Groupe de travail STAPH : Statistique Fonctionnelle et Opératorielle

Partie IV : Recueil de résumés 2002-2003

Coordinateurs
A. Boudou, H. Cardot, F. Ferraty, Y. Romain,
P. Sarda, P. Vieu et S. Viguier-Pla

Résumé

Ce document a pour objectif de présenter les résumés (plus ou moins détaillés selon les souhaits de leurs auteurs) des divers exposés qui ont eu lieu lors des séances du groupe de travail STAPH durant l'année universitaire 2003-2004.

Rappelons que ce groupe de travail en Statistique Fonctionnelle et Opératorielle, créé il y a quatre ans au sein du Laboratoire de Statistique et Probabilités de Toulouse, s'inscrit dans la dynamique actuelle autour des divers aspects fonctionnels de la statistique moderne. Les exposés qui sont présentés traitent de divers aspects de la Statistique Fonctionnelle (estimation nonparamétrique, statistique opératorielle, modèles de réduction de dimension, modèles pour variables fonctionnelles, ...) ; ils sont de nature différentes (exposés didactiques ou bibliographiques, exposés de résultats nouveaux en Statistique Appliquée et/ou Théorique, ...) ; ils témoignent enfin de l'ouverture de la démarche par la grande diversité des exposants.

En préambule de ce document, un court texte est présenté afin de tirer le bilan des deux premières années de travail et afin surtout de mieux préparer l'avenir en faisant perdurer cette dynamique de recherche.

Pour terminer signalons que l'intégralité des activités de ce groupe de travail est disponible sur notre page web :

$$http://www.lsp.ups-tlse.fr/Fp/Ferraty/staph.html$$

Abstract We present the abstracts (of size more or less important according to the wishes of their authors) of the several different talks given during the sessions of the working group STAPH along the academic year 2003-2004. This group in Functional and Operatorial Statistics is born four years ago at the Laboratoire de Statistique et Probabilités of the Université Paul Sabatier de Toulouse, and its aim was to participate at the actual dynamic existing around the different functional features of modern statistics.

These talks were about different functional topics (nonparametric estimation, statistics of operators, models for functional data, models for dimension reduction, ...). They were of different kinds (didactic, bibliografic, applied, theoric, ...) and were presented by a large variety of statisticians.

As a foreword, a short text is presented to take the stock of the activities of this group during its two first years of existency in order to make an efficient preparation of the future.

As a conclusion, note that all the activities of this group are reachable through the following web adress : $http://www.lsp.ups-tlse.fr/Fp/Ferraty/staph.html$.

Sumario This documento presenta resumens (mas o menos cortos segun los deseos de sus autores) de charlas que han sido presentadas durante las sesiones de trabajo del grupo STAPH durante el ano academico 2002-2003. Este grupo de trabajo en el campo de Estadistica Funcional y Operatorial ha sido creado hace cuatros anos en el Laboratoire de Statistique et Probabilités de l'Université Paul Sabatier de Toulouse, para animar investigaciones en varios aspectos funcionales de la estdistica moderna.

Estas conferencias fueran sobre temas variados (estimacion noparametrica, estadistica de operatores, modelos para variables funcionales, modelos de reduccion de dimension, ...) y fueran de tipos differentes (conferencias didacticas o bibliograficas, presentacion de resultados nuevos en estadistica teorica o/y applicada, ...).

Al principio del documento, empezamos con un corto texto de presentacion en lo cual hacemos un chequeo de las actividades de este grupo desde dos anos y en lo cual planteamos los fundamentos para el proximo futuro.

Al final, queremos apuntar que todas nuestras actividades pueden ser consultadas a la direccion : $http://www.lsp.ups-tlse.fr/Fp/Ferraty/staph.html$.

TABLE DES MATIERES

[1] STAPH : Bilan de l'année 2002-2003 et perpectives

Alain Boudou, Hervé Cardot, Frédéric Ferraty
Yves Romain, Pascal Sarda, Philippe Vieu et Sylvie Viguier-Pla

Coordinateurs du groupe de travail STAPH
Laboratoire de Statistique et Probabilités

boudou@cict.fr, cardot@toulouse.inra.fr, ferraty@cict.fr, romain@cict.fr
sarda@cict.fr, vieu@cict.fr, viguier@cict.fr

Créé voilà bientôt 4 ans au sein du Laboratoire de Statistique et Probabilités de Toulouse, notre groupe de travail STAPH qui est (faut-il le rappeler ?) axé sur les différents aspects fonctionnels de la Statistique moderne, a continué cette année ses activités sur le rythme désormais habituel de (environ) une séance chaque quinze jours.

Bien sûr, la plupart de nos séances ont porté sur des thèmes qui nous sont chers ici à Toulouse (statistique des opérateurs, modèles pour variables fonctionnelles et estimation non paramétrique), mais nous avons aussi consacré quelques séances à des questions *a priori* plus éloignées de nos préoccupations immédiates (réseau de neurones, dépendance entre tableaux, statistique spatiale, choix de modèles ...) qui ont été autant d'occasion de tisser ou renforcer des liens scientifiques avec d'autres équipes de statisticiens. Enfin, nous avons essayé de réaliser un équilibre entre des séances centrées sur des applications statistiques et d'autres plus axées sur les aspects théoriques, avec aussi d'autres séances de nature plus philosophique. Ce souhait de ne jamais privilégier les aspects théoriques ou les aspects appliqués des choses, mais plutôt de les faire inter-agir en permanence, est inhérent à notre conception de la recherche en Mathématiques Appliquées et continuera à nous guider dans nos activités futures.

Finalement, comme l'année dernière, nous avons clôturé cette année universitaire, par la tenue de la deuxième édition de nos Journées de Statistique Fonctionnelle et Opératorielle, qui ont réuni une quarantaine de participants (dont presque une dizaine venus de l'étranger) et qui ont été l'occasion à la fois de diffuser/partager des développements récents sur ce thème et de nouer/renforcer des collaborations.

[1] Désormais toutes nos activités sont accessibles sur la page web

$$http://www.lsp.ups-tlse.fr/Fp/Ferraty/staph.html$$

Par ailleurs, les sommaires de nos activités passées sont présentés à la fin de ce document.

Outre ces aspects directement liés aux activités de notre groupe, nous avons constaté avec plaisir l'engouement de plus en plus affiché par la comunauté statistique internationale pour ces aspects fonctionnels de notre discipline, en particulier dans notre laboratoire. C'est donc tout naturellement que nous poursuivrons nos activités l'année prochaine, en nous souhaitant toujours plus ouverts sur des thèmes connexes ou sur de nouvelles collaborations.

En remerciant tous nos intervenants et participants et en leur souhaitant de bonnes vacances,

Toulouse le 1er Juillet 2003

Un exemple d'analyse de données fonctionnelles en océanologie : le cas de l'Etang de Berre

David NERINI*
En collaboration avec Claude MANTÉ

* Adresse pour correspondance :
UMR LOB 6535 CNRS, Campus de Luminy, Case 901
Centre d'Océanologie de Marseille
13288 Marseille Cedex 09
e-mail : nerini@com.univ-mrs.fr

Exposé du 04 Novembre 2002

Résumé

Nous proposons une méthode d'analyse fonctionnelle de séries spatio temporelles de variables physico-chimiques faiblement échantillonnées sur une grille variable. Le développement méthodologique est effectué dans le cadre spécifique de la surveillance en continu d'un écosystème lagunaire : l'Etang de Berre. Ce milieu initialement marin est fortement perturbé par des rejets d'eau douce d'origine anthropique (centrale hydroélectrique de St Chamas). Pour identifier et quantifier les processus qui mènent à une amélioration de la qualité des eaux de l'étang, on cherche à résumer la dynamique de la colonne d'eau en un petit nombre d'états caractéristiques. Ils sont déterminés en regroupant les journées qui se ressemblent le plus du point de vue de l'évolution simultanée de profils horaires de p variables physico-chimiques (température, salinité, oxygène dissous, ...). La détermination de cette typologie journalière des profils est à l'origine de la construction d'un modèle statistique de prévision à 24 heures par arbre de classification (Nerini *et al*, 2000).

Depuis 1996, les profils de variables physico-chimiques sont échantillonnés sur 5 niveaux de la colonne d'eau à l'aide d'une station automatique de prélévements. La première étape de la méthode consiste à considérer chacun des profils comme une fonction échantillonnée en 5 points. Ces fonctions sont estimés à l'aide de splines cubiques dont le paramètre de lissage est déterminé relativement à la variance des appareils de mesure (Wahba 1990). Cette approche permet de s'af-

franchir des problèmes liés au faible taux d'échantillonnage.

Nous montrons ensuite qu'il est possible d'exprimer chaque profil de manière fonctionnelle dans une base de polynômes de Legendre : les paramètres des polynômes sont estimés, pour chaque profil, à l'aide d'un rééchantillonnage optimal en des points choisis à partir du lissage spline. Ceci permet de déterminer la taille optimale de la base polynômiale et d'obtenir des estimateurs des paramètres des polynômes de manière plus aisée que par la méthode des moindres carrés, sans sortir du cadre fonctionnel (Gautschi, 1997, et Laurent, 1972).

Une ACP sur les profils bruts montre que la dynamique du système est nettement conditionnée par un effet saisonnier : deux profils de même forme seront considérés différents s'ils sont fortement décalés en moyenne alors qu'ils isolent le même phénomène (mélange du au vent, stratification de la colonne d'eau, ...). La détermination d'une typologie originale des profils ne doit pas uniquement dépendre de cette forte tendance saisonnière. Une ACP fonctionnelle (Ramsay and Silverman, 1997, et Besse and Ramsay, 1986) des profils est alors réalisée dans la base des polynômes de Legendre, en équilibrant l'effet de la moyenne (matrice des premiers coefficients de Fourier) de celui lié à la forme des profils (autres coefficients de Fourier).

La dernière étape concerne la classification journalière de la masse d'eau. Les profils successifs sont regroupés par unité de 24 heures (une journée) et une classification sur tableau de distances orbitales est réalisée dans l'espace des premiers facteurs de l'ACP (Manté, 1989). Nous montrons que cette méthode appliquée sur des données de salinité, permet de suivre la dynamique quotidienne de la masse d'eau comme une succession dans le temps d'états caractéristiques, chacun associé à un épisode physique particulier (mélange par le vent, mise en place d'une stratification verticale du bassin, ...).

Références

Besse Ph., Ramsay (1986). Principal components analysis of sampled functions, *Psychometrika*, **51**, 285-311.

Gautschi W.,(1997).*Numerical Analysis, An Introduction*, Birkhaüser, Boston.

Laurent P. J. (1972). *Approximation et optimisation*, Hermann, Paris.

Manté C. (1989). ACP d'un processus multiple non stationnaire : application à des données météorologiques, *Statistique et analyse de données*, **14**(2), 25-53.

Nerini D., Manté C., Durbec J. P. (2000). Forecasting physicochemical variables by a classification tree method. Application to the Berre lagoon, Acta Biotheoretica.

Ramsay J. O., Silverman B. W.(1997). *Functional data analysis*, Springer Verlag.

Wahba G. (1990). *Spline models for observational data*, SIAM, Philadelphia.

12

.

Sur la decomposition en valeurs singulieres complexes et applications

Simplice DOSSOU-GBETE *
en collaboration avec Antoine de FALGUEROLLES

* Adresse pour correspondance :
Laboratoire de Mathematiques Appliquees
Universite de Pau et Pays de l'Adour
64012 Pau
e-mail : simplice.dossou-gbete@univ-pau.fr

Exposé du 18 Novembre 2002

Résumé

Références

Besse Ph., Ramsay (1986). Principal components analysis of sampled functions, *Psychometrika*, **51**, 285-311.

Gautschi W.,(1997).*Numerical Analysis, An Introduction*, Birkhaüser, Boston.

.

