1 - Curriculum Vitae

Laurent DELSOL

Né le 25 mai 1982 à Cherbourg, France.
Apt. 83, 210 rue des Poilus, 45160 Olivet, France.
Tél : 00332.38.49.26.96.
E-mail : laurent.delsol@univ-orleans.fr

Formation

Septembre 2009	Recrutement en tant que Maître de Conférence au sein du labratoire MAPMO de l'Université d'Orléans.
2008-2009	Postdoctorat au sein de l'institut de statistique de l'Université Catholique de Louvain. Collaboration avec Ingrid Van Keilegom dans le cadre du projet E.R.C. intitulé "M- and Z-estimation in semiparametric statistics : applications in various fields".
2005-2008	-Moniteur à l'Université Paul Sabatier (Toulouse III). -Allocataire de recherche : doctorat en Mathématiques Appliquées, spécialité statistique, sous la direction de Frédéric FERRATY et Philippe VIEU au sein de l'Institut de Mathématiques de Toulouse (équipe L.S.P.) à l'Université Paul Sabatier. Date de soutenance : 17 juin 2008. Titre : Régression sur variable fonctionnelle : Estimation, Tests de structure et Applications. Mention : Très Honorable. Jury : Denis Bosq (rapporteur). Frédéric Ferraty (directeur). Peter Hall (rapporteur). Pascal Sarda (examinateur). Winfried Stute (examinateur). Ingrid Van Keilegom (examinatrice) Philippe Vieu (directeur).
2004-2005	Master 2 Recherche de Mathématiques Appliquées à l'Université Toulouse III (mention : TB).
2003-2004	Maîtrise de Mathématiques Fondamentales à l'Université Paul Sabatier (mention : B).
2002-2003	Licence de Mathématiques Fondamentales à l'Université Paul Sabatier (mention : B).
2000-2002	DEUG MIAS à l'Université Paul Sabatier (mention : TB).
2000	Baccalauréat série Scientifique.

Langues et Compétences informatiques

- Français : langue maternelle.
- Anglais : lu, parlé, écrit.
- Espagnol : niveau Terminale.
- Logiciels utilisés : R, Splus, Maple.
- Production de documents : LaTeX, Word, HTML.
- Environnements : Unix et Windows.

Activités d'enseignement

2009-2014	MCF à l'université d'Orléans. L1 Biologie : Cours et TD, Analyse de données. M2 SPA et ATI : Cours-TP, Statistique et Image (partagé avec Cécile Louchet). M2 SPA : Cours-TP, Estimation non-paramétrique. M1 SPA : TD, Statistique théorique. M1 SIME et MPME : Cours et TD-TP, Analyse de données. L3 MIAGE : Cours-TP, Statistique. L1 Mathématiques : Cours-TP, Initiation au calcul formel (avec Maple). 1ère année IUT Informatique : TD, Mathématiques discrètes.
2007-2008	Moniteur à l'Université Toulouse III. L2 SVS : TD de Probabilités et Statistique (Statistique descriptive, Probabilités, Statistique inférentielle). IUP SID (Statistique et Informatique Décisionnelle) : TD de Probabilités en première année.
2006-2007	Moniteur à l'Université Toulouse III. L2 SVT BIP : TP Maple et R, "Méthodes et outils de calculs" (Analyse, Statististique, Probabilités). L1 filière PC : TD d'Analyse "Fonctions de la variable réelle. Intégrales. Equations différentielles".
2005-2006	Moniteur à l'Université Toulouse III. L2 SVT BIP : TP Maple et R, "Méthodes et outils de calculs" (Analyse, Statististique, Probabilités). L3 Pluridisciplinaire à dominante mathématique : TP Word et Excel. L1 filière PC : TD d'Analyse "Fonctions de la variable réelle. Intégrales. Equations différentielles".

Encadrement de projets et de stages en Master SPA I et II.

J'ai suivi au cours des années 2005-2008 la formation proposée par le C.I.E.S. de la région Midi-Pyrénées. Dans le cadre de cette formation j'ai notament suivi des stages :

- de communication
- d'informatique
- de secourisme.

J'ai également aidé à la préparation et l'encadrement des Défis Solaires en 2007 à Toulouse dans le cadre d'un atelier-projet proposé par le C.I.E.S..

Activités scientifiques

Séjours dans des laboratoires étrangers

2008	Un mois à l'Université Catholique de Louvain (Belgique) sur invitation du Professeur Ingrid Van Keilegom.
2008-2009	Une année à l'Université Catholique de Louvain (Belgique) dans le cadre d'un stage de post-doctorat sur invitation du Professeur Ingrid Van Keilegom.
2010	Une semaines à l'Université Catholique de Louvain (Belgique) pour travailler avec Ingrid Van Keilegom.
2011	Deux semaines à l'Université Catholique de Louvain (Belgique) pour travailler avec Ingrid Van Keilegom et Catherine Timmermans.

Thèmes de recherche :

Je travaille depuis le début de ma thèse sur des modèles de régression sur variable fonctionnelle. J'ai en particulier travaillé sur les propriétés asymptotiques (normalité, erreurs $\mathbb{L}^p$) de l'estimateur non-paramétrique à noyau de l'opérateur de régression, notamment dans le cas d'un échantillon alpha-mélangeant. J'ai également proposé et étudié de nouvelles procédures de test de structure adaptées au cas de la régression sur variable fonctionnelle ainsi que sur l'utilisation de méthodes de type bootstrap dans ce contexte.
J'ai également collaboré avec Christophe Crambes et Ali Laksaci sur des problématiques concernant les propriétés asymptotiques de Z-estimateurs conditionnels.
Je travaille également avec Ingrid Van Keilegom sur l'étude de M-estimateurs définis à partir de critères semi-paramétriques non continus par rapport au paramètre sur lequel on veut maximiser.
Mes travaux de recherche m'ont conduit au cours des dernières années à des collaborations avec Christophe Crambes, Frédéric Ferraty, Wenceslao Gonzalez-Manteiga, Ali Laksaci, Catherine Timmermans, Ingrid Van Keilegom, Philippe Vieu et Rainer Von Sachs. Et des collaborations se mettent en place (ou sont déjà en cours) avec des collègues orléanais (Cécile Louchet, Didier Chauveau et Nicolas Debarsy).

Liste de publications

Articles parus dans des revues à comité de lecture

- L. Delsol (2007) CLT and L^q errors in nonparametric functional regression, *Comptes Rendus de l'Académie des Sciences, Ser. I,* **345**, (7), pp 411-414.
- L. Delsol (2007) Régression non-paramétrique fonctionnelle : Expressions asymptotiques des moments. *Annales de l'I.S.U.P.*, **LI**, (3), pp 43-67.
- L. Delsol (2008) Tests de structure en régression sur variable fonctionnelle *Comptes Rendus de l'Académie des Sciences, Ser. I,* **346**, (5-6), pp 343-346.
- C. Crambes, L. Delsol and A. Laksaci (2008) Robust nonparametric estimation for functional data. *Journal of Nonparametric Statistics,* **20**, (7), pp 573-598.
- L. Delsol (2009) Advances on asymptotic normality in nonparametric functional Time Series Analysis. *Statistics*, **43**, (1), pp 13-33.
- L. Delsol, F. Ferraty and P. Vieu (2011) Structural test in regression on functional variables. *J. Multivariate Analysis* **102**(3), 422-447.
- L. Delsol (2013) No effect tests in regression on functional variable and some applications to spectrometric studies, *Computational Statistics*, **4**, 1775-1811.
- C. Timmermans, L. Delsol and R. von Sachs (2013) Using Bagidis in nonparametric functional data analysis : Predicting from curves with sharp local features. *J. Multivariate Analysis* (MA) **115**, 421-444.

Articles parus dans des Actes ou Proceedings avec comité de lecture

- C. Crambes, L. Delsol and A. Laksaci (2008) Robust Nonparametric Estimation for Functional Data. *Functional and Operatorial Statistics.* [Ed S ; DABO-NIANG et F. FERRATY]. Springer.
- L. Delsol (2008) Nonparametric Regression on Functional Variable and Structural Tests. *Functional and Operatorial Statistics.* [Ed S ; DABO-NIANG et F. FERRATY]. Springer.

Chapitre de Livres

- L. Delsol (2011) Nonparametric methods for α-mixing functional random variables. *Oxford Handbook of Functional Data Analysis.* Oxford University Press.

- L. Delsol, F. Ferraty and A. Martinez (2011) Functional data analysis : an interdisciplinary statistical topic. *Statistical learning and Data Science, Eds. B. Goldfarb, C. Pardoux,M. summa, M. Touati,* Chapman & Hall.

