Modèle linéaire mixte avec segmentation : application à la détection de changements dans les dates de vendanges

Franck Picard, Eva Budinska, Stephane Robin

$42^{èmes}$ Journées de la Société Française de Statistique

Marseille 24-28 Mai 2010

Résumés des communications

24 mai 2010

09:00-12:30 : *Accueil - Tutoriel "Environnement",* département de Mathématiques, salle 9009.
Organisateur: Liliane Bel

13:30-17:30 : *Tutoriel "Environnement" (suite),* département de Mathématiques, salle 9009.
Organisateur: Liliane Bel

13:30-17:30 : *Demi Journées STID*
Organisateur: J.M. Poggi

Introduction à la statistique spatiale,
Edith Gabriel ***16***
Enseignement à distance : l'experience du Master statistique et économétrie de Toulouse,
Anne Ruiz-Gazen, Christine Maurel ***16***
Analyse de données avec R - Complémentarité des méthodes d'analyse factorielle et de classification,
François Husson, Julie Josse, Jérôme Pagès ***17***

25 mai 2010

09:00-10:30 : *Accueil-Réception,* bât A

10:30-11:00 : *Ouverture des journées,* Amphi 2

11:00-11:50 : *Séance inaugurale : Paul Champsaur*

11:50-12:40 : *Conférence Le Cam : Emmanuel Candès*

14:30-15:15 : *John P. Nolan*
Stable distributions: models for heavy tailed data.

14:30-15:15 : *J. Y. Dauxois*
Statistique semi-paramétrique ou non paramétrique des risques concurrents.

14:30-16:00 : *Hommage D. Schwartz 1 (en partenariat avec Sanofi-Aventis)*

15:15-16:00 : *Tilman Gneiting*
Matern Cross-Covariance Functions for Multivariate Random Fields.

15:15-16:00 : *Ingo Steinwart*
Theory and practice of kernel-based statistical learning methods.

16:20-18:00 : *Statistique spatiale 1*

Régression quantile spatiale localement linéaire (40min),
Marc Hallin, Zudi Lu, Keming Yu **21**
Robust quantile estimation and prediction for spatial processes,
Baba Thiam, Sophie Dabo-Niang **22**
Régression et prédiction non-paramétrique spatiale,
Sophie Dabo-Niang, Anne-Françoise Yao **22**
Modélisation des dépassements de seuils pour un processus observé à des temps irréguliers,
Nicolas Raillard **22**

16:20-18:00 : *Hommage D. Schwartz 2*

16:20-18:00 : *Chaînes de Markov - Processus*

Estimation de l'ordre d'une chaîne de Markov cachée à émissions de la famille exponentielle,
Cécile Low-Kam, André Mas **19**
Vitesse de convergence en M-estimation de données markoviennes,
Loïc Hervé, James Ledoux, Valentin Patilea **19**
Estimation non paramétrique pénalisée de la dérivée de la densité d'un processus de diffusion,
Emeline Schmisser **20**
Tester si un processus de Poisson est modifié en admettant que certains événements ponctuels se regroupent en classes,
Franz Streit **20**
Modèles de comptage appliqués aux décisions de candidature aux offres d'emploi sur le web,
Julie Séguéla, Gilbert Saporta **69**

16:20-18:00 : *Données censurées - Survie*

Inférence statistique dans des modèles de moments conditionnels en présence de censure,
Pascal Lavergne, Olivier Lopez, Valentin Patilea **25**
Prédiction de la fonction de survie par sélection de modèle,
Ion Grama, Jean-François Petiot **25**
Extensions du test du logrank aux données censurées par intervalle,
Fanny Leroy, Ludovic Trinquart **26**
Analyse bayésienne de données de survie discrètes sujettes à censure informative avec un modèle à effets aléatoires partagés reparamétré.,
Sophie Ancelet-Enjalric, Elise de la Rochebrochard, Jean Bouyer **26**
Modèle de fragilité et algorithme EM stochastique,
Charles El-Nouty, Estelle Kuhn **27**

16:20-18:00 : *Econométrie*

Les prix à terme de l'électricité sur le marché britannique, une approche par la cointégration,
Jérémy Froger, Muriel Renault 23
La tendance de l'économie souterraine, au niveau d'un pays pendant la période de transition,
Andrei Tudorel, Andreea Iluzia Iacob, Claudiu Herteliu 23
La méthode d'évaluation contingente appliquée au ruissellement érosif : une nouvelle approche de l'estimation du consentement à payer,
Dimitri Laroutis, Patrice Lepelletier 24
A test of endogeneity in quantiles,
Kim Tae-Hwan, Christophe Muller 24
Méthodes de détection des unités atypiques : cas des enquêtes structurelles ukrainiennes,
Michel Grun-Rehomme, Olga Vasechko 24

16:20-18:00 : *Biostatistique 1 : génome*

Détection de sélection darwinienne sur un gène par une approche sans vraisemblance,
Aude Grelaud, Christian Robert, François Rodolphe 17
Processus de tests de rapports de vraisemblance pour la détection de QTL,
Charles-Elie Rabier, Jean-Marc Azaïs, Céline Delmas 18
Tests multiples pour la comparaison des probabilités de survenue d'une infidélité de transcription dans des ARNm sains et cancéreux,
Olivier Collignon, Marie Brulliard, Jean-Marie Monnez, Pierre Vallois, Benoit Thouvenot, Sandrine Jacquenet, Virginie Ogier, Olivier Roitel, Bernard E. Bihain 18
Détection d'interaction de gènes à l'échelle du génome,
Mathieu Emily 18
Inférence sur réseaux géniques par Analyse en Facteurs,
Yuna Blum, Chloé Friguet, Sandrine Lagarrigue, David Causeur 19

26 mai 2010

09:15-10:00 : *Soumendra Lahiri*
Frequency Domain Empirical Likelihood Method for Irregularly Spaced Spatial Data.

09:00-17:00 : *Tutoriel "Flux de données",* département de Mathématiques, salle 9015.
Organisateur: Christian Derquennes

09:00-17:00 : *Tutoriel "Incertitudes",* département de Mathématiques, salle 9017.
Organisateur: Christian Derquennes

10:20-11:40 : *Recensements*

Recensement en France : le point de vue de l'INSEE,
François Clanché, Session "Recensements" organisée par le groupe SES 31
Recensements de population : nouvelles méthodes, nouveaux défis,
Jean-François Royer 31
Recensement en France : le point de vue des communes,
Marie-Hélène Boulidard, Session "Recensements" organisée par le groupe SES 32
Recensement en Italie : le point de vue de l'ISTAT,

10:20-12:20 : *Plans d'expérience - Qualité - Fiabilité*

10:20-12:00 : *Séries temporelles 1*

10:20-11:40 : *Statistique spatiale 2 : Champs de Markov*

10:20-12:00 : *PLS*

Analyse canonique généralisée régularisée et approche PLS,
Arthur Tenenhaus, Michel Tenenhaus 28
Kernel Generalized Canonical Correlation Analysis,
Arthur Tenenhaus 28
The Non-Metric Partial Least Squares approach,
Giorgio Russolillo 28
An integrated PLS Regression-based approach for multidimensional blocks in PLS path modeling,
Vincenzo Esposito Vinzi, Giorgio Russolillo, Laura Trinchera 28
La régression multivariée contrainte PLS,
Philippe Casin, Francois Marque 29

10:20-11:40 : *Biostatistique 2 : épidémiologie*

Approche variationnelle pour la cartographie spatio-temporelle du risque en épidémiologie à l'aide de champs de Markov cachés,
David Abrial, Lamiae Azizi, Myriam Charras-Garrido, Florence Forbes 33
Sur les modèles de comptage à inflation de zéro avec application à la modélisation de la tendance dans la prévalence de certaines maladies allergiques liées au travail en France de 2001 à 2007,
Joseph Ngatchou-Wandji, Christophe Paris 33
Surveillance de la contamination chimique en Méditerranée basée sur les capacités accumulatrices de la moule : détermination d'une réponse universelle de capteur,
Marc Bouchoucha, Michel Lhermine, Jean-Claude Franc, François Galgani, Bruno Andral, Pierre Boissery 34
Essais de modélisation de l'épilepsie en Tunisie : théorie et application basées sur des modèles de régression logistique,
Abdelwaheb Daouthi, Mohamed Dogui, Abdeljelil Farhat 34

27 mai 2010

09:00-09:45 : *Prix Simon Laplace : Anestis Antoniadis et Irène Gijbels*

09:45-10:30 : *Catherine Matias*
Réseaux en biologie moléculaire : inférence, analyse, évolution.

09:45-10:30 : *Stéphane Loisel*
Dépendance stochastique en théorie de la ruine et en actuariat.

10:50-12:50 : *Partenariat AXA : Finance - Assurance - Actuariat*

Extremal events in a bank operational losses (40min),
Hela Dahen, Georges Dionne, Daniel Zajdenweber 39
Risques extrêmes en assurance vie,
Johan Attal 39
Modélisation multivariée des comportements à risque des conducteurs d'automobile,
Mériem Maatig 40
Facteurs explicatifs du rachat en Assurance-Vie : classification et prévisions du risque de rachat,

10:50-11:50 : *Statistique des processus de type fractal*

10:50-12:30 : *Sondages et statistique publique*

10:50-12:30 : *Apprentissage - Classification 1*

10:50-12:10 : *Biostatistique 3 : modèles à effets mixtes*

Les modèles de Markov cachés à effets mixtes,
Maud Delattre **54**

Evaluation et optimisation des protocoles de prélèvements dans les études pharmacocinétiques en crossover analysées par des modèles non linéaires à effets mixtes,
Thu Thuy Nguyen, Caroline Bazzoli, France Mentré **55**

Incorporation de fonction de coûts pour l'optimisation de protocoles dans les modèles non linéaires à effets mixtes : application à la pharmacocinétique de la zidovudine et de son métabolite actif,
Caroline Bazzoli, Emanuelle Comets, Sylvie Retout, France Mentré **55**

Amélioration des propriétés de mesure d'un questionnaire de satisfaction des patients hospitalisés : application d'un modèle de mesure à variable latente centrale,
Sophie Tricaud-Vialle, Alain Morineau **56**

14:00-14:45 : *Markus Reiss*

Asymptotic equivalence and sufficiency for volatility estimation under microstructure noise.

14:00-14:45 : *Hein Putter*

Frailties in multi-state models: Are they identifiable? Do we need them?

14:00-14:45 : *Table Ronde (Vie Privée)*

14:45-16:25 : *Image*

Contribution à la squelettisation en niveaux de gris,
Rabaa Youssef, Sylvie Sevestre-Ghalila, Anne Ricordeau **51**

Décisions en environnement non stationnaire par méthodes d'ensembles via One-class SVM - application à la segmentation d'images texturées,
Pierre Beauseroy, André Smolarz, Xiyan He **51**

Réduction de dimension par un nouvel estimateur de la distance de Patrick Fisher à l'aide des fonctions orthogonales,
Faouzi Ghorbel, Wissal Drira **52**

Estimation du paramètre du champ de Ising et fonction de partition,
Jean-François Giovannelli **52**

14:45-16:25 : *Statistique fonctionnelle 1*

Méthode de Lissage Bayésienne Tempérée Pour Estimer les Paramètres d'un Modèle d'Equation Différentielle (40min),
David Campbell, Russell Steele **57**

Découpage de courbes de densité : application au dépistage du cancer,
Fabrice Morlais, Frédéric Ferraty, Philippe Vieu **57**

Prédiction en régression linéaire fonctionnelle avec variable d'intérêt fonctionnelle,
Christophe Crambes, André Mas **57**

Semiparametric models with functional responses in a survey sampling setting : model assisted estimation of electricity consumption curve,
Herve Cardot, Etienne Josserand **58**

14:45-16:25 : *Apprentissage : Classification 2*

14:45-16:25 : *Biostatistique 4 : médecine*

14:45-16:25 : *Environnement - Changement climatique 1*

14:45-16:25 : *Partenariat Danone : Biostatistiques 4 : Tests multiples - Risques alimentaires*

Forces et faiblesses de différentes approches statistiques pour l'analyse d'évènements récurrents,
Jérôme Tanguy, Anissa Elfakir, Sébastien Marque **53**
Application des cartes de Kohonen aux bases de données nutritionnelles,
Sonia Fortin, Pascale Rondeau, Sébastien Marque **53**
Extraction de systèmes de consommation alimentaire utilisant la Nonnegative Matrix Factorization (NMF) pour l'évaluation des choix alimentaires,
Mélanie Zetlaoui, Stéphan Clémençon, Max Feinberg, Philippe Verger **54**

16:45-18:05 : *Extrêmes*

Estimation de courbes de niveaux extrêmes pour des lois à queues lourdes,
Abdelaati Daouia, Laurent Gardes, Stéphane Girard, Alexandre Lekina **65**
Une méthode de folding pour des extrêmes multivariés avec application à l'estimation de la mesure de probabilité spectrale,
Armelle Guillou, Philippe Naveau, Alexandre You **66**
Estimation de quantiles extrêmes et de probabilités d'événements rares,
Arnaud Guyader, Nicolas Hengartner, Eric Matzner-Lober **66**
Estimation de queues bivariées,
Elena Di Bernardino, Véronique Maume-Deschamps, Clémentine Prieur **66**

16:45-18:05 : *Environnement - Changement climatique 2*

Homogénisation de séries climatiques,
Olivier Mestre, Victor Venema **68**
Modélisation statistique des changements climatiques, détection et attribution,
Aurélien Ribes **69**
Modèle linéaire mixte avec segmentation : application à la détection de changements dans les dates de vendanges,
Emilie Lebarbier, Franck Picard, Eva Budinska, Stéphane Robin **69**
Sélection bayésienne de variables pour les modèles d'état dans le cadre de reconstructions climatiques,
Ophélie Guin, Philippe Naveau **69**

16:45-18:25 : *Apprentissage 3 : sélection de modèle*

Courbes Principales et Sélection de Modèle,
Aurélie Fischer **64**
Estimation de la fonction de répartition conditionnelle à partir de données censurées par intervalle, cas 1, par sélection de modèles,
Sandra Plancade **64**
Clustering et sélection de variables sur des données génétiques,
Dominique Bontemps, Wilson Toussile **64**
Bornes de Risque pour CART en Classification,
Servane Gey **65**
Bornes de risque pour les forêts purement uniformément aléatoires,

Robin Genuer **65**

16:45-18:25 : *Séries temporelles 2 : mémoire longue*

Test de comparaison du paramètre de longue mémoire (40 min),
Frédéric Lavancier, Anne Philippe, Donatas Surgailis **67**
Time varying fractionally integrated model,
Adnen Ben Nasr, Mohamed Boutahar **67**
Structural change and long memory in the dynamic of U.S. inflation process,
Mustapha Belkhouja, Mohamed Boutahar **67**
Réalisation d'un modèle de prévision à court, moyen et long terme de l'activité d'un transporteur,
Wilfried Despagne **68**

16:45-18:25 : *Statistique fonctionnelle 2*

Utilisation de tests de structure en régression sur variable fonctionnelle (40min),
Laurent Delsol, Frédéric Ferraty, Philippe Vieu **62**
A Functional Regression Approach for Prediction in a District-Heating System,
Aldo Goia **63**
Functional common principal components models,
Graciela Boente, Daniela Rodriguez, Mariela Sued **63**
Sélection de modèle incluant des composantes principales,
Alois Kneip, Pascal Sarda **63**

16:45-18:25 : *Graphes - Modèles graphiques*

Gaussian Faithful Markov Trees,
Dhafer Malouche, Bala Rajaratnam **60**
Modèles de graphe aléatoire à classes chevauchantes pour l'analyse des réseaux,
Pierre Latouche, Etienne Birmelé, Christophe Ambroise **61**
Accuracy of Variational Estimates for Random Graph Mixture Models,
Steven Gazal, Jean-Jacques Daudin, Stéphane Robin **61**
Consistance des estimateurs variationnels pour un modèle de graphe aléatoire,
Alain Celisse, Jean-Jacques Daudin **62**
Convergence de la constante de Cheeger de graphes de voisinage,
Ery Arias-Castro, Bruno Pelletier, Pierre Pudlo **62**

28 mai 2010

09:00-09:45 : *Omiros Papaspiliopoulos*
Builidng bridges: an overview of data augmentation methods for estimation of diffusion processes.

09:00-09:45 : *John Einmhal*
Estimating extreme quantile regions for multivariate regularly varying distributions.

09:45-10:30 : *Axel Munk*
Statistical Multiscale Analysis: From Biomembranes to Biomolecular Microsopy.

09:45-10:30 : *Christian-Yann Robert* Processus ponctuels de dépassement : convergence et estimation.

10:50-12:50 : *Fiabilité - Incertitudes (session du groupe)*

Estimation de modèles markoviens discrets dans un cadre industriel fiabiliste à données manquantes,
Alberto Pasanisi, Shuai Fu, Nicolas Bousquet ***74***
Stratification Directionnelle adaptative,
Miguel Munoz Zuniga, Josselin Garnier, Emmanuel Remy, Etienne de Rocquigny ***75***
Évaluation d'un risque d'inondation fluviale par planification séquentielle d'expériences,
Aurélie Arnaud, Julien Bect, Mathieu Couplet, Alberto Pasanisi, Emmanuel Vazquez ***75***
Caractérisation des coefficients de Strickler d'un fleuve par inversion probabiliste,
Mathieu Couplet, Laurent Lebrusquet, Alberto Pasanisi ***76***
Polynômes de chaos sous Scilab via la librairie NISP,
Michaël Baudin, Jean-Marc Martinez ***76***
Analyse de sensibilité d'un robinet à soupape à l'aide de développements sur chaos polynomial,
Marc Berveiller, Géraud Blatman, Jean-Marc Martinez ***76***

10:50-12:50 : *Statistique non (semi) paramétrique 1*

Estimation dans un modèle défini par des équations estimantes conditionnelles pour des données fonctionnelles,
Matthieu Saumard, Valentin Patilea ***81***
Efficacité semi-paramétrique pour la méthode des moments généralisée,
Paul Rochet, Jean-Michel Loubes ***81***
Estimation récursive en régression inverse par tranche (sliced inverse regression),
Thi Mong Ngoc Nguyen, Jérôme Saracco ***81***
Régression semi-paramétrique de variables explicatives de dénombrement,
Belkacem Abdous, Célestin Kokonendji, Tristan Senga Kiessé ***81***
Régression inverse par tranches pour une population stratifiée,
Marie Chavent, Vanessa Kuentz, Benoit Liquet, Jérôme Saracco ***82***
Estimation non-paramétrique des ensembles de niveaux de la régression,
Thomas Laloe ***82***

10:50-12:30 : *Séries temporelles 3 - Processus*

Blind forecasting for Gaussian time-series,
Thibault Espinasse, Fabrice Gamboa, Jean-Michel Loubes ***73***
Estimation des modèles VARMA structurels avec innovations linéaires non corrélées mais non indépendantes,
Yacouba Boubacar Mainassara, Christian Francq ***73***
Estimation adaptative des modèles vectoriels autorégressifs avec une variance dépendant du temps,

10:50-12:30 : *Statistique mathématique 1*

10:50-12:30 : *Logiciels : Enseignements*

10:50-12:30 : *Statistique bayésienne*

Soleiman Khazaei, Judith Rousseau 72

14:00-14:45 : *Jean-Michel Loubes*

14:00-14:45 : *Alexandre Tsybakov*
Estimation de matrices de faible rang en grande dimension.

14:45-16:25 : *Statistique semi(non) paramétrique 2*

Bootstrap dans l'estimation de la densité par la méthode du noyau,
Smail Adjabi, Mouloud Cherfaoui 91
Réduction itérative du biais pour des lisseurs multivariés,
Pierre-André Cornillon, Nick Hengartner, Eric Matzner-Lober 91
Sur l'estimation du support d'une densité,
Gérard Biau, Benoît Cadre, Bruno Pelletier 92
Aspect brownien d'un test semi-paramétrique d'indépendance,
Bernard Colin, Ernest Monga 92
Tests d'hypothèses dans un modèle de régression non paramétrique,
Zaher Mohdeb 93

14:45-16:45 : *Mélanges -Modèles latents*

Modèles de mélange tronqués pour l'écologie microbienne. Estimation du nombre d'espèces manquantes,
Sebastien Li-Thiao-Te, Jean-Jacques Daudin, Stéphane Robin, Émilie Lebarbier 85
Approche bayésienne variationnelle pour l'agrégation de modèles en classification,
Stevenn Volant, Marie-Laure Martin-Magniette, Stéphane Robin 85
Classification basée sur des mélanges de modèles hiérarchiques bivariés,
Vera Georgescu, Nicolas Desassis, Samuel Soubeyrand, André Kretzschmar, Rachid Senoussi 86
Estimation d'un modèle à blocs latents par l'algorithme SEM,
Christine Keribin, Gérard Govaert, Gilles Celeux 86
Partition latente et dégénérescence dans les mélanges gaussiens,
Christophe Biernacki 87
Modèles linéaires généralisés à facteurs : une estimation par algorithme EM local,
Xavier Bry, Christian Lavergne, Mohamed Saidane 87

14:45-16:05 : *Statistiques et applications / Prix du meilleur stage ENSAI*

Estimation indirecte de l'âge en paléodémographie : approche bayésienne,
Henri Caussinus, Daniel Courgeau, Isabelle Séguy, Luc Buchet 84
Masses des lois multinomiales négatives. Application au traitement d'images polarimétriques,
Phlippe Bernardoff, Florent Chatelain, Jean-Yves Tourneret 84
Air-Conditioning Effect Estimation For Mid-Term Forecasts of Tunisian Electricity Consumption,
Farouk Mhamdi, Mouhamed Ould Mahmoud Mouhamed, Mériem Jaïdane, Jomaa Souissi 84

Estimation de densité via un algorithme EM-Kernel,
Catherine Aaron **85**
Prix du meilleur stage de l'ENSAI,

14:45-16:25 : *Statistique mathématique 2 : rupture - algorithme*

Détection de rupture dans un modèle exponentiel,
Olivier Lopez, Vladimir Spokoiny **89**
Algorithme rapide pour la détection optimale des ruptures,
Guillem Rigaill **90**
Echantillonnage de champs gaussiens de grande dimension,
Olivier Feron, François Orieux, Jean-François Giovannelli **90**
Méthodes non linéaires pour des problèmes statistiques inverses,
Pierre Barbillon, Gilles Celeux, Agnès Grimaud, Yannick Lefebvre, Etienne De Rocquigny **90**
Modèles génératifs de rangs relatifs à un algorithme de tri par insertion,
Christophe Biernacki, Julien Jacques **91**

14:45-15:45 : *Ondelettes*

Classification non supervisée des données fonctionnelles avec ondelettes,
Anestis Antoniadis, Xavier Brossat, Jairo Cugliari, Jean-Michel Poggi **82**
Trajectory prediction by functional regression in Sobolev space,
Kairat tastambekov, Stéphane Puechmorel, Daniel Delahaye, Christophe Rabut **83**
Méthodes multivariées combinant ondelettes et analyse en composantes principales pour le débruitage de données issues de spectrométrie de masse,
Elise Mostacci, Caroline Truntzer, Hervé Cardot, Patrick Ducoroy **83**

14:45-16:25 : *Analyse factorielle et analyse des données*

Analyse en axes principaux de variables symboliques de type histogramme,
Sun Makosso Kallyth, Edwin Diday **87**
Application de l'Analyse des Correspondances Ordinales au suivi d'espèces végétales aquatiques,
Claude Manté, Guillaume Bernard, Patrick Bonhomme **88**
ACP projetée de données séquentielles,
Jean-Marie Monnez **88**
Variabilité des dimensions en ACP : cas complet et incomplet,
Julie Josse, François Husson **89**
L'analyse exploratoire multidimensionnelle d'un modèle structurel fondée sur une classe de critères de covariance généralisée,
Xavier Bry, Patrick Redont, Thomas Verron **89**

16:45-17:00 : *Cloture des journées*

Liste des résumés

Introduction à la statistique spatiale

Edith Gabriel

Demi Journées STID

La géologie, la météorologie, l'épidémiologie, la foresterie, les sciences du sol, l'écologie, ..., sont autant de domaines de recherche où les données ont pour point commun d'être localisées dans l'espace géographique et de n'être ni indépendantes, ni identiquement distribuées. Il s'agit d'observations d'un processus aléatoire $\{Z(s); s \in D\}$, où D est un sous-ensemble du d-espace euclidien, s est une localisation spatiale et $Z(s)$ est une quantité aléatoire. Leur modélisation nécessite de caractériser la dépendance (spatiale) entre différentes observations et leur caractère non homogène : moyenne non constante (variations à grande échelle) et/ou hétéroscédasticité (variations à petite échelle). On s'intéressera au cadre géostatistique où la variable d'étude se déploie continûment sur le domaine D et où $Z(s)$ est un vecteur aléatoire en s. Il s'agira dans un premier temps de justifier le choix d'outils de statistique spatiale pour faire de l'estimation, prédiction à partir de telles données. Nous verrons ensuite comment caractériser l'organisation spatiale des variables étudiées (analyse variographique) et présenterons la méthode de krigeage qui permet de prévoir la valeur prise par une variable en un site non échantillonné à partir d'observations ponctuelles en des sites voisins.

Enseignement à distance : l'experience du Master statistique et économétrie de Toulouse

Anne Ruiz-Gazen, Christine Maurel

Demi Journées STID

Depuis septembre 2003, l'équipe toulousaine d'enseignants du master 2 statistique et économétrie propose une version entièrement à distance et diplômante des cours offerts en présentiel. Depuis la rentrée 2009, l'université Toulouse 1 Capitole a aussi ouvert la première année de master voie statistique ainsi qu'un diplôme d'université en statistique appliquée. L'objectif de la présentation est de partager notre expérience et de témoigner des difficultés de la mise en place de ces cursus à distance au niveau de l'organisation mais aussi de la pédagogie.

Analyse de données avec R - Complémentarité des méthodes d'analyse factorielle et de classification

François Husson, Julie Josse, Jérôme Pagès

Demi Journées STID

L'objectif de cet exposé est de présenter les fonctionnalités disponibles sur R en analyse de données. R est un langage gratuit en pleine expansion et de plus en plus utilisé tant par le monde de l'entreprise, de l'enseignement que de la recherche.

La spécificité de l'analyse des données à la française est présente dans ce logiciel grâce à plusieurs bibliothèques de fonctions. Dans une première partie de l'exposé, on montrera comment mettre en oeuvre, sous R, des méthodes d'analyses factorielles classiques (ACP, AFC ou ACM) et plus avancées (Analyse Factorielle Multiple ou AFM Hiérarchique) à partir du package FactoMineR. Dans une seconde partie, on se focalisera sur la complémentarité des méthodes d'analyses factorielles et de classification pour visualiser des données.

On s'appuiera sur plusieurs exemples et on mettra l'accent sur l'intérêt d'un tel outil dans le cadre de l'enseignement : gratuité et facilité de téléchargement du logiciel pour les étudiants en stage, mais également facilité d'accès du logiciel par des menus interactifs et performance du logiciel pour des méthodes évoluées.

Détection de sélection darwinienne sur un gène par une approche sans vraisemblance

Aude Grelaud, Christian Robert, François Rodolphe

Biostatistique 1 : génome

En génétique des populations, les modèles sont souvent complexes ce qui rend l'évaluation de la vraisemblance difficile, si ce n'est impossible. En revanche, des mécanismes de génération de données sont parfois disponibles ce qui explique pourquoi les méthodes d'inférence bayésienne sans vraisemblance sont si utilisées dans ce domaine. Nous nous intéressons ici à la détection des effets de la sélection darwinienne sur un gène. Les données sont alors constituées d'un ensemble de séquences homologues. Notre objectif est d'estimer la distribution a posteriori de certains paramètres, les autres étant considérés comme des paramètres de nuisance. Ces derniers incluent un arbre phylogénétique qui est ici représenté par un arbre de coalescence, ce qui correspond au modèle de Moran. Sous cette hypothèse, nous pouvons construire un mécanisme de génération de données. Notre procédure d'estimation repose sur l'algorithme ABC-SMC (Del Moral et al., 2009). Nous montrerons ici les résultats obtenus sur le gène *env* du virus HIV.

Références

Del Moral, P. and Doucet, A. and Jasra, A. (2009) An Adaptive Sequential Monte Carlo Method for Approximate Bayesian Computation, Annals of Applied Statistics.

Processus de tests de rapports de vraisemblance pour la détection de QTL

Charles-Elie Rabier, Jean-Marc Azaïs, Céline Delmas

Biostatistique 1 : génome

On considère le processus de tests de rapport de vraisemblance (LRT) en référence au test d'absence de QTL sur un intervalle [0,T] représentant un chromosome (QTL désigne un gène à effet quantitatif). On étudie la distribution asymptotique du processus de LRT sous l'hypothèse nulle d'absence de QTL sur [0,T], et sous l'alternative générale qu'il existe m QTL sur [0,T]. On suggère d'estimer le nombre de QTL, leurs positions, et leurs effets par vraisemblance pénalisée. Les résultats seront généralisés au cas où les individus sont structurés en familles.