ACP dans le domaine des fréquences appliquée à la météorologie

Olivier CAUMONT

Ecole de la Météorologie de Toulouse
e-mail : olivier.caumont@meteo.fr ou olicau@wanadoo.fr

Exposé du 25 Novembre 2002

Résumé

Soit X_n une série p-dimensionnelle d'éléments de L_p vérifiant :
- X_n est stationnaire au sens strict, ergodique et centrée,
- la fonction d'autocovariance est absolument sommable,
- pour chaque fréquence, les valeurs propres de la fonction de densité sont simples.

L' Analyse en Composantes Principales dans le domaine des Fréquences (ACPF) de X_n décrite ici produit une série Y_n q-dimensionnelle ($q < p$) d'éléments de $L^2_{\mathbb{C}^q}(\Omega, \mathcal{A}, P)$. La série Y_n est la série filtrée de X_n qui en constitue un résumé optimal du point de vue de la reconstitution de X_n à partir de Y_n par le filtre inverse.

Les fondements théoriques de ce type de méthode s'appuient sur les travaux de Boudou et Dauxois (1988) et Boudou (1995).

Les propriétés de cette méthode la rendent susceptible d'applications en météorologie. Il est ainsi possible de compresser des données, comparer des séries et détecter des périodicités particulières dans X_n. L'objectif de cete exposé est de présenter certaines de ces applications.

Références

Boudou, A. (1995). Mise en oeuvre de l'analyse en composantes principales d'une serie stationnaire multidimensionnelle, *Publications de l'Institut de Statistique de l'Université de Paris*, **XXXIX**, fasc. 1.

Dauxois, J. et Boudou, A. (1988). Analyse des séries multidimensionnelles stationnaires, *Publications du Laboratoire de Statistique et Probabilité de Toulouse*, **02-88**.

Consistance de l'estimation des paramètres d'un modèle non linéaire sur données fonctionnelles

Fabrice ROSSI

LISE/CEREMADE
UMR CNRS 7534, Paris Dauphine
e-mail : rossi@ufrmd.dauphine.fr

Exposé du 02 Décembre 2002

Résumé

Quand on généralise le perceptron multi-couches (PMC) au cas des données fonctionnelles (Rossi et Conan, 2002, et Conan et Rossi, 2002), on obtient un modèle non linéaire semi-paramétrique de la forme suivante :

$$Y = H\left(w_0, \int F_1(w_1,.)Xd\mu, \ldots, \int F_k(w_k,.)Xd\mu\right) + \mathcal{E}, \tag{1}$$

où μ désigne une mesure finie (en général une probabilité) sur $\mathbb{R}^p$, X une v.a. à valeurs dans $L^p(\mu)$, F_j une fonction de $W_j \times \mathbb{R}^p$ dans $\mathbb{R}$, telle que $F(w_j,.) \in L^q(\mu)$, $\mathcal{E}$ un bruit centré, Y une variable aléatoire à valeurs dans $\mathbb{R}t$ et H une fonction de type perceptron multicouches. On peut noter la ressemblance du modèle avec le SIR fonctionnel proposé par (Ferré et Yao, 2002). Il y a cependant deux différences majeures : H est un modèle semi-paramétrique et l'estimation des paramètres se fait en bloc ce qui implique une optimisation non linéaire. En effet, on se donne une fonction de coût qui mesure l'adéquation entre Y et le modèle (notée c), par exemple la norme quadratique, et on cherche les paramètres qui minimisent :

$$\mathcal{D}(w_0, \ldots, w_k) = E\left(c\left(Y, H\left(w_0, \int F_1(w_1,.)Xd\mu, \ldots, \int F_k(w_k,.)Xd\mu\right)\right)\right).$$

Dans un premier temps, on estime $\mathcal{D}(w_0, \ldots, w_k)$ par

$$\widehat{\mathcal{D}}(w_0,\ldots,w_k)_n = \frac{1}{n}\sum_{i=1}^{n} c\left(Y_i, H\left(w_0, \int F_1(w_1,.)X_i d\mu, \ldots, \int F_k(w_k,.)X_i d\mu\right)\right), \tag{2}$$

où les (X_i, Y_i) sont des couples de même loi que (X, Y). Dans le cas indépendant, la consistance de l'estimateur $\hat{w}_n = \arg\min \widehat{\mathcal{D}}(w_0,\ldots,w_k)_n$ s'obtient simplement par une loi des grands nombres uniforme (Andrews, 1987), avec des hypothèses raisonnables (en particulier X est à valeurs dans $L^p(\mu)$, sans autre restriction).

Quand les fonctions observées (les X_i) ne sont pas connues parfaitement, le problème se complique. Dans Rossi *et al.* (2002), et dans Rossi et Conan (2002), on choisit un *design* aléatoire, c'est-à-dire des suites de v.a. à valeurs dans $\mathbb{R}^p$, les T_{ij}, qu'on suppose identiquement distribuées (la mesure de probabilité correspondante étant μ). En pratique, on travaille sur

$$\widetilde{\mathcal{D}}(w_0,\ldots,w_k)_n^m =$$

$$\frac{1}{n}\sum_{i=1}^{n} c(y_i, H(w_0, \frac{1}{m_i}\sum_{j=1}^{m_i} F_1(w_1,t_{ij})x_i(t_{ij}), \ldots, \frac{1}{m_i}\sum_{j=1}^{m_i} F_k(w_k,t_{ij})x_i(t_{ij}))),$$

avec $m = \sup_i m_i$. On obtient de nouveau la consistance de l'estimateur des paramètres, mais au prix d'hypothèses très fortes sur X (à valeurs dans un espace compact de fonctions continues définies sur un compact de $\mathbb{R}^p$).

Dans Conan et Rossi (2002) et (2002b) et, on aborde le problème différemment, en projetant les fonctions observées sur les premières fonctions d'une base topologique de $L^2(\mu)$. On obtient de nouveau la consistance des estimateurs, mais avec les mêmes hypothèses restrictives sur X (toujours dans le cadre d'un *design* aléatoire). On peut aussi appliquer des résultats de Abraham *et al.* (2000) pour obtenir d'autres hypothèses sur X, elles aussi assez restrictives (X à valeurs dans l'ensemble des fonctions continues à variations bornées définies sur un intervalle de $\mathbb{R}$).

Nous travaillons à l'amélioration des résultats obtenus dans les directions suivantes :

- le cas traité pour l'échantillonnage (*design* aléatoire) est assez défavorable (hypothèses fortes) et n'est pas toujours très satisfaisant (les T_{ij} sont i.i.d). Nous travaillons actuellement sur un *design* déterministe ;
- l'extension au cas non indépendant pour (Y_i, X_i) est indispensable pour traiter des séries temporelles. De plus, nous avons obtenu des résultats sur l'extension du modèle à une réponse fonctionnelle (dans le cas i.i.d.) et une modélisation auto-régressive non linéaire pourrait être intéressante ;

– l'obtention d'une distribution asymptotique pour les paramètres du PMC fonctionnel pourrait permettre la simplification du modèle par des tests de nullité de certains coefficients.

Références

Abraham, C., Cornillon, P.A. et Matzner-Lober, E. (2001). Unsupervised curve clustering using B-splines, *Technical Report, ENSAM–INRA–UM II–Montpellier*, **00-04**.

Andrews, D.W. (1987). Consistency in nonlinear econometric models : A generic uniform law of large numbers, *Econometrica*, **55**(6), 1465-1471.

Conan-Guez, B. et Rossi, F. (2002). Approche régularisée du traitement de données fonctionnelles par un perceptron multi-couches. In *Actes des neuvièmes journées de la SFC*, Toulouse, France, Septembre 2002, 169-172.

Conan-Guez, B. et Rossi, F. (2002b). Multi-layer perceptrons for functional data analysis : a projection based approach. In José R. Dorronsoro, editor, *Artificial Neural Networks – ICANN 2002*, Madrid, August 2002, Springer, 667-672.

Ferré, L. and Yao, A.F. (2000). Functional sliced inverse regression analysis. *Technical Report, Laboratoire de Statistique et Probabilités, Université Paul Sabatier, Toulouse*, **0248-3289**.

Rossi, F. et Conan-Guez, B. (2002). Modélisation supervisée de données fonctionnelles par perceptron multi-couches. In *Actes des neuvièmes journées de la SFC (conférence invitée)*, Toulouse, France, Septembre 2002, 93-100.

Rossi, F., Conan-Guez, B. and Fleuret, F. (2002). Functional data analysis with multi layer perceptrons. In *Proceedings of IJCNN 2002 (WCCI 2002)*, Honolulu, Hawai, USA, May 2002, 2843-2848.

Rossi, F., Conan-Guez, B. and Fleuret, F. (2002). Theoretical properties of functional multi layer perceptrons. In *Proceedings of ESANN 2002*, , Bruges, Belgium, April 2002, 7-12.

20

.

Méthodes multivariées pour caractériser la distribution spatiale et saisonnière de la population de merlan de la mer Celtique

Marion VERDOIT

IUT STID, Domaine d'Auriac 11000 Carcasonne
e-mail : mverdoit@yahoo.fr

Exposé du 13 Janvier 2003

Résumé

La plupart des populations de poissons benthiques et démersales présentent des caractéristiques spatiales et saisonnières liées à leur cycle de vie annuel particulièrement à travers les migrations des zones de reproduction vers les zones de nutrition. Par conséquent, on observe souvent une séparation spatiale et/ou saisonnière entre les adultes et les juvéniles (Nicholsky, 1968). Malgré cela, la plupart des modèles de dynamique des populations ne sont pas structurés spatialement et temporellement. De tels modèles seraient cependant nécessaires pour évaluer l'impact de mesures de gestion spatio-temporelles, qui sont de plus en plus évoquées comme outil de gestion supplémentaires afin de réduire la surexploitation. La construction de tels modèles nécessite une meilleure connaissance des cycles biologiques des populations et des patterns de distribution spatiaux et saisonniers associés. Comme pour beaucoup d'autres espèces, les principaux traits du cycle de vie du merlan (Merlangius merlangus, L. 1758) sont bien identifiés, mais la connaissance des zones préférentielles des stades démographiques, et les saisons associées est généralement approximative. En général, les enregistrements des campagnes scientifiques et de la pêche commerciale sont les deux sources de données disponibles pour déterminer les distributions spatiales et saisonnières des différents stades démographiques de populations exploitées. Les campagnes scientifiques n'ont pas une couverture saisonnière suffisante pour permettre de délimiter les périodes qui caractérisent les principaux stades démographiques. Au contraire les données commerciales représentent beaucoup d'information avec une bonne couverture spatiale et temporelle mais souvent ne fournissent pas d'information sur les plus jeunes individus de taille inférieure à la taille légale. L'objectif de cette étude est d'utiliser de façon complémentaire les caractéristiques