Manuscrit de Thèse

- L. Delsol (2008) Régression sur variable fonctionnelle : Estimation, Tests de structure et Applications. *Thèse de doctorat de l'Université de Toulouse.*

Articles soumis

- L. Delsol and I. Van Keilegom, M-estimators when the criterion function is not smooth.

Articles en préparation

- L. Delsol, and C. Louchet, Hyperspectral Image Segmentation using functional kernel density estimation.

Participation à des congrès :

Communications orales lors de congrès internationaux :

- Asymptotic normality in nonparametric time series analysis, V^e *Workshop of the IAP research network, Louvain-la-Neuve, 2006.*
- Regression on functional variable : Estimation, Structural tests and Applications, *International Workshop on Functional and Operatorial Statistics, Toulouse, 2008.*
- Structural tests in regression on functonal variable, *Joint Statistical Meetings, Denver, 2008.* [Sur invitation]
- Structural tests in regression on functional variable : Theoretical background, Bootstrap methods and Applications., *Annual meeting of the Statistical Belgian Society, Namur, 2008.*
- Testing for linearity in regression on functional variable, *IASC Conference, Yokohama, 2008.* [Sur invitation]
- Testing for specific models in regression on functional variable, *ASMDA, Vilnius, 2009.* [Sur invitation]
- Testing structural assumptions in regression when the covariate is functional, *S. Co., Milan, 2009.* [Sur invitation]
- Testing structural assumptions in regression on functional variable, 10th International Vilnius Conference on Probability and Mathematical Statistics, du 28 juin au 2 juillet 2010, Vilnius, Lithuanie. [Sur invitation]
- Recent Advances on Structural Tests for Regression Models with Functional Covariate, Joint Statistical Meeting, du 31 juillet au 5 août 2010, Vancouver, Canada. [Sur invitation]
- Focusing on Structural Asumptions in Regression on Functional Variable, 58th ISI congress, 21-26 août 2011, Dublin. [Sur invitation]
- From time series study to regression models with functional data. International Workshop on Recent Advances in Time Series Analysis, 9-12 juin 2012, Protaras, Chypre. [Sur invitation]
- Selecting informative BAGIDIS coefficients in nonparametric functional regression, First ISNPS Conference, 15-19 juin 2012, Chalkidiki, Grece. [Sur invitation]
- Functional kernel smoothing methods applied to segment hyperspectral images, Workshop "Dependent functional data", 24-26 janvier 2013, Göttingen, Germany. [Sur invitation]
- Functional kernel smoothing methods applied to segment hyperspectral images, 6th International Conference of the ERCIM WG on Computational and Methodological Statistics ; London, England. 14-16 décembre 2013,[Sur invitation]

– Hyperspectral image segmentation based on functional kernel density estimation, 2nd ISNPS Conference, 12-16 juin 2014, Cadiz, Spain.

Communications orales lors de congrès nationaux :

– Normalité asymptotique de la régression sur variable fonctionnelle avec application à la construction d'I.C., *IV^e Journées STAPH, Grenoble, 2006.*
– Régression non-paramétrique fonctionnelle : expressions asymptotiques des moments et des erreurs L^p, *39^e Journées de Statistique, Angers, 2007*
– Régression non-paramétrique fonctionnelle, propriétés asymptotiques et tests, *Deuxièmes Rencontres des Jeunes Statisticiens, Aussois, 2007.*
– *Journées Franco-Tchèques*, le 5 mai 2008, Université Toulouse III.
– Tests de modèles particuliers en régression sur variable fonctionnelle, *VI^e Journées STAPH*, Dijon, 2009. [Sur invitation]
– Utilisation de tests de structure en régression sur variable fonctionnelle. *42èmes JdS*, Marseille, 2010. [Sur invitation]
– Segmentation d'images hyperspectrales à partir d'estimation à noyau fonctionnel de la densité, *45èmes JdS*, Toulouse, 2013.

Présentations de poster :

– Robust nonparametric estimation for functional data. *International Workshop on Functional and Operatorial Statistics*, les 19-21 juin 2008 ; Toulouse FRANCE.
– Robust nonparametric estimation for functional data : Lp errors and applications. *International Seminar on Nonparametric Inference*, les 5-7 novembre 2008, Vigo, ESPAGNE.
– Structural tests in regression on functional data *International Workshop on Functional and Operatorial Statistics*, 16-18 Juin 2011, Santander, ESPAGNE.

Présentations à un séminaire ou un groupe de travail :

– *Séminaire étudiant de Statistique et Probabilités*, le 11 avril 2006 à l'Université Toulouse III.
– *Groupe de Travail STAPH*, le 13 novembre 2006 à l'Université Toulouse III.
– *Séminaire étudiant de Statistique et Probabilités*, le 13 mars 2007 à l'Université Toulouse III.
– *Séminaire étudiant de Statistique et Probabilités*, le 2 octobre 2007 à l'Université Toulouse III.
– *Séminaire de Statistique de Montpellier*, le 20 octobre 2008, à l'INRA-SUPAGRO de Montpellier.
– *Séminaire Application des Mathématiques*, le 22 octobre 2008, à l'Université de Bourgogne (Dijon).
– *Séminaire de Statistique*, le 25 novembre 2008, à l'Université Toulouse III.
– *Séminaire de Statistique*, L.J.K, le 27 novembre 2008, Université Joseph Fourier, Grenoble.
– *Séminaire SIUTE*, le 16 décembre 2008, Université Lille 3.
– *Séminaire de Statistique et Probabilités*, le 12 janvier 2009, Université d'Angers.
– *Séminaire de Statistique de Rennes*, le 23 janvier 2009, Université Rennes 1.
– *Séminaire de Statistique et Probabilités du L.M.A.*, le 27 janvier 2009, université de Pau et des pays de l'Adour.
– *Séminaire de recherche de l'université d'Informatique de Namur*, le 9 février 2009, Université de Namur.
– *Séminaire de l'Institut de Science Financière et d'Assurances*, le 23 février 2009, Université Claude Bernard (Lyon 1).
– *Séminaire du G.R.E.M.A.Q.*, Toulouse School of Economics, le 9 mars 2009.
– *Séminaire de Statistique de l'Institut de statistique*, le 8 mai 2009, Université Catholique de Louvain.
– *Séminaire de Statistique de l'Institut de Mathématiques de Bordeaux*, le 18 février 2010.
– *Rencontres Mathématiques de Rouen*, les 01-02 juin 2010.
– *Séminaire de Statistique de l'Université Libre de Bruxelles*, le 10 octobre 2014.
– *Séminaire de Statistique du Laboratoire de Mathématiques d'Avignon*, le 28 novembre 2014.

Rapporteur de revues :

- *Journal of Statistical Planning and Inference*,
- *Applied Mathematics and Information Science*,
- *Journal of Multivariate Analysis*,
- *Journal of Nonparametric Statistics*,
- *Journal of the Royal Statistical Society, Series B*,
- *Statistic and Probability letters*,
- *Journal of the Korean Statistical Society*,
- *Statistical Papers*,
- *Communications in Statistics : Theory and Methods*,
- *Engineering Structures*,
- *Electronic Journal of Statistics*.

2 - Travaux de recherche

Introduction

De nombreux problèmes concrets font intervenir des phénomènes de nature fonctionnelle (c'est à dire de dimension grande ou infinie). Il peut par exemple s'agir de l'évolution d'une quantité au cours du temps (courbes de croissance, évolution du prix d'une action en bourse, courbes de température, mesures du taux d'ozone, ...), de courbes spectrométriques, d'enregistrements sonores, de relevés satellites, d'images, ... D'autre part, les améliorations effectuées au niveau des dispositifs de mesure permettent de plus en plus facilement de collecter des données sur des grilles assez fines pour bien rendre compte de la nature fonctionnelle de ces phénomènes. Les outils classiques de statistique multivariée sont souvent inadaptés pour considérer ce type de jeux de données. Ils se heurtent à des problèmes provenant de la grande dimension des variables, de la corrélation de leur composantes et ne permettent pas de bien prendre en compte la régularité et la structure du phénomène étudié. Une autre approche consiste à considérer ces données comme la discrétisation (de la réalisation) de variables aléatoires, dites fonctionnelles, à valeurs dans un espace de dimension infinie. Cette modélisation, déjà envisagée dans les travaux précurseurs de Rao (1958) et Tucker (1958), permet souvent de représenter de manière plus synthétique les données en prenant compte de la régularité et de la structure du phénomène observé. L'étude statistique de variables fonctionnelles s'est ensuite abondamment développée sous l'impulsion des travaux de Grenander (1981), Dauxois *et al.* (1981), et Ramsay (1982) qui mettent en avant la manière dont certaines méthodes classiques en statistique multivariée peuvent être adaptées et s'intéressent à de nouveaux problèmes que ce type de données peut engendrer. Ce domaine de la statistique, popularisé au travers des monographies de Ramsay et Silverman (1997, 2002, 2005), Bosq (2000) ainsi que Ferraty et Vieu (2006), est actuellement en plein essor autant pour la diversité de ses domaines d'application (agronomie, biologie, génétique, médecine, chimie, études environnementales, économie, linguistique, télédétection, ...) que pour l'intérêt des problèmes et des méthodes qu'il reste encore à considérer.