Tests multiples pour la comparaison des probabilités de survenue d'une infidélité de transcription dans des ARNm sains et cancéreux

Olivier Collignon, Marie Brulliard, Jean-Marie Monnez, Pierre Vallois, Benoit Thouvenot, Sandrine Jacquenet, Virginie Ogier, Olivier Roitel, Bernard E. Bihain

Biostatistique 1 : génome

L'implication d'erreurs de transcription dans l'hétérogénéité du cancer avait jusqu'alors été peu considérée. En effet, la transcription est supposée fidèle et contrôlée par un système complexe de vérification. Afin d'étudier l'hétérogénéité des séquences d'ARNm issus de tissus sains et cancéreux de 17 gènes d'intérêt, les probabilités de survenue d'une substitution de base ont été comparées à chaque position des séquences des transcrits à l'aide d'une procédure de tests multiples. Pour cela, les séquences Expressed Sequences Tags, qui sont des copies partielles des ARNm d'un gène, ont été utilisées et un modèle prenant en compte l'erreur de séquençage inhérente à ces données a été proposé. Enfin, l'estimateur Location Based Estimator du nombre moyen de tests faux positifs a été étendu au cas de statistiques de tests discrètes. Cette étude préliminaire a ainsi permis de mettre en évidence les positions des ARNm plus fréquemment sujettes à des substitutions dans les tissus cancéreux que dans les tissus sains et d'introduire la notion d'infidélité de transcription chez l'Homme.

Détection d'interaction de gènes à l'échelle du génome

Mathieu Emily

Biostatistique 1 : génome

Cet article propose une étude comparative entre deux méthodes de détection d'interaction de gènes à l'échelle du génome : un premier modèle de régression logistique et un second de fouille de données. Les résultats montrent que les tests utilisés sont peu puissants notamment lorsque les marqueurs d'intérêt ont une fréquence d'allèle faible (< 0.2). Cependant, dans ce cas les méthodes de fouille de données sont préférables. A l'inverse lorsque la fréquence d'allèle augmente et que la taille d'échantillon est supérieure à 20000, la méthode de régression semble plus performante. Une procédure efficace de correction pour les

tests multiples est également proposée. Cette correction utilise la structure de réseaux biologiques et semble prometteuse pour détecter des gènes responsables.

Inférence sur réseaux géniques par Analyse en Facteurs

Yuna Blum, Chloé Friguet, Sandrine Lagarrigue, David Causeur

Biostatistique 1 : génome

La technologie des puces à ADN permet l'analyse simultanée du niveau d'expression de plusieurs milliers de gènes. Un des enjeux de l'analyse de ce type de données est de comprendre la structure de dépendance, qui rend compte des relations biologiques entre les gènes. En particulier, on s'intéresse ici à la modélisation du réseau de régulation des gènes impliqués dans le contrôle d'un caractère phénotypique. Dans un premier temps, on définit un cadre général pour la prise en compte de la dépendance par l'identification de facteurs latents, modélisant la variation commune à l'ensemble des gènes. On montre que l'introduction de ces facteurs dans les procédures d'analyse différentielle en améliore la puissance ainsi que la stabilité des taux d'erreurs. De plus, dans le contexte des modèles graphiques gaussiens pour la modélisation des réseaux d'interactions entre gènes, on présente une méthode d'estimation des corrélations partielles s'appuyant sur la réduction de la dimension des données par les variables latentes. La méthode est illustrée par son application à une étude visant à identifier les gènes impliqués dans le métabolisme des lipides chez le poulet (UMR INRA Génétique Animale de Rennes).

Estimation de l'ordre d'une chaîne de Markov cachée à émissions de la famille exponentielle

Cécile Low-Kam, André Mas

Chaîne de Markov - Processus

Nous cherchons à estimer l'ordre, i.e. le nombre d'états cachés, d'un modèle de Markov caché (HMM), quand aucune borne supérieure sur cet ordre n'est connue. Nous nous intéressons aux HMM dont la distribution des états observables appartient à la famille exponentielle. Deux estimateurs pour l'ordre sont présentés : l'un est basé sur l'estimateur du maximum de vraisemblance, et l'autre sur un mélange bayésien. Tous deux sont pénalisés. Nous prouvons la consistance de ces estimateurs, et des simulations sont effectuées afin de comparer leur performance.

Vitesse de convergence en M-estimation de données markoviennes

Loïc Hervé, James Ledoux, Valentin Patilea

Chaîne de Markov - Processus

Let $\{X_n\}_{n\geq 0}$ be a V-geometrically ergodic Markov chain with $V \geq 1$ some fixed unbounded real-valued function and consider $M_n(\alpha) = n^{-1}\sum_{k=1}^{n} F(\alpha, X_{k-1}, X_k)$, $\alpha \in \mathcal{A} \in \mathbb{R}$

for some real-valued functional $F(\cdot,\cdot,\cdot)$. Define the $M-$estimator $\widehat{\alpha}_n$ such that $M_n(\widehat{\alpha}_n) \leq \min_{\alpha\in\mathcal{A}} M_n(\alpha) + c_n$ with c_n, $n \geq 1$ some sequence of real numbers decreasing to zero. Under some standard regularity assumptions, close to that of the i.i.d case, and under the moment assumption

$$\left(\left|\frac{\partial F}{\partial \alpha}(\alpha,x,y)\right| + \left|\frac{\partial^2 F}{\partial \alpha^2}(\alpha,x,y)\right|\right)^{3+\varepsilon} \leq C\left(V(x)+V(y)\right)$$

for some constants $\varepsilon > 0$ and $C > 0$, the estimator $\widehat{\alpha}_n$ satisfies a Berry-Esseen theorem uniformly with respect to the underlying probability distribution of the Markov chain.

Estimation non paramétrique pénalisée de la dérivée de la densité d'un processus de diffusion

Emeline Schmisser

Chaîne de Markov - Processus

Nous considérons un processus de diffusion $(X_t)_{t\geq 0}$ de fonction de dérive b et de coefficient de diffusion σ. Nous supposons que le processus est strictement stationnaire et β-mélangeant. Ce processus est observé à des instants discrets $t_k = k\Delta$ pour k variant de 0 à $n+1$. Nous calculons deux estimateurs non paramétriques de la dérivée f' de la densité f de ce processus. Pour cela, nous considérons une famille de sous-espaces vectoriels et, sur chacun de ces sous-espaces, nous calculons deux estimateurs de f', $\hat{f'}_m^{(d)}$ et $\hat{f'}_m^{(b)}$. Nous choisissons ensuite les deux meilleurs estimateurs possibles, $\hat{f'}_{\hat{m}}^{(d)}$ et $\hat{f'}_{\hat{m}}^{(b)}$ en introduisant deux fonctions de pénalité $pen^{(d)}(m)$ et $pen^{(b)}(m)$. Le premier estimateur, $\hat{f'}_m^{(d)}$, est obtenu en dérivant un estimateur de f. Cet estimateur minimise une fonction de contraste. L'estimateur adaptatif $\hat{f'}_{\hat{m}}^{(d)}$ est convergent si $n\Delta \to \infty$ (que le pas Δ soit fixé ou qu'il tende vers 0). Pour calculer le second estimateur, nous supposons que le coefficient de diffusion σ est constant égal à 1. Dans ce cas, la fonction f' vérifie l'égalité $f'(x) = 2b(x)f(x)$. L'estimateur $\hat{f'}_m^{(b)}$ minimise une fonction de contraste basée sur cette égalité. Lorsque le pas Δ est petit, le risque de l'estimateur $\hat{f}_{\hat{m}}^{(b)}$ est plus petit que le risque de $\hat{f'}_{\hat{m}}^{d}$. Cependant, l'estimateur $\hat{f}_{\hat{m}}^{(b)}$ ne converge que si $\Delta \to 0$ et $n\Delta \to \infty$.

Tester si un processus de Poisson est modifié en admettant que certains événements ponctuels se regroupent en classes

Franz Streit

Chaîne de Markov - Processus

En observant la réalisation d'un processus ponctuel on est amené à se poser la question si cette réalisation a été engendrée par un processus homogène de Poisson ou par un modèle stochastique plus compliqué. Quand on envisage comme alternative un processus de Poisson à événements regroupés en classes, on trouve peu d'indications utiles et explicites dans la littérature pour aider à faire le choix entre ces hypothèses. Cela s'explique probablement par le fait que la fonction de vraisemblance du processus de Poisson à événements

ponctuels regoupés en classes ('Poisson cluster process') est une expression assez complexe, parce que le nombre des regroupements possibles augmente rapidement avec le nombre des événements ponctuels. Il s'avère cependant que d'habitude, seulement relativement peu de termes de cette fonction de vraisemblance fournissent une contribution non-nulle à la statistique du score efficace. Cette constatation permet dans certains cas d'indiquer d'une manière explicite des tests localement les plus puissants pour distinguer un processus homogène de Poisson des alternatives susmentionnées.

Modèles de comptage appliqués aux décisions de candidature aux offres d'emploi sur le web

Julie Séguéla, Gilbert Saporta

Chaîne de Markov - Processus

Les modèles pour données de comptage sont des techniques largement étudiées dans la littérature. Nous proposons une application de trois types de modèles poissonniens pour modéliser les candidatures aux offres d'emploi sur le web en deux temps : alerte en cas de retour très faible, sinon estimation de l'effectif. Pour cela, en plus de la régression de Poisson standard, nous mettons en oeuvre des modèles construits en deux étapes : hurdle et zero-inflated, adaptés pour la modélisation des zéros en excès et les problèmes de sur-dispersion. Efficaces pour la prédiction des alertes, ces méthodes s'avèrent moins performantes pour la modélisation de l'effectif, à l'inverse de la régression de Poisson standard.

Régression quantile spatiale localement linéaire

Marc Hallin, Zudi Lu, Keming Yu

Statistique spatiale 1

Soit $\{(Y_{\mathbf{i}}, \mathbf{X_i}), \mathbf{i} \in \mathbb{Z}^N\}$ un processus spatial stationnaire de dimension $(d+1)$. On note $\mathbf{x} \mapsto q_p(\mathbf{x})$, $p \in (0,1)$, $\mathbf{x} \in \mathbb{R}^d$, la fonction de régression quantile spatiale d'ordre p, caracterisée par $\mathrm{P}\{Y_{\mathbf{i}} \leq q_p(\mathbf{x}) | \mathbf{X_i} = \mathbf{x}\} = p$. On suppose que le processus a été observé sur un domaine spatial rectangulaire N-dimensionnel, de la forme $\mathcal{I}_{\mathbf{n}} := \{\mathbf{i} = (i_1, \ldots, i_N) \in \mathbb{Z}^N | 1 \leq i_k \leq n_k,\ k = 1, \ldots, N\}$, où $\mathbf{n} = (n_1, \ldots, n_N) \in \mathbb{Z}^N$. Nous proposons un estimateur localement linéaire de q_p, généralisant aux champs aléatoires avec dépendance spatiale complexe et non spécifiée les méthodes de régression quantile considérées dans le contexte des observations indépendantes ou des séries temporelles. Sous des conditions très générales, nous obtenons une représentation de Bahadur pour les estimateurs de q_p et de ses dérivées premières, à partir de laquelle nous établissons leur convergence et leur normalité asymptotique. Le processus spatial est soumis à des hypothèses de mélange qui généralisent les concepts chronologiques habituels. La taille du domaine rectangulaire $\mathcal{I}_{\mathbf{n}}$ peut tendre vers l'infini selon des taux non isotropes. La méthode fournit sur la dépendance entre Y et les régresseurs $\mathbf{X}$ des renseignements beaucoup plus riches que les traditionnelles méthodes de régression linéaire généralement adoptées dans le domaine de la modélisation spatiale.

Ce travail a été réalisé en collaboration avec Zudi Lu (University of Adelaide) et Keming Yu (Brunel University).

Références

Hallin, M., Lu, Z. and Yu, K. (2009). Local linear spatial quantile regression. *Bernoulli* **15**, 659-686.

Robust quantile estimation and prediction for spatial processes

Baba Thiam, Sophie Dabo-Niang

Statistique spatiale 1

Spatial quantile estimation is an interesting and crucial problem in statistical inference for a number of applications where the influence of a vector of covariates on some response variable is to be studied in a context of spatial dependence. In this paper, we present a statistical framework for modeling conditional quantiles of spatial processes assumed to be strongly mixing in space. We are mainly concerned in this work with L_1-consistency as well as asymptotic normality of the kernel conditional quantile estimator in the case of random fields. We propose a spatial nonparametric predictor based on the conditional quantile estimator. We illustrate the proposed methodology with some simulations.

Régression et prédiction non-paramétrique spatiale

Sophie Dabo-Niang, Anne-Françoise Yao

Statistique spatiale 1

Nous nous intéressons à l'estimation de la fonction de régression $r(x) = E(Y|X = x)$ à partir d'observations d'un processus $\{Z_{\mathbf{i}} = (X_{\mathbf{i}}, Y_{\mathbf{i}}), \mathbf{i} \in \mathbb{Z}^N\}$. On suppose que les variables $Z_{\mathbf{i}}$ sont de même distribution que $Z = (X, Y)$, où Y est une variable réelle, intégrable et X un vecteur aléatoire à valeurs dans un espace séparable $\mathcal{E}$ muni (éventuellement de dimension infinie). Dans ce travail, la convergence nos estimateurs est étudiée sous conditions de mélange à partir d'observations dans une région rectangulaire de $\mathbb{Z}^N$. Nous illustrerons nos résultats par des simulations. L'application de nos méthodes à la prédiction spatiale sera également abordée.

Modélisation des dépassements de seuils pour un processus observé à des temps irréguliers

Nicolas Raillard

Statistique spatiale 1

La modélisation des événements extrêmes est de première importance pour garantir la pérennité des constructions humaines, et de manière plus générale pour assurer la continuité des activités humaines, tâche souvent dévolue à la théorie des valeurs extrêmes qui a pour but, entre autres, d'estimer la probabilité d'événements rarement observés, notamment celles de dépassement de seuils élevés. L'objectif de cette étude est d'estimer des telles quantités pour des phénomènes naturels échantillonnés de façon irrégulière.

Les prix à terme de l'électricité sur le marché britannique, une approche par la cointégration

Jérémy Froger, Muriel Renault

Econométrie

En présence de séries non stationnaires, ce qui est souvent le cas pour les séries de prix étudiées en économétrie, l'utilisation de la régression classique par MCO peut conduire à une régression fallacieuse ("'spurious regression"'). Une des solutions possibles pour pallier ce problème est l'utilisation de la cointégration. Il devient alors possible, dans le cadre d'un modèle à correction d'erreur, d'estimer les paramètres d'une relation de long terme entre ces variables. Dans les modèles à correction d'erreur introduits par Engel et Granger (1987) la relation entre les séries se décompose en une relation de long traduisant le lien économique entre les variables en niveaux sur laquelle s'exerce une force de rappel et en une relation de court terme exprimant les oscillations autour de cette relation de long terme. Nous avons appliqué cette méthodologie à l'étude des relations entre les prix à terme de l'électricité sur le marché britannique et les prix à terme de certaines commodités, i.e. le gaz, le charbon et le CO2.

Références

Engle, R.F., Granger, C.W.J., (1987). Co-integration and error correction: representation, estimation, and testing, Econometrica, 55(2), 251-276.

La tendance de l'économie souterraine, au niveau d'un pays pendant la période de transition

Andrei Tudorel, Andreea Iluzia Iacob, Claudiu Herteliu

Econométrie

L'article présente certaines directions de la recherche dans le domaine de l'économie souterraine, basées sur les résultats les plus récents présentés dans la littérature au sujet de sa définition, de sa quantification, de l'identification des causes qui la produisent etc. Cette étude présente une analyse de la causalité entre la taille de l'économie souterraine et le taux de croissance de l'économie formelle en utilisant un test de type Granger à partir de données portant sur 110 pays et qui concernent l'économie souterraine et le taux de croissance annuel de PIB. En utilisant les séries de données de 22 pays en transition nous estimons également les paramètres de certains modèles pour analyser la relation qui existe entre la taille de l'économie souterraine et le taux de croissance de l'économie officielle. Les paramètres de tous les modèles sont estimés en utilisant la méthode des moindres carrés et la méthode des moindres carrés généralisée. Le source de données pour l'analyse de la causalité et pour l'estimation des modèles est Schneider (2005).

Références

Schneider, F. (2005) Shadow economies around the world: what do we really know?, European Journal of Political Economy, 21(3), 598-642.

La méthode d'évaluation contingente appliquée au ruissellement érosif : une nouvelle approche de l'estimation du consentement à payer

Dimitri Laroutis, Patrice Lepelletier

Econométrie

La base de données sur laquelle nous travaillerons est le résultat d'une enquête menée en Haute Normandie, plus précisément dans la vallée du commerce, portant sur la méthode d'évaluation contingente et notamment sur le consentement à payer des individus pour une réduction du risque de ruissellement érosif. La finalité de ce travail est d'obtenir une bonne estimation du prix consenti à payer par les habitants de la vallée du commerce afin de réduire les impacts de ce phénomène naturel. La base de données est constituée de 47 facteurs explicatifs et du consentement à payer pour 221 personnes interrogées. Nous sélectionnons les facteurs les plus significatifs suite à une analyse de la variance. Nous mettons en place un estimateur du consentement à payer le plus efficace possible. Notre approche consiste à estimer chaque combinaison possible issue des différentes modalités des facteurs les plus significatifs. Dans cette optique nous n'établissons pas de modèle prédictif donnant le consentement à payer en fonction des facteurs significatifs, mais nous construisons un arbre d'estimation sur lequel nous sommes capables au bout de chaque branche de donner une prédiction du consentement à payer.

A test of endogeneity in quantiles

Kim Tae-Hwan, Christophe Muller

Econométrie

In this paper we develop a test to detect the presence of endogeneity in different quantiles in the conditional distribution of a variable of interest. This Hausman test type is based on one estimator consistent only under no endogeneity at the examined quantile and another estimator consistent in both the null and the alternative hypotheses. We derive the asymptotic distribution of the test statistic. Moreover, we study the finite sample properties of this test with Monte Carlo simulations of which results exhibit substantial power in the studied cases. Finally, we apply our test to Engel curve estimation with UK data. We find that the pattern of the endogenenity of the total expenditure for various commodities (food, alcohol, fuel, transport, services) is complex when examining it across quantiles.

Méthodes de détection des unités atypiques : cas des enquêtes structurelles ukrainiennes

Michel Grun-Rehomme, Olga Vasechko

Econométrie

Le problème de la détection d'unités atypiques ou de valeurs extrêmes est général et très ancien. Il se pose à tous ceux qui ont à analyser des données réelles. De nombreuses méthodes algébriques, graphiques ou probabilistes existent pour détecter les valeurs extrêmes ("'outliers"') sur une variable. Dans cette présentation, deux nouvelles méthodes

non paramétriques simples de détection des unités atypiques dans le cadre univarié sont proposées. La première dans le domaine de la V-robustesse permet de tenir compte de l'éloignement de l'observation par rapport au centre de la distribution et de la forme de la fin de la distribution. Et la seconde, basée sur une statistique du coefficient de variation, permet d'accentuer le caractère atypique des unités et de mesurer l'effet d'une unité atypique sur la qualité de l'information. De plus, une troisième méthode, obtenue comme combinaison convexe de deux techniques de la théorie des valeurs extrêmes, en minimisant la variance, est également proposée, permettant ainsi de décider entre différentes stratégies. Ces différentes méthodes sont comparées sur des données des enquêtes structurelles ukrainiennes. Il n'existe pas de méthode universelle. Les différentes méthodes paramétriques ou non paramétriques ne permettent pas toujours de trancher sur le nombre d'unités atypiques à retenir, mais elles peuvent déterminer très rapidement une stratégie haute (très peu d'entreprises atypiques) et une stratégie basse de détection des unités atypiques. Il est nécessaire d'utiliser une technique facile à mettre en production.

Inférence statistique dans des modèles de moments conditionnels en présence de censure

Pascal Lavergne, Olivier Lopez, Valentin Patilea

Données censurées - Survie

Une large littérature statistique, biostatistique, économétrique, ... étudie l'inférence statistique (estimation et tests) dans des modèles semi-paramétriques définis par des moments conditionnels lorsque les données sont censurées. La régression paramétrique classique ou quantile en présence d'une censure à droite sur la variable à expliquer sont des exemples de référence. Nous proposons une nouvelle classe d'estimateurs pour des modèles définis par l'espérance conditionnelle d'une fonction connue dépendante de vecteurs observés Y et X et d'un paramètre inconnu θ, sachant le vecteur aléatoire X, lorsque les observations Y sont soumises à un mécanisme de censure. L'estimateur proposé minimise une distance pondérée à l'aide d'un noyau et des poids qui prennent en compte la censure. Les résultats théoriques sont obtenus uniformément par rapport à la fenêtre du noyau et sous de conditions générales sur le mécanisme de censure. Plusieurs exemples de mécanismes de censure sont proposés.

Prédiction de la fonction de survie par sélection de modèle

Ion Grama, Jean-François Petiot

Données censurées - Survie

Nous proposons une estimation semi-paramétrique d'une fonction de survie $S(t) = P(T \geq t)$ où T est la durée de vie d'un individu, éventuellement censurée à droite. Classiquement, nous observons des variables $Zi = \min(Ti, Ci)$ où les Ci sont les temps de censure indépendants des durées de vie Ti. Notre but est d'obtenir une prédiction des probabilités de survie $S(t)$ au-delà des durées observées, c'est-à -dire pour $t > \max Zi$. Le modèle présenté est choisi dans une famille de modèles qui ont de bonnes propriétés pour cette prédiction tout en étant assez flexibles pour bien ajuster les données observées. Ce modèle peut aussi permettre d'améliorer la qualité de l'estimateur de S(t) pour les

valeurs de t inférieures mais relativement "'proches"' de ce maximum. L'idée principale du modèle est de choisir de façon automatique un seuil u à partir duquel les prévisions pour les durées de vie sont encore fiables. En dessous de ce seuil $S(t)$ est estimée par une méthode complètement non-paramétrique, comme celle de Kaplan-Meier. Au dessus de ce seuil un modèle paramétrique est choisi, nous utiliserons ici la loi exponentielle. Le choix du seuil u sera assuré par une suite de tests d'ajustement. La méthode est appliquée à des données de "ré-hospitalisation", la durée de vie étant ici le délai écoulé entre une sortie d'un hôpital et une ré-admission pour la même cause médicale.

Extensions du test du logrank aux données censurées par intervalle

Fanny Leroy, Ludovic Trinquart

Données censurées - Survie

Lorsque les données sont censurées par intervalle (type II), la fonction de survie est couramment estimée par la méthode non-paramétrique de Peto-Turnbull ou par celle de Groeneboom & Wellner. Ces méthodes sont basées sur une partition de l'axe des temps puis sur la maximisation de la vraisemblance via un algorithme itératif. La méthode de Peto-Turnbull suppose qu'en dehors de certains intervalles fermés de la partition l'estimation de la fonction de survie est constante. La méthode de Groeneboom & Wellner relaxe cette hypothèse mais requiert l'estimation d'un plus grand nombre de paramètres. Après un rappel sur ces méthodes, nous nous intéressons à la comparaison des fonctions de survie entre deux groupes. Nous rappelons l'extension du test du logrank aux données censurées par intervalles basée sur la méthode de Groeneboom & Wellner, déjà décrite dans la littérature. Puis nous proposons l'extension du test du logrank basée sur la méthode de Peto-Turnbull. Dans les deux cas, la statistique de test repose sur les estimations du nombre d'événements et du nombre de sujets à risque. Dans le cas de la généralisation via la méthode de Groeneboom & Wellner, on estime ces quantités à chaque pseudo-temps d'événement. Dans le cas de la généralisation via la méthode de Peto-Turnbull, on les estime pour chaque intervalle fermé de la partition de l'axe des temps sur lequel la fonction de survie n'est pas constante. Ces méthodes sont appliquées à des données issues d'un essai contrôlé randomisé en neurologie vasculaire.

Analyse bayésienne de données de survie discrètes sujettes à censure informative avec un modèle à effets aléatoires partagés reparamétré.

Sophie Ancelet-Enjalric, Elise de la Rochebrochard, Jean Bouyer

Données censurées - Survie

Les mécanismes de censure dits informatifs viennent souvent complexifier l'analyse de données de survie. L'inférence de modèles standards ne tenant pas compte de ce type de données manquantes peut mener à des conclusions biaisées. De nombreux travaux se sont axés autour des modèles à effets aléatoires partagés pour l'analyse de données longitudinales avec censure informative. Nous présentons une extension à cette classe de modèles hiérarchiques pour l'analyse de données de survie discrètes sujettes à censure informative. Notre modèle est basé sur la combinaison de deux modèles à hasards

proportionnels en temps discret et tient compte de l'existence possible d'une variabilité résiduelle non-partagée. Puis, nous proposons d'utiliser une reparamétrisation du modèle proposé, par ajout de paramètres redondants, ainsi qu'un choix spécifique de lois a priori afin d'améliorer les propriétés de convergence des algorithmes MCMC implémentés pour mener l'inférence bayésienne du modèle. Enfin, nous proposons de valider notre modèle à partir de calculs de facteurs de Bayes partiels en vue de tester l'hypothèse d'existence d'une censure informative intervenant sur un mécanisme de survie d'intérêt et le calcul de p-valeurs bayésiennes "mixtes" en vue de quantifier l'adéquation entre notre modèle et des données observées. Nous illustrons la pertinence de notre approche bayésienne par des simulations ainsi que sur un jeu de données réelles issues de l'enquête "Devenir Après Initiation de la Fécondation-In-Vitro".

Modèle de fragilité et algorithme EM stochastique

Charles El-Nouty, Estelle Kuhn

Données censurées - Survie

Le modèle de Cox (1972) joue un rôle essentiel en analyse de survie. Toutefois, il est souvent peu réaliste du fait de l'hétérogénéité intrinsèque à la population observée. Cette hétérogénéité est prise en compte via un effet aléatoire dans le modèle de fragilité introduit dans l'article de Vaupel et al. (1979).

Estimer les paramètres d'un modèle de fragilité repose principalement sur la vraisemblance observée et sur l'estimation du maximum de vraisemblance associé. Mais le calcul direct est souvent difficile, voir impossible. Puisque les variables aléatoires de fragilité ne sont pas observées, on se situe dans le cadre plus général des modèles à variables latentes. Pour approcher l'estimateur du maximum de vraisemblance observée dans de tels modèles, on a généralement recours à l'algorithme Expectation Maximization (EM) proposé dans l'article de Dempster et al. (1977). Cependant, il n'est pas applicable dans beaucoup de cas pratiques et doit souvent faire l'objet d'approximations.

Notre objectif est de proposer un algorithme convergent pour l'estimation par maximum de vraisemblance dans les modèles de fragilité. A cet effet, nous considérons une approximation stochastique de l'algorithme EM couplée à une méthode de Monte Carlo par chaînes de Markov (SAEM-MCMC) introduite dans l'article de Kuhn et Lavielle (2004). La suite générée par l'algorithme converge presque sûrement vers un maximum local de la vraisemblance observée sous des hypothèses peu contraignantes. Bien que faisant appel à de la simulation, du fait de l'imbrication astucieuse des deux méthodes SAEM et MCMC, cet algorithme est rapide. Nous comparons nos résultats à ceux obtenus par d'autres algorithmes dans la littérature.

Références

Cox, D.R. (1972) Regression models and life-tables, J.R.S.S. Serie B, 34, 187-220.

Dempster, A.P., Laird, N.M. et Rubin, D.B. (1977) Maximum likelihood from incomplete date via the EM algorithm, J.R.S.S. Serie B, 39(1), 1-38.

Kuhn, E. et Lavielle, M. (2004) Coupling a stochastic approximation version of EM algorithm with an MCMC procedure, ESAIM P&S, 8,115-131.

Vaupel, J.W., Manton, K.G. et Stallard, E. (1979) The impact of heterogeneity in individual frailty on the dynamics of mortality, Demography, 16, 439-454.

Analyse canonique généralisée régularisée et approche PLS

Arthur Tenenhaus, Michel Tenenhaus

PLS

Nous donnons dans cette communication une définition de l'analyse canonique généralisée au niveau de la population (ACG-population) qui constitue le cadre théorique de l'approche PLS proposée par Herman Wold et à ses extensions proposées par Jan-Bernd Lohmöller et Nicole Krämer. En écrivant les équations stationnaires de l'ACG-population au niveau de l'échantillon et en utilisant des estimations régularisées (shrinkage estimations) des matrices de covariance des blocs, nous obtenons de nouvelles équations stationnaires au niveau de l'échantillon. Ces équations stationnaires sont également celles d'un problème d'optimisation que nous appelons analyse canonique généralisée régularisée (ACGR). En recherchant un point fixe de ces équations stationnaires au niveau de l'échantillon nous obtenons un algorithme très similaire à l'approche PLS de Wold-Lohmöller-Krämer. De plus, nous démontrons la convergence monotone de l'algorithme proposé.

Kernel Generalized Canonical Correlation Analysis

Arthur Tenenhaus

PLS

Un problème classique en statistique est d'étudier les liens entre plusieurs blocs de variables. L'objectif est de trouver les variables d'un bloc influençant les variables d'autres blocs. L'Analyse Canonique Généralisée Régularisée (ACGR) est un cadre très attractif pour traiter ce type de problématique. Cependant, l'ACGR ne capture que des relations linéaires entre blocs et pour accéder à des liens non linéaires nous proposons une extension à noyau de l'ACGR.