des données scientifiques et commerciales afin de déterminer les distributions spatio-temporelles du recrutement et du stock reproducteur d'une population exploitée. L'analyse est basée sur des méthodes descriptives multivariées d'ordination qui tiennent compte des proximités spatiales et temporelles et des méthodes de classifications. On présente brièvement l'approche utilisé qui est appliquée à la population de merlan de mer Celtique, une importante espèce commerciale pour la pêcherie française. Les données des campagnes scientifiques de chalutages expérimentaux provenant de 2 sources. : des données françaises issues des campagnes de l'Ifremer et des données anglaises issues des campagnes programmées par le Cefas. La saison de la campagne est pour les années considérées l'automne pour l'Ifremer et le printemps pour le Cefas. Les années considérées vont de 1997 à 1999 pour les données françaises et sont les années 1998 et 1999 pour les données anglaises. Les CPUE (captures par unité d'effort) scientifiques de merlan sont calculées comme les captures en poids par âge par heure de chalutage. Les données commerciales de captures et d'effort ont été extraites de la base de statistiques de pêche de l'Ifremer. L'échelle spatiale et celle du rectangle statistique CIEM (Conseil international pour l'Exploration de la Mer). La capture par rectangle n'est connue que si le bateau ne pêche qu'exclusivement dans un seul rectangle. Ainsi nous avons analysé ces marées spécifiques appelées marées monorectangles. La période considérée s'étend de 1993 a 1997. Du fait que la correspondance entre les catégories commerciales et les stades démographiques de la population n'est pas très marquée, les CPUE par catégorie commerciale ont été converties en CPUE par groupes d'âge en utilisant les données d 'échantillonnages commerciaux. Finalement nous avons analysé des données de CPUE / rectangle /mois/âges. Les méthodes d'analyses sont les suivantes : Tout d 'abord pour le calcul des indices de CPUE : l 'échelle temporelle utilisée est celle du mois. Concernant l'échelle spatiale, pour être cohérent avec la résolution spatiale des données commerciales qui celle du rectangle CIEM, les données scientifiques ont été moyennées par rectangle CIEM. Les CPUE ont par ailleurs été moyennées sur les années pour chaque type de données. La méthode d'analyse est une typologie qui consiste pour chaque type de données en une Analyse en Composantes Principales (ACP) normée suivie d'une Classification Hiérarchique Ascendante (CHA). Dans l'ACP, les individus sont les rectangles CIEM, dans le cas des données scientifiques, et les rectangles-mois dans le cas des données commerciales. Les variables sont les classes d'âge de merlan dans les deux cas. Chaque individus est décrit par un jeu de CPUE par groupe d'âge (en kg par h de chalutage). Dans un deuxième temps la procédure de Classification Hiérarchique Ascendante permet de grouper les individus en classes selon l'abondance de chaque groupe d'âge, en utilisant leurs coordonnées sur les axes principaux issus de l'ACP. Du fait qu'il est probable d'observer de la corrélation spatiale et temporelle (Legendre, 1993 ; Moran, 1948 ; Cliff and Ord, 1973), nous avons effectué en plus de l'ACP conventionnelle, une ACP de contiguïté (Thioulouse et al., 1995 ; Le Foll, 1982 ; Wartenberg, 1985 ; Geary, 1954 ; Lebart, 1969) qui tient compte des proximités

spatiales et temporelles. Cette méthode consiste à définir une relation de voisinage entre les individus. Basée sur cette relation, la variance totale du tableau de données est décomposée en une première composante liée aux individus voisins (appelée variance locale) et une seconde liée aux individus qui ne sont pas voisins. En plus des analyses locale et globale, une analyse totale et également réalisée. Cette analyse est similaire à une ACP classique dans laquelle les individus sont pondérés par leur nombre de voisins. L'intérêt de ces analyses est de comparer les structures observées aux échelles locale et globale et celles observées sans prendre en compte une échelle particulière (l'analyse totale). Résultats de l'ACP classique sur les données scientifiques : Dans le cas des données scientifiques on ne présente que les résultats d'une ACP classique. Dans les deux cas (automne et printemps), le premier axe indique un effet taille, opposant les rectangles présentant de fortes abondances pour tous les groupes d'âge, des rectangles à faibles abondances. Les axes suivants sont expliqués par des groupes d'âge particuliers. Pour les deux tableaux de données, les partitions issues de la CHA sélectionnées correspondent à 4 classes. Dans les 2 cas, la première classe correspond à des rectangles de plus faibles abondances pour tous les groupes d'âge en comparaison aux autres rectangles, on a nommé la classe " merlans rares ". La distribution spatiale des rectangles de chaque classe a été représenté sur une carte. Ainsi, pour chaque jeu de données, les cartes montrent que de fortes abondances de merlans sont trouvées uniquement au nord de la mer Celtique (au nord de 50°N). Dans le reste de la mer Celtique, le merlan n'est pas absent mais surtout rare. Ensuite on observe des différences entre les campagnes d'automne et du printemps qui illustrent le caractère saisonnier de la distribution spatiale. Dans le cas des données du printemps, la deuxième classe est caractéristique de très fortes abondances du groupe d'âge 2, tandis que la classe 3 correspond à un rectangle avec les groupes d'âge 2 à 5+, mais non caractéristiques. La classe 4 est caractérisée par des pics d'abondance pour tous les groupes d'âge y compris l 'âge 1. Ainsi, en mars, de très fortes abondances des groupes d'âge 2 à 5+ sont observées dans le nord de la mer Celtique, en relation avec la reproduction. Pour le groupe d'âge 1, de fortes abondances sont également trouvées dans le Canal de Bristol à la même période de l'année. En automne, le merlan apparaît plus largement distribué, bienqu'il se situe toujours dans le nord de la mer Celtique. La classe 2 comprend des rectangles avec de plus fortes abondances des groupes d'âge 0 et 1 qui sont concentrés dans le centre de la mer Celtique sur la région des Smalls et à l'Est du Canal de Bristol. Les classes 3 et 4 sont caractéristiques des adultes, on voit que des adultes peuvent être présents dans cette zone centrale mais en moindre abondance. Ces derniers sont plus abondants dans les rectangles situés autour de cette zone. Résultats issus de l'ACP de contiguïté sur les données commerciales : Les projections des variables sur les trois premiers axes factoriels, pour les analyses totale, locale et globale sont présentés. Les projections pour l'analyse totale sont très similaires à celles obtenues à partir d'une ACP classique. Comme pour les données scientifiques, le premier axe résulte d'un effet taille, pour les 3 ana-

lyses. Le groupe d'âge 1 ne semble pas être corrélé aux groupes d'âge supérieurs et est bien représenté sur les axes 2 des analyses totale et locale. Dans l'analyse globale l'axe 2 oppose les groupes d'âge 2 et 5+. Des différences entre les plus vieux groupes d'âge apparaissent sur les axes 2 et 3. Dans l'analyse totale, l'axe 2 oppose les plus vieux groupes d'âge (5+ et 4) aux plus jeunes groupes d'âge (1 et 2). L'analyse totale montre des oppositions immatures/matures et jeunes adultes/vieux adultes. Ces oppositions sont accentuées dans l'analyse locale. Dans l'analyse globale, l'axe 3 sépare les groupes d'âge 2 et 3, mais cet axe n'explique qu'une très faible part de la variabilité du jeu de données. Comme précédemment, les résultats de la CHA sont représentés sur des cartes mensuelles. La compositions des classes des 3 analyses sont similaires ce qui montre que les structures spatiales et saisonnières sont similaires aux échelles locale et globale. On ne présente donc que les cartes relatives à l'analyse totale pour les 12 mois. Comme pour les données scientifiques, la CHA amène à 4 classes. Une classe où le merlan est rare, une classe caractéristique des vieux adultes, une autre des jeunes adultes et enfin une 4ème classe caractéristique de l'âge 1 donc des immatures. L'analyse des données commerciales indique 4 périodes dans l'année associées à des zones particulières. On observe de fortes abondances des adultes les plus vieux durant la période de reproduction (Janvier à Avril) dans les régions côtières (Sud de l'Irlande et Cornouaille). Des abondances intermédiaires d'adultes sont observées dans la même zone en Mai et Juin. Durant la troisième saison, de Juillet à Septembre, tous les groupes d'âges sont observés sur la zone des "Smalls", où en particulier, de fortes abondances du groupe d'âge 1 sont trouvées en relation avec le recrutement. Durant le dernier trimestre, la distribution des adultes est plus dispersée, surtout entre 51°N et 52°N. Le groupe d'âge 1 est principalement trouvé de juillet à septembre et en moindre abondance en décembre. Cependant les résultats du groupe d'âge 1 doivent être pris avec précaution, du fait des variations potentielles des taux de rejets. Le merlan mature doit se disperser sur toute la zone après la saison de reproduction, du fait que de plus fortes abondances peuvent être observées au nord de 50°N (août et septembre). Les cartes mensuelles issues de l 'analyse des CPUEs commerciales sont en relative conformité avec les connaissances préalables du cycle de vie du merlan, qui postule entre autre que les merlans matures se concentrent au nord de la mer Celtique en zones côtière pour se reproduire. L'apport de ces analyses par rapport à la littérature est de décrire explicitement les rectangles statistiques selon les concentrations des différents groupes d'âges. Cependant, les compartiments spatiaux ne sont pas aussi précis que nous aurions pu l'espérer. Temporellement, on observe des mois tels que juin et décembre qui correspondent à des distributions transitoires. Spatialement, il semble que l'échelle du rectangle CIEM soit trop grande. Les analyses de contiguïté sur les données commerciales donnent des résultats très similaires entre les 3 analyses, montrant ainsi que les structures observées sont consistantes selon les échelles En conclusion, cette étude basée sur une analyse statistique de donnée fournit un moyen rationnel, de construire un modèle spatio-temporel de

dynamique de population. Ce modèle est en cours de développement pour tester des mesures de gestion telles que des fermetures de zone et de saisons. Ensuite, les résultats doivent être comparés avec des résultats d'expériences de marquage de façon à estimer des coefficients de migration entre les compartiments. Finalement, les compartiments spatiaux et les saisons définis ici peuvent constituer une base pour définir des estimations d'abondance qui sont explicitement spatiales et saisonnières, par exemple à partir de données de marées ciblées, des données de captures scientifiques ou des données acoustiques.

Références

Cliff, A. D. and J. K. Ord, 1973. *Spatial autocorrelation. Pion, London*, 178 pp.

Geary, R. C., 1954. The contiguity ratio and statistical mapping. *The Incorporated Statistician*, **5** (3) : 115-145.

Lebart L., 1969. Analyse statistique de la contiguïté. *Publication de l'Institut de Statistiques de l'Université de Paris*, **28** : 81-112.

Le Foll, Y., 1982. Pondération des distances en analyse factorielle. *Statistique et Analyse de données*, **7** : 13-31.

Legendre, P., 1993. Spatial autocorrelation : trouble or new paradigm ? *Ecology*, **74**, 6 : 1659-1673.

Moran, P. A. P., 1948. The interpretation of statistical maps. *Journal of the Royal Statistical Society B*, **10** : 243-251.

Nicholsky, G. V., 1968. *The ecology of fishes*. Academic Press London and New York, 351 pp.

Thioulouse, J., D. Chessel, S. Champely, 1995. Multivariate analysis of spatial pattern : a unified approach to local and global structures. *Environmental and Ecological Statistics*, **2** : 1-14.

Wartenberg, D., 1985. Multivariate spatial correlation : a method for exploratory geographical analysis. *Geographical Analysis*, **17** : 263-283.

.

Contribution à l'étude des modèles de régression pour variables aléatoires fonctionnelles

Aldo GOIA

Università degli Studi di Torino
Dipartimento di Matematica, Italie
e-mail : goia@econ.unito.it ou algoia@tin.it

Soutenance de Doctorat de l'Université Paul Sabatier
en co-tutelle avec l'Université de Turin

22 Janvier 2003

Mots Clés : Modèles fonctionnels de régression - Régression non paramétrique - Estimateur à noyau - Variables aléatoires fonctionnelles mélangeantes - Prévision - Modèle linéaire fonctionnel - Test de permutation - Pseudo-rapport de vraisemblance.

Résumé

Ce travail est une contribution à l'étude des modèles fonctionnels de régression dans lesquels la variable réponse Y est réelle et le régresseur X est une fonction aléatoire définie sur un sous-ensemble compact de $\mathbb{R}$. Ainsi, le modèle s'écrit : $Y = \Psi(X) + \epsilon$, où $\Psi(.)$ est un opérateur à valeurs réelles et ϵ une v.a. réelle, centrée et de variance finie, que nous supposons non corrélée avec X.

Notre travail est axé sur deux sujets principaux. Un premier domaine d'étude concerne l'analyse d'un estimateur de $\Psi(.)$ de type Nadaraya-Watson lorsque le modèle de régression est non paramétrique et les observations sont dépendantes. Sous des conditions très générales, on donne des résultats de convergence presque complète pour cet estimateur. Cette étude s'applique à la prévision non paramétrique des séries chronologiques : nous nous intéressons plus particulièrement à la prévision d'une série de données de nature économique.

