	Statistique mathématique 3	STID/ Enseignement	Apprentissage 4	Sélection de variables	Plan d'expériences et quantification optimale (15h15-15h55) Image (15h55-16h35)	
16h35-16h55	Pause café					
16h55-17h55	Table ronde enseignement	Fiabilité et incertitudes 2	Statistique d'enquête 2	Climat 1	Classification non supervisée 2	
18h00	Rencontre Jeunes Statisticiens / Conférenciers invités					
20h00	Café de la Statistique					

Vendredi 31 mai 2013

8h30-9h30	**Anthony Davison**			**Mark van de Wiel**		
9h30-10h30	**Petra Friederichs**			**David Haziza**		
10h30-10h50	Pause café					
10h50-12h10	Statistique d'enquête 3	Données de survie 2	Climat 2	Statistique mathématique 4	Enseignement 3	Biopharmacie
12h15 -12h30	Clôture des journées					
12h30 -14h00	Déjeuner					

Table des matières

Résumés du lundi 27 mai 2013

Conférence Le Cam - Peter J. Bickel, 10h15-11h15.

Topics in inference for unlabelled graphs

Peter Bickel (Berkeley University)

Network data are becoming common in many areas beyond the social sciences where their study was focused. In B. and Chen PNAS (2009), we introduced a "nonparametric" model for unlabeled graphs with average degrees ranging from $\ll \log(n)$ to $O(n)$ and studied it asymptotically in relation to "block model" introduced by Holland and Leinhardt (1983) which can be viewed as "histogram" approximations and their relation to fitting methods called modularities arising in the physics literature. In B. Chen and Levina (2012), we studied quantities variously called motifs, "moments" which can be used for goodness of fit testing and in principle for fitting block and other parametric moments in the same asymptotic range. Maximum likelihood and variational likelihood fitting have been studied by Celisse and Daudin (2012) and spectral methods by Rohe et al (2011) and Priebe et al (2013), among others. All of these methods except for spectral and variational are NP complete. We will give an overview of work in the asymptotic regime mentioned focusing on new results on the equivalence of maximum, variational, and profile likelihood for block models. These results are partly based on work of Le Cam and Yang (1988). We will also touch on properties of spectral methods, the most successful fitting methods.

Finance 1, 11h20-12h20.

Logit emboité vs logit multinomial. Analyse empirique des stratégies de gestion des risques

Hassen Raïs (Université Toulouse 1)

Cette communication cherche à expliquer les choix des stratégies de gestion des risques par les entreprises non financières. Ces choix peuvent être modélisés par un logit multinomial ou un logit emboitée. Le deuxième modèle a pour avantage de relâcher l'hypothèse d'indépendance des alternatives non pertinentes liée au premier modèle. A travers les données d'une enquête empirique les deux modèles sont développés et leurs résultats comparés. Cette analyse montre que la supériorité du modèle logit emboité est relative et l'évaluation des effets marginaux des variables explicatives sur les choix de stratégies issue des deux modèles "concurrents" montre que le choix de l'un ou de l'autre peut modifier les conclusions de cette analyse.

High-growth enterprises in Portugal : evidence from micro data

Homero Gonçalves (Banco de Portugal), Vítor Silveira (Banco de Portugal)

Portuguese economy has been facing a severe recession over the last few years during which unemployment has grown to record levels. Restoring growth is then a major goal for policy makers nowadays and to that end, high-growth enterprises can play a very important role. Indeed, besides being usually associated with innovation, their extraordinary growth makes the largest contribution to net employment creation, despite typically representing a small proportion of the business population. Better familiarity with these firms would allow policy makers to develop appropriate approaches to maximize the chances of potential high-growth firms to develop. Using micro data derived from two administrative databases managed by the Banco de Portugal we highlight some high-growth enterprises' characteristics and show how the economic and financial crisis has affected these companies. Additionally making use of individual balance-sheet data we compare their performance with that of the remaining Portuguese non-financial corporations. Finally, we examine their relationship with the financial system, namely how their access to credit compares with the remaining enterprises. Some of the main findings reveal that the share of high-growth enterprises decreased with the recent economic and financial crisis. Nonetheless, some activities are showing more resilience to the negative environment, namely NACE 72 - Scientific research and development which, for its close link to innovation, is a positive prospect for the future. High-growth enterprises also present better economic and financial performance and that seems to be recognized by the banking sector reflecting in better access to credit.

Mesures du kurtosis en économie et en finance

Hélène Honoré (Eurofidai - CNRS)

L'intérêt de la littérature économique pour les moments d'ordres supérieurs naît des limites de la variance à modéliser toutes les facettes du risque. Ainsi, dans le modèle d'espérance d'utilité, à l'aversion pour le risque et à la prudence, certains auteurs ajoutent la tempérance qui refléterait une aversion pour les queues épaisses. Dans le but de couvrir un large spectre de comportements des investisseurs face au risque, la variance et les moments d'ordres supérieurs sont simultanément pris en compte dans la stratégie des hedge funds pour lesquels les investisseurs averses au risque préfèrent également un faible kurtosis. Le kurtosis correspond à un double mouvement : des épaules vers le centre d'une part et vers les queues d'autre part. Dans ces travaux économiques, le kurtosis est très fréquemment formalisé par le quatrième moment central standardisé. Or, il existe une large gamme de mesures du kurtosis alternatives qui subissent moins fortement l'influence des valeurs extrêmes et dont l'interprétation peut ainsi en être facilitée. Cet article a pour objectif, dans un premier temps, de comparer, d'un point de vue statistique, ces différentes mesures du kurtosis et, dans un second temps, de décrire l'effet que le choix de se porter sur telle ou telle mesure peut avoir sur l'interprétation économique d'une application empirique.

Apprentissage 1, 11h20-12h20.

Cobra : agrégation non-linéaire de prédicteurs

Benjamin Guedj (Laboratoire de Statistique Théorique et Appliquée), Gérard Biau (Laboratoire de Statistique Théorique et Appliquée et Institut Universitaire de France), Aurélie Fischer (Laboratoire de Probabilités et Modèles Aléatoires), James Malley (National Institutes of Health)

Nous présentons une nouvelle méthode d'agrégation non-linéaire d'estimateurs initiaux $r_1, \ldots, r_M$ de la fonction de régression. L'estimateur résultant, combiné par un procédé de moyenne locale utilisant les estimateurs initiaux comme indicateurs de distance, s'avère asymptotiquement plus performant que le meilleur des $r_1, \ldots, r_M$. Nous présentons le package R COBRA implémentant notre méthode, et ses excellentes performances sur données réelles et simulées dans le contexte de la prédiction. COBRA est disponible sur http ://cran.r-project.org/web/packages/COBRA/index.html. Pré-publication : http ://arxiv.org/abs/1303.2236

Convergence des forêts aléatoires sous-échantillonnées

Erwan Scornet (Université Paris 6)

Les forêts aléatoires, proposées par L. Breiman (2001), comptent parmi les méthodes les plus utilisées dans les problèmes d'estimation en grande dimension bien que certaines de leurs propriétés demeurent incomprises d'un point de vue théorique. Nous présentons dans cet exposé un cadre général pour l'étude de ces méthodes d'estimation de la régression. Nous montrerons en particulier qu'elles peuvent être envisagées sous la forme d'estimateurs à moyenne locale que l'on détaillera dans deux cas particuliers, les forêts à découpes centrées d'une part et les forêts à découpes uniformes d'autre part. Par ailleurs, nous montrerons qu'un sous-échantillonnage adéquat permet d'assurer la convergence des forêts complètement développées (qui sont proches de l'algorithme original envisagé par L. Breiman).

Résultats sur les algorithmes de L^2-boosting pour les régressions parcimonieuses

Magali Champion (IMT, Univ. Paul Sabatier), Christine Cierco-Ayrolles (INRA Toulouse), Matthieu Vignes (INRA Toulouse), Sébastien Gadat (IMT, Univ. Paul Sabatier)

Nous nous intéressons à l'estimation des paramètres d'une régression multi-tâches parcimonieuse par algorithmes de Boosting. Le modèle considéré est le suivant : soit $X_1, ..., X_p$ p variables explicatives, on suppose que la réponse Y observée de taille m dépend linéairement de ces variables suivant l'équation $Y = XA + e$, où A désigne l'ensemble des coefficients de régression inconnus et e est une variable aléatoire décrivant la présence de bruit dans le modèle. L'objectif de cette étude consiste évidemment à retrouver la structure (éléments non nuls) de A dans le cadre de la grande dimension. Pour cela, nous proposons d'étudier les algorithmes de L^2-Boosting qui permettent de construire une approximation globale d'une fonction f donnée, en l'occurrence, $f(X) = XA$, à l'aide d'approximations locales. Après une étude détaillée des Weak Greedy Algorithms, versions déterministes de ces algorithmes, nous présentons un résultat déjà connu de consistance dans le cas univarié. Nous proposons également un résultat nouveau concernant la stabilité du support sous réserve que les coefficients non nuls de A soient suffisamment grands. Un des enjeux de ce travail consiste ensuite à étendre ces résultats dans le cas où les observations

sont multiples. Pour cela, nous proposons deux adaptations de l'algorithme précédent, qui vérifient aussi des résultats de stabilité, sous des hypothèses concernant notamment la parcimonie du modèle.

Statistique médicale 1, 11h20-12h20.

Modélisation conjointe de données longitudinales et de temps d'événement avec application à la prédiction de rechute de cancer de la prostate

Mbéry Séne (INSERM U897, ISPED,)

Dans la dernière décennie, la modélisation conjointe s'est rapidement développée dans le domaine de la recherche médicale. Ils permettent d'étudier un marqueur longitudinal et un temps d'événement corrélés. Parmi eux, les modèles à effets aléatoires partagés ont reçu plus d'attention. Ces modèles étendent naturellement le modèle de survie avec variables explicatives dépendantes du temps et offrent un cadre flexible pour explorer le lien entre le biomarqueur longitudinal et le risque d'événement. L'objectif de ce travail est d'évaluer à travers un exemple réel d'étude de progression de cancer de la prostate après une radiothérapie. En particulier, différentes spécifications de la dépendance entre le biomarqueur longitudinal, l'antigène spécifique de la prostate, et le risque de rechute clinique sont investiguées pour bien comprendre le lien entre ces deux processus. Ces différents modèles conjoints sont comparés en termes de qualité d'ajustement et d'adéquation aux hypothèses du modèle conjoint mais aussi en termes de pouvoir prédictif en utilisant la cross-entropie pronostique. En effet, en plus de mieux comprendre le lien entre la dynamique de PSA et le risque de rechute clinique, la perspective dans les études sur le cancer de la prostate est de fournir des outils pronostiques dynamiques de rechute clinique basés sur l'historique du biomarqueur.

Analyse des critères dérivés de données longitudinales intensives : application à une évaluation quotidienne de la douleur

Pierre Bunouf (Pierre Fabre Biométrie), Jean-Marie Grouin (Université de Rouen)

Cette communication propose une stratégie d'analyse des données longitudinales dites intensives, c'est à dire mesurées fréquemment, après réduction de l'information en critères d'évaluation. Une telle analyse nécessite des règles rigoureuses pour dériver les critères et une gestion adaptée des valeurs manquantes. Notre cas d'étude est un essai clinique placebo-contrôlé pour évaluer l'efficacité d'un traitement dans la douleur. Des visites mensuelles chez l'investigateur étaient planifiées mais l'intensité de la douleur était rapportée quotidiennement sur la durée de l'essai par les sujets via des agendas électroniques. Une fois le mécanisme des valeurs manquantes élucidé, les méthodes statistiques pour analyser les critères dérivés aux visites sont basées sur une modélisation directe ou bien sur une imputation multiple. En outre, certaines de ces méthodes peuvent utiliser l'information provenant des critères dérivés comme celle provenant des données brutes. Nous montrons que les méthodes fondées sur l' imputation multiple sont particulièrement bien adaptées à notre contexte et que l'imputation des valeurs manquantes au niveau des données brutes permet de tirer le meilleur parti de l'information disponible. Nous montrons également comment des estimations individuelles par sujet permettent de caractériser de manière simple l'influence des profils incomplets dans l'estimation globale.

Modèle mixte à fonction spline pénalisée pour prédire la maladie d'Alzheimer

Abderazzak Mouiha (I.U.S.M.Q.), Simon Duchesne (IUSMQ et Dép. Radiologie Université Laval)

Les biomarqueurs longitudinaux provenant de la neuroimagerie, en particulier le volume hippocampique (HC) et le métabolisme cérébral mesuré par tomographie par émission de positon et marqueur Fluoro-Desoxy-Glucose (FDG-PET), ont été identifiés comme ayant une forte valeur prédictive pour le diagnostic et le prognostic de la maladie d'Alzheimer (AD) (Doody, R. et al.). Dans cette communication, nous avons étudié l'évolution du volume hippocampique et du FDG-PET en fonction des tests neuropsychologiques (Mini-Mental-State-Examination (MMSE) et Alzheimer's Disease Assessment Scale-Cognitive (ADAS-Cog)). En se basant sur les recherches de C. Jack et al., A. Caroli et al. et A. Mouiha et al., nous proposons un modèle mixte à fonction spline pénalisée qui ajuste mieux les données longitudinales sur la maladie d'Alzheimer sélectionnées à partir de l'étude ADNI (Alzheimer's Disease Neuroimaging Initiative http ://adni.loni.ucla.edu) . Les résultats montrent qu'il existe, d'une part, une relation non linéaire entre la progression de la maladie et le diagnostic, et d'autre part la variation des biomarqueurs en fonction des tests cognitifs.

Régression 1, 11h20-12h20.

La régression logistique fonctionnelle avec dérivée

Aziza Ahmedou (Laboratoire Angevin de REcherche en MAthématiques), Jean-Marie Marion (Institut de Mathématiques Appliquées - UCO), Besnik Pumo (Université d'Angers)

Cet article concerne l'étude d'un modèle de régression logistique avec une variable explicative fonctionnelle et sa dérivée. Dans le cadre de l'estimation des paramètres, nous proposons une estimation des paramètres fonctionnels à partir de la décomposition dans deux bases Splines qui permettent de transformer notre modèle en un modèle linéaire généralisé (GLM) multivarié et d'estimer les paramètres par la méthode du maximum de vraisemblance en construisant un algorithme afin d'estimer les paramètres du modèle. Nous montrons la convergence des estimateurs Splines du maximum de vraisemblance et illustrons notre modèle sur un exemple.

Régression linéaire généralisée sur composantes supervisées

Xavier Bry (Université Montpellier 2), Catherine Trottier (Université Montpellier 3), Thomas Verron (ITG-SEITA), Frédéric Mortier (CIRAD)

Dans l'estimation courante d'un GLM, la structure de corrélation des régresseurs n'est pas utilisée pour trouver des structures prédictives fortes. La recherche de combinaisons linéaires des régresseurs qui maximisent simplement la vraisemblance du GLM a deux conséquences majeures : 1) la colinéarité des régresseurs est un facteur d'instabilité de l'estimation, et 2) le modèle pouvant s'ajuster à des dimensions de bruit, ses pouvoirs explicatif et prédictif sont fragilisés. Des adaptations de la régression PLS au GLM ont été proposées pour le cas d'un modèle univarié. Nous proposons ici pour le cas multivarié une méthode, la Régression Linéaire Généralisée sur Composantes Supervisées (Supervised Component Generalized Linear Regression, SCGLR). Elle étend l'algorithme des scores de Fisher de sorte à combiner la régression PLS avec l'estimation du GLM. SCGLR est testée sur des données simulées, puis des données réelles.

Modèle logistique ordinal à variable latente : logits cumulés ou adjacents ?

Petan Dossar (Université Paris 6), Ndiogou Seck (Université Paris 6), Mounir Mesbah (Université Paris 6)

Dans le choix d'un modèle logistique ordinal, deux principaux types de lien sont en concurrence, les "logits cumulés" et les "logits adjacents". MacCullagh (1980) et Anderson (1984) avaient abordé la question de ce choix, avec des points de vue différents. Dans le cas où la réponse ordinale est multivariée, et en particulier, dans le cas des mesures répétées d'un même item, où les réponses de chaque individu aux différents items peuvent être naturellement considérés comme fonction d'une même variable latente unidimensionnelle, le modèle à réponses graduées (GRM) de Samejima (1969) est construit à partir des "logits cumulés", alors que le modèle à crédit partiel (PCM) de Masters, (1982) l'est à partir des "logits adjacents". Dans ce travail, nous comparons les deux approches, surtout dans le cas latent, où le PCM s'avère supérieur au GRM, grâce à deux propriétés intéressantes que nous démontrons : l'exhaustivité de la somme des réponses aux items notée S_i pour le paramètre individuel θ_i et une propriété de "stochastic ordering" des distributions du trait latent, conditionnellement à S. La deuxième propriété peu connue, et est, à notre connaissance, nulle part démontrée de manière satisfaisante. La portée pratique de ces résultats est illustrée à l'aide de données réelles et de simulations intensives que nous présenterons.

Grande dimension, 11h20-12h20.

Comparaison de méthodes basées sur SIR pour des cas sous-déterminés ($n < p$)

Raphaël Coudret (Université de Bordeaux et INRIA), Benoit Liquet (Université de Bordeaux et University of Cambridge), Jérôme Saracco (IMB, INRIA)

Parmi les méthodes pour analyser des données de grande dimension, la régression inverse par tranches (sliced inverse regression ou SIR en anglais) est particulièrement intéressante si des relations non-linéaires existent entre la variable à expliquer et des combinaisons linéaires des prédicteurs. Lorsque la dimension de ces prédicteurs est plus grande que le nombre d'observations, les versions classiques de SIR ne peuvent plus être utilisées. Des améliorations diverses comme RSIR et SR-SIR ont été proposées pour résoudre ce problème et estimer les paramètres du modèle sous-jacent. Dans cette présentation, nous introduisons une nouvelle procédure d'estimation qui utilise l'algorithme QZ (SIR-QZ). Nous en présentons également une autre basée sur l'inverse généralisé de Moore-Penrose (SIR-MP). Ces approches sont ensuite comparées avec RSIR et SR-SIR par le biais de simulations. Enfin, nous illustrons, sur un jeu de données génétiques, l'intérêt de l'approche SIR-QZ proposée pour trouver des eQTL.

Moment generating function of linear spectral statistics in a spiked population model

Damien Passemier (Hong Kong University of Science and Technology (HKUST)), Matthew R. McKay (Hong Kong University of Science and Technology (HKUST)), Yang Chen (University of Macau)

Dans cet exposé, nous considérons des statistiques spectrales linéaires (LSS) des valeurs propres d'une matrice de Wishart complexe qui possède une matrice de population avec une variance isolée. Nous établissons une relation entre la fonction génératrice des moments de cette LSS et son analogue dans le cas blanc, en utilisant des polynômes orthogonaux. Nous utilisons ensuite des résultats issus d'une

interprétation des fluides de Coulomb afin de donner des caractérisations des quantités qui apparaissent dans cette expression. Nous appliquons enfin cette théorie générale à un exemple concret : celui de la capacité de coupure (outage capacity) dans un système de communication multi-antennes.

Minimax adaptive dimension reduction for regression

Quentin Paris (ENS Cachan, Bretagne)

Soit (X, Y) une variable aléatoire à valeurs dans $\mathcal{X} \times \mathbb{R}$, où $\mathcal{X} \subset \mathbb{R}^p$. Nous étudions l'estimation de la fonction de regression $r(x) := \mathbb{E}(Y|X = x)$ d'un point de vue minimax, en supposant que r appartient à une classe de fonctions $\mathcal{F}$ de la forme

$$\mathcal{G} \circ \mathcal{H} := \Big\{ g \circ h : g \in \mathcal{G}, h \in \mathcal{H} \Big\}.$$

Ici, $\mathcal{G}$ est une classe de fonctions β-Hölderiennes $g : \mathbb{R}^p \to \mathbb{R}$ et $\mathcal{H}$ une classe de fonctions $h : \mathcal{X} \to \mathbb{R}^p$ dont on ne suppose ni qu'elles sont linéaires ni qu'elles sont régulières. Nous définissons la *dimension réduite* comme la plus petite valeur de ℓ pour laquelle $r = g \circ h$ pour $g \in \mathcal{G}$ et $h \in \mathcal{H}$ o'u $h(\mathcal{X})$ engendre un espace de dimension ℓ. Ensuite, nous construisons un estimateur adaptatif de r convergeant à une vitesse ne dépendent que de la dimension réduite sous des conditions d'entropie de la classe $\mathcal{H}$.

Statistique non paramétrique 1, 11h20-12h20.

Estimation non paramétrique des composantes d'un modèle de mélange par une approche clustering

Laurent Rouvière (Ensai), Stéphane Auray (Ensai), Nicolas Klutchnikoff (Ensai)

Nous étudions les performances d'estimateurs non paramétriques de la densité combinés avec une phase de clustering pour estimer les composantes d'un modèle de mélange. On considère un n-échantillon i.i.d. $(Y_1, X_1), \ldots, (Y_n, X_n)$ à valeurs dans $R \times R^d$ à partir duquel on cherche à estimer la densité de Y qui se met sous la forme $f(y) = \sum_{i=1}^{M} \alpha_i f_i(y)$. L'approche proposée consiste à effectuer une étape de clustering sur l'échantillon $X_1, \ldots, X_n$ afin de prédire le groupe de chaque observation Y_k. Les densités f_j sont ensuite estimées à l'aide d'un estimateur à noyau qui utilise les prévisions de l'étape préliminaire de clustering. Nous étudions le risque L_1 de l'estimateur proposé et nous prouvons que, sous certaines hypothèses concernant la performance de la procédure de clustering utilisée, cet estimateur atteint les vitesses de convergence optimales sur les classes de densités classiques telles que les classes de Hölder. Nous proposons enfin des méthodes de clustering qui vérifient les hypothèses requises.

Estimation du support de la densité et de son contour a l'aide de polyèdres

Catherine Aaron (Université Blaise Pascal)

Soit une densité (inconnue) f à support S compact dans $\mathbb{R}^d$ et $\mathcal{X}_n$ un $n-$échantillon i.i.d. issu de f. On s'interresse à l'estimation de S et de sa frontière ∂S en utilisant des polyèdres (i.e. des unions de simplexes). On présentera plusieurs estimateurs du support ainsi que certaines de leur propriétés. En particulier, on introduira un premier estimateur, union des simplexes de Delaunay conditionnés à un critère de plus proches voisins dont on donnera certaines propriétés (convergence à une vitesse "usuelle" sous

optimale mais avec critères objectifs permettant d'espérer que l'estimation du support est homéomorphe au support). On présentera également d'autres estimateurs, unions de "petits" simplexes avec un critère de restriction "à rayon fixé". Pour ces derniers estimateurs on montrera que la vitesse de convergence est très proche de la vitesse optimale. Pour ces estimateurs aussi quelques indices nous laissent penser qu'on a un homéomorphisme entre les estimateurs et le support.

Sélection de modèles pour l'estimation de la densité relative

Gaëlle Chagny (Laboratoire MAP5, Université Paris Descartes)

L'objectif de ce travail est de proposer un estimateur adaptatif pour une fonction récemment utilisée dans les problèmes à deux échantillons, la densité relative. Cette fonction, outil pour la comparaison des distributions de deux variables X et X_0, est définie comme la densité de $F_0(X)$, où F_0 est la fonction de répartition de X_0. La technique d'estimation choisie s'inspire de la sélection de modèles : un estimateur est sélectionné automatiquement sur la bases des observations par un critère adapté des travaux de Goldenshluger et Lepski (2011), à partir d'une collection d'estimateurs par projection. Le compromis biais-variance est réalisé, et une borne non asymptotique pour le risque quadratique intégré établie. Des vitesses de convergence sont également démontrées. La méthode est illustrée par des simulations.

Janine Illian, 14h00-15h00.

Spatial statistics and the real world – the old, the new and the challenging

Janine Illian (University of St Andrews)

The spatial pattern formed by the individuals in an ecological community reveals the individuals' local interactions in the presence of associations with specific environmental conditions and informs on community dynamics. Hence, point process methodology may be used to help provide answers to concrete ecological questions. However, data sets obtained in relevant studies often suffer from a number of limitations specific to ecological data. A major issue here is that the observation process is complex and non-uniform in space. This is because data sets are often derived from opportunistic sampling regimes or wildlife surveys suffering from incomplete detection. In other cases several data sources exist that aim to describe the same phenomenon but each of these have been observed with different types of error.

Over the years, most of my research has involved interdisciplinary work applying spatial statistics to problems in ecology trying to resolve some of the challenges. This talk will review some of the contributions that this interdisciplinary work has made to ecology, primarily in the context of biodiversity. I will also discuss a number of more recent challenges I have come across ? ?" including several interesting applications from outside ecology, such as geolinguistics and crime modelling.

Sylvia Frühwirth-Schnatter, 14h00-15h00.

Flexible modelling based on sparse finite mixtures

Sylvia Frühwirth-Schnatter (Vienna University of Economics)

There has been a tremendous increase in applied work both in statistics as well as econometric using infinite mixtures based in particular on Dirichlet process priors. However, going to infinity makes the mathematics much more complicated, in particular, for applied statisticians and econometricians.

This talk investigates the concept of sparse finite mixture modelling, which is based on a shrinkage prior on the weights that removes all redundant components automatically. The choice of the hyperparameters of this prior is based on recent asymptotic results by Rousseau and Mengersen (2011). The sparsity prior provides an automatic tool to select the number K of components and avoids the cumbersome computation of the marginal likelihood for each K. Furthermore, it is shown how the label switching problem could be solved using the framework of sparse finite mixtures. In contrast to infinite mixtures, this allow identification of component-specific parameters and classification. This approach is applied to various issues in statistical modelling such as choosing choosing the number of latent classes in a latent class model, model-based clustering based on finite mixtures of normals, switching regression models, and choosing a flexible link function in binary data modelling. Finally, the sparse finite mixture approach is compared to infinite mixtures based on Dirichlet process priors.

Statistique spatiale 1, 15h05-16h25.

Une statistique non-paramétrique de détection d'agrégats spatiaux

Lionel Cucala (Université Montpellier 2)

Nous proposons une nouvelle statistique pour identifier les agrégats de fortes ou de faibles valeurs dans des données spatiales. Contrairement aux statistiques de détection d'agrégats classiques, celle-ci n'est pas basée sur un rapport de vraisemblance et ne nécessite donc pas de choisir de distribution spécifique. Cette statistique semble puissante contre tout type d'agrégats, quelle que soit la distribution sous-jacente. Nous illustrons cette méthode en analysant spatialement les revenus des foyers fiscaux en France.

Accuracy of areal interpolation methods

Van Huyen Do (GREMAQ, TSE, Toulouse), Christine Thomas-Agnan (GREMAQ, TSE), Anne Vanhems (ESC Toulouse)

L'analyse de données socio-économiques nécessite souvent de combiner des bases de données provenant de différentes sources administratives, données collectées sur plusieurs partitions différentes de la zone d'intérêt. Il est donc nécessaire de transformer les données provenant d'unités spatiales d'origine ('sources') en données associées aux unités spatiales 'cibles'. Par exemple, on peut s'intéresser à un quadrillage commun régulier d'une zone d'intérêt et souhaiter adapter toutes les informations initiales à cette nouvelle partition cible unique. Cette option est actuellement à l'étude en France à l'INSEE et en Europe avec la directive de l'UE 'INSPIRE' (INfrastructure for SPatial InfoRmation). Il existe trois méthodes principales : la méthode des poids proportionnels, les techniques de lissage, et l'interpolation basée sur des méthodes de régression. A l'aide d'un modèle basé sur des processus de Poisson ponctuels et dans le cas d'une

partition cible régulière, nous étudions et comparons la précision de ces différentes méthodes, en nous intéressant plus particulièrement aux méthodes de poids proportionnels et aux méthodes de régression. L'erreur statistique dépend alors de la nature de la variable d'intérêt, et de sa correlation avec des variables auxiliaires. Nous montrons qu'il n'y a pas de méthode unique toujours plus performante que les autres.

Segmentation d'images hyperspectrales à partir d'estimation à noyau fonctionnel de la densité

Laurent Delsol (MAPMO, Université d'Orléans), Cécile Louchet (MAPMO, Université d'Orléans)

Décomposer une image en un ensemble de régions homogènes est un problème classique, appelé segmentation, en traitement d'image. La détection de telles régions est habituellement une manière pertinente d'identifier des éléments spécifiques de la scène. De nombreuses méthodes ont été proposées pour segmenter des images en niveaux de gris ou multispectrales. L'approche du maximum a posteriori, utilisant un champ de Potts comme a priori et une estimation de la densité sur chaque région, constitue un exemple intéressant d'utilisation de la statistique bayésienne dans ce domaine. D'autre part, de nombreuses méthodes de statistique fonctionnelle sont maintenant proposées pour permettre l'étude de données correspondant à des courbes. L'estimateur à noyau de la densité a notamment été adapté à de telles données. Nous considérons dans cet exposé des images hyperspectrales pour lesquelles chaque pixel est décrit au travers d'une courbe (discrétisée en un grand nombre de points) et discutons la manière dont l'estimation à noyau fonctionnel de la densité et l'approche par maximum a posteriori peuvent être combinées.

Deux algorithmes pour la classification non supervisée de données géostatistiques

Thomas Romary (Mines ParisTech)

Avec le développement des plateformes de télédétection, et l'évolution des moyens d'échantillonnage des compagnies minières ou pétrolières, les jeux de données spatiales deviennent de plus en plus grands. Il devient souvent nécessaire de séparer le domaine d'étude en différentes zones homogènes afin de simplifier l'étape de modélisation. La définition de ces zones peut se voir comme un problème de classification où l'on cherche à découper le domaine d'étude en zones homogènes. L'application des méthodes de classification pour des observations indépendantes ne permet généralement pas de conserver une cohérence spatiale dans les classes. Les algorithmes de segmentation d'image basés sur des champs de Markov, ne sont quant à eux pas applicables lorsque le plan d'échantillonnage n'est pas régulier. Les approches existantes, basées sur une estimation de mélange de fonctions aléatoires gaussiennes par l'algorithme EM, sont limitées à des tailles d'échantillon raisonnables et pour un faible nombre de variables. Nous proposons dans ce travail deux algorithmes basés sur des adaptations d'algorithmes classiques, qui permettent de traiter un large volume de données. Le premier procède par classification ascendante hiérarchique tandis que le second est basé sur la méthode de classification spectrale. Les deux algorithmes sont appliqués à des jeux de données synthétiques et à un jeu de données minières.

Statistique mathématique 1, 15h05-16h25.

Vraisemblance empirique pour l'estimation par substitution de paramètres fonctionnels

Davit Varron (Université de Franche Comté), Alexis Flesch (Université Technologique de Belfort Montbelliard)

Cet exposé traite de l'estimation par substitution basée sur un échantillon i.i.d. de loi P0. Il est possible de voir la méthode de vraisemblance empirique de la façon suivante pour construire une région de confiance autour de l'estimateur suivant de $\theta_0 := T(P0) : \theta_n := T(P_n)$, où P_n est la mesure empirique, en faisant varier continument les poids de cette mesure empirique dans un 'voisinage' de poids uniformes. La valeur de l'estimateur décrit ainsi une région, qui semble intéressante, pour peu que la notion de 'voisinage' soit définie de façon pertinente. La vraisemblance empirique classique choisit des voisinages associés à la divergence de Kullback. Lorsque T est linéaire et de dimension finie, alors θ_0 est l'espérance de variables observées, dans ce cas les propriétés asymptotiques de cette méthode sont bien connues [1]. Bertail [2] a montré que ces propriétés étaient conservées lorsque T est nonlinéaire, mais satisfait des conditions de différentiabilité dans le contexte de processus empiriques. Cependant, un tel résultat théorique reste insatisfaisant en pratique, pour des problèmes de temps de calculs. Nous proposons une linéarisation qui contourne ce problème de temps de calcul, et dont nous montrons la validité asymptotique. Par ailleurs, nous étendons ce résultat aux statistiques basées sur plusieurs échantillons indépendants, et aux paramètres θ_0 de dimension infinie.

Estimation de la densité et de la fonction derépartition par itérations de l'opérateur Debernstein

Claude Manté (CNRS/ AMU)

On propose une méthode originale d'estimation de la fonction de répartition et de la densité avec des polynômes de Bernstein. On y met à profit la décomposition spectrale de l'opérateur de Bernstein pour raffiner un algorithme d'accélération de convergence. On obtient ainsi une méthode dont les performances sont meilleures que celles de l'estimateur de Bernstein.

Minimiser le risque empirique pour des pertes à queue lourde

Emilien Joly (ENS), Gábor Lugosi (Pompeu Fabra University), Christian Brownless (Pompeu Fabra University)

Le sujet de ce papier est une discussion autour de la minimisation du risque empirique pour des pertes non nécessairement bornées et possiblement à queues lourdes. Dans ces situations, les moyennes empiriques classiques peuvent misérablement échouer. Pourtant, un estimateur de la moyenne robuste, proposé dans la littérature, peut être utilisé pour remplacer la moyenne empirique usuelle. Dans ce papier, nous étudions une minimisation du risque empirique reposant sur un estimateur robuste proposé par O. Catoni. Nous développons des bornes de performance basées sur un argument de chainage et une formule de Taylor appliquée à l'estimateur de la moyenne de Catoni.

Goodness-of-fit test in semiparametric transformation models

Benjamin Colling (Université catholique de Louvain), Ingrid Van Keilegom (UCL)

$$\Lambda_{\theta_0}(Y) = m(X) + \epsilon \quad ,$$

où Y est une variable dépendante univariée, X est une covariable de dimension 1 et ϵ est un terme d'erreur indépendant de X et de moyenne 0. On suppose que $\{\Lambda_\theta : \theta \in \Theta\}$ est une famille paramétrique de fonctions strictement croissantes, alors que m est une fonction inconnue. On utilisera l'estimateur profile likelihood pour le paramètre θ_0 proposé par Linton, Sperlich et Van Keilegom (2008). On veut tester l'hypothèse

$$H_0 : m \in \mathcal{M} \quad H_1 : m \notin \mathcal{M} \quad ,$$

où $\mathcal{M} = \{m_\beta : \beta \in \mathcal{B}\}$ est une certaine classe de fonctions de régression et $\mathcal{B} \subset \mathbb{R}^p$. On utilisera une statistique de type Kolmogorv-Smirnov et une statistique de type Cramer-von Mises dont l'idée est de comparer la fonction de distribution de ϵ estimée de manière nonparamétrique $\widehat{F}_\epsilon(y)$ à la fonction de distribution de ϵ estimée sous l'hypothèse nulle $\widehat{F}_{\epsilon_0}(y)$. On étudie la convergence du processus $n^{1/2}(\widehat{F}_\epsilon(y) - \widehat{F}_{\epsilon_0}(y))$, $y \in \mathbb{R}$ ainsi que les distributions asymptotiques des deux statistiques de tes sous l'hypothèse nulle et sous une alternative locale. Une procédure bootstrap est utilisée pour approximer les valeurs critiques des statistiques de test sous H_0.

Sélection de modèles, 15h05-16h25.

Un critère de validation croisée approximé universel

Daniel Commenges (INSERM, ISPED, Bordeaux), Cécile Proust-Lima (INSERM, ISPED, Bordeaux), Cécilia Samieri (INSERM, ISPED, Bordeaux), Benoit Liquet (Université de Bordeaux et University of Cambridge)

La sélection d'estimateurs est une tâche essentielle en modélisation. Un cadre général est que les estimateurs de la distribution sont obtenus par la minimisation d'une fonction (la fonction d'estimation) et ils sont évalués par une autre fonction (la fonction d'évaluation). Les fonctions d'estimation et d'évaluation estiment généralement des risques. Un cas classique est que les deux fonctions estiment un risque l'entropie croisée (cross-entropy) ; dans ce cas le critère d'information d'Akaike (AIC) est pertinent. Dans des cas plus généraux, l'évaluation des risques peut être estimée par validation croisée. La validation croisée est très exigeante en calcul : une formule d'approximation peut donc être très utile. Un critère de validation croisée approximée universel (UACV) est proposée. Ce critère peut être adapté à différents types d'estimateurs (vraisemblance pénalisée, maximum a posteriori) et de fonctions d'évaluation. Cette formule se réduit au critère d'information de Takeuchi(TIC) lorsque l'entropie croisée est choisie pour définir les fonctions d'estimation et d'évaluation. La distribution asymptotique de UACV et d'une différence de UACV peut être obtenue.

Sélection de variables de calage par une méthode de bootstrap

Guillaume Chauvet (Ensai (Crest)), Camelia Goga (Institut de Mathématiques de Bourgogne)

Lors d'une enquête en population finie, les estimateurs sont généralement calés sur des totaux auxiliaires afin de réduire leur variance. La sélection des variables utilisées lors du calage est un problème important en pratique, car les variables de calage les plus explicatives font généralement diminuer la variance, mais un trop grand nombre de variables de calage peut conduire au contraire à la réaugmenter. Nous proposons ici un critère Bootstrap d'arrêt dans l'inclusion de variables de calage. Une courte étude par simulations montre que la méthode proposée conduit à une plus grande parcimonie dans le nombre de variables de calage, avec des poids calés moins dispersés, et sans inflation pour la variance.