The Non-Metric Partial Least Squares approach

Giorgio Russolillo

PLS

Dans ce travail, nous montrons comment les algorithmes PLS, correctement ajustés, peuvent travailler comme des algorithmes de Codage Optimal. Cette nouvelle fonctionnalité du PLS, qui avait été jusqu'à maintenant totalement inexplorée, nous a permis de mettre au point une nouvelle série de méthodes PLS: les méthodes Non-Metric PLS (NM-PLS).

An integrated PLS Regression-based approach for multidimensional blocks in PLS path modeling

Vincenzo Esposito Vinzi, Giorgio Russolillo, Laura Trinchera
PLS

L'approche PLS aux modèles à équations structurelles (PLS Path Modeling, PLS-PM) est couramment considérée comme une approche basée sur les composantes. Cette méthode a été récemment revisitée en tant que cadre général pour l'analyse des tableaux multiples. Nous proposons ici deux nouvelles méthodes d'estimation des poids externes dans le cadre de la PLS-PM: le Mode PLScore et le Mode PLScow. Chaque mode est fondé sur l'utilisation de la régression PLS pour l'étape d'estimation externe. Toutefois, en Mode PLScore une régression PLS est exécutée sous les contraintes classiques de la PLS-PM de variance unitaire pour les scores des variables latentes ; tandis que dans le Mode PLScow les poids externes sont contraints d'avoir une norme unitaire. Cette dernière contrainte est la contrainte classique de normalisation dans le cadre de la régression PLS. Nous montrons comment les deux nouveaux modes sont liés aux méthodes d'estimation externe classiques de la PLS-PM, c.-à -d. au Mode A et au Mode B, ainsi qu'au Nouveau Mode A récemment proposé par Tenenhaus & Tenenhaus (2009).

Références

Tenenhaus M. and Tenenhaus M. (2009), A criterion based PLS approach to structural equation modelling, presented at 6th International Conference on Partial Least Squares Methods.

La régression multivariée contrainte PLS

Philippe Casin, Francois Marque
PLS

La régression multivariée contrainte (RMC) présente de nombreuses applications en économétrie, notamment dans le cas de données de panels ou lorsqu'il s'agit de mettre en évidence des relations de cointégration entre les variables. Le modèle s'écrit : Y=XB+ZC+U, où Y, B et Z désignent des ensembles variables quantitatives et U un vecteur gaussien, et il est contraint car B=ab , le rang de a étant inférieur au nombre de variables X. Les variables Xa sont alors obtenues par l'analyse canonique entre l'espace engendré par les résidus des régressions de X par rapport à Z et l'espace engendré par les résidus de Y par rapport à Z ; or, les résultats de cette analyse canonique peuvent être très sensibles à de faibles variations des données de départ. En effet, d'une part, les résidus d'une régression, lorsque celle-ci a un fort pouvoir explicatif, ont une faible variance et sont donc peu robustes ; d'autre part, les composantes canoniques sont elles même très sensibles à des variations des variables engendrant les espaces. Cette communication met en évidence les propriétés que doivent remplir les variables Xa pour être robustes ; un compromis entre ces propriétés débouche sur une nouvelle méthode de calcul des variables Xa proche des techniques PLS.

Analyse asymptotique des processus autoregressifs de bifurcation avec données manquantes

Benoîte De Saporta, Anne Gégout-Petit, Laurence Marsalle
Séries temporelles 1

Nous étudions le comportement asymptotique de l'estimateur des paramètres d'un processus autorégressif de bifurcation dans le contexte de données manquantes. Les données manquantes sont modélisées par un processus de Galton-Watson multi-type à deux catégories. Sous des hypothèses faibles sur le bruit de l'autorégression, (indépendance conditionnelle paire par paire et conditions de moments), nous établissons la convergence presque sûre de notre estimateur. Notre travail repose sur des résultats asymptotiques non-standard pour les martingales.

Modélisation et prévision des prix du CPC

Delphine Blanke, Denis Bosq

Séries temporelles 1

Dans cet exposé, nous étudions la modélisation des prix du CPC (coke petroleum calcination) et utilisons les modèles obtenus pour la prévision. La difficulté du problème vient de la non-homogénéité des observations. Nous considérons d'abord des techniques empiriques (régression polynomiale) et nous remarquons qu'elles sont des cas particuliers de la méthode SARIMA. Nous utilisons alors la méthode SARIMA et nous la comparons à une méthode non paramétrique.

Méthodes récursives en estimation et prévision non paramétriques

Aboubacar Amiri

Séries temporelles 1

Nous introduisons d'abord une famille paramétrique d'estimateurs récursifs de la densité indexée par un paramètre l appartenant à [0, 1]. Leur comportement asymptotique en fonction du paramètre l va nous amener à introduire des critères de comparaison basés sur les biais, variance et erreur quadratique asymptotiques. Pour ces critères, nous comparons nos estimateurs entre eux et aussi comparons notre famille d'estimateurs à l'estimateur non récursif de la densité de Parzen-Rozenblatt. Ensuite, nous définissons à partir denotre famille d'estimateurs de la densité une famille d'estimateurs récursifs à noyau de la fonction de régression. Nous étudions également ses propriétés asymptotiques en fonction du paramètre l. Ces résultats permettent ainsi de construire une famille de prédicateurs non paramétriques qui permettent de réduire le temps calcul.

Ségmentation bayésienne hiérarchique de processus MA constant par morceaux

S. Suparman

Séries temporelles 1

On utilise dans ce travail une méthode bayésienne pour traiter une problème de segmentation de processus MA par morceaux. La complexité des lois a posteriori ainsi que la

structure particulière de l'espace des paramètres amène a utiliser la méthode de simulation de type Monte Carlo par Chaînes de Markov à sauts réversibles. Les sorties de l'agorithme sont utilisées pour obtenir plusieurs types d'estimateur des paramètres d'intéret : Maximum Marginal a Posteriori, et Moyenne Marginale a Posteriori. La délicat problème du réglage des hyperparamètres est contourné en munissant des hyperparamètres de loi, utilisant ainsi une structure bayèsienne hiérarchique.

Une méthode de segmentation pour le traitement de séries temporelles

Christian Derquenne

Séries temporelles 1

La méthode proposée permet de segmenter une série temporelle. Elle offre une démarche originale contenant une phase essentielle de préparation des données afin de produire la structure la plus adéquate possible pour initialiser la phase de modélisation selon un modèle linéaire hétéroscédastique incluant des tendances, des constantes et des dispersions différentes. Cette méthode peut être utilisée dans de nombreux domaines d'applications, mais aussi pour de nombreux objectifs : stationnarisation, recherche de segments, construction de différents modèles sur une même série ayant des comportements différents, simplification de séries dans le but de réaliser de la classification de courbes, etc.

Recensement en France : le point de vue de l'INSEE

François Clanché, Session "Recensements" organisée par le groupe SES

Recensements

François Clanché, chef du département de la démographie à l'INSEE, et à ce titre responsable du recensement français, fera le point de la situation dans notre pays, du point de vue de l'Institut statistique national. La nouvelle méthode de recensement en France est unique au monde. Elle repose sur un cycle glissant de 5 années : la première collecte ayant eu lieu en 2004, les résultats du premier cycle 2004-2008 sont désormais complètement disponibles, les résultats 2005-2009 sont en cours d'établissement. Après avoir rapidement rappelé les principales caractéristiques de cette méthode de recensement, François Clanché expliquera les avantages que l'INSEE a constatés en appliquant cette méthode année après année, tant du point de vue de l'organisation des travaux dans l'Institut que du point de vue de l'extension et de la fiabilité des résultats. Face à ces avantages, il subsiste des difficultés et des défis : il expliquera lesquels, et esquissera les innovations projetées pour y faire face dans les prochaines années.

Recensements de population : nouvelles méthodes, nouveaux défis

Jean-François Royer

Recensements

Les recensements de population sont les plus anciennes des opérations statistiques. Leurs méthodes ont évolué avec les changements des sociétés. Depuis la fin du XX° siècle, les méthodes de recensement utilisées depuis des décennies ont été confrontées à de graves difficultés, dans un grand nombre de pays : refus de répondre, impossibilité de réunir les budgets nécessaires, délais croissants, incertitudes sur la qualité des résultats... Face à cela, les différents pays ont réagi différemment, en fonction de leurs particularités administratives. Ainsi les pays du Nord de l'Europe ont-ils abandonné le recensement traditionnel au profit de "recensements virtuels" obtenus par l'utilisation ou la fusion de répertoires administratifs. Les pays disposant de registres de population, dont l'Italie est un exemple, ont cherché à profiter de l'existence de ces registres pour alléger et améliorer la collecte exhaustive. La France a choisi à la fin du siècle une voie originale associant sondage et recensement traditionnel. En même temps qu'il se diversifie du point de vue des méthodes, l'objet "recensement" se précise et s'homogénéise sur le plan international. Dans l'Union Européenne, les recensements sont désormais prescrits à tous les pays dans le cadre d'un règlement qui fixe les dates ainsi que la nature des résultats statistiques attendus, mais qui laisse la possibilité de différences importantes dans les méthodes. L'ONU pour sa part émet des recommandations sur les mêmes sujets. Loin d'être un champ figé, la méthodologie des recensements est donc un terrain vivant, un de ceux où se joue l'usage social des statistiques. Où en est-on en 2010, justement à la veille d'un nouveau cycle de recensements dans le monde ? Pour éclairer cette question, le groupe "Statistique et société" de la SFdS organise une session d'information et de débat aux Journées de statistique de Marseille, comme il l'avait fait en 2003-2004-2005, lors du lancement du nouveau recensement français. Trois interventions sont prévues. Ces trois exposés, et le débat qui suivra, devraient permettre aux participants de saisir, au-delà des particularités nationales, quelles sont les principales contraintes affectant ces grandes opérations statistiques publiques dans le monde contemporain, et dans quelle mesure les progrès techniques et l'ingéniosité des statisticiens permettent d'y faire face.

Recensement en France : le point de vue des communes

Marie-Hélène Boulidard, Session "Recensements" organisée par le groupe SES

Recensements

Marie-Hélène Boulidard, qui est expert démographe-ingénieur territorial, membre de la Commission nationale d'évaluation du recensement, et rapporteur du groupe de travail du Conseil national de l'information statistique sur la "Diffusion des résultats du recensement", abordera les mêmes questions, cette fois du point de vue des collectivités territoriales. Les 36 000 communes françaises sont dans notre nouveau système coproductrices du recensement ; elles sont également des utilisatrices très importantes de ses résultats, qu'il s'agisse des chiffres de population légale (importants pour la vie administrative locale, et notamment pour le montant des subventions reçues de l'Etat), des statistiques décrivant la population de la commune, ou encore des outils nouveaux créés du fait de la nouvelle méthode de recensement dans le communes de 10 000 habitants ou plus, comme le "répertoire des immeubles localisés". Marie-Hélène Boulidard expliquera ce que la nouvelle méthode de recensement a changé pour les collectivités de différentes tailles, et décrira les difficultés qui restent à surmonter pour que cette opération favorise le plus possible une bonne administration locale du pays.

Recensement en Italie : le point de vue de l'ISTAT

Fabio Crescenzi, Session "Recensements" organisée par le groupe SES

Recensements

Fabio Crescenzi, qui travaille à la Direction générale du recensement de l'ISTAT, Institut national de statistiques d'Italie, apportera un témoignage permettant de mettre en perspective l'expérience française par rapport à la démarche d'un grand pays voisin. L'Italie a réalisé un recensement en 2001, et s'apprête à renouveler cette opération en 2011. Comme en France, les statisticiens y sont soumis à une forte pression pour abaisser les coûts et la pénibilité de l'opération pour la population, tout en fournissant des données exhaustives et rapides. Comment un registre de population peut-il être mis à profit dans ce contexte ? Des méthodes de sondage peuvent-elles être envisagées ? Quel est le rôle des municipalités, et notamment des plus grandes ? Y a-t-il des possibilités de recourir au téléphone, ou à Internet, pour rendre l'opération plus facile pour la population ? Fabio Crescenzi fera le point des solutions que l'ISTAT a envisagées, et de celles qu'il a finalement retenues.

Approche variationnelle pour la cartographie spatio-temporelle du risque en épidémiologie à l'aide de champs de Markov cachés

David Abrial, Lamiae azizi, Myriam Charras-Garrido, Florence Forbes

Biostatistique 2 : épidémiologie

L'analyse spatio-temporelle d'une épidémie permet aux épidémiologistes de comprendre son étiologie et fournit des suggestions pour planifier de nouvelles études pour examiner les causes sous-jacentes. Cette analyse donne lieu à une estimation du risque épidémiologique dans différentes unités géographiques et produit ainsi des cartes de risque permettant la détection de différences de niveau de risque à différents pas de temps. Nous proposons d'adapter une méthode par champs de Markov cachés discrets(issue de l'analyse d'images) dans le cadre spatial, pour permettre une classification intrinsèque des risques en vue du tracé des cartes de niveaux de risque et de l'étendre par la suite à un contexte spatio-temporel. Afin d'estimer les paramètres du modèle et de définir les classes, l'algorithme EM-champ moyen est utilisé.

Sur les modèles de comptage à inflation de zéro avec application à la modélisation de la tendance dans la prévalence de certaines maladies allergiques liées au travail en France de 2001 à 2007

Joseph Ngatchou-Wandji, Christophe Paris

Biostatistique 2 : épidémiologie

Les modèles de comptage à inflation de zéro sont souvent utilisés pour modéliser des données de comptage surdispersées et/ou comportant une grande proportion de zéro. Ces modèles ont été utilisés avec succès en econometrie, en épidémiologie, en santé publique,

en biologie et dans bien d'autres domaines. Dans cette note, nous faisons le survol de la littérature sur ces modèles, puis les appliquons à la modélisation de la tendance dans la prévalence de certaines maladies allergiques liées au travail telles que l'asthme, la rhinite et la dermite. Les données sont françaises, collectées de janvier 2001 à décembre 2007 par le Réseau National de Vigilence et de Prévention des Pathologies Professionnelles (RNV3P).

Surveillance de la contamination chimique en Méditerranée basée sur les capacités accumulatrices de la moule : détermination d'une réponse universelle de capteur

Marc Bouchoucha, Michel Lhermine, Jean-Claude Franc, François Galgani, Bruno Andral, Pierre Boissery

Biostatistique 2 : épidémiologie

La mesure des contaminants chimiques directement dans la colonne d'eau est coûteuse, difficilement interprétable et peu applicable à de nombreux échantillons prélevés le long d'un important linéaire côtier. Pour surmonter ces difficultés, les contaminants peuvent être mesurés dans la chair de bioaccumulateurs naturels comme les moules. Depuis 1994, le réseau RINBIO cherche à évaluer les niveaux de contamination chimique à l'échelle de la façade méditerranéenne en utilisant une technique de caging de moules. Or, les caractéristiques du capteur représenté par la moule varient en fonction du milieu et notamment de ses capacités trophiques. L'objectif de l'étude est de déterminer la réponse de capteur qui peut s'écrire : $C^{(k)}(t)/X^0(t) = F[\vec{\alpha}, \vec{\Phi}^{(k)}(t)] + \epsilon^{(k)}(t)$ avec avec k le site, t la campagne, $C^{(k)}(t)$ concentration dans le coquillage, $X^0(t)$ mesure du bruit de fond dans l'eau de la campagne, F réponse de capteur de forme paramétrique simple, $\vec{\alpha}$ vecteur des paramètres et $\vec{\Phi}^{(k)}(t)$ vecteur des paramètres de capteur et $\epsilon^{(k)}(t)$ facteur de bruit. Les identifications sont réalisées par une procédure d'optimisation dans laquelle est imbriquée une étape de régression robuste. La détermination de cette réponse permet de calculer les valeurs en contaminants dans la colonne d'eau quelle que soit la croissance des coquillages et donc de les comparer.

Essais de modélisation de l'épilepsie en Tunisie : Théorie et application basées sur des modèles de régression logistique

Abdelwaheb Daouthi, Mohamed Dogui, Abdeljelil Farhat

Biostatistique 2 : épidémiologie

Le but de ce travail est de faire une modélisation de l'épilepsie en Tunisie. Nous avons essayé à travers cette modélisation de montrer la pertinence des modèles de régression logistique comme outils d'analyse. Pour atteindre cet objectif, nous avons commencé par la présentation des modèles de régression logistique. Ensuite, nous avons étudié la méthode d'estimation du maximum de vraisemblance. Enfin et après avoir construit une base de données, nous avons estimé plusieurs modèles de régression logistique appliqués à un groupe de variables quantitatives et qualitatives relatives à 5000 patients épileptiques. Les principaux de ces modèles sont le logit et le probit multinomiaux. L'étude de ces modèles

nous a permis de montrer dans quelle mesure les paramètres sélectionnés sont significatifs.

Étude statistique du coefficient de covariation symétrique signé

Bernédy Kodia, Bernard Garel

Statistique spatiale 2 - Champs de Markov

Le coefficient de covariation symétrique signé est une mesure permettant de capter la dépendance entre variables aléatoires α-stables symétriques. Dans le cas des vecteurs aléatoires sous-gaussiens, la matrice des coefficients de covariation symétriques signés coïncide avec la matrice de corrélation du vecteur gaussien sous-jacent. Deux estimateurs de ce coefficient sont proposés : le premier basé sur les moments fractionnaires d'ordre inférieur et le second sur les "screened ratio". Nous procédons à l'analyse de la robustesse de ces estimateurs, en utilisant des données normales contaminées.

Mise en oeuvre du krigeage sur arbre

Edwige Polus, Chantal De Fouquet

Statistique spatiale 2 - Champs de Markov

Un modèle de fonctions aléatoires définies sur une topologie d'arbre (De Fouquet & Bernard-Michel, 2006) a été développé pour l'estimation de concentrations le long d'un réseau hydrographique. Le principe consiste à décomposer le réseau en filets élémentaires joignant chaque "source" à "l'exutoire". La concentration Z au point x s'exprime comme une combinaison linéaire $Z(x) = \sum w_i Y_i(x)$ de variables aléatoires élémentaires Y_i définies sur ces filets, dont les coefficients w_i dépendent de la position sur l'arbre. Le krigeage des concentrations au point x revient donc à l'estimation d'une combinaison linéaire des $Y_i(x)$ à partir d'autres combinaisons linéaires, les concentrations mesurées aux points expérimentaux x_α : $Z(x_\alpha) = \sum w_i(x_\alpha) Y_i(x_\alpha)$. Le krigeage dépend des hypothèses sur les concentrations, qui se ramènent à des hypothèses sur les variables élémentaires Y_i et sur les coefficients w_i. Ces hypothèses déterminent les conditions requises pour l'estimation : le nombre minimum de mesures et leur répartition par arête. Ce modèle est appliqué aux concentrations en nitrates sur un réseau constitué d'une petite portion de la Seine (de l'amont de Paris à l'estuaire) et de la Marne. Tout d'abord, la simulation déterministe ProSe, disponible en tout point (Even et al., 1998), permet de tester les hypothèses. Ensuite, le krigeage est effectué à partir des mesures aux stations effectivement disponibles.

Références

De Fouquet, C and Bernard-Michel, C. (2006). Modèles géostatistiques de concentrations ou de débits le long des cours d'eau. Comptes rendus Géoscience, 338, 5, 307-318.

Even, S., Poulin, M., Garnier, J. [et al.] (1998) River ecosystem modelling. Application of the PROSE model to the Seine river (France). Hydrobiologia; 373/374:27-45.

Les dimensions Fractals : Mythes et réalités écologiques

Nicolas Bez

Statistique spatiale 2 - Champs de Markov

La notion de fractal proposée par Mandelbrot a séduit la communauté écologique pour les notions d'indépendance d'échelle et de transfert d'échelle qu'elle recèle. L'objectif de ce travail est de préciser qu'en réalité, la notion de fractal cache deux concepts différents de rugosité d'un part et d'auto-homothétie d'autre part. Les cas où ces notion se rejoignent sont très particuliers (processus Gaussiens) et rarement rencontrés dans la nature. En s'appuyant sur la définition rigoureuse de la dimension fractal qui repose sur la dimension de Haussdorf-Besicovitch, on montre que la dimension fractal quantifie une propriété locale des processus à savoir la rugosité. De plus, lorsque son estimation est basée sur le variogramme, on montre que la formule retenue par la bibliographie écologique qui utilise la demie-pente du log-variogramme est erronée. L'utilisation de la pente du log-madogramme est recommandée. La différence est illustrée à l'aide de données réelles. La dimension d'auto-homothétie doit être distinguée de la dimension fractal. On montre des cas où la dimension fractal est non entière alors que le processus n'est pas auto-homothétique. Il est donc important, dans la pratique, de ne pas confondre ces deux notions souvent cachées derrière le même vocable de dimension fractal. Enfin, parce que précisément la dimension fractal peut être associée à des indépendances d'échelles, il importe de préciser l'effet du support d'information, qui ici, comme dans l'ensemble des statistiques spatiales existe.

Schéma d'échantillonnage pour les campagnes de mesures de la qualité de l'air : une approche par optimisation

Thomas Romary, Chantal De Fouquet, Laure Malherbe

Statistique spatiale 2 - Champs de Markov

Dans ce travail, nous présentons une méthode d'échantillonnage optimal dans un contexte spatial, pour une application aux campagnes de mesure de la concentration de benzène dans l'air ambiant, à l'échelle de l'agglomération. Dans un premier temps, nous nous fondons sur l'analyse des données de campagnes antérieures, réalisées sur deux agglomérations différentes, pour définir une modélisation a priori. Précisément, la modélisation retenue est une modélisation en dérive externe comprenant une dérive et un résidu corrélé spatialement. L'analyse conduite nous permet de choisir les variables auxiliaires pertinentes et de déterminer un modèle de variogramme a priori pour le résidu. Dans un second temps, nous nous proposons d'optimiser l'implantation des appareils de mesure de la concentration de benzène sur une troisième agglomération. Pratiquement, nous cherchons à minimiser la moyenne sur l'agglomération de la variance de krigeage universel, dont les paramètres sont fondés sur la modélisation a priori. Cette optimisation est effectuée par recuit simulé. Nous présentons les résultats obtenus sur trois agglomérations différentes. Enfin, nous discutons des avantages et inconvénients de cette méthodologie en la comparant notamment à une démarche déterministe mise en oeuvre dans un travail précédent. Ce travail a été réalisé pour le Laboratoire Central de Surveillance de la Qualité de l'Air. Il a été financé par le Ministère de l'Energie, de l'Ecologie, du Développement durable et de la Mer. Les données ont été fournies par les associations AIRAQ, ATMO Nord-Pas-de-Calais et ATMO Champagne-Ardenne.

Querelle de voisinage dans les plans en cross-over

Pierre Druilhet

Plan d'expérience - Qualité - Fiabilité

L'article fondateur de Kushner (Ann. Stat. 1997) a révolutionné l'étude des plans en cross-over, permettant d'obtenir des plans optimaux ou ayant une bonne efficacité pour des modèles plus riches que précédemment. Nous présentons une généralisation de ces méthodes et nous nous comparons les résultats obtenus dans différentes situations.

Qualité des plans d'expériences SFD de grande dimension pour l'analyse de sensibilité

Olivier Vasseur, Magalie Claeys-Bruno, Michelle Sergent

Plan d'expérience - Qualité - Fiabilité

Dans le domaine de l'expérimentation numérique, lorsque les relations entre la réponse et les entrées du code de calcul sont complexes, les plans d'expériences Space Filling Designs (SFD) sont utilisés pour l'exploration du code ou la construction de métamodèles. A partir de l'utilisation de critères basés sur l'Arbre de Longueur Minimale (ALM) qui permettent d'apprécier la qualité intrinsèque de ces plans, nous comparons la qualité des résultats obtenus par ces SFD dans le cas d'analyses de sensibilité de filtres optiques interférentiels en grande dimension. En conclusion, ces travaux permettent de proposer des pistes de recherches pour relier les qualités intrinsèque et extrinsèque des plans SFD.

Utilisation d'un Modèle d'Equations Structurelles de type PLS à la validation d'un questionnaire de Culture de Sécurité

Marion Izotte, Pauline Occelli, Sandrine Domecq, Jean-Luc Quenon

Plan d'expérience - Qualité - Fiabilité

Les modèles d'équations structurelles (MES) sont des modèles statistiques complexes qui permettent de mettre en relation des concepts non observables. Ils ont été développés pour examiner des rapports de causalité multiple mais leur usage s'est aujourd'hui étendu à la validation d'instrument. Leur principe est fondé sur l'articulation d'analyses factorielles et de régressions.

Le concept de culture de sécurité (CS) est utilisé pour décrire les façons de penser, agir et sentir d'un collectif de travail en matière de sécurité. En santé, une CS développée est considérée comme un pré-requis à l'amélioration de la sécurité des soins.

L'objectif était d'utiliser un MES avec l'approche PLS (Partial Least Square) pour valider un questionnaire de mesure de la culture de sécurité (CS) des soins.

Une version française du questionnaire américain Hospital Survey On Patient Safety Culture (HSOPSC), a été choisie. La validation a été menée en deux étapes : exploratoire (corrélation, Analyse en Composantes Principales, alphas de Cronbach) et confirmatoire (MES).

Une structure en dix dimensions et quarante items a été mise en évidence. Elle a été confirmée par le MES qui a permis de hiérarchiser les dimensions selon leur influence sur la CS. L'impact de chaque item sur sa dimension a été quantifié.

Les analyses réalisées ont permis d'adapter la structure du questionnaire au contexte français. L'utilisation du MES de type PLS a permis de déterminer les éléments à cibler en priorité pour améliorer la CS.

Encore peu utilisés en santé, le MES de type PLS offre un apport intéressant dans le domaine de la validation d'instrument.

Évaluation de performances des systèmes d'attente fiables

Smail Adjabi, Karima Lagha

Plan d'expérience - Qualité - Fiabilité

Dans ce travail, on étudie le problème d'évaluation de performances d'un système d'attente de type GI/GI/1, en utilisant la propriété qualitative des distributions non paramétriques du temps des inter-arrivées. Nous obtenons ainsi les bornes pour la caractéristique de performance : temps moyen d'attente. La propriété qualitative est utilisée à défaut d'informations suffisantes, elle apporte souvent une première approximation dans les modèles de fiabilité. Nous avons utilisé deux approches, l'approche des bornes et l'approche de simulation. Dans la première approche on a déterminé les bornes de la caractéristique de performance pour le système considéré, dans lequel, on a utilisé les modèles d'Erlang, de Weibull et des échantillons validés par des tests pour les différentes classes de distributions non paramétriques. Dans la deuxième approche, on simule les caractéristiques pour ce système afin de vérifier les résultats obtenus avec ceux de la première approche.

Régression pour les évènements récurrents : application à la maintenance imparfaite des systèmes réparables

Vincent Couallier, Genia Babykina

Plan d'expérience - Qualité - Fiabilité

L'objectif de la communication est de proposer un modèle particulier, initialement proposé par Yves Legat (2009), ainsi que l'estimation statistique et les outils de prédiction associés, pour traiter les évènements récurrents de défaillance d'un système réparable. Le modèle s'appuie sur les outils fournis par l'approche probabiliste des processus ponctuels. Son intensité stochastique dépend à la fois du nombre d'évènements passés, du temps t, et de covariables pouvant dépendre du temps. Notons que le modèle proposé se place dans une classe de modèles très large proposée par Pena et Hollander (2006) mais, de part sa spécificité, nous obtenons pour notre modèle des résultats intéressants pour la prédiction. Les résultats théoriques concernent la consistence et la normalité asymptotique des estimateurs d'une part, des méthodes de prédiction du nombre de défaillances à venir d'autre part. Ce dernier résultat provient du fait qu'une loi explicite est disponible pour le nombre d'évènements à venir dans un intervalle de temps donné, une fois l'estimation par maximum de vraisemblance obtenue. La difficulté vient du shéma d'observation du processus, qui peut être censuré à droite et à gauche. Ainsi, l'histoire du processus, qui entre en jeu dans l'écriture de son intensité, n'est pas disponible à l'observateur et le processus de vraisemblance repose sur une version filtrée du processus de comptage.

Références

Y. Le Gat. (2009) Une extension du processus de Yule pour la modélisation stochastique des évènements récurrents. PhD thesis, ENGREF, 2009.

Genèse et estimation d'un modèle de fiabilité en baignoire et à taux de défaillance borné

Edwige Idee, Lambert Pierrat

Plan d'expérience - Qualité - Fiabilité

La pertinence d'un modèle de fiabilité en baignoire résulte d'un compromis entre sa représentativité et sa complexité. Nous présentons la genèse et la procédure d'estimation d'un modèle comportant un nombre minimal de paramètres et doté d'une relative souplesse d'adaptation. Celui-ci résulte de l'adaptation d'un modèle initial, basé sur une loi exponentielle dont le paramètre de forme unitaire est affecté d'une perturbation aléatoire. On obtient ainsi un taux de défaillance en baignoire (Idée, 2006), comportant deux paramètres et un taux asymptotique fini égal à l'inverse du paramètre d'échelle. L'instant correspondant au minimum de la courbe est toujours supérieur au paramètre d'échelle, ce qui dans certains cas d'application peut conduire à des limitations. Afin de ramener cet instant à une valeur inférieure au paramètre d'échelle et d'accroître l'acuité du minimum, ce modèle est modifié de la façon suivante. On fait d'abord disparaître le paramètre de perturbation en le fixant à une valeur unitaire, puis on mélange la loi correspondante à une loi gamma. On montre que si le paramètre de forme de cette loi gamma est un entier positif fixé, on obtient un modèle explicite, défini par deux paramètres : le paramètre d'échelle et un paramètre de mélange, le taux de défaillance étant encore asymptotiquement borné à l'infini. Nous présentons les principales caractéristiques de ce modèle original ainsi que la procédure d'estimation de ses deux paramètres par la méthode du maximum de vraisemblance sous certaines conditions et dans le cas d'un échantillon complet. Partant d'un échantillon de 60 cycles de défaillance relatif à un dispositif technique, les performances de ce modèle à deux paramètres sont comparées à celles d'un modèle classique à cinq paramètres, basé sur le mélange de deux lois de Weibull. On montre l'intérêt du modèle proposé en termes de meilleure conformité aux données pour une complexité moindre.