Problèmes concrets, modèles et problématiques considérés

Prenons tout d'abord un exemple provenant de l'industrie agro-alimentaire. Lors du conditionnement d'émincés de viande, il est obligatoire d'indiquer sur l'emballage le taux de graisse de chaque portion. Il est possible d'obtenir cette valeur de manière précise par analyse chimique mais cela est assez coûteux, c'est pourquoi on aimerait pouvoir prédire le taux de graisse à partir de mesures spectrométriques que l'on peut obtenir plus facilement. On dispose de données, accessibles sur le site de StatLib (http ://lib.stat.cmu.edu/datasets/tecator), provenant de l'étude de 215 morceaux de viande. Pour chaque morceau de viande, on obtient par analyse chimique la teneur en graisse et on collecte des données spectrométriques correspondant à l'absorbance du morceau de viande pour 100 longueurs d'onde comprises entre 850nm et 1050nm. Comme le soulignent Leurgans *et al.* (1993), ces données spectrométriques sont intrinsèquement de nature fonctionnelle et sont assimilables à des courbes (voir Figure 1), appelées courbes spectrométriques. Au travers de ce problème concret apparait clairement le problème de la prédiction. Toutefois avant même de considérer celui-ci il semble pertinent de s'interroger sur l'existence même et la nature du lien reliant la courbe spectrométique et le taux de graisse.
Mon travail de recherche concerne principalement l'étude de modèles où une variable d'intérêt réelle Y dépend d'une courbe (ou plus généralement d'une variable explicative fonctionnelle) $\mathcal{X}$. Au cours de ces dernières années, je me suis principalement intéressé à l'étude de ce lien au travers du modèle de régression sur variable fonctionnelle

$$Y = r(\mathcal{X}) + \epsilon, \tag{1}$$

dans lequel Y est une variable aléatoire réelle, $\mathcal{X}$ une variable aléatoire à valeurs dans un espace semi-métrique $(\mathcal{E}, d)$, r est un opérateur inconnu que l'on cherche à étudier tandis que ϵ représente le terme d'erreur et vérifie $\mathbb{E}[\epsilon|\mathcal{X}] = 0$. Parmi les modèles de ce type considérés dans la littérature, on trouve divers

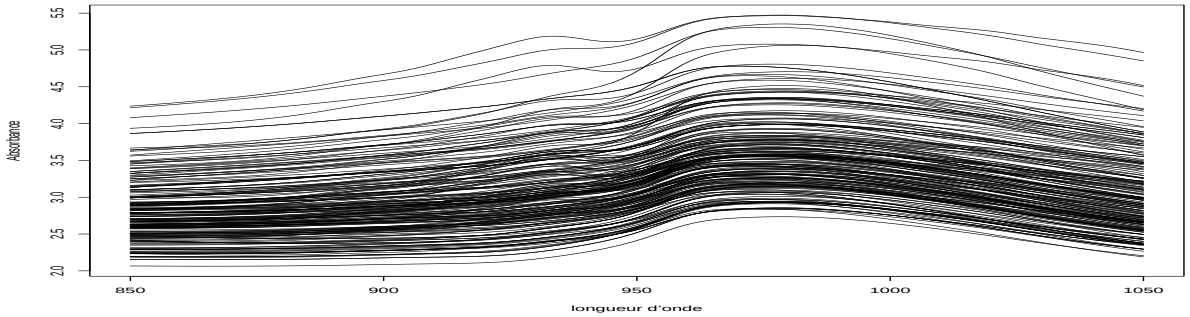

FIGURE 1 – 215 courbes spectrométriques

modèles basés sur des hypothèses de structure tels que le modèle linéaire fonctionnel (Ramsay and Dalzell, 1991, Cardot *et al.*, 1999, Crambes *et al.*, 2008), le modèle à indice simple fonctionnel (Ait Saidi *et al*, 2008), le modèle partiellement linéaire fonctionnel (Aneiros-Perez et Vieu, 2008), ... D'autre part, plusieurs auteurs se sont également intéressés à l'étude de modèles non-paramétriques dans lesquels on ne fait pas d'hypothèse sur la structure de r mais simplement sur sa régularité (voir Ferraty et Vieu, 2000, 2006).
Supposons maintenant que l'on dispose d'un échantillon composé de N couples indépendants $(\mathcal{X}_i, Y_i)_{1\leq i\leq N}$ de même loi que le couple $(\mathcal{X}, Y)$ et que l'on souhaite prédire la valeur de la variable d'intérêt correspondant à une valeur x de la variable explicative. On dispose de diverses méthodes de prédiction basées sur méthodes d'estimation de r adaptées aux différents modèles que nous venons d'évoquer. Avant d'essayer une de ces méthodes, il semble raisonnable de considérer le problème de la construction de tests de structure portant sur l'opérateur r. Ils permettent notamment de s'interroger sur l'existence d'un effet de la variable explicative sur la variable d'intérêt au travers du test de l'hypothèse $\mathcal{H}_0 : \{\exists C \in \mathbb{R},\, r \equiv C\}$, mais aussi de se demander si cet effet a une structure particulière en testant par exemple si r est linéaire, à indice simple, multivarié, ... On peut ensuite, à partir des informations obtenues sur la nature du modèle de régression sous-jacent, proposer une méthode d'estimation pertinente.

Discutons à présent d'un problème de prédiction un peu différent lié à l'étude de séries temporelles au travers d'une approche fonctionnelle. On s'intéresse pour cela à des données environnementales liées à l'étude du phénomène El Nino. On dispose de mesures mensuelles de température éffectuées autour du courant El Nino de mai 1950 et avril 2008 (voir Figure 2.a). Ces données sont disponibles en ligne (http ://www.cpc.ncep.noaa.gov/data/indices/) et sont régulièrement mises à jour. On souhaite prédire les valeurs des températures de la période de 12 mois suivante (i.e. de mai 2008 à avril 2009) à partir de ces données. Afin de prendre en compte la structure et la "périodicité" de l'évolution de la température au cours des années, on propose de regarder nos données non plus seulement au travers de leur forme discrétisée mais aussi comme provenant de l'évolution d'un processus à temps continu Z_t observé sur une période $[0, \tau N]$ (ici $\tau = 12$ et $N = 58$). On propose ensuite de suivre l'approche introduite par Bosq (1991) et de découper la trajectoire observée en N courbes annuelles $\mathcal{X}_i(t) = Z_{\tau(i-1)+t}$ (voir Figure 2.b). Cette approche a l'avantage de donner une représentation plus synthétique des données ainsi que de bien prendre en compte la régularité et la structure de l'évolution des températures. Le problème de prédiction de la température pour le mois m peut alors être considéré au travers de l'étude du modèle de régression

$$Y_i^m := G_m(\mathcal{X}_i) = r(\mathcal{X}_{i-1}) + \epsilon_i,$$

où G_m est l'opérateur qui à la courbe $\mathcal{X}$ associe la valeur correspondant au mois m. De manière plus générale on peut considérer d'autres problèmes de prédiction (extrema, effets cummulés, appartenance à une classe, ...) en remplaçant G_m par d'autres types d'opérateurs.