Un deuxième sujet d'étude nous a amené à définir des tests de nullité pour l'opérateur $\Psi(.)$ dans le cas où il est linéaire : outre un test de permutation basé sur l'opérateur de covariance entre le régresseur et la variable réponse, on présente un test basé sur la statistique du pseudo-rapport de vraisemblance dans le cas d'erreurs gaussiennes, puis dans un cas plus général. L'étude est complétée par des simulations qui permettent d'évaluer le niveau et la puissance de ces tests.

Références

Cardot, H., Goia, A., Sarda, P. (2002). Testing for no effect in functional linear regression models, some computational approaches. Soumis à *Communications in Statistics : Simulations and Computation*.

Ferraty, F., Goia, A., Vieu, P. (2002a). Functional nonparametric model for time series : a fractal approach for dimension reduction. *Test*, **11 (2)**, 317-344.

Ferraty, F., Goia, A., Vieu, P. (2002b). Lebart L., 1969. Analyse statistique de la contiguïté. *Comptes Rendus de l'Académie des Sciences de Paris*, Ser. I, **334** : 217-221.

Ferraty, F., Goia, A., Vieu, P. (2002c). *Statistica funzionale. Modelli di regressione non parametrici*. Franco Angeli, Milano. Goia, A. (2000).

Goia, A. (2000). Techniques non-paramétriques pour l'étude de données longitudinales fonctionnelles. *Quaderno n. 4, Serie A*. Dipartimento di Statistica e Matematica applicata alle scienze umane "Diego de Castro". Torino.

Analyse de la dépendance simultanée entre deux ensembles

Gabriel KISSITA

Institut Supérieur de Gestion
Université Marien Ngouabi
BP 15020, Brazzaville, (CONGO)
e-mail : gakissita@yahoo.fr

Exposé du 27 Janvier 2003

Résumé

L' Analyse Concor proposée par Lafosse et Hanafi [1] recherche la dépendance simultanée entre un tableau de référence et K autres tableaux. Une modification de cette méthode est proposée par Vivien et Sabatier [2] en vue de réaliser une régression PLS nommée ACIMO.

Dans le but de généraliser l 'Analyse de la CO-Inertie Multiple ý deux ensembles d'ensembles Vivien [3] propose l'Analyse de la CO-Inertie Multiple Généralisée (ACIMOG) , ici encore dans le but de faire de la régression PLS.

Nous proposons ici plutôt des généralisations de l'Analyse Concor que nous nommons respectivement Analyse Concor Généralisée (Concor G) et Analyse Concor Généralisée Non Linéaire (Concor GNL) . L'optimisation du critère associé à Concor GNL aboutit en effet à un problème non linéaire qui nécessite pour sa résolution un algorithme. Ensuite, on applique Concor G aux variables qualitatives, en obtenant ainsi à l'ordre un l'Analyse canonique sur variables qualitatives de Cazes et coll. [4].Enfin pour montrer l'intérèt pratique de nos méthodes, on les applique à un même jeu de données (vins de Cahors) .

Références

Lafosse, R. et Hanafi, M. (1997). Concordance d'un tableau avec K tableaux :

Définition de K+1 uples synthétiques. *Revue de Statistique Appliquée*, **XLV**(4), 111-126.

Vivien, M. et Sabatier, R. (2001). Une extension multi-tableaux de la régression PLS.*Revue de Statistique Appliquée*, **XLIX**(1), 31-54.

Vivien, M. (2002). *M - Approches PLS linéaires et non -linéaires pour la modélisation de multi-tableaux : théorie et applications*, ThËse 3 décembre 2002 , Université Montpellier I.

Cazes et Coll. (1977). *Codage et analyse des tableaux logiques - Introduction ý la pratique des variables qualitatives*, Cahiers du B.U.R.O.

Optimal Classification Trees

Jarumir ANTOCH*
en collaboration avec Jan KLASCHKA et Petr SAVICKÝ

* Adresse pour correspondance :
Charles University, Sokolovska 83
CZ-186 75 Prague 8, République Tchèque
e-mail : antoch@karlin.mff.cuni.cz

Exposé du 3 Février 2003

Résumé

Les arbres de classification et de régression sont traditionnellement obtenus à l'aide d'une partition récursive, c'est-à-dire en procédant "de haut en bas", utilisant des partitions localement optimales. Utilisant la puissance actuelle des ordinateurs il semble que cette optimalité locale (ou partition récursive à un pas) puisse être remplacée par une optimisation globale. Pendant cet exposé il sera présenté deux algorithmes qui optimisent en suivant une procédure allant "de bas en haut". Des résultats d'expérience par simulations seront ensuite présentés, afin d'illustrer les comportements de ces algorithmes.

1. Introduction

Since the sixties, when the first tree-based methods appeared, recursive partitioning (binary or k-ary) has kept the position as a state-of-art method to search for tree-structured classification models, though it has been known, that "local" optimization of individual splits does not yield, in general, "globally" optimal trees. In their classical monograph on classification and regression trees (section 2.5.8), Breiman et al. (1984) advocate recursive partitioning (compare also with scepticism of Gelfand et al. (1991)) : *"At this stage of computer technology, an overall optimal tree growing procedure does not appear feasible for any reasonably sized data set."* The aim of this contribution is to show that with the *current* po-

wer of computer technology, optimal tree growing is possible *to some extent*. Two novel algorithms are presented that allow, when the number of predictors is not too large, to construct trees with the smallest classification error in training data within the class of all possible binary classification trees of a specified size. The size is specified in different ways in the two algorithms. Our limited experimental experience suggests that such trees may have good generalization properties for problems with a complex relationship between the dependent variable and the predictors. The work was inspired by both classical (namely CART by Breiman et al. (1984)) and recent developments as, e.g., Siciliano (1998) in the tree-based methodologies as well as by recent works on RAM-resident solutions to complex data-analytical problems Pijls and Bioch (1999).

2. Basics and notations

We consider tree-structured classification models with $P \geq 1$ binary $\{0,1\}$-valued predictors (independent variables) $X_1, \ldots, X_P$ and $K \geq 2$ classes $C_1, \ldots, C_K$. An extension to nonbinary predictors via, e.g., dummy variables is also possible.

Data of an individual case consist of a *predictor vector* $\mathbf{x} = (x_1, \ldots, x_P) \in \{0,1\}^P$ and a *class label* $C \in \{C_1, \ldots, C_K\}$. The set $\{0,1\}^P$ is the *predictor space*. Let *cubes* be such subsets $B = A_1 \times \cdots \times A_P$ of the predictor space which, for each index i, have either $A_i = \{0\}$, or $A_i = \{1\}$, or $A_i = \{0,1\}$. Further, let $\mathcal{B}_P$ denote the set of all 3^P cubes $B \subseteq \{0,1\}^P$. The *dimension* of the cube $B = A_1 \times \ldots \times A_P$ is the number of those indices i for which $A_i = \{0,1\}$. The cube B is a *subcube* of the cube B' if $B \subseteq B'$. Cubes of dimension 0 are singletons. For data set $\mathcal{L}$ and cube B, let $\mathcal{L}_B$ be the set of those cases from $\mathcal{L}$ whose predictor vectors belong to B.

The optimization algorithms will be described in terms of assigning trees to cubes. A binary classification tree for a cube B is a description of a partition of B into its subcubes, together with a function that is a constant from $\{C_1, \ldots, C_K\}$ on each of the used subcubes. For a tree T, the function is denoted by f_T and its domain by D_T. More specifically, a binary classification tree T is either

(i) a *trivial* tree, when D_T is a cube and $f_T \in \{C_1, \ldots, C_K\}$ is a constant function defined on D_T;

(ii) a *compound* tree, i.e. a pair of trees $T = (T_L, T_R)$, such that D_{T_L} and D_{T_R} are disjoint and $D_{T_L} \cup D_{T_R}$ is a cube. Then, $D_T = D_{T_L} \cup D_{T_R}$ and f_T equals f_{T_L} on D_{T_L} and f_{T_R} on D_{T_R}.

The tree diagram corresponding to a trivial tree consists of only one node, which is a leaf and also the root. The root of a tree diagram corresponding to a compound tree is a split on the predictor that is constant on the two subcubes and

nonconstant on the whole cube.

We say that a predictor vector $\mathbf{x} = (x_1, \ldots, x_P) \in B$ is classified by T to a class C, if $f_T(\mathbf{x}) = C$. The size of a trivial tree T is $|T| = 1$. For a compound tree $T = (T_L, T_R)$, let $|T| = |T_L| + |T_R|$. Note that $|T|$ is equal to the number of leaves of the tree diagram corresponding to T.

The set of all trees T for a cube B, i.e. trees satisfying $D_T = B$, will be denoted $\mathcal{T}(B)$. Let $\mathcal{T}$ be a shorthand for $\mathcal{T}(\{0,1\}^P)$. By $\mathcal{T}_m(B)$ (leaving argument B, when B is the whole predictor space) we denote the set of all trees for cube B, whose size is equal to m.

Let a data set $\mathcal{L}$, cube $B \in \mathcal{B}_P$ and a classifier $T \in \mathcal{T}(B)$ be given. Let $N_i > 0$ for $i = 1, \ldots, K$ be the number of those cases from $\mathcal{L}$ that belong to class C_i. Let $N_{ij}(B)$ for $i, j = 1, \ldots, K$ be the number of those cases from $\mathcal{L}_B$ that belong to class C_i, and are classified into C_j by T. For $i, j = 1, \ldots, K$ we denote by Z_{ij} the cost of classifying (i.e., misclassifying for $i \neq j$) one case from class C_i to C_j. Let $\pi_1, \ldots, \pi_K$ be the prior probabilities of classes. See, e.g., Breiman et al. (1984) for various modes of prior probabilities usage.) The misclassification cost $R(T|\mathcal{L}_B)$ of the classifier T on $\mathcal{L}_B$ is defined as

$$R(T|\mathcal{L}_B) = \sum_{i=1}^{K} (\pi_i / N_i) \sum_{j=1}^{K} Z_{ij} N_{ij}(B).$$

Note, that for $T = (T_L, T_R) \in \mathcal{T}(B)$, $T_L \in \mathcal{T}(B_L)$ and $T_R \in \mathcal{T}(B_R)$, the equality $R(T|\mathcal{L}_B) = R(T_L|\mathcal{L}_{B_L}) + R(T_R|\mathcal{L}_{B_R})$ holds.

3. Tree optimization algorithms

Traditional tree-growing by recursive partitioning consists of a search for the best split of the predictor space in the root, followed by a search for the best split of the obtained subcubes, etc. At the moment a split for a cube is chosen, the properties of further splits of the considered subcubes are typically unknown. For a rare example of an attempt to "look more steps ahead", see Friedman (1979).

The optimization approach proposed is characterized, instead, by searching for the best tree among *all possible trees* of some size. The algorithms build the optimal trees in a bottom-up manner. In fact, first to complete an optimal tree, we assign (as an intermediate step) an optimal tree to each cube in $\mathcal{B}_P$. At the moment a split of a cube is being chosen, the optimal trees for all subcubes of the cube are already known.

While optimizing a split, the algorithms can optimize the two candidate subtrees independently due to the following simple fact.$\mathcal{T}_m(B)$ by $T^* = (T_L, T_R)$

where $|T_L| = m_L$ and $|T_R| = m_R$. Let $B_L = D_{T_L}$ and $B_R = D_{T_R}$. Then T_L and T_R minimize $R(T|\mathcal{L}_{B_L})$ in $\mathcal{T}_{m_L}(B_L)$ and $R(T|\mathcal{L}_{B_R})$ in $\mathcal{T}_{m_R}(B_R)$, respectively.

3.1. Cost-complexity minimization (Algorithm I)

We have designed and implemented a novel algorithm that, based on a learning data set $\mathcal{L}$, can find, for a given *complexity parameter* $\alpha \geq 0$, a classification tree minimizing the *cost-complexity measure*

Proposition 1. Let B be a cube, $\mathcal{L}$ a data set and m an integer. Let the misclassification cost $R(T|\mathcal{L}_B)$ be minimized in $\mathcal{T}_m(B)$ by $T^* = (T_L, T_R)$ where $|T_L| = m_L$ and $|T_R| = m_R$. Let $B_L = D_{T_L}$ and $B_R = D_{T_R}$. Then T_L and T_R minimize $R(T|\mathcal{L}_{B_L})$ in $\mathcal{T}_{m_L}(B_L)$ and $R(T|\mathcal{L}_{B_R})$ in $\mathcal{T}_{m_R}(B_R)$, respectively.