Régression gaussienne à poids logistiques et maximum de vraisemblance pénalisé

Lucie Montuelle (Université Paris Sud 11/ Inria Saclay Idf), Erwan Le Pennec (Université Paris Sud 11/ Inria Saclay Idf)

Cette communication s'inscrit dans le cadre général de l'estimation de densités. Nous souhaitons estimer des densités conditionnelles à l'aide de mélanges gaussiens, ce qui revient à estimer les différents paramètres de ces mélanges, ainsi que le nombre de composantes, dépendants d'une covariable. Cette dépendance rend l'estimation des paramètres plus difficile que dans le cadre traditionnel des mélanges gaussiens à paramètres fixes (McLachlan et Peel). Par conséquent, peu de résultats théoriques ont été établis pour des paramètres conditionnés par une covariable. Nous nous sommes concentrés sur des poids logistiques et des moyennes dépendants de la covariable. Les seuls résultats à notre connaissance, correspondant à cette situation, sont de Chamroukhi et al., qui proposent des simulations numériques basées sur l'EM et le critère BIC, avec des poids logistiques affines et des moyennes polynomiales. En nous appuyant sur les outils théoriques fournis par Cohen et le Pennec, nous présenterons une inégalité d'oracle, pour une stratégie de maximum de vraisemblance pénalisé, permettant d'estimer les différents paramètres (variables) du mélange, ainsi que le nombre de composantes. Nous proposerons un choix de pénalités, proportionnel à la dimension du modèle, permettant d'assurer une convergence rapide de l'erreur entre estimateur du maximum de vraisemblance pénalisé et densité cible. Nous illustrerons enfin nos résultats théoriques par des simulations numériques.

Critères icl pour la sélection de modèle pour la classification croisée de données continues

Aurore Lomet (UTC), Gérard Govaert (HEUDIASYC, UMR CNRS 7253), Yves Grandvalet (UTC)

La classification croisée a pour objectif de partitionner simultanément les lignes et les colonnes d'un tableau de données pour révéler la structure en blocs homogènes. L'une des méthodes proposées se base sur le modèle probabiliste des blocs latents. Pour un même jeu de données, plusieurs classifications croisées peuvent être proposées : elles peuvent différer par leur nombre de classes par exemple. La sélection du nombre de classes devient alors un problème fondamental afin d'obtenir une classification des données pertinente. Pour résoudre ce problème, nous proposons une nouvelle procédure utilisant des critères de sélection de modèle basés sur la vraisemblance classifiante intégrée (ICL). Nous développons un critère asymptotique ICL-BIC, un critère dérivé BIC et un critère exact ICL se basant sur les distributions conjuguées. Les résultats obtenus par le critère ICL exact et les deux critères asymptotiques ICL-BIC

et BIC sur des jeux de données simulées montrent que ceux-ci sont performants et robustes pour des tableaux suffisamment grands quant à la sélection du nombre de classes et du type de modèle.

Études de cas 1, 15h05-16h25.

Split plot et strip plot - applications industrielles

Pascale Rondeau (Danone Research), Olivier Brack (K.S.I.C.), Maïna Kerbrat (Aixial For Danone Research)

Les Plans d'expériences sont aujourd'hui largement utilisés dans le milieu industriel. Ils contribuent au développement et à l'optimisation de nouveaux procédés. Dans le cadre de la fabrication d'un produit laitier, le procédé est fait de plusieurs unités process, celles-ci soumises à des contraintes. Les contraintes principales sont le nombre d'essais réalisables par jour, et la gestion de la configuration qualitative des unités de process. L'expérimentateur cherchera à rassembler les essais de façon pertinente dans le but de réduire la perte de matières premières donc le coût. Les contraintes liées au process ne sont, dans la plupart des cas, que partiellement prises en compte. Les plans de la famille des split plot et strip plot, suggérés ces dernières années par D.K. Sehgal, B. Jones et Christopher J. Nachtsheim (oct2009) semblent tout à fait appropriés pour remédier à cette limitation... Une approche comparative des solutions classiques et alternatives est proposée au travers d'exemples d'application en milieu industriel. Des indicateurs (en termes d'écriture du modèle, d'efficacité, de nombre d'essais,...)sont comparés entre matrices développées pour ces exemples (matrice de criblage, split plot, strip plot). Des recommandations seront proposées pour définir un usage adéquat de ces matrices strip plot et split plot. JMP 10.0 sera utilisé pour générer et traiter ces types de plans.

Analyse par la méthode des données de panel estimation des modèles du parc auto : cas de l'Algérie

Rachid Toumache (LASAP-ENSSEA), Khaled Rouaski (LASAP-ENSSEA)

L'évolution du parc automobile en Algérie est due à la variation non linéaire du revenu présenté par la richesse nationale des pays (PIB) plutôt quo d'autres facteurs infrastructures tels que le prix de véhicule, le prix du carburant, le réseau routier, la densité de la population, l'étendu du pays... etc. Cette étude prévoit l'image future du parc automobile algérien, elle est basée sur la technique d'une série chronologique de coupe instantanée. L'évolution du parc auto est modélisée en utilisant trois modèles utilitaires fournis par la littérature à savoir : la fonction Gompertz, la fonction Quasi-logistique et la fonction Logistique. En outre, ces modèles ont été calibrés par l'usage des données de panels ou bien des données regroupées. 46 pays du monde (sections) ont été captés durant 32 ans allant de 1971 au 2002 (source : banque mondiale) pour construire quatre panels qui comprennent l'Algérie, la Chine, l'Inde et les Etats-Unis. Pour ce choix, il a été pris en considération la tendance internationale de l'évolution du parc en fonction du PIB ainsi que les pays ayant les mêmes caractéristiques que notre pays. Comme résultat des travaux, un ensemble de scénarios futurs du parc auto. algérien ont été dégagés par les différents modèles statistiquement significatifs.

Élaboration d'un score géomarketing

Ronan Le Gleut (ENSAI), Suzanne Michel (ENSAI), Éléonore Prévost (ENSAI)

Ce projet a pour but l'élaboration d'un score géomarketing pour La Banque Postale dans les villes de plus de 15 000 habitants (en France métropolitaine hors Corse et Île-de-France). Dans un premier temps, il s'agit de décrire le parc des agences. Deux méthodes d'analyse de données ont été appliquées : l'analyse en composantes principales, puis l'analyse factorielle multiple. A la suite de celles-ci nous avons procédé à une classification ascendante hiérarchique. Nous distinguons alors trois profils d'agences suivant la situation géographique et le profil sociodémographique de la population : les agences à "la campagne", en zones balnéaires, voire plus globalement "touristiques", et dans les grandes agglomérations. Dans un second temps, il est intéressant de modéliser le bénéfice potentiel réalisable sur chaque zone. Les agences sont différenciées selon leur statut : agence mère (importante) ou agence satellite (dépendante de son agence mère). Cependant, seul un modèle sur les agences satellites était possible. Nous avons étudié différents modèles suivant plusieurs méthodes de sélection de variables. En raison de l'hétérogénéité des données, la construction d'échantillons d'apprentissage/test n'a pas été possible. La modélisation sur des échantillons bootstrap ne nous a pas permis de proposer un modèle pleinement satisfaisant mais la régression non paramétrique apporte des pistes pour l'améliorer.

La prévision de la détresse financière des entreprises tunisiennes par le modèle régression logistique semi paramétrique et les réseaux de neurones

Sami Mestiri (), Abdeljelil Farhat ()

L'objectif de cet article est de comparer deux techniques de classification des entreprises : la régression logistique semi paramétrique et les réseaux de neurones dans le but de prévoir le risque de crédit des banques tunisiennes. L'échantillon utilisé comporte 528 firmes tunisiennes de différents secteurs d'activités dont nous disposons des bilans et des comptes financiers des exercices 1999-2006. Une différence a été constaté entre le modèle régression logistique semi paramétrique et celui basé sur les réseaux de neurones en terme de performance de distinction entre les entreprises saines et celles en détresse. En fait, nous avons démontré que les modèles basés sur les réseaux de neurones donnent des meilleurs résultats des prévisions en termes de bon classement ainsi que par les résultats obtenus de la courbe ROC.

Données de survie 1, 15h05-16h25.

Inférence sur l'effet de régression temporel en analyse de survie

Cécile Chauvel (LSTA, UPMC), John O'Quigley (LSTA, UPMC)

Dans le contexte de l'étude des durées de vie, le coefficient de régression du modèle à risques proportionnels de Cox est supposé constant au cours du temps. Une méthode graphique facilement implémentable et interprétable pour évaluer l'évolution de ce coefficient de régression temporel, et par extension, la validité du modèle, est présentée. La méthode repose sur la théorie des processus empiriques. Si le coefficient de régression est constant au cours du temps, le processus converge vers un mouvement brownien avec dérive. Des tests d'hypothèse sur la valeur moyenne du coefficient de régression temporel peuvent également être dérivés.

Un nouveau test pour l'analyse de données de prévention en recherche clinique

Valérie Gares (INSERM), Sandrine Andrieu (INSERM), Jean-François Dupuy (INSA de Rennes), Nicolas Savy (Université Paul Sabatier)

Il est bien connu que la statistique du logrank est la meilleure statistique pour tester l'hypothèse des risques proportionnels. Cependant, dans le cadre de la prévention en recherche clinique, il est naturel de considérer qu'il y a ou qu'il peut y avoir des effets tardifs. Précédemment, les auteurs ont considéré une statistique du logrank pondéré développée par Fleming et Harrington et donnent des recommandations pour l'utilisation de ce test dans le contexte de recherche clinique. Cependant, l'utilisation de ce test fait l'hypothèse d'un effet tardif. Or, quand on met en place un essai clinique, nous ne sommes pas certains d'un effet et/ou nous ne voulons pas prendre le risque de faire certaines hypothèses. Dans cet exposé, nous considèrerons une statistique définie comme le maximum entre la statistique du logrank et la statistique de Fleming-Harrington. Cette statistique nous permet de ne pas faire l'hypothèse d'effet tardif. Nous présenterons ici la performance de ce test et nous donnerons des recommandations pour son utilisation dans les essais cliniques, notamment le calcul du nombre de sujets nécessaires.

Inférence pour des modèles de survie paramétriques

Damien Bousquet (I3M, Université de Montpellier)

Dans ce travail, nous présentons une nouvelle méthode d'estimation des paramètres ajoutés dans les distributions de probabilités introduites par Bousquet, Daurès et Marin (2011). La construction de ces nouvelles distributions de probabilités se base sur une généralisation du procédé de Marshall et Olkin (1997). La méthode d'estimation des paramètres ajoutés est de type moments. Cette méthode intègre la présence de données censurées à droite. Il s'agit d'une alternative à l'autre méthode des moments présentée par Bousquet, Daurès et Marin (2012).

Modèle illness-death pour données censurées par intervalle avec effets aléatoires : application à la cohorte paquid

Célia Touraine (Inserm, Isped, Univ Bordeaux), Pierre Joly (Inserm, Isped, Univ Bordeaux)

Le modèle illness-death permet aux individus de passer d'un état dit "sain" à un état dit "décédé", soit directement, soit en passant par un état intermédiaire dit "malade". Des modèles de régression tels que des modèles à intensités de transition proportionnelles sont utilisés pour évaluer l'influence de facteurs individuels sur chaque transition. Nous proposons d'intégrer à ces modèles des effets aléatoires associés à des facteurs qui seraient partagés par les individus appartenant à un même groupe. Nous appliquerons un tel modèle illness-death à l'étude de la démence pour des données tronquées à gauche et censurées par intervalle.

Segmentation, 15h05-16h25.

Apports de la segmentation pour construire des indices de similarité entre séries temporelles

Christian Derquenne (EDF)

Le pré-traitement des données est essentiel quelle que soit la complexité de l'application et des méthodes statistiques qui seront mises en oeuvre. Notamment l'utilisation de séries temporelles demande une attention particulière si l'on désire les relier entre elles afin de rechercher des similarités globales et/ou effectuer des analyses locales. La méthode de segmentation de courbes que nous avons développée par ailleurs est alors un outil performant pour exhiber de l'information exhaustive. Cela peut faciliter la découverte de points de rupture entre deux séries, la modélisation de liaisons non linéaires, l'identification de l'évolution locale (par segment) de deux phénomènes, ... Nous proposons quatre indices de similarité incluant l'information apportée par la segmentation. Le premier permet de mesurer l'influence statistique temporelle d'une variable explicative sur une réponse, le deuxième exhibe la corrélation entre deux variables, le troisième teste le nombre de points de rupture en commun entre les deux séries, alors que le dernier évalue leur écart moyen. Ces indices, appliqués sur des données réelles, fournissent une aide précieuse pour faire apparaitre des informations cachées qui peuvent se révéler très fructueuses pour l'expert du domaine d'application. Nos futures recherches seront dédiées au développement d'autres indices reposant également sur notre méthode de segmentation, afin de réaliser de la classification de courbes, par exemple.

Détection de ruptures à partir de méthodes à noyaux

Morgane Pierre-Jean (CERIM, Univeristé Lille 2), Guillemette Marot (CERIM, Univeristé Lille 2), Guillem Rigaill (URGV, Université Evry Val d'Essonne), Alain Célisse (Laboratoire de mathématique Painlevé, Université Lille 1)

Nous nous plaçons dans le cadre de détection de ruptures dans des signaux tels que des profils génomiques provenant de cellules tumorales. Les profils génomiques mesurés par des expériences à haut débit ont la particularité d'être constants par morceaux mais également d'être de grande dimension du fait du grand nombre de loci le long du génome. Ces profils génomiques ont aussi la caractéristique d'être bivariés et l'une des deux dimensions est un signal multimodal, ce qui rend la détection de ruptures dans la moyenne difficile (méthodes classiques pour la détection de ruptures). L'algorithme proposé pour segmenter ces signaux utilise une méthode à noyau qui permet de détecter les changements dans toute la distribution du signal. Cet algorithme utilise la programmation dynamique ce qui le rend très efficace. Un autre problème se pose sur la sélection de modèle après la segmentation. Nous proposons de comparer les différentes méthodes de calibration de constantes de pénalités sur des simulations. Une autre partie sera ensuite consacrée à l'évaluation de la qualité des segmentations trouvées par la méthode. Pour finir, nous allons également évaluer l'influence du choix du noyau sur la qualité des segmentations.

Une approche robuste pour la segmentation d'un processus AR(1)

Souhil Chakar (AgroParisTech/INRA)

Nous considérons le problème de la détection de ruptures dans l'espérance d'un processus gaussien AR(1). La prise en compte de la structure de dépendance ne permet pas d'utiliser la démarche d'inférence du cas indépendant. En particulier, l'algorithme de programmation dynamique permettant d'obtenir la solution optimale ne peut pas être utilisé. Nous proposons ici un estimateur robuste du paramètre d'auto-corrélation dont nous montrons la consistance. Nous proposons alors de reprendre la démarche d'inférence classique en utilisant cet estimateur dans les différents critères d'estimation des ruptures et de leur nombre. Nous montrons que les propriétés asymptotiques de l'estimation dans le cadre classique sont conservées, et illustrons par des simulations l'intérêt de la prise en compte de la structure de dépendance.

A penalized maximum likelihood estimator for the segmentation of RNA-seq data

Alice Cleynen (AgroParisTech), Emilie Lebarbier (AgroParisTech)

Nous considérons ici un modèle de segmentation pour des données de comptage provenant d'expériences de séquençage du génome (nombre de reads débutant à chaque position du génome) que nous modélisons à l'aide de la loi binomiale négative. Nous proposons un estimateur de log-vraisemblance pénalisée pour le choix du nombre de segments dans un contexte non-asymptotique. Le choix de notre pénalité est inspiré des articles de Birgé et Massart, de sorte que l'estimateur correspondant vérifie une inégalité oracle. Nous illustrons les performances de notre estimateur sur des données simulées et de RNA-Seq.

Graphes, 16h45-17h45.

Log-linear models on non-product spaces

Tamas Rudas (Eotvos Lorand University), Anna Klimova (Institute of Science and Technology)

Log-linear models on non-product spaces arise naturally in machine learning (feature selection), in many problems of official statistics (e.g., the analysis of congenital abnormalities) or in market research (market basket analysis). In these cases, not all combinations of the 'Yes' – 'No' categories of the variables are observed : newborns with no congenital abnormalities are not recorded, each purchase consists of, at least, one item. A classical variant of independence, applicable when the sample space is not a Cartesian product, is the Aitchison – Silvey independence, which assumes that there is no overall effect, that is there is no common effect that would apply to all category combinations or cells. Relational models are common generalizations of these and of standard log-linear models. Some of the properties of the families of distributions in this model class are quite surprising, both from the algebraic and the stochastic points of view. Algebraically, the models can be expressed using a set of generalized odds ratios, and if there is no overall effect present, there is exactly one out of these generalized odds ratios, that is non-homogeneous. Poisson and multinomial likelihoods under models without the overall effect are not equivalent, sufficient statistics are not preserved in the maximum likelihood estimates, and the existence of a factorization of a model does not necessarily imply likelihood independence of the components. The talk will present these properties and give simple illustrations of the applications of relational models.

Consensus LASSO : inférence conjointe de réseaux de gènes dans des conditions expérimentales multiples

Nathalie Villa-Vialaneix (Unité MIAT, INRA de Toulouse), Magali San Cristobal (INRA-INSA de Toulouse)

Nous présentons ici une méthode pour l'inférence de réseaux de co-expression génique à partir de données d'expression obtenues dans des conditions expérimentales différentes. Cette approche est basée sur une double pénalité, permettant d'une part l'obtention d'une solution parcimonieuse et, d'autre part, la proximité entre les divers réseaux inférés dans les diverses conditions et un réseau consensus commun à toutes les conditions.

Impact du réseau social dans un modèle d'échange en population finie

Pierre Barbillon (AgroParisTech INRA), Mathieu Thomas (AgroParisTech INRA, Université de Groningen)

Les échanges de graines entre paysans est un sujet d'étude important dans la mesure où ils ont une influence sur l'évolution de la diversité des variétés cultivées. Ces échanges sont structurés par un réseau social entre paysans. Afin de mieux comprendre ce processus dynamique d'échange et son importance, nous proposons d'étudier un modèle stochastique d'extinction-colonisation prenant en compte la complexité du réseau social : les échanges ne sont supposés possibles qu'au travers un réseau social fixé tandis qu'un phénomène d'extinction peut intervenir aléatoirement pour chaque paysan à chaque génération. Il est alors possible d'explorer l'influence des propriétés topologiques du réseau sur la persistance d'une variété donnée dans le réseau au bout d'un nombre de génération fixé. Nous nous concentrons principalement sur trois types de réseaux sociaux afin de décrire des systèmes d'organisation différents. Nous prenons en compte le fait que le nombre de fermes est fini et donc responsable de variabilité dans les résultats. Ceux-ci sont obtenus par calcul exact lorsque le nombre de fermes est petit et par simulation autrement. La précision des résultats de simulation est améliorée par un filtre particulaire ou par des techniques de splitting.

Modèles de mélange, 16h45-17h45.

Modèle de mélange poissonien pour l'analyse de données de déplacements

Andry Randriamanamihaga (ifsttar), Etienne Côme (ifsttar), Latifa Oukhellou (IFSTTAR), Gérard Govaert (UTC)

Les systèmes de Vélos en Libre Service déployés dans plusieurs grandes métropole ces dernières années ouvrent des perspectives intéressantes et nouvelles en termes d'analyse de la mobilité. Dans cet article, nous proposons un modèle génératif pour extraire des données générées par ces systèmes des motifs pertinents. Le modèle proposé s'appuie sur un modèle de mélange poissonien et prend en compte différentes spécificités de l'application. Celui-ci permet d'extraire des groupes d'origine et de destination ayant des profils d'usage similaire au cours du temps, ce qui permet de mieux cerner les causes sous jacentes à la mobilités des habitants.

Mise en garde sur l'utilisation des mélanges gaussiens avec données manquantes

Vincent Vandewalle (IUT Roubaix, Université Lille 2), Christophe Biernacki (INRIA Lille - Université Lille 1)

Les données manquantes sont un problème bien connu des statisticiens mais leur fréquence d'occurrence s'accroît avec l'augmentation de la taille des jeux de données modernes. En classification non supervisée dans un cadre de modèles de mélanges gaussiens, l'algorithme EM permet en principe de traiter facilement ce type de données en introduisant deux niveaux de données manquantes : la classe et les variables. Cependant, le phénomène bien connu de dégénérescence dans les mélanges gaussiens est ici particulièrement sensible lors de la mise en pratique de l'algorithme EM. En effet, les expériences numériques montrent clairement que la dégénérescence est relativement lente et plus fréquente qu'avec des données complètes. En pratique, des situations de ce type sont dangereuses car difficiles à détecter. À la clé, le risque est de considérer comme valable une solution proche d'une solution dégénérée ou encore de perdre beaucoup de temps à itérer inutilement avant de détecter que le chemin suivi mène à une solution dégénérée.

Détection non-asymptotique de mélanges à moyennes inconnues

Béatrice Laurent (IMT, INSA Toulouse), Clément Marteau (Inst. de Math. de Toulouse), Cathy Maugis-Rabusseau (Inst. de Math. de Toulouse)

Ce travail s'intéresse à la détection de distributions de type 'mélanges' dans un cadre uni-dimensionnel. Plus précisément, l'objectif est de déterminer si la distribution d'un échantillon $X_1, \ldots, X_n$ suit (à une translation près) une loi de référence ϕ ou bien est définie comme un mélange à deux composantes. Nous proposons une procédure de test non-asymptotique et établissons des conditions pour lesquelles la puissance du test est contrôlable. Dans un second temps, nous comparons les performances de notre algorithme aux méthodes existantes dans la littérature dans un cadre asymptotique de référence.

Données fonctionnelles 1, 16h45-17h45.

Plans de sondage à grande entropie pour données fonctionnelles : estimation de la variance de l'estimateur de horvitz-thompson et construction de bandes de confiance pour la moyenne

Hervé Cardot (Université de Bourgogne), Camelia Goga (Institut de Mathématiques de Bourgogne), Pauline Lardin (EDF, La Poste)

Pour des plans de sondage de taille fixe à grande entropie, la variance de l'estimateur de Horvitz-Thompson peut être approchée par la formule de Hàjek. L'intérêt de cette approximation asymptotique de la variance est qu'elle ne fait intervenir que les probabilités d'inclusion du premier ordre. Nous étendons cette formule au cas ou la variable d'intérêt est fonctionnelle et montrons sous des hypothèses sur l'entropie du plan de sondage et sur la régularité des trajectoires qu'elle fournit un estimateur uniformément convergent de la variance de l'estimateur de Horvitz-Thompson de la trajectoire moyenne. Les vitesses de convergence sont obtenues pour le cas particulier du plan réjectif. Nous en déduisons qu'il est possible de construire des bandes de confiance dont la couverture est asymptotiquement celle souhaitée par simu-

lations de processus gaussiens dont la variance est la variance estimée. Cette méthode est illustrée sur l'estimation de la consommation électrique moyenne à partir d'un échantillon de courbes de consommation individuelles mesurées chaque demi-heure pendant une semaine.

Test sur la significativité de variables fonctionnelles en régression

Samuel Maistre (CREST-Ensai), Valentin Patilea (CREST- Ensai)

Nous considérons un problème de régression avec une variable expliquée à valeurs dans un espace de Hilbert de dimension finie ou infinie et des variables explicatives 'hybrides' : une partie à valeurs dans un espace de dimension finie, une autre partie fonctionnelle à valeurs dans un espace de Hilbert de dimension infinie. Le problème étudié est le test de significativité de régresseurs fonctionnels. Ce type de problème apparait dans nombreuses situations : test de significativité d'une variable fonctionnelle dans un modèle semi paramétrique partiellement linéaire fonctionnel avec une variable expliquée réelle, test de significativité de variables fonctionnelles dans une régression non paramétrique avec de régresseurs 'hybrides', test de l'effet d'une variable fonctionnelle sur une autre variable fonctionnelle ou de dimension finie,... Nous proposons une nouvelle méthode de test basée sur un lissage à noyau sur une variable de dimension finie et fixée q. En laissant tendre le paramètre de lissage h vers zéro, la statistique de test proposée a une loi asymptotique normale centrée réduite sous l'hypothèse nulle. Le test associé il détecte des alternatives locales à la Pitman approchant l'hypothèse nulle moins vite que $n^{-1/2}h^{-q/4}$ où n est la taille de l'échantillon i.i.d. Ainsi les dimensions de la variable expliquée et de régresseurs fonctionnelles n'ont pas d'effet sur le comportement asymptotique du test.

Classification bayésienne non supervisée de données fonctionnelles

Damien Juery (Montpellier SupAgro), Christophe Abraham (Montpellier SupAgro), Bénédicte Fontez (Montpellier SupAgro)

Nous nous intéressons à la classification non supervisée de données fonctionnelles dans un cadre statistique bayésien. Nous généralisons un modèle de classification de données, basé sur le processus de Dirichlet, aux données fonctionnelles. Le caractère innovant tient au fait que, contrairement à d'autres articles qui font usage de la dimension finie en projetant les courbes dans des bases de fonctions, les calculs sont ici réalisés sur les courbes complètes, c'est-à-dire en restant en dimension infinie. Le cadre des espaces de Hilbert à noyaux reproduisants nous permet alors d'exprimer les densités, en dimension infinie, des courbes par rapport à une mesure gaussienne. Nous proposons un algorithme qui généralise l'algorithme Gibbs with Auxiliary Parameters (Neal, 2000) dans le cas de processus. Les performances sont comparées à celles d'autres méthodes déjà existantes, puis discutées.

Études de cas 2, 16h45-17h45.

Présentation des logiciels Wolfram Research

Martin Hadley (Wolfram)

Mathematica est un environnement de programmation et de développement de haut niveau utilisé pour l'enseignement, la recherche et le monde de l'entreprise dans des domaines interdisciplinaires. Notre philo-

sophie de développement tout-en-un et hybride numérique/symbolique est en adéquation avec l'évolution constante des méthodes et applications de la statistique. Enfin nous évoquerons Wolfram—Alpha, résultat de nos efforts dans les domaines du Data Science, du Web sémantique et du langage naturel. Moteur de connaissances permettant de résoudre un large éventail de problèmes mathématiques et d'accèder à de nombreuses collections de données via une requête en langage naturel. Nous démontrerons l'avancée des travaux sur l'analyse automatique des données.

Echantillon essaimé : une methode pour augmenter les petits échantillons

Alain Morineau (MODULAD), Thi Minh Thao Huynh (DEENOV), Roland Marion-Gallois (MEDTRONIC)

On dispose d'un petit échantillon (taille 50 par exemple) sur un grand nombre de variables qualitatives présentant un système complexe de liaisons. Comment simuler, sans hypothèses sur les distributions, un échantillon de grande taille (disons, 500) possédant l'essentiel des propriétés statistiques du petit échantillon, c'est-à-dire semblant extrait de la même population ? La procédure comporte plusieurs étapes : (1) ACM du petit échantillon et sélection d'un sous-ensemble d'axes servant de support ; (2) simulation de coordonnées (500 par exemple), axe par axe, dans des lois normales indépendantes de variances égales aux valeurs propres ; (3) pour chacune des 500 lignes du tableau simulé, chercher ses k plus proches voisins (k-PPV) dans le petit fichier ; (4) utiliser une moyenne des k-PPV doublement corrigée pour tenir compte des fréquences des modalités et de la distance des voisins ; on obtient 500 points moyens à coordonnées quantitatives ; (5) passer au codage disjonctif le plus proche par "vote majoritaire". La moyenne des k-PPV permet de transférer sur les simulations le système des liaisons entre les variables. Les formules de correction comportent des paramètres (dont k) optimisés pour minimiser le carré des écarts entre les distributions. Les tests classiques permettent d'évaluer la qualité de l'essaimage. De même que le Bootstrap cherche à évaluer des variabilités sans hypothèse sur les distributions, l'essaimage d'un échantillon cherche à réaliser des simulations en s'appuyant sur les données, et rien que les données.

La maintenance prédictive au service de l'industrie avec ibm

Serge Retkowsky (IBM), Peggy Vaugard (IBM)

Maîtriser les coûts tout en assurant le niveau de qualité désiré constitue le défi clé de la production industrielle. Directeurs d'usine, responsables de la chaîne logistique, bureaux des méthodes, spécialistes de la maintenance et du contrôle qualité, ingénieurs de maintenance et spécialistes du contrôle qualité : tous cherchent à éviter les coûts induits par les indisponibilités non planifiées et les pannes des équipements. Mais ils doivent aussi maîtriser le coût des opérations de maintenance, de réparation et de rénovation (MRO)... Il existe désormais un moyen d'aller au-delà de la maintenance préventive, planifiée à intervalles réguliers. Grâce aux solutions d'analyse prédictive comme celles développées par IBM SPSS, les industriels peuvent mettre en place de nouvelles normes de qualité et réaliser des économies en réduisant le temps d'indisponibilité résultant des opérations de maintenance non planifiées. Ils peuvent ainsi éliminer quasiment toute maintenance inutile.

Analyse de données 1, 16h45-17h45.

Staqis : modélisation de multitableaux à quatre entrées d'après une généralisation de statis

Robert Sabatier (Université Montpellier), Christelle Reynes (Université Montpellier 1), Myrtille Vivien (Université Montpellier 1)

Dans la vaste littérature consacrée aux analyses de tableaux à plusieurs entrées, rares sont les méthodes qui traitent de structure à quatre dimensions, si celle-ci n'est pas un cube. Nous proposons une nouvelle méthode appelée STAQIS (Structuration des Tableaux à Quatre Indices de la Statistique) qui, à travers la généralisation naturelle de l'un des critères optimisés par STATIS, fournit une telle solution. Après avoir mis au point un algorithme, nous verrons sur un exemple réel, de grande dimension, l'applicabilité de cette méthode. Enfin, nous pourrons, à l'aide d'un test de permutation, vérifier si la structure du multitableaux à quatre dimensions est pertinente.

Integration of longitudinal biological data sets

Kim-Anh Le Cao (University of Queensland), Kathy Ruggiero (University of Auckland)

High-throughput biotechnology platforms enable rapid acquisition of enormous amounts of data. Analysing each data-type in isolation to understand the molecular mechanisms associated with complex physiological processes yields only one perspective on cellular responses to a given physiological state. However, a holistic understanding can only be achieved through a fully integrated systemic approach by combining information not only across multiple data types but also at multiple time points enabling the capture of temporal physiological changes. Efforts have been made towards developing methodologies to integrate two biological data sets and to identify the relationships between them at a given time point, or to analyse a single data set measured across several time points to identify correlated profiles. However, no attempts have been made to integrate data from two platforms or more whilst also taking account of temporal dependencies. We propose to integrate different data types measured from multiple high-throughput platforms collected at critical times throughout disease or state progression with a focus on how to select correlated biological entities across time points. The proposed methodology is based on the use of projection to latent structure multivariate approaches and smoothing splines to model the trajectories. The methodology is implemented in the R package mixOmics which proposes many visualisation tools for data integration.

Soft subpace clustering pour données multi-blocs basé sur les cartes topologiques auto-organisées SOM : 2S-SOM

Mory Ouattara (CSTB), Ndèye Niang (CEDRIC), Fouad Badran (CEDRIC), Corinne Mandin (CSTB)

Dans cette communication, nous proposons une méthode de soft subspace clustering basée sur les cartes topologiques pour la classification d'individus décrits par des variables structurées en blocs homogènes. L'algorithme nommé Soft subspace SOM (2S-SOM) consiste a optimiser la fonction de coût de SOM modifée en introduisant des poids adaptatifs sur les blocs et sur les variables de chaque bloc. Cette double pondération permet de distinguer les blocs les plus importants prenant ainsi en compte de la structuration en blocs, et d'identifier pour chaque bloc les variables les plus informatives. Le système de poids résultant

permet alors de déterminer simultanément les groupes d'individus et leurs sous espaces caractéristiques optimaux. La méthode proposée est illustrée sur des données réelles issues des bases de l'UCI repository of machine learning.

Enseignement 1, 16h45-17h45.

Motivating students to use business statistics beyond the classroom : a group problem solving approach

Nadine Galy (Toulouse Business School), Cameron Guthrie (Toulouse Business School), Anne Vanhems (ESC Toulouse)

Les enseignants en statistiques dans les écoles de commerce ont souvent des difficultés à motiver les élèves pour utiliser des techniques statistiques au-delà de la salle de classe. Des techniques d'apprentissage actives comme la résolution collective de problèmes ont été proposées pour surmonter ces obstacles. Le but de cette étude est de présenter et de tester un modèle motivationnel dans le domaine de l'enseignement des statistiques. La théorie de l'autodétermination de la motivation des élèves nous permet de faire le postulat suivant : la volonté des enseignants de statistiques à encourager l'autonomie des étudiants a un impact positif sur l'auto-perception des étudiants de leur autonomie et de leur compétence. Bien que nos résultats ne supportent que partiellement cette théorie de l'autodétermination, nous constatons que la motivation intrinsèque a un impact positif sur les intentions de poursuivre des études statistiques et d'utiliser les statistiques dans un contexte professionnel. Nous identifions aussi d'autres influences positives sur la motivation des élèves pour étudier les statistiques dans les écoles de commerce.

Projets tuteurés autour de la simulation aléatoire de jeux pour enfants

Frédérique Letué (LJK Université de Grenoble), Alain Birebent (IREM de Grenoble), Nathalie Catinot (IREM de Grenoble), Philippe Garat (LJK Université de Grenoble), Florent Girod (IREM de Grenoble), Damien Jacquemoud (IREM de Grenoble)

La plupart du temps, quand des activités de simulations sont présentées dans les chapitres 'probabilités et statistique' des manuels de mathématiques, celles-ci sont surtout des illustrations de théorèmes mathématiques que l'on sait prouver ou de calculs que l'on sait faire. Or, en statistique, on utilise les simulations à d'autres fins : établir la loi de variables aléatoires qu'on ne peut calculer 'à la main', proposer des valeurs d'hyper-paramètres, etc. Dans cet exposé, nous proposons des activités de simulation et de statistique pour des élèves de Terminale et des étudiants de DUT STID, dans des situations où le calcul n'est pas possible. Les activités sont basées sur des jeux de société pour enfants de 3 à 6 ans, où le déroulement du jeu réside essentiellement dans l'aléa. Apres avoir présenté les règles des jeux étudiés, nous montrerons comment les élèves/étudiants les ont simulés, les analyses statistiques qu'on peut tirer des simulations et discuterons l'intérêt pédagogique de tels projets.

Essai d'analyse statistique du parcours des étudiants : enquête par sondage

Hanya Kherchi Medjden (LASAP-ENSSEA), Khadidja Sadi (LASAP-ENSSEA)

Aujourd'hui le développement et la croissance d'une nation semblent dépendre de son niveau culturel et scientifique, et par la même, de la valeur de son enseignement. Depuis plus de deux décennies, l'enseignement supérieur algérien traverse une crise profonde et complexe. Cela conduit à une détérioration de la qualité, de la rentabilité et de l'efficacité de l'enseignement supérieur. Vu l'importance de la réussite et de l'échec dans le domaine de l'enseignement supérieur qui se manifestent surtout au niveau du tronc commun et vu la nécessité de l'orientation en spécialité. Cette recherche consiste à suivre le parcours des étudiants ingénieurs depuis la première année à l'ENSSEA (Ecole Nationale Supérieure de Statistique et d'Economie Appliquée) jusqu'à leur quatrième année. Les données sont recueillies d'une part, à partir de la bases de données disponible au service pédagogique et à partir d'une enquête par sondage réalisée sur les étudiants en fin de cycle de la même école. Le but de cette recherche est de faire d'abord, une analyse statistique de l'échantillon et ensuite l'élaboration d'un modèle de classification en trois classes (C0 : aucun doublement, C1 : un doublement et C2 : deux doublements) à l'aide de l'analyse discriminante (ADL) en prenant en compte des caractéristiques pédagogiques et non pédagogiques.

Résumés du mardi 28 mai 2013

Prix du Docteur Norbert Marx - Guy S. Mahiane, 08h30-09h30.

Modélisation des Interactions entre deux agents sexuellement transmissibles : le cas de l'Herpès Simplex Virus type-2 et du Virus de l'Immunodéficience Humaine

S. Guy Mahiame (Johns Hopkins University)

Nous présentons des modèles mathématiques pour l'étude des interactions entre deux infections sexuellement transmissibles. Ces modèles permettent d'estimer les probabilités de transmission et/ou d'acquisition par rapport et par partenariat sexuel en tenant compte des facteurs pouvant affecter ces probabilités, parmi lesquels la variabilité de la prévalence en fonction des âges des partenaires.

En modélisant la suite des états sérologiques de chacun des individus par une chaîne Markov, on peut calculer la vraisemblance dont la maximisation fournit une estimation des probabilités de transmission et des risques relatifs (pour la circoncision, l'usage du condom) et de l'effet de chacune des infections sur l'infectivité de l'autre. Nous proposons d'utiliser la méthode du Bootstrap estimer les intervalles de confiance des paramètres.