Il est également possible de prendre comme variables explicatives plusieurs courbes annuelles précédentes car leur ensemble est encore une variable fonctionnelle. D'autre part, les résultats obtenus à partir de cette approche fonctionnelle des séries temporelles ne se détériorent pas si on a une discrétisation plus

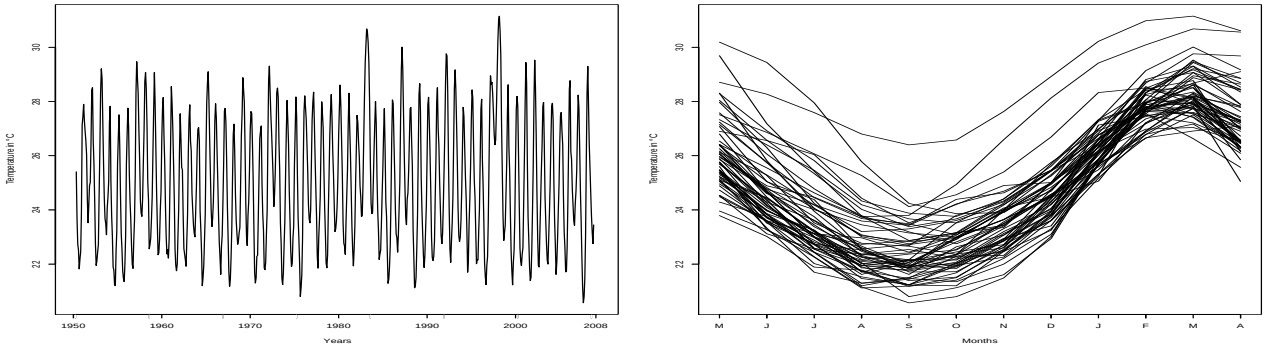

FIGURE 2 – 2.a) 696 mesures mensuelles 2.b) 58 courbes annuelles

fine des observations et s'appliquent directement à des situations où le nombre ou la nature des points de discrétisation sont différents d'une période à l'autre. Enfin, on propose de prendre en compte la dépendance qu'il existe entre les courbes $(\mathcal{X}_i)_{1\leq i\leq N}$ en considérant des échantillons de variables α-mélangeantes. Cette condition de mélange fort introduite par Rosenblatt (1956) est l'une des plus générales et est notament vérifiée par différents types de processus (voir par exemple Bradley, 2005). Une discussion de l'utilisation des méthodes à noyau fonctionnel pour étudier ce type de données fait l'objet d'un chapitre que l'on m'a invité à rédiger pour un livre sur l'état de l'art en statistique fonctionnelle (voir Ferraty et Romain, 2011, I.4).

Résultats obtenus

Estimation en régression non-paramétrique sur variable fonctionnelle

Dans les modèles non-paramétriques de régression sur variable fonctionnelle, on ne suppose pas que l'opérateur r a une structure particulière mais simplement qu'il est régulier (par exemple hölderien) par rapport à une pseudo-métrique d. Dans ce type de modèles, on utilise principalement l'estimateur à noyau fonctionnel défini par Ferraty et Vieu (2000) qui constitue une généralisation de l'estimateur de Nadaraya-Watson au cas de variables fonctionnelles :

$$\forall x \in \mathcal{E},\ \hat{r}(x) = \begin{cases} \dfrac{\sum_{i=1}^{N} Y_i K\left(\frac{d(X_i,x)}{h_n}\right)}{\sum_{i=1}^{N} K\left(\frac{d(X_i,x)}{h_n}\right)} & \text{si } \sum_{i=1}^{N} K\left(\frac{d(X_i,x)}{h_n}\right) > 0, \\ 0 & \text{sinon,} \end{cases} \tag{2}$$

où K est un noyau dont le support est $[0;1]$, d est la pseudo-métrique que l'on considère et $(h_n)_{n\in\mathbb{N}}$ est la séquence de paramètres de lissage. Ici, la pseudo-métrique d est considérée comme un nouveau paramètre de notre méthode. En dimension infinie le choix de la topologie que l'on considère est crucial. L'utilisation de topologies induites par des pseudo-métriques offre des possibilités nouvelles de trouver une topologie qui ait à la fois de bonnes propriétés de concentration par rapport aux variables explicatives et qui fasse ressortir de ces variables les traits importants pour faire de la prédiction. En pratique, on se fixe à partir des connaissances que l'on a sur les données une famille de pseudo-métriques (pseudo-métriques de projection, basées sur les dérivées, ...) et on choisit dans cette famille celle qui est la plus adaptée par validation croisée. Afin de mieux comprendre et interpréter les résultats obtenus à l'aide de l'estimateur (2), on s'intéresse à l'étude de ses propriétés asymptotiques. Ces propriétés asymptotiques s'expriment en fonction de quantités importantes connues sous le nom de probabilités de petites boules et définies par

$$F_x(h) = P(d(\mathcal{X}, x) \leq h).$$

Ces quantités jouent le même rôle que h_n^p dans le cas multivarié, ont l'avantage de ne pas nécessiter l'existence d'une densité par rapport à une mesure de référence donnée et dépendent clairement du choix de la

pseudo-métrique d.

Normalité asymptotique

Je me suis tout d'abord intéressé à l'étude de la normalité asymptotique. De premiers résultats dans ce sens ont été obtenus par Masry (2005) dans le cas α-mélangeant sans donner l'expression des termes dominants du biais et de la variance de la loi asymptotique. Leurs expressions ont été données par Ferraty *et al* (2007) dans le cas indépendant à l'aide des constantes

$$\begin{aligned} M_0 &= K(1) - \int_0^1 (sK(s))' \tau_0(s)\, ds, \\ M_1 &= K(1) - \int_0^1 K'(s) \tau_0(s)\, ds, \\ M_2 &= K^2(1) - \int_0^1 \left(K^2\right)'(s) \tau_0(s)\, ds, \end{aligned}$$

où $\forall s \in [0;1]$, $\tau_0(s) = \lim_{h\to 0} F_x(hs) F_x^{-1}(h)$ (cette limite existe pour un grand nombre de processus, voir Ferraty *et al*, 2007). Je me suis intéressé à la manière dont on pouvait étendre leur approche au cas de variables α-mélangeantes. L'hypothèse de régularité que l'on fait sur r se fait au travers de la fonction ϕ_x définie par

$$\forall s \in \mathbb{R}^+,\ \phi_x(s) = \mathbb{E}\left[r(\mathcal{X}) - r(x) \,|\, d(\mathcal{X}, x)\right]_{|d(\mathcal{X},x)=s}$$

que l'on suppose nulle et dérivable au point 0.
Sous des conditions assez générales, on montre alors que

$$\frac{M_1}{\sqrt{M_2\sigma_\epsilon^2}} \sqrt{n\hat{F}_x(h_n)} \left(\hat{r}(x) - r(x) - h_n \phi_x'(0) \frac{M_0}{M_1}\right) \to \mathcal{N}(0,1) \tag{3}$$

où $\sigma_\epsilon^2 = \mathbb{E}\left[\epsilon^2 | \mathcal{X}\right]_{|\mathcal{X}=x}$ et $\hat{F}_x(t) = \frac{1}{n}\sum_{i=1}^n 1_{[d(X_i,x),+\infty[}(t)$. Ce résultat est donné pour différentes conditions de α-mélange dans le cadre d'une publication parue dans la revue Statistics (Delsol, 2009). On y discute également la construction d'intervalles de confiance asymptotiques ponctuels pour l'opérateur r à l'aide d'estimations des différentes constantes au travers de quelques simulations. Cet article contient aussi une application de ces méthodes aux données environnementales liées à l'étude du phénomène El Nino présentées précédemment qui permet d'illustrer l'intérêt de notre approche non-paramétrique fonctionnelle par rapport à une approche non-paramétrique multivariée (modèle g.a.m.) ou linéaire fonctionnelle.

Convergence en norme $\mathbb{L}^p$

Je me suis ensuite intéressé à l'étude des propriétés de convergence en norme $\mathbb{L}^p$ de l'estimateur (2). Des premiers résultats, concernant les vitesses de convergence des erreurs $\mathbb{L}^p$, sont proposés par Dabo-Niang et Rhomari (2003) dans le cas d'échantillons indépendants. Il serait intéressant de donner l'expression des termes asymptotiquement dominants des erreurs $\mathbb{L}^p$ car cela ouvrirait des perspectives notamment en ce qui concerne le choix du paramètre de lissage. L'idée principale de notre approche est d'utiliser le résultat de normalité asymptotique précédent et des arguments d'uniforme intégrabilité.
On note $B = \phi'(0)\frac{M_0}{M_1}$ et $V = \sqrt{\frac{M_2\sigma_\epsilon^2}{M_1^2}}$. Sous des conditions générales, on montre dans le cas d'échantillons indépendants ou α-mélangeants que pour tout $q \leq \ell$ (ℓ dépend des hypothèses) on a :

$$\mathbb{E}\left[|\hat{r}(x) - \mathbb{E}\left[\hat{r}(x)\right]|^q\right] = \frac{V}{\left(nF(h_n)\right)^{\frac{q}{2}}} \left(\mathbb{E}\left[|W|^q\right] + o(1)\right)$$

et

$$\mathbb{E}\left[|\hat{r}(x) - r(x)|^q\right] = \mathbb{E}\left[\left|h_n B + W\frac{V}{\sqrt{nF(h_n)}}\right|^q\right] + o\left(\frac{1}{\left(nF(h_n)\right)^{\frac{q}{2}}}\right),$$

où W est une variable gaussienne centrée réduite. On peut également obtenir le même type de résultat sans les valeurs absolues. De plus, l'expression des erreurs $\mathbb{L}^p$ peut être donnée de manière plus explicite lorsque l'exposant est entier. Cependant, le cas où l'exposant est impair nécessite l'introduction de notations que nous n'avons pas la place de présenter dans ce résumé. Tous ces résultats ont donné lieu à une publication dans la revue des Annales de l'ISUP (Delsol, 2007). Ils ouvrent des perspectives intéressantes en terme du choix optimal du paramètre de lissage et semblent également innovants dans le cadre multivarié où les seuls résultats de ce type pour l'erreur $\mathbb{L}^1$ sont donnés par Wand (1990) dans le cas réel et à design fixé. A noter également qu'une note consacrée à une présentation succinte des résultats de normalité asymptotique et l'expression des erreurs $\mathbb{L}^p$ a été publiée dans la revue des Comptes Rendus de l'Académie des Sciences.

