

Curriculum vitae

Nom, prénoms : LEONTE (ép. CIUPERCA) Gabriela

Année et lieu de naissance : 1966, Roumanie

Statut Marital : mariée, deux enfants (8 et 13 ans)

Nationalité : française

Fonction actuelle : Maître de conférences à l'Université Lyon 1

ADRESSE PROFESSIONNELLE :

Université Lyon 1, UMR 5208, Institut Camille Jordan,

Bat. Braconnier, 43, blvd du 11 novembre 1918, F - 69622 Villeurbanne, France,

tél : 33(0)4.72.43.16.90, fax : 33(0)4.72.43.16.87

Mail : Gabriela.Ciuperca@univ-lyon1.fr

Page web : <http://math.univ-lyon1.fr/~gciuperca/>

CURSUS UNIVERSITAIRE

- 27 novembre 2009 : **Habilitation à Diriger des Recherches**, Discipline : Statistique, Titre : “Inférence statistique pour des modèles non-identifiables, Estimation du taux d'entropie, Applications des Statistiques”. Université Lyon 1.
- 1992-1996 : **Thèse de Doctorat**, Spécialité : Statistique Appliquée, Titre : “Modélisation du métabolisme de la glucose” Mention “très honorable avec félicitations”, Institut National Agronomique Paris. Sous la direction du Professeur Richard Tomassone.
- 1991-1992 : **DEA Modélisation Stochastique et Statistique**, Mention “Assez Bien”, Université Paris XI Orsay. Mémoire sous la direction du Professeur Camille Duby.
- 1984-1988 : **Maîtrise de Mathématique et d'Informatique**, Mention “Très Bien”, Faculté de Mathématique, Université de Iasi, Roumanie .

CARRIÈRE

- **à partir du 1.09.2001** Maître de Conférences à l'Univ. Lyon 1
- **1.09.1998-31.08.2001** Maître de Conférences à l'IUT de Poitiers
- **1997-1998** ATER à l'Université Paris 13
- **1990-1991** Maître Assistant en Roumanie

RECONNAISSANCE INTERNATIONALE :

- Membre élue en août 2005 de l'*International Statistical Institut (ISI)*. On peut faire partie de cet institut si on s'est distingué par sa contribution au développement ou à l'application des méthodes statistiques.
 - Membre de l'*ERCIM Working Group* (The European Research Consortium for Informatics and Mathematics)-*Mixture Models*. Ce groupe de travail, qui inclut des chercheurs de toute l'Europe, se focalise sur l'aspect numérique des modèles de mélange de densités (populations) avec un intérêt particulier sur les applications (économie, médecine et épidémiologie, physique, chimie, ...).
-

Les principaux thèmes de recherche abordés depuis la thèse de doctorat :

- **statistique mathématique** :
 - inférence statistique pour des mélanges de densités ;
 - estimation paramétrique dans un modèle non-linéaire avec plusieurs points de rupture ;
 - estimation paramétrique et non-paramétrique des taux d'entropie(Shannon, Rényi, Tsallis, ...) pour des chaînes de Markov ;
 - modèles avec des variables à longue mémoire ;
 - méthodes de type LASSO ;
 - estimation non-paramétrique, maximum de vraisemblance empirique.
 - **applications des statistiques** : en médecine, à la pollution de l'air,
-

Table des matières

1 Publications, Contrats de Recherche, Encadrement	3
1.1 Publications	3
1.1.1 <i>Articles Publiés, acceptés</i>	3
1.1.2 <i>Articles soumis, en révision</i>	4
1.2 Communications dans des Conférences avec actes	4
1.3 Contrats de recherche	5
1.4 Encadrement	5
1.5 Autres activités scientifiques	6
2 Présentation des thèmes de recherche	7
2.1 Mélanges de densités	7
2.2 Modèle avec rupture, variables indépendantes	8
2.3 Estimation des taux d'entropie	10
2.4 Modèles à longue mémoire	12
2.5 Estimation semi- ou non-paramétrique	12
2.5.1 Maximum de vraisemblance(MV) empirique	12
2.5.2 Estimation semi-paramétrique pour des lois des valeurs extrêmes .	13
2.6 Applications des statistiques	13
2.7 Perspectives	14