Nous illustrons notre procédure avec l'étude des interactions entre le Virus de l'Immunodéficience Humaine (VIH) et l'Herpès Simplex Virus type-2 (HSV-2) en utilisant les données issues d'une étude menée à Orange Farm (Afrique du Sud). Il ressort de cette étude que la présence de l'HSV-2 augmente le risque de transmission du VIH de la femme à l'homme. La circoncision masculine réduit aussi bien le risque d'acquisition du VIH que celui de l'HSV-2 par les individus de sexe masculin. En utilisant le critère d'information d'Akaike, on montre que le modèle par-partenariat s'ajuste mieux aux données.

Données fonctionnelles 2, 09h35-10h35.

Prédiction non-asymptotique et adaptative dans le modèle linéaire fonctionnel

Elodie Brunel (Université Montpellier 2), André Mas (Université Montpellier 2), Angelina Roche (Université Montpellier 2)

Nous présentons une procédure de sélection de la dimension pour l'estimateur de la régression en composantes principales fonctionnelle. Cette méthode consiste à minimiser un critère des moindres carrés sur l'espace engendré par les m vecteurs propres associés au plus grandes valeurs propres de l'opérateur de

covariance empirique. La dimension m de l'espace d'approximation est sélectionnée en minimisant le critère des moindres carrés usuel pénalisé. L'estimateur obtenu vérifie une inégalité-oracle non-asymptotique pour le risque lié à l'erreur de prévision et atteint la vitesse de convergence minimax sur une certaine classe d'ellipsoïdes. Les résultats numériques qui seront présentés montrent que cette méthode de sélection de la dimension possède de réels avantages en pratique par rapport aux méthodes usuelles de validation croisée.

Tests minimax adaptatifs pour les modèles fonctionnels linéaires

Nadine Hilgert (INRA Montpellier), André Mas (Université Montpellier 2), Verzelen Nicolas (INRA Montpellier)

Nous présentons deux nouvelles procédures pour tester la nullité du paramètre de pente dans les modèles linéaires fonctionnels à sortie réelle. Les statistiques de test sont obtenues en combinant une approche par tests multiples et des projections aléatoires des covariables explicatives sous forme d'analyse en composante principale fonctionnelle. Les procédures sont totalement basées sur les données et ne requièrent pas de connaissance a priori sur la régularité du paramètre de pente ni sur celle des covariables fonctionnelles. Les niveaux et puissances par rapport à des alternatives locales sont étudiées dans un cadre non asymptotique. Nous montrons ainsi que les procédures sont minimax adaptatives à la régularité inconnue de la pente, à un terme multiplicatif $\log \log n$ près, inévitable. Comme résultat supplémentaire, nous déduisons les distances de séparation minimax de la pente pour une large gamme de classes de régularité. Les résultats présentés dans cette communication sont issus de [Hilgert, Mas and Verzelen, Minimax adaptive tests for the functional linear model, arXiv :1206.1194v].

Intervalle de confiance pour la prévision dans un modèle fonctionnel

Anestis Antoniadis (Université de Grenoble), Xavier Brossat (EDF R&D), Jairo Cugliari (INRIA), Jean-Michel Poggi (Université Paris Descartes)

Nous construisons un intervalle de confiance pour la prévision d'un processus fonctionnel qui n'est pas nécessairement stationnaire. Le caractère non stationnaire est double : d'une part, le niveau moyen de la série change dans le temps, d'autre part il existe des groupes dans les données qui peuvent être vus comme des classes de stationnarité.

Le modèle de prévision utilisé consiste à trouver dans le passé des contextes similaires à la situation présente. Ainsi, on construit un vecteur de poids à l'aide d'une notion de similarité. Puis, les futurs de ces situations passés sont moyennés en utilisant les poids pour construire la prévision de la situation future.

La construction de l'intervalle de prévision doit prendre en compte la nature non stationnaire du processus. Nous utilisons les poids obtenus lors de la phase de recherche de situations similaires pour estimer la distribution de la prévision. La prévision coïncide avec la moyenne de la distribution ainsi obtenue. Enfin, les quantiles de cette distribution estimée sont utilisés comme les bornes de l'intervalle de prévision.

Statistique mathématique 2, 09h35-10h35.

Adaptive Bayesian estimation in Gaussian sequence space models

Rudolf Schenk (Université Catholique Louvain), Jan Johannes (Université catholique Louvain), Anna Simoni (Université de Cergy-Pontoise)

Nous considérons l'estimation bayésienne non paramétrique dans un modèle Gaussien séquentiel. Nous étudions la procédure d'un point de vue fréquentiste, plus précisément, nous sommes intéressés au taux de concentration de la loi a posteriori se rétrécissant vers la loi des observations. Nous dérivons d'abord des bornes inférieures et supérieures pour des taux de concentration pour des lois a priori Gaussiennes dependantes d'un paramètre de réglage. Ce résultat établit la consistance a posteriori mais le taux de concentration dépend du paramètre d'intérêt et de celui de réglage. Sous un choix approprié du paramètre de réglage nous dérivons le taux de concentration uniformément sur une classe de paramètres et nous montrons que ce taux correspond au taux minimax. Car le choix du paramètre de réglage dépend de la classe, nous introduisons une loi a priori hiérarchique et nous prouvons que le taux de concentration de la loi a posteriori dans un modèle Gaussien séquentiel direct équivaut le taux minimax et que l' estimateur bayesien qui ne dépend que des données est minimax-optimal.

Consistance et vitesse de contraction de l'a-posteriori bayésien dans le modèle de forme invariante

Dominique Bontemps (IMT - Univ. Paul Sabatier), Sébastien Gadat (IMT - Univ. Paul Sabatier)

Dans ce papier, le modèle de forme invariante désigne l'estimation d'une fonction f0 soumise à une translation aléatoire de loi g0 dans un modèle de bruit blanc. Nous nous intéressons à un tel modèle lorsque la loi des déformations est inconnue. Notre but est d'estimer la loi du processus P(f0,g0) ainsi que f0 et g0 eux-mêmes. Dans cette optique, nous adoptons un point de vue bayésien et décrivons un a-priori sur f et g tel que la distribution a-posteriori concentre autour de P(f0,g0) à une vitesse polynomiale quand n tend vers l'infini. Nous obtenons des vitesses de contractions logarithmiques pour la forme f0 et la distribution g0. Nous établissons également des minorations pour l'estimation de f0 et g0 dans le cadre fréquentiste.

Constructions probabilistes pour les copules discrètes

Olivier Faugeras (GREMAQ, Université Toulouse 1)

Soit X un vecteur aléatoire de fonction de répartition F discontinue. On propose deux approches probabilistes pour construire une fonction de copule C associée à X. Dans la première approche, des représentants $\mathbf{U}$ de copule C associée à X sont construits de façon explicite par des tranformations aléatoires. Ces couplages multivariés permettent de caractériser de façon probabiliste l'ensemble des fonctions de copules associées à un vecteur X discret. Les propriétés de dépendance de ces représentants de copule U sont comparées à celles du X initial et l'impact de la structure de randomisation est élucidée. Des formules explicites sont obtenues dans des cas particuliers (données bivariées, reconnaissance de forme). Enfin, une application de ces constructions à la copule empirique montre qu'elles sont aisément simulables. Dans la seconde approche, on donne une autre preuve purement probabiliste du théorème de Sklar par simple addition d'un vecteur continu et passage à la limite. Lorsque F est inconnue et continue mais

qu'on dispose d'un échantillon de données, on construit une séquence de représentants de copule associés à la fonction de répartition empirique qui converge presque sûrement vers celui de la copule associée à F. Ce résultat est étendu au cas discontinu et ce dernier théorème est ensuite interprété d'un point de vue inférentiel.

Analyse de données 2, 09h35-10h35.

Analyse factorielle discriminante multi-voie

Laurent Le Brusquet (SUPELEC), Arthur Tenenhaus (Supélec Sciences des Systèmes)

L'analyse factorielle discriminante est étendue au cas des données multi-voie, c'est-à-dire des données sur lesquelles ont été observées plusieurs modalités pour chaque variable. Les données multi-voie sont ainsi structurées en tenseur. L'extension proposée repose sur une modélisation des axes discriminants. Cette modélisation prend en compte la structure tensorielle des données. Les gains attendus par rapport aux méthodes consistant à construire un classifieur à partir de la matrice obtenue par dépliement du tenseur, sont une meilleure interprétabilité et un meilleur comportement vis-à-vis du sur-apprentissage, phénomène d'autant plus présent dans le contexte multi-voie que le nombre de modalités est grand. Un algorithme de directions alternées permet d'obtenir les axes discriminants. Les performances obtenues sur données simulées permettent de confirmer ces gains.

Imputation multiple à l'aide des méthodes d'analyse factorielle

Vincent Audigier (Agrocampus Ouest), François Husson (Agrocampus Ouest), Julie Josse (Agrocampus Ouest)

Les données manquantes constituent un problème incontournable dans la pratique de la statistique. Une solution commune pour gérer ces données manquantes consiste à remplacer chacune d'entre elles par une valeur plausible. On parle d'imputation simple. Néanmoins appliquer une méthode statistique sur un tableau imputé simplement pose un problème majeur : les données imputées jouent le même rôle que les données observées alors qu'elles sont incertaines. Pour rendre compte de cette incertitude, on peut proposer plusieurs imputations pour chaque donnée manquante. On parle alors d'imputation multiple.

Cette présentation a pour objet de nouvelles méthodes d'imputation multiple pour des données quantitatives, qualitatives et mixtes dans le cadre de données manquantes au hasard. L'idée est d'étendre les méthodes d'imputation simples basées sur l'emploi de techniques de réduction de la dimension telle que l'analyse en composantes principales.

Après avoir présenté les principes de l'imputation multiple, nous détaillerons les propriétés de nos différents algorithmes. Nous proposerons ensuite des simulations pour les situer par rapports aux méthodes existantes telles que l'imputation multiple par équations enchaînées, ou l'imputation reposant sur l'hypothèse d'une distribution jointe à l'ensemble des données.

Une extension de l'analyse factorielle multiple pour des groupes de variables mixtes

Amaury Labenne (IRSTEA), Marie Chavent (IMB, INRIA), Vanessa Kuentz-Simonet (IRSTEA), Tina Rambonilaza (IRSTEA), Jérôme Saracco (IMB, INRIA)

L'Analyse Factorielle Multiple (AFM) initialement proposée par Escofier et Pagès en 1982 est une méthode dédiée à l'étude d'un ensemble de n individus décrits par des groupes de variables quantitatives. Plus tard, cette méthode a été étendue pour prendre en compte des groupes de variables qualitatives (Pagès, 1983) puis simultanément des groupes quantitatifs et des groupes qualitatifs (Pagès, 2002). Cependant, cette méthode ne permet pas à l'heure actuelle de prendre en compte des groupes mixtes, c'est-à-dire contenant à la fois des variables quantitatives et qualitatives. Le but de notre étude étant de confectionner des indicateurs de développement durable en intégrant l'aspect de la qualité de vie, nous avons été confrontés à l'analyse de groupes de variables comportant des variables quantitatives et qualitatives. Dans ce travail, nous proposons une extension de l'AFM, appelée MFAMIX, pour l'analyse factorielle multiple de groupes de variables mixtes. Cette approche s'appuie sur une combinaison de l'AFM et de la méthode PCAMIX qui permet l'analyse de données mixtes. La méthode MFAMIX sera présentée à l'aide d'une décomposition en valeurs singulières et sera illustrée sur des données socio-économiques relatives à la qualité de vie.

Statistique médicale 2, 09h35-10h35.

Evaluation de protocoles pour des essais de bioéquivalence en crossover analysés par des modèles non linéaires à effets mixtes

Thu Thuy Nguyen (UMR 738 INSERM, Université Paris Diderot), France Mentré (INSERM UMR738, Paris 7), Anne Dubois (Département de Pharmacocinétique Clinique, Institut de Recherches Internationales Servier)

Les modèles non linéaires à effets mixtes (MNLEM) peuvent être utilisés pour analyser les essais de bioéquivalence en crossover. Une approche pour évaluer des protocoles des essais, basée sur la matrice d'information de Fisher, a été étendue aux MNLEM, tenant compte de la variabilité intra-sujet et les covariables discrètes, et implémentée dans PFIM3.2. Les objectifs du présent travail sont d'évaluer cette approche par simulation d'essais en crossover avec deux ou quatre périodes, deux séquences et de prédire la puissance du test de bioéquivalence pour plusieurs protocoles. Nous avons considéré différents scénarios, avec plusieurs protocoles et niveaux de variabilité. 1000 réplications de chaque scénario avec différents nombres de sujets ont été simulées sous l'hypothèse nulle du test de bioéquivalence, et analysées avec MONOLIX2.4. Nous avons comparé les erreurs standard (SE) prédites par PFIM à celles obtenues par simulation. Nous avons prédit par PFIM la puissance attendue et le nombre de sujets nécessaire (NSN) pour plusieurs protocoles, avec différentes hypothèses alternatives du test. Malgré une sous-prédiction de la SE de la variabilité intra-sujet, PFIM prédit correctement les SE de tous les autres paramètres comprenant les effets traitement, période, séquence ainsi que la puissance des tests et le NSN. PFIM est un outil efficace pour planifier des essais de bioéquivalence en crossover analysés par des MNLEM.

Détection et identification des clusters du cancer de la thyroïde en algérie

Oumelkheir Moussi (ENSSEA), Naima Boudrissa (ENSSEA), Mourad Semrouni (EHS CPMC), Fella Hasbellaoui (EHS CPMC)

Les cancers en Algérie sont devenus émergents et constituent une préoccupation majeure. Le cancer de la thyroïde augmente de manière spectaculaire depuis une quinzaine d'années, c'est le troisième cancer chez la femme d'après le registre des tumeurs d'Alger. Entre 2007 et 2010 il a été multiplié par 6 et ce seulement au service endocrinologie du centre Pierre et Marie Curie d'Alger. Ces observations nous amènent à se poser plusieurs questions : - Certaines wilayas ont-elles un nombre de cas de cancers excessif ? - Les cas de cancer de la thyroïde sont-ils anormalement concentrés ? - La distribution spatiale de ces cas, est-elle aléatoire ? Répondre à ces questions revient à décrire l'hétérogénéité Spatiale. Dans ce travail nous allons utiliser des méthodes de détection d'agrégats de cas et dont les statistiques sont souvent basées sur les distances afin d'analyser l'existence des clusters de cancers de la thyroïde en Algérie. Ce sont le test du coefficient de corrélation de Moran, le test de Tango et la méthode de balayage de Kulldorff. Le test d'ajustement a été utilisé pour vérifier l'hypothèse des risques constants de l'incidence du cancer de la thyroïde.

Analyse des données symboliques du trachome

Christiane Guinot (Université de Tours François Rabelais), Denis Malvy (Département de Médecine Interne, CHU St-André and Université Victor Segalen Bordeaux 2,), Jean-François Schemann (3 Institut de Recherche pour le Développement, Dakar, Sénégal), Filipe Afonso (Syrokko), Raja Haddad (Syrokko), Edwin Diday (CEREMADE Université Paris-Dauphine)

Résumé. Le trachome, causé par des infections oculaires répétées avec Chlamydia trachomatis dont le vecteur est une mouche, est une cause importante de cécité dans le monde. On présente ici une application des méthodes d'analyse des données symboliques à une étude d'intervention sur le trachome menée au Mali pour choisir parmi trois stratégies d'antibiothérapie celle présentant le meilleur rapport coût-efficacité et découvrir quels sont les paramètres sociodémographiques et environnementaux sur lesquels on pourrait tenter d'intervenir. L'Analyse des Données Symboliques se caractérise par l'étude de classes d'individus considérées comme nouvelles unités statistiques décrites par des variables dont les valeurs expriment pour chaque classe, la variation des valeurs prises par ses individus. Cette variation s'exprime dans cet article par des diagrammes de fréquences ou des intervalles de variation. Finalement, les résultats obtenus sont discutés à la lumière de ceux fournis antérieurement par la méthode de régression logistique multiple. L'analyse des Données Symboliques fournit effectivement un nouvel éclairage sur cette étude et suggère que certains paramètres sociodémographiques, économiques et environnementaux seraient liées à la maladie et son évolution au cours du traitement, quelle que soit la stratégie.

Ingénierie, 09h35-10h35.

Une démarche qualité au service de la préparation de données de recherches en sciences sociales

Cédric Gendre (INRA, US-ODR)

Les données sont des composantes stratégiques de nombreuses recherches. Avant d'entrer dans le processus de recherche proprement dit, ces données font l'objet de différents traitements (ou prétraitements) décrits au travers des étapes dites du "cycle de vie" des données. La fiabilité et la traçabilité de ces étapes, autrefois souhaitables, sont désormais nécessaires non seulement par souci d'efficacité, mais aussi parce que la profession l'exige. Par ailleurs, les données traitées peuvent être mises à disposition de différents partenaires qui, au même titre que la recherche, sont en droit d'exiger une qualité et une reproductibilité des processus et des traitements qui ont servi à l'élaboration de ces données. Nous proposons ici trois outils pour en promouvoir la diffusion : - une charte au travers de laquelle le gestionnaire des données s'engage à respecter un certain nombre de principes, - un guide des bonnes pratiques qui, pour chacune des étapes du cycle de vie des données, détaille différents outils et leur mise en oeuvre pratique, - un pense-bête qui permet une auto-évaluation en repérant les étapes qui mériteraient d'être améliorées.

Spatial exploratory analysis of the Guerry's data with GeoXp

Thibault Laurent (Toulouse School of Economics)

En 1833, André-Michel Guerry a écrit son célèbre Essai sur la statistique morale de la France. Récemment, Friendly (2007), Dray et Jombard (2011) et Filzmoser et al. (2012) ont appliqué sur ces données des méthodes statistiques modernes, issues de l'analyse multivariée ou de l'analyse de données spatiales. L'objectif de ce document est d'appliquer sur ces données historiques, les outils de l'analyse exploratoire interactive de données spatiales, en utilisant le package GeoXp (Laurent et al., 2012), disponible sur le logiciel R.

Des outils pour la recherche reproductible : Sweave et Statweave

Valérie Orozco (Toulouse School of Economics), Christophe Bontemps (Toulouse School of Economics)

En statistique, comme en économétrie, l'écriture de programmes, qu'ils soient destinés à la manipulation de données, aux traitements statistiques ou aux estimations économétriques, est nécessaire pour établir des résultats et les diffuser. Il est important que les programmes ayant engendré les résultats soient reliés, voire intégrés, aux résultats eux-mêmes. Le concept de recherche reproductible (Schwab, Karrenbach et Claerbout, 2000) qui encourage cette pratique et l'intégration des "codes" dans les documents de recherche permet aux lecteurs de comprendre, vérifier et reproduire les résultats. Cette rigueur (fiabilité, traçabilité, reproductibilité), et la maîtrise de ces programmes qui génèrent les résultats de recherche sont désormais de plus en plus exigés par notre profession ainsi que par les éditeurs de revues scientifiques. Nous présentons ici deux outils, Sweave et StatWeave, permettant d'intégrer programmes, résultats et éventuels commentaires dans un même document de travail ou un article de recherche écrit sous LaTeX ou OpenOffice.

Modèles semi-paramétriques, 11h00-12h20.

Choix optimal de fenêtre pour un estimateur semi-paramétrique des données de comptage

Tristan Senga Kiessé (Université de Nantes)

Dans cette communication, nous nous intéressons a un estimateur discret semi-paramétrique de distribution de denombrement. Des methodes de choix de paramètre de lissage optimal sont etudiées comme, par exemple, la validation croisee. Une nouvelle expression defenêtre optimale est proposee pour l'estimateur semi-paramétrique utilisant des noyaux associés discrets triangulaires ; de plus, des propriétes asymptotiques de la fonction de validation croisee sont etablies. La performance de l'estimateur semi-paramétrique selon le type de fenêtre optimale est evaluée sur des donnees réelles en appliquant une méthode de reechantillonnage .

Estimation de déformations entre distributions avec la distance de Wasserstein

Hélène Lescornel (Institut de Mathématiques de Toulouse), Jean-Michel Loubès (IMT Toulouse)

Nous présentons l'étude d'un modèle où les données proviennent de déformations d'une même distribution. Plus précisément, on observe des réalisations d'une variable epsilon à travers J différentes fonctions de déformation. La loi de epsilon, mu, et les fonctions de déformations sont inconnues. Nous nous plaçons dans un cadre semi paramétrique où l'on suppose que la forme de la déformation est connue mais que son importance, représentée par un paramètre d-dimensionnel θ_j^* est inconnue. Dans l'exposé nous proposerons des estimateurs pour les quantités $\theta^* = (\theta_1^*, ..., \theta_J^*)$ et μ puis nous présenterons leurs propriétés asymptotiques. L'estimateur des paramètres de déformation est obtenu en alignant les distributions des observations. Plus exactement, on minimise un critère empirique défini avec la distance de Wasserstein entre les mesures associées aux observations. Cette quantité permet de définir un estimateur de la densité mu. Nous obtenons tout d'abord la consistance des estimateurs principalement sous des hypothèses de régularité sur les fonctions de déformation. On utilise ici les résultats classiques de M estimation, avec des contraintes relativement faibles sur la loi structurelle mu. Dans un second temps, nous présenterons un résultat de convergence en loi pour l'estimateur des paramètres de déformation. Ce résultat, basé sur une Delta-Méthode requiert des conditions plus fortes sur les distributions étudiées.

Model equivalence tests for overidentification restrictions

Pascal Lavergne (Toulouse School of Economics)

Un nouveau cadre théorique est proposé pour tester la validité em approximative de conditions de moment sur-identifiantes. Cette validité est mesurée par une divergence entre la vraie distribution de probabilité des données et la mesure la plus proche qui impose les restrictions sur-identifiantes. La divergence peut être choisie parmi celles de la famille de Cressie-Read. L'hypothèse alternative est que cette divergence est plus petite qu'un seuil choisi par l'utilisateur. Je dérive l'enveloppe semi-paramétrique des tests invariants pour cette hypothèse et je propose des tests qui atteignent cette enveloppe.

Estimation adaptative dans le modèle single-index par l'approche d'oracle

Oleg Lepski (Aix-Marseille Université), Nora Serdyukova (Georg August Universitaet)

Dans le cadre de l'estimation non paramétrique d'une fonction multidimensionnelle nous nous intéressons à l'adaptation structurelle. Nous supposons que la fonction à estimer possède la structure ≪single-index≫ dans laquelle ni fonction de lien ni vecteur d'indice ne sont connus. Nous proposons une nouvelle procédure qui s'adapte simultanément à l'indice inconnu ainsi qu'à la régularité de la fonction de lien. Nous présentons une inégalité d'oracle ≪locale≫ (définie par la semi-norme ponctuelle) pour la procédure proposée, qui est ensuite utilisée pour obtenir la borne supérieure du risque maximal sous une hypothèse de régularité sur la fonction de lien. D'après la borne inférieure obtenue pour le risque minimax l'estimateur construit est un estimateur adaptatif optimal sur l'ensemble de classes considérées. Pour la même procédure on établit également une inégalité d'oracle ≪globale≫ (en norme Lr) et étudie sa performance sur les classes de Nikol'skii. Cette étude montre que la méthode proposée peut être appliquée à l'estimation de fonctions ayant une régularité inhomogène.

Statistique spatiale 2, 11h00-12h20.

On the random coefficient of AR models on Z^2 : structure and estimation

Abdelouahab Bibi (Université Mentouri, Constantine), Soumia Kharfouchi (Université Mentouri, Constanti)

Dans cette communication, nous analysons la structure probabiliste d'une classe de modèles indexés par Z^2 et gouvernée par des processus AR à coefficients stochastiques notés DS-AR. Ces processus sont proposés pour modéliser certaines données financières récoltées en plusieurs différentes dates. Notre étude, envisage les conditions sous lesquelles les modèles DS-AR obéirent certaines stabilités, nous donnons cependant des conditions suffisantes pour l'existence et l'unicité de solution stationnaire non anticipative. Une approche GMM (non reporté sur la version longue à cause de la limitation des contributions) et ses propriétés asymptotiques est proposée pour estimer les paramètres de modèle.

Prédiction de courbes de chlorophylle-a dans l'océan antarctique par régression linéaire fonctionnelle

Séverine Bayle (INRA), Pascal Monestiez (INRA), David Nerini (MIO)

Afin d'étudier les processus biogéochimiques de l'Océan Austral, des balises posées sur des éléphants de mer ont permis de récolter des profils de variables physiques (fluorimétrie, température, salinité, lumière) dans la zone sub-Antarctique autour des îles Kerguelen. Nous nous intéressons particulièrement aux données de fluorimétrie qui donne une estimation du contenu pigmentaire en chlorophylle-a contenue dans les organismes photosynthétiques, jouant un rôle essentiel dans le cycle océanique du carbone. Les profils verticaux de chlorophylle-a permettent d'évaluer la quantité de carbone par des mesures directes et d'obtenir ainsi une cartographie de cette zone de l'océan Antarctique. Notre objectif est d'utiliser les autres variables physiques plus faciles à mesurer que la chlorophylle-a (en particulier la luminosité) pour reconstruire de manière indirecte les profils de fluorescence. Dans un premier temps, un modèle linéaire fonctionnel a été utilisé, permettant de prédire des profils de chlorophylle-a à partir des dérivées de profils

de lumière en journée. Ce modèle intervient dans le cas particulier où la variable à prédire est une fonction et où la variable prédictive est également une courbe. Nous montrons que l'utilisation d'un tel modèle permet d'obtenir une bonne qualité de reconstruction pour accéder aux variations hautes fréquences des profils de chlorophylle-a à sub-mésoéchelle. Puis, afin de prévoir des profils de chlorophylle-a de nuit, nous utilisons une interpolation par krigeage fonctionnel.

Méthode de génération de données fictives spatialement et temporellement dépendantes. Contribution à la modélisation du péril sécheresse dans le cadre du régime d'indemnisation des catastrophes naturelles

Jean Ardon (Caisse centrale de réassurance)

Dans le cadre des études menées par CCR pour modéliser les événements catastrophes naturelles (inondations, séismes, sécheresse, . . .), nous proposons une méthode de génération de données fictives. Ces données sont dépendantes entre elles spatialement et temporellement. La méthode choisie consiste à étudier et modéliser séparément ces deux types de dépendances. La dépendance temporelle est modélisée site par site par des processus autorégressifs à coefficients potentiellement variables. La dépendance spatiale est modélisée par une copule. La difficulté à ajuster des copules non gaussiennes dans un problème de grande dimension nous a poussé à proposer une méthode de simulation de la copule empirique. Nous mettons cette méthode en oeuvre sur une variable appelée Soil Wetness Index (SWI) issue du modèle SIM de Météo France et obtenue de manière décadaire sur un maillage régulier France entière de résolution 8 km par 8 km.

Spatial modelling of plant diversity from high-throughput environmental DNA sequence data

Angelika Studeny (INRIA Rhone-Alpes), Florence Forbes (INRIA Rhone-Alpes), Eric Coissac (Laboratoire d'Ecologie Alpine, Universite Joseph Fourier Grenoble), Alain Viari (INRIA Rhone-Alpes), Lucie Zinger (Laboratoire d'Ecologie Alpine, Universite Joseph Fourier Grenoble), Céline Mercier (Laboratoire d'Ecologie Alpine, Universite Joseph Fourier Grenoble), Aurelie Bonin (Laboratoire d'Ecologie Alpine, Universite Joseph Fourier Grenoble), Frederic Boyer (Laboratoire d'Ecologie Alpine, Universite Joseph Fourier Grenoble), Pierre Taberlet (Laboratoire d'Ecologie Alpine, Universite Joseph Fourier Grenoble)

Cet article présente une approche statistique pour modèliser les correlations spatiales entre espèces dans un écosystème. L'originalité reside dans la particularité des données, génerées par des sequençages à haut-débit de l'AND environnemental d'échantillons de sol. Les données utilisées dans cet etude étaient recueillis à la station biologique CNRS des Nouragues, en Guyane Française. L'étude décrit les relations spatiales bivariées de ces données par un modèle linéaire de co-régionalisation separable où l'on estime un parameter de cross-correlation. Sur la base de cette estimation, nous visualisons le modèle de co-occurrences sous forme de graphes d'interactions. Les limites de cette approche sont discutés ainsi que les alternatives possible.

Statistique bayésienne, 11h00-12h20.

Modélisation des transferts foliaires et racinaires de métaux issus de particules fines (pm10) et de leur phytotoxicité

Thomas Puechlong (CNRS), Camille Dumat (ECOLAB), Christophe Laplanche (ECOLAB), Annabelle Austruy (ECOLAB), Tian Tian Xiong (ECOLAB)

L'accumulation de métaux lourd, issus des rejets industriels, dans les plantes comestibles entraîne des risques sanitaires. Les mécanismes de ce transfert-effet étant peu connus, une approche par modélisation est effectué. Cette modélisation du transfert et de l'effet des microparticules de métaux lourd est inspiré des techniques utilisées en Pharmacocinétique/Pharmacodynamique. La mise en place de ce modèle est réalisé dans un cadre bayésien, implémenté dans le logiciel libre OpenBUGS. Une description des avantages que propose cette approche est présentée, notamment en ce qui concerne la facilité d'implémentation d'un modèle non linéaire appliqué sur des données de structure hiérarchisée complexe.

Approche bayésienne pour le postraitementstatistique de prévisions d'ensemble

Eric Parent (AgroParisTech), Jacques Bernier (EDF), Vincent Fortin (Environnement Canada), Anne-Catherine Favre (ENSE3)

Les prévisions d'ensemble fournissent le moyen d'évaluer empiriquement l'incertitude des sortie d'un modèle déterministe de prévision. Ses conditions initiales sont perturbées et de nombreux runs du modèle sont lancés. On les appelle, dans le langage météo, les membres de la prévision d'ensemble. Ils peuvent également provenir de plusieurs modèles deterministes structurellement différents. Malheureusement, les systèmes actuels de prévision d'ensemble sous-estiment généralement la variabilité naturelle du phenomène. De nombreuses approches statistiques ont été proposées pour le post-traitement statistique des prévisions d'ensemble, notamment le Bayesian Model Averaging de Raftery, le Bayesian Processor of Outputs de Krzystzfowicz et le Best Member Dressing de Fortin. Dans cet exposé, le cadre statistique bayesien sera adopté pour toutes ces techniques, afin de discuter de leurs avantages et de leurs inconvenients dans un même cadre décisionnel formel. L'application porte sur la prévision d'ensemble de la température à l'aéroport de Québec durant l'été 2008.

Détection de matériaux nucléaires en temps réel par un algorithme smc

Aude Grelaud (Crest-Ensai), Rong Chen (Rutgers University), Minge Xie (Rutgers University), Priyam Mitra (Rutgers University)

Nous proposons un une procédure d'inférence statistique dédiée à l'analyse de données provenant d'un réseau de détecteurs mobiles dans le but de détecter la présence de matériaux nucléaires en temps réel. Ce travail est motivé par la nécessité d'avoir d'un système de détection fiable afin d'éviter les attaques nucléaires dans les grandes villes aux États-Unis, comme New-York par exemple. L'idée est d'installer des capteurs nucléaires et des appareils GPS dans un grand nombre de voitures circulant dans la zone d'intérêt, la lecture du capteur ainsi que la position GPS correspondant à chaque véhicule sont envoyées à un centre de surveillance toutes les 30 secondes et traitées immédiatement. Nous avons développé une méthodologie de détection en temps réel visant à détecter la présence d'une cible nucléaire et éventuellement évaluer son emplacement et le rayon d'émission de la bombe. Nous travaillons dans un Cadre bayésien et utilisons un

algorithme Monte-Carlo séquentiel (SMC) pour estimer les paramètres du modèle. Le critère de détection finale repose sur une combinaison des sorties de plusieurs chaînes. Une étude par simulation évaluer les performances de la méthode, que la bombe soit à l'arrêt ou en mouvement, et nous permet de fournir des recommandations sur la façon d'interpréter les résultats.

Apport de la connaissance experte pour la classification spatiale de communes des alpes françaises en deux zones climatiques

Aurore Lavigne (UMR 518 INRA/AgroParisTech), Nicolas Eckert (Instea), Eric Parent (AgroParisTech), Liliane Bel (UMR 518 INRA/AgroParisTech)

Dans un contexte de changement climatique analyser l'évolution de l'activité avalancheuse dans le passé est une première étape primordiale pour prédire correctement l'évolution future à long terme. Deux modèles ont été proposés pour modéliser l'évolution des fréquences d'avalanches dans les Alpes en tenant compte des deux zones climatiques présentes (une au nord et l'autre au sud). Les deux modèles utilisent les même données, mais leur approche tout comme leurs résultats diffèrent. Alors que le premier, considère les zones climatiques comme fixes, dans le second elles sont le résultat d'un modèle de classification spatial des communes sur la base de leur série temporelle. Nous proposons ici de concilier ces deux approches, en ajoutant au modèle de classification spatiale de la seconde approche, la connaissance d'expert sur la localisation des zones, sous la forme d'un prior, sous le paradigme bayésien. L'élicitation des paramètres est réalisé en proposant des cartes à l'expert, qu'il juge en faisant appel à des éléments climatiques mais aussi à la réponse de l'activité avalancheuse à des conditions météorologiques données. Cette approche nous permet de retrouver des classes spatialement cohérentes avec des évolutions temporelles distinctes : une zone nord-ouest d'activité décroissante et une zone sud-est d'activité croissante. En ajoutant l'altitude dans le modèle, nous montrons que ces zones résultent de l'interaction du climat local avec l'altitude.

Séries temporelles, 11h00-12h20.

Calcul et estimation de la matrice de variance asymptotique de modèles armafaibles multivariés

Yacouba Boubacar Maïnassara (Université de Franche-Comté)

Dans ce travail, nous considérons le problème de l'estimation de la matrice de variance asymptotique de l'estimateur du quasi-maximum de vraisemblance (QMV) des paramètres d'un modèle ARMA multivarié avec innovations linéaires non corrélées mais non nécessairement indépendantes (em i.e. VARMA faible). Dans un premier temps, nous proposons une expression des dérivées des résidus en fonction des résidus passés et des paramètres du modèle VARMA. Ceci nous permettra ensuite de donner une expression explicite de la variance asymptotique de l'estimateur du QMV, en fonction des paramètres des polynômes VAR et MA, et des moments d'ordre deux et quatre du bruit. Enfin nous en déduisons un estimateur convergent de la matrice de variance asymptotique.

Test de la causalité instantanée en présence d'une variance non conditionnelle non constante

Hamdi Raissi (IRMAR-INSA), Quentin Giai Gianetto (IRMAR-INSA)

Le problème du test de la causalité instantanée entre deux variables avec une structure de covariance non constante dans le temps est étudié. Nous montrons que les tests qui reposent sur l'hypothèse de stationnarité des processus ne sont pas valides dans notre cadre non standard : nous établissons que le test classique ne contrôle pas l'erreur de première espèce, alors que les tests reposant sur des corrections de type White (1980) ou HAC peuvent avoir une perte de puissance conséquente. Ainsi un test modifié de type bootstrap est proposé pour prendre en compte les covariances dépendant du temps.

Sélection de modèles autorégressifs par le critère ϕ_β

Freedath Djibril Moussa (Université de Fès), Abdelaziz El Matouat (Université du Havre), Hassania Hamzaoui (Université de Fès)

L'ordre d'un modèle autorégressif peut être estimé par plusieurs critères d'information dont le critère ϕ_β. Ce critère dépend d'un paramètre β dont le choix n'est pas spécifié en pratique. Dans ce travail, nous proposons une procédure de choix de β conduisant à une bonne estimation de l'ordre par le critère ϕ_β, pour un échantillon de taille finie. L'intérêt de la méthode est mis en évidence sur des données simulées.

The k-factors garma process with infinite variance innovations

Mor Ndongo (), Abdou Kâ Diongue ()

We develop the theory of k-factors Gegenbauer Autoregressive Moving Average (GARMA)process with infinite variance innovations. We establish conditions for existence and invertibility of the model. We also discuss the parameter estimation by using two methods. The first one is the Conditional Sum of Squares (CSS) approach and the second is the Markov Chains Monte Carlo (MCMC) Whittle method. For comparison purpose, Monte Carlo simulations are used to evaluate the finite sample performance of these estimation techniques.

Analyse de données 3, 11h00-12h00.

Visualisation et débruitage de données par ACP régularisée

Marie Verbanck (Agrocampus Ouest), Julie Josse (Agrocampus Ouest), François Husson (Agrocampus Ouest)

L'Analyse en Composantes Principales (ACP) est communément utilisée pour explorer et visualiser des données. Un modèle classique en ACP est le modèle à effets fixes (Caussinus, 1986) dans lequel les données sont générées comme une structure fixe ayant une représentation en rang inférieur, entachée d'erreur. Sous ce modèle, l'ACP ne fournit pas la meilleure estimation du signal sous-jacent en termes d'erreur quadratique moyenne. En suivant le même principe qu'en régression ridge, nous proposons une version régularisée de l'ACP qui revient à seuiller chaque valeur singulière par un terme qui correspond au ratio de la variance du signal sur la variance totale de la dimension associée. Le terme de régularisation

est dérivé analytiquement puis justifié comme un traitement bayésien du modèle à effets fixes. L'ACP régularisée est comparée à l'ACP classique ainsi qu'à une méthode dite de seuillage doux des valeurs singulières (Candès it et al., 2012), à la fois sur des jeux de données simulés et sur un jeu de données génomiques réel en grande dimension. L'ACP régularisée fournit des résultats prometteurs en termes d'estimation de la structure sous-jacente et permet ainsi d'obtenir des représentations graphiques plus proches de celles qui auraient été obtenues à partir du signal seulement.