Régression robuste

Les résultats précédents concernent l'estimation de l'opérateur de régression r défini dans le modèle (1) qui correspond à la moyenne conditionnelle de la variable d'intérêt Y sachant la variable explicative fonctionnelle $\mathcal{X}$. Afin de proposer des méthodes de prédiction plus robustes à la présence de données aberrantes il peut être intéressant de proposer des méthodes d'estimation basées sur des caractéristiques différentes de la loi conditionnelle de Y sachant $\mathcal{X}$. Les estimateurs robustes introduits par Azzedine *et al.* (2006) sont une manière générale de répondre à ces problèmes. On s'intéresse à un opérateur t défini par l'équation

$$C\left(t\left(\mathcal{X}\right),\mathcal{X}\right) := \mathbb{E}\left[\Psi_{\mathcal{X}}\left(Y,t\left(\mathcal{X}\right)\right)|\mathcal{X}\right] = 0,$$

où Ψ est une fonction connue. Ce type de problème est une généralisation du modèle (1) que l'on retrouve en prenant $\Psi_x\left(y,s\right) = y - s$. Il permet de s'intéresser à des problèmes d'estimation plus variés tels que les quantiles conditionnels par exemple. Bien entendu, le critère C est inconnu et doit être remplacé par sa version empirique. On se fixe un élément x de $\mathcal{E}$ et on définit sur l'ensemble $A_n\left(x\right) := \left\{\sum_{i=1}^n K\left(\frac{d(\mathcal{X}_i,x)}{h_n}\right) > 0\right\}$ (dont la probabilité tend vers 1) l'estimateur $\hat{t}\left(x\right)$ comme solution (en s) de l'équation

$$\hat{M}\left(s,x\right) := \frac{\sum_{i=1}^n \Psi_x\left(Y,s\right) K\left(\frac{d(\mathcal{X}_i,x)}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{d(\mathcal{X}_i,x)}{h_n}\right)} = 0.$$

Certaines propriétés asymptotiques (convergence presque sûre, normalité asymptotique) de cet estimateur ont été considérées par Azzedine *et al.* (2006) et Attouch *et al.* (2007). J'ai eu l'occasion de travailler avec Christophe Crambes et Ali Laksaci sur les propriétés de convergence en norme $\mathbb{L}^p$ de ce type d'estimateurs. Nous avons obtenu des résultats tout à fait similaires à ceux présentés précédemment pour l'estimateur $\hat{r}$ à la fois dans le cas indépendant et dans le cas α-mélangeant. Cette collaboration a donné lieu à une publication dans la revue Journal of Nonparametric Statistics (Crambes *et al.*, 2008) ainsi qu'à une publication dans les actes du congrès IWFOS de Toulouse en 2008.

Tests de structure en régression sur variable fonctionnelle

Revenons à présent au modèle de régression sur variable fonctionnelle (1) et supposons que nous disposons d'un échantillon de variables indépendantes. Avant d'essayer d'estimer d'une manière ou d'une autre l'opérateur inconnu r, il semble pertinent de se poser la question de l'existence d'un lien entre la variable explicative et la variable d'intérêt mais aussi de se demander si ce lien a une structure particulière. En effet, ces informations supplémentaires semblent utiles et importantes pour mieux choisir la méthode d'estimation que l'on va utiliser. De manière assez étonnante, s'il existe de nombreux résultats concernant les tests de structure en régression multivariée et de nombreuses méthodes d'estimation en régression sur variable fonctionnelle, il semble que le problème des tests de structure en régression sur variable fonctionnelle n'avait pas reçu une grande attention avant ces dernières années. En effet, avant 2009, la littérature semble réduite aux travaux de Cardot *et al.* (2003, 2004) concernant les tests d'hypothèse dans le cas particulier du modèle linéaire fonctionnel, un article de Gadiaga et Ignaccolo (2005) qui propose des tests de non effet basés sur

des méthodes de projection et une approche heuristique permettant de construire des tests d'adéquation basés sur l'analyse en composante principale fonctionnelle proposée par Chiou et Müller (2007). Il n'existait donc pas de résultat théorique général permettant de s'intéresser au problème de tester si le modèle est linéaire, à indice simple fonctionnel, si l'effet de la variable explicative fonctionnelle peut se réduire à l'effet d'un vecteur de points particuliers de celle-ci, ...

Notre objectif est de proposer une manière générale de construire des tests portant sur la structure de l'opérateur r. En d'autres termes, on se fixe une famille fermée $\mathcal{R}$ d'opérateurs de carré intégrable (par rapport à la loi de $\mathcal{X}$: $P_{\mathcal{X}}$) et on souhaite tester l'hypothèse nulle

$$\mathcal{H}_0 : \{\exists r_0 \in \mathcal{R}, P\left(r\left(\mathcal{X}\right) = r_0\left(\mathcal{X}\right)\right) = 1\},$$

contre des alternatives locales de la forme

$$\mathcal{H}_{1,n} : \inf_{r_0 \in \mathcal{R}} \mathbb{E}\left[\left(r - r_0\right)^2\left(\mathcal{X}\right) w\left(\mathcal{X}\right)\right] \geq \eta_n^2,$$

où w est un opérateur de poids à support W borné. Je me suis intéressé, en collaboration avec Frédéric Ferraty et Philippe Vieu, à la manière dont l'approche introduite par Härdle et Mammen (1993) dans le cas multivarié peut être adaptée au cas de variables fonctionnelles. Lorsque la famille $\mathcal{R}$ est convexe, on peut définir la projection r_0 de r sur $\mathcal{R}$. Pour tester si r appartient à la famille $\mathcal{R}$, on a alors envie de comparer r à sa projection r_0. Plaçons-nous pour commencer dans le cas le plus simple où $\mathcal{R}$ est réduite à un opérateur r_0 fixé a priori. La première idée que l'on peut avoir est de construire une statistique de test basée sur la comparaison d'un estimateur à noyau fonctionnel $\hat{r}$ et de r_0. Cependant, en construisant ce genre de statistique de test on ajoute un terme de biais. Si on peut le rendre négligeable en supposant assez de régularité de r et en considérant un noyau d'ordre plus élevé dans le cas multivarié, cela n'est plus vrai lorsque la variable explicative est fonctionnelle car on ne dispose pas de noyaux d'ordre supérieur. On propose donc, comme dans Härdle et Mammen (1993) de considérer un estimateur à noyau de la différence $\left(r - r_0\right)(x)$:

$$\frac{\sum_{i=1}^n \left(Y_i - r_0\left(\mathcal{X}_i\right)\right) K\left(\frac{d(\mathcal{X}_i, x)}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{d(\mathcal{X}_i, x)}{h_n}\right)}.$$

Pour des raisons techniques liées à la fois au fait que l'on veut éviter de se restreindre à un support compact pour w et qu'il est difficile d'estimer la densité de variables fonctionnelles, on choisit de supprimer le dénominateur. Comme dans la plus part des cas l'opérateur r_0 est inconnu, on le remplace par un estimateur $\hat{r}_0^{\mathcal{R}}$ spécifique à la famille $\mathcal{R}$. On aboutit ainsi à la statistique de test

$$T_n^{\mathcal{R}} = \int \left(\sum_{i=1}^n \left(Y_i - \hat{r}_0^{\mathcal{R}}\left(\mathcal{X}_i\right)\right) K\left(\frac{d\left(x, \mathcal{X}_i\right)}{h_n}\right)\right)^2 w\left(x\right) dP_{\mathcal{X}}\left(x\right).$$

On suppose également que $n < N$ et que l'estimateur $\hat{r}_0^{\mathcal{R}}$ est construit sur l'échantillon $\left(\mathcal{X}_i, Y_i\right)_{n+1 \leq i \leq N}$. Les termes dominants du biais et de la variance de la statistique de test sous l'hypothèse nulle sont apportés par les variables

$$\begin{aligned} T_{1,n} &= \int \sum_{i=1}^n K^2\left(\frac{d\left(X_i, x\right)}{h_n}\right) \epsilon_i^2 w\left(x\right) dP_X\left(x\right), \\ T_{2,n} &= \int \sum_{1 \leq i \neq j \leq n} K\left(\frac{d\left(X_i, x\right)}{h_n}\right) K\left(\frac{d\left(X_j, x\right)}{h_n}\right) \epsilon_i \epsilon_j w\left(x\right) dP_X\left(x\right), \end{aligned}$$

dont la distribution est la même sous l'hypothèse nulle et sous l'alternative.