1 Publications, Contrats de Recherche, Encadrement

1.1 Publications

1.1.1 *Articles Publiés, acceptés*

Statistique Mathématique

1. Ciuperca G.(2012b), "Empirical likelihood for nonlinear model with missing responses", accepté à *Journal of Statistical Computation and Simulation*.
2. Ciuperca G.(2011e), "A general criterion to determine the number of change-points", *Statistics and Probability Letters*, Vol. 81, No. 8, p. 1267–1275, 2011.
3. Ciuperca G.(2012a), "The S-estimator in change-point random model with long-memory", à paraître dans *Statistics*.
4. Ciuperca G.(2011d), Girardin V., Lhote L., "Computation and estimation of generalized entropy rates for denombrable Markov chains", *IEEE Transactions on Information Theory*, Vol. 57, No. 7, p. 4026–4034, 2011.
5. Ciuperca G.(2011c), "Asymptotic behaviour of the LS estimator in a nonlinear model with long memory", *Journal of the Korean Statistical Society*, Vol. 40, p. 193-203, 2011.
6. Ciuperca G.(2011b), "Penalized least absolute deviations estimation for nonlinear model with change-points", *Statistical Papers*, Vol. 52, No. 2, p. 371-390, 2011.
7. Ciuperca G.(2011a), "Estimating nonlinear model with and without change-points by the LAD method", *Annals of the Institute of Statistical Mathematics*, Vol. 63, No. 4, p. 717-743, 2011.
8. Ciuperca G., Mercadier C., (2010), "Semi-parametric estimation for heavy tailed distributions", *Extremes*, Vol. 13, No. 1, p. 55-87, 2010.
9. Ciuperca G.(2009), "The M-estimator in a multi-phase random regression model", *Statistics and Probability Letters*, Vol. 75, No. 5, p. 573–580, 2009.
10. Ciuperca G., Dapzol N.(2008), "Maximum Likelihood Estimator in a Multi-phase Random Regression Model", *Statistics*, Vol. 42, No. 4, p. 363–381, 2008.
11. Ciuperca G., Girardin V.(2007), "Estimation of the entropy rate of a countable Markov Chain", *Communications in Statistics- Theory and Methods*, Vol. 36, No. 14, p. 2493–2508, 2007.
12. Ciuperca G.(2004), "Maximum Likelihood Estimator in a two-phase Nonlinear Random Regression Model", *Statistics and Decision*, Vol. 22, No. 4, p. 335–349, 2004.
13. Ciuperca G., Ridolfi A., Idier J.(2003), "Penalized Maximum Likelihood Estimator for Normal Mixtures", *Scandinavian Journal of Statistics*, Vol. 30, No. 1, p. 45–59, 2003.
14. Ciuperca G.(2002), "Likelihood Ratio Statistic for Exponential Mixtures", *An-*

nals of the Institute of Statistical Mathematics, Vol. 54, No. 3, p. 585–594, 2002.

15. Ciuperca G.(1999), "Sur le test de maximum de vraisemblance pour le mélange de populations", *C.R.A.S.*, Série I, Tome 328, No. 4, Février 1999.

Applications des statistiques

16. Rouby, C., Thomas-Danguin, T., Vigouroux, M, **Ciuperca G.**, Jiang T., Alexanian, J., Barges, M., Gallice, I., Demolis, M., Degraix, J.L., Sicard, G. "The Lyon Clinical Olfactory Test : validation and measurement of hyposmia and anosmia in healthy and demented populations", accepté à *International Journal of Otolaryngology*. **17.** Chambon V., Franck N., Koechlin E., Fakra E., **Ciuperca G.**, Azorin J.-M., Farrer C.(**2008**), "The architecture of cognitive control in schizophrenia", *Brain*, 131, p. 962–970, 2008.

18. Plesa A., **Ciuperca G.**, Louvet V., Pujo-Menjouet L., Génieys S., Dumontet C., Thomas X., Volpert V.(**2006**), "Diagnostics of the AML with immunophenotypical data", *Mathematical Modelling of Natural Phenomena*, Vol. 1 No. 2, p. 104–123, 2006.

19. Bel L., Bellanger L., Bonneau V, **Ciuperca G.**, Dacunha-Castelle D., Deniau C., Ghattas B., Misiti M., Misiti Y., Oppenheim G., Poggi J.M., Tomassone R.(**1999**), "Eléments de comparaison de prévisions statistiques des pics d’ozone", *Revue de Statistique Appliquée*, XLVII(3), p. 7–25, 1999.