Analyse en composantes principales sparse pour données multiblocs et extension à l'analyse des correspondances multiples sparse

Anne Bernard (CNAM de Paris), Gilbert Saporta (CNAM de Paris)

L'Analyse en Composantes Principales pour des données quantitatives, et l'Analyse des Correspondances Multiples pour des données qualitatives, sont des techniques de réduction de dimension bien connues. Cependant, les composantes obtenues à l'issue de ces méthodes sont des combinaisons de toutes les variables de départ, ce qui rend l'interprétation des résultats difficile pour des données de grande dimension. Pour pallier ces difficultés, nous proposons deux nouvelles méthodes de sélection de groupes de variables quantitatives et qualitatives : la 'Group Sparse Principal Component Analysis' (GSPCA) et l'ACM sparse, respectivement. La GSPCA est une extension de la SPCA-rSVD de Shen et Huang pour des données structurées par bloc. Elle utilise les liens entre l'ACP et la décomposition en valeurs singulières d'une matrice, afin d'extraire les composantes en résolvant un problème d'approximation de matrice de rang inférieur. Une contrainte de type 'Group Lasso' est introduite dans ce problème de minimisation afin d'obtenir des composantes étant combinaison d'un petit nombre de groupes de variables. Les loadings d'un groupe de variables sont mis à zéro permettant de réduire le nombre de variables sélectionnées. Puisque l'ACM est un cas particulier de l'ACP pour des blocs de variables indicatrices, l'ACM sparse est définie comme une extension de la GSPCA. Une application de cette méthode sera présentée sur un jeu de données bien connu comportant 27 races de chiens, décrites par 6 variables qualitatives.

Analyse en composantes principales partielle de données séquentielles d'espérance et de matrice de covariance variables dans le temps

Romain Bar (Université de Lorraine, IECL), Jean-Marie Monnez (Institut Elie Cartan de Lorraine)

On suppose que des vecteurs de données pouvant être de grande dimension et arrivant séquentiellement dans le temps sont des observations indépendantes d'un vecteur aléatoire d'espérance mathématique et de matrice de covariance variables dans le temps. On définit alors une méthode récursive d'estimation en ligne de vecteurs directeurs des r premiers axes principaux d'une analyse en composantes principales (ACP) partielle de ce vecteur aléatoire. On applique ensuite ce résultat au cas particulier de l'analyse canonique généralisée (ACG) partielle après avoir défini un processus d'approximation stochastique de type Robbins-Monro de l'inverse d'une matrice de covariance.

Extrêmes 1, 11h00-12h20.

Estimation bayésienne non-paramétrique de fonctions de survies pour des données d'avalanches censurées et sous-estimées

Ophélie Guin (AgroParisTech), Liliane Bel (AgroParisTech), Eric Parent (AgroParisTech), Nicolas Eckert (Instea)

La question des avalanches et en particulier de leurs distances d'arrêts est primordiale dans la gestion des risques. En effet, ces dernières peuvent entrainer des pertes humaines et matérielles considérables. Nous avons donc besoin de méthodes permettant de prédire leur fréquence et leur amplitude. Pour cela nous pouvons nous appuyer sur une base de données, la base EPA (Enquête Permanente sur les Avalanches). Cependant, si ces données nous permettent d'avoir des informations, elles ne sont fiables que sur une période assez récente ce qui rend une inférence sur certains paramètres nous intéressant difficile. Afin d'avoir un jeu de données plus conséquent, une idée est de faire appel à la dendrochronologie. Plus précisément il s'agit, après avoir échantillonné un certain nombre d'arbres, de déterminer en fonction des impacts trouvés sur ces derniers les années où il y a eu ou non des avalanches. L'avantage de telles données est que nous pouvons remonter plusieurs siècles dans le passé mais en contrepartie elles sont censurées et souvent sous-estimées. Le but de ce travail est donc d'être capable d'estimer des probabilités de dépassements d'avalanches pour différents seuils à partir de données dendrochronologiques. Pour cela nous proposons un modèle basé sur l'estimation bayésienne non-paramétrique d'une fonction de survie pour des données censurées et sous-estimées.

Estimation de l'indice des valeurs extrêmes conditionnel par un estimateur de Hill local lissé

Gilles Stupfler (Université d'Aix-Marseille), Laurent Gardes (Université de Strasbourg)

On s'intéresse à l'estimation de l'indice des valeurs extrêmes d'une fonction de répartition F à queue lourde. Ce paramètre contrôle le comportement de F à l'infini, ce qui implique que son estimation est nécessaire notamment lorsqu'on souhaite estimer des quantiles extrêmes. L'estimateur le plus utilisé de l'indice des valeurs extrêmes a été proposé par Hill en 1975. En pratique, la variable Y d'intérêt est souvent reliée à une covariable X. Dans ce cas, l'indice des valeurs extrêmes dépend de la valeur de la covariable et est appelé indice des valeurs extrêmes conditionnel. Dans la plupart des travaux sur ce sujet, son estimation a été considérée dans le cas où la covariable X n'est pas aléatoire. Le cas où X est aléatoire n'a été considéré que très récemment. Notre but est de proposer un estimateur qui est une adaptation de l'estimateur de Hill en présence d'une covariable aléatoire. Notons que la consistance uniforme en probabilité de l'estimateur proposé est établie alors que dans la plupart des travaux récents sur ce sujet, les auteurs ne démontrent que la convergence ponctuelle de leur procédé. On donne aussi la normalité asymptotique ponctuelle de l'estimateur proposé et ses performances sont évaluées sur simulations.

Estimation de quantiles extrêmes et probabilités d'évènements rares d'un processus stochastique

Gilles Durrieu (Université de Bretagne Sud), Ion Grama (Université de Bretagne Sud), Quang-Khoai Pham (Université de Bretagne Sud), Jean-Marie Tricot (Université de Bretagne Sud)

Nous considérons un processus stochastique à temps continu $X(t)$ à incréments indépendants de distribution F_t. Nous proposons un estimateur adaptatif non paramétrique de quantiles d'ordre élevé. L'idée de notre approche consiste à ajuster la queue de la distribution F_t, avec une distribution de Pareto de paramètre $\theta_{t,\tau}$ à partir d'un seuil τ. Le paramètre $\theta_{t,\tau}$ est estimé en utilisant un estimateur non paramétrique à noyau de taille de fenêtre h basé sur les observations plus grandes que τ. Sous certaines hypothèses de régularité, nous montrons que l'estimateur adaptatif proposé de $\theta_{t,\tau}$ est consistant et nous donnons sa vitesse de convergence. Nous proposons également une procédure de tests séquentiels pour déterminer le seuil τ et le paramètre h. Enfin, nous étudions les propriétés de cette méthode sur des simulations et sur des données réelles dans le but d'estimer des changements globaux (pollution, changement de température) et ainsi d'aider à la surveillance de systèmes aquatiques.

Estimation de mesures de risque extrêmes

Jonathan El Methni (Université Joseph Fourier), Stéphane Girard (INRIA Rhone-Alpes), Laurent Gardes (Université de Strasbourg)

Des mesures de risques classiques sont la Value-at-Risk, la Conditional Tail Expectation, la Conditional Value-at-Risk et la Conditional Tail Variance. En termes statistique, la Value-at-Risk est le quantile de la distribution des pertes. On s'intéresse aux propriétés de ces mesures de risque dans le cas de pertes extrêmes (où l'ordre du quantile n'est plus fixé mais tend vers 0) qu'on supposera modélisées par des lois à queues lourdes. On considèrera aussi ces mesures de risque avec la présence d'une covariable. On ajoute ainsi deux difficultés dans l'estimation de mesures de risque. Par conséquent, le but principal de cette communication est de proposer des estimateurs de toutes les mesures de risque énoncées ci-dessus pour des pertes extrêmes dans le cas de lois à queues lourdes en présence d'une covariable. On établira les propriétés asymptotiques de nos estimateurs et on illustrera leurs comportements sur des données simulées et sur un jeu de données pluviométriques.

Christian P. Robert, 14h00-15h00.

Relevant statistics for Bayesian model choice

Christian P. Robert (CEREMADE)

The choice of the summary statistics used in Bayesian inference and in particular in Approximate Bayesian Computation (ABC) algorithms has bearings on the validation of the resulting inference. Those statistics are nonetheless customarily used in ABC algorithms without consistency checks. We derive necessary and sufficient conditions on summary statistics for the corresponding Bayes factor to be convergent, namely to asymptotically select the true model. Those conditions, which amount to the expectations of the summary statistics to asymptotically differ under both models, can be exploited in ABC settings to infer whether or not a choice of summary statistics is appropriate, via a Monte Carlo validation.

Juan Cuesta-Albertos, 14h00-15h00.

The random projection method in functional data analysis

Juan A. Cuesta-Albertos (Universidad de Cantabria)

In this talk we present a review of some applications of the random projection method in functional data analysis. This procedure consists of, instead testing a given null hypothesis in a functional space, to test (the transformation of) this hypothesis on an one-dimensional randomly chosen projection, thus taking advantage of the multiple procedures which are available in the one dimensional case. We will see how this idea can be applied to some problems including two sample goodness of fit tests, normality tests and multiway ANOVA models including covariables. We will also analyze a test for the linear model in functional spaces.

Études de cas 3, 15h05-16h05.

Voiture personnelle ou voiture partagée ? les apports d'une modélisation logit multinomiale

Amandine Chevalier (IFPEN-BIPE-Mines ParisTech), Frédéric Lantz (IFPEN-IFP School)

Nous cherchons à expliquer les choix modaux, et plus particulièrement l'importance de la voiture, ainsi que le choix entre la voiture possédée et la voiture partagée (louée ou de quelqu'un d'autre). A partir de données d'enquête (Observatoire des Mobilités et Arbitrages Automobiles, BIPE, octobre 2010), nous nous sommes basés sur les activités des ménages, leur motorisation, ainsi que les variables socio-économiques pouvant impacter leurs choix modaux pour estimer un modèle logit multinomial. Sur notre échantillon, les variables les plus significatives sont les besoins de déplacement, la motorisation, l'âge, la zone de la commune de résidence et la situation de famille. Le modèle a notamment révélé l'importance du poids de la motorisation sur les choix modaux en général et l'usage de la voiture en particulier. Enfin nous avons pris en considération le coût de transport, ainsi que le taux d'effort qu'il représente en estimant un modèle logit conditionnel intégrant également des constantes spécifiques à chaque mode destinées à prendre en compte leurs caractéristiques (confort par exemple). Les résultats montrent, d'une part, que la voiture personnelle est préférée à tous les autres modes, et d'autre part, que la mobilité est un besoin arbitrable sur la base de son coût monétaire, et plus encore lorsque sa pression sur le budget augmente.

Comment estimer la probabilité de vente suivant l'ordre de la mise en page de chaque items sur le web ?

Virginie Braido (SAS)

Dans le domaine du e-marketing, la mise en page des produits en ligne et leur positionnement est une question récurrente : comment les disposer sur chaque page pour les rendre attractifs (aux acheteurs)et obtenir un revenu maximal (un nombre maximum de ventes.) sur le produit des ventes. L'idée principale consiste sur la base de données de recherches et de ventes d'hôtels, de définir à partir de leurs attributs un ordre d'affichage et suivant un modèle logistique une probabilité de vente. Ainsi sur la base des hôtels achetés faire une extrapolation pour les clients qui naviguent sur le site sans acheter et proposer les meilleurs offres en termes de prix/qualité dès les premières pages.

Big data - how big is big ?

Volker Kraft (SAS Institute / JMP Devision), Ian Cox (SAS Institute / JMP Devision)

The analysis of 'Big Data' is normally associated with large-scale hardware and expensive software. But multi-processor hardware and multi-threaded software are now also ubiquitous in personal computing devices. So, in 2013, just what is possible with a laptop that is within the budget of a serious analyst ? Most real-world analysis efforts that are given the 'big data' tag start by trying to understand issues of data quality, and by trying to identify interesting and potentially useful relationships and structure. Indeed, in such situations the data is often under scrutiny for the first time, so the investigation is necessarily somewhat open-ended. In addition, there are clear advantages to offering the analyst the capability to do significant data manipulation and data discovery without the need to call on an IT group. This presentation shows some real-world examples of what's currently possible using JMP, and positions this within the spectrum that covers the range from analysis of text book examples that are amenable to hand calculation to genuinely large-scale problems that can only be handled with investments in hardware and software of m or more.

Trafic routier (avec le soutien de l'AMIES), 15h05-16h05.

Complétion spatiale de données du trafic routier avec des processus Gaussiens sur des réseaux routiers

Jean-Noël Kien (IMT), Philippe Goudal (Mediamobile), Fabrice Gamboa (Institut Mathématiques de Toulouse), Jean-Michel Loubès (Institut de Mathématiques de Toulouse)

La société Mediamobile/V-Traffic, Filiale de TDF, est l'acteur français leader dans le domaine de la collecte, la diffusion de l'information trafic et dans la conception de services enrichis dédiés aux usagers de la route. Elle a mis en place, depuis 2005, en partenariat avec l'IMT, une stratégie de R et D. Une partie de son activité est de collecter le suivi GPS de flottes de véhicules pour fournir un flux de vitesse sur des points du réseau routier. Ce flux est partiel et nécessite une complétion afin de l'extrapoler sur l'ensemble du réseau. Des processus Gaussiens sur des réseaux routiers se révèlent être les outils adéquats en prenant en compte les corrélations spatiales. La complétion est alors modélisée comme étant la prédiction d'un tel processus qui aurait été observé partiellement sur le réseau routier. Dans un premier temps, nous réalisons l'évaluation théorique de ce type de processus sur des réseaux réels : comment simuler, estimer et prédire. Ensuite, nous utilisons cette modélisation sur données réelles en présentant les performances obtenues en validation par rapport à un ensemble de benchmarks et la façon dont les problèmes liés au cas concret ont été surmontés. Enfin, quelques étapes d'implémentation de cette modélisation sur des données de grande dimension posent des contraintes de faisabilité qui seront discutées.

Statistique et sécurité routière :l'observation de la conduite en situation naturelle

Guillaume Saint-Pierre (IFSTTAR), Cindie Andrieu (IMT Toulouse III)

Nous vivons une période de rupture, celle de l'acquisition de données tout azimut. Les déplacements quotidiens n'échappent pas à cette règle. L'exploitation des"traces" numériques enregistrées par des véhicules instrumentés s'est développée récemment en Europe, et ouvre de nouvelles possibilités pour

la recherche en sécurité routière. L'objectif de cet article est de montrer comment cette évolution méthodologique et technique offre un nouveau champ d'application de la statistique. Notre propos sera illustré sur deux exemples que nous avons développés dans des travaux récents.

Champs et processus gaussiens indexés par un graphe : application au trafic routier

Thibault Espinasse (ICJ-Lyon 1), Fabrice Gamboa (IMT Toulouse), Jean-Michel Loubès (IMT Toulouse)

Nous nous intéressons dans ce travail à un processus de vitesses sur le réseau routier. Le but est d'utiliser la structure du réseau (connue) pour spécifier la structure de dépendance spatiale du processus. Un champ aléatoire sur un graphe est un processus spatial indexé par les sommets de ce graphe, c'est à dire de la forme $(X_i)_{i\in G}$, où G désigne l'ensemble des sommets d'un graphe $\mathbf{G}$. Dans cette présentation, nous étendons quelques résultats classiques pour les séries chronologiques, au cas de processus gaussiens indexés par un graphe. Une série chronologique peut en effet être considérée comme un processus spatial indexé par le graphe $\mathbb{Z}$. En particulier, nous étudions les processus généraux type $ARMA$ sur un graphe, générés à l'aide de l'opérateur d'adjacence. L'objectif de cette étude est de fournir un algorithme simple permettant de prendre en oeuvre la structure spatiale du traffic routier (par exemple champ de vitesse) pour faire de la complétion de données manquantes, ou de la prédiction gaussienne.

Apprentissage 2, 15h05-16h05.

A multivariate HMM with dependency structure for the detection of copy number variations

Xiaoqiang Wang (INRA/AgroParisTech), Emilie Lebarbier (AgroParisTech), Julie Aubert (AgroParisTech/INRA), Stéphane Robin (AgroParisTech/INRA)

Les modèles de Markov caché fournissent un cadre statistique naturel pour la détection de variation du nombres de copies (CNV) en génomique. Dans cet article, nous considérons un modèle de Markov caché, prenant en compte simultanément plusieurs processus cachés dépendants. De plus, pour un grand nombre de séries, l'inférence par maximum de vraisemblance est intraitable. Dans ce cadre, nous présentons un algorithme d'inférence approché fondé sur une approche variationnelle pour déduire des variations structurales dans des génomes de plantes.

Une nouvelle approche multivariée pour la détection de défauts en semi-conducteur

Ali Hajj Hassan (STMicroelectronics), Sophie Lambert-Lacroix (UJF-Grenoble 1 / CNRS / UPMF / TIMC-IMAG), Francois Pasqualini (STMicroelectronics (Crolles2)/ Process Control)

Dans le domaine des semi-conducteurs, la détection de défauts dans les puces microélectroniques regroupées en wafers (galettes de silicium rangées latéralement) est une étape clé pour garantir aux clients le niveau de qualité exigé. En sortie de fabrication, le Test Paramétrique (PT) est réalisé. Il permet de détecter les wafers défectueux en utilisant des mesures de paramètres électriques statiques effectuées sur plusieurs sites du wafer. L'objectif de notre étude est de développer une approche statistique multivariée pour la détection de défauts au niveau du PT. Le but de cette approche est de maximiser le taux de détection des wafers défectueux tout en minimisant le taux de fausses alarmes. Notre approche est basée

sur les Machines à Vecteurs de Support à une classe (OC-SVM pour One Class Support Vector Machines), une extension du SVM original introduite dans le cadre d'une classification à une classe . Nous avons utilisé OC-SVM dans un scénario réel, où les étiquettes des wafers ne sont pas disponibles. Nous avons donc construit le modèle en utilisant tous les wafers disponibles. Nous avons ensuite amélioré la performance du modèle en introduction une nouvelle méthode de sélection des variables de type filtrage. Cette méthode ne nécessite pas aussi la connaissance des étiquettes des wafers. Notre approche a été validée sur un jeu de données réel en semi-conducteur.

Une comparaison de quatre techniques d'inférence des refusés dans le processus d'octroi de crédit

Asma Guizani (), Besma Souissi (), Salwa Ben Ammou (), Gilbert Saporta (CNAM de Paris)

Le credit scoring est une méthode d'évaluation du niveau du risque associé à un dossier de crédit potentiel. Cette méthode implique l'utilisation de différentes techniques statistiques pour aboutir à un modèle de scoring. Basé sur les caractéristiques du dossier du client, ce modèle estime le risque de crédit en prévoyant la solvabilité du demandeur de crédit. Les institutions financières utilisent ce modèle pour estimer la probabilité de défaut qui va être utilisée pour affecter chaque client à la catégorie qui lui correspond le mieux : bon payeur ou mauvais payeur. Les seules données disponibles pour construire le modèle de scoring sont les dossiers acceptés dont la variable à expliquer est connue (bon/mauvais). Ce modèle ne tient pas compte des demandeurs de crédit rejetés ce qui implique qu'on ne pourra pas estimer leur probabilité de défaut. Donc ce modèle donne des résultats biaisés à cause de la non-représentativité de l'échantillon (biais de sélection). L'inférence des refusés tente de remédier au problème de biais de sélection en réintégrant les dossiers refusés à l'échantillon initial. La repondération, le parceling, la classification mixte et la reclassification itérative sont les quatre techniques que nous appliquons dans notre cas. Nous avons comparé la performance des modèles de score obtenus par ces différentes techniques par leurs courbes ROC.

Statistique médicale 3, 15 :05-16 :05.

Evaluation de la complexite genomique comme facteur pronostique dans le cancer du sein

Eléonore Gravier (Institut Curie), Anne Vincent-Salomon (Institut Curie), Vanessa Benhamo (Institut Curie), Guillem Rigaill (Institut Curie)

Aujourd'hui, un des défis majeurs dans le traitement du cancer du sein de petite taille et sans envahissement ganglionnaire est d'identifier les patientes pour lesquelles une chimiothérapie adjuvante pourrait être évitée. Les paramètres cliniques usuels ne permettant pas d'identifier clairement les patientes à haut risque de rechute, nous proposons d'étudier la valeur pronostique de la complexité génomique de ces tumeurs par la technologie de puce Single Nucleotide Polymorphism (SNP). La complexité génomique d'une tumeur est ici définie comme le nombre de pertes et de gains du nombre de copies d'ADN sur l'ensemble de son génome. Les profils de nombre de copies d'ADN de 214 tumeurs du sein sont générés par la technologie de puce SNP. Ils sont ensuite segmentés afin d'estimer leur complexité génomique. Le pouvoir pronostique de ce paramètre est évalué sur une population cas / témoin constituée par 109 de ces 214 tumeurs puis validée sur la cohorte des 105 tumeurs restantes. Sur le jeu de validation, la complexité

génomique présente une valeur pronostique supérieure à celle apportée par les paramètres cliniques usuels utilisés en cancérologie (risque relatif de 3.49, p=0.012, modèle de Cox).

Calcul de taille d'échantillon dans le cadre de critères de jugements multiples avec un contrôle de la r-power et du gfwer

Philippe Delorme (Université de Montreal), Pierre Lafaye de Micheaux (Université de Montreal), Benoit Liquet (MRC, Université de Cambridge), Jérémie Riou (Danone Research)

A l'heure actuelle en recherche clinique, un nombre croissant de plans expérimentaux utilisent des critères de jugement principaux multiples. Dans ce contexte, l'étude sera perçue comme un succès par les promoteurs s'il est possible de rejeter au moins r hypothèses nulles parmi l'ensemble des m hypothèses nulles testées. Dans ce contexte, le statisticien se doit de prendre en compte la multiplicité induite par cette pratique. Pour cela, il peut s'appuyer sur une littérature abondante sur les méthodes d'analyse ainsi que pour le calcul de taille d'échantillon quand r = 1 ou r = m. L'objectif de notre travail consiste ici à développer une méthodologie permettant le calcul de tailles d'échantillon pour les procédures les plus couramment utilisées en recherche clinique, à savoir les procédures 'single-step' et 'step-wise', et ce quelle que soit la valeur de r. Afin de valider et d'illustrer l'intérêt de ce travail, nous présentons une étude par simulations ainsi qu'une application réelle dans le cadre d'un essai clinique.

De l'importance de la méthode statistique pour évaluer la reproductibilité d'un score médical

Caroline Le Gall (Methodomics), Pierre-Antoine Gourraud (UCSF School of Medicine University of California)

L'indice de localisation et de gravité du psoriasis (score de PASI : Psoriasis Area and Severity Index) est l'un des scores de sévérité les plus utilisés dans les essais cliniques pour évaluer les patients, décider de l'usage de biothérapies et en mesurer l'efficacité. Malgré son statut de score de référence dans le psoriasis, et à l'image d'autres scores de sévérité des maladies multifactorielles, le degré de validation statistique de ce score est mal compris. Sa reproductibilité généralement évaluée entre praticiens du même service d'un CHU est souvent estimée par de simples corrélations et sans prendre en compte le domaine de valeurs cruciales à la pratique clinique. L'objectif de cette communication est de montrer que l'usage des coefficients de corrélation classiques entraîne une surestimation de la reproductibilité du score de PASI. La surestimation est d'autant plus problématique qu'elle apparaît sur la plage de valeurs critiques pour la décision de l'usage de biothérapies. Il est préférable d' utiliser une approche par modélisation mixte en estimant le coefficient de corrélation intra-classe (ICC). Notre étude illustre la difficile adaptation des modélisations mathématiques à la réalité des problématiques biomédicales et la nécessité d'utiliser les bons indicateurs statistiques. Elle ouvre la perspective de construire un nouveau score plus robuste à la diversité des présentations cliniques du psoriasis et mieux adapté à la prise de décision concernant les traitements.

Régression 2, 15h05-16h05.

Tests d'hypothèses linéaires dans un modèle de régression non paramétrique

Zaher Mohdeb (Université de Constantine 1)

Une procédure de test d'hypothèse linéaire sur la fonction de régression f dans un modèle de régression non paramétrique est proposée. Plus précisément, on teste l'hypothèse que f est un élément de E, où E est un espace vectoriel de dimension finie. Nous proposons une statistique de test basée sur une approximation de la distance dans l'espace L2. La statistique de test est obtenue en estimant cette distance par la distance empirique des observations à l'espace E. En supposant que les fonctions considérées sont höldériennes d'ordre plus grand que 1/2 et on obtient le comportement asymptotique de la statistique de test proposé, on a donc ainsi le niveau et la puissance asymptotique du test.

Modèle de régression pour des probabilités cumulées en présence de risques concurrents et de censure par intervalles

Pierre Joly (Inserm, Isped, Univ Bordeaux), Paul Blanche (Inserm, Isped, Univ Bordeaux), Célia Touraine (Inserm, Isped, Univ Bordeaux)

En analyse des données de survie, la présence de risques concurrents est commune. Dans ce cas, avec un modèle multi-états, on modélise souvent l'influence de variables explicatives sur les intensités de transition qui représentent les 'risques instantanés' de subir un évènement, par des modèles de régression à intensités de transition proportionnelles. Cependant, il peut aussi être intéressant d'étudier les facteurs de risque qui influent sur la 'probabilité cumulée' de subir un évènement, en particulier pour faire du pronostic. Nous proposons ici d'utiliser un modèle de régression qui permet d'estimer directement l'effet des variables explicatives sur les probabilités cumulées. Ceci permet une interprétation simple de l'influence des facteurs de risque sur la probabilité de subir un évènement. Nous nous plaçons dans le cadre d'un modèle illness-death avec des données tronquées à gauche et censurées par intervalles. Une application sera présentée sur les données d'une cohorte sur le vieillissement cérébral.

Méthodes de construction d'un groupe de contrôle pour un groupe traité

Leslie Hatton (EDF et Agrocampus Ouest), Philippe Charpentier (EDF), Eric Matzner-Lober (Agrocampus Ouest et Université Rennes 2)

Pour répondre aux enjeux environnementaux et réduire la pointe de consommation électrique, EDF est amené à activer différents mécanismes de Gestion Active de la Demande sur le marché résidentiel, consistant à piloter la charge des clients. Pour optimiser son parc de production et assurer l'équilibre Offre-Demande, EDF doit quantifier cette réduction de consommation appelée effacement. Pour cela, on estime la consommation moyenne s'il n'y avait pas eu d'effacement, la baseline. L'effacement s'obtient par différence entre la baseline et la courbe réalisée. Parmi les méthodes proposées dans la littérature et testées sur des données réelles, les résultats démontrent l'avantage des méthodes d'estimation de baseline utilisant un groupe de contrôle. Sans contraintes opérationnelles, la solution optimale est de répartir les clients recrutés aléatoirement entre le groupe traité et le groupe de contrôle. Malheureusement, ce type d'expérience n'est pas adapté au contexte du groupe EDF. Nous disposons toutefois, sur la même région géographique, d'un ensemble de courbes individuelles de clients (non traités) issus d'un panel représentatif

du portefeuille résidentiel ainsi que des variables explicatives communes au groupe traité. Nous souhaitons sélectionner parmi cet ensemble, des individus pour obtenir un groupe de contrôle. Nous proposons alors trois méthodes : un algorithme séquentiel et les deux méthodes de régression sous contraintes, Ridge et Lasso.

Statistique et sciences humaines, 16h25-17h25.

Un modèle de graphes aléatoires pour l'analyse d'un réseau ecclésiastique dans la Gaule mérovingienne

Yacine Jernite (New York University), Pierre Latouche (SAMM, Université Paris 1), Charles Bouveyron (SAMM, Université Paris 1), Patrick Rivera (LAMOP, Université Paris 1), Laurent Jégou (LAMOP, Université Paris 1), Stéphane Lamassé (LAMOP, Université Paris 1)

Au cours des deux dernières décennies, de nombreux modèles de graphes aléatoires ont été proposées pour extraire des connaissances à partir des réseaux. La plupart d'entre eux recherchent des collectivités ou plus généralement des groupes de sommets avec des profils de connexion homogènes. Alors que les premiers modèles ont mis l'accent sur des réseaux avec des arêtes binaires, leurs extensions permettent désormais de traiter des réseaux valués. Ce travail a été motivé par la nécessité d'analyser un réseau historique où une partition des sommets est donné et dont les arêtes sont de type discrêt. La partition connue des sommets est considérée comme une décomposition d'un réseau en sous-graphes que nous proposons de modéliser en utilisant un modèle stochastique avec des groupes latents inconnus. Chaque sous-graphe a son propre vecteur de mélange et voit ses sommets associés aux clusters. Les sommets sont alors connectés avec une probabilité dépendant des sous-graphes seulement, alors que les types des arêtes sont supposés être échantillonnés à partir des groupes latents. L'algorithme variational Bayes EM est proposé pour l'inférence ainsi qu'un critère de sélection de modèles pour l'estimation du nombre de cluster. La méthodologie proposée a été appliquée à un réseau ecclésiastique dans la Gaule mérovingienne.

Inférence de dates d'activité à partir d'un réseau d'interactions datées

Pierre Latouche (SAMM, Université Paris 1), Fabrice Rossi (SAMM, Université Paris 1)

Nous proposons dans cet article un nouveau modèle génératif pour les graphes qui s'appuie sur une approche à espace latent pour expliquer un ensemble d'interactions datée. L'objectif du modèle est de fournir des estimations globales pour les dates d'activité d'un ensemble d'acteurs dont les dates d'interaction sont connues avec une précision raisonnable, par opposition à une estimation locale simple. Nous montrons sur des données artificielles que le modèle proposé produit une meilleure estimation des dates d'activité que les moyennes locales si le réseau étudié est suffisamment dense.

Estimation de la loi a posteriori de la fonction graphon d'un w-graphe. Application au réseau de la blogosphere politique française

Pierre Latouche (SAMM, Université Paris 1), Stéphane Robin (AgroParisTech/INRA)

Les réseaux sont aujourd'hui utilisés dans de nombreux domaines scientifiques afin de représenter les interactions entre entités d'intérêt. Depuis les premiers travaux de Moreno en 1934, de nombreux modèles

de graphe aléatoire ont été proposés dans le but d'extraire des informations pertinences à partir de ces données structurées. Le modèle à blocs stochastiques, stochastic block model (SBM) en anglais, permet par exemple de rechercher des groupes de noeuds ayant des profiles de connexions homogènes. Nous nous intéressons ici au modèle de W-graphe qui présente l'intérêt de généraliser la plupart des modèles de graphe aléatoire existants mais pour lequel peu de méthodes existes pour réaliser l'inférence du modèle sur données réelles. Dans un premier temps, nous rappelons comment le modèle SBM peut être représenté sous la forme d'un W-graphe avec une fonction graphon bloc-constante. A l'aide d'un algorithme de type variational Bayes expectation maximization, nous approchons ensuite la loi a posteriori des paramètres d'un modèle SBM et nous montrons comment l'incertitude sur les paramètres, caractérisée par cette approximation variationnelle, peut être intégrée de manière analytique afin d'obtenir une estimation de la loi a posteriori de la fonction graphon du W-graphe. Ces travaux sont testés sur données simulées et sur un extrait du réseau de la blogosphere politique française.

Études de cas 4, 16h25-17h25.

Politiques céréalières en Algérie

Hanya Kherchi Medjden (LASAP-ENSSEA), Bahia Bouchafaa (LASAP- ENSSEA)

La filière blés est une filière stratégique pour l'Algérie, elle est considérée comme le fer de lance pour l'industrie agro-alimentaire mais elle reste toujours sous la proposition de reconstitution pour les raisons suivantes : la production, reste insuffisante pour satisfaire la demande nationale ; par ailleur, la consommation de blé par la population ne cesse d'augmenter. Cependant, La production des blés en Algérie présente une caractéristique fondamentale depuis l'indépendance à travers l'extrême variabilité du volume des récoltes. Cette particularité témoigne d'une maîtrise insuffisante de cette culture et de l'indice des aléas climatiques. Cette production est conduite en extensif et elle est à caractère essentiellement pluvial. La demande en blé est couverte, en partie par la production nationale qui oscille, selon les campagnes entre 0,9 et 4,9 millions de tonnes tandis que le reste est satisfait par les importations. Les prix minimums garantis (PMG) à la production payés aux producteurs par sont fixés annuellement par décret, et sont souvent peu incitatifs pour les producteurs. Depuis 1983 Les ajustements à la hausse des PMG se sont accélérés et permettent maintenant au prix national d'être notablement supérieur au prix mondial. Dans ce travail, nous proposons de faire une analyse statistique de la politique céréalière en Algérie.

Les logiciels ibm spss pour le marketing prédictif

Serge Retkowsky (IBM), Peggy Vaugard (IBM)

L'analyse prédictive avec IBM SPSS offre aux entreprises les moyens d'affiner leurs stratégies marketing, de mieux cibler leurs clients et d'être plus réactif. Quel message adresser à tel client à quel moment ? Quelle offre lui proposer et par quel canal ? Quelle est la probabilité que tel client passe une nouvelle commande ? Historique des commandes, comportements d'achat, données de navigation sur les sites web, caractéristiques sociodémographiques et autres informations relevant du service clients (réclamations...) sont étudiées à la loupe à l'aide des logiciels IBM SPSS pour fidéliser les clients et obtenir rapidement de meilleurs résultats.

Méthodologie relsys® de calcul des paramètres de fiabilité d'un système par le calcul scientifique

Jérome de Reffye (Pi-Ramses)

Un modèle de faisabilité de la méthodologie RELSYS*copyright* est développé. Le but principal de cette présentation est de prouver que cette voie pour résoudre tous les calculs des paramètres de fiabilité d'un système est la plus complète qu'il soit possible de réaliser. On obtient une synthèse des précédentes méthodes et le problème du calcul de la fiabilité dynamique est résolu. Par contre il faudra fournir un effort sur la qualité des données d'alimentation.

Apprentissage 3, 16h25-17h25.

Un regret unifié pour l'optimisation convexe en ligne

Pierre Gaillard (EDF, ENS Paris), Nicolò Cesa-Bianchi (Università degli studi di Milano), Gábor Lugosi (Pompeu Fabra University), Gilles Stoltz (ENS Paris, CNRS)

Le problème d'optimisation convexe en ligne consiste à prévoir séquentiellement les valeurs d'une certaine suite à valeur dans un ensemble convexe. L'objectif du joueur est de s'assurer une performance similaire à la meilleure stratégie (oracle) d'un ensemble de stratégies de référence. Plus l'ensemble de référence est grand, plus l'erreur de l'oracle est potentiellement faible, mais plus le joueur a du mal à s'en rapprocher et plus son regret sera grand. C'est le compromis entre erreur d'estimation et erreur d'approximation, que l'on retrouve régulièrement en statistiques. Dans la littérature, l'objectif initial n'est pas toujours la minimisation de la perte cumulée, de plus le compromis biais-variance est géré de façons variées et différents ensembles de références sont considérés. Cela a mené à diverses notions de regret, comme le regret en ruptures, le regret adaptatif, ou le regret escompté. Cet exposé propose une notion de regret plus puissante, qui tend à unifier les trois précédentes. Nous en profiterons pour montrer que des algorithmes comme celui d'Herbster et Warmuth (1998) ou de Zinkevich (2003) permettent d'obtenir des bornes satisfaisantes sur ce regret généralisé.

Learning when high an low accuracy observations are available

Federico Zertuche (Université Joseph Fourier, Grenoble)

The problem studied is the prediction of an output given an input by learning from different types of observations produced by an experiment with adjustable precision. Under Gaussian hypothesis, we study the effect of the relationship between the different precisions on the regression function. Then, we present a model in which this relationship is estimated by local polynomials. We present an alternative model that does not use the Gaussina hypothesis based on adaptative spline-wavelets.

Sur l'estimateur des plus proches voisins mutuels

Arnaud Guyader (Université rennes 2), Nicolas Hengartner (Los Alamos National Laboratory)

On s'intéresse ici à l'estimation de la fonction de régression $m(x)$ associée à un couple aléatoire (X, Y), à partir d'un échantillon i.i.d. $D_n = (X_i, Y_i)_{1 \leq i \leq n}$ de même loi que (X, Y). Dans ce contexte, l'estimateur

des plus proches voisins mutuels au point x consiste à identifier les k plus proches voisins de x, puis à ne conserver que ceux dont x est lui-même l'un des k plus proches voisins, et enfin à calculer la moyenne pour prédire. Nous montrons dans ce travail que cet estimateur est non seulement consistant mais de vitesse de convergence optimale.

Génétique/Génomique 1, 16h25-17h25.