Pour obtenir la normalité asymptotique de notre statistique de test on demande que sous l'ypothèse nulle l'estimateur $\hat{r}_0^{\mathcal{R}}$ converge assez vite vers r. La divergence sous l'alternative demande que la suite η_n ne converge pas trop vite vers 0, et seulement que l'estimateur $\hat{r}_0^{\mathcal{R}}$ soit régulier et reste assez proche de la famille $\mathcal{R}$. Sous ces hypothèses ainsi que d'autres conditions techniques assez générales on obtient que

– Sous $(\mathcal{H}_0)$, $\frac{1}{\sqrt{Var(T_{2,n})}}\left(T_n^* - \mathbb{E}\left[T_{1,n}\right]\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$,

– Sous $(\mathcal{H}_1)$, $\frac{1}{\sqrt{Var(T_{2,n})}}\left(T_n^* - \mathbb{E}\left[T_{1,n}\right]\right) \xrightarrow{P} +\infty$.

Ce résultat théorique général offre d'intéressantes perspectives pour construire des tests de structure généraux. Les hypothèses que nous considérons sont vérifiées par de nombreux estimateurs et permettant notamment de tester un modèle a priori, si l'opérateur r est constant (test de non effet) mais aussi si le modèle est linéaire, à indice simple fonctionnel, si l'effet de la variable explicative fonctionnelle se réduit à l'effet d'un vecteur de points particuliers de celle-ci, ... Le résultat théorique général et l'ensemble de ces exemples sont présentés dans une note publiée dans la revue des Comptes Rendus de l'Académie des Sciences, sous une version un peu plus longue (dans laquelle sont aussi évoquées les méthodes de bootstrap décrites plus bas) dans les actes de la conférence IWFOS de Toulouse en 2008 et sont détaillés dans article publié dans la revue Journal of MultiVariate Analysis (Delsol *et al.*, 2011).

Etant donné la nature complexe des termes dominants du biais et de la variance de notre statistique de test, on peut s'attendre à de mauvais résultats si on cherche à utiliser directement la loi asymptotique pour calculer la valeur de seuil de notre test. On propose d'utiliser des méthodes de rééchantillonnage (voir par exemple Efron, 1979, Mammen, 1993, Härdle *et al.*, 2005) permettant de générer à partir de l'échantillon original B échantillons de même distribution que celui-ci mais pour lesquels l'hypothèse nulle est approximativement vérifiée puis de choisir comme valeur seuil (pour un test de niveau α) le quantile empirique d'ordre $1-\alpha$ (que l'on note $\tau^*_{[B(1-\alpha)]}$) des valeurs prises par notre statistique de test sur ces échantillons. Pour générer ces écantillons on propose de laisser les $\mathcal{X}_i$ inchangés, de générer par rééchantillonnage de nouveaux résidus ϵ_i^* puis de générer des réponses $Y_i^* = \hat{r}_0^{\mathcal{R}}(\mathcal{X}_i) + \epsilon_i^*$ pour lesquelles l'hypothèse nulle est approximativement vérifiée. Différentes méthodes de rééchantillonnage (naïf, naïf et lissé, sauvage) sont considérées et comparées au travers de simulations dans le cas des tests de non effet. On observe notament le bon comportement général de nos procédures de test ainsi que la manière dont le choix des paramètres (paramètre de lissage, nombre de boucles de rééchantillonnage) influence leurs performances. Tout cela, ainsi qu'une application des tests de non effet à l'étude de données spectrométriques dont je vais parler plus en détails par la suite, a fait l'objet d'un article publié dans la revue Computational Statistics (Delsol, 2012). Enfin, on propose d'approcher l'intégrale qui apparaît dans la définition de la statistique de test par une moyenne empirique sur un troisième sous-échantillon. Il serait intéressant d'étudier par la suite d'autres approches ne nécessitant pas de découper l'échantillon original en trois sous-échantillons.

Application à des données spectrométriques

Revenons à l'étude des données spectrométriques présentées au début de ce résumé. L'étude de ces jeux de données en chimie quantitative est souvent précédée d'une transformation des données qui revient en quelque sorte à considérer des dérivées des courbes spectrométriques. Il a également été remarqué lors de précédentes études statistiques que les dérivées des courbes spectrométriques sont de bons prédicteurs dans ce type de situation. On peut donc se poser la question de savoir quelles sont les dérivées qui ont un effet significatif pour prédire le taux de graisse des morceaux de viande. Afin de considérer ce problème, on propose de faire des tests de non effet en prenant comme variables explicatives les dérivées successives (d'ordre 0 à 4) des courbes spectrométriques. Une fois enlevé du taux de graisse l'effet des dérivées d'ordre 2 et 3, on observe que les autres dérivées n'ont pas d'effet significatif. D'autre part, des tests de linéarité font apparaître que certaines dérivées ont un effet significativement non linéaire et qu'il vaut mieux utiliser un estimateur non-paramétrique tandis que pour les autres un estimateur linéaire donne des résultats meilleurs concernant l'erreur de prédiction.

On s'est également intéressé à l'étude d'un échantillon de plus petite taille provenant de l'étude spectrométrique de 80 échantillons de maïs pour lesquels on désire prédire le taux de moisissure à partir de l'observation de la courbe spectrométrique (observée en 700 longueurs d'onde réparties entre 1100nm et 2500 nm). Cet échantillon est disponible sur internet à l'adresse http ://software.eigevector.com/Data/Corn/index.html. Sur cet échantillon on se pose également la question de tester si des portions de la courbe ont un effet significatif. On fait ainsi ressortir que certaines parties n'ont pas d'effet significatif. On observe également qu'en utilisant comme prédicteur une portion significative de la courbe au lieu de toute la courbe on obtient de

meilleurs résultats de prédiction. Les tests de linéarité effectués ne rejettent pas l'hypothèse de linéarité. On peut donc utiliser un estimateur linéaire ce qui permet d'obtenir de meilleurs résultats. Comme indiqué précédemment, ce travail a fait l'objet d'un article publié dans la revue Computational Statistics (Delsol, 2012).

Afin de poursuivre l'étude de ces données, il serait intéressant de considérer des modèles semi-paramétriques de régression sur variable fonctionnelle permettant d'inclure dans la construction de notre modèle les connaissances que l'on a de ces problèmes de prédiction à partir de courbes spectrométriques dans d'autres disciplines (chimie, agronomie, biologie, ...).

Choix de la pseudo-métrique par validation croisée

Comme nous l'avons expliqué précédemment en (2), la construction d'estimateurs à noyaux pour des données fonctionnelles fait intervenir une pseudo-métrique permettant plus de flexibilité dans la topologie utilisée. Son choix ne doit pas être fait à la légère car il peut avoir une grande importance dans la qualité de l'estimateur qui en découle. L'étude des propriétés asymptotiques de l'estimateur défini en (2), mettent en évidence un terme de biais du type $O\left(h^{\beta_d}\right)$ (avec β_d tel que r est β_d-Hölderienne par rapport à d) et un terme de variance de la forme $O\left(\sqrt{nF_x(h)}^{-1}\right)$. L'utilisation de métriques usuelles conduit dans certains cas (par exemple pour les processus gaussiens) à des probabilités de petites boules $F_x(h)$ très petites et donc à une variance forte. On pourrait alors proposer d'utiliser des pseudo-métriques de type projection pour obtenir des probabilités de petites boules $F_x(h)$ plus grandes mais on demande alors davantage de régularité à l'opérateur de régression r. Il s'agit, comme souvent en statistique, de faire un compromis entre les termes de biais et variance, c'est à dire de chercher une pseudo-métrique qui dégage de la variable explicative $\mathcal{X}$ les caractéristiques influançant Y. Une manière d'y parvenir, est de choisir parmi une famille de pseudo-métriques (choisie en fonction de la nature des courbes) celle qui est optimale par validation croisée. Nous nous sommes intéressés, avec Catherine Timmermans et Rainer von Sachs, à cette problématique dans le cas d'une pseudo-métrique définie à partir de la projection des variables fonctionnelles sur une base d'ondelette (le problème étant alors de sélectionner les coefficients à retenir pour expliquer Y). Le résultat majeur de ce travail, montre que, sous certaines hypothèses,

$$\frac{MISE(\hat{d})}{min_{d\in\mathcal{D}_n} MISE(d)} \overset{n\to+\infty}{\to} 1,$$

où $\hat{d}$ est la pseudo-métrique choisie par validation croisée et $\mathcal{D}_n$ la famille de pseudo-métriques considérée. Ce résultat étant établi dans un cadre général, il peut être utilisé avec d'autres familles de pseudo-métriques (projection sur d'autres bases, pseudo-métriques de nature variées, ...). Ce travail, publié dans la revue Journal of MultiVariate Analysis (Timmermans *et al.*, 2013), constitue une première justification théorique du choix de la pseudo-métrique par validation croisée (approche courament utilisée en pratique).