20. Bel L., Bellanger L., **Ciuperca G.**, Dacunha-Castelle D., Gilibert E., Jakubowicz P., Oppenheim G., Tomassone R.(**1998**), "On Forecasting Ozone Episodes in the Paris Area", *Biometrical Letters*, Vol. 35, No. 1, p. 37–66, 1998.

21. Ciuperca G.(1998), "A method to treat the dynamical Statistical Models", *Journal of Biological Systems*, Vol. 6, No. 4, p. 357–375, 1998.

22. Ciuperca G.(1998), "Influence de la matrice de covariances dans des modèles décrits par un système d’équations différentielles", *Revue de Statistique Appliquée*, XLVI(2), p. 59–81, 1998.

1.1.2 Articles soumis, en révision

23. Ciuperca G., "Empirical likelihood for nonlinear model with missing responses", soumis.

1.2 Communications dans des Conférences avec actes

- 1. Ciuperca G.**, "Estimation robuste dans un modèle paramétrique avec rupture", *41èmes Journées de Statistique, 25-29 mai 2009, Bordeaux*.
- 2. Ciuperca G., Girardin V.**, "On the estimation of the entropy rate of finite Markov chains ", *XIth International Symposium on ASMDA (Applied Stochastic Models and Data Analysis), 17-20 May 2005, Brest, France*.

3. **Ciuperca G.**, "Estimateur du maximum de vraisemblance pénalisé pour des mélanges de densités dégénérées", *33èmes Journées de Statistique A.S.U.*, 14-18 mai 2001, Nantes.
4. **Ciuperca G.**, "Prévision de la concentration d'ozone dans la région parisienne", *XXIXèmes Journées de Statistique*, 26-30 mai 1997, Carcassonne.

1.3 Contrats de recherche

1) **Contrat de recherche entre l'Université Paris XI Orsay et AIRPARIF (privé)** : (période septembre 1995 et août 1997) sur la prévision des pointes de pollution dans la région parisienne. A la suite de ce contrat, les articles no. 12 et 13 ont été publiés et surtout, la prévision de la pollution d'Ile de France est depuis opérationnelle.

2) **BQR "Diagnostic et modélisation des leucémies"** (période 2005-2006) basé sur la collaboration entre des mathématiciens de l'UMR 5208 (UCBL-ECL-INSA) et des médecins du service d'Hématologie Clinique (UCBL). A la suite de ce BQR l'article **no. 11** a été publié.

3) **GDR "Mascot NUM"** (Méthodes d'Analyse Stochastique pour les Codes et Traitements numériques). L'objectif principal de ce groupe de recherche (national) est de coordonner les efforts de recherche mathématique dans la modélisation et dans l'informatique. Ce GDR est mise en place *depuis le mois de mai 2008*.

4) **Partenariat Univ. Lyon 1/ Volvo Truck**, *depuis avril 2010*.

1.4 Encadrement

- **Encadrement doctoral.** J'ai codirigé, pendant la période janvier 2003-avril 2006, avec Hélène Tattegrain-Veste(INRETS) la thèse en Statistique de Nicolas DAPZOL "Analyse de l'activité de conduite par les chaînes de Markov et les modèles de rupture multi-phase. Méthodologie et applications".

La thèse de N. Dapzol a reçu le prix : "Young Researcher Seminar of the European Conference of Transport Research Institute". Comme le nom l'indique, ce prix a été décerné à la Conférence : "European Conference of Transport Research Institute", La Haye, Pays-Bas, du 11 au 13 mai 2005.

Après sa thèse Nicolas Dapzol a été recruté en tant que Chargé de Recherche à INRETS. A la suite de ce co-encadrement, l'article **Ciuperca et Dapzol(2008)** a été publié.

- **Encadrement Mémoire Master Recherche(DEA).**

1) Pendant l'année universitaire 2003/2004, j'ai encadré le mémoire "Vitesse de conver-

gence dans le Théorème Central Limite pour des suites de variables aléatoires faiblement dépendantes”, de Delphine RIOLI élève de l’ENS-Lyon, DEA de Mathématiques.

2) Pendant l’année universitaire 2009/2010, j’ai encadré le mémoire Détection d’un changement dans un modèle à longue mémoire. Application À l’étude de l’inflation”, de Rima HADDAD étudiante à l’ISFA LYON, M2R.