Impact du choix de la méthode de prédiction d'identité entre les gènes sur la précision en cartographie de QTL

Laval Jacquin (SAGA - INRA Toulouse), Jean-Michel Elsen (INRA UR0631), Hélène Gilbert (INRA-LGC)

On dit que deux segments chromosomiques sont identiques par descendance (IBD) s'ils ont été hérités d'un même chromosome ancestral. Plusieurs méthodes ont été développées, durant ces dernières années, afin de calculer des prédictions d'états IBD pour des couples de segments chromosomiques. Ces méthodes sont utiles pour localiser les zones chromosomiques ayant un effet sur la variabilité d'une variable réponse mesurée chez des individus. Le problème statistique à résoudre est donc un problème inverse. La zone ayant un effet sur la variabilité de la variable réponse est aussi connue sous le nom de "Quantitative Trait Locus' (QTL). La variable réponse quant à elle est une quantité réelle mesurable, aussi appelé phénotype, associée aux individus. Un QTL peut être détecté s'il est en association non-aléatoire avec une ou des zones chromosomiques testées. Cette association non-aléatoire est appelée déséquilibre de liaison (DL). L'objectif de cette communication est de montrer que les meilleures méthodes de prédiction d'états IBD, pour la cartographie de QTL, sont celles qui exploitent le mieux le DL entre les différentes zones testées et le QTL. Afin d'illustrer ce propos on examinera des exemples d'applications sur un jeu de données réelles.

Distances génomiques

Clément Carré (INRA), Eduardo Manfredi (INRA), Fabrice Gamboa (Institut Mathématiques de Toulouse)

Ce travail rentre dans le cadre de la prédiction de phénotypes à partir de données génomiques des animaux de rente. Pour des raisons liées au nombre croissant de données accessibles par individu, ainsi que la combinatoire explosive des possibles interactions génétiques, nous avons choisi d'explorer un modèle non paramétrique : la régression à noyaux. Pour le calcul de cette régression, il est nécessaire de définir une distance (au sens mathématique) génomique entre individus. Cette étude porte sur la définition de 8 distances, et de leur comparaison en termes d'efficacité à prédire des jeux de données simulées. Une semi-distance (corrélation) produit les meilleurs résultats et laisse espérer des améliorations par l'étude future de distances de la même famille.

Approche par groupe de gènes pour les données longitudinales d'expression génique avec une application dans un essai vaccinal contre le VIH

Boris Hejblum (Inserm U897, ISPED, VRI), Rodolphe Thiébaut (Inserm U897, ISPED, VRI)

Les mesures répétées d'expression génique sont de plus en plus courantes. Appliquée dans des études transversales, l'analyse par groupe de gènes a démontré sa puissance en termes de sensibilité et d'interprétation. Nous étendons ici cet outil aux données longitudinales d'expression génique, en tenant compte de la possible hétérogénéité des groupes de gènes. Notre approche, TcGSA (Time-course Gene Set Analysis), teste si les gènes appartenant à un groupe donné sont stables au cours du temps, grâce aux estimations du maximum de vraisemblance. Cette approche peut s'appliquer dans le cas de données déséquilibrées dues à des données manquantes aléatoirement. Une classification non supervisee des dynamiques estimées pour les gènes appartenant à un groupe de gènes est ensuite réalisée, afin d'y exhiber les principales tendances. Nous avons applique TcGSA dans l'essai DALIA-1, un essai de vaccin thérapeutique contre le VIH au cours duquel des patients infectes par le VIH-1 ont reçu un vaccin à base de cellules dendritiques, avant d'arrêter temporairement leur traitement antirétroviral. La méthode TcGSA, appliquée à 260 modules fonctionnels, a nettement amélioré les résultats par rapport à une analyse gène-par-gène de l'expression différentielle au cours de la phase de vaccination (qui n'avait révèlé aucun changement significatif après correction pour la multiplicité des tests). Après l'interruption du traitement antirétroviral (une perturbation importante pour le système immunitaire), TcGSA détecte un changement significatif de nombreux modules, notamment de ceux lies à la production d'interferon, qui étaient attendus.

ENBIS (avec le soutien de l'AMIES), 16h25-17h25.

Modélisation de simulateurs industriels multifidélités » : apport d'une approche bayésienne globale

Céline Helbert (Institut Camille Jordan)

Les simulateurs industriels sont aujourd'hui largement utilisés pour représenter des systèmes physiques complexes. La compréhension du phénomène réel passe alors par l'étude de ces simulateurs. La difficulté vient du coût prohibitif en temps calcul. Dans notre travail, nous nous intéressons à des simulateurs 'multifidélités'. Des simulations peu coûteuses en temps calcul mais potentiellement moins représentatives du phénomène modélisé peuvent être obtenues en dégradant le schéma numérique. L'objectif de cet exposé est de présenter deux modélisations différentes qui permettent d'intégrer plusieurs niveaux de fidélité de la réponse étudiée. Nous commencerons par montrer comment l'information apportée par des simulations dégradées peut être intégrée au sein de la loi a priori du krigeage. L'approche bayésienne est alors plus pertinente que l'approche classique, notamment pour des études de propagation d'incertitude, d'analyse de sensibilité ou d'optimisation globale basée sur la variance et dans un contexte 'multifidélités' . Ensuite, nous présenterons l'approche "multifidélités" alternative basée sur le modèle de cokrigeage de Kennedy et O'Hagan et largement développée ces dernières années. Quelques pistes d'amélioration de ce modèle seront alors proposées. Enfin nous comparerons ces deux modélisations "multifidélités" sur une étude de cas issue du domaine pétrolier.

Process monitoring of large scale systems

Marco Seabra dos Reis (University of Coimbra)

Current processes are characterized by a large amount of variables that evolve along time in a dynamical way. Monitoring this high-dimensional space is a challenge, as not only the relationships between variables must be considered, but also the dynamics of the system. Several approaches have been developed for handling this reality, and in this talk we will address representatives of two such classes of methods. In one case, one tries to efficiently model the cross- and auto-correlation present in data using a parsimonious dynamical latent variable model. This will provide the basic model for deriving the monitoring procedure. In the second case, one adopts a multiscale framework for process monitoring that takes advantage of several features of wavelet-like transforms, such as their decorrelation and energy compaction ability.

Analyse de l'exposition aux ondes électro-magnétiques via la dosimétrie stochastique

Joe Wiart (Orange Labs), Emmanuelle Conil (Orange Labs), Abdelhamid Hadjem (Orange Labs), Azedine Gati (Orange Labs), Nadège Varsier (Orange Labs)

L'utilisation intensive des moyens de communication sans fil s'est accompagnée d'une perception de risque dans la population. Les technologies et usages évoluant rapidement, la quantification de l'exposition est devenue un axe de recherche important. Les études qui ont été menées ces 20 dernières années ont d'abord analysé et spécifié les méthodes permettant de garantir la conformité aux limites des systèmes mis sur le marché. Pour cela des configurations "pire cas" ont été mises en place. Prenant avantage des progrès dans le domaine du calcul numériques, la méthode des différences finies dans le domaine temporel (FDTD), a été de plus en plus utilisée pour évaluer l'exposition des personnes. En dépit des progrès actuels, les temps de calcul restent importants et ne permettent pas l'utilisation de méthodes telle que Monté Carlo pour analyser l'influence de la variabilité des usages et des sources sur l'exposition. Pour répondre à ce challenge une approche nouvelle basée sur des modèles "réduits" est nécessaire. Dans le domaine de l'électromagnétisme l'analyse de la propagation des incertitudes dans les codes de calcul est un domaine d'investigation relativement récent. Les analogies existantes entre l'expansion polynomiale et l'analyse modale ont favorisé, dans cette communauté l'utilisation du chaos polynomial. Dans ce papier une approche utilisant la troncature de l'expansion, une estimation des coefficients par régression, une qualification des modèles via l'erreur quadratique moyenne et une sélection des polynômes les plus influents est utilisée pour analyser l'influence de la position du téléphone près de la tête.

Assemblée générale de la SFdS, 17h30-18h30.

Résumés du mercredi 29 mai 2013

Christian Francq, 08h30-09h30.

Risk-parameter estimation in volatility models

Christian Francq (CREST (CNRS)), Jean-Michel Zakoïan (CREST (CNRS))

Nous introduisons le concept de paramètre de risque dans un modèle de volatilité conditionnelle et nous proposons plusieurs estimateurs de ce paramètre. Pour une mesure de risque donnée r, le paramètre de risque conditionnel est une fonction du coefficient de la volatilité et du risque du processus des innovations. Une méthode en deux étapes permet d'estimer successivement ces quantités. Nous proposons également une méthode alternative en une seule étape, qui est fondée sur une reparamétrisation du modèle et sur l'utilisation d'un estimateur du quasi-maximum de vraisemblance non gaussien. La théorie asymptotique, établie pour des mesures de risque générales, comprenant notamment la valeur à risque (VaR) ou l'ES (Expected Shortfall), permet de quantifier le risque d'estimation. Dans le cas GARCH standard, la comparaison asymptotique montre la supériorité de la méthode en une étape quand les innovations sont à queues lourdes. Les résultats théoriques sont illustrés sur des séries de rendements d'indices boursiers.

Johan Segers, 08h30-09h30.

Semiparametric Gaussian copula models : geometry and efficient rank-based estimation

Johan Segers (Université catholique de Louvain), Ramon van den Akker (Tilburg University), Bas Werker (Tilburg University)

Un estimateur simple, basé sur les rangs, et semiparamétriquement efficace est proposé pour des modèles de copules gaussiennes multivariées à marges inconnues et des structures de corrélation générales. Une représentation algébrique des sous-espaces pertinents de l'espace tangentiel est construite permettant d'étudier facilement l'adaptivité par rapport aux marges inconnues et l'efficacité de l'estimateur de maximum de pseudo-vraisemblance. Quelques exemples bien connus sont étudiés en détail : matrices de corrélation circulaires, modèles à facteur, et matrices de Toeplitz, comprenant des structures à moyenne mobile et des modèles autorégressifs. Pour certains de ces exemples, l'efficacité asymptotique relative de l'estimateur de maximum de pseudo-vraisemblance est inférieure à 20%. Ces résultats sont confirmés sur des échantillons de taille finie par des simulations de Monte Carlo.

Finance 2, 09h35-10h35.

Value at risk confidence intervals estimation

Fedya Telmoudi (ISG Tunis), Mohamed El Ghourabi (ISG Tunis)

Le but de cet article est de construire des intervalles de confiance pour la Valeur en Risque (VaR). Une méthode hybride est développée et appliquée au modèle GJR-GARCH (1,1) avec des innovations à queue lourd, combinée avec la méthode régression à vecteur de support (SVR) et l'estimateur de Dekker pour l'estimation des quantiles extrêmes. La distribution limite est étudiée, où la méthode d'approximation normale est appliquée. Une analyse de la performance est établit en comparant l'utilisation de SVR et le maximum de vraisemblance pour la modélisation du modèle GJR-GARCH (1,1).

Problème d'estimation non-paramétrique pour des processus stochastiques périodiques

Dominique Dehay (Université Rennes 2), Khalil El Waled (Université Rennes 2)

Ce travail porte sur un modèle d'équation stochastique différentielle linéaire où les coefficients du drift et de la diffusion sont des fonctions de IR dans IR ,continues, périodiques de période P. Dans un premier temps nous nous intéressons aux propriétés de la solution de cette équation. En suite nous nous focalisons sur l'estimation non-paramétrique du coefficient du drift sur une trajectoire continue du processus observée sur un intervalle fini [0, T]. Nous allons construire un estimateur à noyau périodique et nous établirons sa consistance : convergence en moyenne quadratique et la convergence presque sûre quand T tend vers l'infini, ainsi que la normalité asymptotique.

Modèle hybride pour l'évaluation et la couverture des produits dérivés de crédit avec des paramètres stochastiques

Majdi Elhiwi (Université de SFAX, Tunisie)

Dans le cadre de l'évaluation des produits financiers qui peuvent subir des risques de crédit, nous avons construit un model nommé IDBM 'modèle d'intensité de défaut de la barrière', basé sur la probabilité conditionnelle de survie appelée aussi fonction du hasard associée à la barrière. Notre modèle s'inscrit dans l'approche hybride qui combine les modèles classiques structurels et ceux à forme réduite. Le IDBM a réussi de prouver la décomposition de Doob-Meyer du processus de défaut de la barrière. En effet, nous avons présenté le processus de défaut associé à la barrière comme la somme de son compensateur, qu'est un processus prévisible croissant, et une martingale relativement à la filtration construite et qui rend la barrière aléatoire un temps d'arrêt. La puissance de notre résultat de point de vue probabiliste s'accroit puisque notre modèle repose uniquement sur les données de marché observables. Cette caractérisation permet au modèle IDBM de ne plus s'intéresser à l' instant de défaut lui même, comme celui le cas des modèles classiques, mais à sa probabilité instantanée modélisée par l'intensité de défaut de la barrière. Par ailleurs, le modèle IDBM peut être facilement généraliser pour évaluer plusieurs produits dérivés comme les (CDS) Credit Default Swaps et les CDO Collateralized Debt Obligations.

Statistique d'enquête 1, 09h35-10h35.

Modélisation de données d'enquêtes par une approche basée sur la vraisemblance empirique

Yves Berger (University of Southampton), Valentin Patilea (CREST- Ensai)

Berger et De la Riva Torres (2012) ont proposé une approche basée sur la vraisemblance pour des plans d'échantillonnage à probabilités inégales. Nous proposons d'adapter l'approche proposée par Berger et De La Riva Torres (2012) pour l'estimation de paramètres définis par des équations estimantes conditionnelles. L'approche proposée permet de tenir compte de l'effet du plan d'échan-til-lon-nage pour l'estimation ponctuelle et l'estimation d'intervalles de confiance. L'approche proposée dépend de l'information du plan d'échantillonnage ; c'est-à-dire les probabilités d'inclusion d'ordre un et la stratification. Il offre également une justification de type vraisemblance pour le calage. L'approche proposée ne repose pas sur des estimations de variance, des effets d'échantillonnage, des méthodes de ré-échantillonnage ou de linéarisation, même lorsque le paramètre d'intérêt n'est pas linéaire.

Calage sur information auxiliaire incertaine : proposition d'algorithme de redressement ridge

Flavien Alleaume (Médiamétrie), Lorie Dudoignon (Médiamétrie)

Le redressement d'échantillons s'opère généralement par une méthode de calage sur marges et à l'aide d'algorithmes qui proposent le plus souvent un calage exact de l'échantillon sur les marges avec lesquelles on souhaite être en adéquation. Si, les informations auxiliaires proviennent d'une enquête exhaustive, cette approche se justifie pleinement. Or dans la pratique, on est parfois amené à estimer les marges à partir d'une enquête par sondage. Dans ce contexte, il peut être intéressant d'autoriser un relâchement des contraintes, i.e. de permettre une tolérance dans le respect des marges associées aux variables de calage. L'adaptation de la logique de la régression ridge aux redressements d'échantillons permet de répondre à cette problématique. Nous proposons ici un algorithme qui fait la synthèse entre les approches de Beaumont et Bocci et celle de Singh et Mohl. Cette méthode est appliquée au Panel Multi-Écrans de Médiamétrie, nouveau dispositif de mesure d'audience de la télévision et d'Internet. Une discussion en termes de performance des différents algorithmes testés est proposée.

Risque d'amplification de biais de l'estimateur par calage généralisé en présence de non-réponse

Eric Lesage (CREST-ENSAI), David Haziza (Université Montréal et CREST-ENSAI)

Dans cette présentation, il sera question de l'utilisation du calage généralisé comme méthode de pondération en une étape. Le calage poursuit alors simultanément trois objectifs : réduire le biais de non-réponse, assurer la cohérence entre les estimations de l'enquête et les totaux connus sur la population et, si possible, réduire la variance. Nous examinons les propriétés de l'estimateur par calage généralisé dans le cas où les variables instrumentales (variables explicatives de la probabilité de répondre) ne sont disponibles que pour les répondants à l'enquête. Nous mettons en évidence les risques d'amplification de biais de l'estimateur par calage généralisé en présence de non-réponse. Ce type de phénomène a été étudié en épidémiologie ; Pearl (2010) and Myers et al. (2011).

Enseignement 2, 09h35-10h35.

Refonte du cours de statistique dans une école de commerce — expérience commentée

Gilles Stoltz (ENS Paris, CNRS)

Cet exposé retrace cinq années d'enseignement du cours de statistique à HEC Paris, de sa refonte totale à la rentrée 2007 jusqu'à l'obtention d'une version stable et satisfaisante au cours de l'année universitaire 2011–12. Le contenu retenu est le contenu typique d'un premier cours de statistique : modélisation statistique, intervalles de confiance, tests d'hypothèses, régression linéaire. La difficulté était de faire face à un public très hétérogène, tant du point de vue du bagage mathématique que de la motivation à étudier la statistique, discipline presque unanimement jugée technique et peu digne d'intérêt. La solution adoptée consiste en un cours magistral proposant des parenthèses mathématiques signalées comme explicitement culturelles, et un polycopié permettant une progression et un investissement à plusieurs vitesses. Par ailleurs, le livre de Florence Noiville, 'J'ai fait HEC et je m'en excuse', publié courant 2008, a également attiré mon attention sur la nécessité de présenter des exemples et illustrations en lien avec la vie publique et citoyenne, pas seulement avec l'entreprise. Cela a conduit en particulier à renommer le cours en 'Statistique pour citoyens d'aujourd'hui et managers de demain', qui résume bien le nouveau point de vue adopté. Dans cet exposé, je m'attacherai surtout à détailler les essais, erreurs, et succès rencontrés, sur le mode : Voilà ce que j'aurais aimé que l'on me dise sur les étudiants en école de commerce il y a six ans !

Enseignement de statistique pour des ingénieurs des sciences du vivant

Liliane Bel (INRA/ Agroparistech)

L'utilité de l'enseignement de statistique pour des ingénieurs en sciences du vivant est une évidence qu'il n'est plus besoin de démontrer. Néanmoins la nécessaire formalisation de cet enseignement rebute souvent un public qui a choisi une filière d'étude axée sur la biologie et dont l'intérêt pour les mathématiques est moindre voire nul. De plus la diversité du recrutement induit de fortes disparités dans les compétences acquises au préalable par les étudiants d'une même promotion. La démarche pédagogique adoptée pour motiver ces étudiants s'appuie sur des illustrations d'applications en lien étroit avec leurs centres d'intérêt. Cela reste insuffisant néanmoins pour convaincre les étudiants des deux premières années du bien-fondé de cet enseignement et une certaine défiance est notée à son encontre. Toutefois cette incompréhension diminue voire disparait dès lors que les étudiants sont confrontés à des traitements de données lors de stages ou de projets.

Les enseignements de statistique en licence de sciences humaines et sociales

Cécile Hardouin (Université Paris Ouest Nanterre)

Cet exposé trouve sa place dans une série de communications orales dédiées à des expériences commentées d'enseignements de service en statistique. Il a pour objet l'enseignement des statistiques dans des disciplines de sciences humaines ou sociales comme la psychologie ou l'ethnologie. Le public de ce cours est souvent issu de filières littéraires, le bagage en mathématiques est très faible et pourtant le contenu est le même que dans des filières plus scientifiques : statistiques descriptives et statistiques inférentielles (estimation, intervalles de confiance, tests d'hypothèses, régression, ACP...). Au problème posé par le niveau des étudiants s'ajoute souvent une grande réticence de leur part pour cet apprentissage, voire un

'blocage en maths'. Dans cet exposé je présenterai les différents essais pour motiver l'auditoire, ainsi que quelques méthodes basiques pour présenter les notions statistiques

Processus 1, 09h35-10h35.

Consistency results for the kernel density estimate on continuous time stationary processes

Sultana Didi (LSTA Paris 6), Djamal Louani (LSTA Paris 6)

Nous présentons des propriétés asymptotique de consistance presque sûre, ponctuelles et uniformes, avec vitesse de convergence pour l'estimateur à noyau f_T, basé sur un processus stationnaire à temps continu admettant une densité f. Nous obtenons nos résultats, en faisant usage d'une approche basée sur l'utilisation des différence de martingales en liaison avec les techniques de projection des tribus.

Estimation du noyau de transition d'un processus markovien déterministe par morceaux

Romain Azaïs (INRIA Bordeaux - Sud-Ouest)

Les processus markoviens déterministes par morceaux sont une classe générale de processus stochastiques non diffusifs, faisant intervenir des trajectoires déterministes ponctuées par des sauts aléatoires. Cette classe de modèles stochastiques couvre un grand nombre d'applications, en biologie (mécanisme de production d'un antibiotique par une bactérie) ou en fiabilité (propagation de fissure). Dans ce cadre, on propose d'estimer de manière non paramétrique et récursive le noyau de transition d'un tel processus à partir de l'observation d'une trajectoire en temps long. On montre la consistance de l'estimateur considéré. Des simulations illustrent les résultats asymptotiques.

Une généralisation de la notion de fonctions aléatoires stationnairement corrélées

Alain Boudou (IMT-Toulouse), Sylvie Viguier-Pla (IMT-Toulouse - UPVD)

Nous introduisons et examinons la notion de fonctions aléatoires continues (f.a.c.) stationnaires qui commutent. Celles-ci pourraient être définies par la commutativité des opérateurs de décalage. Cette notion généralise la notion de stationnarité corrélée. Lorsque deux fonctions sont stationnairement corrélées, elles commutent, la réciproque n'est pas vraie. Par exemple, lorsque $(X_n)_{n\in\mathbb{Z}}$ est une série stationnaire, la série stationnaire $(X_{3n})_{n\in\mathbb{Z}}$ commute avec $(X_n)_{n\in\mathbb{Z}}$, bien que les deux ne soient pas nécessairement stationnairement corrélées. Nous donnons plusieurs définitions équivalentes de cette notion. Nous étudions différentes propriétés asymptotiques associées à cette notion. Nous examinons ensuite comment, à partir de deux f.a.c. stationnaires qui ne commutent pas, on peut "extraire" des parties qui commutent. Nous explorons, sur des exemples simulés, chacune de ces propriétés, et voyons comment celles-ci peuvent être exploitées sur des exemples concrets.

Statistique non paramétrique 2, 09h35-10h35.

Loi limite pour les estimateurs à noyau du taux de hasard avec fenêtre adaptative

Sarah Ouadah (Université Paris Ouest Nanterre la Défense)

Nous présentons des lois limites pour les estimateurs à noyau de la densité des temps de survie et du taux de hasard dans un modèle de censure à droite. Ces lois limites ont la particularité d'être uniformes relativement au paramètre de lissage des estimateurs considérés. Ceci a l'intérêt d'assurer la convergence des estimateurs sous des conditions générales, même lorsque le paramètre de lissage est aléatoire. Nous établissons cette propriété d'uniformité pour toutes les valeurs du paramètre de lissage pour lesquelles ces estimateurs sont convergents. De plus, nous obtenons des évaluations explicites asymptotiques des erreurs aléatoires correspondantes.

Estimation par noyaux associés mixtes d'un modèle de mélange

Francial G. Libengué (Université de Franche-Comté), Sobom M. Somé (Université de Franche-Comté), Célestin C. Kokonendji (Université de Franche-Comté)

Nous nous intéressons dans cette communication à l'estimation non-paramétrique de densités gouvernant les variables aléatoires réelles à support connu constitué à la fois des intervalles et des ensembles discrets deux à deux disjoints. Nous décrivons d'abord le modèle étudié qui peut être le cas d'un processus (fonction) de suivis d'un malade à p phases au cours desquelles le patient subit alternativement des soins intensifs avec des prélèvements quasi-instantanés (i.e. continus) correspondant a une période d'hospitalisation, et des soins externes où les prélèvements sont périodiques ou séquentiels dans le temps (i.e. discrets). Puis, nous définissons le noyau associé mixte correspondant . Ensuite, nous utilisons les outils unifiant les analyses discrètes et continues pour montrer les principales propriétés de ces estimateurs à savoir le biais, la variance ainsi que l'erreur quadratique moyenne intégrée. En particulier nous mettons un accent sur les points situés aux frontières lors de passage du continu au discret. Enfin, vient la discussion sur le choix des paramètres de lissage global et local ainsi que la méthode optimale appropriée pour ceci.

Approximation de la fonction de répartition d'une distribution composée via une développement polynomial

Pierre-Olivier Goffard (IML, Marseille), Stéphane Loisel (ISFA, Lyon), Denys Pommeret (IML, Marseille)

L'objectif de ce travail est de mettre en oeuvre une méthode numérique pour approcher la fonction de répartition associée à une distribution composée. Nous proposons une expression de la densité de la distribution composée par rapport à une mesure de probabilité appartenant à une Famille Exponentielle Naturelle Quadratique (FENQ). L'expression est une série infinie issue de la projection sur un système de polynome orthogonaux par rapport à cette mesure. La fonction de répartition s'en déduit par intégration puis une approximation est finalement obtenue par troncature de la série infinie. Des simulations numériques illustrent les performances de la méthode et l'application à des modèles actuarielles montrent la pertinence d'un tel travail.

Analyse de données 4, 09h35-10h35.

Spécification et estimation d'un MES avec des variables latentes formatives endogènes

Mouloud Tensaout (Université du Maine,GAINS-ARGUmans), Hervé Guyon (Université Paris Sud,PESOR)

La littérature sur la spécification des variables latentes formatives ou composite (VLC) a mis en en évidence diverses difficultés d'interprétation et d'estimation de telles variables dans un modèle d'équations structurelles (MES), (absence de l'erreur de mesure, exhaustivité des mesures formatives, interprétation confondante). Néanmoins ces critiques sont menées dans un MES dont lequel la variable latente VLC est exogène. Dans ce papier nous prolongeons ces discussions en examinant le cas d'un MES comportant des VLC endogènes. A l'aide d'exemples simulés nous testons plusieurs hypothèses : H1 : l'effet causal d'une variable exogène sur la VLC endogène est entièrement médiatisé par les mesures exhaustives de la VLC. En conséquence l'effet causal direct de la variable exogène sur la VLC est nul.H2 : Dans le cas le contraire le modèle formatif de la VLC est incorrectement spécifié. Les mesures de la VLC endogène ne sont pas exhaustives. H3 : L'omission des relations entre la variable exogène est les indicateurs de la VLC endogène conduit a une estimation biaisée de l'effet causal sur la VLC endogène. Nous montrons alors, que l'effet d'une variable latente exogène sur la VLC endogène est médiatisé par les relations entre les indicateurs Xi de la VLC endogène.

Méthodes simples, sparses et algorithmes génétiques

Christelle Reynes (Université Montpellier 1), Robert Sabatier (Université Montpellier), Guilhem Kister (Université Montpellier)

Les méthodes sparses sont des techniques de sélection de variables appliquées à des méthodes factorielles qui conduisent à un seuillage des coefficients. Cette approche se heurte à différentes limites : utilisation brute de coefficients dont la valeur n'est pas toujours pertinente pour le but visé, difficulté à éliminer les variables redondantes,... Nous proposons une méthode basée sur les algorithmes génétiques utilisant des coefficients entiers et favorisant l'apparition du coefficient nul pour pallier à ces limites. Le principe est appliqué à l'Analyse en Composantes Principales, à la régression Partial Least Squares et à l'Analyse Factorielle Discriminante. Des applications dans le domaine des -omics seront proposées.

Extended regularized generalized canonical correlation analysis to multi-group data analysis

Arthur Tenenhaus (Supélec Sciences des Systèmes), Michel Tenenhaus (HEC, Paris)

L'analyse de données structurées par blocs concerne l'analyse de plusieurs ensembles de variables observées sur un seul ensemble d'individus. L'analyse de données structurées par groupes concerne l'analyse d'un seul ensemble de variables mesurées sur plusieurs groupes d'individus. De nombreuses méthodes existent pour l'analyse de données structurées par blocs. Il apparaît que l'analyse canonique généralisée régularisée (RGCCA) proposée par [1] qui est une approche générale pour l'analyse de données multi-bloc peut être étendue à l'analyse de données multi-groupe.

Fiabilité et incertitudes 1, 11h00-12h00.

Évaluation des incertitudes associées à la mesure granulométrique d'un aérosol par technique smps

Loic Coquelin (LNE), Laurent Le Brusquet (SUPELEC), Nicolas Fischer (LNE), Charles Motzkus (LNE), Francois Gensdarmes (IRSN), Tatiana Mace (LNE), Séverine Demeyer (LNE), Gilles Fleury (SUPELEC)

La détermination de la granulométrie en nombre d'un aérosol (concentration en nombre de particules en fonction du diamètre) à partir de mesures effectuées par un SMPS (Scanning Mobility Particle Sizer) est un problème mathématiquement mal posé. Une procédure d'inversion pour l'estimation de ce mesurande fonctionnel est proposée ainsi qu'une méthodologie pour propager l'incertitude résultant à la fois des erreurs de mesure et du manque de connaissances sur la physique sous-jacente au processus de mesure. L'inversion consiste en la décomposition du signal sur une base d'ondelettes discrètes couplée à des techniques de régularisation. Une comparaison entre la méthode développée et une technique de régularisation standard avec contraintes de lissage lorsque l'on considère une distribution de taille simulée avec des pics larges et étroits est proposée. Les résultats montrent un meilleur accord entre la reconstruction moyenne calculée par simulations de Monte-Carlo et la granulométrie originale pour la nouvelle procédure d'inversion.

Optimisation de la maintenance d'un equipement optronique

Camille Baysse (Thales Optronique/INRIA), Didier Bihannic (Thales Optronique), Benoîte de Saporta (INRIA Bordeaux, université Bordeaux 4, IMB), Anne Gégout-Petit (INRIA Bordeaux, Université Bordeaux 2, IMB), Michel Prenat (Thales Optronique), Jérôme Saracco (INRIA Bordeaux, université Bordeaux)

Dans le cadre de l'optimisation de la fiabilité, Thales intègre dans ses équipements, des systèmes d'observation de leur état de santé. Cette fonction est réalisée par des HUMS (Health Usage Monitoring System). Nous souhaitons mettre en place dans le HUMS, un programme capable à partir de ses observations, de déterminer l'état du système et de proposer une stratégie de maintenance. Nous décomposons ce problème en deux étapes : la première étape consiste à détecter l'état dégradé du système en utilisant une variable informative et des chaînes de Markov cachées pour estimer l'instant de rupture dans l'évolution de cette variable. La seconde étape consiste à proposer une politique de maintenance optimale et dynamique, adaptée à l'état du système et qui prenne en compte à la fois les pannes aléatoires et les pannes liées à un phénomène d'usure. Nous voulons estimer le meilleur instant pour effectuer une maintenance. Pour cela nous modélisons l'état de notre système par un processus Markovien déterministe par morceaux (PDMP). Nous considérons ce problème comme un problème d'arrêt optimal dont l'objectif est de maximiser une fonction de performance qui tienne compte du temps de fonctionnement, du coût de maintenance, de réparation et d'immobilisation. Nous présentons ces résultats sur des données simulées. Cette méthodologie peut s'étendre à des cas plus complexes.

Le mouvement brownien géométrique non-homogène comme modèle de dégradation pour la propagation de fissure

Christian Paroissin (Université de Pau)

Depuis quelques décennies, des modèles de dégradation ont été étudiés afin de mieux comprendre le vieillissement de composants. Ces modèles sont particulièrement utiles par rapport à des modèles de durées pour des composants hautement fiables, par exemple. Par ailleurs, les modèles de dégradation permettent de construire des politiques de maintenance plus sophistiquées. Les modèles classiques de dégradation appartiennent à la famille des processus de Lévy. Les trois principaux processus sont : (a) le processus gamma, (b) les processus de Poisson composés et (c) le mouvement brownien avec tendance. Ces processus sont dit homogènes car leurs accroissements sont indépendants et stationnaires. Cela implique que la dégradation moyenne est nécessairement linéaire en temps. Donc on ne peut pas modéliser n'importe quel phénomène ; par exemple, la propagation de fissure est plutôt exponentielle en moyenne. C'est pourquoi des modèles non-homogènes sont de plus en plus étudiés. Dans ces modèles, on relâche l'hypothèse de stationnarité des accroissements. Dans cet exposé, nous étudions le mouvement brownien géométrique non-homogène, proposé dans ce contexte par Ebrahimi (2005). Pour ce modèle, on considère deux problèmes : (i) l'estimation semi-paramétrique des paramètres du modèle ; (ii) la loi du temps de défaillance du composant (défini comme le premier temps de passage d'un seuil critique par le processus). Des exemples classiques seront traités.

Processus 2, 11h00-12h00.

Problèmes à deux échantillons : tests à noyaux multiples basés sur des approches bootstrap non asymptotiques

Magalie Fromont (CREST Ensai, IRMAR), Béatrice Laurent (IMT, INSA Toulouse), Matthieu Lerasle (CNRS Université de Nice Sophia-Antipolis), Patricia Reynaud-Bouret (CNRS Université de Nice Sophia-Antipolis)

Considérant soit deux processus de Poisson soit deux échantillons indépendants, nous nous intéressons au problème de test d'égalité de leurs lois respectives. Nous proposons tout d'abord des procédures de test dont les statistiques de test sont des U-statistiques basées sur un noyau unique, qui peut être choisi comme un noyau de projection, un noyau d'approximation ou un noyau reproduisant. Les valeurs critiques correspondantes sont construites à partir d'approches bootstrap non asymptotiques spécifiques, de telle façon que les tests obtenus soient de niveau alpha choisi. Nous exhibons des conditions sur l'alternative, exprimées en termes de vitesses de séparation, qui garantissent un contrôle exact du risque de deuxième espèce, et qui sont optimales dans certains cas pour un noyau bien choisi. Nous introduisons ensuite des procédures de test agrégées ou à noyaux multiples, qui nous permettent de contourner la difficulté relative au choix du noyau et/ou de ses paramètres, et d'importer des idées issues de l' estimation adaptative par sélection de modèles, par seuillage ou par noyaux d'approximation. Les valeurs critiques induites dans ces tests à noyaux multiples sont calibrées pour qu'ils soient toujours de niveau alpha et qu'ils vérifient aussi des inégalités de type oracle conduisant à des propriétés d'adaptativité au sens du minimax dans certains cas.

Apprentissage de données multi-fidélités par mélange de processus gaussiens

Matthias de Lozzo (ONERA), Loïc Le Gratiet (CEA)

Afin d'approcher un phénomène physique par apprentissage statistique de données entrées/sorties, nous disposons d'une source d'observations haute-fidélité (HF) dont l'utilisation est coûteuse et de plusieurs sources de données basse-fidélités (BF) plus économiques. Nous construisons un mélange de processus gaussiens multi-fidélités basé sur ces observations dont l'espérance fournit le modèle de substitution de la source HF et la variance l'erreur quadratique du modèle. Les processus gaussiens multi-fidélités sont des co-krigeages construits via une "agrégation par validation croisée", chacun étant associé à une des sources BF et à la source HF. On observe qu'un co-krigeage moyenné est plus robuste qu'un co-krigeage classique en présence d'un faible nombre d'observations HF. De plus autour des points HF, leur erreur est proche de celle de validation croisée ce qui localement fait de cette dernière un estimateur de l'erreur de prédiction. Par la suite, une classification dure basée sur l'erreur de généralisation est effectuée de sorte à associer à tout point du domaine des paramètres d'entrée un unique co-krigeage moyenné. Cette classification est ensuite considérée comme la réalisation d'une variable aléatoire afin de rendre compte de la continuité du phénomène physique. Le modèle de substitution retenu est l'espérance du modèle issu de la classification dure et sa variance représente son erreur quadratique ; ces quantités sont connues sous forme analytique.

Tests d'adéquation pour les processus de Poisson et les processus de Hawkes

Christine Tuleau-Malot (Université Nice Sophia-Antipolis), Patricia Reynaud-Bouret (Université Nice Sophia-Antipolis), Vincent Rivoirard (Université Paris Dauphine), Franck Grammont (Université Nice Sophia-Antipolis)

En neurosciences, le vecteur permettant l'étude de l'activité cérébrale est le potentiel d'action. C'est pourquoi une majorité des études sur le cerveau considère les trains de spikes, soit l'enregistrement temporel des différents potentiels d'actions émis lors d'une tâche. Cependant, dans ces différentes études, un certain nombre d'hypothèses sont faites quant à la modélisation sous-jacente des trains de spikes. Ainsi, les modèles considérés sont le processus de Poisson homogène, le processus de Poisson inhomogène et le processus de Hawkes. Ces modélisations, qui ont principalement évoluées avec les découvertes biologiques sur le fonctionnement des neurones, sont généralement faites sans aucune réelle justification mathématique. L'objectif ici est, après de courts rappels sur les différents processus considérés et sur les méthodes permettant d'estimer leur intensité, de présenter deux procédures statistiques, essentiellement basées sur le test de Kolmogorov-Smirnov, qui permettront pour l'une de tester l' adéquation au processus de Poisson et pour l'autre l'adéquation au processus de Hawkes. En sus de ces considérations purement théoriques, une application sur données simulées permettra de valider pratiquement chacune des procédures. Par ailleurs, afin d'illustrer le fait qu'une modélisation unique semble peu probable pour les trains de spikes, des données réelles seront également considérées.

Enseignement 3, 11h00-12h00.