Références

Idée, E (2006), Loi de probabilité obtenue par perturbation aléatoire simple de l'un des paramètres de la loi de référence, Prépublication du LAMA-Université de Savoie, 06-2007.

Extremal events in a bank operational losses

Hela Dahen, Georges Dionne, Daniel Zajdenweber

Partenariat AXA : Finance - Assurance - Actuariat

Les pertes opérationnelles sont des dangers importants pour les banques car leurs valeurs maximales sont difficiles à prédire, ce qui est différent du risque de crédit. Par exemple, nos données indiquent qu'une très grande banque américaine peut subir, en moyenne, plus de quatre pertes majeures durant une année. Cette banque a plusieurs pertes excédant des centaines de millions de dollars sur les 52 pertes documentées de plus de 1 million de dollars sur la période 1994-2004. La queue de distribution de Pareto $(0,95 \leq alpha \leq 1)$ indique que cette banque peut atteindre des pertes entre 1 milliard et 11 milliards de dollars à des probabilités situées respectivement à 1 % et 0,1 %. Les primes d'assurance correspondantes sont évaluées entre 245 millions et près de milliard.

Risques extrêmes en assurance vie

Johan Attal
Partenariat AXA : Finance - Assurance - Actuariat

Les évènements extrêmes sont des évènements d'occurrence très faible et de coût très important. En Vie individuelle, le risque pour AXA est d'une part un risque technique lié aux garanties de prévoyance (garanties en cas de décès, invalidité et arrêts de travail) et d'autre part financier (garanties en cas de décès sur les contrats d'épargne). L'enjeu de la cartographie des différents risques auquel nous sommes exposés, ainsi que de la mesure de ces risques, est majeur car il permet de définir notre besoin en fonds propres.

La connaissance de l'historique des sinistres en France n'est pas suffisante pour appréhender notre exposition. Il n'y a en effet pas eu de catastrophes majeures en France et pourtant on ne peut pas considérer que la probabilité d'un tel évènement est nulle. En l'absence de distribution de probabilités, il est donc nécessaire de définir des "worst case scenarios" et d'en mesurer l'impact sur notre portefeuille. Nous calibrons donc un ou plusieurs scénarios de nombre de décès, en supposant que leur probabilité d'occurrence sur un an est inférieure à 0.05.

Modélisation multivariée des comportements à risque des conducteurs d'automobile

Mériem Maatig
Partenariat AXA : Finance - Assurance - Actuariat

L'objet de cette étude est d'avoir une spécification économétrique qui modélise simultanément les comportements à risque des conducteurs. En effet, l'existence d'un biais d'endogénéité peut nous amener à retenir, à tort, des variables explicatives dans une équation séparée. Pour cela nous nous appuyons sur des données collectées à partir d'une enquête. Nous avons réalisé l'enquête par questionnaire sur un échantillon représentatif de la région parisienne en termes de sexe et âge. Un modèle probit trivarié a été effectué pour tester l'existence de liens de causalité entre les trois comportements à risques des conducteurs. Dans cette modélisation, les caractéristiques qui expliquent ces trois comportements sont corrélées. Dans ce cas, l'estimation autonome de l'une de trois équations peut comporter un biais d'endogénéité. Les caractéristiques individuelles qui expliquent l'utilisation des places de parking réservées aux handicapés et la conduite sous l'emprise de l'alcool, expliquent aussi négativement la probabilité de l'utilisation du téléphone portable au volant.

Facteurs explicatifs du rachat en Assurance-Vie : classification et prévisions du risque de rachat

Xavier Milhaud, Stéphane Loisel, Véronique Maume-Deschamps
Partenariat AXA : Finance - Assurance - Actuariat

En Assurance Vie, certaines caractéristiques de contrats jouent un rôle majeur dans la décision de l'assuré de racheter son contrat. Les conditions de souscription, son âge, sa profession ainsi que d'autres facteurs propres à sa situation lors de la souscription influencent ses décisions. Il est également possible que l'environnement économique et financier soit important. Deux modèles de segmentation nous ont permis de développer ces idées :

les arbres de classification et de régression, et la régression logistique. Les contrats de type Prévoyance ainsi que ceux d'Epargne sont impactés, et les résultats montrent clairement que la garantie de participation au bénéfice est très discriminante. Nous nous focalisons dans cette étude sur des produits de type Mixte. Nous présentons dans un premier temps les fondamentaux de chacun des modèles ainsi que leurs hypothèses et limites. Puis nous testons différents facteurs comme possibles déclencheurs de la décision de rachat, dans le but de segmenter le portefeuille en classe de risque : la duration du contrat et l'option de participation au bénéfice sont des éléments essentiels. En dernière partie, nous discutons des différences entre les deux modélisations en termes de résultats numériques et d'un point de vue opérationnel.

Modélisation bayésienne hiérarchique pour l'estimation de matrice de covariance - Application à la gestion actif-passif de portefeuilles financiers

Mathilde Bouriga, Olivier Féron, Jean-Michel Marin, Christian. P Robert

Partenariat AXA : Finance - Assurance - Actuariat

Ce papier concerne l'estimation de matrices de covariance dans le cas où le nombre de données utilisées pour l'estimation est faible par rapport à la dimension du problème et où les méthodes d'estimation classiques fondées sur le Maximum de Vraisemblance sont peu robustes. Nous proposons une méthode d'estimation non supervisée fondée sur une modélisation bayésienne hiérarchique du problème d'estimation de matrice de covariance : on pose une loi Inverse Wishart a priori pour la matrice, conditionnellement aux hyper-paramètres sur lesquels on pose des a priori de référence. On considère une matrice cible de structure diagonale. L'estimateur bayésien associé au coût entropique sera approché par des méthodes de type Monte Carlo par Chaînes de Markov. On comparera empiriquement notre estimateur avec celui obtenu par Maximum de Vraisemblance. L'aspect régularisant de la méthode sera étudié en l'appliquant sur données financières dans un cadre de gestion actif-passif, où le nombre de données est faible par rapport à la taille de la matrice.

Les transformations de Schoenberg : propriétés et applications en Analyse des Données

François Bavaud

Apprentissage - Classification 1

la classe de toutes les fonctions transformant, composante par composante, une distance euclidienne en une autre distance euclidienne a été établie par Schoenberg en 1938. Ce résultat, assez méconnu en Analyse des Données, élargit le champ d'application du "multidimensional scaling", dans sa version ordinaire ou pondérée. Ainsi que l'illustrent quelques exemples, son potentiel semble prometteur pour les applications faisant intervenir des dissimilarités euclidiennes.

Condition nécessaire et suffisante de convergence en loi de l'estimateur des plus proches voisins

Alain Berlinet, Rémi Servien

Apprentissage - Classification 1

L'estimateur des plus proches voisins de la densité est un estimateur simple et facile à mettre en oeuvre. Sa normalité asymptotique a été établie par Moore et Yackel (1977) sous des hypothèses faisant intervenir les dérivées de la densité. Sans faire d'hypothèse de continuité sur la densité, nous donnons une condition nécessaire et suffisante de convergence en loi de cet estimateur. Nous utilisons pour cela l'indice de régularité d'une mesure de probabilité (Beirlant, Berlinet et Biau (2008)) qui intervient de fait dans la loi limite.

Références

Beirlant, J., Berlinet, A. et Biau, G. (2008) Higher order estimation at Lebesgue points, Annals of the Institute of Statistical Mathematics, 60, 651-677.

Moore, D.S. et Yackel, J.W. (1977) Large sample properties of nearest neighbour density function estimates, Statistical Decision Theory and Related Topics II, Gupta, S.S. et Moore, D.S., Academic Press, New-York.

Choix d'un indice de capabilité basé sur des objectifs industriels: proportion de non-conformes et centrage

Daniel Grau

Apprentissage - Classification 1

Les indices de capabilité ont été introduits pour mesurer la performance d'un processus de production industrielle. Le concept de performance, initialement lié à la proportion de pièces non conformes fabriquées, a rapidement évolué pour tenir compte aussi de la position de la moyenne du processus par rapport à sa cible. Si les liens entre les indices et le centrage ont déjà été étudiés, ceux entre les indices et la proportion de non conformes ne l'ont été que partiellement. Dans cet exposé nous clarifions ces liens et montrons sur un exemple réel comment ces résultats peuvent être utilisés.

Exploitation of Unsupervised Cumulative Precision Measures for Efficient Clustering Quality Estimation

Jean-Charles Lamirel

Apprentissage - Classification 1

In the context of unsupervised classification, or clustering, the fact of not having a reference classification represents a heavy handicap to evaluate the performance of the algorithms. On their own side, traditional quality indexes (Inertia, DB...) do not allow to properly estimate the quality of the clustering in several cases, as in that one of the textual data. We thus present an alternative approach for clustering quality evaluation based on unsupervised measures of Recall, Precision and F-measure exploiting the descriptors of the

data associated with the obtained clusters. The Recall makes it possible to measure the exhaustiveness of the contents of the clusters in terms of peculiar descriptors specific to each cluster. The Precision measures the homogeneity of the clusters in terms of proportion of the data containing the associated peculiar descriptors. This paper especially focuses on the construction of a new cumulative Micro precision index that makes it possible to evaluate the overall quality of a clustering result while clearly distinguishing between homogeneous and heterogeneous results. The experimental comparison of the behavior of the classical indexes with our new index is performed on a dataset of bibliographical references issued from the PASCAL database.

Algorithme des k plus proches voisins pondérés et application en diagnostic

Eve Mathieu-Dupas

Apprentissage - Classification 1

La méthode des k plus proches voisins pondérés est une méthode de classification supervisée offrant des performances très intéressantes dans la recherche de nouveaux biomarqueurs pour le diagnostic. Nous présentons le fondement théorique de cette méthode et illustrerons cette technique d'apprentissage statistique au travers du problème diagnostique d'une pathologie complexe.

Etudes des lésions pulmonaires chez le porc : une analyse symbolique sur des concepts issus de l'approche classique

Christelle Fablet, Carole Toque, Stéphanie Bougeard, Edwin Diday

Biostatistique 3 : médecine

L'objectif de l'étude est de décrire les liens qui peuvent exister entre des variables quantitatives et qualitatives caractérisant les lésions pulmonaires chez les porcs, et plus particulièrement les deux maladies que sont la pneumonie et la pleurésie. Les données présentent une structure emboitée et consistent en 125 élevages comprenant chacun 30 porcs. Une approche classique, nécessitant la création de nombreuses variables agrégées telles que des moyennes et des médianes au niveau des élevages, basée sur l'analyse en composantes principales (ACP) suivie d'une classification ascendante hiérarchique (CAH), conduit à la formation de quatre classes d'élevage. Parallèlement, une analyse de données symboliques dont une ACP, a été conduite principalement sur des variables à valeur histogramme correspondant directement aux fréquences des porcs de chaque élevage. Nous concluons à la pertinence de l'analyse de données symboliques appliquée grâce au logiciel SYR en épidémiologie animale. En particulier, l'analyse de données symboliques permet de prendre en compte la variabilité des données initiales, et les différences entre classes sont identifiées grâce au logiciel qui permet de trier les élevages et les variables sur la base des fréquences de leurs modalités en s'affranchissant de la création de nombreuses variables de synthèse.

Procédures d'inférence bayésienne pour des dispositifs adaptatifs de recherche de doses dans les études cliniques

Bruno Lecoutre, Gérard Derzko

Biostatistique 3 : médecine

Dans le développement d'un nouveau médicament, c'est au cours de la phase 2 que la relation dose-réponse est évaluée et que les doses les plus prometteuses sont sélectionnées pour la phase 3. A côté du dispositif en groupes parallèles de doses, il existe des dispositifs adaptatifs visant à réduire le nombre de patients soumis aux doses les moins efficaces. En particulier lorsque la réponse est binaire (succès/échec) plusieurs dispositifs adaptatifs ont pu être proposés, qui de plus permettent que les réponses soient différées, notamment les modèles d'urne de Freedman généralisés, d'allocation linéaire, ou encore les dispositifs dits "drop-the-loser" et "doubly adaptive biased coin design". La complexité des distributions d'échantillonnage de ces dispositifs est une source de difficulté pour les méthodes d'inférence fréquentistes. L'approche bayésienne apparaît plus simple et plus générale, car elle est basée sur la fonction de vraisemblance, qui pour ces plans est simplement proportionnelle à la fonction de vraisemblance associée à la comparaison de proportions binomiales indépendantes. Cependant, dans le contexte des essais cliniques, il est nécessaire que ces méthodes répondent aux critères fréquentistes standard. Dans le cas de deux traitements, l'étude détaillée des performances fréquentistes de procédures bayésiennes non informatives pour différents dispositifs adaptatifs conduit à une conclusion favorable à ces méthodes (Lecoutre et al, 2010). Nous étendrons cette étude au cas de plus de deux traitements, situation plus usuelle dans les études de phase 2.

Références

Lecoutre, B., Derzko, G. et Elquazyr, K. (2010). Frequentist performance of Bayesian inference with response-adaptive designs.

Proposition et étude longitudinale d'un indicateur de la performance avec un implant cochelaire

Julie Bestel, Pierre-Louis Gonzalez, Nathalie Noel-Petroff, Thierry Van Den Abbeele

Biostatistique 3 : médecine

Nous proposons une étude rétrospective de la base de données d'enfants porteurs d'un implant cochléaire de la même marque, et suivis dans le même hôpital. Comme le but de l'implant cochléaire est de restaurer l'audition, les patients sont évalués sur différents aspects : acceptation du système, niveau de détection des sons, compréhension du message oral, expression, et compréhension par autrui. Notre base de données comporte donc cinq évaluations, notées chacune de 0 à 5, et réalisées à différentes sessions : de 3 mois à 4 ans d'utilisation. Le fichier contient également des données médicales et matérielles, possiblement explicatives des résultats. Nous proposons de construire un indicateur unique noté K, à l'aide d'une typologie par arbre de classification, sur l'ensemble des patients ayant 2 ans d'implant, recul pertinent en clinique. La règle de construction de K est alors appliquée à tous les sujets, pour toutes les sessions. Nous étudions ensuite la relation entre K et les variables possiblement prédictives, à l'aide d'une régression logistique multinomiale. Parmi les nombreux modèles testés, le modèle utilisant la fonction de lien logit généralisé, construit à l'aide des trois variables explicatives - durée d'utilisation, niveau d'audition

pré-implant et stratégie de codage du signal - est intéressant en clinique. Il rend compte du fait que la probabilité d'être dans une classe de "mauvais" résultats diminue avec la durée d'utilisation, diminue si le patient avait une expérience auditive avant l'implant, et diminue avec la dernière génération de codage du son introduite par le constructeur.

Risque de Lentigines et comportement d'exposition et protection face au soleil

Emmanuelle Mauger, Khaled Ezzedine, Randa Jdid, Olivier Nageotte, Julie Latreille, Pilar Galan, Serge Hercberg, Christiane Guinot

Biostatistique 3 : médecine

Le but de cette recherche était d'étudier le risque de lentigines chez des femmes adultes en fonction de l'âge et de leurs caractéristiques phénotypiques, génétiques, et comportementales. Une typologie d'exposition et protection a d'abord été recherchée à partir d'un questionnaire administré à un premier échantillon de femmes. Un arbre de décision a ensuite été construit et des règles de décision définies. Ces règles ont ensuite été appliquées à un second échantillon de femmes pour lesquelles le comportement face au soleil était connu, ainsi que la présence éventuelle de lentigines au niveau du visage, et pour lesquelles des données phénotypiques et génétiques étaient disponibles. Finalement, les facteurs de risque de lentigines ont été recherchés grâce à une série de régressions logistiques. La première étude a permis d'identifier quatre types de comportement et d'établir des règles de décision reposant sur trois items. Ces règles ont ensuite permis d'affecter chacune des femmes de la seconde étude à la typologie. Le risque de lentigines a été trouvé significativement lié à l'âge, à la couleur de la peau et des cheveux, à certains polymorphismes du gène MC1R, à la pratique de sports nautiques et au comportement d'exposition et de protection face au soleil. Cette recherche met en avant le lien complexe qui existe entre la couleur de la peau, l'indice de protection des produits solaires utilisés et le comportement individuel d'exposition.

Fonction d'influence pour la reconstruction de phylogénies robustes

Mahendra Mariadassou, Avner Bar-Hen

Biostatistique 3 : médecine

Les arbres phylogénétiques sont une bonne façon de représenter les liens de parenté entre espèces et sont couramment utilisés dans de nombreux domaines de la biologie : génomique comparative, épidémiologie, biologie de la conservation, etc. Pour toutes ces applications, l'arbre reconstruit doit être le plus proche possible du vrai arbre. A erreur de reconstruction donnée, l'arbre doit en particulier être le plus robuste possible. Une source majeure de non robustesse est la contamination du jeu de données par des données aberrantes. En conséquence, détecter et isoler les données aberrantes constitue une stratégie qui permet de robustifier l'arbre. Nous montrons ici comment les fonctions d'influence empirique permettent de détecter les données influentes, susceptibles d'être aberrantes, et comment supprimer les données les plus exceptionnelles permet de reconstruire un arbre robuste. L'application à deux jeux de données (mammifères à placentas et zygomycètes) montre que la reconstruction par maximum de vraisemblance n'est pas robuste aux données aberrantes et que la suppression des quelques données les plus influentes améliore sensiblement la robustesse de l'arbre reconstruit.

Résultats asymptotiques pour la méthode systématique de Deville

Guillaume Chauvet, Jean-Claude Deville
Sondages et statistique publique

Nous nous intéressons à un algorithme d'échantillonnage connu dans la littérature sous le nom de méthode systématique de Deville. Cet algorithme permet de sélectionner sans remise des échantillons de taille fixe, avec des probabilités d'inclusion fixées. Dans ce travail, nous montrons que cet algorithme peut être bien approché par une autre procédure de tirage où des individus sont sélectionnés indépendamment. Cette comparaison permet d'obtenir des résultats asymptotiques pour l'estimateur de Horvitz-Thompson.

Propriétés asymptotiques d'estimateurs non paramétriques model-based de la fonction de répartition sur un petit domaine

Sandrine Casanova, Eve Leconte
Sondages et statistique publique

Nous nous intéressons à l'estimation de la fonction de répartition (f.d.r.) en sondage sur des sous-populations (domaines). Si un domaine est de taille suffisante, l'estimation de la f.d.r. se base uniquement sur les individus du domaine et les estimateurs produits sont de précision acceptable. Cependant, dans la plupart des applications, les tailles d'échantillons correspondant à des petits domaines ne sont pas suffisantes. L'estimation se fonde alors sur une information auxiliaire fournie par une covariable et de l'information est "empruntée" aux autres domaines. Dans ce contexte, Chambers & Tzavidis (2006) ont proposé un estimateur paramétrique de la f.d.r. sur un domaine, basé sur des quantiles. Aragon & Casanova (2007) et Casanova (2010) ont adapté cet estimateur au cas non paramétrique et ont proposé un autre estimateur basé sur les quantiles. Ces estimateurs se placent dans un cadre model-based où le problème est de prédire la variable d'intérêt pour les individus non échantillonnés du domaine. Pour un individu fixé, sa variable d'intérêt peut toujours être vue comme le quantile conditionnel à la valeur de sa covariable pour un certain ordre appelé ordre-quantile. Les ordres-quantiles des individus de l'ensemble des échantillons sont estimés et on prédit ensuite à l'aide de polynômes locaux la variable d'intérêt d'un individu hors échantillon par les quantiles conditionnels associés aux ordres qui décrivent ou résument le domaine de l'individu. Nous nous focalisons ici sur les propriétés asymptotiques de ces estimateurs avec une approche model-based : nous étudions leur biais asymptotique sous le modèle ainsi que leur convergence en moyenne quadratique.

Références

Aragon Y. et Casanova S. (2007). Estimation de la fonction de répartition sur un domaine à l'aide des quantiles et des M-quantiles conditionnels, xxxix èmes Journées de Statistique, Angers.
Casanova S. (2010). Using M-quantiles to estimate a cumulative distribution function in a domain. En révision aux Annales d'Economie et de Statistique.
Chambers, R. L. and Tzavidis, N. (2006). M-quantiles models for small area estimation, Biometrika, 93, 255-268.

Inégalités de concentration pour le sondage aléatoire simple

Daniel Bonnery
Sondages et statistique publique

En théorie des sondages, la loi de l'estimateur de Horvitz Thompson de la moyenne d'un caractère d'une population est, en pratique, approximée par une loi normale. Cette approximation est justifiée par un théorème de Hàjek, qui démontre la convergence vers une loi normale de l'estimateur de Horvitz Thompson sous certaines conditions. Il existe aussi des résultats à distance finie, notamment ceux de Hoeffding et Serfling qui permettent d'obtenir des inégalités de concentration étant donné la dispersion du caractère sur la population. Toutefois, les inégalités obtenues s'avèrent très larges et sont peu utilisées en pratique. Lors de l'exposé, on s'attachera à montrer comment obtenir des inégalités de concentration plus fines pour ce problème particulier en utilisant des outils moins puissants, mais aussi moins généraux. Dans un premier temps en utilisant les symétries de l'espace des échantillons de taille n, et dans un second temps, en munissant l'espace des échantillons de la distance de Hamming, et en considérant la mesure des boules centrées en un échantillon. Enfin, on comparera les inégalités obtenues avec les inégalités de Hoeffding et Serfling.

Multilevel models and small area estimation in the context of Vietnam living standards surveys

Phong Nguyen, Dominique Haughton, Irene Hudson, John Boland
Sondages et statistique publique

L'exposé présente une méthode pour obtenir des estimateurs pour petites régions dans le contexte des Enquêtes sur le Niveau de Vie au Vietnam. On introduit brièvement ces enquêtes, puis on rappelle les concepts principaux en estimation pour petites régions, notamment l'utilisation de données auxiliaires, et on contraste les modèles simples avec ceux de régression. On traite les effets aléatoires dans ces modèles et on propose un modèle multi-niveaux pour une estimation au niveau de la commune au Vietnam, à notre connaissance le premier modèle de ce type construit à partir de données sur le niveau de vie au Vietnam. Notre modèle pour la moyenne au niveau communal du logarithme des dépenses familiales par personne utilise des variables indépendantes disponibles par le Recensement de 1999 et l'enquête aux ménages de 2002 et suit des idées exposées par Moura (1994, 1999); on discute comment mesurer la précision du modèle.

Les statistiques ethniques sont-elles éthiques ?

Stéphane Jugnot
Sondages et statistique publique

Depuis 2006, la question des statistiques ethniques fait débat en France, sans qu'aucun consensus n'existe, ni parmi les acteurs sociaux, ni parmi les chercheurs, dont les praticiens de la statistique. Sous la question de l'outil, c'est en fait celle de son impact social qui est en cause. D'un côté, la crainte de racialiser et fragmenter la société. De l'autre, celle de ne pas se donner les moyens de lutter contre les discriminations. Pour cette raison, le débat des statistiques ethniques n'est pas un débat purement technique ou scientifique, il est d'abord politique et concerne l'ensemble des acteurs sociaux. Pour cette raison, les statistiques ethniques posent des questions éthiques et déontologiques que cette contribution se proposera d'aborder. Après avoir rappelé le contexte récent, la contribution abordera la question des finalités de l'outil, dans la mesure où la pertinence, donc la légitimité, de

l'outil dépend des usages poursuivis. Nous tenterons de montrer que les statistiques ethniques ne sont légitimes que pour accompagner une politique de promotion de la diversité utilisant les mêmes catégories, nécessitant donc un choix politique préalable. Une fois le cadre clarifié, nous aborderons enfin les questions éthiques soulevées par le débat récent.

Estimation du paramètre de longue mémoire de séries temporelles non linéaires

Marianne Clausel, François Roueff, Murad Taqqu, Ciprian Tudor

Statistique des processus de type fractal

On considère une série temporelle à mémoire longue de la forme G(X) où X est une série temporelle gaussienne. En utilisant la décomposition en chaos de Wiener du processus, nous étudions le comportement de l'estimateur du paramètre de longue mémoire basé sur les coefficients d'ondelettes. Nous montrons notamment que différents comportements peuvent se produire selon le rang de Hermite de la fonction Hermite G considérée mais aussi les coefficients de G dans sa décomposition en série de Hermite. Nous illustrons ces différents comportements par des exemples montrant que suivant les cas le comportement asymptotique de l'estimateur peut être assez proche ou non de ce qui est connu dans le cas gaussien.

On the identification of hidden pointwise Hölder exponents

Antoine Ayache, Qidi Peng

Statistique des processus de type fractal

L'exposant de Hölder ponctuel (EHP) d'un processus stochastique X permet de mesurer la régularité locale de X. Nombre d'auteurs se sont déjà intéressès au problème de l'estimation de cet exposant à partir de l'observation d'une trajectoire discrétisée de X. Cependant, il ne semble pas toujours réaliste de supposer qu'une telle observation est directement accessible mais seulement une version corrompue de celle-ci. Est-il alors possible d'effectuer l'estimation ?

A notre connaissance, cette question a été assez peu étudiée dans la littérature et les articles qui l'abordent se placent dans un cadre qui est essentiellement celui de processus Gaussiens à accoissements stationnaires; l'EHP a alors une structure relativement simple puisqu'il reste contant au cours du temps. L'objectif de notre exposé est d'étudier cette question dans un cadre nouveau, celui du mouvement Brownien multifractionnaire; l'EHP de ce processus a généralement une structure assez complexe parce qu'il varie d'un instant à l'autre.

Modélisation d'une série financière par mouvement Brownien multi-fractionnaire parcimonieux

Pierre R. Bertrand, Abdelkader Hamdouni, Nabiha Haouas, Samia Khadraoui

Statistique des processus de type fractal

Dans ce travail, nous introduisons un modèle parcimonieux de processus de type fractal. En premier, nous rappelons le passage du mouvement brownien fractionnaire (fBm) au mouvement brownien multi-fractionnaire (mBm) et proposons de sélectionner un modèle parcimonieux. En second, nous étayons notre point de vue par des expériences statistiques et l'observation d'un artefact numérique: quand nous estimons l'index de Hurst dépendant du temps H(t) pour un mouvement brownien fractionnaire, la fluctuation de l'échantillonnage donne l'impression que H(t) est lui-même un processus stochastique, même lorsque H(t) est constant. Nous appliquons cette modélisation à des données financières réelles: la série des prix journaliers du Nasdaq de 1971 jusqu'à la fin 2009.

Etude comparée de classifications sur matrices très creuses et de grandes dimensions

Mireille Gettler Summa, Francesco Palumbo, Cristina Tortora

Apprentissage - Classification 2

Les méthodes de classification non supervisée ont pour but de révéler une structure entre des éléments, selon les associations qu'on peut y détecter par leurs valeurs sur un ensemble de variables. Lorsque l'on s'intéresse à des grands ensembles d'unités, il est nécessaire d'en réduire la dimensionnalité avant le processus de classification. Quand les variables présentent des liens non linéaires, les approches classiques sont inopérantes. Les classifications de variables qualitatives soulèvent dans ce sens de nombreux problèmes ; les associations sont en général non linéaires. Avec un recodage binaire de l'ensemble des modalités des variables, on obtient le plus souvent des matrices très creuses et de grande dimension. Pour contourner la situation, quand le nombre de variables est important, l'approche plus utilisée est de transformer les variables qualitatives en variables continues, puis de faire la classification sur les valeurs de ces dernières. Notre travail s'attache à classifier de façon non supervisée des variables qualitatives dans le contexte général suivant : il n'y a pas de liens linéaires entre les variables et elles sont en grand nombre. Nous proposons une approche en plusieurs étapes: Analyse factorielle, redéploiement des coordonnées des premiers axes factoriel dans un espace de dimension supérieure, construction des classes dans ce dernier espace, enfin visualisation des classes obtenues dans l'espace des facteurs. On appliquera cette approche sur les données "epub" du "CRAN-R", et nous nous intéresserons sur cet exemple à la comparaison entre l'approche par le détour des vecteurs de support et celle classique d'un arbre hiérarchique.

Inégalités d'oracle exactes pour la prédiction d'une matrice en grande dimension

Stéphane Gaiffas, Guillaume Lecué, Alexandre Tsybakov

Apprentissage - Classification 2

Nous considérons le problème de prédiction d'une matrice de taille $m \times T$ en "grande dimension", c'est-à-dire où mT est bien plus grand que la taille de l'échantillon n. Pour cela, nous considérons l'algorithme de minimisation de la norme trace, ou des versions "élastiques" de la norme. Cet algorithme est maintenant bien connu et utilisé pour les problèmes de complétion de matrice ou d'apprentissage multi-tache, voir notamment [1], [2], [3], [4], parmi d'autres. Dans ce travail, nous proposons des inégalités d'oracle "exactes" pour ces algorithmes.