M-estimation avec critère non-lisse

Je travaille avec Ingrid Van Keilegom dans le cadre du projet ERC "M- and Z-estimation in semiparametric statistics : applications in various fields". Plus précisément, on s'intéresse à l'étude de problèmes de M-estimation où l'objectif est d'estimer un paramètre d'intérêt θ_0 qui maximise un critère semi-paramétrique

$$M(\theta, h_0) = \mathbb{E}\left[m\left(X_i, \theta, h_0\right)\right]$$

dans lequel h_0 est un paramètre de nuisance inconnu et la fonction m n'est pas dérivable par rapport à θ. On considère alors comme estimateur de θ_0 une valeur qui maximise le critère empirique

$$M_n\left(\theta, \hat{h}\right) := \frac{1}{n}\sum_{i=1}^{n} m\left(X_i, \theta, \hat{h}\right).$$

L'objectif est de faire le lien entre l'approche considérée par Chen *et al.* (2003) dans le cas de Z-estimateurs et celle proposée par Van der Vaart et Wellner (1996) pour étudier des M-estimateurs lorsqu'il n'y a pas de paramètre de nuisance. On se sert notament d'hypothèses portant sur l'entropie de la famille à laquelle appartient le paramètre de nuisance ainsi qu'à des résultats fins sur les processus empiriques pour obtenir sous des conditions assez générales la consistance ($\hat{\theta}_n \xrightarrow{P} \theta_0$), la vitesse de convergence ($\exists r_n,\, r_n \left\|\hat{\theta}_n - \theta_0\right\| = O_p(1)$), et la distribution asymptotique de ces estimateurs :

$$\exists \mathbb{G} \text{ processus gaussien, } r_n\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{\mathcal{L}} \arg\max_{\gamma} \mathbb{G}(\gamma).$$

L'article présentant ces résultats théoriques et quelques exemples d'application de problèmes que nos résultats permettent de considérer est sur le point d'être soumis. L'étude de Z- ou M-estimateurs conditionnels en régression sur variable fonctionnelle est une perspective de travail pour les années à venir.

Travaux en cours

Segmentation d'images hyperspectrales

Décomposer une image en un ensemble de régions (c'est à dire groupes de pixels) homogènes est un problème classique, appelé segmentation, en traitement d'image. Ces régions ont habituellement une signification concrète et correspondent à des éléments spécifiques de la scène que l'on cherche à identifier (par exemple fond, pulpe, écorce dans l'image Fig 3.a)). De nombreuses méthodes ont été proposées pour segmenter des images en niveaux de gris ou multispectrales (voir par exemple Mumford et Sha, 1989, Maître, 2003 ou Aubert et Kornprobst, 2006 et les références qu'ils contiennent). Suite au travail précursuer de Besag (1989), différentes méthodes de segmentation ont été construites à partir d'approches bayésiennes (voir par exemple les travaux récents de Deng and Clausi, 2004, Orbanz et Buhmann, 2008, Chen *et al.*, 2010, et les références qu'ils contiennent) L'approche par maximum a posteriori (M.A.P.) est une méthode de segmentation par détection de régions qui a fait ses preuves. Elle consiste à déterminer l'image segmentée x la plus vraisemblable conditionnellement à l'image originale y. Cela revient à trouver (via la formule de Bayes)

$$\begin{aligned} x_{MAP} &= \operatorname{argmax}_x \ \mathbb{P}(\mathcal{X} = x | \mathcal{Y} = y). \\ &= \operatorname{argmax}_x f_{\mathcal{Y}|\mathcal{X}=x}(y)\, \mathbb{P}(\mathcal{X} = x). \end{aligned}$$

Un a prioiri de type champs de Potts est introduit sur $\mathcal{X}$ pour modéliser la régularité spatiale au sein de l'image segmentée. Tandis que $f_{\mathcal{Y}|\mathcal{X}=x}(y)$ peut être estimée (sous certaines hypothèses) à partir des densitées estimées (souvent au travers de densitées gaussiennes) sur chacunes des régions définies par x.

Je travaille depuis quelques mois avec Cécile Louchet (MAPMO) sur la généralisation de ces méthodes de segmentation dans le cas d'images hyperspectrales (ou plus généralement d'images pour lesquelles à chaque pixel est associé une courbe - discrétisée en un grand nombre de points). Voici un exemple (disponible à l'url `www.cs.columbia.edu/CAVE/databases/multispectral/real\_and\_fake`) pour lequel chaque pixel de l'image est décrit au travers de 31 mesures de réflectance correspondant à des longueurs d'onde allant de 400 à 700 nm (avec un pas de 10 nm).

En combinant des travaux récents en statistique non paramétrique fonctionnelle concernant l'estimation de la densité de variables fonctionnelles (voir notamment Dabo-Naing, 2002) avec l'approche décrite ci-dessus, nous proposons une méthode innovante. La recherche du maximum a posteriori se fait au travers d'un algorithme de type Mode Conditionnel Itéré. Les premiers tests que nous avons effectués sur des données simulées et sur des images réelles sont prometteurs (voir par exemple Figure 4).

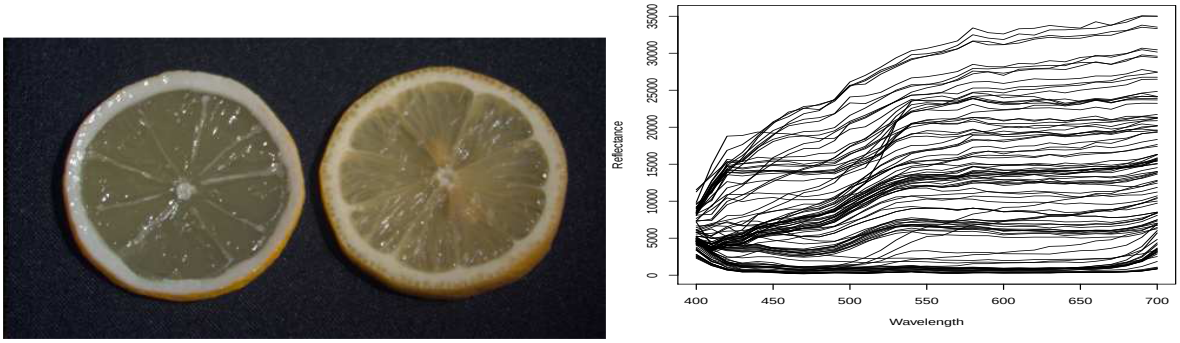

FIGURE 3 – a) Fausse et véritable tranches de citron. b) Échantillon de 100 courbes de réflectance.

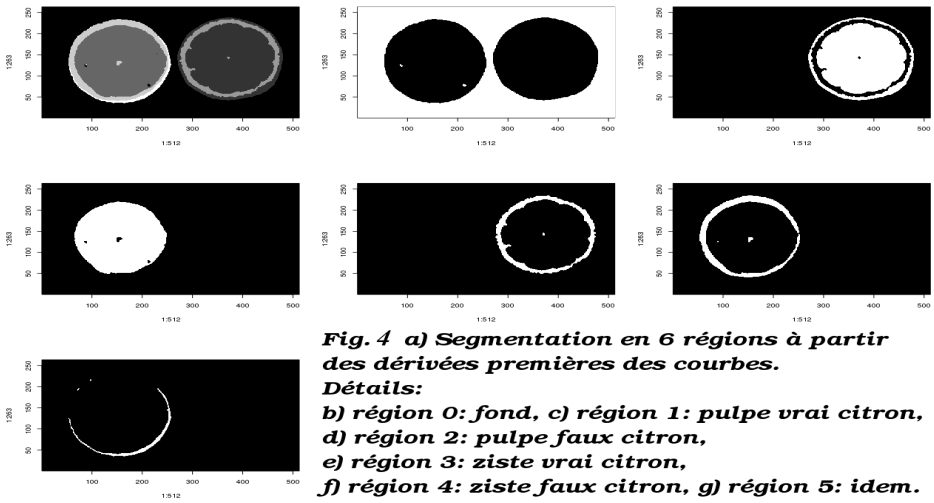

Fig. 4 a) Segmentation en 6 régions à partir des dérivées premières des courbes. Détails: b) région 0: fond, c) région 1: pulpe vrai citron, d) région 2: pulpe faux citron, e) région 3: ziste vrai citron, f) région 4: ziste faux citron, g) région 5: idem.