- **Encadrement Master Professionnel.**

- J’assure l’encadrement de nombreux stages industriels des étudiants de la deuxième année du Master Professionnalisant : “Statistique, informatique et techniques numériques” ;
- J’encadre des projet TER de la première année du Master Professionnalisant MAIM.

1.5 Autres activités scientifiques

Expertises et conseils scientifiques :

- *Referee* pour des revues internationales de statistique :
 - Communications in Statistics - Theory and Methods,
 - Communications in Statistics - Simulation and Computation
 - Scandinavian Journal of Statistics,
 - Journal of Statistical Planning and Inference,
 - Journal of Multivariate Analysis
 - Journal of Nonparametric Statistics
 - Mathematical Review.
- Entre septembre 2004 et septembre 2008 j’ai été membre de la Commission des Spécialistes, section 25/26, Université Lyon 1. Depuis décembre 2008 je fais partie du Comité Consultatif de l’Université Lyon 1, section 25/26.

- *septembre 2010*, membre du jury, en qualité d’expert, pour un concours externe d’Ingénieur de recherche CNRS.

- *27 octobre 2010*, membre jury de thèse de doctorat ”Modélisation de l’incertitude sur les trajectoires d’avions”, soutenue par Nobert FUEMKEU à INRETS/Lyon 1.

- *novembre 2010*, expertise scientifique d’un dossier de bourse CIFRE.
-

2 Présentation des thèmes de recherche

2.1 Mélanges de densités

Soit Y une variable aléatoire définie sur l'espace de probabilité $(\Omega, \mathcal{B}, \mathbb{P}_\theta)$, avec le paramètre $\theta \in \Theta$.

La théorie classique d'inférence statistique pour des modèles paramétriques est construite sur l'hypothèse d'identifiabilité : si $\theta_1 \neq \theta_2$ alors $\mathbb{P}_{\theta_1} \neq \mathbb{P}_{\theta_2}$. Un exemple de lois non-identifiables est le mélange de densités. Soit $\mathcal{F} = \{f_\gamma(x); \gamma \in \Gamma\}$ une famille de densités. Pour un p connu, l'ensemble des p -mélanges de densités de \mathcal{F} est :

$$\mathcal{G}_p := \left\{ g_{\pi, \alpha} = \sum_{i=1}^p \pi_i f_{\gamma_i}(\cdot) \quad / \quad \pi = (\pi_1, \dots, \pi_p), \alpha = (\gamma_1, \dots, \gamma_p), \right. \\ \left. \forall i = 1, \dots, p, \gamma_i \in \Gamma, 0 \leq \pi_i \leq 1, \sum_{i=1}^p \pi_i = 1 \right\}$$

Un problème important pour ce type de modèle, est de déterminer le nombre p de composantes ; en fait, de tester si le modèle est un mélange de p -densités ou un mélange de q -densités, avec $q < p$. Dans le cas des modèles identifiables, le test du rapport de vraisemblance est une des solutions les plus utilisées. Pour le cas de mélanges de lois, la loi asymptotique du rapport de vraisemblance a été longtemps un problème ouvert, puisque sous l'hypothèse nulle, le modèle n'est pas identifiable et donc le théorème classique de χ^2 est inapplicable.

Si on teste une densité contre un mélange de deux densités, un changement de variables est d'abord nécessaire pour trouver la loi asymptotique du rapport de vraisemblance. Pour cette reparamétrisation, l'ensemble des dérivées de la log-vraisemblance, par rapport à un de ces paramètres, doit être une classe de Donsker. On note cet ensemble par \mathcal{D} . Dans l'article **Ciuperca(1999)** j'ai donné des conditions suffisantes, relativement faibles pour les densités f_γ , pour que \mathcal{D} soit une classe de Donsker.

J'ai construit le premier exemple de non-consistance du test de vraisemblance pour un ensemble compact de paramètres (**Ciuperca(2002)**). Il s'agit de la famille des densités exponentielles : $f_\gamma(x) = e^{-(x-\gamma)} \mathbb{1}_{x>\gamma}$ pour $\gamma \in \Gamma = [0, G]$. Pour tester une densité contre un mélange simple, je démontre d'abord que l'ensemble \mathcal{D} des fonctions scores n'est pas relativement compact dans $L^2(f_0)$, donc il n'est pas Donsker.