Références

[1] Francis R. Bach. Consistency of trace norm minimization. J. Mach. Learn. Res., 9:1019-1048, 2008.
[2] E. J. Candès and Y. Plan. Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. Technical report, Department of Statistics, Stanford Univeristy, 2009.
[3] E. J. Candès and Y. Plan. Matrix completion with noise. Proceedings of the IEEE, to appear.
[4] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. Found. of Comput. Math., 9:717-772, 2008.

Vitesse minimax du regret interne en prédiction de suites individuelles

Sebastien Gerchinovitz

Apprentissage - Classification 2

Le problème de la prédiction séquentielle avec avis d'experts consiste à prédire tour après tour les valeurs d'une certaine suite (températures, pics d'ozone journaliers), à l'aide de prédictions de base qu'on peut combiner pour former une seule prédiction. Dans cette communication, nous nous intéressons à une formalisation générique de ce problème de décision séquentielle, et étudions la vitesse minimax d'un critère de performance, le regret interne. D'après les travaux de Stoltz (2005), Stoltz et Lugosi (2005), ainsi que Blum et Mansour (2007), cette vitesse est comprise entre $\Omega(\sqrt{n})$ et $\mathcal{O}(\sqrt{n \ln N})$, où n désigne le nombre de tours de prédiction et N le nombre d'actions. Nous montrons que le terme $\sqrt{\ln N}$ est absent dans deux quantités maximin et minimax associées, où les pertes (stochastiques) sont supposées indépendantes et i.i.d. respectivement.

Références

Blum, A. and Mansour, Y. (2007) From External to Internal Regret, Journal of Machine Learning Research 8, 1307-1324.
Stoltz, G. and Lugosi, G. (2005) Internal regret in on-line portfolio selection, Proceedings of the 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop, 403-417, Springer.
Stoltz, G. (2005) Incomplete information and internal regret in prediction of individual sequences, PhD dissertation, Université Paris-Sud.

Discrimination et Classification supervisée en référence à des prototypes

Stéphane Verdun, Véronique Cariou, El Mostafa Qannari

Apprentissage - Classification 2

L'objectif de la classification supervisée est d'affecter des individus à des groupes définis a priori à partir des mesures effectuées sur des variables. Dans ce contexte, les analyses discriminantes linéaire et quadratique sont parmi les méthodes les plus populaires. Elles sont fondées sur des hypothèses de multinormalité. Dans certaines situations, cette règle s'avère inappropriée en particulier dans le cas d'un groupe multimodal ou en présence d'éléments atypiques. Pour pallier ce problème, la méthode proposée consiste à pondérer les individus de manière à déterminer des statistiques robustes. Dans une communication précédente, nous avons introduit une méthode de classification (non supervisée) basée sur la détermination d'une matrice stochastique. Un des intérêts de cette approche est d'exhiber des

barycentres pondérés (appelés prototypes) au sein des différents groupes. Cette démarche est étendue au cadre de la classification supervisée et au cadre de la discrimination. Pour la classification supervisée, nous adoptons une démarche similaire à celle préconisée dans le cadre des réseaux de neurones probabilistes. Pour l'analyse factorielle discriminante, nous utilisons le système de pondération pour l'estimation des paramètres de localisation des matrices de variance covariance à l'intérieur de chaque groupe, ainsi que de la matrice de variance covariance totale.

Contribution à la squelettisation en niveaux de gris

Rabaa Youssef, Sylvie Sevestre-Ghalila, Anne Ricordeau

Image

La sensibilité au bruit de la squelettisation d'images en niveaux de gris peut être gérée par l'amincissement paramétré. Son principe est d'ajouter au critère de pixel "simple" un critère basé sur le contraste local qui permet de lisser le squelette gris en fusionnant les régions sombres se différenciant par un faible contraste. Pour ce faire, un écart de niveaux représentatif de la zone est comparé à un paramètre fixé par l'utilisateur. Par un ensemble de simulations, on exhibe ici la croissance de ce paramètre avec l'amplitude du bruit. On propose ensuite une première régulation de ce paramètre de contraste en fonction des moments de l'étendue des zones sombres à fusionner.

Décisions en environnement non stationnaire par méthodes d'ensembles via One-class SVM - application à la segmentation d'images texturées

Pierre Beauseroy, André Smolarz, Xiyan He

Image

Nous proposons une méthode de décision basée sur un ensemble d'espaces de représentation, dans le but d'optimiser ou de préserver les performances d'un système décisionnel en présence de bruit, de perte d'information ou de non-stationnarité. Le principe de l'approche proposée consiste à apprendre un ensemble de règles de décision. Chaque règle est bâtie sur un sous-espace de représentation extrait d'une projection de l'espace initial. La décision finale est obtenue par combinaison de l'ensemble des règles. La spécificité de la méthode réside dans le fait que seules les règles définies dans des espaces non perturbés par la non stationnarité sont prises en compte pour la décision finale. Sélectionner les sous-espaces non perturbés et prendre simultanément la décision peut s'apparenter à un problème connu sous l'appellation détection d'anomalies ou de nouveautés. Il est alors possible de l'aborder au moyen de méthodes de classification à une classe de type One-class SVM, consistant à définir une règle de décision pour chaque classe et chaque sous-espace de représentation. La segmentation d'images texturées constitue une application tout à fait appropriée pour illustrer cette méthode et évaluer ses performances. Nous présentons ici quelques résultats obtenus sur une application de segmentation d'images à deux classes de textures et nous nous intéressons plus particulièrement au comportement de la méthode proposée aux frontières inter-régions.

Réduction de dimension par un nouvel estimateur de la distance de Patrick Fisher à l'aide des fonctions orthogonales

Faouzi Ghorbel, Wissal Drira

Image

Here, we intend to introduce new estimators of the Patrick Fisher distance and of the Euclidian probabilistic dependence measure by using orthogonal series. Therefore non parametric multiclass dimension reduction could be easily generalized to the multivariate case reduction. An asymptotical study will be presented to show that the estimate is consistent. The performance of the proposed estimates will be studied by simulations in the case of mixtures of non Gaussian multivariate distributions. Such estimators will be applied in Face recognition.

Estimation du paramètre du champ de Ising et fonction de partition

Jean-François Giovannelli

Image

Le papier propose de nouveaux estimateurs du paramètre du champ de Ising. Ils sont fondés sur une expression explicite et relativement simple de la fonction de partition du champ. Celle-ci est connue de la physique statistique depuis longtemps (Onsager, 1944), mais, à notre connaissance, elle n'a jamais été exploitée pour des méthodes d'estimation du paramètre. Tirant partie de ce résultat, le papier propose plusieurs estimateurs fondés sur la vraisemblance exacte et concurrents de stratégies existantes fondées sur la pseudo-vraisemblance. Une étude numérique en terme de biais et variance en fonction de la vraie valeur du paramètre montre que les estimateurs proposés offrent des performances largement meilleures que celui fondé sur la pseudo-vraisemblance.

Références

L. Onsager, "A two-dimensional model with an order-disorder transition", Phys. Rev., vol. 65, n°3 & 4, pp. 117-149, February 1944.

Comparaisons généralisées par paires pour la comparaison de deux groupes

Marc Buyse

Partenariat Danone : Biostatistiques 4 : Tests multiples - Risques alimentaires

Dans cet exposé, nous généralisons la statistique U de Wilcoxon-Mann-Whitney pour comparer deux groupes d'observations. Les observations, éventuellement répétées dans le temps, peuvent provenir d'un seul critère de jugement ou de plusieurs critères de jugement quels que soient leur nature (variables binaires, variables continues, temps jusqu'à un événement, etc.) Dans le cas de multiples observations par patient, un ordre de priorité doit être défini pour les mesures répétées et / ou les différents critères d'intérêt. L'approche proposée généralise plusieurs tests non paramétriques, parmi lesquels le test exact de Fisher

pour une variable binaire, le test de Wilcoxon-Mann-Whitney pour une variable continue, et le test de Gehan pour un temps jusqu'à un événement. L'approche mène naturellement à une mesure universelle de la différence entre les groupes, appelée la "différence de paires favorables". Il existe des relations simples entre cette mesure et les mesures habituelles de différence entre deux groupes, telles la différence de risque pour une variable binaire, la taille d'effet (ou différence standardisée) pour une variable continue et le risque relatif ("hazard ratio") pour un temps jusqu'à un événement. Nous utiliserons plusieurs exemples provenant d'essais cliniques randomisés pour montrer la très grande versatilité de l'approche proposée.

Forces et faiblesses de différentes approches statistiques pour l'analyse d'évènements récurrents

Jérôme Tanguy, Anissa Elfakir, Sébastien Marque

Partenariat Danone : Biostatistiques 4 : Tests multiples - Risques alimentaires

La modélisation d'événements récurrents est une problématique rencontrée dans une large variété de disciplines telles les sciences sociales, la recherche médicale, la santé publique, l'épidémiologie ou l'économie. Dans le cadre d'une étude clinique en nutrition, un événement de nature récurrente peut être est abordé selon deux approches différentes : une approche considérant l'événement comme une donnée de comptage, et une approche considérant l'évènement comme une donnée type survie pour occurrence multiple. Dans les deux cas plusieurs modélisations sont présentées et comparées d'un point de vue clinique et statistique (tests de comparaison pour des modèles emboités et non emboités. Forces, faiblesses et évolutions possibles de chaque approche sont soulignées.

La structure complexe de l'événement étudié offre une occasion d'aborder une grande variété de modèles statistiques. Pour l'approche type données de comptage, des modèles classiques tels que la régression binomiale négative sont appliqués, mais aussi des modèles à mélange de distribution comme les modèles modifiés en zéro, ou les modèles dit de "barrière" utilisant des distributions tronquées. Une autre approche utilisant un modèle logit cumulatif est également testée pour son interprétation simple. Enfin, les mêmes données seront abordées sous l'angle de l'analyse de survie via un modèle à fragilités partagées, permettant entre autre de prendre en compte de la dépendance des événements entre eux.

L'exploration à travers ces multiples méthodes statistiques permet d'obtenir des résultats divers et complémentaires, et contribue à améliorer les connaissances cliniques concernant l'événement étudié.

Application des cartes de Kohonen aux bases de données nutritionnelles

Sonia Fortin, Pascale Rondeau, Sébastien Marque

Partenariat Danone : Biostatistiques 4 : Tests multiples - Risques alimentaires

Ces dernières années, les modes de consommation alimentaires ont considérablement évolué, aussi bien au sein des sociétés occidentales qu'au sein des pays en développement.

Propre à chacun, leur impact direct en Santé est indéniable. La compréhension des pratiques de consommation est donc naturellement au coeur des problématiques nutritionnelles actuelles. L'utilité d'identification de groupes de consommateurs n'est ainsi plus à démontrer ; mais aujourd'hui aucune classification ne fait consensus. C'est pourquoi il convient d'en proposer de nouvelles. S'affranchissant de tout a priori, l'approche non-supervisée s'impose à l'approche supervisée de façon évidente. D'autre part, si les méthodes hiérarchiques peuvent être envisagées, les méthodes non-hiérarchiques présentent l'avantage de donner des groupes distincts les uns des autres sans intégrer la notion de formation séquentielle (pas à pas) de ces groupes et laisser supposer que certains soient seulement le résultat de scissions de "sur-groupes". Au contraire, les consommateurs peuvent être considérés simultanément. La méthode des cartes de Kohonen, particulièrement adaptée pour les problématiques de grandes dimensions comme les problématiques nutritionnelles, réunit ces deux conditions : elle est non-supervisée et non-hiérarchique. L'algorithme repose en effet sur un apprentissage non supervisé et elle est, de plus, caractérisée par l'association de réponses voisines. On la définit aussi comme une méthode d'auto-organisation. Appliquée à des bases de données nutritionnelles accessibles, les cartes de Kohonen ont permis de partitionner les consommateurs selon leurs habitudes alimentaires et de fait d'identifier des profils de consommateurs. La classification de Kohonen : un outil précieux dans la compréhension des problématiques nutritionnelles.

Extraction de systèmes de consommation alimentaire utilisant la Nonnegative Matrix Factorization (NMF) pour l'évaluation des choix alimentaires

Mélanie Zetlaoui, Stéphan Clémençon, Max Feinberg, Philippe Verger

Partenariat Danone : Biostatistiques 4 : Tests multiples - Risques alimentaires

Dans les pays occidentaux où l'approvisionnement alimentaire est satisfaisant, les consommateurs agencent leur régime au moyen d'un grand nombre d'aliments. L'objectif de ce travail est d'étudier comment une technique récente en analyse de variables latentes, la "Nonnegative Matrix Factorization"(NMF), peut être appliquée aux données de consommation pour comprendre cet agencement. De telles données sont positives par nature et de grande dimension. Le modèle statistique NMF ici construit fournit alors une représentation des données par des variables latentes positives, appelées systèmes de consommation, qui sont en nombre très petit. L'approche NMF favorisant la sparsité, les systèmes de consommations obtenus sont de plus facilement interprétables. En application, des résultats numériques à partir d'une enquête française de consommation, sont donnés. Une méthode de clustering, basée sur la méthode des k-means dans le sous-espace des systèmes de consommation, permet de construire des groupes de consommateurs facilement interprétables par les nutritionnistes.

Les modèles de Markov cachés à effets mixtes

Maud Delattre

Biostatistique 4 : modèles à effets mixtes

Les modèles de Markov cachés à effets mixtes sont des modèles très récents. Ils se définissent comme l'extension des modèles de Markov cachés classiques aux études de population. Dans les cas où l'on peut dissocier plusieurs états dans une maladie, ces nouveaux modèles sont particulièrement adaptés à l'analyse de données longitudinales recueillies lors d'essais cliniques. Toutefois, l'estimation pour ces modèles est une problématique complexe. En particulier, les modèles de Markov cachés à effets mixtes présentent une structure hautement non linéaire et certaines données ne sont pas observées, ce qui complique grandement l'expression de la vraisemblance et sa maximisation. Nous proposons une méthodologie complète d'apprentissage de ces modèles en trois étapes. Un algorithme EM stochastique combiné à l'algorithme de Baum-Welch permettra d'abord d'estimer les paramètres de population de nos modèles. Ensuite, nous estimons les paramètres des modèles de Markov cachés individuels par maximisation de leur distribution a posteriori. Enfin, les séquences d'états les plus probables au vu des données sont obtenues par l'algorithme de Viterbi. Des études de Monte Carlo ainsi qu'une première application à des données d'épilepsie renvoient des résultats encourageants du point de vue de la qualité des estimateurs obtenus.

Evaluation et optimisation des protocoles de prélèvements dans les études pharmacocinétiques en crossover analysées par des modèles non linéaires à effets mixtes

Thu Thuy Nguyen, Caroline Bazzoli, France Mentré

Biostatistique 4 : modèles à effets mixtes

Les modèles non linéaires à effets mixtes peuvent être utilisés pour analyser les essais pharmacocinétiques de bioéquivalence ou d'interaction en crossover. En amont de la modélisation se pose la question du recueil des données. Il s'agît de trouver une balance entre le nombre de sujets nécessaires et le nombre de mesures par sujet ainsi que de définir des temps de prélèvement appropriés. En effet, le choix du protocole a un impact important sur les résultats des études comme sur la précision d'estimation des paramètres et sur la puissance des tests. Pour évaluer et optimiser les protocoles dans les essais en crossover, nous proposons une extension du calcul de la matrice d'information de Fisher dans les modèles non linéaires mixtes prenant en compte la variabilité intra sujet en plus de la variabilité inter sujet par linéarisation du modèle au premier ordre autour de la moyenne des effets aléatoires. Nous y incluons également les effets des covariables discrètes changeant entre les périodes. Nous utilisons les erreurs standards prédites pour calculer la puissance du test Wald de comparaison ou d'équivalence et le nombre de sujet nécessaires pour une puissance donnée. Ces extensions sont évaluées par simulations et implémentées dans PFIM 3.2.

Incorporation de fonction de coûts pour l'optimisation de protocoles dans les modèles non linéaires à effets mixtes : application à la pharmacocinétique de la zidovudine et de son métabolite actif

Caroline Bazzoli, Emanuelle Comets, Sylvie Retout, France Mentré

Biostatistique 4 : modèles à effets mixtes

Des études pharmacocinétiques sont souvent réalisées dans le développement et lors de l'utilisation clinique des médicaments. Les essais réalisés chez les patients requièrent souvent l'utilisation des modèles non linéaires à effets mixtes. Ces modèles peuvent être plus complexes lorsqu'ils décrivent des observations provenant de différentes réponses biologiques telle que la pharmacocinétique d'une molécule et de son métabolite. On parle alors de modèles à réponses multiples. Cependant, en amont de l'étape d'estimation, la détermination du protocole pour le recueil des données est une tache importante et complexe. La théorie pour l'évaluation et l'optimisation de tels protocoles est basée sur la matrice d'information de Fisher et a été récemment étendue aux modèles à réponses multiples. L'algorithme de Fedorov-Wynn est un outil approprié particulièrement dans le cadre de l'optimisation de protocoles statistiques. Usuellement, l'optimisation de protocoles est réalisée pour un nombre total de prélèvements et ne permet pas de prendre en considération la pénibilité des temps ou des types de prélèvements dans la composition du protocole. Nous avons ainsi introduit des fonctions de coûts dans une extension de l'algorithme de Fedorov-Wynn permettant l'optimisation de protocoles pour plusieurs réponses et pour un coût total fixé. Nous avons appliqué ce développement au premier modèle pharmacocinétique conjoint de la zidovudine et de son métabolite actif intracellulaire, coûteux à mesurer. Ce travail permet de montrer que l'algorithme de Fedorov-Wynn couplé aux fonctions de coûts est un outil puissant pour déterminer des protocoles optimaux en fonctions de différentes contraintes cliniques.

Amélioration des propriétés de mesure d'un questionnaire de satisfaction des patients hospitalisés : application d'un modèle de mesure à variable latente centrale

Sophie Tricaud-Vialle, Alain Morineau

Biostatistique 4 : modèles à effets mixtes

Les modèles d'équations structurelles peuvent être utilisés pour valider questionnaires et indicateurs de satisfaction. Les modèles de mesure de la satisfaction ont très souvent la particularité de présenter, parmi les variables latentes, une variable privilégiée qui constitue le centre du modèle : cette variable latente, la satisfaction, est la variable endogène d'une équation dont toutes les autres latentes sont les variables exogènes. Ces autres variables latentes sont les composantes de la satisfaction : elles sont liées directement à la variable centrale tout en présentant d'autres liaisons "causales" entre elles. Nous présentons ici l'application d'un tel modèle à la mesure de la satisfaction des patients dans les établissements de santé français. La particularité de variable latente centrale du modèle est prise en compte dans le mode de résolution du modèle : on assure la cohérence, autour de la variable centrale, entre les estimations du modèle interne et celles du modèle externe ; en particulier il faut s'assurer que toute modification d'une composante de la satisfaction va se répercuter sur les scores individuels de la variable centrale. Nous présenterons l'algorithme utilisé pour l'estimation des équations structurelles, une adaptation de l'approche PLS. Enfin, nous utilisons une procédure spécifique pour l'estimation des coefficients d'impact, en utilisant l'équation liant la satisfaction à ses composantes : la procédure BIS (Best Impact Solution) a pour résultat une équation dont les coefficients permettent d'évaluer directement l'effet de la modification d'une variable exogène sur la variable endogène, en laissant les autres variables exogènes évoluer en fonction de la modification faite sur cette variable endogène.

Méthode de Lissage Bayésienne Tempérée Pour Estimer les Paramètres d'un Modèle d'Equation Différentielle

David Campbell, Russell Steele

Statistique fonctionnelle 1

L'utilisation répandue des modèles d'équations différentielles ordinaires (EDO) a depuis longtemps été sous-représentée dans la littérature statistique. Les méthodes les plus communes pour estimer les paramètres des modèles d'EDO sont les moindres carrés non-linéaires et une méthode basée sur les MCMC. Ces méthodes dépendent d'une vraisemblance basée sur la solution numérique de l'EDO. Le défi relevé par ces méthodes est que les espaces de paramètres sont difficiles à naviguer, aggravé par la grande variaité de formes fonctionnelles qu'un modèle d'EDO peut produire avec des petits changements de valeurs des paramètres. Ce travail décrit la méthode de lissage bayésienne tempérée (LBT). Ce méthode emploient une expansion de bases pour approximer la solution d'EDO dans la vraisemblance, où la forme de l'expansion est guidée par le modèle d'EDO. Cette approximation de l'EDO lisse la surface de vraisemblance, réduisant ainsi les restrictions de mouvement des paramètres. La méthode de LBT utilise une suite de densités postérieures basée sur des approximations lisses à la solution d'EDO. Le niveau de l'approximation est déterminé par la valeur du paramètre de lissage qui contrôle le niveau de rugosité dans la surface de vraisemblance. Dans un algorithme semblable au tempérant parallèle, des chaînes MCMC parallèles sont utilisées pour échantillonner de la suite de densités postérieures, tout en permettant aux paramètres d'EDO de permuter entre les chaînes. Cette méthode est présentée et examinée contre une variété de modèles alternatifs.

Découpage de courbes de densité : Application au dépistage du cancer

Fabrice Morlais, Frédéric Ferraty, Philippe Vieu

Statistique fonctionnelle 1

Le dépistage actuel du cancer broncho-pulmonaire est effectué à l'aide d'une radiographie pulmonaire, d'un scanner thoracique et d'un examen cytologique des expectorations. La cytologie automatisée des expectorations est une méthode permettant l'analyse informatique des cellules d'un crachat sur la lame d'un microscope. Comme une personne est représentée par l'ensemble des cellules de sa lame, il nous a paru intéressant d'utiliser la densité de probabilité comme unité statistique. La modélisation fonctionnelle des données, méthode pour laquelle l'unité statistique est à valeurs dans un espace infini, répond bien à cette problématique statistique puisque, par définition, une densité de probabilité est une fonction. Lors de cet exposé nous présenterons la méthode de classification supervisée de courbes de densité que nous avons développée, pour discriminer des personnes ayant un cancer et des personnes saines, et nous vous donnerons quelques résultats issus de données réelles.

Prédiction en régression linéaire fonctionnelle avec variable d'intérêt fonctionnelle

Christophe Crambes, André Mas

Statistique fonctionnelle 1

Ce travail concerne l'étude de la prédiction dans le modèle linéaire fonctionnel lorsque la variable d'intérêt est elle aussi fonctionnelle. Nous introduisons un prédicteur basé sur

les décompositions de Karhunen-Loève des courbes X (variable explicative) et Y (variable d'intérêt). Les résultats obtenus permettent de fournir un développement asymptotique de la moyenne quadratique de l'erreur de prédiction. Nous donnons également un résultat d'optimalité pour ces vitesses dans un sens minimax, ainsi qu'un théorème de la limite centrale du prédicteur.

Semiparametric models with functional responses in a survey sampling setting : model assisted estimation of electricity consumption curve

Herve Cardot, Etienne Josserand

Statistique fonctionnelle 1

Ce travail adopte une approche de type sondage quand le but est d'estimer une courbe moyenne d'une grande base de données de données fonctionnelles. Lorsque les capacités de stockage sont limitées, grâce aux techniques de sondage, une petite partie des observations est une alternative intéressante par rapport aux techniques de compression. Nous proposons ici de prendre en considération une information auxiliaire réelle ou multivariée obtenu à moindre coût sur la population toute entière, avec une approche semi-paramétrique de type modèle assisté, dans le but d'améliorer les estimateurs d'Horvitz-Thompson de la courbe moyenne. D'abord, nous estimerons les composantes principales afin de réduire la dimension des signaux, et ensuite nous utiliserons des modèles semi-paramétriques pour estimer les courbes qui n'ont pas été observées. Cette technique se montre vraiment efficace sur une base de données réelle de 18902 courbes de consommation électrique mesurée toutes les demi heures pendant deux semaines.

Changements climatiques passés reconstruits à partir du pollen : Vers une modélisation statistique basée sur les mécanismes

Vincent Garreta

Environnement - Changement climatique 1

La reconstruction des changements climatiques survenus dans les derniers milliers d'années est cruciale pour comprendre la dynamique du climat dans des conditions différentes de celles enregistrées de nos jours (e.g insolation, concentration en CO_2). De telles reconstructions peuvent être obtenues en modélisant les relations entre le climat et les pollens retrouvés dans les sédiments lacustres. Ces modèles, appelés Fonction de Transfert (FT), sont des modèles statistiques purement descriptifs qui, malgré la diversité de leur forme, se basent tous sur un même groupe d'hypothèses. Cela réduit la confiance que l'on peut avoir dans des reconstructions uniquement obtenues à l'aide de ces FT. Au contraire, des FT basées sur les mécanismes liant l'environnement à la végétation et au pollen fourniraient des reconstructions basées sur des hypothèses différentes et soutenues par les recherches récentes en écologie. Pour obtenir ce type de FT nous proposons de coupler un modèle mécaniste de végétation et un modèle bayésien hiérarchique. Le cadre bayésien est naturellement approprié pour l'inférence d'un modèle dont la vraisemblance est implicitement définie par un simulateur mécaniste. Cependant, ici, elle est compliquée par les dimensions spatio-temporelles considérées. Nous montrons comment le problème peut être abordé en combinant des algorithmes de Monte Carlo pour un cas réel de reconstruction de l'Holocène en Suède. Ce type de modèle forme la prochaine génération de FT mais nécessite de poursuivre le développement d'outils statistiques pour l'inférence de modèles composites dans

la lignée de ce qui est proposé en Calcul Bayésien Approché (ABC) et émulation de modèle.

Processus max-stables pour extrêmes climatiques. Application aux hauteurs de neige extrêmes en Suisse

Juliette Blanchet

Environnement - Changement climatique 1

Une meilleure connaissance des événements extrêmes est un point clé dans la gestion des risques naturels. Or la grande majorité des processus naturels, notamment climatiques (température, pluie, vent...), sont des processus spatiaux. Un cadre naturel à la modélisation des extrêmes spatiaux est offert par les processus max-stables qui peuvent être vus comme une extension en dimension infinie des extrêmes multivariés, généralisant les structures de dépendance extrêmes aux espaces continus. Néanmoins, contrairement au cas univarié avec la distribution GEV, il n'existe pas de modèle fermé pour les processus max-stables. Différentes représentations ont été proposées dans la littérature notamment par Schlather (2002) et Smith (1991, non publié), et récemment appliquées à des données climatiques (température et précipitation). Malheureusement, aucune de ces applications ne tient explicitement compte du fait que les données traitées sont des données climatiques. Utilisées comme tels, les représentations de Schlather et Smith peuvent se révèler trop simplistes au regard de la complexité des processus physiques et climatiques sous-jacents. Cette présentation vise à étendre les processus max-stables de Schlather et Smith pour le cas particulier des extrêmes climatiques. Le modèle est basé sur une transformation climatique de l'espace euclidien, laquelle permet entre autre la prise en compte des effets directionnels et régionaux induits par les processus météorologiques en jeu. La vraisemblance complète n'étant pas disponible pour les processus max-stables, une vraisemblance composite est utilisée pour l'estimation. La méthodologie est illustrée et validée sur l'analyse de données de hauteurs de neige maximales en Suisse.

Références

Schlather, M. (2002). Models for stationary max-stable random fields. Extremes, 5(1) :33-44.

Smith, R. L. (1991). Max-stable processes and spatial extremes. Unpublished.

Analyse bayésienne hierarchique de données d'avalanches et de bilans de masse glaciaires pour l'obtention de proxys climatiques en haute montagne

Nicolas Eckert, Emmanuel Thibert

Environnement - Changement climatique 1

En haute montagne, la collecte de données climatiques directes est difficile, particulièrement en saison hivernale. Les bilans de masses glaciaires sont depuis longtemps utilisés comme des indicateurs robustes des changements de précipitations, température et rayonnement. L'activité avalancheuse, conditionnée par la quantité et la qualité (humidité, taille de grains... etc.) de la neige disponible peut également être utilisée à cet effet. Outre

leur intérêt climatique, l'étude des fluctuations temporelles de ces phénomènes présente un intérêt en matière de prévention des risques associés, avalanches exceptionnelles par exemple. Ce travail analyse différents jeux de données d'avalanches et de bilan de masse glaciaires recueillies au cours des 60 dernières années dans les Alpes Françaises. Pour chacun d'eux, un modèle d'observation adapté à la structure des données est proposé. Ensuite, la composante temporelle annuelle est isolée puis analysée à l'aide de modèles de séries temporelles explicites permettant de distinguer des évolutions systématiques en moyenne et variance de la variabilité interannuelle. Modélisation et inférence sont effectuées dans le cadre bayésien hiérarchique. Les différentes séries d'indicateurs obtenus, globalement cohérents, corroborent l'idée d'une rupture climatique au début des années 80. Celle-ci se traduit par des bilans de masse glaciaires fortement négatifs au cours des 25 dernières années, et contraint les avalanches à descendre à des altitudes moins basses que par le passé. Par contre, la fréquence des avalanches ne semble pas être affectée, les fluctuations semblant plutôt de nature cycliques que reliées à une évolution systématique du climat.