Une unité d'enseignement "études de cas en statistique" en 3e année de licence MASS

Anne Gégout-Petit (IMB, Univ. Bordeaux), Vincent Couallier (IMB, Bordeaux)

Nous présentons le contenu et le déroulement de l'unité d'enseignement "Etudes de cas en statistique" que nous avons mise en place à la rentrée 2011 en 3e année de licence MASS (Mathématiques Appliquées et Sciences Sociales) à l'université Bordeaux Segalen. Il s'agit de confronter les étudiants à différentes problématiques réelles à résoudre avec ou sans données avec l'outil logiciel (Excel et SPSS) utilisé pour analyser les données ou simuler dans un objectif d'aide à la décision. Il semble que la diversité des cas, associée à la différence d'approche des deux enseignants (les auteurs) a permis d'accrocher un grand nombre d'étudiants qui dans l'ensemble ont apprécié cette UE et sont intéressés par des poursuites d'études en statistique ou dans des domaines qui utilisent cet outil (économétrie, épidémiologie,..)

Cours d'homogénéisation de statistique en licence professionnelle

Adeline Samson (MAP5, Univ Paris Descartes)

Cet exposé se place dans le cadre d'une série d'exposés ayant trait aux enseignements à un public varié. Il retrace cinq années d'enseignement du cours d'homogé-néisation de statistique en licence professionnelle 'Statistique et Informatique Décisionnelle pour la Santé' au département STID de l'IUT de Paris. Le public de ce cours est très hétérogène (ce qui en fait sa richesse) : étudiants de formation initiale, continue, étudiants de formation mathématique, biologie, biochimie, pharmacie. Le contenu retenu est le contenu typique d'un premier cours de statistique : modélisation statistique, intervalles de confiance, tests d'hypothèses (de comparaison à une valeur de référence, entre deux populations, tests du χ^2). L'accent est mis sur les TP afin d'aider à faire passer les notions de statistique inférentielle et pour ancrer l'enseignement dans un parcours professionnel. Dans cet exposé, je présenterai les différents essais pédagogiques réalisés au cours de ces dernières années, les erreurs et succès rencontrés.

Revue statistique et enseignement : numéro spécial évolutif sur l'interdisciplinarité

Jeanne Fine (Statisticienne, Toulouse), Dominique Lahanier-Reuter (Laboratoire Théodile CIREL, Université Lille 3), Claudine Schwartz (Statisticienne, Paris)

"L'enseignement de la statistique en interdisciplinarité" est le thème du dernier numéro, publié fin 2012, de la revue Statistique et Enseignement (www.statistique-et-enseignement.fr). Ce numéro spécial ouvre plusieurs débats auxquels les lecteurs sont invités à contribuer : 1) sur la notion (très interdisciplinaire) de preuve statistique, utilisée tant dans les sciences expérimentales que dans les sciences humaines, ainsi que sur les images et le vocabulaire à employer pour l'approche de la notion de test 2) sur la statistique en tant que discipline scolaire, sur la statistique comme entrée dans des activités interdisciplinaires en milieu scolaire : qu'en pensent les professeurs ? 3) sur la question : quelle(s) discipline(s) devrai(en)t être en charge de l'enseignement de la statistique dans l'enseignement secondaire ? Pour certains, cet enseignement doit être introduit en mathématiques, pour d'autres, les notions telles que intervalle de confiance et différence significative, lien statistique significatif, pourraient être introduites en physique par le biais de la mesure, en biologie par le biais de l'influence des niveaux d'un facteur, en sciences économiques et sociales avec l'étude de déterminants sociaux, ... L'objet de la communication est de présenter la genèse

de ce numéro spécial et d'apporter un éclairage sur le contenu de ces débats à partir des trois articles déjà publiés.

Classification non supervisée 1, 11h00-12h00.

Un modèle dynamique à variables latentes pour le partitionnement de données temporelles

Hani El Assaad (IFSTTAR), Allou Samé (IFSTTAR), Gérard Govaert (HEUDIASYC, UMR CNRS 7253), Patrice Aknin (IFSTTAR)

Cet article aborde le problème de la classification de données temporelles en utilisant un mélange dynamique de lois gaussiennes dont les moyennes sont considérées comme des variables latentes qui évoluent suivant des marches aléatoires. L'objectif final est de suivre l'évolution dynamique de certains composants ferroviaire critiques à l'aide des données acquises par des capteurs embarqués. Les paramètres de l'algorithme proposé sont estimés par la méthode du maximum de vraisemblance via l'algorithme Expectation-Maximization (EM). Contrairement à d'autres approches comme l'estimation bayésienne par la méthode du maximum a posteriori dans laquelle les hyperparamètres de lissage doivent être fixées par l'utilisateur, les résultats des simulations montrent la capacité de l'algorithme proposé à estimer correctement ceux-ci tout en gardant un faible taux d'erreur de classification.

Comparison of linear modularization criteria of networks using relational metric

Patricia Conde Céspedes (LSTA-Paris VI), Jean-François Marcotorchino (Thales - Université Paris VI)

Un graphe est un ensemble d'objets liés par une certaine relation typée. Le problème de clustering des graphes peut, alors, être modélisé via l'Analyse Relationnelle Mathématique. L'écriture relationnelle permet de comparer sur les mêmes bases un certain nombre de critères de modularisation. Nous proposons une ré-écriture relationnelle des critères de modularisation linéaires tels le critère de Girvan-Newman, Zahn-Condorcet et Owsinski-Zadrozny. Nous introduirons aussi, deux nouveaux critères de modularisation : la modularité équilibrée et l'indice de Janson-Marcotorchino. Le premier constitue une version équilibrée de la modularité de Girvan-Newman et le second se base sur la structure de l'écart à l'indétermination. Les résultats obtenus avec les critères mentionnés seront testés avec l'algorithme de Louvain générique.

Modèle de classification de données qualitatives par modes de dépendance conditionnelle

Matthieu Marbac (DGA - INRIA Lille), Christophe Biernacki (INRIA Lille - Université Lille 1), Vincent Vandewalle (IUT Roubaix, Université Lille 2)

Nous proposons un modèle de mélange pour la classification non-supervisée de données qualitatives. Dans cette approche, les variables sont regroupées par blocs conditionnellement indépendants. La distribution de chaque bloc est rendue parcimonieuse en estimant uniquement la valeur des probabilités pour les croisements de modalités les plus importants, alors que la masse de probabilité restante est uniformément répartie parmi les autres croisements de modalités. Un algorithme EM est utilisé pour estimer simul-

tanément le nombre de modes par bloc et les paramètres du modèle tandis que le problème combinatoire dû à la recherche des blocs est résolu par un algorithme de Gibbs.

Histoire, 11h00-12h00.

Daniel Encontre (1762-1818) : enseignant de mathématiques transcendantes à Montpellier et de dogme protestant à Montauban

Antoine de Falguerolles (Enseignant-chercheur retraité)

Encontre appartient à une famille de ministre du culte protestant. Pasteur formé à l'étranger, il quitte rapidement le ministère pastoral pour enseigner à Anduze puis à Montpellier. La création de l'université impériale (loi du 10 mai 1806 complétée par le décret du 17 mars 1808)lui ouvre une carrière universitaire. Nommé professeur de mathématiques transcendantes à la faculté des sciences de Montpellier, il en devient le premier doyen. Démissionnaire et nommé professeur de dogme à la faculté de théologie protestante de Montauban (1814), il en devient le second doyen (1816). Cette faculté, rattachée à l'université de Toulouse lors de sa recréation en 1896, en sera exclue en application de la loi de 1905. NUMDAM le crédite de 9 publications dans les "Annales de Gergonne" entre 1810 et 1814. TOLOSANA permet de prendre connaissance de ses "Éléments de géométrie plane". Certains de ses travaux concernent le calcul des probabilités. Deux publications du "Recueil des Bulletins publiés par la Société des Sciences et Belles-Lettres de Montpellier" ont retenu l'attention de Richard Pulskamp (Sources in the History of Probability and Statistics). Dans son livre (Le système du Monde, Pierre Simon Laplace un itinéraire dans la science, NRF, 2004), Roger Hahn rappelle que Laplace fut chargé en 1800 par l'Académie d'examiner un projet de livre adressé à l'Institut par "cet instructeur de mathématiques à Montpellier" ... Cet exposé est l'occasion d'essayer de faire le point sur ses talents de probabiliste.

La médienne : une idée de laplace (1818)

Delphine Blanke (Université d'Avignon), Edith Gabriel (Université d'Avignon et des Pays de Vaucluse), Didier Josselin (Université d'Avignon et des Pays de Vaucluse)

La première partie de l'exposé consiste à présenter des résultats issus de Laplace (1818). Dans la dernière partie du second supplément (1818) de sa célèbre 'Théorie Analytique des Probabilités', Pierre-Simon de Laplace compare la variance asymptotique de la moyenne empirique $\bar{x}$ et celle de la médiane M_n dans un contexte de régression. De plus, en se basant sur la normalité asymptotique jointe de $(\bar{x}, M_n)$, il cherche à établir comment un estimateur, basé sur une combinaison linéaire de $\bar{x}$ et M_n, $(1-\alpha)\bar{x}+\alpha M_n$, peut avoir une meilleure efficacité asymptotique. Il conclut cependant ses travaux de manière pessimiste en notant que son approche se heurte à la connaissance précise de la loi des erreurs, et que pour des erreurs de loi normale, $\bar{x}$ ne peut pas être amélioré... En se basant sur cette idée, Josselin et Ladiray (2002) étudient des propriétés théoriques de cet estimateur qu'ils appellent "médienne" et l'appliquent à l'analyse de la pullulation du campagnol terrestre dans le Doubs. Dans la deuxième partie de l'exposé, nous proposons une famille d'estimateurs adaptatifs dérivés de l'estimateur de Laplace Blanke et al., 2012). Enfin, nous présentons les résultats d'une étude Monte Carlo sur la robustesse de ces estimateurs que nous comparons aux estimateurs classiques d'un paramètre de localisation.

Caractérisations de lois probabilistes via le maximum de vraisemblance

Christophe Ley (Université Libre de Bruxelles), Duerinckx Mitia (Université libre de Bruxelles (ULB)), Yvik Swan (Université du Luxembourg)

Un célèbre théorème de caractérisation dû à Gauss dit que l'estimateur du maximum de vraisemblance (MLE pour Maximum Likelihood Estimator) du paramètre de position dans une famille de position est la moyenne empirique pour tous les échantillons de toutes tailles d'échantillon possibles si et seulement si la famille est gaussienne. Il existe beaucoup d'extensions de ce résultat (entre autres par Poincaré et von Mises) dans diverses directions, la plupart se focalisant sur des familles de position et d'échelle univariées ainsi que sur des familles de position sphérique. Dans cet exposé, nous présentons une vision unifiée de cette littérature en proposant des théorèmes de caractérisation MLE généraux pour des familles à un paramètre. En agissant de la sorte, nous donnons les outils nécessaires pour déterminer si une famille donnée est MLE-caractérisable par rapport à son paramètre (position, échelle, asymétrie, ...), et, si tel est le cas, nous définissons le concept fondamental de Minimal Necessary Sample Size (MNSS) qui donne les conditions minimales à partir desquelles une caractérisation a lieu. Pour simplifier la présentation, nous allons nous focaliser sur les familles de position. Notre résultat nous permet de retrouver la plupart des références-clés de cette littérature et de discuter ces divers résultats dans un cadre unifié, et de nombreux nouveaux résultats de caractérisation seront donnés.

Société Française de Biométrie, 11h00-12h00.

Classification de gènes co-exprimés par modèles de mélange. Des puces à ADN au séquençage haut-débit

Gilles Celeux (INRIA Saclay Ile-de-France), Marie-Laure Martin-Magniette (INRA/ AgroParisTech / URGV), Cathy Maugis-Rabusseau (IMT), Andréa Rau (INRA Jouy en Josas)

La détection de groupes de gènes co-exprimés est une question biologique importante car de tels gènes sont de bons candidats pour être co-régulés.

Les progrès de la technologie des puces à ADN au milieu des années 1990 a permis de mesurer le niveau d'expression de milliers de gènes simultanément dans différentes conditions expérimentales. Cette technologie fournissant une mesure de type continu de l'expression des gènes, nous avons considéré des mélanges gaussiens multivariés pour mettre en évidence des groupes de gènes co-exprimés.

Ces dernières années, les avancées significatives du séquençage haut-débit a fait du séquençage des ARN (RNA-seq) un nouvel outil pour l'étude de l'expression des gènes. Bien que les puces à ADN et le RNA-Seq ont pour but de caractériser l'activité transcriptionnelle, les outils statistiques pour l'analyse des données issue de la première technologie ne sont pas adaptées pour la seconde technologie, cette dernière fournissant des comptages. Nous proposons des mélanges de lois de Poisson pour classer les gènes à partir des comptages observés. Ces mélanges doivent prendre en compte les caractéristiques des données de RNA-seq. Nous avons étudié les performances de notre procédure sur des données simulées et des données réelles.

Modèles à variables latentes pour des données issues de tiling arrays

Caroline Bérard (Université de Rouen)

Les puces tiling arrays sont des puces à haute densité permettant l'exploration des génomes à grande échelle. Elles sont impliquées dans l'étude de l'expression des gènes et de la détection de nouveaux transcrits grâce aux expériences de transcriptome, ainsi que dans l'étude des mécanismes de régulation de l'expression des gènes grâce aux expériences de ChIP-chip. Dans l'objectif d'analyser des données de ChIP-chip et de transcriptome, nous proposons une modélisation fondée sur les modèles à variables latentes, en particulier les modèles de Markov cachés. Les caractéristiques biologiques du signal issu des puces tiling arrays telles que la dépendance spatiale des observations le long du génome et l'annotation structurale sont intégrées dans la modélisation. Nous proposons un mélange de régressions pour la comparaison de deux échantillons dont l'un peut être considéré comme un échantillon de référence (ChIP-chip), ainsi qu'un modèle gaussien bidimensionnel avec des contraintes sur la matrice de variance lorsque les deux échantillons jouent des rôles symétriques (transcriptome). Enfin, une modélisation semi-paramétrique autorisant des distributions plus flexibles pour la loi d'émission est envisagée. Les différents modèles sont illustrés sur des jeux de données réelles de ChIP-chip et de transcriptome issus d'une puce NimbleGen couvrant le génome entier d'Arabidopsis thaliana.

Méthodes d'analyses de microbiote en lien avec son environnement, exemple du microbiote de Daphnia

Mahendra Mariadassou (INRA), Samuel Pichon (University of Basel), Dieter Ebert (University of Basel)

Les techniques de séquençage haut débit permettent désormais de caractériser des écosystèmes microbiens, ou microbiotes, en les résumant à un inventaire : une liste d'OTUs (Operational Taxonomic Units), et leurs abondances. Il est nécessaire de comparer ces inventaires et de les relier à des variables d'intérêt. Outre la difficulté de reconstruire ces inventaires, les fortes similarités qui existent entre différents OTUs nécessitent de comparer les inventaires de façon astucieuse. Une grande classe de méthodes s'appuie sur un arbre phylogénétique des OTUs. La distance entre inventaires est alors la fraction de l'arbre qui est spécifique à un des deux inventaires (UniFrac) ou la distance entre "barycentres" de chaque inventaire (DPCoA). Ces différentes méthodes ont des emphases différentes : sur les OTUs rares (UniFrac) ou au contraire abondants (DPCoA) et nécessitent des efforts différents en terme de reconstruction des OTUs. Une autre grande classe voit l'inventaire comme une distribution de probabilité sur un arbre de référence et considère la distance de Wassertein, ou l'effort nécessaire pour transformer une distribution en une autre. Une fois les distances obtenues, on peut s'intéresser à l'effet de covariables environnementales (localisation, type d'hôte, etc.) sur les distances entre microbiotes pour détecter des forçages extérieurs. Nous passerons en revue les méthodes d'ordination et les tests d'association avec covariables environnementales avant de les illustrer sur le microbiote de la Daphnée.

Résumés du jeudi 30 mai 2013

Prix Pierre Simon de Laplace - Christian Gouriéroux, 09h00-10h00.

Granularity theory

Christian Gouriéroux (CREST et University of Toronto)

The risk analysis in portfolio of credits or life insurance contracts is made difficult by the nonlinearities of risk models, the dependencies between the individual risks and the large sizes of the portfolios, which can include several hundred thousands of contracts. The granularity principle has been introduced in the Basel 2 regulation for credit risk to solve these difficulties when computing the reserves. The principle requires three steps. First the modeling step considers a Risk Factor Model (RFM), which distinguishes the systematic and unsystematic risks. Second this model is applied to a virtual portfolio of infinite size, leading to the so-called Asymptotic Risk Factor Model (ARFM). This gives in general explicit formulas for the Value-at-Risk and thus for the required capital. Third for a portfolio of large, but finite size, closed form approximations are derived from an expansion around the ARFM. This provides the granularity adjustment for the required capital. The granularity principle can be applied to a variety of other problems. We focus in this presentation on the question of the efficient estimation in panel factor models with both micro-and macro dynamics.

Michael Goldstein, 10h15-11h15.

Bayesian uncertainty analysis for complex physical systems modelled by computer simulators

Michael Goldstein (Durham University)

Most large and complex physical systems are studied by mathematical models, implemented as high dimensional computer simulators. While all such cases differ in physical description, each analysis of a physical system based on a computer simulator involves the same underlying sources of uncertainty. These are : condition uncertainty (unknown initial conditions, boundary conditions and forcing functions), parametric uncertainty (as the appropriate choices for the model parameters are not known), functional uncertainty (as models are typically expensive to evaluate for any choice of parameters), stochastic uncertainty (arising from intrinsic randomness in the system equations), solution uncertainty (as solutions to

the system equations can only be assessed approximately), structural uncertainty (as the model is different from the physical system), multi-model uncertainty (as there often is a family of models, at different levels of resolution, possibly with different representations of the underlying physics), decision uncertainty (as the links between decisions that we can make in the model and those things that we can influence in the world are uncertain) and measurement uncertainty (in the data used to calibrate the model).

There is a growing field of study which aims to quantify and synthesise all of the uncertainties involved in relating models to physical systems, within the framework of Bayesian statistics, and to use the resultant uncertainty specification to address problems of forecasting and decision making based on the application of these methods. This methodology is concerned both with practical and methodological issues (how can we work out the likely behaviour of the physical system given our evaluations of the computer simulator ?) and foundational issues (why should our methods work and what do our answers mean ?). This talk will give an overview of the status of this emerging methodology, with particular emphasis on Bayesian multivariate, multi-level, multi-model emulation, careful structural discrepancy modelling and iterative history matching. The methodology will be illustrated with examples of current areas of application, in particular asset management for oil reservoirs, galaxy modelling, and assessment of rapid climate change..

David Hunter, 10h15-11h15.

Maximum Smoothed Likelihood for Multivariate Mixtures

David Hunter (Penn State University), Xiaotian Zhu (Penn State University), Michael Levine (Purdue University), Didier Chauveau (Université d'Orléans)

We introduce an algorithm for estimating the parameters in a finite mixture of completely unspecified multivariate components in at least three dimensions under the assumption of conditionally independent coordinate dimensions. We prove that this algorithm, based on a majorization–minimization idea, possesses a desirable descent property just as any EM algorithm does. We also demonstrate via simulation studies that the new algorithm gives very similar results to another algorithm that does not satisfy any descent algorithm. We provide and demonstrate code for implementing the new algorithm in a publicly-available R package.

Extrêmes 2, 11h20-12h40.

Estimation de niveaux de retour : comparaisons et discussion

Cécile Mercadier (Institut Camille Jordan, UCBL), Juan Juan Cai (TU Delft), Anne-Laure Fougères (Institut Camille Jordan, UCBL)

L'estimation de niveaux de retour est un problème statistique classique en hydrologie et plus généralement dans de nombreuses applications environnementales. Dans cet exposé, on reviendra sur la définition du niveau de retour. On présentera ensuite trois méthodes de la théorie univariée des valeurs extrêmes pour estimer un niveau de retour sous l'hypothèse stationnaire. Une étude comparative sera menée à l'aide de simulations. Une discussion sera engagée sur l'extension du problème et des procédures dans un cadre non stationnaire en s'appuyant sur la bibliographie récente.

Données environnementales : la théorie multivariée des valeurs extrêmes en pratique

Anne-Laure Fougères (Institut Camille Jordan, UCBL), Juan Juan Cai (TU Delft), Cécile Mercadier (Institut Camille Jordan, UCBL)

L'analyse détaillée de données climatiques est une étape importante dans la gestion des risques environnementaux. Nous nous intéressons au cas où deux variables quantitatives sont mesurées au cours du temps, formant ainsi une série temporelle bivariée (X_t, Y_t) supposée stationnaire. Au cours de cette présentation, nous évoquerons le problème de l'estimation d'une probabilité de défaillance, définie comme la probabilité $P(X_t > x, Y_t > y)$, où x et y sont deux valeurs extrêmes (trop grandes pour être observées, ou presque). La théorie multivariée des valeurs extrêmes fournit des réponses à cette question. Un aspect particulièrement important dans l'étape de modélisation est celui de la prise en compte de la dépendance, pouvant subsister ou au contraire s'effacer lorsque l'on se focalise sur les valeurs extrêmes. Nous présenterons plusieurs méthodes d'estimation, fondées sur des approches introduites par Draisma et al. (2004) d'une part, et par Heffernan et Tawn (2004) d'autre part. Nous mettrons ensuite en concurrence les estimateurs déduits sur des simulations dans un premier temps, puis sur des données climatiques. Les résultats obtenus seront finalement discutés.

Le processus t-extrêmal : construction et domaine d'attraction elliptique

Thomas Opitz (Université Montpellier 2)

La structure de dépendance de type t-extrêmal apparaît asymptotiquement pour les lois de probabilité de type elliptique dans le cadre de la dépendance asymptotique. Une extension au cadre spatial est possible ; la structure de dépendance est alors paramétrisée par un degré de liberté positif et par une fonction de corrélation. Ici nous présentons la construction dite spectrale du processus max-stable correspondant, basée sur le processus Gaussien avec la même fonction de corrélation. En outre, nous démontrons que tout processus à lois fini-dimensionnelles elliptiques a pour limite max-stable le processus t-extrêmal.

Modélisation et simulation dans un cadre spatio-temporel max-stable de processus climatiques

Aurélien Bechler (INRA/LSCE), Liliane Bel (INRA/ Agroparistech), Mathieu Vrac (LSCE)

Depuis peu, les processus max-stables sont devenus des outils incontournables pour la modélisation statistique des extrêmes des processus spatiaux. Néanmoins, de nombreuses questions restent en suspens en particulier concernant les algorithmes permettant de les simuler de façon précise et rapide. Dans ce travail on utilise la modélisation donnée par la construction spectrale du processus t-extremal (Opitz, 2012) qui est une généralisation du processus de Schlather (Schlather, 2002). Elle permet de modéliser des processus de maxima avec des structures de dépendance asymptotique présentant une grande variabilité. On adapte pour la simulation conditionnelle de ces processus la méthodologie introduite par Dombry et al. (2011). On a appliqué cet algorithme pour simuler des données de précipitations extrêmes dans le sud de la France, région dans laquelle les conséquences (entre autres les inondations) peuvent être désastreuses.

Processus ponctuels spatiaux, 11h20-12h40.

Processus ponctuels spatio-temporels : analyse et simulations

Edith Gabriel (Université d'Avignon et des Pays de Vaucluse)

Les données ponctuelles spatio-temporelles sont de plus en plus nombreuses et d'un large spectre scientifique. Elles correspondent souvent à une unique réalisation du processus sous-jacent dans un domaine fini. Dans de nombreux cas, les analyses séparées des composantes spatiale et temporelle sont d'intérêt limité, essentiellement parce que les objectifs scientifiques de l'étude sont de comprendre et de modéliser les mécanismes stochastiques d'interaction spatio-temporelle. Les caractéristiques d'ordre deux sont utilisées pour analyser la structure spatio-temporelle du processus ponctuel. La fonction de corrélation de paire et la fonction K de Ripley étendues au cadre non-homogène et spatio-temporel permettent en particulier de mesurer l'agrégation versus la régularité et l'interaction spatio-temporelle, guidant ainsi le choix de modèle. La simulation de processus ponctuels spatio-temporels est ensuite un outil utile, d'une part pour comprendre le comportement de modèles et d'autre part comme composante nécessaire aux méthodes d'inference de type Monte Carlo. Dans cet article, nous présentons les résultats sur les caractéristiques d'ordre deux de processus ponctuels spatio-temporels inhomogènes, puis nous discutons l'influence des méthodes de correction des effets de bord sur des estimateurs non-paramétriques de ces caractéristiques. Enfin, nous présentons un ensemble de modèles de processus ponctuels spatio-temporels implémentés dans le package R 'stpp'.

Estimation de l'interaction du premier ordre d'un processus ponctuel spatial d'interaction de paires de portée finie

Nadia Morsli (LJK, Grenoble)

À partir d'une réalisation d'un processus ponctuel d'interaction de paires homogène observée sur un domaine borné de R^d ($d \geq 1$), on décrit la procédure d'estimation du terme d'interaction du premier ordre (qui peut être aussi appelé l'intensité de Poisson) de son intensité conditionnelle de Papangelou. L'idée sur laquelle l'estimation est basée permet, sous l'hypothèse d'une portée d'interaction finie, de négliger les termes d'interaction d'ordre supérieur. Les propriétés asymptotiques de l'estimateur sont prouvées et une étude par simulations illustre la performance de l'estimateur sur une fenêtre d'observation finie.

Sismicité et modélisation pour l'arc des petites antilles

Larissa Valmy (Université Antilles-Guyane), Jean Vaillant (Université des Antilles-Guyane)

Les répartitions d'occurrences d'événements sismiques peuvent être modélisées à l'aide de processus ponctuels spatio-temporels marqués (Ogata, 1998 ; Ogata et Zhuang, 2006). Parmi les modèles proposés dans la littérature, le plus utilisé est le ETAS (Epidemic Type Aftershock Sequence) suggéré par Ogata en 1998 et ayant fait l'objet de diverses extensions. Il est en partie basé sur sur la loi de Gutenberg-Richter et celle d'Omori modifiée. Les justifications selon la région sismique dépendent des types de faille ou plaques tectoniques présents dans la région étudiée. L'étude du lien entre séismes majeurs, magnitudes et nombre attendu de répliques se fait par le biais de fonctions dites 'déclenchantes'. En utilisant les techniques d'inférence statistique appropriées, nous menons une étude comparative de diverses fonctions déclenchantes, dont la loi d'Omori modifiée, dans le cas de la sismicité de l'arc des Petites Antilles.

Algorithme VBEM pour le processus de Cox log gaussien

Julia Radoszycki (INRA Toulouse), Nathalie Peyrard (INRA Toulouse), Régis Sabbadin (INRA Toulouse)

Le processus de Cox log gaussien est un modèle classiquement utilisé pour représenter les interactions spatiales dans les cartes de comptages. Il a été utilisé, par exemple, pour la modélisation de comptages de plantes adventices (mauvaises herbes des parcelles cultivées) sur des parcelles divisées en quadrats. Un algorithme de type MCMC a été proposé récemment pour estimer le mode a posteriori du vecteur de paramètres et du champ gaussien caché qui contient l'information spatiale sur la répartition de communautés d'espèces adventices. L'inconvénient des algorithmes MCMC est qu'ils sont coûteux en temps de calcul. A l'inverse, les méthodes variationnelles sont connues pour conduire à des algorithmes plus rapides et efficaces en pratique. L'algorithme VBEM (Variational Bayesian Expectation Maximization), par exemple, exploite ce principe. En pratique, il est nécessaire de spécifier la mise en oeuvre des étapes E et M de l'algorithme VBEM pour le processus de Cox log gaussien. Nous proposons une spécification de cet algorithme dans le cas d'une fonction de covariance exponentielle, basée sur une hypothèse de type champ moyen et des simulations de Monte-Carlo. Des expériences sur données simulées montrent que l'algorithme VBEM proposé est aussi performant (sauf pour l'estimation du paramètre de covariance) et beaucoup plus rapide que l'algorithme MCMC existant.

Computer experiments, 11h20-12h40.

Maximum de vraisemblance et validation croisée pour l'estimation des hyper-paramètres de covariance pour le krigeage

François Bachoc (CEA et Université Paris VII), Josselin Garnier (Universite Paris VII), Jean-Marc Martinez (CEA)

Dans le cadre de l'estimation de la fonction de covariance pour un modèle de Krigeage, nous étudions les estimateurs du Maximum de Vraisemblance (MV) et de la Validation Croisée (VC). Dans un premier temps, nous montrons que, lorsque la famille paramétrique de fonctions de covariance est mal spécifiée, la VC est plus robuste que le MV. Nous étudions d'abord analytiquement le cas de l'estimation d'un unique paramètre de variance, puis nous étudions numériquement le cas général. Dans un second temps, nous considérons, dans le cas où la famille paramétrique est bien spécifiée, l'impact asymptotique de la régularité du plan d'expériences sur les estimateurs par MV et VC. Le plan d'expériences est une grille régulière parfaite aléatoirement perturbée, le degré de perturbation étant fonction d'un unique paramètre scalaire de régularité. Nous prouvons alors la consistance et la normalité asymptotique, pour les estimateurs du MV et de la VC. Les matrices de covariance asymptotique sont des fonctions déterministes du paramètre de régularité. Par une étude exhaustive de ces matrices, nous montrons que l'irrégularité du plan d'expériences est souvent un avantage pour l'estimation, mais nous identifions des cas pour lesquels cela est faux. Ainsi, nous répondons par la négative à l'assertion selon laquelle un plan d'expériences irrégulier est toujours meilleur qu'un plan d'expériences régulier pour l'estimation des hyper-paramètres de covariance.

NorMalité asymptotique d'un estimateur des indices de Sobol dans un contexte de krigeage avec bruit d'observations

Loïc Le Gratiet (Université Paris VII), Josselin Garnier (Université Paris VII)

Les gros codes de calcul sont souvent utilisés en science et en ingénierie pour étudier des systèmes physiques complexes. Ces codes ont dans de nombreux cas un grand nombre de paramètres d'entrée. L'analyse de sensibilité a pour objectif de hiérarchiser leur influence sur la sortie du modèle. Un outil souvent utilisé pour faire une telle analyse est la décomposition de la variance de Hoeffding-Sobol. Cependant, cette méthode requiert un très grand nombre de simulations. Les codes étant souvent extrêmement coûteux en temps de calcul, une approximation de leur relation entrée/sortie est généralement construite. Cette approximation, appelée méta-modèle, est ensuite utilisée pour effectuer l'analyse de sensibilité.

Approche bayésienne pour l'estimation d'indices de Sobol

Benoit Jan (Supélec), Julien Bect (Supélec), Emmanuel Vazquez (Supélec), Pierre Lefranc (Grenoble Electrical Engineering Lab (G2Elab), Power Electronics group)

Le problème considéré est l'estimation des indices de Sobol du premier ordre d'une fonction réelle f coûteuse à évaluer, à partir d'un nombre réduit d'évaluations. Nous nous intéressons à la loi a posteriori de ces indices, lorsque f est modélisée par un processus gaussien. Nous montrons qu'il peut être risqué de procéder à une estimation de ces distributions par une approche de type plug-in pour les hyperparamètres du processus gaussien - l'incertitude sur ces hyperparamètres pouvant constituer une part importante de l'incertitude sur les indices de Sobol - et qu'il est préférable d'employer une approche complétement bayésienne. Nos propos sont illustrés sur un exemple académique, puis sur un cas-test issu de l'électronique de puissance.

Analyse de sensibilité pour modèles à variables d'entrée dépendantes

Gaelle Chastaing (Laboratoire Jean Kuntzmann, Grenoble), Fabrice Gamboa (Université Paul Sabatier, Toulouse), Clémentine Prieur (Laboratoire Jean Kuntzmann, Grenoble)

Dans un modèle de régression, les paramètres d'entrée peuvent être à l'origine d'une importante variabilité de la sortie. L'analyse de sensibilité permet de repérer les principales sources d'incertitude d'un modèle, c'est-à-dire d'identifier les variables d'entrée les plus influentes. Les indices de Sobol, dont leur construction repose sur la décomposition de la variance de la sortie, sont les plus fréquemment utilisés pour atteindre de tels objectifs. Néanmoins, leur définition se base sur la décomposition ANOVA fonctionnelle de la sortie pour des variables d'entrée indépendantes. Le but de cet exposé est de présenter une généralisation de cette décomposition au cas où les entrées sont dépendantes. Sous cette contrainte, la fonction du modèle peut se décomposer en une unique somme de fonctions dont l'unicité est garantie par des conditions d'orthogonalité spécifiques. De cette décomposition, il en découle la définition d'indices de sensibilité généralisés, permettant de quantifier la contribution des entrées dans le modèle. Nous proposons également une méthode d'estimation de ces nouveaux indices. En utilisant les bases de fonctions usuelles, nous construisons de nouvelles bases de projection pour chaque terme de la décomposition. Ainsi, chaque composante est approximée par une combinaison linéaire de fonctions adaptées, dont les coefficients sont obtenus par moindres carrés. Nous illustrerons notre méthode par des exemples numériques.

Statistique non paramétrique 3, 11h20-12h40.

Un test de convexité du support de la densité

Catherine Aaron (Université Blaise Pascal)

Soit une densité (inconnue) f à support S compact dans $\mathbb{R}^d$ et $\mathcal{X}_n$ un $n-$échantillon i.i.d. issu de f. On intéresse à la construction d'un test de convexité du support. Dans un premier temps on réalisera le test en supposant que la densité est uniforme sur son support et dans un deuxième temps on supposera seulement que la densité est "à falaise". Dans les deux cas on montrera comment majorer la $p-$value à partir des observations.

Débruitage de chaos par ondelettes : théorie et applications

Matthieu Garcin (Université Paris 1), Dominique Guégan (Université Paris 1)

Nous nous intéressons à des signaux chaotiques décrits par un attracteur, lequel est perturbé de manière non-linéaire par du bruit. Ce bruit est un bruit dynamique ou un bruit de mesure et est modélisé par des variables aléatoires alpha-stables. En utilisant des ondelettes, nous pouvons construire des estimateurs du signal pur. Nous proposons plusieurs sortes d'estimateurs, selon que le signal est connu (oracle), qu'il appartient à un ensemble (minimax) ou qu'il est décrit par une loi de probabilité (Bayes). Ces résultats sont obtenus grâce à la description précise de la loi de probabilité de tous les coefficients d'ondelette de l'attracteur bruité. Nous proposons ensuite des exemples et des applications, notamment aux problèmes inverses.

Estimation de densité par noyau bêta bivarié avec structure de corrélation

Sobom M. Somé (Université de Franche-Comté), Francial G. Libengué (Université de Franche-Comté), Célestin C. Kokonendji (Université de Franche-Comté)

L'objet de cette communication est de présenter le noyau bêta bivarié avec la structure de corrélation introduite par Sarmanov. Ce noyau associé est conçu par une variante de la méthode mode-dispersion et est utilisé pour l'estimation de densités sur [0, 1]x[0, 1] . Des propriétés théoriques de l'estimateur sont examinées, en particulier les biais de bordure et ensuite comparées à ceux du cas produit (sans structure de corrélation). Une étude par simulation ainsi qu'une application aux données réelles seront présentées avec une sélection de la matrice des fenêtres optimale par validation croisée.

Evaluation de performances des systèmes d'attente par la méthode non paramétrique du noyau adaptée

Smail Adjabi (LAMOS, Université de Béjaïa), Karima Lagha (LAMOS, Université de Béjaïa)

L'évaluation de performances d'un système de files d'attente lorsque les lois des inter-arrivées et de service sont générales (quelconques) est complexe car on a peu d'information sur ces lois. Pour cela, on estime ces lois par la méthode non paramétrique du noyau de Parzen-Rosenblatt adaptée au cas où la taille de l'échantillon est aléatoire, ensuite, on évalue les performances de ce système. Les résultats obtenus sur

les caractéristiques de performances par simulation sur des systèmes du type G/M/1, M/G/1 et G/G/1 montrent l'intérêt de l'estimateur de la densité par la méthode du noyau adaptée.

Génétique / Génomique 2, 11h20-12h40.

Sélection de marqueurs biologiques pour la détection d'interaction de gènes

Chloé Friguet (LMBA - Univ. de Bretagne Sud), Mathieu Emily (IRMAR - Universite Rennes 2)

Nous proposons une nouvelle méthode de sélection de marqueurs biologiques permettant la détection d'interaction de gènes en association avec le développement de maladies complexes. A partir d'un ensemble de marqueurs, notre méthode extrait un sous-ensemble de marqueurs qui caractérise de façon optimale la variabilité de la totalité des couples de marqueurs. Nous quantifions la corrélation d'un couple quelconque de marqueurs par un couple de marqueurs sélectionnés par l'information mutuelle normalisée. La faisabilité de notre méthode a été démontrée sur un ensemble de jeu de données simulé à partir d'un jeu de données de référence. De plus, la comparaison de notre méthode avec les stratégies existantes démontre d'une part la puissance de notre méthode et d'autre part un meilleur contrôle de la proportion de faux positifs.