Ensuite je montre que la statistique du maximum de vraisemblance vaut 0 avec une probabilité 1/2 et elle converge fortement vers $+\infty$ avec une probabilité 1/2. Donc, pour les densités exponentielles, on ne peut pas utiliser la méthode du maximum de vraisemblance pour tester le nombre de composantes. Une étude numérique est réalisée pour conforter le résultat théorique obtenu.

Dans l'article **Ciuperca et al.(2003)**, le nombre de composantes p est supposé connu, par contre l'espace des paramètres n'est plus compact. Plus précisément, on considère un mélange de p densités normales univariées. Si la variance d'une des com-

posantes est proche de zéro, la densité $g_{\pi,\alpha}$ peut dégénérer et donc l'estimateur du maximum de vraisemblance (MV) peut ne pas exister. Dans ce cas, l'algorithme EM dégénère. Pour palier à cet inconvénient, on a proposé un estimateur du MV pénalisé, qui est borné pour toute valeur de la variance, pour une taille d'échantillon fixée. Évidemment, l'algorithme EM pénalisé correspondant ne dégénère plus. Nous avons démontré que cet estimateur est fortement convergent, asymptotiquement efficace. Sa loi asymptotique est gaussienne centrée de matrice de covariance, l'inverse de l'information de Fisher. Les résultats obtenus sont généralisés pour un mélange de toute densité qui dégénère sur \mathbb{R} . Des simulations numériques mettent en évidence les avantages et les performances de l'estimateur proposé.

2.2 Modèle avec rupture, variables indépendantes

Un autre modèle non-identifiable, plus complexe, est le modèle avec rupture :

$$Y = h_1(x)\mathbb{1}_{t \leq \tau_1} + h_2(x)\mathbb{1}_{\tau_1 < t \leq \tau_2} + \dots + h_{K+1}(x)\mathbb{1}_{\tau_K < t} + \varepsilon$$

pour $x \in \mathbb{R}^p$, $t \in \mathbb{R}$. Les paramètres τ_1, \dots, τ_K s'appellent *points de rupture*. Ces paramètres sont estimés à partir de n observations de $(Y_i, X_i, t_i)_{1 \leq i \leq n}$. Des techniques paramétriques, non-paramétriques ou sémi-paramétriques peuvent être utilisées pour estimer ces points de rupture mais aussi pour estimer les fonctions h . Nous avons considéré seulement des techniques paramétriques. Dans ce cas, les fonctions h ont la forme $h_k(x) \equiv h_{\beta_k}(x)$, $\beta_k \in \Gamma$, avec Γ compact et $x \in \mathbb{R}^p$.

Les problèmes d'estimation dans un modèle avec points de rupture sont :

- déterminer d'abord le nombre K de changements. Évidemment, si à la suite de cette étape on obtient $K = 0$, alors le modèle n'a pas de points de rupture et on doit considérer un modèle de régression "classique".
- une fois le nombre K fixé, trouver la localisation des points de rupture τ_1, \dots, τ_K ;
- estimer les fonctions h_1, \dots, h_{K+1} du modèle, dans deux cas : quand elles sont totalement inconnues ou partiellement inconnues.

Une difficulté majeure dans l'estimation de la localisation des points de rupture est la non-régularité de la fonction à optimiser aux points $(\tau_1, \tau_2, \dots, \tau_K)$ considérés comme paramètres. Notons aussi que l'inférence statistique est influencée par la continuité ou la discontinuité du modèle aux points de rupture, mais aussi par le caractère déterministe ou aléatoire de la variable explicative X et du temps de mesure t . Dans le cas où les temps de mesure $(t_i)_{1 \leq i \leq n}$ sont aléatoires, l'estimation des points τ_1, \dots, τ_K est plus difficile dans le cas $K > 1$ puisqu'il y a des segments avec des bornes complètement inconnues.

La plupart des modèles de rupture traités auparavant par d'autres auteurs n'ont qu'un seul point de rupture ($K = 1$) et surtout avec des fonctions linéaires $h_{\beta}(x) = \beta^t x$, ce qui

facilite beaucoup les démonstrations.