Gaussian Faithful Markov Trees

Dhafer Malouche, Bala Rajaratnam

Graphes - Modèles graphiques

Dans ce travail on s'intéresse à deux types de modèles graphiques : les modèles de concentration et les modèles de covariance. Un modèle graphique peut être défini comme un graphe associé à la distribution de probabilité d'un vecteur aléatoire. Chaque sommet dans ce graphe correspond à une variable du vecteur aléatoire. Dans un modèle de concentration l'absence d'une arête liant deux variables indique que celles-ci sont indépendantes sachant toutes les autres variables. Par contre, dans le modèle de covariance, l'absence d'une arête indique une indépendance marginale entre ces deux variables. Ces deux types de modèles permettent aussi de lire beaucoup d'autres relations d'indépendances conditionnelles entre les variables. Mais certaines d'entre elles peuvent être omises par le graphe. Dans le cas où le graphe permet de lire toutes les indépendances conditionnelles existantes dans la distribution de probabilité, on dira que cette distribution de probabilité est fidèle à son graphe. On présente ici deux résultats théoriques concernant les modèles graphiques représentés par un arbre qui est un graphe où chaque paire de sommets est liée exactement par une seule trajectoire. Il a été montré dans les deux cas et dans le cadre des modèles graphiques gaussiens que l'hypothèse de fidélité est nécessairement satisfaite. Ainsi, si le graphe de concentration associé à une loi gaussienne est arbre, alors ce graphe va permettre de lire toutes les indépendances conditionnelles existantes dans la loi probabilité (voir Becker et al., 2005) et le même résultat a été aussi observé dans le cas de modèles de covariances (voir Malouche & Rajaratnam, 2009). Cependant, les méthodes utilisées pour la démonstration des deux résultats sont complètements différentes.

Références

Becker, A., Geiger, D., Meek, C., 2005. Perfect tree-like markovian distributions. Probability and Mathematical Statistics 25 (2), 231-239.

Malouche, D., Rajaratnam, B., 2009. Gaussian Covariance Faithful Markov Trees, Technical report, Department of Statistic, Stanford University.

Modèles de graphe aléatoire à classes chevauchantes pour l'analyse des réseaux

Pierre Latouche, Etienne Birmelé, Christophe Ambroise

Graphes - Modèles graphiques

Les réseaux sont largement utilisés dans de nombreux domaines scientifiques afin de représenter les intéractions entre objets d'intérêt. Ainsi, en biologie, les réseaux de régulation s'appliquent à décrire les mécanismes de régulation des gènes, à partir de facteurs de transcription, tandis que les réseaux métaboliques permettent de représenter des voies de réactions biochimiques. En sciences sociales, ils sont couramment utilisés pour réprésenter les intéractions entre individus. Dans ce contexte, de nombreuses méthodes non-supervisées de clustering ont été développées afin d'extraire des informations, à partir de la topologie des réseaux. La plupart d'entre elles partitionne les noeuds dans des classes disjointes, en fonction de leurs profils de connection. Récemment, des études ont mis en évidence les limites de ces techniques. En effet, elles ont montré qu'un grand nombre de réseaux contenaient des noeuds connus pour appartenir à plusieurs groupes simultanément. Pour répondre à ce problème, nous proposons l'Overlapping Stochastic Block Model (OSBM). Cette approche autorise les noeuds à appartenir à plus d'une classe et généralise le très connu Stochastic Block Model (SBM), sous certaines hypothèses. Nous montrons que le modèle est identifiable dans des classes d'équivalence et nous proposons un algorithme d'inférence basé sur des techniques variationnelles globales et locales. Finalement, en utilisant des données simulées et réelles, nous comparons nos travaux avec d'autres approches.

Accuracy of Variational Estimates for Random Graph Mixture Models

Steven Gazal, Jean-Jacques Daudin, Stéphane Robin

Graphes - Modèles graphiques

L'analyse des réseaux exerce depuis quelques années un attrait croissant. Les données qui sont sous la forme de mesures de relations entre items sont de plus en plus disponibles, et abandonnent la structure usuelle d'un jeu de données de type individus-variables pour une structure de type individus-individus. Ces données "relationnelles" sont très souvent présentées sous la forme d'un graphe, même si cette représentation a ses limites, notamment quand le nombre d'individus dépasse la centaine. La représentation graphique des données des réseaux est alors attractive, mais nécessite un modèle synthétique. Le modèle de graphe le plus ancien et le plus utilisé est le modèle de Erdös-Rényi, dont les propriétés moyennes ou asymptotiques sont connues. L'écriture littérale de la vraisemblance de ce modèle est très simple, mais son temps de calcul croit de façon exponentielle avec le nombre d'individu. Une utilisation des algorithmes d'estimation usuels comme E-M n'est pas envisageable. Une approche variationnelle a été utilisée comme alternative pour implémenter un algorithme d'estimation des paramètres du modèle, et cela pour des réseaux de très grande taille (Daudin & al 2008). Les propriétés statistiques des estimateurs produits par cette approche sont cependant mal connues. L'objectif est de mener une étude sur la qualité de ces estimateurs et d'en prouver la convergence.

Références

Daudin, J.-J., Picard, F., Robin, S. (2008) A mixture model for random graphs. Stat Comput, 18, 173-183.

Consistance des estimateurs variationnels pour un modèle de graphe aléatoire

Alain Celisse, Jean-Jacques Daudin

Graphes - Modèles graphiques

Les modèles statistiques pour les graphes aléatoires hétérogènes sont utilisés dans beaucoup de domaines : réseaux sociaux, réseaux écologiques, réseaux biologiques. Les modèles de mélanges sur les noeuds servent à identifier les sous-groupes de noeuds ayant une connectivité similaire, ce qui permet d'analyser la topologie du graphe. La vraisemblance n'est pas calculable et l'algorithme EM doit être remplacé par une approximation variationnelle dont les propriétés n'étaient pas établies jusqu'à présent. Dans cet exposé nous démontrons la consistance des estimateurs variationnels.

Convergence de la constante de Cheeger de graphes de voisinage

Ery Arias-Castro, Bruno Pelletier, Pierre Pudlo

Graphes - Modèles graphiques

Nous nous intéressons dans ce travail aux ensembles minimisant la constante de Cheeger d'un sous-ensemble $\mathcal{M}$ de $\mathbb{R}^d$. Cette dernière minimise le rapport d'un périmètre à un volume parmi tous les sous-ensembles de $\mathcal{M}$. Étant donné un n-échantillon issu de la mesure uniforme sur $\mathcal{M}$, nous introduisons une version régularisée de la conductance du graphe de voisinage construit sur l'échantillon. Nous établissons alors la convergence de la conductance régularisée vers la constante de Cheeger de $\mathcal{M}$. En outre, nous montrons la convergence des suites de partitions optimales du graphe vers les ensembles de Cheeger de $\mathcal{M}$ pour la topologie de $L^1(\mathcal{M})$.

Utilisation de tests de structure en régression sur variable fonctionnelle

Laurent Delsol, Frédéric Ferraty, Philippe Vieu

Statistique fonctionnelle 2

Ce travail s'intéresse à la construction et à l'utilisation de tests de structure en régression sur variable fonctionnelle. Nous proposons, de manière générale, de construire notre statistique de test à partir d'un estimateur spécifique au modèle particulier dont nous voulons tester la validité et de méthodes d'estimation à noyau fonctionnel. Un résultat théorique montre, sous des hypothèses générales, la normalité asymptotique de notre statistique de test sous l'hypothèse nulle (c'est à dire lorsque l'hypothèse sur la structure du modèle est valide) et sa divergence sous des alternatives locales. Ce résultat permet

d'envisager la construction de tests de structure de nature très variée permettant par exemple de tester si la variable explicative n'a pas d'effet, si cet effet est linéaire, ou bien si l'effet de la variable explicative fonctionnelle se résume par l'effet de quelques caractéristiques réelles associées à celle-ci. Différentes méthodes de rééchantillonnage sont proposées pour calculer la valeur seuil du test. La méthode la plus adaptée (au vu de simulations) est ensuite utilisée dans le cadre de l'étude de données spectrométriques. L'utilisation de différents tests construits à partir de l'approche que nous proposons permet d'apporter des éléments de réponses à des questions concrètes liées à ces données. Nous discutons finalement les points qui peuvent être améliorés et présentons brièvement des perspectives intéressantes qu'offre l'utilisation de tests de structure dans le cadre de procédures s'intéressant à l'extraction de caractéristiques importantes pour la prédiction au sein de la courbe explicative mais aussi au choix de la semi-métrique.

A Functional Regression Approach for Prediction in a District-Heating System

Aldo Goia

Statistique fonctionnelle 2

Nous considérons le problème de la prédiction à court terme de pics de demande dans un système de chauffage urbain. Notre dataset consiste en quatre périodes séparées, avec 198 jours pour chaque période et 24 observations horaires dans chaque jour relatifs à la consommation de chaleur et le climat. Nous tenons en considération la nature fonctionnelle des données et proposons une méthodologie de prédiction basée sur la régression fonctionnelle. L'influence de variables explicatives exogènes est modelée d'une façon appropriée. Le résultats "out-of-sample" de l'approche proposée sont évalués.

Functional common principal components models

Graciela Boente, Daniela Rodriguez, Mariela Sued

Statistique fonctionnelle 2

Dans cet exposé, nous discutons l'extension au cas fonctionnel du modèle de composantes principales communes, qui a été largement étudié lorsqu'on s'intéresse à des observations multivariées. Nous proposons des estimateurs pour les composantes principales communes et nous étudions leur distribution asymptotique. Ces résultats font partie d'un travail conjointement écrit avec Graciela Boente et Daniela Rodriguez.

Sélection de modèle incluant des composantes principales

Alois Kneip, Pascal Sarda

Statistique fonctionnelle 2

Nous considérons un modèle de régression linéaire de grande dimension et plus précisément le cas d'un modèle factoriel pour lequel le vecteur des variables explicatives se décompose en la somme de deux termes aléatoires décrivant respectivement la variabilité spécifique et commune des prédicteurs. Nous montrons tout d'abord que les procédures de sélection de variables et d'estimation usuelles telles que le lasso ou le sélecteur Dantzig sont performantes dans ce contexte et sous l'hypothèse additionnelle que le vecteur des paramètres est sparse. Cette hypothèse peut être cependant restrictive. Nous introduisons

ainsi un modèle de régression augmenté qui inclut les composantes principales. Nous montrons que ces composantes peuvent être convenablement estimées à partir de l'échantillon et nous nous concentrons ensuite sur les propriétés théoriques du modèle augmenté.

Courbes Principales et Sélection de Modèle

Aurélie Fischer

Apprentissage 3 - Sélection de modèle

Les courbes principales sont une généralisation non linéaire de la notion de première composante principale. Intuitivement, une courbe principale est une courbe de R^d passant au "milieu" d'une distribution de probabilité en dimension d ou d'un nuage de données de R^d. Plusieurs définitions ont été proposées, dont l'une repose sur la minimisation d'un critère des moindres carrés. Nous nous intéressons au choix de la classe sur laquelle minimiser ce critère pour obtenir une courbe principale qui résume au mieux la forme des données sans pour autant conduire à de l'interpolation, et adoptons le point de vue de la sélection de modèle par pénalisation.

Estimation de la fonction de répartition conditionnelle à partir de données censurées par intervalle, cas 1, par sélection de modèles.

Sandra Plancade

Apprentissage 3 - Sélection de modèle

Considérons une variable aléatoire positive Y appelée temps de survie, dépendant d'une covariable X. Dans le cadre de la censure par intervalle, cas 1 (ou "current status data"), la variable Y n'est pas directement observée. La seule information dont on dispose est la donnée du triplet (X,U,d) ou U est un temps de mesure indépendant de Y conditionnellement à X, et d est égal à 1 si Y est inférieur à U et 0 sinon.

Le but de cet exposé est de construire un estimateur adaptatif de la fonction de distribution conditionnelle de Y sachant X à partir d'un échantillon de (X,U,d), par sélection de modèles.

Après avoir construit une collection de modèles pour des fonctions de deux variables par produits tensoriels de fonctions d'une variable, on calcule dans chacun de ces modèles l'estimateur des moindres carrés. On obtient ainsi une collection d'estimateurs. Enfin, une procédure de sélection de modèles pénalisée fournit un estimateur adaptatif.

Clustering et sélection de variables sur des données génétiques

Dominique Bontemps, Wilson Toussile

Apprentissage 3 - Sélection de modèle

Nous nous intéressons au problème d'estimer les variables pertinentes et le nombre de composantes d'une loi de mélange pour des données génotypiques multilocus. Un critère du maximum de vraisemblance pénalisé est proposé, et une inégalité oracle non-asymptotique est obtenue. En outre, sous des conditions faibles portant sur la distribution qui a généré les observations, le modèle sélectionné est asymptotiquement consistent.

D'un point de vue pratique, la pénalité est définie à une constante multiplicative près, et celle-ci est calibrée par l'heuristique de pente. Sur des données simulées la procédure

de sélection fait mieux que des critères classiques tels que BIC et AIC. Le nouveau critère apporte une réponse à la question : "Quel critère choisir en fonction de la taille de l'échantillon ?".

Bornes de Risque pour CART en Classification

Servane Gey

Apprentissage 3 - Sélection de modèle

Des bornes de risque pour les arbres de classification CART sont obtenues sous une condition de marge dans le cadre de la classification supervisée binaire. Ces bornes sont prises conditionnellement à la construction de l'arbre de plus grande profondeur construit sur un échantillon d'apprentissage. Elles permettent de valider le choix de la pénalité dans l'algorithme d'élagage d'une part, et de montrer que la sélection d'un sous-arbre dans la suite de sous-arbres élagués à l'aide d'un échantillon-témoin n'altère pas trop la qualité du classificateur sélectionné d'autre part. Dans le cadre de la classification binaire, et sous une condition de marge, ces bornes de risque, obtenues par des techniques de sélection de modèles, permettent de valider l'algorithme CART.

Bornes de risque pour les forêts purement uniformément aléatoires

Robin Genuer

Apprentissage 3 - Sélection de modèle

Introduites par Leo Breiman en 2001, les forêts aléatoires sont une méthode statistique très performante. D'un point de vue théorique, leur analyse est difficile, du fait de la complexité de l'algorithme. Pour expliquer ces performances, des versions de forêts aléatoires simplifiées (et donc plus faciles à analyser) ont été introduites : les forêts purement aléatoires. Dans cet article, nous introduisons une autre version simplifiée, que nous appelons forêts purement uniformément aléatoires. Dans un contexte de régression avec une seule variable explicative, nous montrons que les arbres aléatoires ainsi que les forêts aléatoires atteignent la vitesse de convergence minimax. Et plus important, nous prouvons que les forêts aléatoires améliorent les performances des arbres aléatoires, en réduisant la variance des estimateurs associés d'un facteur trois quarts.

Estimation de courbes de niveaux extrêmes pour des lois à queues lourdes

Abdelaati Daouia, Laurent Gardes, Stéphane Girard, Alexandre Lekina

Extrêmes

Le problème d'estimation des courbes de niveaux extrêmes est équivalent à l'étude des quantiles conditionnels quand l'ordre du quantile tend vers un. Nous montrons que sous certaines conditions, il est possible d'estimer de telles courbes au moyen d'un estimateur à noyau de la fonction de survie conditionnelle. En conséquence, ce résultat nous permet d'introduire deux versions lisses de l'estimateur de l'indice de queue conditionnel indispensable lorsque l'on veut extrapoler. Nous établissons la loi limite des estimateurs ainsi construits. Pour conclure, une illustration sur données simulées est présentée.

Une méthode de folding pour des extrêmes multivariés avec application à l'estimation de la mesure de probabilité spectrale

Armelle Guillou, Philippe Naveau, Alexandre You

Extrêmes

Nous commencerons notre exposé par introduire une nouvelle approche d'estimation dans le contexte de la théorie des valeurs extrêmes, appelée "folding". Celle-ci a ete developpee par You et al. (2010) dans un cadre univarie. Elle consiste en une " transformation des données initiales" et a été initialement introduite dans un contexte de "perfect sampling"(Corcoran et Schneider, 2003). Cette nouvelle approche a permis d'améliorer de façon significative l'estimation de quantiles extrêmes en adaptant la méthode pic au delà d'un seuil (POT). Notre objectif dans cet exposé sera donc double : dans un premier temps, nous étendrons le concept de folding dans le cas multivarié ; dans un second temps, nous appliquerons cette extension du folding à un problème d'estimation classique en théorie des extrêmes multivariés, à savoir l'estimation de la mesure de probabilité spectrale S. Nous illustrerons, par le biais de simulations et sur des données réelles nos résultats.

Références

Corcoran, J.N., Schneider, U. (2003). Shift and scale coupling methods for perfect simulation. Probability in the Engineering and Informational Sciences, 17, 277-303.
You, A., Schneider, U., Guillou, A., Naveau, P. (2010). Improving extreme quantile estimation via a folding procedure. A paraître dans Journal of Statistical Planning and Inference.

Estimation de quantiles extrêmes et de probabilités d'événements rares

Arnaud Guyader, Nicolas Hengartner, Eric Matzner-Lober

Extrêmes

Etant donnée une probabilité μ sur $\mathbb{R}^d$ (d grand), on note X un vecteur aléatoire générique de loi μ et $\Phi : \mathbb{R}^d \to \mathbb{R}$ une application "boîte noire". Un réel q étant fixé, le but est de générer un échantillon i.i.d. $(X_1, \ldots, X_N)$ tel que pour tout i : $X_i \sim \mathcal{L}(X|\Phi(X) > q)$. Lorsque q est grand comparé aux valeurs typiques de la variable $\Phi(X)$, la méthode Monte-Carlo classique devient trop coûteuse. Dans ce travail nous présentons et analysons une nouvelle approche pour ce problème. Celle-ci procède en plusieurs étapes, s'inspirant de l'algorithme de Metropolis-Hastings et des méthodes dites multi-niveaux en estimation d'événements rares. Deux problèmes peuvent être traités très facilement via cette nouvelle méthode : estimation de quantiles extrêmes et estimation d'événements rares. Les idées présentées seront illustrées sur un problème de tatouage numérique.

Estimation de queues bivariées

Elena Di Bernardino, Véronique Maume-Deschamps, Clémentine Prieur

Extrêmes

Ce papier traite de l'estimation de la queue d'une distribution bivariée. Nous développons une approche bidimensionnelle de la méthode de dépassement de seuil (Peaks Over Threshold) et une version bivariée du Théorème de Pickands-Balkema-de Hann. Nous démontrons des propriétés de convergence pour l'estimateur ainsi construit. La structure de

dépendance entre les marges est modélisée par une copule. Enfin, nous proposons quelques simulations.

Test de comparaison du paramètre de longue mémoire

Frédéric Lavancier, Anne Philippe, Donatas Surgailis

Séries temporelles 2 : mémoire longue

Nous proposons un test pour comparer le paramètre de longue mémoire de deux processus éventuellement corrélés. Le test est construit à partir de la statistique V/S basée sur deux estimations de la variance asymptotique des sommes partielles. Nous établissons la consistance asymptotique du test. Des simulations illustrent les performances du test sur des petits échantillons et sa sensibilité au paramètre de type fenêtre de la statistique V/S. A partir d'un développement asymptotique de la statistique V/S, nous obtenons un critère adaptatif pour le choix de ce paramètre.

Time varying fractionally integrated model

Adnen Ben Nasr, Mohamed Boutahar

Séries temporelles 2 : mémoire longue

Nous proposons un nouveau modèle afin de modéliser les séries temporelles avec de la persistance variable dans le temps. Ce modèle permet au paramètre de mémoire longue d'avoir un ou plusieurs changements structurels continus. Nous développons un test de Multiplicateur de Lagrange (LM) afin de tester le changement structurel, et une stratégie permettant de déterminer le nombre de changements dans le paramètre de mémoire longue. Des études de simulation montrent que le test a de bonnes propriétés au niveau de la taille et la puissance; elles montrent aussi que la stratégie proposée identifiant le nombre de régimes est très satisfaisante. Une application empirique sur les valeurs absolues des rentabilités du CAC40 indique que de telles données sont caractérisées par une instabilité dans leur persistance expliquée par un changement structurel dans le paramètre de mémoire longue.

Structural change and long memory in the dynamic of U.S. inflation process

Mustapha Belkhouja, Mohamed Boutahar

Séries temporelles 2 : mémoire longue

Long range dependence and regime switching are very intimately related effects. In this paper we consider the problem of spuriously detecting break dates in hypothesis of long memory data generating processes. For this purpose, we address the issue of estimating the number of breaks using several techniques, namely, the information criteria, Bai and Perron's sequential selection procedure (1998), and the automatic procedure of Lavielle (2004). By means of Monte Carlo experiments, we investigate the effect of increasing the long memory parameter on selecting the number of breaks and their locations, and show that the Lavielle's method is the best technique since its frequency of choosing the true number of changes is the highest particularly when the order of integration is close to 0.5.

As it seems that inflation rates contains long memory and structural breaks, an application to the U.S. inflation process is presented to illustrate the usefulness of these procedures. The results show that the Lavielle's method (2004) selects only two breaks, however, the number of breaks detected by the information criteria and the sequential procedure of Bai and Perron (1998) are superior or equal to three.

Références

Bai, J. and P. Perron (1998): "Estimating and testing linear models with multiple structural changes". Econo- metrica, 66: 47-78.
Lavielle, M. (2004): "Using penalized contrasts for the change-point problem". Elsevier.

Réalisation d'un modèle de prévision à court, moyen et long terme de l'activité d'un transporteur

Wilfried Despagne

Séries temporelles 2 : mémoire longue

L'article décrit une activité de recherche qui vise à répondre à une problématique industrielle. Un transporteur sous température dirigée cherche à optimiser la planification de ses ressources humaines et matérielles à travers la prévision à court et moyen terme de son activité. Le challenge réside dans le fait de trouver un modèle de prévision unique s'adaptant, sans intervention humaine, aux spécificités des 70 agences de messagerie du transporteur. La matière première est l'information récoltée par le transporteur depuis plus de six ans. Les outils sont des algorithmes mathématiques éprouvés, utilisés pour la prévision des séries temporelles. Le travail vise à combiner ces outils pour qu'ils extraient le maximum d'information déterministe capable d'être anticipée. L'article pose la problématique et son contexte économique. Il poursuit par un descriptif des procédures utilisées et un argumentaire pour défendre leur choix. Les solutions informatiques adoptées sont inventoriées. Aujourd'hui, sans intervention humaine, les données réelles se mettent à jour quotidiennement, son présenté par le biais d'une interface web et de nouvelles prévisions sont recalculées toutes les semaines, tout ceci sur un serveur central.

Homogénisation de séries climatiques

Olivier Mestre, Victor Venema

Environnement - Changement climatique 2

Dans un grand nombre d'endroits, on peut constituer des séries d'observations climatologiques relativement complètes, remontant au XIXème siècle. Cependant, les conditions de mesure ont été profondément modifiées au cours du temps. Les changements d'emplacement, d'instrumentation... se traduisent par autant de biais dans les séries de données. Or ces ruptures artificielles peuvent être du même ordre de grandeur que les phénomènes climatiques que l'on étudie. Leur détection et leur correction sont donc indispensables avant toute étude climatique : c'est ce que l'on appelle l'homogénéisation des séries.

Le problème se traite en deux phases : détection des ruptures, et correction, pour lesquelles toute une variété de méthodes existent.

On présentera les résultats de l'Action Européenne COST ES0601 : intercomparaison des méthodes d'homogénéisation des séries de données, qui permet une synthèse des meilleurs aspects des différentes procédures.

Modélisation statistique des changements climatiques, détection et attribution

Aurélien Ribes

Environnement - Changement climatique 2

Dans le cadre des études de détection et d'attribution, différents outils statistiques sont utilisés afin d'étudier les changements climatiques en cours. La détection d'un changement, tout d'abord, consiste à montrer qu'un phénomène est effectivement un changement ; on montre, statistiquement, que ce phénomène est incohérent avec la seule variabilité interne du système climatique, considérée comme aléatoire. L'attribution d'un changement à une ou plusieurs causes consiste à établir un lien de causalité entre différents facteurs explicatifs physiquement plausibles et le changement étudié. Nous présentons ici un bref descriptif des modèles et de la démarche statistiques mis en oeuvre dans ce cadre, qui accordent une place importante aux tests d'hypothèses. Nous introduisons ensuite quelques unes des problématiques statistiques pouvant être rencontrées pour mener à bien ces études.

Modèle linéaire mixte avec segmentation : application à la détection de changements dans les dates de vendanges

Emilie Lebarbier, Franck Picard, Eva Budinska, Stéphane Robin

Environnement - Changement climatique 2

Nous nous intéressons à la détection de changements dans les dates de vendanges de plusieurs stations qui seraient dûs à des changements de pratiques et non à des changements climatiques. Ces séries sont analysées simultanément à l'aide d'un modèle linéaire mixte avec ruptures qui permet de prendre en compte à la fois des covariables et des corrélations entre series. Pour obtenir les paramètres du maximum de vraisemblance, nous utilisons un algorithme EM et proposons un nouvel algorithme de programmation dynamique pour l'étape de segmentation. Cependant, se pose la question du choix du nombre de segments. Ici nous généralisons trois critères de sélection de modèles, qui avaient été proposés dans le cas de la segmentation d'une serie, à la segmentation jointe de plusieurs series. Nous comparons ces critères par une étude de simulation.

Sélection bayésienne de variables pour les modèles d'état dans le cadre de reconstructions climatiques

Ophélie Guin, Philippe Naveau

Environnement - Changement climatique 2

De nombreuses variantes sur la sélection de variables pour un modèle de régression sous l'approche bayésienne ont été proposées dans la littérature. Dans cette présentation, nous adaptons cette méthode de sélection de variables à un modèle d'état, ce qui revient à ajouter une équation à notre modèle de régression. On applique cette méthode à un problème de reconstruction climatique. En effet, afin de comprendre si le réchauffement climatique actuel est plus important que la variabilité climatique naturelle, il est nécessaire de d'avoir de longues séries climatiques. Seulement, des mesures directes de températures et précipitations manquent, en particulier pour les période les plus anciennes, et il est nécessaires d'utiliser des proxis climatiques afin de reconstruire des chronologies passées. Un proxy bien connu est la croissance des cernes d'arbres. Afin de comprendre les relations existantes entre ce proxy est des variables climatiques on essaye d'expliquer la croissance des cernes d'arbres avec la meilleures combinaison possible de séries de températures et de précipitations.

Approche bayésienne des modèles à équations structurelles

Séverine Demeyer, Nicolas Fischer, Gilbert Saporta

Statistique bayésienne

Les modèles à équations structurelles (SEMs) sont des modèles multivariés à variables latentes utilisés pour modéliser les structures de causalité dans les données. Une approche bayésienne d'estimation et de validation des modèles SEMs est proposée et l'identifiabilité des paramètres est étudiée. Cette étude montre qu'une étape de réduction des variables latentes au sein de l'algorithme de Gibbs permet de garantir l'identifiabilité des paramètres. Cette heuristique permet en fait d'introduire les contraintes d'identifiabilité dans l'analyse. Pour illustrer ce point, les contraintes d'identifiabilité sont calculées dans une application en marketing, dans laquelle les distributions des contraintes sont obtenues par combinaison des tirages a posteriori des paramètres.

Bayesian variable selection for probit mixed models

Meïli Baragatti

Statistique bayésienne

In computational biology, gene expression datasets are characterized by very few individual samples compared to a large number of measurments per sample. Thus, it is appealing to merge these datasets in order to increase the number of observations and diversify the data, allowing a more reliable selection of genes relevant to the biological problem. This necessitates the introduction of the dataset as a random effect. Extending previous work of Lee et al. (2003), a method is proposed to select relevant variables among tens of thousands in a probit mixed regression model, considered as part of a larger hierarchical Bayesian model. Latent variables are used to identify subsets of selected variables and the collapsing technique of Liu (1994) is combined with a Metropolis-within-Gibbs algorithm (Robert and Casella, 2004). The method is applied to a merged dataset made of three individual gene expression datasets, in which tens of thousands of measurements are available for each of several hundred human breast cancer samples. Even for this large

dataset comprised of around 20000 predictors, the method is shown to be efficient and feasible. As a demonstration, it is used to select the most important genes that characterize the estrogen receptor status of the cancer patients.

Références

Lee, K., Sha, N., Dougherty, E., Vannucci, M., Mallick, B., 2003. Gene selection: a bayesian variable selection approach. Bioinformatics 19 (1), 90-97.

Liu, J., 1994. The collapsed gibbs sampler in bayesian computations with application to a gene regulation problem. Journal of the American Statistical Association 89 (427), 958-966.

Robert, C., Casella, G., 2004. Monte Carlo statistical methods, second edition. Springer.

Estimation de la variance généralisée

Thu Pham-Gia

Statistique bayésienne

En analyse statistique multidimensionnelle la matrice des variances-covariances Σ joue un rôle fondamental. On peut la mesurer soit par ses racines propres, soit par son déterminant qui s'appelle alors la variance généralisée. Estimer la variance généralisée d'une population normale dans R^k , $k \geq 2$, a fait l'objet de plusieurs travaux. Nous approchons le problème ici à travers la fonction de Meijer G, qui a connu plusieurs applications importantes dernièrement, en calculs numériques et en intégration. Ici, elle permet le calcul précis de l'intervalle de confiance à $(1-\alpha)100\%$ de $|\Sigma|$ à partir de la variance généralisée $|\mathrm{S}|$ d'un échantillon aléatoire. Des résultats relatifs à la distribution de $|\mathrm{S}|$ seront présentés et permet de raffiner des approches où $|\mathrm{S}|$ est présente dans des statistiques pour tester certains paramètres. Une application basée sur le rapport de 2 variances généralisées, sera donnée.