3.2. Cost-complexity minimization (Algorithm I)

We have designed and implemented a novel algorithm that, based on a learning data set $\mathcal{L}$, can find, for a given *complexity parameter* $\alpha \geq 0$, a classification tree minimizing the *cost-complexity measure*

$$R_\alpha(T|\mathcal{L}) = R(T|\mathcal{L}) + \alpha|T|.$$

Note that the cost-complexity-based pruning in CART – see Breiman et al. (1984) – minimizes the same quantity. However, the search for the best tree in CART is confined to the set of *trees obtained by pruning from one specific tree*, while our algorithm looks for the tree minimizing the cost-complexity measure in the set of *all possible trees*.

Assume, we have $R_\alpha(T^*|\mathcal{L}) = \min_{T\in\mathcal{T}} R_\alpha(T|\mathcal{L})$, where $\alpha > 0$ and $|T^*| = m$. Then, at the same time, we have $R(T^*|\mathcal{L}) = \min_{T\in\mathcal{T}_m} R(T|\mathcal{L})$. Thus, by varying the α-values, we can calculate $\min_{T\in\mathcal{T}_m} R(T|\mathcal{L})$ for some values of m, along with finding the trees that realize the minima.

The trees minimizing $R_\alpha(T|\mathcal{L})$ for some α have to satisfy the following geometrical condition.

Proposition 2. For all m, if $\mathcal{T}_m$ is nonempty, let $\varphi(m) = \min_{T\in\mathcal{T}_m} R(T|\mathcal{L})$. Let H be the convex hull of the set $\{(m, y); \varphi(m) \text{ well-defined}, y \geq \varphi(m)\}$. Let T^* minimize $R(T|\mathcal{L})$ within $\mathcal{T}_m$. Then, there is an $\alpha \geq 0$ such that T^* minimizes $R_\alpha(T|\mathcal{L})$ within $\mathcal{T}$, if and only if the point $(m, \varphi(m))$ lies on the boundary of H.

Let us call an ordering of $\mathcal{B}_P$ a *subcube-consistent* order, if every cube is preceded by all its proper subcubes.

Algorithm I

Input : Description of the variables, data set $\mathcal{L}$, α.
Output : A tree T^* minimizing $R_\alpha(T|\mathcal{L})$.
Description :

- For each cube B in $\mathcal{B}_P$ containing at least one training case, where the cubes are taken in any subcube-consistent order, do :
  1. Choose a class $C \in \{C_1, \ldots, C_K\}$ such that $R_\alpha(T|\mathcal{L}_B)$ of the trivial tree T with $D_T = B$ and $f_T = C$, is minimized.
  2. For every pair B_L,B_R of disjoint subcubes of B, such that both contain at least one case from $\mathcal{L}$ and $B = B_L \cup B_R$ (if any exists), consider the tree $T = (T_L, T_R)$, where T_L and T_R are the trees assigned to B_L and B_R.
  3. Among the trees considered in the two preceding steps, choose a tree T which minimizes $R_\alpha(T|\mathcal{L})$ (breaking possible ties deliberately). Store the information about which tree has been selected. The tree is guaranteed to minimize R_α in $\mathcal{T}(B)$.
- After finishing the above loop, trace the tree T^* assigned to $B = \{0,1\}^P$ using the stored information. This is the output.

Note that by setting the parameter α close to zero, the tree minimizing the size among trees with the lowest possible misclassification cost is found.

3.3. The cost minimization for a given size (Algorithm II)

The second algorithm can find for any given $M \geq 1$ a sequence of trees minimizing the misclassification cost $R(T|\mathcal{L})$ on the learning data set $\mathcal{L}$ within the sets $\mathcal{T}_m$ for $m = 1, 2, \ldots, M$. Unlike the former algorithm, the new algorithm can yield trees that do not fulfill the condition from Proposition 2. We pay for this advantage with increased space and time complexity.

The algorithm is similar to Algorithm I. The difference is that Algorithm II selects and stores a cost-minimal tree for each of the sizes $1, 2, \ldots, M$, if they exist, for each cube with at least one training case.

The cubes are processed in the same order as by Algorithm I. The description differs only in parts 2 and 3, which should be replaced by :

2′. For every pair B_L,B_R of disjoint subcubes of B such that both contain at least one case from $\mathcal{L}$ and $B = B_L \cup B_R$ (if any exists) and every pair m_L, m_R such that $m_L + m_R \leq M$, consider tree $T = (T_L, T_R)$, where T_L and T_R are the trees assigned to B_L for size m_L and to B_R for size m_R, resp. The resulting tree is a candidate for size $m_L + m_R$.

3′. Use the tree from step 1 for size 1. For each $m = 2, 3, \ldots, M$, choose among the candidates for size m from step 2′ the tree minimizing $R(T|\mathcal{L})$ (breaking possible ties deliberately). Store the information about which tree has been selected for each size.

The M trees assigned to $\{0,1\}^P$ are the output of the algorithm.

3.4. Software and algorithmic complexity

Algorithms I and II have been implemented in C++ by Petr Savický. The space required by Algorithm I is $O(3^P K)$ with a moderate constant depending on the implementation. For Algorithm II, the space requirement is $O(3^P(K + 5M))$ with a similar constant. The time required by both algorithms is linear in the corresponding space requirement.

4. Conclusions

In some cases, bottom-up strategies can be more successful than the traditional top-down "one-step" search for the best split. In order to establish a method of more general applicability based on our algorithms, one has to look for various ways to find a heuristic compromise between the full tree optimization and the traditional top-down methods. Moreover, the simple construction/selection strategy using two halves of the training data should be supplemented by other well-known approaches as, e.g., cross-validation.

Acknowledgement : This research was supported by grants GAČR 201/00/1482 (for the second and third author) and GAČR 201/03/0945 (for the first author).

Références

Antoch, J. and Klaschka, J. (2002). *Classification and regression trees.* In : Encyklopedia EOLSS, eds. Shaarawi A. et al. UNESCO :Paris. (in print).

Antoch, J. and Mola, F. (2002). *Software for classification and regression tree based methods.* In : Encyklopedia EOLSS, eds. Shaarawi, A. et al. UNESCO :Paris. (in print).

Breiman, L. et al. (1984). *Classification and Regression Trees.* Belmont CA : Wadsworth.

Friedman, J.H. (1979). *A tree-structured approach to nonparametric multiple regression.* In : Lecture Notes in Mathematics **757**, eds. Gasser, T. and Rosenblatt, M., 5 – 22. Berlin : Springer-Verlag.

Gelfand S.B., Ravishankar C.S. and Delp, E.J. (1991). *An iterative growing and pruning algorithm for classification tree design.* IEEE Transactions on Pattern Analysis and Machine Intelligence **13**, 163 – 174.

Klaschka, J. and Antoch, J. (1996). *How to grow trees fast.* In : ROBUST'96, eds. Antoch, J. and Dohnal, G., 91 – 106. Praha :JČMF.

Klaschka, J., Siciliano, R. and Antoch, J. (1998). *Computational enhancements in tree growing methods.* In : Advances in Data Science and Classification, eds. Rizzi A. et al., 295 – 302. Heidelberg :Springer-Verlag.

Pijls, W. and Bioch, J.C. (1999). *Mining frequent itemsets in memory-resident databases.* Manuscript, `http ://www.few.eur.nl/few/people/pijls/`.

Quinlan, J.R. (1986). *Induction of decision trees.* Machine Learning **1**, 81-106.
Savický, P., Klaschka, J. and Antoch, J. (2000). *Optimal classification trees.* In : COMPSTAT'2000, Proceedings in Computational Statistics (eds. J.G. Bethelem & P.G.M. van der Heiden), 427 – 432. Heidelberg : Physica-Verlag.

Savický, P., Klaschka, J. and Antoch, J. (2001). *Optimal classification trees.* In : ROBUST'2000, eds. J., Antoch and G., Dohnal, 267 – 283. Praha : JČMF.

Siciliano R. (1998). *Exploratory versus decision trees.* In : COMPSTAT 98, Proceedings in Computational Statistics eds. R. Payne and P. Green, 113 – 124. Heidelberg : Physica-Verlag.

.

Kernel estimates of hazard functions and their application

Ivana HOROVÁ

Masaryk University
Department of Applied Mathematics
66295 Brno
République Tchèque.
e-mail : horova@math.muni.cz

Exposé du 24 Mars 2003

Résumé

In recent years the considerable attention has been paid to methods for analyzing data on events observed over time and to the study of factors associated with occurence rates for these events. In summarizing survival data, there are two functions of central interest, namely the survival and the hazard function. The well-known product - limit estimator of the survival function was proposed by Kaplan and Meier in 1958. A single sample of survival data may also be summarized through the hazard function which shows the dependence of the instaneous risk of on time. We focus on kernel estimate of hazard functions and their derivatives under random censoring based on the Nelson estimator in 1972. As far as the biomedical application is concerned the dynamics of the underlying curve is of a great interest. From this reason the attention is paid to the estimate of the points of the most rapid change. The application to the carcinoma data set is also presented.

Références

Granovsky B.L., Müller H.G. (1991). Optimizing Kernel Methods. A unifying Variational Principle. *Int. Stat. Review*, **59**, 3, 378–388.

Granovsky B.L., Müller H.G., Pfeifer C. (1995). Som Remarks of Optimal Kernel

Functions. *Statistics & Decision*, **13**, 101-116.

Horová I., Vieu P., Zelinka J. (2002). Optimal Choice of Nonparametric Estimates of a Density and of its Derivatives. *Statistics & Decision*, **20**, 355-378.

Isaacson E., Keller H.B. (1966). *Analysis of Numerical Methods*. John Wiley & Sons, Inc. New York, London, Sydney.

Kaplan E.I., Meier P.V. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the Am. Stat. Asoc.*, **53**, 282, 457-481.

Müller H.G., Wang J.L. (1990a). Nonparametric Analysis of Changes in Hazard Rates for Censored Survival Data : A Alternative Change–Point Models. *Biometrika*, **77**, 2, 305-14.

Müller H.G., Wang J.L. (1990b). Locally Addaptive Hazard Smoothing. *Prob. Th. Rel. Fields*, **85**, 523-538.

Müller H.G., Wang J.L. (1994). Hazard Rate Estimation under Random Censoring with Varying Kernels and Bandwidths. *Biometrics*, **50**, 61-76.

Nelson W. (1972). Theory and Applications of Hazard Plotting for Censored Data. *Technometrics*, **14**, 945-966.

Stoer J., Bulirsch R. (1980). *Introduction to Numerical Analysis*. Springer-Verlag, New York, Heidelberg, Berlin.

Tanner M.A., Wong W.H. (1983). The Estimation of the Hazard Function from Randomly Censored Data by the Kernel Method. *The Annals of Statistics*, **11**, 3, 989-993.

Tanner M.A., Wong W.H. (1984). Data-Based Nonparametric Estimation of the Hazard Function with Applications to Model Diagnostis and Exploratory Analysis. *Journal of the Am. Stat. Assoc.*, **79**, 35, 174-182.

Centrages et disparités : peut-on "négocier" avec les données ?

Jacques VANPOUCKE

Université Paul Sabatier
Toulouse
e-mail : vpk@cict.fr

Exposé du 7 Avril 2003

Mots clés : Centrages, paramètres de "dispersion", statistiques d'ordre, PMC (Pitman), EDA (Tukey, sans c !), robustesse, fonctions d'influence.

Résumé

On introduit brièvement la problématique de la description de tableaux de données, et de l'usage de l'estimation qui y est fait.