Comparison of approaches for metagenomic biomarker discovery

Remi Brazeilles (Danone Research), Pascale Rondeau (Danone Research), Kim-Anh Le Cao (University of Queensland)

The recent advances in high-throughput assays and the decreasing cost of sequencing technologies enable clinical diagnostics through the comparison of microbial communities. Metagenomic biomarker discovery involves the identification of microorganisms of any uncultured sample whose relative abundances differ between groups of samples. The statistical analysis of such data include many of the challenges encountered in genomics data analysis, such as high dimensionality, small sample size and noisy data. In addition, metagenomic data are characterized by a large amount of null or very high presence counts, high inter-subject variability and comply underlying biology. We propose to compare recently proposed methodologies especially developped for metagenomics analysis (LEfSe and Metastats) with existing approaches based on multivariate analysis and projection to latent structure (Partial Least Squares Discriminant Analysis) and Machine Learning tools (Random Forests) on four publicly available microbial stud- ies. Our study focuses on the feature selections and robustness of the approaches as well as the biological relevance of the selected organisms.

Etude de la répartition des mutations le long du génome à l'aide de chaînes de markov cachées : exemple de la bactérie borrelia sp

Sylvain Coly (INRA), Myriam Charras-Garrido (INRA), David Abrial (INRA)

La borréliose de Lyme est une maladie dont le nombre de cas augmentent sensiblement d'année en année. Cependant, son étude est complexe en raison de symptômes variés et d'espèces-hôtes nombreuses. On a séquencé 63 génomes de son agent causal, la bactérie Borrelia burgdorferi, pour obtenir des informations sur sa dispersion, son adaptation à l'hôte, sa dynamique d'évolution. L'objectif de l'étude est de déterminer le long du génome des segments de variabilité différente, c'est-à-dire des zones avec des

concentrations distinctes de mutations. En effet, cette classification des zones du génome renseigne sur la fonction des gènes qui y sont situés. On utilise pour cela un modèle de Markov caché. Cette méthode a recours aux algorithmes de Baum-Welch, pour estimer les paramètres du modèle markovien, et de Viterbi, pour donner la chaîne des états cachés la plus probable en phase avec notre jeu de séquences. De plus, les paramètres de lois estimés par l'algorithme de Baum-Welch permettent de caractériser les différents états obtenus. L'analyse des séquences de comparaisons a permis de mettre en évidence la forte variabilité du gène ospC, ainsi que d'autres sections du génome. Ce procédé paraît pertinent pour obtenir une partition a priori d'une chaîne ADN. Il fournit également des indications sur les fonctionnalités des gènes qu'elle contient.

The effect of molecular information on the estimation of genetic variance components by linear mixed model

Fabrizio Assenza (INRA UR0631), Jean-Michel Elsen (INRA UR0631), Andres Legarra (INRA UR0631), Clément Carré (INRA UR0631), Guillaume Sallé (INRA UMR1282), Christele Robert-Granié (INRA UR0631), Carole Moreno (INRA UR0631)

Recent developments both in genotyping technology and computational power made access to genomic information affordable enough to be accounted among the sources of information routinely available for the computation of the relatedness between individuals. Which is needed for the estimation of genetic variance components by linear mixed models. We present hereby the effect of including molecular information on the precision of some genetic parameter estimates from real data. The heritabilities and the genetic correlations between growth traits and resistance to gastrointestinal parasites traits were estimated, together with their approximate standard errors and their empirical sampling distributions, according to two different models : one including pedigree information only and the other including both pedigree and molecular information. The estimates obtained by including molecular information were more precise than those obtained by pedigree only, furthermore the two models didn't always converge on the same estimate.

Finance / Économétrie, 11h20-12h40.

Estimation de rangs conditionnels et tests d'exogénéité dans des modèles nonparamétriques et nonséparables

Ingrid Van Keilegom (UCL), Frédérique Fève (Toulouse School of Economics), Jean-Pierre Florens (Toulouse School of Economics)

Considérons un modèle de régression nonparamétrique et nonséparable Y=g(Z,U), où g(Z,U) est croissante en U, et U est uniforme sur l'intervalle [0,1]. Nous supposons qu'il existe un instrument W indépendant de U. Les variables observables sont Y, Z et W, toutes univariées. Le but de ce papier est double. D'abord, nous étudions les propriétés asymptotiques d'un estimateur à noyau de la distribution de V = F(Y—Z), qui est égale à U quand Z est exogène. Nous montrons que cet estimateur converge vers une loi uniforme à une vitesse plus rapide que la vitesse paramétrique. Ensuite, nous construisons des statistiques de test pour l'hypothèse que Z est exogène. Les statistiques de test sont basées sur le fait que Z est exogène si et seulement si V est indépendant de W, et donc elles ne nécessitent pas l'estimation de la fonction g. Les propriétés asymptotiques des tests proposés sont établies, et nous montrons via des simulations qu'une approximation par bootstrap des valeurs critiques des tests fonctionne bien pour petits

échantillons. Nous donnons aussi un exemple empirique en utilisant les données venant du 'U.K. Family Expenditure Survey'.

Une méthode itérative pour estimer les paramètres de modèles définis par des moments conditionnels

Weiyu Li (CREST-Ensai), Valentin Patilea (CREST- Ensai)

Nombreux modèles statistiques et économétriques peuvent s'écrire sous la forme d'une équation de moment conditionnel. Une approche usuelle pour estimer les paramètres de tels modèles consiste à remplacer le moment conditionnel par un ensemble fini suffisamment riche de moments non conditionnels (marginaux) et appliquer la méthode de moments généralisée (GMM en anglais). Cette approche n'assure pas la consistance de l'estimateur car la vraie valeur du paramètre n'est pas toujours identifiée. Motivés par ce problème, nombreux travaux récents sont parus dans la littérature statistique et économétrique proposant des approches consistantes alternatives à la méthode GMM. Voir par exemple Lavergne et Patilea (2008) pour un panorama de la littérature. Dans leur article, Lavergne et Patilea ont également proposé une nouvelle approche, qu'ils ont appelé emphsmooth minimum distance (SMD), basée sur la minimisation d'un contraste non linéaire. Dans notre travail, nous proposons une version itérative de la méthode SMD basée sur une approximation quadratique du contraste et nous étudions ses propriétés. Cette version itérative permet le calcul explicit de l'estimateur à chaque itération et par conséquent elle peut s'implémenter facilement. A l'aide des simulations, nous comparons la nouvelle méthode itérative avec de méthodes classiques d'estimation (moindres carrés, maximum de vraisemblance, GMM).

Approximations of distributions using the scaled Laplace transform

Robert Mnatsakanov (West Virginia University), Khachatur Sarkisian (National Institute for Occupational Safety and Health)

Le problème de reconstitution d'une fonction de répartition et d'une densité de probabilité d'une variable aléatoire positive via la transfomée inverse de Laplace, est étudié. La convergence uniforme des approximations correspondante est dérivée. En utilisant une simulation nous reconstituons les distributions dans le problème du decomposé et dans la probabilité de ruine dans le modèle classique du risque.

Agrégation des comportements individuels dans le contexte des modèles à tirages représentatifs polynomiaux

Frédéric Verschueren (IWEPS)

Pour capter au mieux l'hétérogénéité dans les choix des agents économiques, la macroéconomie moderne fait largement appel aux fondements de la théorie microéconomique dans ses propositions de modélisation. Toutefois, deux obstacles freinent l'usage systématique de ces macromodèles récents dans l'analyse appliquée : la non-disponibilité de statistiques individuelles appropriées et la complexité mathématique de certains modèles microéconomiques. A cet égard, peu de travaux se sont intéressés aux implications empiriques de l'agrégation des comportements individuels. De plus, les équations macroéconomiques proposées possèdent certaines limites (nécessité de données individuelles, hypothèse d'un nombre d'acteurs infini). Pour surmonter ces difficultés, notre approche introduit les modèles à tirages représentatifs

polynomiaux. Chaque décision individuelle est ainsi reliée au rang de l'individu dans l'échantillon via un polynôme. L'agrégation de ces décisions aboutit à une forme exacte pour le macromodèle, dont l' estimation à l'aide de séries temporelles agrégées permet de reconstituer l'évolution de la distribution empirique des données individuelles. Une première application de la méthodologie concerne l'évaluation du taux de pauvreté des travailleurs. La seconde se rapporte à la théorie de l'investissement irréversible qui prévoit pour certaines firmes un régime de non investissement. Les séries chronologiques proviennent des comptes nationaux belges.

Christophe Ambroise, 14h10-15h10.

Statistical Analysis of biological Networks

Christophe Ambroise (Université d'Evry)

Networks are a straightforward formalism for representing interactions between objects of interest and are thus used in many scientific fields. For instance, in Biology, regulatory networks allows to describe the regulation of gene expression through transcriptional factors, while metabolic networks focus on representing pathways of biochemical reactions. Besides, the binding procedures of proteins are often described as protein-protein interaction networks.

In this presentation we discuss two related aspects of network study : the statistical inference of networks using biological high-throughput data and the use of these inferred network for gaining biological insight via clustering.

We will discuss inference through sparse Gaussian Graphical Models (GGM), which give a sounded representation of direct relationships between biological objects (gene, protein, reaction...) and are accompanied with sparse inference strategies well suited to the high dimensional setting. They are also versatile enough to include prior structural knowledge to drive the inference. Still, GGM are now in need for a second breath after showing some limitations : among other questionings, the state-of-the-art reconstruction strategies often suffer a lack of robustness.

One possibility to further explore networks consists in studying their group structure. This means that nodes can be spread into latent classes having similar connectivity patterns which can provide insight ! . We present and discuss mixture model based statistical tools dedicated to uncover latent network structures.

Jan Johannes, 14h10-15h10.

Linear inverse problems with noise in the operator :Minimax-optimal estimation and adaptation

Jan Johannes (Université catholique Louvain)

Statistical ill-posed inverse problems are becoming increasingly important in a diverse range of disciplines, including geophysics, astronomy, medicine and economics. Roughly speaking, in all of these applications the observable signal g = T f is a transformation of the functional parameter of interest f under a linear operator T. Statistical inference on f based on an estimation of g which usually requires an inversion of T is thus called an inverse problem. The lecture course focuses on statistical ill-posed inverse problems with noise in the operator where neither the signal g nor the linear operator T are known in advance, although they can be estimated from the data. Our objective in this context is the construction

of minimax-optimal fully data-driven estimation procedures of the unknown function f. Special attention is given to four models and their extensions, namely Gaussian inverse regression, density deconvolution, functional linear regression and nonparametric instrumental regression, which lead naturally to statistical ill-posed inverse problems with noise in the operator.

Statistique mathématique 3, 15h15-16h35.

Asymptotic behavior of the whittle estimator for the increments of a Rosenblatt process

Jean-Marc Bardet (Université Paris 1), Ciprian Tudor (Université Lille 1)

The purpose of this paper is to estimate the self-similarity index of the Rosenblatt process by using the Whittle estimator. Via chaos expansion into multiple stochastic integrals, we establish a non-central limit theorem satisfied by this estimator. We illustrate our results by numerical simulations.

Estimation de la loi stationnaire des chaînes semi-markoviennes

Vlad Stefan Barbu (LMRS, Université de Rouen), Jan Bulla (Laboratoire de Mathématiques Nicolas Oresme, Université de Caen), Antonello Maruotti (Dipartimento di Istituzioni Pubbliche, Economia e Società, Università di Roma Tre)

Dans cette présentation nous nous intéressons à l'estimation de la loi stationnaire d'un processus semi-markovien à temps discret. Après avoir présenté brièvement le cadre semi-markovien à temps discret, nous proposons un estimateur de la loi stationnaire associée. Le résultat principal concerne les propriétés asymptotiques de cet estimateur, lorsque la taille de l'échantillon augmente. Nous illustrons les propriétés asymptotiques des estimateurs par un exemple numérique. Estimer cette loi stationnaire est une question importante, au moins pour deux raisons. D'abord, d'un point de vue théorique, on est toujours intéressés par le comportement à l'équilibre d'un processus (lorsque cet équilibre existe). Ensuite, quand un certain phénomène a commencé depuis un moment suffisamment lointain, on peut toujours considérer qu'il ait atteint cet état d'équilibre lorsqu'on commence effectivement à l'observer. Dans ce cas-là, il est justifié de prendre la loi stationnaire comme étant la loi initiale du processus. Comme cette loi initiale intervient dans la plupart des quantités liées au processus, il est important d'être en mesure d'estimer cette loi stationnaire, par des estimateurs ayant de bonnes propriétés asymptotiques.

Le processus de Ornstein-Uhlenbeck engendré par le processus de Ornstein-Uhlenbeck

Frédéric Proïa (INRIA Bordeaux Sud-Ouest), Bernard Bercu (Institut Mathématiques de Bordeaux), Nicolas Savy (Université Paul Sabatier)

Nous étudions le processus de Ornstein-Uhlenbeck lorsque la perturbation aléatoire est elle-même engendrée par un processus de Ornstein-Uhlenbeck. Nous analysons l'ergodicité du processus et nous établissons ensuite la convergence presque sûre ainsi que la distribution asymptotique des estimateurs du maximum de vraisemblance du modèle dans le cas stable, puis dans le cas instable où la variance résiduelle est explosive. Nous proposons pour conclure une procédure statistique permettant de tester l'hypothèse

que les résidus d'un processus de Ornstein-Uhlenbeck ont un comportement Brownien, ou bien qu'ils sont eux-mêmes issus d'un processus de Ornstein-Uhlenbeck stable.

Comportement asymptotique de l'estimateur non paramétrique de la fonction de renouvellement associée à des variables aléatoires positives stationnaires béta-mélangeantes

Michel Harel (IMT, Toulouse), Fy Ravelomanantsoa (Université d'Antananarivo)

Nous estimons la fonction de renouvellement, associée à des variables aléatoires positives stationnaires béta-mélangeantes, par une somme finie de fonctions de répartition empiriques. Nous étudions ses propriétés asymptotiques : normalité asymptotique, convergence presque sûre ainsi que sa convergence faible. Cette dernière sera traitée dans l'espace des fonctions définies sur [0,1] continues à droite et admettant une limite à gauche muni de la topologie de Skorokhod.

STID / Enseignement, 15h15-16h35.

Quel est le bagage statistique de nos futurs étudiants ?

Vincent Vandewalle (IUT Roubaix, Université Lille 2)

Le mot statistique apparaît désormais dès les programmes de mathématiques de collège. Sa présence a également connu un essor important dans les programmes de lycée avec l'introduction de la loi normale en terminale. Les bacheliers de 2013 sortiront donc avec un bagage statistique plus important que leurs prédécesseurs. Ils auront déjà étudié des aspects de statistique descriptive univariée mais aussi des aspects inférentiels à travers le calcul d'intervalles de confiance. Ils auront déjà utilisé des tableurs et recouru à de nombreuses simulations pour illustrer les notions introduites. Nous retracerons ici les enseignements dispensés du collège jusqu'au lycée, puis nous mettrons en relief les implications des nouveaux programmes de collèges et lycées pour l'enseignement en DUT STatistique et Informatique Décisionnelle (STID).

Enseignement statistique et monde professionnel : illustration d'un lien fort au travers d'un cours de scoring

Jean-Philippe Kienner (BPCE)

Depuis 10 ans, en parallèle d'une carrière dans des grands groupes (télécom, banque), je m'occupe d'enseignements appliquant les méthodes statistiques aux problématiques marketing (IUT de Paris, Ensai). C'est le cas en particulier d'un cours de scoring que je vais utiliser comme fil rouge dans cet exposé afin de présenter des exemples concrets de liens entre enseignement statistique et monde professionnel.

Nouveaux outils pour la visualisation de données

François-Xavier Jollois (Université Paris Descartes)

Depuis longtemps, les aspects de visualisation sont utiles pour décrire des situations, des événements ou tout autre aspect du monde qui nous entoure. Dernièrement, nous avons vu apparaître un grand nombre d'outils disponibles en ligne pour permettre la réalisation de graphiques, voire d'infographies.

Parallèlement, un grand nombre de personnes se sont mises à produire des visualisations sur des domaines très variés : socio-économie, éducation, environnement, loisirs, ... Nous arborerons dans l'exposé ces nouveaux outils et leurs usages, dans lesquels la statistique a toute sa place.

Stage sur la comparaison de méthodes de redressement

Amélie Rodrigues (Etudiante)

L'étude réalisée durant ce stage montre que la nouvelle méthode de redressement parue en 2010 fait ses preuves. Comparée à la méthode traditionnelle, elle permet de réduire le nombre de variables dans le calage afin de prendre l'information la plus pertinente. Pour cela, elle utilise les axes principaux de l'analyse en composante principale des données. Elle se sert des coordonnées de chaques individus sur les premiers axes, et non plus de leurs valeurs pour chaques variables quantitatives. Cette méthode atteint son objectif en limitant le nombre de variables de calage sans augmenter la variance, mais aussi elle nous apporte de meilleurs résultats en matière de précision.

Apprentissage 4, 15h15-16h35.

Un algorithme de classification et de sélection de variables simultanées pour discriminer une variable qualitative avec un grand nombre de modalités

Olivier Collignon (CRP Santé), Jean-Marie Monnez (Institut Elie Cartan de Lorraine)

En apprentissage supervisé, le nombre de modalités d'une variable qualitative que l'on souhaite expliquer peut être grand. Lorsque certaines de ces modalités ont des effectifs faibles, les regrouper en un nombre plus petit de classes peut s'avérer utile afin de conduire des analyses discriminantes pertinentes. Par ailleurs, sélectionner les variables discriminantes est une étape cruciale pour construire des méthodes de classement efficaces, en particulier lorsque le nombre de variables mesurées est grand comparé à la taille de l'échantillon. Afin de résoudre ces problèmes, nous avions proposé dans une étude précédente un algorithme qui regroupe simultanément les modalités à discriminer en un nombre réduit de classes et qui cherche de manière ascendante les variables qui expliquent au mieux ce regroupement par une minimisation alternée du Lambda de Wilks. Dans cet article, nous proposons deux améliorations de cet algorithme. Tout d'abord, l'AIC (Akaike Information Criterion) a été ajouté comme nouvelle mesure d'optimalité afin de pouvoir appliquer une régression logistique à la classification et aux variables sélectionnées par l'algorithme. Ensuite, la sélection ascendante des variables a été convertie en une sélection pas-à-pas de manière à retirer des prédicteurs qui deviendraient inutiles au cours de la progression de l'algorithme. Nous illustrons ensuite notre algorithme sur un jeu de données simulées.

Error rate control for classification rules in multi-class mixture models

Tristan Mary-Huard (INRA-AgroParisTech), Vittorio Perduca (Paris Descartes), Marie-Laure Martin-Magniette (INRA), Blanchard Gilles (Universität Potsdam)

Consider a sample of observations stemming from P populations. The distribution of this sample can be modelled as a mixture of distributions, each population being described by its own probability density distribution f_p and weight π_p. In many cases the labels associated with each observation are unobserved,

and the goal is to find a suitable division of the sample. This is obtained by applying to the sample a classification rule, i.e. a function that maps an observation X into $1, ..., P$. The Maximum A Posteriori classification rule is the most popular classification rule, known to minimize the prediction error. While optimality in the sense of prediction error is a desirable property, two drawbacks of the MAP rule should be mentioned. First, optimality does not prevent against a high level of misclassification. Second, it does not account for the possible asymmetry between classes. In real applications, a small number among the P classes may be of major importance for the experimenter, meaning that one should focus about the misclassification rate of the rule to be used on these classes of interest. Based on the control of Type 1 error rate in testing methodology, we propose a methodology to obtain classification rules that guaranty that as many observations as possible are classified in the classes of interest, under the constraint that the proportion of misclassified observations in these classes is controlled. Optimal rules and heuristic procedures to estimate these rules are exhibited, and applied to the analysis of ChIP-chip data.

Stabilité des classifieurs neuronaux relativement au classifieur de bayes

Ibtissem Ben Othman (ENSI, CRISTAL, GRIFT), Faouzi Ghorbel (Laboratoire CRISTAL, pôle GRIF)

Malgré l'utilisation massive des classifieurs basés réseaux de neurones dans les applications industrielles, leur formulation mathématique reste difficile à expliciter. Cela explique le peu de travaux modélisant de manière formelle de l'instabilité de leurs résultats de classifications bien reconnue. En se référant au point de vue statistique de la classification, nous tenterons dans le présent travail d'évaluer le degré de stabilité des réseaux de neurones Bayésiens et cela en les comparant aux méthodes statistiques et neuronales classiques au sens du biais et de la variance des taux d'erreur.

Diagnostic par fusion de décisions binaires corrélées

André Smolarz (UTT - Troyes), Pierre Beauseroy (UTT - Troyes), Yuan Dong (UTT - Troyes)

En considérant le diagnostic d'un système comme un problème de décision à deux classes, ≪ fonctionnement normal ≫ et ≪ fonctionnement anormal ≫, nous proposons d'étudier la performance globale d'une règle de décision en modélisant et en fusionnant les sorties d'un ensemble de classifieurs en fonction des performances individuelles de chacun d'entre eux et de leurs corrélations. Dans ce contexte nous considérons des mesures issues d'un ensemble de capteurs et chaque classifieur est alors défini sur un sous-espace de mesures issues d'un sous ensemble de capteurs choisis au hasard parmi tous les capteurs.

Sélection de variables, 15h15-16h35.

Modèles mixtes en génétique animale : sélection de variables par optimisation combinatoire

Julie Hamon (INRIA Lille - Gènes Diffusion), Clarisse Dhaenens (LIFL / Université Lille 1 / Inria), Julien Jacques (Painlevé / CNRS / Université Lille1 / Inria), Gaël Even (Gènes Diffusion)

La sélection génomique est basée sur un grand nombre de marqueurs couvrant l'ensemble du génome. Grâce notamment aux nouvelles technologies de séquençage haut-débit, nous sommes maintenant capable de lire l'information génomique sur près de 800 000 marqueurs. Un des enjeux consiste à identifier un

sous-ensemble de marqueurs génomiques explicatifs pour un trait d'intérêt quantitatif (phénotype). La spécificité des études animales nécessite l'utilisation de modèles mixtes, du fait des liens de parenté entre individus. Nous proposons d'effectuer, dans ce cadre, une sélection des marqueurs d'intérêt à l'aide de méthodes d'optimisation combinatoire. Des premiers résultats sur simulations montrent l'intérêt de l'utilisation de modèles mixtes comparés aux méthodes ne prenant pas en compte les relations familiales.

Sélection d'effets fixes dans les modèles linéaires mixtes de grande dimension

Florian Rohart (Institut Mathématiques de Toulouse), Magali San Cristobal (INRA-INSA de Toulouse), Béatrice Laurent (IMT, INSA Toulouse)

On se place dans le cadre du modèle linéaire mixte dans lequel les observations sont structurées. On propose l'ajout d'une pénalisation l1 portant sur les effets fixes dans la log-vraisemblance complétée, obtenue en considérant les effets aléatoires comme des données manquantes. Un algorithme "multicycle ECM" est utilisé pour résoudre le problème d'optimisation ; cet algorithme peut être combiné à n'importe quelle méthode de sélection de variables développée pour le modèle linéaire classique. La méthode proposée fonctionne lorsque le nombre de paramètres p est plus grand que le nombre d'observations n ; elle est plus rapide que le lmmLasso (Schelldorfer (2011)) puisque ne nécessitant pas l'inversion d'une matrice de taille $n \times n$ à chaque itération du processus de convergence. Des résultats théoriques sont fournis dans le cas où les variances des effets aléatoires et de la résiduelle sont connues. La combinaison de l'algorithme avec la méthode procbol (Rohart (2011)) donne de très bons résultats sur l'estimation de l'ensemble des effets fixes ainsi que l'estimation des variances ; ces résultats sont meilleurs que ceux du lmmLasso, en petite dimension $p< n$ mais aussi en grande dimension $p \ll n$.

Group lasso et inégalités oracles dans le cadre du modèle linéaire généralisé

Mélanie Blazère (Institut de Mathématiques de Toulouse), Jean-Michel Loubès (Institut de Mathématiques de Toulouse), Fabrice Gamboa (Institut Mathématiques de Toulouse)

Aujourd'hui nous avons accés à des bases de données de plus en plus grandes ce qui nous amène à travailler avec des modèles de régression en très grandes dimensions. Sous une hypothèse de parcimonie on peut espérer estimer correctement le paramètre de régression. Cette hypothèse est assez naturelle en génomique de même que l'hypothèse qui nous permet de considérer des blocs de variables ayant des comportements similaires. Il existe une riche littérature concernant le modèle linéaire gaussien avec des estimateurs tels que le Lasso ou le Group Lasso (qui permet de prendre en compte des groupes de covariables). Nous étendrons ici cette procédure Group Lasso au modèle linéaire généralisé et nous étudierons les propriétés asymptotiques de cet estimateur dans le cadre d'un modèle linéaire généralisé parcimonieux en grande dimension. Nous en déduirons des inégalités oracles pour l'erreur de prédiction ainsi que pour celle d'estimation sous des hypothèses concernant la loi jointe des variables observées, la matrice d'incidence ainsi que sous une hypothèse de parcimonie et nous montrerons la capacité de cet estimateur à construire de bonnes approximations parcimonieuses du vrai modèle. Nous détaillerons plus particulièrement ces inégalités dans le cas du modèle de régression poissonien qui contrairement au modèle logistique n'a pas été étudié dans le cadre parcimonieux.

Stabilité de la sélection de variables pour la classification de données en grande dimension

Emeline Perthame (Agrocampus Ouest), Chloé Friguet (LMBA - Univ. de Bretagne Sud), David Causeur (Agrocampus Ouest)

Les données à haut-débit ont motivé le développement de méthodes statistiques pour la sélection de variables. Ces données sont caractérisées par leur grande dimension et par leur hétérogénéité car le signal est souvent observé simultanément à plusieurs facteurs de confusion. Les approches habituelles sont ainsi remises en question car elles peuvent conduire à des décisions erronées. Plusieurs articles récents montrent l'impact négatif de l'hétérogénéité des données sur le nombre de faux-positifs des tests multiples. On s'intéresse ici aux performances de classification de la sélection de variables, via la procédure LASSO (Tibshirani (1996)) en termes de reproductibilité des ensembles de variables sélectionnés. Des simulations montrent que l'ensemble des variables sélectionnées par le LASSO n'est pas nécessairement celui des prédicteurs attendus. Aussi, d'intéressantes performances de classification ne sont atteintes que pour un nombre trop important de variables sélectionnées. Notre méthode s'appuie sur la description de la dépendance entre covariables grâce à un petit nombre de variables latentes (Friguet et al. (2009)). La stratégie proposée consiste à appliquer les procédures sur les données conditionnellement à cette structure de dépendance. Cette stratégie permet de stabiliser les variables sélectionnées : d'intéressantes performances de classification sont atteintes pour de plus petits ensembles de variables et les variables les plus prédictives sont détectées.

Plan d'expériences et quantification optimale, 15h15-15h55.

Application de la quantification optimale à l'estimation de quantiles conditionnels

Isabelle Charlier (ULB-ECARES/Bordeaux 1-Inria), Davy Paindaveine (ULB - ECARES), Jérôme Saracco (IMB, INRIA)

Les applications les plus courantes des méthodes non-paramétriques concernent l'estimation d'une fonction de régression (i.e. de l'espérance conditionnelle). Cependant, il est souvent intéressant de modéliser les quantiles conditionnels, en particulier lorsque la moyenne conditionnelle ne permet pas de représenter convenablement l'impact des covariables sur la variable dépendante. Dès les années 1950, la 'quantification' était utilisée en ingénierie et permettait de discrétiser un signal continu grâce à un nombre fini de 'quantifieurs'. En mathématiques, le problème de la quantification optimale consiste à trouver la meilleure approximation d'une distribution continue d'une variable aléatoire par une loi discrète. Le but de ce travail est d'appliquer la quantification optimale en norme Lp à l'estimation des quantiles conditionnels. La convergence de l'estimateur proposé est étudiée d'un point de vue théorique et celui-ci a été implémenté dans R afin d'en évaluer le comportement numérique et de le comparer à des méthodes existantes.

Plans optimaux pour l'estimation des effets totaux en présence d'autovoisinages et de blocs à structure non-circulaire

Walter Tinsson (Université de PAU), Pierre Druilhet (Université Blaise Pascal,)

On considère ici des plans pour effets de voisinage tenant compte aussi d'un effet supplémentaire d'autovoisinage. Une méthode de construction de plans d'expérience universellement optimaux pour les effets totaux lorsque les blocs n'ont pas de structure circulaire est proposée.

Image, 15h55-16h35.

Traitement d'image par régression non paramétrique

Benoît Thieurmel (Greenwich Statistics), Pierre-André Cornillon (Université Rennes 2), Nicolas Hengartner (Los Alamos National Laboratory), Eric Matzner-Lober (Agrocampus Ouest et Université Rennes 2)

Les estimateurs non paramétriques sont impactés par le fléau de la dimension. Les estimateurs adaptatifs peuvent exploiter la régularité inconnue de la fonction à estimer et si cette régularité est élevée peuvent partiellement diminuer le fléau de la dimension. Nous présentons une procédure itérative basée sur l'estimateur non paramétrique à noyau classique qui peut être utilisée quand le nombre de variable explicative est élevée. Nous considèrerons l'application de cette méthode au débruitage d'images.

Analyse parcimonieuse des données d'irm fonctionnelle dans un cadre bayésien variationnel

Christine Bakhous (INRIA Rhône-Alpes), Florence Forbes (INRIA Rhone-Alpes), Farida Enikeeva (INRIA Rhône-Alpes), Thomas Vincent (INRIA Rhône-Alpes), Michel Dojat (INSERM U836), Philippe Ciuciu (CEA/INRIA)

L'analyse des données d'Imagerie par Résonance Magnétique fonctionnelle (IRMf) est principalement effectuée à travers le modèle linéaire général (GLM) dans lequel l'activité d'une région cérébrale est supposée dépendre des différents types de stimuli (moteur, visuel, etc.) or la spécialisation fonctionnelle cérébrale indique que l'activation d'une région donnée n'est induite que par certains de ces stimuli. Inclure des conditions non pertinentes peut dégrader les résultats, en particulier quand la fonction de réponse hémodynamique (FRH) est conjointement estimée. De plus la sélection a priori des conditions pertinentes pour chaque région cérébrale n'est pas toujours possible (e.g. pathologie). Afin de faire face à ces difficultés, nous proposons une procédure variationnelle efficace permettant la sélection automatique des conditions selon l'activité cérébrale qu'elles suscitent. Une amélioration de la détection d'activation ainsi que de l'estimation de la FRH sont illustrées sur données réelles.

Table ronde enseignement, 16h55-17h55.

"Quelles données accessibles pour des applications pédagogiques ? Quid des open data ?" Table ronde du groupe enseignement de la statistique de la SFdS

Marthe-Aline Jutand (ISPED - Univ Bordeaux Segalen), Frédérique Letué (IUT2/STID UPMF et LJK, université de Grenoble)

L'enseignement de la statistique ne peut pas être un enseignement purement théorique et nécessite donc d'utiliser des fichiers de données pour les applications. Il s'avère que pour l'enseignant de statistique intervenant en économie, en santé, en biologie ... avoir accès à des données ayant un intérêt pour les étudiants et avec lesquelles il pourra illustrer ou introduire les concepts de son enseignement n'est pas toujours un exercice facile. L'objet de la table ronde est d'aider l'enseignant dans sa quête aux données. La table ronde débutera donc par une présentation d'un certain nombre d'espaces ouverts proposant des fichiers de données, en soulignant leurs spécificités, ainsi que leurs points forts pour un usage pédagogique. Il sera intéressant de s'interroger dans le cadre de cette table ronde sur l'impact de l'accessibilité libre de ces données, publiques ou non, sur l'usage de fichiers de données dans l'enseignement universitaire. Peut-on repérer des changements de pratiques pédagogiques suite à la mise en place de ces open-data ? Comment les enseignants utilisent ces sources de données ? Quelles sont les pratiques rencontrées et sont-elles différentes selon les domaines d'applications ? Les participants pourront alors réfléchir à la possibilité de créer une communauté de pratiques et d'échanges de fichiers à visée pédagogique.

Fiabilité et incertitudes 2, 16h55-17h55.

Cartes de contrôles non paramétriques adaptées à des distributions asymétriques et fondées sur les statistiques des prédécesseurs

Jean-Christophe Turlot (Université de Pau (UPPA)), Christian Paroissin (Université de Pau), Narayanaswamy Balakrishnan (Mc Master University)

On propose des cartes de contrôle non paramétriques, fondées sur la statistique des prédécesseurs et plus généralement sur les statistiques de placement, pour des problèmes de tests unilatéraux. Ces cartes de contrôle sont adaptées à des distributions asymétriques, aux distributions de durée de vie par exemple. Il s'agit de tester $G = F$ contre $G\<F$, où F est la distribution des observations sous contrôle et G la distribution des observations d'un échantillon à tester. De telles cartes de contrôle ne semblent pas encore avoir été proposées dans le cadre normatif. Cette communication a pour objet de présenter deux cartes de contrôle fondées sur les statistiques de placement : ces deux types de cartes se ramènent à des versions pondérées de statistiques de placement. On calcule les valeurs critiques ou limites de contrôle, ainsi que la période opérationnelle moyenne de ces deux types de cartes sous l'hypothèse nulle, mais aussi leur puissance pour les deux familles d'alternatives proposées par Lehmann. Des résultats numériques complètent la présentation.

Modélisation de la fiabilité de matériels exposés aux surtensions atmosphériques

Lise Guérineau (EDF - Université de Bretagne Sud), Evans Gouno (Université de Bretagne Sud)

L'objectif de ce travail est d'évaluer la fiabilité d'un matériel électrique implanté en différents points d'une zone géographique pour laquelle on a recueilli des informations concernant l'activité orageuse (localisation et intensité des impacts de foudre). Ce matériel est donc exposé à des surtensions liées à des épisodes orageux qui vont dégrader sa durée de bon fonctionnement. Dans un premier temps, à partir de données météorologiques, les chroniques des impacts de foudre sont modélisées. Puis, on associe à la durée de bon fonctionnement du matériel un taux de défaillance constant par morceaux. On propose alors des estimations par maximum de vraisemblance des paramètres de cette fonction. Les propriétés des estimateurs sont étudiées sur des données simulées et des résultats sur des données de terrain sont présentés.

A Chi-squared type goodness-of-fit test for the Kumaraswamy-log-logistic distribution.

Noureddine Saaidia (Badji Mokhtar University - Univ. de Bordeaux), Xuan Quang Tran (University of Bordeaux, IMB, UMR 5251, F-33400 Talence), Ramzan Tahir (Department of social and preventive medicine University of Montreal)

In this paper a modified Chi-square goodness-of-fit test based on Rao-Robson-Nikulin statistic is developed for the Kumaraswamy-log-logistic distribution by using the maximum likelihood estimators.

Statistique d'enquête 2, 16h55-17h55.

Les redressements dans les enquêtes auprès des entreprises : spécificités et pratiques à l'Insee

Emmanuel Gros (Insee)

La statistique institutionnelle distingue traditionnellement statistique auprès des entreprises et statistique auprès des ménages. Cette séparation, qui apparaît clairement dans les organigrammes des instituts nationaux statistiques, et se retrouve, plus ou moins explicitement, dans les colloques et les publications, découle de l'existence de spécificités fortes, propres à chaque univers. Dans la sphère entreprise, ces spécificités - univers très hétérogène et de taille relativement restreinte, problème de charge statistique, collecte par voie postale ou par Internet, etc. - affectent le processus d'enquête aussi bien en amont - lors de la procédure d'échantillonnage, cf. Fizzala (2013) pour plus de détails - qu'en aval - à l'occasion des procédures de contrôles et de redressement des données - de la phase de collecte. Ceci fera d'ailleurs l'objet d'un article de Demoly, Fizzala et Gros (2013) dans un numéro spécial du journal de la SFdS sur le sujet "Enquêtes et Échantillonnage". Dans cette présentation, nous nous focalisons sur les spécificités relatives aux redressements dans les enquêtes entreprises : caractérisation des unités n'ayant pas retourné de questionnaire, correction de la non-réponse, gestion des unités atypiques et calage.

Une méthode de détermination du seuil pour la winsorisation

Cyril Favre-Martinoz (Crest-Ensai), Jean-François Beaumont (Statistique Canada), David Haziza (Université de Montréal, Statistique Canada)

Dans les enquêtes entreprises, il est courant d'observer des variables dont la distribution est fortement asymétrique. Les unités ayant une valeur élevée et une probabilité d'inclusion faible ont une grande influence sur les estimateurs classiques. En présence d'unités influentes, les estimateurs classiques demeurent sans biais mais leur variance peut être élevée. La winsorisation est une méthode souvent employée dans les enquêtes entreprises dans le but de traiter les valeurs influentes. Cette méthode consiste à réduire la valeur et/ou le poids d'une unité afin de réduire l'impact de cette unité sur l'estimation, et rendre l'estimateur utilisé plus robuste. Le fait de réduire le poids ou la valeur d'une unité engendre un léger biais, mais contribue à une diminution de la variance. La winsorisation requiert la détermination d'une tuning constante, qui correspond au seuil auquel les unités avec une grande valeur sont réduites. Au delà du type de winsorisation considéré, le choix de ce seuil est crucial, car il permet de réaliser le compromis biais-variance. Kokic et Bell (1994) et Rivest et Hurtubise (1995) ont proposé des méthodes pour déterminer le seuil optimal dans le cas d'un sondage aléatoire simple stratifié. Nous développerons au cours de cet exposé, l'approche de type 'min-max',développée dans l'article de Beaumont, Haziza et Ruiz-Gazen (2013).