D’abord, nous mettons en oeuvre la méthode d’estimation du maximum de vraisemblance (**Ciuperca(2004)**, **Ciuperca et Dapzol(2008)**) pour le modèle :

$$Y_i = h_{\beta_1}(X_i)\mathbb{1}_{0 < X_i \leq \tau_1} + h_{\beta_2}(X_i)\mathbb{1}_{\tau_1 < X_i \leq \tau_2} + \dots + h_{\beta_{K+1}}(X_i)\mathbb{1}_{\tau_K < X_i < 1} + \varepsilon_i$$

avec X_i variables aléatoires de densité absolument continue positive et ε_i centrés i.i.d., de variance finie. Pour ce modèle, on impose la condition d’un saut non-nul dans chaque point de rupture. Les estimateurs du MV des paramètres sont fortement consistants. La vitesse de convergence est d’ordre n^{-1} pour les estimateurs des points de rupture et d’ordre $n^{-1/2}$ pour les estimateurs des paramètres de régression $(\beta_1, \dots, \beta_{K+1})$. Ils convergent en loi vers le *argmax* de K processus de Poisson composés indépendants pour les premiers et vers un processus gaussien centré pour les seconds.

Ces résultats ont été généralisés dans l’article **Ciuperca(2009)**, en utilisant la méthode de M-estimation. Pour $\theta_1 = (\beta_1, \dots, \beta_{K+1})$, $\theta_2 = (\tau_1, \dots, \tau_K)$, $\theta = (\theta_1, \theta_2)$, $\rho : \mathbb{R} \rightarrow \mathbb{R}_+$ une fonction convexe, le *M-estimateur* est :

$$\arg \min_{\theta} \sum_{i=1}^n \rho(Y_i - h_{\beta_1}(X_i)\mathbb{1}_{0 < X_i \leq \tau_1} - h_{\beta_2}(X_i)\mathbb{1}_{\tau_1 < X_i \leq \tau_2} - \dots - h_{\beta_{K+1}}(X_i)\mathbb{1}_{\tau_K < X_i < 1})$$

Si $\rho(x) = x^2$, on obtient l’estimateur des moindres carrés et pour $\rho(x) = |x|$ l’estimateur des moindres déviations. L’estimateur du MV est obtenu en considérant $\rho(x) = -\log \varphi(x)$ avec φ la densité de l’erreur ε .

On montre que tous les résultats obtenus pour l’estimateur du MV restent valables pour le M-estimateur. La loi asymptotique du M-estimateur des points de rupture est le *argmin* de K processus de Poisson composés indépendants et la distribution du saut est une fonction de ρ .

Si les erreurs d’un modèle contiennent des outliers, alors l’estimateur des moindres carrés a une grande variance. Les outliers peuvent créer également des problèmes dans la détection des points de rupture, un outlier pouvant être ”confondu” avec un point de rupture dans le modèle. Dans ce cas, il est préférable de considérer la méthode d’estimation LAD (“Least Absolute Deviations”). Par contre l’étude théorique de ses propriétés et le calcul numérique associé sont plus difficiles à cause de la nondifférentiabilité de la fonction objectif.

Les modèles de rupture considérés dans les papiers **Ciuperca(2011a)**, **Ciuperca(2011b)** ont la forme :

$$Y_i = h_{\beta_1}(X_i)\mathbb{1}_{i \leq \tau_1} + h_{\beta_2}(X_i)\mathbb{1}_{\tau_1 < i \leq \tau_2} + \dots + h_{\beta_{K+1}}(X_i)\mathbb{1}_{\tau_K < i} + \varepsilon_i$$

$i = 1, \dots, n$. La variable X_i est supposée déterministe.

Par contre, la contrainte de discontinuité du modèle dans les points de rupture n’est plus

imposée. D'autre part, pour les erreurs ε_i on n'impose plus les conditions d'existence pour les deux premiers moments, mais $\mathbb{E}[\text{sign}(\varepsilon)] = 0$.

L'estimateur LAD du paramètre θ est la variable aléatoire qui minimise la somme des déviations absolues des erreurs :

$$\hat{\theta}_n = (\hat{\theta}_{1n}, \hat{\theta}_{2n}) = \arg \min_{\theta} \sum_{i=1}^n |Y_i - h_{\beta_1}(X_i)\mathbb{1}_{i \leq \tau_1} - h_{\beta_2}(X_i)\mathbb{1}_{\tau_1 < i \leq \tau_2} - \dots - h_{\beta_{K+1}}(X_i)\mathbb{1}_{\tau_K < i}|$$

Comme pour les estimateurs précédents, on montre la convergence forte et on étudie leur vitesse de convergence. La vitesse de convergence de l'estimateur LAD des paramètres de régression est plus lente que celle du M-estimateur. Ceci s'explique par le fait qu'on n'exige plus la discontinuité aux points de rupture. En ce qui concerne les estimateurs LAD des points de rupture, ils ont la même vitesse de convergence, d'ordre n^{-1} , que celle obtenue par la M-estimation. La loi asymptotique de $\hat{\theta}_{1n}$ est gaussienne, comme pour le M-estimateur, tandis que celle de $\hat{\theta}_{2n}$ est le *argmin* d'un processus aléatoire de forme analytique connue.