Une dialectique s'instaure entre la tendance à l'universalisme, inhérente aux techniques de centrage (et plus généralement, à toute pratique modélisatrice) et la nécessaire prise en compte des disparités dont la "réalité" des données se fait le témoin. Sur le champ de la "gestion" de ces disparités, le point de vue reste universaliste, et d'un utilitarisme "technocratique".

En fonction de la culture scientifique locale, des moyens affectés, et des contraintes de mise en oeuvre, la diversité des "résultats" produits (pour un même tableau de données) est déjà légendaire. Cette diversité porte aussi bien sur la "précision" des estimations/descripteurs, que sur leur choix, leurs qualités (vraies ou supposées) affichées, ... voire, tout simplement, leur pertinence (le concurrent n'est jamais pertinent).

Une posture est proposée, prônant de pervertir les techniques, les méthodes, et de détourner le cours habituel des opérations auxquelles on s'astreint dans une

telle situation "de routine". En bref, elle repose sur l'utilisation d'outils robustes, ou robustifiés, au coeur d'une procédure évolutive d'estimation des paramêtres sensibles ; rien là de bien neuf !

L'innovation, s'il en est, est au niveau épistémologique, puisqu'il s'agit, en quelque sorte, de proposer une négociation aux données, pour les "aider" à se choisir "consensuellement" (ou presque !) de "bons" représentants. L'humain est ici seul juge de la qualité du résultat ; il s'en assure à travers la maîtrise du protocole de la négociation conduisant au consensus. Il suffit pour cela de se "doter" d'une des procédures (quelque peu censitaires) de pondération des votes dont nous avons le secret.

Un exemple très simple est présenté, démontant les mécanismes de la démarche à travers l'emploi du Biweight de Tukey.

Artificial neural network to classify ECG beats

***M.A. CHIKH**
en collaboration avec
N. BELGACEM et F. BERESKI-REGUIG

* Adresse pour correspondance :
Laboratoire de Génie Biomedical. Département d'informatique.
Faculté des Sciences de l'Ingénieur.
Université Abou Bekr Belkaid.
Tlemcen B.P 230, ple Chetouane. 13000 Algerie.
e-mail : mea_chikh@mail.univ-tlemcen.dz

Exposé du 11 Avril 2003

Abstract

In this study, electrocardiogram (ECG) beat classification was performed by using artificial neural networks (ANNS). Three neural classifiers have been developed to classify normal and abnormal premature ventricular (P.V.C) beats in ECG signal. The neural network classifiers with different input vectors are comparatively investigated to classify ECG beats. The comparative performance results of the three classifiers are reported using MIT-BIH database.

.

Data-driven rate-optimal specification testing in non-parametric regression models

Pascal LAVERGNE *
en collaboration avec Emmanuel GUERRE

* Adrese pour correspondance :
INRA-ESR, B.P. 27
31326 CASTANET-TOLOSAN
e-mail : lavergne@toulouse.inra.fr

Exposé du 14 Avril 2003

Résumé

Consider n i.i.d. observations (Y_i, X_i) in $\mathbb{R} \times \mathbb{R}^p$ and the heteroscedastic regression model

$$Y_i = m(X_i) + \epsilon_i \ , \ \mathbb{E}[\epsilon_i | X_i] = 0 \ \text{ and } \ \text{Var}[\epsilon_i | X_i] = \sigma^2(X_i) \, ,$$

where the regression function $m(\cdot)$ and the variance function $\sigma^2(\cdot)$ are unknown. We are interested in testing the hypothesis that the regression belongs to some parametric family $\{\mu(\cdot;\theta); \theta \in \Theta\}$, that is

$$H_0 \ : Y_i = \mu(X_i;\theta) + \epsilon_i \ , \ \text{for some } \theta \in \Theta. \tag{3}$$

Tests of H_0 are called lack-of-fit tests or specification tests. A popular approach for testing H_0, developed by Ramsey (1969) among others, consists in regressing the residuals from the estimated parametric model on some functions of the X_i's and in testing the fit of this auxiliary regression. To get a nonparametric consistent test, the approach has been extended to the case where the number of auxiliary functions grows with the sample size, see Hart (1997) for a review.

The fundamental issue for implementing a test of H_0 that we address here is the choice of the auxiliary model dimension. Since this is analogous to a model selection problem, some authors have proposed to use the well-known Schwarz (1978) information criterion (hereafter BIC), see e.g. Ledwina (1994), Hart (1997, Chapter 7) and Aerts, Claeskens and Hart (2000). This amounts to choose among n auxiliary models the one that gives the better measure of fit penalized by the

dimension of the model times $\ln n$. The measure of fit from the selected model is then used to built a test statistic.

We propose a data-driven procedure based on a new penalized criterion. Like the BIC, our criterion is designed to select a model corresponding, in some sense, to the true underlying regression function. We calibrate our new criterion to obtain an adaptive and rate-optimal test, i.e. a test that detects alternatives of unknown smoothness that get closer to the null hypothesis at the fastest possible rate when the sample size grows. This property does not hold for the BIC-based test. The central finding of our work is that $\ln n$ statistics and a penalization of much lower order than in the BIC are sufficient to obtain such an optimal test.

A competing approach that yields adaptive rate-optimal tests consists in choosing as a test statistic the maximum of studentized measures of fit from a sequence of auxiliary models, see Horowitz and Spokoiny (2001), Spokoiny (2001), and the related work of Baraud, Huet and Laurent (2002). Our approach has several distinctive advantages compared to this maximum approach. First, in this approach, critical values must be computed by simulations, which can be very time-consuming in large samples, unless one imposes supplementary restrictions as normality and homoscedasticity of the error terms, see Fan (1996) and Fan and Huang (2001). By contrast, our approach allows simple implementation in large samples as well as use of bootstrap critical values in small samples. Second, as will be shown, the maximum approach induces some loss in power, especially for alternatives of high regularity. Our test proved to be consistent against Pitman local alternatives converging to the null at a faster rate than the one obtained by Horowitz and Spokoiny (2001). Third, our test was found to be adaptive and rate-optimal against alternatives with low smoothness that are not allowed for by existing procedures. A simulation experiment also illustrates the good relative performances of our test.

Our general framework also improves upon previous work in two other ways. First, we consider general nonlinear nonlinear regression model with multdimensional covariates, heteroscedasticity of unknown form, and non-normal errors. Second, while previous work focused on specific smoothing techniques, such as regression on Fourier series (Fan and Huang, 2001) or kernel methods (Horowitz and Spokoiny, 2001), our general formulation allows for the use of various smoothing techniques. An important consequence is that our test can easily be adapted to be powerful against some specific alternatives such as additive nonparametric models.

The paper is organized as follows : the rest of this section is devoted to the presentation of our procedure. In Section 2, we present our main assumptions and details of the construction of the test. In Section 3, we derive the asymptotic properties of the test under the null hypothesis. In Section 4, we detail its consistency properties against smooth alternatives and against Pitman local alternatives. In Section 5, we prove the validity of a bootstrap method and compare the small sample performances of our test with a test based on the maximum approach.

In Section 6, we consider two main extensions of our results. We first consider general linear smoothing methods. We then propose a modification of our test whose power against nonparametric additive alternatives is not affected by the curse of dimensionality. In Section 7, we comment on how the approach could be extended to accomodate other testing problems than the one discussed in this paper. Proofs of the results are given in Section 8.

Let $\widehat{\theta}_n$ be the nonlinear least-squares estimator of θ in Model (3), and $\widehat{U}$ be the vector of estimated parametric residuals $\widehat{U}_i = Y_i - \mu(X_i; \widehat{\theta}_n), i = 1, \ldots, n$. A statistic $\widehat{T}_h$ is designed to estimate $\min_{\theta\in\Theta} \sum_{i=1}^n (m(X_i) - \mu(X_i;\theta))^2$. Since the estimated residuals write $\widehat{U}_i = Y_i - \mu(X_i;\widehat{\theta}_n) = m(X_i) - \mu(X_i;\widehat{\theta}_n) + \epsilon_i$, it follows that $\sum_{i=1}^n \widehat{U}_i(m(X_i) - \mu(X_i;\widehat{\theta}_n))$ should be close at first-order to $\sum_{i=1}^n (m(X_i) - \mu(X_i;\widehat{\theta}_n))^2$, which is of expected order $\min_{\theta\in\Theta} \sum_{i=1}^n (m(X_i) - \mu(X_i;\theta))^2$. The items $(m(X_i) - \mu(X_i;\widehat{\theta}_n))$ are not observed, but can nevertheless be estimated through a leave-one out linear nonparametric estimator with smoothing parameter h of the form

$$\widehat{\delta}_h(X_i) = \sum_{j\neq i} \nu_{ij}(h)\widehat{U}_j \ .$$

Typical examples considered in the literature and in our paper comes from regression on polynomials of order $1/h$ at most, regression on piecewise polynomials with binwidth h, or kernel regression with bandwidth h. This leads to consider the statistic

$$\widehat{T}_h = \sum_{i=1}^n \widehat{U}_i\widehat{\delta}_h(X_i) = \sum_{1\leq i\neq j\leq n} \frac{\nu_{ij}(h) + \nu_{ji}(h)}{2}\widehat{U}_i\widehat{U}_j = \widehat{U}'W_h\widehat{U} \ ,$$

where the generic elements of W_h are $w_{ij}(h) = (\nu_{ij}(h) + \nu_{ji}(h))/2$ for $i \neq j$ and $w_{ii}(h) = 0$. It is known that the suitably normalized statistic $\widehat{T}_h/\widehat{v}_h$ converges under H_0 to a standard normal when h goes to zero as n grows, see De Jong (1987).

Consider now a set $\mathcal{H}_n$ of admissible smoothing parameters. Following Horowitz and Spokoiny (2001) and Lepski, Mammen and Spokoiny (1997), we chose a geometric grid of $J_n + 1$ smoothing parameters as

$$\mathcal{H}_n = \{h_j = h_0 a^{-j}, j = 0, \ldots, J_n\} \ , \ a > 1 \ . \tag{4}$$

This parsimonious choice is discussed later on in Section 3. Let $\widehat{v}_{h,h_0}$ be a non-negative estimate of the standard deviation of $\widehat{T}_h - \widehat{T}_{h_0}$ under H_0. We select h as

$$\widetilde{h} = \arg\max_{h\in\mathcal{H}_n} \left\{ \widehat{T}_h - \widehat{T}_{h_0} - \gamma_n \widehat{v}_{h,h_0} \right\} \ . \tag{5}$$

Note that, when $\widehat{T}_h$ comes from a regression on auxiliary functions of X, the exact order of $\widehat{v}^2_{h,h_0}$ equals the number of auxiliary functions, i.e. the dimension

of the auxiliary model. Thus our penalization term is proportional to the square root of this dimension. This contrasts with the BIC, which uses the dimension itself. Our test writes

$$\text{Reject } H_0 \text{ when } \frac{\widehat{T}_{\tilde{h}}}{\widehat{v}_{h_0}} \geq z_\alpha \ , \tag{6}$$

where z_α is the quantile of order $(1-\alpha)$ of a standard normal and $\widehat{v}_{h_0}$ is an estimate of the standard deviation of T_{h_0} under H_0.

Under the null hypothesis, the test has asymptotic level α. Indeed, as the order of the $\widehat{T}_h - \widehat{T}_{h_0}$'s is $\widehat{v}_{h,h_0}$ under H_0, our data-driven $\tilde{h}$ yields $\widehat{T}_{\tilde{h}}/\widehat{v}_{h_0} = \widehat{T}_{h_0}/\widehat{v}_{h_0}$ asymptotically for a large enough γ_n. Hence, the test statistic converges to a standard normal for a well-chosen γ_n. We showed that this holds when γ_n is comparable to $\sqrt{2\ln\ln n}$, a surprisingly low order compared to the BIC. On the contrary, when H_0 does not hold, $\tilde{h}$ can clearly differ from h_0. Combining a growing number of statistics then allows to detect alternatives of unknown smoothness. Moreover, our test has always better power than the one based upon $\widehat{T}_{h_0}$ only, because our selection procedure implies that $\widehat{T}_{\tilde{h}} \geq \widehat{T}_{h_0}$ in any case.