Processus Stick-Breaking et extensions pour le traitement bayésien de processus ponctuels

Florencia Chimard, Jean Vaillant

Statistique bayésienne

Beaucoup de phénomènes sont de nature stochastique et nécessitent pour être modélisés et prédits des techniques statistiques appropriées. Un processus ponctuel (Daley et Veres Jones (1988); Karr (1991) ; van Lieshout (2000)) est un mécanisme stochastique qui modélise des localisations de points dans un espace donné. Une répartition de points dans l'espace ou dans le temps peut-être considérée comme la réalisation d'un processus ponctuel (Vaillant (1991,1992)). Elle se présente sous la forme d'un ensemble de coordonnées et/ou de dates d'occurrences. On ne dispose en général que d'une réalisation de ce processus (Vaillant et al. (1997)). La loi d'un processus ponctuel est complètement caractérisée par son processus intensité conditionnelle et la modélisation se fait donc uniquement à travers ce dernier. Notre étude concerne l'apport des modèles de mélange de processus ponctuels dans de telles modélisations. Elle fait également apparaître l'utilité de l'approche bayésienne pour apprécier la variabilité associée à certains facteurs environnementaux. Le lien

avec les lois a priori Stick-Breaking (Ishwaran et James (2001)) telles que le processus de Dirichlet (Ferguson (1973,1974)) est présenté. Des modèles de mélange de processus de Poisson non homogènes ont été présentés par différents auteurs, par exemple, Kottas et Sanso (2007) ont considéré pour la loi mélangeante un processus de Dirichlet. Il est à noter cependant, que lorsque les données sont sous forme de comptages régionalisés, la modélisation concerne plutôt la loi de dénombrement (Green et Richardson (2002)). Pour notre part, nous nous intéressons à des processus Stick-Breaking à noyaux pour la prise en compte des localisations locales dans les mélanges spatiaux de processus ponctuels. Des algorithmes de type MCMC pour le traitement approprié de données incomplètes sont présentés et discutés puis sont testés sur des données artificielles.

Références

Daley, D., and Vere-Jones, D. An Introduction to the Theory of Point Processes. New York, 1988.

Ferguson, T. A bayesian analysis of some nonparametric problems. The annals of statistics 1 (1973), 209-230.

Ferguson, T. Prior distributions on spaces of probability measures. The annals of statistics 2 (1974), 615-629.

Green, P., and Richardson, S. Hidden markov models and disease mapping. American Statistical Association 97, 460 (2002).

Ishwaran, H., and James, L. F. Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association 96, 453 (2001), 161-173.

Karr, A. F. Point Processes and their statistical inference. New York, 1991.

Kottas, A., and Sanso, B. Bayesian mixture modeling for spatial poisson process intensities, with applications to extreme value analysis. Statistical Planning and Inference 137 (Mars 2007), 3151-3163.

Vaillant, J. Negative binomal distributions of individuals and spatio-temporal cox processes. Scan J. Statist 18 (1991), 235-248.

Vaillant, J. Echantillonnage et étude statistique de populations en milieu héhétrogène. Statistique Aplliquée 4 (1992), 15-26.

Bayesian Nonparametric Inference of decreasing densities

Soleiman Khazaei, Judith Rousseau

Statistique bayésienne

Abstract In this paper we discuss consistency of the posterior distribution in cases where the Kullback-Leibler condition is not verified. This condition is stated as : for all $\epsilon > 0$ the prior probability of sets in the form $f; KL(f0, f) \leq \epsilon$ where $KL(f0, f)$ denotes the Kullback-Leibler divergence between the true density $f0$ of the observations and the density f , is positive. This condi- tion is in almost cases required to lead to weak consistency of the posterior distribution, and thus to lead also to strong consistency. However it is not a necessary condition. We therefore present a new condition to replace the Kullback-Leibler condition, which is usefull in cases such as the estimation of decreasing densities. We then study some specific families of priors adapted to the estimation of decreasing densities and provide posterior concentration rate for these priors, which is the same rate a the convergence rate of the maximum likelihood estimator. Some simulation results are provided.

Blind forecasting for Gaussian time-series

Thibault Espinasse, Fabrice Gamboa, Jean-Michel Loubes

Séries temporelles 3 - Processus

Le but de cet exposé est de fournir, dans le cadre des séries chronologiques, un estimateur "aveugle" de l'opérateur de projection sur le passé infini. Il s'agit de donner un prédicteur, lorsque la structure de covariance est inconnue, et qu'un unique échantillon est disponible pour simultanément estimer la covariance et prédire le processus. Nous obtenons la vitesse de convergence en erreur quadratique, en utilisant un résultat de concentration sur la covariance empirique, et une astucieuse décomposition de Schur donnant une forme alternative de ce projecteur. La vitesse est alors obtenue en fonction de la régularité de la densité spectrale.

Estimation des modèles VARMA structurels avec innovations linéaires non corrélées mais non indépendantes

Yacouba Boubacar Mainassara, Christian Francq

Séries temporelles 3 - Processus

Pour la modélisation des séries temporelles multivariées, les modèles VARMA (Vector AutoRegressive Moving-Average) occupent une place centrale. Ils sont généralement utilisés avec des hypothèses fortes sur le bruit qui en limitent la généralité. Dans ce travail, nous nous intéressons à l'analyse statistique de modèles vectoriels ARMA (VARMA) pour des processus qui peuvent avoir des dynamiques non linéaires très générales. Nous appelons VARMA forts les modèles standard dans lesquels le terme d'erreur est supposé être une suite iid, et nous parlons de modèles VARMA faibles quand les hypothèses sur le bruit sont moins restrictives.

Dans un premier temps, nous étudions les propriétés asymptotiques du quasi-maximum de vraisemblance (QMLE) des paramètres d'un modèle VARMA sans faire l'hypothèse d'indépendance sur le bruit, contrairement à ce qui est fait habituellement pour l'inférence de ces modèles. Relâcher cette hypothèse permet aux modèles VARMA faibles de couvrir une large classe de processus non linéaires. Nous faisons des hypothèses d'ergodicité et de mélange afin d'établir la convergence forte et la normalité asymptotique de l'estimateur du QMLE.

Ensuite, nous accordons une attention particulière à l'estimation de la matrice de variance asymptotique qui a une forme "sandwich", et qui peut être très différente de la variance asymptotique standard. Nous établissons la convergence d'un estimateur de cette matrice. Enfin, des versions modifiées des tests de Wald, du multiplicateur de Lagrange et du rapport de vraisemblance sont proposées pour tester des restrictions linéaires sur les paramètres libres du modèle.

Estimation adaptative des modeles vectoriels autoregréssifs avec une variance dependant du temps

Valentin Patilea, Hamdi Raissi

Séries temporelles 3 - Processus

Nous analysons les modèles Vectoriels AutoRegressifs (VAR) quand les innovations sont non conditionnellement hétéroscédastiques. La structure de la volatilité est détermin-iste et générale, incluant des discontinuités ou des tendances comme cas particuliers. Dans ce cadre nous proposons des estimateurs des Moindres Carrés Ordinaires (MCO) et des estimateurs des Moindres Carrés Adaptatifs (MCA). L'estimateur des MCA est calculé en estimant la volatilité de façon non paramétrique. Nous obtenons la distribution asymptotique des estimateurs et comparons leur propriétés. En particulier nous montrons que l'estimateur des MCA est asymptotiquement équivalent à l'estimateur des Moindres Carrés Généralisés (MCG) obtenus en supposant que la volatilité des erreurs est connue.

Technique de rééchantillonnage et estimation de l'ordre d'un modèle ARMA avec des données incomplètes

Abdelaziz El Matouat, Hassania Hamzaoui, Freedath Djibril Moussa

Séries temporelles 3 - Processus

La sélection de l'ordre d'un modèle est souvent effectuée à l'aide des critères d'information fondés sur une pénalisation de la log-vraisemblance relative aux observations. Les critères usuels sont les critères AIC, BIC, et ϕ_β. Pour traiter le problème de données incomplètes, une reformulation des critères avec l'algorithme EM a été nécessaire, et les critères AIC_{cd}, BIC_{cd} et $\phi_{\beta_{cd}}$ correspondants ont ainsi été établis. Mais en pratique, l'estimation de l'ordre par ces critères peut ne pas se révéler satisfaisante. De plus, les pénalisations de ces différents critères ne dépendent que de la taille de l'échantillon et de la fonction qui décrit les itérations de l'algorithme EM. Nous proposons une amélioration de ces critères dans le cas des modèles ARMA, par l'introduction d'une pénalisation calculée à partir des données avec la technique de rééchantillonnage. Nous obtenons alors de nouveaux critères, dont les performances sont étudiées sur des échantillons simulés et comparées aux critères AIC_{cd}, BIC_{cd} et $\phi_{\beta_{cd}}$.

Well solved cases of probabilistic traveling salesman problem

Monia Bellalouna, Vangelis Paschos, Walid Khaznaji

Séries temporelles 3 - Processus

Le problème du voyageur de commerce est généralement NP-difficile. Notre but est de chercher des cas qui sont "faciles", càd résolvables en temps polynomial. Le Problème du Voyageur de Commerce Petit illustre bien ces cas. Notre étude comporte deux volets, d'abord nous donnons la relation entre le problème déterministe et son homologue probabiliste, ensuite nous montrons un résultat remarquable, à savoir que ce problème est stable : la perturbation par l'absence de certaines données n'influe pas sur l'optimalité de la solution.

Estimation de modèles markoviens discrets dans un cadre industriel fiabiliste à données manquantes

Alberto Pasanisi, Shuai Fu, Nicolas Bousquet

Fiabilité - Incertitudes (session du groupe)

Les modèles markoviens sont particulièrement utiles pour décrire des systèmes qui, au long de leur vie, passent à travers différents états. Les paramètres de ces modèles sont les probabilités de transition entre les états. Normalement, les données disponibles pour l'inférence statistique sont des séquences temporelles d'états pour un nombre donné d'individus. Quand les séquences sont incomplètes, l'estimation de la matrice de transition n'est pas triviale et demande l'utilisation de techniques plus avancées. Dans cette communication, nous nous focalisons sur l'estimation bayésienne des probabilités de transition. Premièrement, nous présentons différentes méthodes MCMC, en fonction de la structure des données manquantes. Ensuite, nous proposons une manière d'accélérer les calculs MCMC en tenant compte de la dépendance entre les lignes de la matrice de transition. Finalement, nous montrons les résultats d'essais simulés menés sur des matrices typiquement rencontrées dans les problèmes de fiabilité industrielle.

Stratification Directionnelle adaptative

Miguel Munoz Zuniga, Josselin Garnier, Emmanuel Remy, Etienne De Rocquigny

Fiabilité - Incertitudes (session du groupe)

L'estimation d'une probabilité de défaillance, ou autrement dit l'estimation d'une intégrale multidimensionnelle, est une problématique classique en fiabilité des structures et de nombreuses méthodes de calculs existent déjà dans la littérature. Cependant, très peu de méthodes répondent simultanément aux contraintes suivantes couramment rencontrées dans un cadre industriel : une probabilité à estimer très faible, des ressources computationnelles limitées et une estimation contrôlée. Dans cette optique, nous avons développé une nouvelle méthode de Monte-Carlo accélérée, nommée SDA : Stratification Directionnelle Adaptative. Celle-ci couple la méthode de stratification, permettant la mise en place d'une méthode adaptative à tirages préférentiels, et la simulation directionnelle, adaptée à l'estimation de faibles probabilités et possédant un bon rapport "précision/temps de calculs". Nous proposons tout d'abord un estimateur avec deux étapes, apprentissage puis estimation, SDA-2, généralisé ensuite avec un nombre quelconque d'étapes, SDA-L. Dans les deux cas, nous étudions les comportements asymptotique et non asymptotique des estimateurs. Nous proposons également une amélioration de SDA-2, pour solutionner le problème d'espace d'exploration trop vaste engendré par une dimension trop élevée. Pour cela, nous introduisons une statistique, évaluée à la fin de l'étape d'apprentissage, permettant une réduction pertinente du nombre de strates et une classification des variables aléatoires selon leurs influences sur la défaillance.

Évaluation d'un risque d'inondation fluviale par planification séquentielle d'expériences

Aurélie Arnaud, Julien Bect, Mathieu Couplet, Alberto Pasanisi, Emmanuel Vazquez

Fiabilité - Incertitudes (session du groupe)

Nous nous intéressons au risque d'inondation d'une zone habitable ou industrielle, située à proximité d'un fleuve. Le risque est évalué à partir d'un modèle de la ligne d'eau du

fleuve en présence d'incertitudes sur le débit et les caractéristiques du lit fluvial. Comme l'évaluation du modèle de la hauteur d'eau, pour un débit et des caractéristiques du lit fixés, est potentiellement coûteux en temps de calcul, l'estimation d'une probabilité de dépassement de seuil ou d'un quantile de la hauteur d'eau doit en pratique être conduite avec un budget réduit de simulations. Dans cet article, nous nous intéressons spécifiquement à l'estimation d'un quantile et nous proposons une méthode de planification d'expériences séquentielle qui construit une approximation du modèle par krigeage en choisissant les points d'évaluation du modèle de manière à réduire la variance d'estimation du quantile.

Caractérisation des coefficients de Strickler d'un fleuve par inversion probabiliste

Mathieu Couplet, Laurent Lebrusquet, Alberto Pasanisi

Fiabilité - Incertitudes (session du groupe)

La caractérisation statistique des coefficients de rugosité du lit et des berges d'un fleuve (coefficients de Strickler) permet à la fois d'évaluer les risques d'inondation et de prévoir le renforcement des ouvrages à proximité. Il s'agit dans cette étude de caractériser ces coefficients à partir de mesures de débit associées à des mesures de hauteurs d'eau. Un logiciel de calcul hydraulique donne les hauteurs d'eau à partir des coefficients de Strickler. Des algorithmes d'inversion probabiliste sont mis en oeuvre afin de modéliser la densité de probabilité des coefficients de Strickler. Deux approches sont testées : l'une est une variante de l'algorithme EM, l'autre est de type MCMC.

Polynômes de chaos sous Scilab via la librairie NISP

Michaël Baudin, Jean-Marc Martinez

Fiabilité - Incertitudes (session du groupe)

Cette note présente les polynômes de chaos, adaptés à la modélisation et propagation d'incertitudes en simulation numérique. Ce projet prend appui sur la librairie NISP (Non Intrusive Spectral Projection) développée sous licence LGPL. Le module NISP est disponible dans Scilab sous Linux et Windows et peut être téléchargé via le portail ATOMS administré par le Consortium Scilab. Ce travail a été réalisé dans le cadre du projet OPUS (Open-source Platform for Uncertainty treatment in Simulation) soutenu par l'Agence Nationale de la Recherche. Dans ce document, nous introduisons les polynômes de chaos, ainsi que leur intérêt en analyse de sensibilité comme modèles adaptés à la décomposition fonctionnelle de la variance. Puis nous présentons un tutoriel d'introduction au module NISP de Scilab.

Analyse de sensibilité d'un robinet à soupape à l'aide de développements sur chaos polynomial

Marc Berveiller, Géraud Blatman, Jean-Marc Martinez
Fiabilité - Incertitudes (session du groupe)

L'objectif de cette communication est de présenter une méthode d'analyse de sensibilité basée sur un métamodèle de type chaos polynomial creux et adaptatif. Cette méthode consiste à ajouter progressivement les termes significatifs du chaos, jusqu'à ce que la précision du métamodèle dépasse une valeur cible. Au final, seul un faible nombre de termes sont retenus (représentation creuse) par rapport à un chaos polynomial classique de type plein. La méthode est utilisée pour analyser la sensibilité du comportement d'un robinet industriel (via trois quantités d'intérêt) à differents paramètres incertains (matériaux, chargement et géométrie).

Le Package R ConvergenceConcepts: Un nouvel outil graphique pour l'étude de quelques modes de convergence de variables aléatoires

Pierre Lafaye de Micheaux, Benoît Liquet
Logiciels - Enseignements

Ce travail tente d'éclairer l'intuition sous-jacente à la convergence de suites de variables aléatoires en utilisant un package R appelé ConvergenceConcepts. Ce package permet de jouer avec de telles suites. Il est ainsi possible de visualiser des trajectoires d'une suite donnée de variables aléatoires et d'étudier, au moyen de représentations dynamiques adaptées, les quatre modes les plus classiques de convergence, à savoir la convergence presque sûre, la convergence en probabilité, la convergence en moyenne r et la convergence en loi.

Graphes, statistiques au format Web

Laurent Barthon
Logiciels - Enseignements

Probabilité - Statistiques sur le web :
- le format web et ses contraintes
- affichage de graphes, d'histogrammes
- gestion des événements : souris, clavier, liens, animation
- quelques applications : affichage dynamique de lois de probabilité, graphe de la VAN d'un investissement (finance), histogramme de table de mortalité (assurances).

Difficultés suscitées par les tests inductifs paramétriques chez des étudiants en sciences humaines

Noelle Zendrera
Logiciels - Enseignements

La présente communication poursuit le but de présenter une partie des résultats de notre recherche en éducation statistique. L'étude examine en particulier les difficultés suscitées par les tests inductifs paramétriques chez des étudiants en sciences humaines. Deux instruments distincts et complémentaires de collecte de données sont utilisés successivement dans notre expérimentation : d'abord l'"épreuve écrite", ensuite l'"entretien clinique individuel", tous deux axés sur la réalisation d'une tâche concrète. En l'occurrence, la tâche proposée consiste en la "résolution complète" d'un problème de test d'hypothèses sur une moyenne. 90 étudiants, tous volontaires, participent à la première phase, par épreuve écrite. Dans la seconde phase, 10 de ces 90 solutionneurs sont sélectionnés, en fonction de certains critères, puis rencontrés longuement en entretien individuel. L'analyse des données recueillies, à la fois qualitative et quantitative, compte un certain nombre de variables construites. Nous dégagerons ici les résultats obtenus au regard des groupes d'unités statistiques impliqués dans le test (échantillon, populations) et de leurs moyennes respectives (moyenne échantillonnale, moyenne de la population parente, norme). Ces résultats montrent, notamment grâce aux entretiens, que les étudiants élaborent des conceptions originales au regard de ces concepts; certaines des "conceptions erronées" décelées sont inédites à notre connaissance ; ils révèlent également et fortement de véritables "absences de conceptualisation" vis-à -vis de certains de ces concepts.

Statistique et compréhension de la vie quotidienne : un objectif pour l'enseignement de la statistique

Alain Bihan-Poudec

Logiciels - Enseignements

Un des objectifs assignés à l'enseignement de la statistique est la formation de futurs citoyens à même de décrypter les informations qu'ils reçoivent, notamment par les médias, et ce afin de favoriser leur esprit critique. Dans les cursus universitaires en sciences humaines et sociales, l'enseignement de la statistique remplit-il cet objectif ? Un exemple concret de validation dans un cursus de Licence en Sciences de l'Éducation nous permettra d'aborder cette question : partant de classements de films sur internet sur des sites cinématophiles, plusieurs questions étaient soumises aux étudiants : à quelle caractéristique de position correspond le nombre d'étoiles attribué aux films ? Quelles limites et quelles suggestions les étudiants pouvaient faire par rapport à cette procédure d'évaluation des films ? Au-delà des notes obtenues, les commentaires font apparaître des pré-conceptions surprenantes : déni de la variabilité des échantillons, nécessité de données quantitatives pour évaluer, croyance dans la nécessité d'une précision maximale, etc. ces verbatim confortent la nécessité pour l'enseignant d'identifier les conceptions préalables des étudiants.

Le conflit "Entropie vs Variance" pour des familles de lois bivariées

Rémy Landri

Statistique mathématique 1

L'analyse de sensibilité cherche à identifier les incertitudes importantes, et donc à accroître l'information sur les modèles. De façon générale, pour mener une analyse de sensibilité, on considère la variance de la sortie du modèle. Pour l'instant assez peu répandue dans le domaine de l'analyse de sensibilité, l'entropie de Shannon représente une mesure de l'incertitude et de la variabilité qui peut être une alternative à la variance. On propose ici une comparaison Entropie vs Variance, à différents niveaux, en termes de variation des paramètres des lois. Après avoir discuté du cas univarié, on se concentre sur une variété de

familles de distributions bivariées, qui inclue les copules, les lois normales, exponentielles, de Pareto, Gamma, de Dirichlet, etc. Dans un tel cadre bivarié, cette étude comparative s'effectue à partir de certains éléments clefs des analyses de sensibilité fondées sur la variance et sur l'entropie. Enfin, on développe l'analyse de la sensibilité fondée sur l'entropie et on l'applique à des cas-types. Ces applications sont comparées avec les résultats obtenus avec l'analyse de la sensibilité fondée sur la variance.

Test d'adéquation pour la loi gaussienne inverse basé sur la propriété de Matsumoto-Yor

Efoevi Angelo Koudou, Severien Nkurunziza

Statistique mathématique 1

Soient X et Y des variables aléatoires positives indépendantes. D'après la propriété de Matsumoto-Yor, les variables U=1/(X+Y) et V=1/X-1/(X+Y) sont indépendantes si et seulement si X suit une loi gaussienne inverse généralisée et Y suit une loi gamma. Nous utilisons cette propriété pour proposer un test d'adéquation pour la loi gaussienne inverse.

Estimation du paramètre de distribution de la distribution binomiale négative : a priori, effort d'échantillonnage, et information

Lise Vaudor

Statistique mathématique 1

La distribution binomiale négative est fréquemment utilisée pour modéliser la distribution des données d'abondance surdispersées. De ce fait, l'ajustement de la binomiale négative aux données d'abondance et l'estimation de ses paramètres est un enjeu majeur pour de très nombreuses études en écologie, épidémiologie, actuariat, etc. qui reposent sur des données de ce type. L'estimation des paramètres de la binomiale négative, et en particulier de son paramètre de dispersion est néanmoins problématique, car les estimateurs disponibles sont biaisés et peu efficaces, notamment lorsque les échantillons sont petits, que l'espérance de l'abondance est faible, et que sa variance est forte. Dans cette étude, nous tentons d'identifier la source de ces problèmes d'estimation. Pour ce faire, nous utilisons l'entropie (Shannon, 1949) pour quantifier l'information qu'apportent les échantillons sur la dispersion, en fonction des caractéristiques de ces échantillons (taille, moyenne), et des valeurs réelles des paramètres de la binomiale négative. On montre ainsi que le nombre total d'individus observés est un facteur clé de la qualité de l'estimation, et que pour un effort d'échantillonnage donné l'estimation de la dispersion est vraisemblablement très peu fiable si la dispersion réelle se situe dans des gammes de valeurs extrêmes (notamment des valeurs fortes).

Références

Shannon, C. E. (1948) A mathematical theory of communication. Bell Syst. Tech. J. 27, 379- 423.

Sur les estimateurs du maximum du vraisemblance dans les modèles multiplicatifs de Poisson et binomiale négatif

Lucien Diégane Gning, Daniel Pierre Loti Viaud
Statistique mathématique 1

Nous nous intéressons à l'existence et à l'unicité des estimateurs du maximum de vraisemblance des paramètres dans les modèles de Poisson multiplicatif et binomial négatif multiplicatif. Suite à ses travaux sur le modèle de Poisson multiplicatif sans réepétition à deux facteurs, Haberman (1973) a donné une condition nécessaire et suffisante pour l'existence et l'unicité de l'estimateur du maximum de vraisemblance de ce modèle, il a fourni en plus une expression explicite de cet estimateur. Dans cet article, nous proposons une généralisation de ces deux résultats à un modèle de Poisson multiplicatif avec répétitions à plus de deux facteurs. Nous montrons également que la condition obtenue est aussi une condition nécessaire et suffisante d'existence et d'unicité de l'estimateur du maximum de vraisemblance dans le modèle binomial négatif multiplicatif à plusieurs facteurs avec ou sans répétitions. We addres to the existence and the uniqueness of maximum likelihood estimate (MLE) of the paramaters in both multiplicative Poisson model and multiplicative binomial negative model. In 1973 Haberman gaves a necessary and sufficient condition to the existence and the uniqueness of the MLE's parameter of a two-factors multiplicative Poisson model. Furthermore he provided an explicit solution for this MLE. We generalize this result for J-factors multplicative Poisson model for clustered data (J $>$ 2). We show too that this condition is necessary and sufficient for the existence and the uniqueness for MLE's parameter in a J-factors multiplicative negative binomial model for clustered (or not clustered) data ($J \geq 2$).

Références

Haberman S.J. (1973). Log-linear models for frequency data : sufficient statistics and likelihood equations. Ann. Statist. 1, 617-632.

Imputation des données manquantes : comparaison de différentes approches

Mélanie Glasson-Cicognani, André Berchtold
Statistique mathématique 1

Les données manquantes constituent un problème majeur, puisque l'information à disposition est incomplète et donc moins fiable. Il est alors nécessaire de traiter correctement les données manquantes avant d'effectuer des analyses statistiques. L'objectif de cette recherche est de comparer par le biais de simulations numériques différentes méthodes existantes pour le traitement des données manquantes. Nous considérons à la fois des méthodes anciennes comme l'analyse des cas complets, le remplacement par la moyenne ou le remplacement par le plus proche voisin, des méthodes d'imputation simple basées notamment sur la régression, et finalement différentes procédures d'imputation multiple. En partant d'un fichier sans aucune donnée manquante, nous avons créé neuf scénarios variant en fonction du nombre de données manquantes et de leur type (complètement aléatoire, aléatoire ou non-aléatoire). Mille ensembles de données ont été générés à partir de chaque scénario, puis les données manquantes ont été traitées selon différentes procédures et les moyennes, écarts-types et corrélations des variables imputées ont été comparés avec le fichier original sans données manquantes. L'influence du traitement des données manquantes sur un modèle de régression a aussi été évaluée. Nos résultats montrent que les méthodes qui permettent globalement d'arriver aux résultats les plus satisfaisants sont des méthodes basées

sur l'imputation multiple. D'autres méthodes, comme par exemple l'imputation simple par régression, permettent aussi l'obtention de résultats intéressants, mais seulement dans certaines situations particulières. Certaines méthodes anciennes, comme l'analyse des cas complets, sont à bannir absolument.

Estimation dans un modèle défini par des équations estimantes conditionnelles pour des données fonctionnelles

Matthieu Saumard, Valentin Patilea

Statistique non (semi) paramétrique 1

Nous considérons ici un modèle défini par des restrictions de moments conditionnels dans lequel les valeurs de la variable dans le conditionnement ainsi que le paramètre d'intérêt sont des éléments d'un espace fonctionnel. Pour estimer le paramètre d'intérêt, nous utiliserons une technique de troncature afin d'appliquer et généraliser la théorie existante dans les modèles finies-dimensionnelles.

Efficacité semi-paramétrique pour la méthode des moments généralisée

Paul Rochet, Jean-Michel Loubes

Statistique non (semi) paramétrique 1

La méthode des moments généralisée est une méthode de régularisation de problèmes inverses, où l'on cherche à reconstruire une mesure qui vérifie une contrainte linéaire. Cette contrainte dépend d'un paramètre inconnu que l'on veut estimer. Chamberlain a montré qu'un estimateur du paramètre ne peut avoir une variance inférieure à une certaine borne. Nous proposons une preuve différente de ce résultat qui fait intervenir la théorie de l'efficacité semi paramétrique. Notre approche permet également de montrer un résultat du à Hansen sur la construction d'un estimateur asymptotiquement efficace.

Estimation récursive en régression inverse par tranche (sliced inverse regression)

Thi Mong Ngoc Nguyen, Jérôme Saracco

Statistique non (semi) paramétrique 1

Dans cette communication, nous nous intéressons à la méthode SIR (Sliced Inverse Regression, que l'on peut traduire par régression inverse par tranches) qui permet d'estimer le paramètre θ dans un modèle semi-paramétrique de régression du type $y = f(x'\theta, \varepsilon)$ sans avoir à estimer le paramètre fonctionnel f ni à spécifier la loi de l'erreur ε. Nous proposons un estimateur récursif de la direction de θ dans le cas particulier où l'on considère $H = 2$ tranches. Nous donnons des propriétés asymptotiques de cet estimateur (convergence et normalité asymptotique). Nous illustrons aussi sur des simulations le bon comportement numérique de la méthode proposée.

Régression semi-paramétrique de variables explicatives de dénombrement

Belkacem Abdous, Célestin Kokonendji, Tristan Senga Kiessé

Statistique non (semi) paramétrique 1

Dans cette communication, nous proposons un estimateur semi-paramétrique d'une fonction discrète de régression. Nous combinons une approche paramétrique et une non-paramétrique pure qui met en oeuvre la méthode des noyaux associés discrets. Le nouvel estimateur semi-paramétrique discret de régression possède un biais qui peut être inférieur à celui de l'estimateur purement non-paramétrique discret de régression et une même variance que ce dernier.