Conclusions et perspectives

Au travers de mon travail de recherche je me consacre à l'étude des aspects théoriques et appliqués de la statistique, et plus particulièrement de la statistique fonctionnelle. Ce domaine de la statistique devient de plus en plus populaire à cause de l'intérêt des problèmes concrets qu'il permet d'envisager mais aussi de la grande variété des perspectives de recherche qui lui sont rattachées. L'enjeu est d'apporter une réponse à différents types de questions concrètes liées à l'étude de grand jeux de données (temporelles par exemple). Cela amène à considérer de nouveaux modèles et à développer des outils théoriques innovants que l'on pourra ensuite appliquer à nos données. Ce mécanisme de va et vient entre théorie et applications fait des échanges que l'on peut avoir avec d'autres disciplines (agronomie, économie, biologie, traitement d'image, finance, environnement, chimie, ...) un point clé d'avancement de la recherche pour les années à venir.

Références

AIT-SAIDI, A., FERRATY, F., KASSA, R. and VIEU, P. (2008) Cross-validated estimations in the single functional index model. *Statistics*, **42**, 475-494.

ATTOUCH, M., LAKSACI, A. and OULD-SAID, E. (2007) Strong uniform convergence rate of robust estimator of the regression function for functional and dependent processes. *Technical Report L.M.P.A.*.

ANEIROS-PEREZ, G. and VIEU, P. (2008) Nonparametric time series prediction : a semi-functional partial

linear modeling. *J. Multivariate Anal.* **99** (5) 834-857.
AZZEDINE, N., LAKSACI, A. and OULD-SAID, E. (2006) On the robust nonparametric regression estimation for functional regressor. *Technical Report L.M.P.A.*
BOSQ, D. (1991) Modelization, nonparametric estimation and prediction for continuous time processes. In Nonparametric functional estimation and related topics (Spetses, 1990), 509-529, *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.* **335**, Kluwer Acad. Publ., Dordrecht.
BOSQ, D. (2000) *Linear Processes in Function Spaces : Theory and Applications* Lecture Notes in Statistics **149** Springer-Verlag, New York.
BRADLEY, R.C. (2005) Basic properties of strong mixing conditions. A survey and some open questions. Update of, and a supplement to, the 1986 original. *Probab. Surv.* **2** 107-144 (electronic).
CARDOT, H., FERRATY, F. and SARDA, P. (1999) Functional Linear Model *Statist. and Prob. Letters* **45** 11-22.
CARDOT, H., FERRATY, F., MAS, A. and SARDA, P. (2003) Testing Hypothesys in the Functional Linear Model *Scandinavian Journal of Statistics* **30** 241-255.
CARDOT, H., GOIA, A. and SARDA, P. (2004) Testing for no effect in functional linear regression models, some computational approaches. *Comm. Statist. Simulation Comput.* **33** (1) 179-199.
CHEN, X., LINTON, O. and VAN KEILEGOM, I. (2003) Estimation of semiparametric models when the criterion function is not smooth. (English summary) *Econometrica* **71** (5) 1591-1608.
CHIOU, J.M. and MULLER H.G. (2007) Diagnostics for functional regression via residual processes *Computational Statistics & Data Analysis* **51**, (10) 4849-4863.
CRAMBES C., DELSOL L. and LAKSACI A. (2008) Robust nonparametric estimation for functional data. *Journal of Nonparametric Statistics,* **20**, (7), pp 573-598.
CRAMBES, C., KNEIP, A. et SARDA, P. (2009) Smoothing splines estimators for functional linear regression. *Annals of Stat.*, **37**, 35-72.
DABO-NIANG, S. and RHOMARI, N. (2003) Estimation non paramétrique de la régression avec variable explicative dans un espace métrique. (French. English, French summary) [Kernel regression estimation when the regressor takes values in metric space] *C. R. Math. Acad. Sci. Paris* **336** (1) 75–80.
DAUXOIS, J., POUSSE, , A. and ROMAIN, Y. (1982) Asymptotic theory for the principal component analysis of a vector random function : some applications to statistical inference *J. Multivariate Anal.* **12** (1) 136-154.
DELSOL L. (2007) Régression non-paramétrique fonctionnelle : Expressions asymptotiques des moments. *Annales de l'I.S.U.P.*, **LI**, (3), pp 43-67.
DELSOL L. (2009) Advances on asymptotic normality in nonparametric functional Time Series Analysis. *Statistics*, **43**, (1), pp 13-33.
DELSOL L. (2012) No effect tests in regression on functional variable and some applications to spectrometric studies, *Computational Statistics*, 1-37.
DELSOL L., FERRATY F. and Vieu P. (2011) Structural test in regression on functional variables. *J. Multivariate Analysis* **102**(3), 422-447.
EFRON, B. (1979) Bootstrap Methods : Another Look at the Jackknife. *Annals Statist.* **7** (1) 1-26.
FERRATY, F., MAS, A. and VIEU, P. (2007) Advances on nonparametric regression for fonctionnal data. *ANZ Journal of Statistics* **49** 267-286.
FERRATY F., and ROMAIN Y. (2011). *Oxford Handbook of Functional Data Analysis* (Eds.). Oxford University Press.
FERRATY, F. and VIEU, P. (2000) Dimension fractale et estimation de la régression dans des espaces vectoriels semi-normés *Compte Rendus de l'Académie des Sciences* Paris, **330**, 403-406.
FERRATY, F. and VIEU, P. (2006) *Nonparametric Functional Data Analysis.* Springer-Verlag, New York.
GADIAGA, D. and Ignaccolo, R.(2005) Test of no-effect hypothesis by nonparametric regression. *Afr. Stat.* **1** (1) 67-76.
GRENANDER, U. (1981) *Abstract inference.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York. ix+526 pp.
HARDLE, W. and MAMMEN, E. (1993) Comparing Nonparametric Versus Parametric Regression Fits *The*

Annals of Statistics **21**, (4) 1926-1947.
HARDLE, W., MAMMEN, E. and PROENCA, I. (2005) A Bootstrap Test for Single Index Models *Econometrics* 0508007 EconWPA.
LEURGANS, S. E., MOYEED, R. A. and SILVERMAN, B. W. (1993) Canonical correlation analysis when the data are curves. *J. Roy. Statist. Soc. Ser. B* **55** (3) 725-740.
MAMMEN, E. (1993) Bootstrap and wild bootstrap for high-dimensional linear models. *Ann. Statist.* **21** (1) 255-285.
MASRY, E. (2005) Nonparametric regression estimation for dependent functional data : asymptotic normality *Stochastic Process. Appl.* **115** (1) 155-177.
RAMSAY, J.O. (1982) When the data are functions. *Psychometrika* **47** (4) 379-396.
RAMSAY, J. and DALZELL, C. (1991) Some tools for functional data analysis *J.R. Statist. Soc. B.* **53** 539-572.
RAMSAY, J. and SILVERMAN, B. (1997) *Functional Data Analysis* Springer-Verlag, New York.
RAMSAY, J. and SILVERMAN, B. (2002) *Applied functional data analysis : Methods and case studies* Springer-Verlag, New York.
RAMSAY, J. and SILVERMAN, B. (2005) *Functional Data Analysis (Second Edition)* Springer-Verlag, New York.
RAO, C. R. (1958) Some statistical methods for comparison of growth curves. *Biometrics* **14** 1-17.
ROSENBLATT, M. (1956) A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci. U. S. A.* **42** 43-47.
YTIMMERMANS C., DELSOL L. and VON SACHS R. (2013) Using Bagidis in nonparametric functional data analysis : Predicting from curves with sharp local features. *J. Multivariate Analysis* (MA) **115**, 421-444.
TUCKER, L.R. (1958) Determination of parameters of functional equations by factor analysis. *Psychometrica* **23** 19-23.
VAN DER VAART, A. W., and WELLNER, J.A. (1996) *Weak convergence and empirical processes. With applications to statistics.* Springer Series in Statistics. Springer-Verlag, New York. xvi+508 pp.
WAND, M.P. (1990) On exact $\mathbb{L}^1$ rates in nomparametric kernel regression. *Scand.J.Statist* **17** (3) 251-256.