To sum up, the characteristics of our procedure are as follows : first, our selection criterion penalizes less the dimension of the model than BIC does and the number of considered smoothing parameters is smaller than in the BIC-based selection, while the test is adaptive and rate-optimal and the test statistic has a pivotal distribution under H_0. Second, our "leave-one-out" construction of the statistics $\widehat{T}_h$'s allows to relax some restrictive conditions found in related work, as fast-rate estimation of the variance function $\sigma^2(\cdot)$, see Fan and Huang (2001), Horowitz and Spokoiny (2001) and Spokoiny (2001). Third, instead of normalizing the statistic $\widehat{T}_{\tilde{h}}$ by a estimate of its standard deviation, we chose the normalization $\widehat{v}_{h_0}$. This increases power without any cost, at least from an asymptotic viewpoint, see Fan (1996) for a similar device.

Références

AERTS, M., G. CLAESKENS and J.D. HART (1999). Testing the fit of a parametric function. *Journal of the American Statistical Association* 94 869–879.

AERTS, M., G. CLAESKENS and J.D. HART (2000). Testing lack of fit in multiple regression. *Biometrika* 87 (2) 405–4242.

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. *Proceedings of the Second International Symposium on Information Theory*, B.N. Petrov and F. Csaki eds., Akademiai Kiodo, Budapest, pp.

267–281.

ARCONES, M.A. and E. GINE (1993). Limit theorems for U-processes. *Annals of Probability* 21 1494–1542.

BARAUD, Y. (1997). Model selection for regression on random design. Ecole Normale Supérieure, Paris. *Probability Theory and Related Fields* 117 467–493.

BARAUD, Y., S. HUET and B. LAURENT (1999). Adaptive tests of linear hypotheses by model selection. Ecole Normale Supérieure, Paris. Available at `www.dma.ens.fr/~baraud/`.

BILLINGSLEY, P. (1968). *Convergence of probability measures*. Wiley, New-York.

BIRGE, L. and P. MASSART (1993). Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields* 97 113-150.

CHEN, J.C. (1994). Testing goodness-of-fit of polynomial models via spline smoothing techniques. *Statistics and Probability Letters* 19 65–76.

CHOW, Y.S. and H. TEICHER (1988). *Probability Theory : Independence, Interchangeability, Martingales*. Springer-Verlag, New-York.

CLEVELAND, W.S. and S.J. DEVLIN (1988). Locally weighted regression : an approach to regression analysis by local fitting. *Journal of the American Statistical Association* 83 (403) 596–610.

DAVIS, P.J (1975). *Interpolation and approximation*. Dover.

DELGADO, M.A. (1993). Testing the equality of nonparametric regression curves. *Statistics and Probability Letters* 17 199–204.

DE JONG, P. (1987). A Central Limit Theorem for Generalized Quadratic Forms. *Probability Theory and Related Fields* 75 261–277.

DETTE, H. (1999). A consistent test for the functional form of a regression based on a difference of variance estimators. *Annals of Statistics* 27(3) 1012–1040.

EUBANK, R.L. (1988). *Spline Smoothing and Nonparametric Regression*. Marcel Dekker, New-York.

EUBANK, R.L. and J.D. HART (1992). Testing goodness-of-fit in regression via order selection criteria. *Annals of Statistics* 20 (3) 1412–1425.

EUBANK, R.L. and J.D. HART (1993). Commonality of cusum, von Neumann and smoothing-based goodness-of-fit tests. *Biometrika* 80 89–98.

EUBANK, R.L. and C.H. SPIEGELMAN (1990). Testing the goodness of fit of a linear model via nonparametric regression techniques. *Journal of the American Statistical Association* 85 (410) 387–392.

FAN, J. and I. GIJBELS (1996). *Local Polynomial Modelling and its Applications*. Chapman et Hall, London.

FAN, J. and L.S. HUANG (2000). Goodness-of-fit tests for parametric regression models. *Journal of the American Statistical Association* forthcoming.

FAN, J., C. ZHANG and J. ZHANG (2001). Generalized lihelihood ratio statistics and wilks phenomenon. *Annals of Statistics* 29 (1).

GOZALO, P.L. (1997). Nonparametric bootstrap anamysis with applicatins to demographic effects in demand functions. *Journal of Econometrics* 81 357–393.

GUERRE, E. and P. LAVERGNE (2002). Optimal minimax rates for nonparametric specification testing in regression models. *Econometric Theory* 18 1139–1171.

GUERRE, E. and O. LIEBERMAN (2000). α-level adaptive testing in nonparametric regression via selection criteria. LSTA, Univ. Paris 6.

GYÖRFI, L., W. HÄRDLE, P. SARDA and P. VIEU (1989). *Nonparametric Curve Estimation From Time Series*. Springer-Verlag, Berlin.

HALL, P. and J.D. HART (1990). Bootstrap test for difference bewteen means in nonparametric regression. *Journal of the American Statistical Association* 85 1039–1049.

HALL, P., J.S. MARRON, M.H. NEUMANN and D.M. TITTERINGTON (1997). Curve estimation when the design density is low. *Annals of Statistics* 25(2) 756–770.

HÄRDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press.

HÄRDLE, W., G. KERKYACHARIAN, D. PICARD and A. TSYBAKOV (1998).

Wavelets, Approximation and Statistical Applications. Lecture notes in Statistics, 129. Springer-Verlag, Berlin.

HÄRDLE, W. and E. MAMMEN (1993). Comparing nonparametric versus parametric regression fits. *Annals of Statistics* 21 (4) 1926–1947.

HART, J.D. and T.E. WEHRLY (1992). Kernel regression when the boundary region is large, with an application to testing the adequacy of polynomial models. *Journal of the American Statistical Association* 87 1018–1024.

HART, J.D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer Verlag, New-York.

HONG, Y. and H. WHITE (1995). Consistent specification testing via nonparametric series regressions. *Econometrica* 63 1133–1160.

HOROWITZ, J.L. and V.G. SPOKOINY (2001). An adaptive, rate-optimal test of a parametric model against a nonparametric alternative. *Econometrica* 69(3) 599–631.

INGLOT, T., W.C.M. KALLENBERG and T. LEDWINA (1997). Data-driven smooth tests for composite hypotheses. *Annals of Statistics* 25 (3) 1222–1250.

INGSTER, Y. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives. (Part I, II and III) *Mathematical Methods of Statistics* 2 85-114 171–189 and 249–268.

LAVERGNE, P. and Q.H. VUONG (1998). An integral estimator of residual variance and a measure of explanatory power of covariates in nonparametric regression. *Journal of Nonparametric Statistics* 9(4) 363–380.

LEDWINA, T. (1994). Data-driven version of a Neyman's smooth test of fit. *Journal of the American Statistical Association* 89 1000–1005.

LI, Q. and S. WANG (1998). A simple consistent bootstrap test for a parametric regression functional form. *Journal of Econometrics* 87 145–165.

LORENTZ, G.G. (1966). *Approximation of functions*. Holt, Rinehart, and Winston.

MÜLLER, H.G. and U. STADTMÜLLER (1987). Estimation of heteroscedasticity in regression analysis. *Annals of Statistics* 15(2) 610–625.

NEWEY, W.K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics* 79 147–168.

RAMSEY, J.B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society, Series B* 31, 350–371.

RICE, J. (1984). Bandwidth choice for nonparametric regression. *Annals of Statistics* 12 1215–1230.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6(2) 461–464.

SERFLING, R.J. (1980). *Approximations Theorems of Mathematical Statistics*. Wiley, New-York.

SILVERMAN B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society, Series B* 47 (1) 1–52.

SPOKOINY, V.G. (1996). Adaptive hypothesis testing using wavelets. *Annals of Statistics* 24 (6) 2477–2498.

SPOKOINY, V.G. (1999). Data-driven testing the fit of linear models. Weierstrass Institute, Berlin. Available at `www.wias.berlin.de/private/spokoiny`.

STUTE, W. (1997). Nonparametric model checks for regression. *Annals of Statistics* 25 (2) 613–641.

WHITE, H. (1981). Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association* 76 419–433.

WU, C. F. J. (1986). Jacknife, Bootstrap and other Resampling Methods in Regression Analysis (with discussion). *Annals of Statistics* 14 1261–1350.

YANAGIMOTO, T. and M. YANAGIMOTO (1987). The use of marginal likelihood for a diagnostic test for the goodness of fit of the simple linear regression model. *Technometrics* 29 95–101.

ZHENG, X. (1996). A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics* 75 263–289.

Modélisation Statistique pour Variables Aléatoires Fonctionnelles : Théorie et Application

Frédéric FERRATY

GRIMM & LSP
Université Toulouse Le Mirail & Université Paul Sabatier,
5, allées Antonio-Machado & 118, route de Narbonne,
31058 Toulouse Cedex & 31062 Toulouse Cedex, France
e-mail : ferraty@univ-tlse2.fr ou Frederic.Ferraty@math.ups-tlse.fr

Habilitation à Diriger des Recherches
soutenue à l'Université Paul Sabatier

16 Juin 2003

Mots clés : Analyse en composantes principales fonctionnelles, Classification supervisée de courbes, Conditionnement par une variable aléatoire fonctionnelle, Dimension fractale ponctuelle d'un processus, Données fonctionnelles, Estimation fonctionnelle, Estimateurs à noyau, Modèle linéaire fonctionnel, Prévision, Probabilités de petites boules, Processus à temps continu, Régression fonctionnelle nonparamétrique, Séries chronologiques fonctionnelles, Splines de régression, Variables aléatoires fonctionnelles, Variables aléatoires hilbertiennes.

Résumé

L'originilité des travaux exposés dans ce document réside dans le fait qu'ils s'inspirent simultanément de plusieurs grands domaines de la statistique à savoir l'estimation fonctionnelle, la statistique des opérateurs, les variables fonctionnelles. De même, les outils nécessités pour les développements théoriques sont aussi de natures variées. En effet, ils relèvent de l'analyse fonctionnelle, de la théorie des opérateurs linéaires, de l'analyse numérique mais aussi d'outils probabilistes tels que la théorie des variables aléatoires hilbertiennes, les inégalités exponentielles pour des sommes de variables aléatoires, fonctionnelles ou non et plus récemment des probabilités de petites boules. Notons aussi l'omniprésence des aspects pratiques dans ces études, témoins des potentialités en terme d'applications

de ce champ de recherche. Pour finir, l'ensemble de ces travaux s'inscrit pleinement dans la dynamique du groupe de travail STAPH (http ://www.lsp.ups-tlse.fr/Fp/Ferraty/staph.html) à laquelle il participe. es.

Variables fonctionnelles et problématique

Mais avant d'aller plus loin, rappelons qu'une variable aléatoire fonctionnelle est tout simplement une variable aléatoire à valeurs dans un espace $\mathcal{F}$ de dimension infinie. Par exemple, cet espace $\mathcal{F}$ peut être un espace de fonctions, d'opérateurs linéaires,.... Ainsi, la principale source des difficultés, tant d'un point de vue théorique que pratique, provient du fait que les observations de ce type de variables sont supposées appartenir à un espace infiniment dimensionné. Le fil conducteur des travaux présents dans ce document réside dans l'apport systématique de solutions à ce problème de dimension infinie en mettant en place un cadre théorique suffisamment général. Par ailleurs, vu les nombreux domaines d'applications concernés par ce champ d'investigations, les aspects pratiques n'ont jamais été négligés.

Variables fonctionnelles et modélisations

Bien que les modèles de régression tiennent une place prépondérante, d'autres types de modélisation ont été abordés, notamment ceux concernant les séries temporelles ainsi que la classification supervisée de courbes. Par ailleurs, on va être amené à estimer des objets mathématiques (appelés génériquement opérateurs) définis sur l'espace $\mathcal{F}$ de dimension infinie et à valeurs dans $\mathbb{R}$. On parlera alors de modèle paramétrique lorsque l'objet à estimer appartient à un ensemble indexable par un nombre fini de paramètres appartenant à $\mathcal{F}$ et de modèle nonparamétrique dans le cas contraire. Ainsi, selon cette définition, le modèle linéaire fonctionnel peut être considéré comme un modèle paramétrique bien que l'objet à estimer soit de nature fonctionnelle.