Régression inverse par tranches pour une population stratifiée

Marie Chavent, Vanessa Kuentz, Benoit Liquet, Jérôme Saracco

Statistique non (semi) paramétrique 1

Dans cette communication, nous considérons un modèle semiparamétrique de régression dans lequel une variable à expliquer Y dépend d'une covariable quantitative X de dimension p et d'une variable qualitative Z. Cette covariable Z définit une stratification de la population. Ce modèle inclut une reduction de sa partie explicative via un indice $X'\beta$. Nous proposons une approche fondée sur la méthode SIR (Sliced Inverse Regression ou régression inverse par tranches en français) afin d'estimer la direction du vecteur de paramètre β. Nous avons obtenu des résultats asymptotiques pour l'estimateur proposé (convergence et normalité asymptotique). Des simulations ont montré le bon comportement numérique de l'estimateur dans les cas homoscédastique et hétéroscédastique.

Estimation non-paramétrique des ensembles de niveaux de la régression

Thomas Laloe

Statistique non (semi) paramétrique 1

Soit (X, Y) un couple aléatoire à valeurs dans $\Lambda \times J$, où $\Lambda \subset \mathbb{R}^d$ et $J \subset \mathbb{R}$ sont supposés bornés. Nous construisons un estimateur plug-in des ensembles de niveaux de la fonction de régression r de Y sur X, à partir d'un estimateur à noyaux de r. Nous obtenons une vitesse de convergence du même ordre que celle obtenue par Cadre dans le cas de la densité. Nous discutons ensuite les résultats obtenus sur des données simulées.

Classification non supervisée des données fonctionnelles avec ondelettes

Anestis Antoniadis, Xavier Brossat, Jairo Cugliari, Jean-Michel Poggi

Ondelettes

Cet article présente une méthode permettant de détecter efficacement des clusters dans des données fonctionnelles avec une structure temporelle. L'algorithme est basé sur une mesure de similarité fondée sur la décomposition en ondelettes. Cette décomposition est idéale pour l'identification de caractéristiques locales en temps et en échelle grâce à laquelle on peut visualiser et regrouper les données fonctionnelles en groupes homogènes.

Nous considérons les contributions de chaque échelle à l'énergie globale de chaque fonction d'entrée pour générer un codage ne perdant pas la capacité de différencier les signaux. La mesure de similarité, en parallèle avec une technique efficace de sélection de variables, est alors utilisée avec des algorithmes plus ou moins classiques d'apprentissage non supervisé.

La performance de cette méthodologie est démontrée par des simulations et de véritables applications de données.

Trajectory prediction by functional regression in Sobolev space

Kairat Tastambekov, Stéphane Puechmorel, Daniel Delahaye, Christophe Rabut
Ondelettes

Le problème de la prédiction de la trajectoire d'un avion à partir des informations de positions passées est fondamental dans le domaine de la gestion du trafic aérien. Nous présentons une approche nouvelle en utilisant une technique de régression linéaire locale en espace de Sobolev. Une décomposition en ondelettes des trajectoires avion permet une implémentation efficace. Un exemple de prédiction de trajectoires sur une base d'apprentissage est donné.

Méthodes multivariées combinant ondelettes et analyse en composantes principales pour le débruitage de données issues de spectrométrie de masse

Elise Mostacci, Caroline Truntzer, Hervé Cardot, Patrick Ducoroy
Ondelettes

L'identification de nouveaux biomarqueurs diagnostiques ou pronostiques est un des objectifs majeurs en recherche clinique. L'utilisation des technologies à haut débit comme la spectrométrie de masse est prometteuse pour l'identification de tels marqueurs. A partir d'un prélèvement de sang ou de tumeur par exemple, cette technologie permet de traduire sous forme de spectres le profil protéique des individus. Le signal biologique observé dans les spectres est masqué par différentes sources de variabilités techniques, qu'une phase préalable de prétraitement doit permettre de retirer. La méthode classique permettant de retirer le bruit aléatoire de mesure de ce signal combine la méthodologie des ondelettes et un seuillage, ceci spectre à spectre. L'utilisation des ondelettes permet la séparation du bruit aléatoire (coefficients de détail) et du signal biologique (coefficients d'approximation). Le seuillage des coefficients de détails permet ensuite d'annuler un certain nombre d'entre eux. Nous proposons dans ce travail d'améliorer le débruitage classique des données en tenant compte de la structure commune des signaux. Nous avons pour cela adapté aux données spectrométriques deux méthodes de débruitage qui combinent les ondelettes, le seuillage et les analyses en composantes principales classique ou creuse. Les méthodes proposées ont été évaluées et comparées à la méthode de seuillage univariée classique, ceci pour des données réelles et simulées. Il a été montré que l'ajout d'une étape de réduction de la dimension sur les approximations par une analyse en composantes principales en plus

d'un seuillage classique sur les détails améliore le débruitage.

Estimation indirecte de l'âge en paléodémographie : approche bayésienne

Henri Caussinus, Daniel Courgeau, Isabelle Séguy, Luc Buchet

Statistiques et applications

En vue d'estimer la structure par âge des populations du passé en ne disposant que d'indicateurs biologiques, les paléodémographes ont développé un certain nombre de méthodes statistiques, utilisant une population de référence pour apprécier les probabilités conditionnelles de l'âge connaissant l'indicateur. Compte tenu du faible nombre de données disponibles et du caractère instable du problème, ces méthodes sont en général décevantes. Nous montrons comment les améliorer en introduisant une méthode bayésienne simple intégrant un maximum d'informations non réductibles aux données proprement dites.

Masses des lois multinomiales négatives. Application au traitement d'images polarimétriques

Phlippe Bernardoff, Florent Chatelain, Jean-Yves Tourneret

Statistiques et applications

Cet article est dérivé d'une nouvelle expression des masses des lois multinomiales négatives multivariées. Cette expression des masses peut être utilisée pour déterminer les estimateurs du maximum de vraisemblance de ses paramètres inconnus. Une application au traitement d'images polarimétriques est étudiée. Plus précisément, les estimateurs du degré de polarisation utilisant la méthode du maximum de vraisemblance avec différentes combinaisons d'images sont comparés.

Air-Conditioning Effect Estimation For Mid-Term Forecasts of Tunisian Electricity Consumption

Farouk Mhamdi, Mouhamed Ould Mahmoud Mouhamed, Mériem Jaïdane, Jomaa Souissi

Statistiques et applications

Nous proposons, dans ce papier, un modèle de prévision tenant compte de l'effet de la climatisation sur la consommation d'électricité. Cette modélisation a nécessité une analyse spécifique de l'élasticité de la consommation journalière par rapport à la température. Ainsi, l'étude de l'évolution de la part estimée liée à la température (chaud/froid), nous a permis, de prévoir l'évolution du parc d'appareils de climatisation. Ce résultat était particulièrement utile, dans la prévision à moyen terme surtout en absence d'une quantification directe du parc de climatiseurs.

Estimation de densité via un algorithme EM-Kernel

Catherine Aaron

Statistiques et applications

On s'intéresse, dans ce papier à la construction d'un algorithme automatique d'estimation de densité. Cet algorithme repose sur un modèle de mélange. Chaque composante de ce mélange peut être estimée via une méthode à densité dépendant d'un paramètre λ . La valeur de λ pour chaque composante du mélange, ainsi que les probabilités d'appartenance seront estimées par une méthode type EM par abus de langage (on maximisera une pseudo vraisemblance en alternant les étapes optimisation sur λ, optimisation sur les probabilités). Le choix d'une telle méthode sera justifié, pour la dimension 1, en partie 1. Des résultats asymptotiques seront explicités en partie 2. Enfin on présentera l'algorithme proprement dit et quelques résultats, en parties 3 et 4.

Modèles de mélange tronqués pour l'écologie microbienne. Estimation du nombre d'espèces manquantes

Sebastien Li-Thiao-Te, Jean-Jacques Daudin, Stéphane Robin, Émilie Lebarbier

Mélanges -Modèles latents

Dans le modèle d'échantillonnage présenté par Fisher en 1943, chaque espèce apporte à l'échantillon un nombre d'individus distribué selon une loi de Poisson. La moyenne de cette loi est spécifique à l'espèce. Cependant, de nombreuses espèces apportent zéro individus et ne sont donc pas observées dans les données.

Nous utilisons des modèles de mélanges tronqués pour l'abondance des espèces et estimons les paramètres par maximum de vraisemblance. A cause des espèces manquantes, l'algorithme EM n'est pas applicable directement. Nous présenterons comment appliquer l'algorithme EM à des mélanges de distributions tronquées et en déduire la solution du problème.

Pour obtenir des intervalles de confiance, nous utilisons l'approche variationnelle dans un cadre Bayésien. De même que précédemment, il est plus facile d'appliquer l'algorithme pour un mélange de distributions tronquées et d'en déduire la solution du problème.

Dans le cadre de la métagénomique, nous montrerons comment estimer le nombre d'espèces manquantes et comment déduire des courbes de raréfaction du modèle de mélange.

Approche bayésienne variationnelle pour l'agrégation de modèles en classification

Stevenn Volant, Marie-Laure Martin-Magniette, Stéphane Robin

Mélanges -Modèles latents

Nous nous intéressons au cas d'un mélange entre deux populations dont l'une est connue et facilement identifiable. Plusieurs modèles ont été développés pour modéliser la distribution inconnue. Nous proposons une alternative qui consiste à prendre un mélange de plusieurs distributions gaussiennes de moyennes et variances inconnues. Chaque modèle apporte une information plus ou moins pertinente sur l'estimation des paramètres. Nous suggérons alors d'utiliser une approche BMA pour prendre en compte l'incertitude relative à chacun des modèles ainsi que de s'affranchir du choix du nombre de composants. En moyennant sur un ensemble de modèles, le BMA permet de calculer un estimateur agrégé

à partir de l'information apportée par la collection de modèles, pondérée par le poids du modèle concerné. Dans la pratique, ce poids est estimé à partir du BIC mais la qualité de l'approximation pour obtenir ce critère est discutable. Ainsi, nous nous intéressons au cadre bayésien variationnel qui permet de définir naturellement une distribution a posteriori des paramètres et d'obtenir un poids pour chacun des modèles. Nous proposons dans ce travail la définition de poids de chaque modèle à partir de la minimisation de la divergence de Kullback-Leibler entre la distribution estimée des poids et la vraie. Une étude de simulation permet d'évaluer le comportement de notre estimateur agrégé.

Classification basée sur des mélanges de modèles hiérarchiques bivariés

Vera Georgescu, Nicolas Desassis, Samuel Soubeyrand, André Kretzschmar, Rachid Senoussi

Mélanges -Modèles latents

Les approches probabilistes basées sur les modèles de mélange sont de plus en plus utilisées en classification automatique car elles fournissent un cadre formel pour résoudre des problèmes pratiques qui se posent en classification, tels que la détermination du nombre de classes, et permettent d'estimer l'incertitude associée à la classification. Les limites de ces méthodes résident principalement dans le choix de la loi de probabilité des composantes du mélange, qui dépend du type de données et va contraindre la forme des classes. Peu de modèles de mélange ont été étudiés dans le cas multivarié, et il est difficile d'adapter la méthode d'estimation d'une distribution à une autre. Nous proposons une généralisation des méthodes de classification basées sur des modèles de mélange qui :
- s'adapte rapidement à des données de type différents (continues, discrètes, binaires, surdispersées),
- permet d'obtenir des formes de classe et des structures de corrélation variées,
- permet de traiter des données multivariées comportant des observations de plusieurs types.

Pour cela nous considérons un modèle hiérarchique, dans lequel la couche cachée est issue d'un mélange de lois gaussiennes bivariées et la couche d'observation est obtenue par une distribution bivariée dont le choix dépend du type de données observées. L'estimation du modèle se fait par un algorithme MCEM.

Estimation d'un modèle à blocs latents par l'algorithme SEM

Christine Keribin, Gérard Govaert, Gilles Celeux

Mélanges -Modèles latents

Les modèles de mélanges peuvent être utilisés pour résoudre le problème de la classification non supervisée simultanée d'un ensemble d'objets et d'un ensemble de variables. Le modèle à blocs latents définit une loi pour chaque croisement de classe d'objets et de classe de variables, et les observations sont supposées indépendantes conditionnellement au choix des classes d'objets et de variables. Mais il n'est pas possible de factoriser la loi jointe conditionnelle des labels et l'étape d'estimation de l'algorithme EM n'est pas calculable directement. Govaert et Nadif (2008) en ont proposé une approximation variationnelle qu'ils ont confrontée à un algorithme CEM. Nous présentons ici, dans le cadre de données binaires, l'utilisation d'un algorithme SEM effectuant l'étape d'estimation par échantillon-

neur de Gibbs, et nous comparons les résultats avec ceux des méthodes précédentes.

Références

Govaert, G. et Nadif M. (2008) Block clustering with Bernoulli mixture models : Comparison of different approaches. Computational Statistics and Data Analysis, 52, 3233-3245.

Partition latente et dégénérescence dans les mélanges gaussiens

Christophe Biernacki

Mélanges -Modèles latents

Dans le cas des mélanges gaussiens, il est notoire que la vraisemblance peut augmenter sans limite si, par exemple, une des gaussiennes est centrée en une observation et simultanément la matrice de variance correspondante tend vers la singularité. Les tentatives actuelles pour résoudre ce problème reposent sur la détention d'information hypothétique sur les matrices de variance elles-mêmes. Notre proposition consiste à introduire une information naturelle sur la partition latente impliquée dans le modèle génératif sous-jacent qui permet de borner la vraisemblance tout en préservant la convergence des estimateurs. Constatant que l'algorithme EM impliqué dans l'optimisation de cette vraisemblance particulière implique des difficultés combinatoires pour plus de deux composantes gaussiennes, l'information de partition peut être alternativement utilisée pour proposer une borne probabiliste inférieure non asymptotique sur les variances généralisées qui est simple à calculer dans la plupart des situations et qui conserve la convergence des estimateurs. Des expériences numériques illustrent les deux propositions.

Modèles linéaires généralisés à facteurs : une estimation par algorithme EM local

Xavier Bry, Christian Lavergne, Mohamed Saidane

Mélanges -Modèles latents

Les modèles à facteurs ont été développés et étudiés dans le cas où les observations sont supposées être de loi normale. Nous considérons ici le contexte plus large où les observations sont supposées suivre une loi de la famille exponentielle. On obtient ainsi une nouvelle classe de modèles à facteurs: les modèles linéaires généralisés à facteurs (GLMF). Les GLMF permettent la modélisation multivariée de données discrètes (binômiale, Poisson...), mais aussi de certaines données continues non normales (gamma, par exemple). Ils permettent notamment l'étude conjointe de variables de types différents, supposées dépendre de facteurs communs. Les GLMF sont, formellement, une synthèse des GLM et des modèles factoriels standards. Nous proposons une méthode d'estimation des paramètres et des facteurs de ces modèles, en combinant l'algorithme des scores de Fisher pour les GLM avec un algorithme itératif de type EM. Nous étudions les performances de cette méthode en l'appliquant sur des données simulées.

Analyse en axes principaux de variables symboliques de type histogramme

Sun Makosso Kallyth, Edwin Diday

Dans cet article nous proposons deux approches susceptibles d'effectuer une analyse en composantes principales ACP des variables symboliques de type histogramme. Ces approches sont applicables même quand le nombre de modalité des histogrammes diffère. On associe à chaque variable de type histogramme un ensemble de modalités qui ont chacune une fréquence relative lorsqu'on considère un individu. La première approche compte trois étapes. Elle procède d'abord par un codage des modalités des variables de type histogramme. L'objet de ce codage est d'attribuer des valeurs numériques appelées "scores" aux modalités des variables. A ce titre, on propose un codage paramétrique et un codage non paramétrique. La seconde étape consiste à effectuer une ACP des moyennes des histogrammes puis projette en éléments supplémentaires les sommets d'hypercubes induit par l'inégalité de Tchebychev. Dans la seconde approche, en plus des considérations précédentes, on utilise une transformation normalisatrice des données en guise de prétraitement des variables.

Application de l'Analyse des Correspondances Ordinales au suivi d'espèces végétales aquatiques

Claude Manté, Guillaume Bernard, Patrick Bonhomme

Analyse factorielle et analyse des données

Les communautés de macrophytes (algues et phanérogames), dont certaines espèces sont protégées, constituent un des indicateurs clés de l'état écologique des lagunes, tel que défini par la Directive Européenne Cadre sur l'Eau (DCE). Par ailleurs, la restauration de ces communautés de macrophytes, et des herbiers de Zostera en particulier, est l'un des objectifs principaux de la réhabilitation de l'Etang de Berre (Bouches-du-Rhône) portée par le Gipreb (Groupement d'intérêt public pour la réhabilitation de l'Etang de Berre). C'est pourquoi celui-ci met en oeuvre, depuis 1996, le suivi des principales espèces de macrophytes encore présentes dans l'étang. La densité de chaque espèce ou groupe d'espèces a été évaluée visuellement par des plongeurs le long de 31 transects (composés chacun de 20 segments de même longueur) répartis sur le pourtour de l'étang, et ce pendant 11 années. L'évaluation consiste en un codage en 6 modalités du taux de couverture du fonds par l'espèce. L'état annuel de la population de chaque espèce est finalement décrit par un tableau de type 6x20, croisant la densité avec la position : deux variables ordinales. La suite des 11 tableaux sera donc ici décrite via l'Analyse des Correspondances Ordinales de Beh (1997). Les résultats obtenus seront comparés avec ceux de l'AFC.

Références

E. J. Beh, Simple Correspondence Analysis of ordinal cross-classifications using orthogonal polynomials, Biometrical Journal, 39, 5, 589-613 (1997).

ACP projetée de données séquentielles

Jean-Marie Monnez

Analyse factorielle et analyse des données

On suppose que des vecteurs de données arrivant séquentiellement dans le temps sont des observations i.i.d. d'un vecteur aléatoire Z = (R, S) partitionné en deux vecteurs R et S. On définit une méthode récursive d'estimation séquentielle des r premiers facteurs de l'ACP projetée de R par rapport à S. On étudie ensuite le cas particulier de l'analyse canonique pour lequel on définit deux autres processus spécifiques, ainsi que celui de l'analyse factorielle discriminante.

Variabilité des dimensions en ACP : cas complet et incomplet

Julie Josse, François Husson

Analyse factorielle et analyse des données

Dans cette présentation, nous nous intéressons à évaluer la stabilité des dimensions en Analyse en Composantes Principales (ACP). La présentation de l'ACP via un modèle à effets fixes permet de proposer une technique de bootstrap des résidus et d'associer des zones de confiance autour de la position des individus et des variables. Cet algorithme est étendu au cas incomplet et permet de prendre en compte l'incertitude supplémentaire due aux données manquantes. Une méthode d'imputation multiple adaptée au cadre de l'ACP est ensuite proposée pour évaluer la variabilité des dimensions due aux données manquantes.

L'analyse exploratoire multidimensionnelle d'un modèle structurel fondée sur une classe de critères de covariance généralisée

Xavier Bry, Patrick Redont, Thomas Verron

Analyse factorielle et analyse des données

Notre but est d'explorer un modèle structurel: plusieurs groupes de variables décrivant les mêmes unités sont supposées structurés autour de dimensions latentes liées entre elles par un modèle linéaire pouvant comporter plusieurs équations. Ce type de modèle est couramment traité par des méthodes ne considérant qu'une dimension latente par groupe. Cependant, les modèles conceptuels relient couramment des concepts structurellement multidimensionnels, sans que l'on sache a priori combien de dimensions interviennent, ni lesquelles. Nous proposons une classe de critères pouvant mesurer la qualité d'un modèle structurel. Cette classe contient le critère de covariance fondant la régression PLS, ainsi que la covariance multiple fondant la méthode SEER. Elle contient également des critères liés à la rotation quartimax. Tous les critères de cette classe sont appelés à être maximisés sous contrainte de norme unité des vecteurs en argument. Nous donnons un programme d'optimisation libre équivalent, ainsi qu'un algorithme pour le résoudre. Cette optimisation est utilisée à l'intérieur d'un algorithme plus général (nommé THEME, pour: Thematic Equation Model Explorer) permettant la recherche dans chaque groupe de toutes les dimensions utiles au modèle. THEME extrait des composantes thématiques localement hiérarchisées. La méthode est appliquée à la modélisation de données chimiométriques multi-tableaux.

Détection de rupture dans un modèle exponentiel

Olivier Lopez, Vladimir Spokoiny

Statistique mathématique 2 : rupture - algorithme

Nous considérons un modèle de rupture sur le paramètre d'un modèle exponentiel, qui est estimé par maximum de vraisemblance. Grâce à de nouvelles bornes exponentielles, nous obtenons des résultats à distance finie sur la qualité d'estimation de l'instant de rupture ainsi que de son amplitude.

Algorithme rapide pour la détection optimale des ruptures

Guillem Rigaill

Statistique mathématique 2 : rupture - algorithme

Dans les modèles de détection de ruptures, les données sont modélisées par un processus aléatoire dont les paramètres sont soumis à des changements brusques en des instants inconnus, appelés instants de ruptures. La recherche exhaustive des positions des ruptures de norme quadratique minimale se fait par un algorithme de programmation dynamique. L'intérêt de cet algorithme est qu'il permet d'obtenir la solution optimale en réduisant la complexité algorithmique de $O(n^K)$ à $O(Kn^2)$, où $K-1$ est le nombre de ruptures fixé et n la taille du signal. Même si le temps de calcul est réduit, cet algorithme ne peut être utilisé sur des signaux de grandes tailles. Nous proposons ici un nouvel algorithme de programmation dynamique permettant d'obtenir la solution optimale en un temps de calcul très nettement réduit. Notamment il permet d'analyser un signal d'un million de points en quelques minutes. Plus précisement, nous d'montrons qu'au pire des cas sa complexité en temps et en espace sont respectivement de $O(Kn^2)$ et de $O(Kn)$ et nous montrons que son temps de calcul est empiriquement de l'ordre de $O(Kn\log(n))$. Par ailleurs, Nous comparons cet algorithme à l'algorithme de programmation dynamique classique. L'algorithme proposé est au pire des cas équivalent et a une complexité empirique bien plus faible.

Echantillonnage de champs gaussiens de grande dimension

Olivier Feron, François Orieux, Jean-François Giovannelli

Statistique mathématique 2 : rupture - algorithme

Nous proposons une nouvelle approche pour l'échantillonnage de champs gaussiens corrélés dans le cas où les approches classiques ne sont pas utilisables : lorsque la dimension du problème est très grande et lorsque la matrice inverse de covariance (ou matrice de précision) n'est pas creuse. Cette approche est valide dans le cas où une structure particulière de la matrice de précision est disponible. Cette structure apparaît dans la résolution de problèmes inverses par des méthodes d'estimation bayésienne. L'algorithme proposé trouve une application directe pour les méthodes d'inversion myopes et/ou non supervisées, fondées sur des méthodes d'échantillonnage de type MCMC. L'efficacité de cette approche est illustrée sur l'inversion non supervisée d'un problème de super résolution.

Méthodes non linéaires pour des problèmes statistiques inverses

Pierre Barbillon, Gilles Celeux, Agnès Grimaud, Yannick Lefebvre, Etienne De Rocquigny

Statistique mathématique 2 : rupture - algorithme

Dans le cadre du traitement des incertitudes étudié ici, la variabilité intrinsèque des entrées d'un modèle physique est modélisée par une loi de probabilité multivariée. L'objectif est d'identifier cette loi de probabilité à partir d'observations des sorties du modèle. Afin de se limiter à un nombre d'appels raisonnable au code de calcul (souvent coûteux) du

modèle physique dans l'algorithme d'inversion, une méthodologie d'approximation non linéaire faisant intervenir le krigeage et un algorithme EM stochastique est présentée. Elle est comparée à une méthode utilisant une approximation linéaire itérative sur la base de jeux de données simulées provenant d'un modèle de crues simplifié mais réaliste. Les cas où cette approche non linéaire est préférable seront mis en lumière.

Modèles génératifs de rangs relatifs à un algorithme de tri par insertion

Christophe Biernacki, Julien Jacques

Statistique mathématique 2 : rupture - algorithme

Les données de rang proviennent d'un processus de tri dont la nature est généralement inaccessible au statisticien. Faisant l'hypothèse que ce tri repose sur la comparaison entre paires d'objets et que par ailleurs le processus retenu vise à en minimiser le nombre, l'algorithme de tri par insertion s'impose comme l'un des meilleurs candidats. Par l'introduction d'une erreur de Bernoulli sur les comparaisons de paires, on obtient une modélisation générative probabiliste des données de rang dont une des originalités est de dépendre de l'ordre de présentation initiale des objets à classer. En fonction des hypothèses d'échantillonnage relatives à cet ordre de présentation (inconnu), plusieurs modèles réalistes sont obtenus. Des expériences numériques sur des données réelles permettent de comparer ces modèles avec le modèle standard Phi de Mallows.

Bootstrap dans l'estimation de la densité par la méthode du noyau

Smail Adjabi, Mouloud Cherfaoui

Statistique semi(non) paramétrique 2

Dans ce travail, on compare par simulation sur des densités cibles selon le critère de l'erreur quadratique moyenne intégrée, trois méthodes de sélection du paramètre de lissage qui interviennent dans l'estimation de la densité par la méthode du noyau : la méthode plug-in de Sheather and Jones et les deux méthodes de validation croisée : la validation croisée non biaisée et la validation croisée biaisée, en utilisant la technique de ré-échantillonnage de bootstrap. On étudie aussi l'influence du noyau selon le même critère sur les performances de l'estimateur de la densité particulièrement pour les densités à support compact.

Réduction itérative du biais pour des lisseurs multivariés

Pierre-André Cornillon, Nick Hengartner, Eric Matzner-Lober

Statistique semi(non) paramétrique 2

La méthode IBR (iterated biased reduction) permet d'estimer une fonction de régression m inconnue lorsque les variables explicatives sont à valeurs dans $\mathbb{R}^d$. Pour estimer la fonction m, les méthodes non-paramétriques classiques souffrent du fléau de la dimension.

En pratique, il faut donc supposer des hypothèses structurelles: modèles additifs, modèles à directions révélatrices... A contrario IBR estime directement la fonction de régression m. Elle concurrence MARS, les directions révélatrices ou les modèles additifs et sur des exemples réels ou simulés et elle apporte des gains significatifs sur l'erreur de prévision. Cette méthode utilise en pratique un lisseur pilote soit de type splines plaque-minces soit de type noyau gaussien. Cet estimateur pilote est utilisé de manière répétée afin d'estimer le biais et permet de l'enlever progressivement. La méthode, à l'instar du L_2 boosting, nécessite donc l'estimation de l'itération optimale. Des résultats de vitesse de convergence (vitesse minimax) de l'erreur quadratique moyenne de l'estimateur (avec itération optimale) ont été obtenus. L'optimalité du critère de choix de l'itération (GCV) a aussi été démontré. Un exemple simulé simple ($d = 2$) et un exemple réel ($d = 8$) seront traités et comparés aux méthodes existantes: GAM, MARS, PPR, ou L_2-boosting. Un package R disponible sur le CRAN permet d'utiliser cette méthode très simplement.

Sur l'estimation du support d'une densité

Gérard Biau, Benoît Cadre, Bruno Pelletier

Statistique semi(non) paramétrique 2

Etant donnée une densité de probabilité multivariée inconnue f à support compact et un n-échantillon i.i.d. issu de f, nous étudions l'estimateur du support de f défini par l'union des boules de rayon r_n centrées sur les observations. Afin de mesurer la qualité de l'estimation, nous utilisons un critère général fondé sur le volume de la différence symétrique. Sous quelques hypothèses peu restrictives, et en utilisant des outils de la géométrie riemannienne, nous établissons les vitesses de convergence exactes de l'estimateur du support tout en examinant les conséquences statistiques de ces résultats.

Aspect brownien d'un test semi-paramétrique d'indépendance

Bernard Colin, Ernest Monga

Statistique semi(non) paramétrique 2

Étant donné n vecteurs aléatoires $X_1, , X_2..., X_n$ de dimensions finies, on considère le test semi-paramétrique d'indépendance entre ces derniers tel que présenté dans Colin et Monga (2009). Après en avoir illustré son usage sur quelques exemples et avoir mis en évidence de façon empirique sa puissance, on se propose alors dans un cadre plus théorique, de montrer que ce test est naturellement associé à un pont brownien, ce qui permet ainsi, dans le cas de certaines formes d'hypothèses alternatives de décrire plus aisément son comportement asymptotique. Enfin, on comparera au niveau de l'efficacité relative, le test proposé au test du rapport de vraisemblance dans le cadre de certains modèles paramétriques donnés.

Références

Bernard Colin et Ernest Monga (2009) : Efficacité d'un test semi-paramétrique d'indépendance entre vecteurs aléatoires, Congrès de la SFdS : 41e Journées de Statistique (Bordeaux).

Tests d'hypothèses dans un modèle de régression non paramétrique

Zaher Mohdeb

Statistique semi(non) paramétrique 2

On considère le modèle de régression non paramétrique de fonction de régression f. Une procédure de test d'hypothèse sur les coefficients de Fourier de f est proposée. On obtient le comportement asymptotique de la statistique de test proposée, on a donc ainsi le niveau et la puissance asymptotique du test. De tels tests peuvent, en particulier, être utilisés pour comparer deux signaux bruités dans une bande de fréquence. Un autre exemple est le test de l'hypothèse "f est un polynôme trigonométrique". Une étude par simulation a été menée, pour des petites tailles d'échantillon, afin de montrer la performance du test proposée.

Index des auteurs

Les Sponsors partenaires des journées