Approximation des probabilités d'inclusion du plan de poisson conditionnel et applications

Anne Ruiz-Gazen (TSE - Université de Toulouse), Hélène Boistard (TSE - Université de Toulouse), Henrik Lopuhaä (Delft University of Technology)

Nous nous intéressons au plan de Poisson conditionnel encore appelé plan de sondage réjectif. Nous proposons d'étendre à un ordre quelconque les résultats de Hajek (1981) sur l'approximation des probabilités d'inclusion d'ordre deux de ce plan en fonction des probabilités d'inclusion d'ordre un. La démonstration suit les mêmes lignes que Hajek (1981) et utilise des expansions d'Edgeworth. Le résultat permet de vérifier des conditions qui conduisent à la consistance et à la normalité asymptotique d'estimateurs complexes en sondages.

Climat 1, 16h55-17h55.

Un générateur stochastique de séries pluviométriques pour la désagrégation de données journalières en données horaires

Marc Bourotte (BioSP INRA Avignon), Denis Allard (BioSP INRA Avignon)

Les générateurs stochastiques de climat (SWG) permettent de simuler des séries de variables météorologiques (ici les précipitations) aussi longues que désiré, ayant des propriétés statistiques similaires aux séries de données mesurées. Ce travail présente un modèle stochastique pour les données de pluviométrie, simple à mettre en ?'uvre, capable de reproduire les statistiques des précipitations observées pour un site donné à différentes échelles temporelles (d'horaire à journalier). Ce modèle permet la désagrégation des précipitations journalières en données horaires. Les séries pluviométriques sont caractérisées par une alternance de périodes sèches et pluvieuses et par l'intensité des événements pluvieux.

Le modèle est basé sur un processus gaussien dont les valeurs supérieures à un certain un seuil, transformées par une fonction exponentielle-puissance, modélisent l'intensité pluvieuse. Les valeurs inférieures à ce seuil correspondent à une pluviométrie nulle. Nous détaillons la procédure d'estimation par maximum de vraisemblance composite puis nous illustrons le générateur et la méthode de désagrégation sur quelques stations pluviométriques.

Un modèle espace-état linéaire gaussien pour les vitesses de vent en atlantique nord-est

Julie Bessac (IRMAR Université Rennes 1), Pierre Ailliot (Université Brest), Valérie Monbet (IRMAR Université Rennes 1)

Ces dernières décennies, la construction de générateurs aléatoires de conditions météorologiques s'est considérablement développée. Un des principaux objectifs de ces derniers est de pallier le manque de disponibilité des données réelles en générant des séquences artificielles de conditions météorologiques réalistes. Ces séquences peuvent être utilisées pour conduire des études de risques liés au climat (érosion côtière, énergies renouvelables). Nous proposons un modèle stochastique spatio-temporel pour les champs de vent à l'échelle régionale en Atlantique Nord-Est. Beaucoup de modèles pour les vitesses de vent sont basés sur des modèles de type AutoRegressive Moving Average, cependant les modèles à espace d'états se sont révélés plus flexibles en terme de modélisation temporelle que les modèles ARMA. Nous utiliserons un modèle linéaire gaussien à espace d'états dans lequel l'état caché représente un vent moyen à l'échelle régionale et l'équation d'observation permet de relier les échelles régionale et locale. Un des objectifs de ce modèle est de reproduire les déplacements en temps et en espace de systèmes météorologiques comme la propagation des tempêtes. La procédure d'estimation des paramètres est basée sur le maximum de vraisemblance et la validation du modèle se fera par simulation et par prédiction.

Améliorer la calibration des prévisions probabilistes de températures extrêmes par la régression quantile

Jérôme Collet (EDF)

La température influence à la fois la demande et l'offre d'électricité, et est de ce fait une cause potentielle de blackout.Comme tout fournisseur d'électricité, EDF doit donc modéliser l'incertitude sur les températures futures, en utilisant les Systèmes de Prévision d'Ensemble (SPE). On constate que la représentation probabiliste fournie par les SPE n'est pas parfaitement fiable, en particulier quant aux températures extrêmes, qui par ailleurs engendrent les situations les plus tendues. Une méthode possible pour résoudre ce problème est la méthode du meilleur membre (BMM) : cette méthode améliore globalement la représentation, mais il reste des défauts dans les extrêmes. Le principe de BMM est de modéliser la distribution de la différence entre prévision et réalisation. Nous améliorons cette modélisation en utilisant la régression quantile, plus efficace que la modélisation habituelle, qui utilise deux régressions par moindres carrés successives.

Classification non supervisée 2, 16h55-17h55.

Classification approach based on association rules mining for unbalanced data : application to in-hospital maternal mortality in Sénégal and Mali

Cheikh Ndour (University Gaston Berger ; University Pau et des pays de l'Ardour), Simplice Dossou-Gbété (University Pau et des pays de l'Ardour, UMR CNRS 5142)

This paper deals with the supervised classification when the response variable is binary and its class distribution is unbalanced. In such situation, it is not possible to build a powerful classifier by using standard methods such as logistic regression, classification tree, discriminant analysis, etc. To overcome this short-coming of these methods that provide classifiers with low sensibility, we tackled the classification problem here through an approach based on the association rules learning. Association rules learning is a well known method in the area of data-mining. It is used when dealing with large database for unsupervised discovery of local patterns that expresses hidden relationships between variables. In considering association rules from a supervised learning point of view, a relevant set of weak classifiers is obtained from which one derives a classification rule that performs well. Moreover this approach has the advantage of allowing the identification of the patterns that are well correlated with the target class.

Comparaison de deux approches classificatoires pour la détermination d'une typologie des journées de l'île de la réunion en fonction du rayonnement solaire

Miloud Bessafi (ILE2P, Université de la Réunion), Francisco de A. T. De Carvalho (CIn/UFPE, Recife-PE), Philippe Charton (LIM, Université de la Réunion), Mathieu Delsaut (LIM, Université de la Réunion), Thierry Despeyroux (INRIA, Paris-Rocquencourt), Patrick Jeanty (LE2P, Université de la Réunion), Jean-Daniel Lan-Sun-Luk (LE2P, Université de la Réunion), Yves Lechevallier (INRIA Paris-Rocquencourt), Henri Ralambondrainy (LIM, Université de la Réunion), Lionel Trovalet (LE2P, Université de la Réunion)

L'objectif de cet article est de montrer les intérêts et les inconvénients de deux approches classificatoires de courbes. La première est basée sur une représentation des courbes sous forme vectorielle. Dans ce cas la dimension de l'espace vectoriel de description dépend du nombre de plages de discrétisation. Nous utilisons les plages de discrétisation au lieu des points de discrétisation car nous voulons utiliser un indicateur lié à la production d'énergie. La seconde propose la distance D'Urso et Vichi qui est basée sur première et seconde dérivées finies. Cette distance permet d'intégrer plusieurs propriétés mathématiques des courbes et en particulier l'ordre total associé aux plages de discrétisation. Ces deux approches seront appliquées à la classification des journées de l'île de La Réunion en fonction d'un ensemble de sources de production d'énergie de type photovoltaïque.

Sélection de variables en classification et application à l'analyse des risques aériens

Baptiste Gregorutti (Safety Line / LSTA, Université Pierre et Marie Curie)

La société Safety Line propose aux compagnies aériennes des solutions statistiques visant à minimiser les risques d'accidents. Dans ce but, les données contenues dans les enregistreurs de vol sont analysées grâce à des méthodes de classification supervisée. Cependant, le nombre d'enregistrement peut varier de quelques centaines à plus d'un millier par seconde. Nous avons donc recours à des algorithmes de

sélection de variables. Plus précisément, nous adaptons la méthode SVM Recursive Feature Elimination (SVM-RFE) de Guyon et al (2002) aux forêts aléatoires dans le cas où les données d'entrée sont fonctionnelles. Néanmoins, il a été observé que la présence de variables fortement corrélées induit une instabilité dans la sélection de variables. Nous discuterons donc des performances des différents algorithmes connus, notamment en les appliquant aux données de vol. Nous proposerons une alternative robuste permettant d'exhiber les variables les plus influentes sur la classification.

Résumés du vendredi 31 mai 2013

Anthony Davison, 08h30-09h30.

Modélisation spatiale des pluies extrêmes

Anthony Davison (EPFL), Emeric Thibaud (EPFL)

L'analyse des risques associés aux changements climatiques a soulevé la question de la modélisation des événements rares complexes, qui est récemment devenu un sujet de recherche très actif. Dans cet exposé je discuterai de la modélisation spatiale des pluies extrêmes avec les processus max-stables, ajustés en utilisant des excès de seuil. Ces modèles sont consistants avec la théorie classique des extrêmes, et donnent la possibilité d'un traitement cohérent de la dépendance spatiale des pluies extrêmes en utilisant des idées analogues à celles de la géostatistique classique. La méthodologie est illustrée par une analyse de données provenantes d'un bassin versant dans les Alpes Suisses, près du col du Grand Saint Bernard. Nous comparons des modèles max-stables et asymptotiquement indépendants pour ces données, et montrons que cette approche pourrait être utile pour la simulation spatiale de pluie extrême et ainsi pour l'estimation du risque associé à de tels événements rares.

Mark van de Wiel, 08h30-09h30.

ShrinkBayes : Bayesian analysis of high-dimensional data using shrinkage priors

Mark A. van de Wiel (VU University Medical Center)

We present a Bayesian framework that allows for a) various likelihood models b) flexible designs c) random effects and d) multi-parameter shrinkage. The latter is implemented using Empirical Bayes principles by several procedures that estimate hyper-parameters of (mixture) priors or nonparametric priors. The framework provides Bayesian multiplicity correction, by means of a Bayesian False Discovery Rate estimate (BFDR ; Ventrucci et al., 2011).

Petra Friederichs, 09h30-10h30.

Probabilistic forecasting using a mesoscale ensemble weather predictions system with special emphasis on extremes

Petra Friederichs (University of Bonn), Sabrina Bentzien (University of Bonn)

Due to large uncertainties, predictions of high-impact weather are probabilistic in nature. Recently developed mesoscale ensemble prediction systems (EPS) provide such predictions. An EPS not only issues a deterministic future state of the atmosphere but a sample of possible future states. Ensemble postprocessing then translates such a sample of forecasts into probabilistic measures. This study presents and discusses approaches of ensemble postprocessing, and states that ensemble postprocessing should be an integral part of any EPS. In order to provide area-wide forecasts of extreme wind and heavy precipitation we employ quantile and logistic regression for precipitation, and a Bayesian hierarchical model for wind gust prediction. We further discuss the issue of forecast verification, where performance is measured using proper scoring rules and their decomposition.

David Haziza, 09h30-10h30.

Inférence robuste en présence de valeurs influentes dans les enquêtes

David Haziza (Université Montréal et CREST-ENSAI)

Dans les enquêtes auprès des entreprises, il est courant de collecter des variables économiques dont la distribution est fortement asymétrique. Dans ce contexte, on est souvent confronté à la présence de valeurs influentes dans l'échantillon tiré. Ces dernières sont habituellement de très grandes valeurs dont la présence dans l'échantillon tend à rendre les estimateurs classiques très instables. La présentation sera essentiellement basée sur deux articles : Beaumont et al. (2013) et Favre Martinoz et al. (2013). On considère une classe d'estimateurs robustes à la présence de valeurs influentes qui comprend les estimateurs winsorisés de type I et II ainsi que l'estimateur robuste proposé par Beaumont et al. (2013). Ces estimateurs requièrent détermination deune constante qui correspond au seuil à partir duquel les valeurs influentes sont réduites. Nous considérons une méthode de détermination de la constante qui consiste à minimiser le plus grand biais conditionnel estimé de l'échantillon. Les résultats d'une étude par simulation suggèrent que les méthodes proposées conduisent à des estimateurs robustes ayant de bonnes propriétés en termes de biais et d'efficacité relative.

Statistique d'enquête 3, 10h50-12h10.

La procédure d'échantillonnage dans les enquêtes auprès des entreprises : spécificités et pratiques à l'insee

Arnaud Fizzala (Insee)

La statistique institutionnelle distingue traditionnellement statistique auprès des entreprises et statistique auprès des ménages. Cette séparation, qui apparaît clairement dans les organigrammes des instituts nationaux statistiques, et se retrouve, plus ou moins explicitement, dans les colloques et les publications,

découle de l'existence de spécificités fortes, propres à chaque univers. Dans la sphère entreprises, ces spécificités - univers très hétérogène et de taille relativement restreinte, problème de charge statistique, collecte par voie postale ou par Internet, etc. - affectent le processus d'enquête aussi bien en amont - lors de la procédure d'échantillonnage - qu'en aval - à l'occasion des procédures de contrôles et de redressement des données, cf. Gros (2013) pour plus de détails - de la phase de collecte. Ceci fera d'ailleurs l'objet d'un article de Demoly, Fizzala et Gros (2013) dans un numéro spécial du journal de la SFdS sur le sujet "Enquêtes et Échantillonnage". Dans cette présentation, nous nous focalisons sur les spécificités relatives à la procédure d'échantillonnage dans les enquêtes entreprises : construction d'une base de sondage, stratification, détermination des taux de sondage et tirage de l'échantillon.

Redressement d'un estimateur de la consommation d'eau potable d'une population

Karim Claudio (IMB / LyRE), Vincent Couallier (IMB, Bordeaux), Yves Le Gat (IRSTEA), Jérôme Saracco (IMB, INRIA)

Les méthodes de redressement permettent d'améliorer la qualité d'un estimateur par l'exploitation d'une variable auxiliaire. Lorsque le sondage initial est un sondage aléatoire simple, la plus-value de ces méthodes a souvent été démontrée, en comparaison à l'estimateur non redressé. En va-t-il de même lorsque le sondage initial est un sondage stratifié ? Cette communication a pour but de comparer trois méthodes de redressement (post-stratification, régression et calage) consécutives à un sondage stratifié fait dans un cadre industriel portant sur l'estimation de la consommation hebdomadaire en eau. Cette estimation doit être la plus précise possible car elle doit permettre une évaluation fine des pertes en réseau d'eau potable.

Les enquêtes multimode : multi-solution ou multi-problème ?

Gaël de Peretti (Insee), Tiaray Razafindranovona (Insee)

L'Insee doit faire face à une demande d'enquêtes toujours plus exigeante en termes de qualité (précision, pertinence, comparabilité, cohérence, clarté, fraîcheur), dans un contexte général de restriction budgétaire. Une solution envisagée est l'utilisation d'internet comme mode privilégié ou complémentaire de recueil des données du fait de son avantage en termes de coût, que ce soit dans ses dimensions financières ou de temps. Il s'agirait de développer le recours à la collecte par internet, et plus généralement multimode, pour faire face aux problèmes de couverture, d'échantillonnage, de non-réponse ou de mesure, tout en respectant des contraintes budgétaires. L'objectif serait de maximiser la qualité d'enquête dans toutes les dimensions énoncées précédemment, en particulier en termes d'erreur d'enquête totale, en limitant voire diminuant les coûts d'enquête. Cependant, si la collecte par internet est un mode peu coûteux, elle pose des problèmes méthodologiques non négligeables : couverture, auto-sélection ou biais de sélection, non-réponse et les difficultés de sa correction, "satisficing", etc. Aussi, avant de développer ou généraliser l'utilisation du multimode, l'Insee s'est lancé dans une vaste opération d'expérimentations afin d'étudier ces différentes questions méthodologiques. Ces premières expérimentations apportent des réponses mais soulèvent de nouveaux problèmes.

Une application de l'analyse harmonique non commutative à l'analyse de saillance des vignettes-étalons : le cas de 3 vignettes

Salim Lardjane (Université de Bretagne Sud)

Les techniques présentées dans ce travail, basées sur la notion de représentation linéaire d'un groupe non commutatif, permettent d'étudier les problèmes de cohérence des réponse dans les enquêtes subjectives de Santé dont le questionnaire comprend trois vignettes-étalons. L'analyse des réponses des individus aux questions associées aux vignettes-étalons permet de définir une distribution de fréquences sur le groupe des permutations de trois éléments. L'analyse harmonique de cette distribution de fréquences permet d'en obtenir une décomposition sur une base orthonormée adaptée à la structure de groupe. L'identification des harmoniques les plus importantes et l'interprétation de celles-ci permet alors d'identifier les vignettes-étalons critiques en termes de problèmes de cohérence des réponses. La méthodologie proposée est illustrée par une application à des données de Santé issues de l'enquête Share 2004.

Données de survie 2, 10h50-12h10.

Estimation non paramétrique d'une fonction de répartition multivariée en présence de censure à droite et de troncature à gauche

Olivier Lopez (Laboraoire de Statistique Théorique et Appliquée, Paris 6)

Nous proposons un nouvel estimateur non paramétrique de la distribution de deux variables aléatoires censurées à droite et tronquées à gauche. L'outil est motivé par l'étude de contrats d'assurance vie portant sur plusieurs têtes. Nous discutons son implémentation, et son utilisation comme outil de validation de modèles de dépendance entre variables. Nous proposons en outre des méthodes bootstrap basées sur cet estimateur qui permettent d'envisager l'incertitude liée à ces méthodes. Les résultats sont illustrés sur des données canadiennes.

Estimation et modélisation par copules pour un modèle de durée simplifié

Svetlana Gribkova (Laboratoire de Statistique Théorique et Appliquée, Paris 6), Olivier Lopez (Laboratoire de Statistique Théorique et Appliquée, Paris 6), Philippe Saint-Pierre (Laboratoire de Statistique Théorique et Appliquée, Paris 6)

On considère un modèle de durée bivarié où la différence entre deux variables de censure est observée. Cette situation est courante en assurance-vie, lorsqu'on étudie la dépendance entre les durées de vie de deux individus qui ont souscrit un contrat de retraite avec une clause de réversion au conjoint. Dans le cadre de ce modèle, nous présenterons des estimateurs nonparametriques de fonction de survie bivariée et de copule qui lie deux durées de vie. On considère des résultats asymptotiques pour les nouveaux estimateurs et une application aux données réelles.

An EM composite likelihood approach based on frailty model for family studies of unknown genetic factors with incomplete genetic data

Laurent Briollais (Samuel Lunenfeld Research Inst), Yun-Hee Choi (Western University)

La vraisemblance composite est une importante méthode inférentielle construite à partir du produit d'une collection de composantes individuelles de la vraisemblance. Récemment, nous avons développé un algorithme EM pour la vraisemblance composite pour l'échantillonnage en plusieurs phases de données familiales. On s'intéresse ici à la modélisation du temps de survenue d'un évènement parmi les porteurs et non-porteurs d'une mutation spécifique d'un gène. On utilise une vraisemblance composite pondérée pour données de survie où les poids sont données par la probabilité inverse d'échantillonnage des familles et les composantes de la vraisemblance sont les unités complètement observées. Les génotypes manquants pour un gène majeur sont incorporés via un algorithme EM. On montre que les équations d'estimation résultantes, obtenues à partir de la dérivée de la log-vraisemblance composite, sont non biaisées. En épimiologie génétique, l'intérêt ne réside pas seulement dans l'estimation des caractéristiques d'un gène majeur mais aussi dans l'estimation des corrélations familiales résiduelles, un effet aléatoire non observé (c.à.d. fragilité). Pour faire une inférence sur le paramètre de fragilité, on propose d'étendre l'algorithme EM pour la vraisemblance composite en utilisant une formulation de vraisemblance basée sur les paires pour tenir compte des corrélations résiduelles entre paires d'apparentés.

Estimation non paramétrique guidée paramétriquement de la fonction de densité et la fonction de hasard avec des données censurées

Majda Talamakrouni (UCL), Ingrid Van Keilegom (UCL)

La méthode d'estimation non paramétrique guidée paramétriquement est une méthode attrayante qui permet de réduire le biais des estimateurs à noyau de la fonction de densité et la fonction de hasard, sans augmenter leur variance. Le but de ce papier est de généraliser cette méthode au cas de données censurées. Les propriétés asymptotiques des estimateurs proposés sont établies et leur performance est évaluée par des simulations.

Climat 2, 10h50-12h10.

Détection et attribution des changements climatiques : problèmes méthodologiques

Aurélien Ribes (CNRM-GAME, Météo France - CNRS), Alexis Hannart (UMI IFAECI, CIMA, Université de Buenos Aires et CNRS), Philippe Naveau (LSCE, IPSL, CNRS), Serge Planton (CNRM-GAME, Météo France - CNRS), Laurent Terray (CERFACS)

Cet exposé a pour but de faire un faire une présentation générale des problèmes de nature statistique rencontrés dans l'estimation des changements climatiques, ainsi que dans la détection et l'attribution de ces changements à différents facteurs externes, parmi lesquels les émissions anthropiques de gaz à effet de serre. On se restreint ici à la description du modèle linéaire le plus couramment utilisé. Dans ce modèle, les principales difficultés concernent la prise en compte des termes d'erreurs sur les prédicteurs, ainsi que la modélisation statistique et l'estimation de la variabilité climatique.

Détection de changements climatiques à l'aide d'un modèle multiplicatif

Jean-Marc Azaïs (IMT Toulouse), Aurélien Ribes (CNRM-Game)

On présente un modèle multiplicatif ou bilinéaire pour décrire l'évolution d'une variable climatique sur une zone donnée et sur une période donnée. L'estimation et les tests dans ce modèle sont abordés à l'aide d'une vraisemblance pénalisée. La convergence des estimateurs est obtenue par des techniques de minimisation alternée, et on propose une méthode de type Monte-Carlo pour apprécier la variabilité. L'application à des données de température montre l'intérêt du modèle multiplicatif spline pour la détection de changement climatique.

Formation de régions homogènes pour l'analyse régionale des aléas maritimes extrêmes

Jérôme Weiss (EDF), Pietro Bernardara (EDF), Michel Benoit (EDF)

L'analyse régionale est une méthode qui permet de réduire les incertitudes sur les estimations des événements extrêmes, en regroupant les sites d'observation dans des régions homogènes. Nous nous intéressons ici à la formation de régions homogènes pour les aléas maritimes extrêmes. Dans le but de donner une interprétation physique aux régions obtenues, nous cherchons à déterminer les groupes de sites dont les extrêmes sont très probablement générés par les mêmes tempêtes. La méthode que nous proposons vise à identifier les empreintes spatiales typiques des tempêtes. Un exemple est donné à partir des séries de hauteurs significatives d'états de mer issues de la base de données numériques ANEMOC.

Statistique mathématique 4, 10h50-12h10.

Procédures optimales à la Le Cam pour le paramètre de kurtosis de lois de Student multivariées

Anouk Neven (Université du Luxembourg), Christophe Ley (Université Libre de Bruxelles)

La loi de Student, fondamentale en inférence statistique et en modélisation de données financières, dépend de trois paramètres : la position, la forme et le kurtosis. En ayant recours à la théorie de Le Cam, nous proposons des tests optimaux (mais simples à implémenter) pour le paramètre de kurtosis de lois de Student multivariées, tout en laissant le centre et la forme non-spécifiés. Ces tests, localement et asymptotiquement maximin et donc aussi puissants que le test du rapport de vraisemblance, l'emportent sur ce dernier par leur simplicité numérique. Nous étudions la distribution asymptotique des statistiques de test sous-jacentes, à la fois sous l'hypothèse nulle et sous des alternatives locales et calculons explicitement la fonction de puissance correspondante. Finalement, nous discutons brièvement les implications de la 'méthodologie de Le Cam' sur l'estimation du paramètre de kurtosis d'une distribution de Student univariée.

Propriété de norMalité asymptotique locale uniforme pour les modèles de loi gaussienne inverse généralisée

Angelo Efoévi Koudou (Université de Lorraine), Christophe Ley (Université Libre de Bruxelles)

Le but de ce travail est d'exhiber une propriété statistique des lois gaussiennes inverses généralisées. Nous démontrons la propriété de normalité asymptotique locale uniforme (propriété ULAN) pour le modèle paramétrique constitué par les lois d'échantillons iid de variables gaussiennes inverses généralisées. La démonstration repose sur la notion de différentiabilité en moyenne quadratique. Cela ouvre la voie, dans des travaux futurs, à l'utilisation de la méthodologie de Le Cam en vue de construire des test optimaux pour les paramètres.

Un modèle d'interactions poissoniennes et détection de dépendance

Laure Sansonnet (Université Paris-Sud), Christine Tuleau-Malot (Université Nice Sophia-Antipolis)

Dans cet exposé, on considère un modèle d'interactions entre deux processus ponctuels, gouverné par une fonction dite de reproduction, que l'on modélise ici par l'intensité d'un processus de Poisson. Ce modèle peut par exemple être utilisé en neurosciences pour étudier les éventuelles interactions entre deux neurones lors de l'activité cérébrale. Les neurobiologistes souhaitant savoir si les deux neurones considérés dans l'étude évoluent indépendamment ou non, on propose de tester la nullité de la fonction d'intensité. On construit alors une procédure de test multiple obtenue par l'agrégation de tests simples basés sur une méthode de seuillage de coefficients d'ondelettes. Ce test a de bonnes propriétés théoriques. En effet, on peut assurer son niveau mais aussi sa puissance sous certaines hypothèses. De plus, le test est adaptatif au sens minimax sur des espaces de Besov faibles. Enfin, on présente des simulations afin de montrer également le bon comportement pratique de la procédure de test.

Test combinatoire d'homogénéité en analyse géométrique des données

Solène Bienaise (Université Paris Dauphine-CEREMADE), Brigitte Le Roux (Université Paris Descartes-MAP5)

Dans ce papier, nous présentons des méthodes d'inférence statistique pour l'analyse géométrique des données (AGD) basées sur des procédures de permutation s'inscrivant dans le cadre de l'inférence combinatoire. Les tests statistiques multivariés s'appuient, la plupart du temps, sur des hypothèses invérifiables et sont donc souvent inapplicables en AGD. C'est pourquoi, il est nécessaire de concevoir des procédures inductives adaptées aux méthodes d'AGD : analyse en composantes principales, analyse des correspondances simples ou multiples. Les méthodes présentées ici s'appliquent à des nuages euclidiens multidimensionnels et traitent de la comparaison des points moyens de plusieurs groupes d'observations (tests d'homogénéité). Nous présenterons d'abord le cas général, puis celui de la comparaison de deux groupes selon qu'ils forment ou non une partition de l'ensemble de départ. Nous présenterons ce test dans le cas général (à plus de deux dimensions) et l'illustrerons pour des nuages uni et bidimensionnel. Pour cela, nous avons écrit des programmes en langage R, qui fournissent les solutions exactes (en utilisant si nécessaire des méthodes de Monte Carlo), mais aussi les solutions approchées.

Enseignement 3, 10h50-12h10.

Tentative d'identification de paradoxes latents dans l'application des probabilités conditionnelles

Léo Gerville-Réache (Université de Bordeaux), Vincent Couallier (IMB, Bordeaux)

L'analyse des "paradoxes" fait partie des problèmes les plus passionnants des sciences. De l'extérieur, on assiste à des discussions vigoureuses entre scientifiques, de l'intérieur, lorsqu'on se prend au jeu, notre esprit est bouleversé entre nos certitudes démontrées et celles, également démontrées, de nos contradicteurs. On se demande souvent comment il est possible que le temps d'attente d'un consensus soit parfois de plusieurs dizaines d'années. A posteriori, lorsqu'un paradoxe est résolu, ou fait clairement consensus, on est souvent étonné, voire stupéfait, de la simplicité de la solution. Les nombreux paradoxes probabilistes sont source d'une production scientifique et non scientifique abondante. Parmi les plus célèbres, on peut citer le paradoxe de St-Petersbourg, le paradoxe du Monty Hall ou encore celui de l'Apocalypse. Les plus déstabilisants, que vous soyez spécialiste ou non, sont sûrement ceux où le calcul d'une probabilité conditionnelle est sous-jacente.

La statistique vue par des étudiants en sciences de l'éducation

Alain Bihan-Poudec (UCO), Jean-Marie Marion (UCO-IMA)

Comme le souligne Fine (2013), définir la statistique ne va pas de soi. Qu'en est-il pour les étudiants en sciences humaines et sociales quand ils abordent l'enseignement de la statistique à l'université ? Nous avons eu l'opportunité de montrer qu'ils associent la statistique aux mathématiques, aux chiffres, aux calculs, aux pourcentages (Bihan-Poudec et Larose, 2010) ; nous avons aussi pu pointer que leurs représentations admettaient des variations en fonction des diplômes préparés (Bihan-Poudec, 2012 ; Bihan-Poudec et Marion, 2012). Pour aller plus avant, nous avons mené une enquête plus approfondie auprès de 147 étudiants en troisième année de Licence en Sciences de l'Éducation, certains étant en formation initiale, d'autres en formation permanente ou reprise d'études. Outre la recherche d'une confirmation d'un positionnement différent au regard de la statistique en fonction du type de formation suivie, nous avons interrogé ces étudiants quant à leur attitude vis-à-vis de la statistique, la définition qu' ils en donnaient, l'éventuel intérêt que cette discipline revêtait pour leurs études, leur exercice professionnel actuel ou futur. Ce sont les résultats les plus prégnants issus de l'analyse de cette enquête qui seront présentés lors de la communication.

Les difficultés de compréhension des risques d'erreur au regard des croyances épistémiques des étudiants

Anne-Cécile Wauthy (Université de Namur), Benoit Bihin (Université de Namur), Marc Romainville (Université de Namur), Eric Depiereux (Université de Namur)

Depuis de nombreuses années, notre dispositif didactique n'a cessé d'évoluer. Il échoue cependant toujours au niveau de la maîtrise de la gestion des risques d'erreur dans les tests d'hypothèses et les probabilités alpha et beta associées. En décembre 2012, 156 étudiants de deuxième année de Baccalauréat ont été interrogés sur leur compréhension des risques d'erreur dans les tests d'hypothèses ainsi que sur leurs croyances épistémiques. Ces étudiants ont complété le questionnaire sur les croyances épistémiques

de M. Schommer ainsi qu'un questionnaire sur les risques d'erreur et les probabilités qui y sont liées. Cet exposé présente l'analyse des résultats à ces questionnaires en insistant sur le lien entre le type de croyance épistémique des étudiants et leur compréhension des risques d'erreurs et des probabilités alpha et beta associées.

Évolution de la conception de la moyenne chez les étudiants en sciences humaines et sociales

Christelle Chevallier-Gaté (LUNAM, UCO, ISCEA, Angers), Véronique Dubreil-Frémont (LUNAM, UCO, IPSA, Angers)

Un champ de recherche dans l'enseignement de la statistique est l'identification des conceptions que les étudiants ont des notions statistiques (pour revue cf. Shaughnessy, 2007). Lors de nos précédents travaux, nous avions montré que la notion de moyenne perçue par les étudiants eux-mêmes se déclinait de façon très diverse et parfois erronée (Dubreil-Frémont et al., 2012). Les résultats présentés alors corroborent ceux de la littérature scientifique (Gattuso et Mary, 1996), notamment la prédominance de la conception algorithmique sur les autres (Pollatsek et al., 1981). Or si les conceptions de la moyenne sont bien identifiées, leur évolution n'a guère fait l'objet d'étude. Tout semble se passer comme si les préconceptions identifiées précédemment disparaissaient suite aux enseignements. Nous avons donc questionné cette évidence et observé comment les conceptions de la moyenne évoluaient au cours d'une année universitaire. Nous avons interrogé des étudiants en sciences humaines et sociales avant leur premier cours de statistique et une seconde fois à l'issue de cet enseignement. La communication portera sur les résultats obtenus.

Biopharmacie, 10h50-12h10.

Equations différentielles stochastiques en pharmacocinétique de population : modèles et méthodologie

Maud Delattre (AgroParisTech, UMR 518), Marc Lavielle (Inria Saclay Île de France)

Nous nous intéressons à des modèles de diffusion à effets mixtes, à observations discrètes et bruitées. Ces modèles sont pertinents en pharmacocinétique de population, où on est amené à prendre en compte différentes sources de variabilité dans les données. Dans ces modèles, la vraisemblance des observations n'est généralement pas explicite, rendant l'estimation des paramètres à partir des observations particulièrement complexe. Nous proposons une méthodologie d'inférence pour ces modèles. En particulier, nous combinons l'algorithme SAEM à un filtre de Kalman étendu pour estimer les paramètres de population. Nous proposons également des outils pour estimer les paramètres individuels et les trajectoires latentes des processus de diffusion individuels. Ces méthodes sont évaluées sur des jeux de données simulés et appliquées à un exemple concret en pharmacocinétique de population.

Inférence par approximation normale de l'a posteriori dans les modèles dynamiques à effets mixtes

Mélanie Prague (INSERM U897, ISPED, Bordeaux), Daniel Commenges (INSERM, ISPED, Bordeaux), Jérémie Guedj (INSERM UMR 738, Paris Diderot), Julia Drylewicz (Department of Immunology, Univ. Med. Center Utrecht), Rodolphe Thiébaut (Inserm U897, ISPED, VRI)

Les modèles basés sur des équations différentielles sont des outils très utilisés pour décrire les systèmes dynamiques. L'usage de modèles à effets mixtes pour estimer les paramètres en population vient naturellement et permet de tirer avantage de la variabilité entre les individus. Depuis 2007, un algorithme basé sur une approximation normale de l'a posteriori permet d'estimer les paramètres dans ces problèmes, mais son utilisation difficile a limité son accessibilité. Dans un contexte Bayésien afin de réduire les problèmes d'identifiabilité, cette méthode consiste à calculer le maximum a posteriori (MAP) par une optimisation de la vraisemblance pénalisée. Le programme NIMROD (Normal approximation Inference in Models with Random effects based on Ordinary Differential equations) écrit en Fortran est désormais disponible. Nous décrirons la méthode et les spécificités de cet algorithme : un nouveau critère de convergence appelé la Distance Relative au Maximum (RDM) et un critère de choix de modèle basé sur l' approximation de la validation croisée (LCVa). Les illustrations et la comparaison avec des méthodes classiques MCMC se fera sur des exemples de pharmacocinétique : sur données simulées puis sur des données réelles de l'essai clinique PUZZLE pour le dosage plasmatique de la concentration d'un agent antirétroviral, l'amprenavir.

Planification adaptative en deux étapes pour des modèles non linéaires à effets mixtes : application en pharmacocinétique

Cyrielle Dumont (Paris 7 - UMR 738 INSERM), Marylore Chenel (Institut de Recherches Internationales Servier), France Mentré (INSERM UMR738 -Paris 7)

Les modèles non linéaires à effets mixtes (MNLEM) sont utilisés en pharmacocinétique pour l'analyse des concentrations de patients lors du développement de médicament, notamment pour les études pédiatriques. Des approches basées sur la matrice de Fisher (MF) sont utilisées pour optimiser le protocole de ces études. L'expression de MF pour les MNLEM provient d'une approche de linéarisation au premier ordre et est implémentée dans PFIM en R. L'optimisation locale de protocoles nécessite des valeurs de paramètres a priori, difficiles à appréhender. En conséquence, des protocoles adaptatifs sont utiles pour permettre une flexibilité et ceux en deux étapes semblent plus efficaces que les protocoles adaptatifs complets. Nos objectifs sont : i) développer et implémenter l'optimisation du déterminant de MF pour des protocoles en deux étapes pour les MNLEM ; ii) évaluer, par une approche de simulation, l'impact des protocoles en deux étapes sur la précision d'estimation des paramètres quand les "vrais" paramètres sont différents des paramètres "a priori". L'implémentation de MF pour une planification en deux étapes est réalisée dans une version de travail de PFIM. Nous montrons une moins bonne précision d'estimation des paramètres avec le protocole en une étape, obtenu à partir de paramètres "a priori" différents des "vrais" paramètres. Le protocole en deux étapes permet de compenser ces imprécisions d'estimation.

Une approche de population pour un modèle complexe de glucose/insuline

Célia Barthélémy (INRIA Saclay-Île-de-France), Marc Lavielle (INRIA Saclay-Île-de-France)

Les modèles physiologiques sont souvent des modèles complexes décrits par un très grand nombre d'équations différentielles ordinaires et faisant intervenir un nombre important de paramètres. L'évaluation de ces modèles est donc très coûteuse en temps de calcul. Les algorithmes comme SAEM évaluant un grand nombre de fois le modèle, le temps d'estimation peut être trop long. Nous présentons ici une méthode spécifique d'estimation, adaptée à ces modèles complexes. En particulier, une méthodologie qui combine l'algorithme SAEM, des méthodes de Monte-Carlo et des méthodes d'interpolation dans le but de limiter le nombre total de résolutions numériques du système d'équations différentielles ordinaires. Cette méthodologie a été appliquée à un exemple réaliste afin d'en démontrer l'intérêt auprès des pharmacométriciens qui ont besoin d'un tel outil. L'exemple choisi est un modèle composé de 29 équations différentielles ordinaires, qui permet de simuler la régulation du glucose dans un corps sain.

Index des auteurs

Avec le soutien de