Deux différentes approches ont été menées pratiquement de front. La première (d'un point de vue chronologique), paramétrique, s'est concentrée autour du modèle linéaire fonctionnel. La seconde aborde la modélisation de variables fonctionnelles de façon purement nonparamétrique.

Approche paramétrique : le modèle linéaire fonctionnel

Initialisé conjointement avec Hervé Cardot et Pascal Sarda, nous nous sommes intéressés au modèle linéaire fonctionnel (que l'on peut considérer paramétrique selon la définition précédente) et qui correspond au modèle de régression linéaire lorsque le régresseur est une variable aléatoire fonctionnelle (v.a.f.). Nous avons commencé par donner des premières propriétés théoriques d'un estimateur empirique pour ce modèle basé sur l'Analyse en Composantes Principales Fonctionnelle.

Une fois le cadre théorique mis en place, nous avons souligné la nécessité d'utiliser des estimateurs offrant un pouvoir lissant plus important. C'est ainsi que nous avons proposé deux nouveaux estimateurs. Le premier consiste à projeter le précédent sur une base B-spline alors que le second fournit un estimateur B-spline directement à partir d'un critère moindres carrés pénalisés. En plus de l'étude asymptotique, nous avons mis en œuvre ces estimateurs et comparé leurs performances respectives selon diverses situations.

Enfin, nous avons complété ces travaux en proposant deux statistiques de test portant sur le coefficient fonctionnel du modèle linéaire fonctionnel (en collaboration avec André Mas). La première peut être approximée par une v.a. suivant un χ^2 alors qu'on montre que la loi de la seconde convergence vers une gaussienne.

Approche nonparamétrique : modèles pour variables fonctionnelles

Parallèlement, en collaboration avec Philippe Vieu, nous avons développé une approche nonparamétrique pour la modélisation de variables fonctionnelles.

Dans un premier travail, on s'est intéressé au modèle de régression nonparamétrique lorque le régresseur est une variable fonctionnelle. La principale difficulté résulte du problème communément appelé "le fléau de la dimension". En effet, il est connu que dans le cas où le régresseur est à valeurs dans un espace de dimension finie p, la vitesse de l'estimateur se dégrade avec p. La principale innovation fut d'avoir solutionné ce problème du fléau de la dimension (puisqu'ici $p = \infty$) en proposant dans une première étape, une hypothèse de fractalité opérant sur la v.a.f. et en supposant $\mathcal{F}$ semi-normé. Nous avons ainsi obtenu des résultats de convergence pour un estimateur à noyau de l'opérateur de régression. Plus récemment, nous avons généralisé ces résultats dans des espaces semi-métriques et sans condition de type fractale.

Une fois le cadre théorique mis en place dans le cas d'un échantillon identiquement et indépendamment distribué, on s'est intéressé à la modélisation de séries temporelles. Pour cela, on a considéré le α-mélange et étendu les résultats obtenus précédemment.

Par la suite, nous avons adapté l'estimateur à noyau proposé dans le modèle de régression fonctionnelle afin de proposer une nouvelle méthode de classification supervisée de courbes.

Parallèlement, nous avons abordé le modèle à indice fonctionnel simple, généralisation au cadre fonctionnel du modèle à indice simple. Nous avons obtenu des premiers résultats à indice fixé.

Enfin, nous avons étudié l'estimation de la fonction de répartition conditionnellement à une v.a.f. ainsi que celle de la densité et de ses dérivées successives (toujours conditionnellement à une v.a.f.).

Conclusion

Pour conclure, je me contenterai de rappeler deux citations qui recadrent le contexte dans lequel se place l'ensemble de ces travaux. La première est due à Bosq

(1990) dans un article traitant de la modélisation des processus autorégressifs hilbertiens : *"These being nonparametric by themselves, it seems rather heavy to introduce a nonparametric model for observation lying in functional space ..."*. La seconde se trouve dans le livre de Ramsay et Silverman (1997) "Functional Data analysis" dans le paragraphe intitulé "Challenges for the future" : *"theoretical aspects of Functional Data analysis have not been researched in sufficient depth, and it is hoped that appropriate theoretical developments will feed back into advances in practical methodology"*.
Les travaux réalisés dans ce document s'inscrivent pleinement dans le prolongement de ces réflexions tout en apportant de nouveaux (et parfois de premiers) éléments de réponse. Les études à venir découlant des nombreuses perspectives offertes par ce champ d'investigations ne manqueront pas de compléter et d'enrichir ceux déjà présents dans ce manuscrit.

Summary

General Presentation

The works presented in this document cover mainly and simultaneously three great parts of Statistics namely the Nonparametric Estimation, the Statistic of Operators and Functional Random Variables (or Functional Data). The tools used for theoretical developments are also of various kinds. Indeed, they come from functional analysis as well as form linear operators algebra or from numerical analysis. Of course, many probabilistic tools can be used : the theory arround hilbertian random variables, exponential inequalities for sums of (functional or not) random variables and recent small balls probabilities theory. In addition, practical aspects are omnipresent, which show the wide field of applications of such research topics.

Functional variables and problematic

Before to go on, we recall that a functional random variable is a random variable taking its values in infinite-dimensional space $\mathcal{F}$. For instance, $\mathcal{F}$ can be a space of functions, linear operators,... Thus, the main difficulty, both in theory and practice, come from the fact that the observations of such variables are assumed to belong to an infinite-dimensional space. So, the guidelines and the main contribution of the jointed works consist in proposing a general theoretical framework able to override systematically the difficulties induced by this functional context. Moreover, applications and practical developments take an important place too.

Modelling and functional variables

A large part of my investigations concerns the regression models but others models have been studied in relation with time series, supervised classification of curves, distributions conditionally to a functional random variable. The common aim consists in estimating some mathematic objects (called generically operators) defined on the infinite-dimensional space $\mathcal{F}$. A model will be called parametric when the object to be estimated belongs to a set indexed by a finite number of parameters of $\mathcal{F}$ and nonparametric in the opposite case. Thus, according to this definition, the linear functional model can be viewed as a parametric model even if the target to be estimated is a functional object.

Two different approaches have been investigated almost simultaneously. The first (chronologically), parametric, focuses on the linear functional model. The second deals with nonparametric models.

Parametric approach : the linear functional model

Investigated jointly with Hervé Cardot and Pascal Sarda, we focused on the linear functional model which corresponds to a linear regression when the explanatory variable is a functional variable. First asymptotic properties have been stated for an empirical estimator based on functional principal components analysis.

Once the theoretical framework well defined, we emphasized on the necessity to obtain an estimator much smoother. Thus, we proposed two new estimators. The first one consists in projecting the previous empirical estimator onto a B-spline basis whereas the second one is a B-spline expansion obtained from a penalized least squares criterion. Finally, we compared these estimators, both theoritically and practically in various situations.

In addition, two statistics have been introduced in order to test the nullity of the functional coefficient of the functional linear model (in collaboration with André MAS). The first is approximated by a χ^2 whereas the second converges in distribution to a gaussian variable.

Nonparametric approach : models for functional variables

In collaboration with Philippe Vieu, we developped, just after the beginning of the previous parametric modelling, a nonparametric approach in order to build and study new models for functional variables.

In a first work, we focus on a nonparametric regression model when the regressor is a functional variable. The main difficulty is due to the so-called *curse of dimensionality*. Indeed, it is well known that the rate of convergence of some nonparametric estimator decreases when the dimension of the space in which the regressor takes its values increases. The principal originality of our work consists in solving this crucial *curse of dimensionality* by introducing a fractal type hypothesis acting on the functional variable. This fractal assumption allowed us to achieve first asymptotic results.

Once the theoretical framework well installed in the context of identical and independently distributed sample, we have been interested on modelling time

series. To do this, the α-mixing have been considered and the previous results have been extended to this dependent case.

Afterthen, our kernel estimator introduced in the nonparametric regression model situation have been adapted in order to propose a nonparametric supervised classification for curves.

An another model seems very interesting to developp in this functional context namely the single index model. Our goal is to generalize this kind of models in the case when the explanatory variable is functional and the reulting model is called the single functional index model. First results have been stated for a fixed functional index but obviously, the most interesting is to obtain asymptotic properties for an unknown functional index (in progress).

The last field of investigations in the nonparametric context deals with the distribution conditionned by a functional variable. The cumulative density function, density function and its successive derivatives have been studied.

Finally, note that the small balls probabilities like $P(X \in d(x, X))$ appear in the asymptotic developments where X is a functional variable and d a metric or semi-metric. In addition, as any process can be viewed as a functional variable, it is interesting to remark that the rate of convergence can be precised for some classical processes.

In conclusion

To conclude, I just recall two citations which are guidelines with respect to the presented works. The first is due to Bosq (1990) in a paper dealing with autoregressive hilbertian processes : *"These being nonparametric by themselves, it seems rather heavy to introduce a nonparametric model for observation lying in functional space ..."*. The second one can be found in the book of Ramsay et Silverman (1997) "Functional Data analysis" in the section entitled "Challenges for the future" : *"theoretical aspects of Functional Data analysis have not been researched in sufficient depth, and it is hoped that appropriate theoretical developments will feed back into advances in practical methodology"*. All the papers contained in this document give new (and sometimes first) answers to these thinkings. In addition, future investigations in this field will certainly contribute to complet and enrich the works already presented in this manuscrit.

Références

F. Ferraty (2003). *Modélisation pour Variables Aléatoires Fonctionnelles : Théorie et Application*. Synthèse de travaux de recherches présentés en vue de l'Habilitation à Diriger des Recherches, Université Paul Sabatier, Toulouse III (16 Juin 2003).

Sommaire des exposés des années précédentes

2èmes Journées de Statistique Fonctionnelle :
Toulouse 12-13 Juin 2003
A paraître dans les publications # **LSP 2003**

- *Using density level sets for nonparametric control charts*, Amparo BAILLO.
- *ACP conditionnelle*, Mohamed E. BAOUCHE, Tawfik BENCHIKH, Fatiha RACHEDI et Abderrahmane YOUSFATE.
- *ACP d'un processus stationnaire sous contrainte de B-mesurabilité*, Tawfik BENCHIKH et Abderrahmane YOUSFATE.
- *Recalage de courbes et analyse de variance fonctionnelle par ondelettes*,Jérémie BIGOT.
- *Modèle mixte de régression multiple à coefficients lisses*, Anestis ANTONIADIS et Noëlle BRU.
- *Comparison of parametric and semiparametric estimates in a degradation model*, Vincent COUALLIER.
- *Modèle linéaire généralisé et Directions Révélatrices*, Michel DELECROIX.
- *Etude asymptotique de l'Analyse Canonique : de l'approche matricielle et analytique à l'approche opératorielle et tensorielle*, Jeanne FINE.
- *Estimation du potentiel d'interaction de paires d'un processus de Gibbs*, Jean-Michel BILLIOT et Michel GOULARD.
- *Vitesse optimale du cas i.i.d. pour l'estimation non-paramétrique de la densité invariante d'un système dynamique chaotique* , Salim LARDJANE.
- *Estimation de la fonction du taux d'occurence d'événements ponctuels*, Christophe BONALDI et Nicolas MOLINARI.
- *Estimation fonctionnelle des modes conditionnels*, Abbes RABHI et Abderrahmane YOUSFATE.
- *Propriétés extrémales des valeurs singulières d'un opérateur compac*, Jean Jacques TÉCHENÉ.

Année 2001-2002
Sommaire de la publication # **LSP 2002-12**

Journées de Statistique Fonctionnelle :
Toulouse 10-11 Juin 2002
Sommaire de la publication # **LSP 2002-09**

Année 2000-2001
Sommaire de la publication # **LSP 2001-07**

Année 1999-2000
Sommaire de la publication # **LSP 2001-05**