Pour estimer le nombre K de points de rupture, on propose un critère de type Schwarz qui fournit un estimateur faiblement convergent.

L'inconvénient de cette méthode d'estimation est que la fonction objectif à minimiser n'est pas dérivable ce qui peut poser des problèmes numériques. Pour y remédier, on approxime la fonction objectif par une fonction différentiable et on peut alors utiliser des algorithmes classiques d'optimisation numérique.

Pour une suite positive (d_n) convergeant vers zéro pour $n \rightarrow \infty$, on considère les estimateurs SLAD ("smoothed least absolute deviations") :

$$\hat{\theta}_n^s = (\hat{\theta}_{1n}^s, \hat{\theta}_{2n}^s) = \arg \min_{\theta} \sum_{i=1}^n \left\{ |Y_i - h_{\beta_1}(X_i)\mathbb{1}_{i \leq \tau_1} - \dots - h_{\beta_{K+1}}(X_i)\mathbb{1}_{\tau_K < i}|^2 + d_n^2 \right\}^{1/2}$$

La vitesse de convergence de ces estimateurs dépend de la suite de pénalisation d_n et elle peut être plus lente que la vitesse des estimateurs LAD.

On donne un critère similaire au cas non-pénalisé pour trouver le nombre de points de rupture.

Des simulations numériques confirment les performances des estimateurs LAD et SLAD si le modèle contient des outliers. En plus, l'algorithme associé à la méthode pénalisée converge plus vite que l'algorithme associé à la méthode LAD. Enfin, si le modèle ne présente pas d'outliers, les méthodes LAD, SLAD et des moindres carrés fournissent des estimateurs avec les mêmes performances.

2.3 Estimation des taux d'entropie

L'entropie renseigne sur le degré d'incertitude du phénomène étudié. Si X est une variable aléatoire discrète à valeurs dans $(x_j)_{j \in I}$, alors l'entropie de Shannon de X est :

$$\mathbb{H}_1(X) = - \sum_{j \in I} \mathbb{P}[X = x_j] \log \mathbb{P}[X = x_j]$$

Soit $\mathbf{X} = (X_i)_{1 \leq i \leq n}$ une suite aléatoire. Si X_i prend des valeurs dans un ensemble discret $E \subset \mathbb{R}$, l'entropie au temps n de $(X_i)_{1 \leq i \leq n}$ est :

$$\mathbb{H}_n(\mathbf{X}) = - \sum_{(x_1, \dots, x_n) \in E^n} \mathbb{P}[X_1 = x_1, \dots, X_n = x_n] \log \mathbb{P}[X_1 = x_1, \dots, X_n = x_n]$$

Si la limite suivante $\lim_{n \rightarrow \infty} n^{-1} \mathbb{H}_n(\mathbf{X})$ existe, elle s'appelle *taux d'entropie (de Shannon)* et on la note $\mathbb{H}(\mathbf{X})$.

Il y a très peu de résultats dans la littérature concernant l'estimation de $\mathbb{H}(\mathbf{X})$ dans le cas d'une suite non i.i.d. Dans l'article **Ciuperca et Girardin(2007)**, nous avons considéré l'estimation de $\mathbb{H}(\mathbf{X})$ pour des chaînes de Markov homogènes ergodiques, pas nécessairement stationnaires, avec espace d'états dénombrable.

Soit une chaîne \mathbf{X} , d'espace d'états E de probabilités de transition $P = (P(i, j))_{i, j \in E}$ et de loi stationnaire $\pi = (\pi(i))_{i \in E}$. Le taux d'entropie de Shannon de la chaîne de Markov, s'il existe, est :

$$\mathbb{H}(\mathbf{X}) = - \sum_{i \in E} \pi(i) \sum_{j \in E} P(i, j) \log P(i, j)$$

Pour estimer $\mathbb{H}(\mathbf{X})$, nous proposons des méthodes non-paramétriques et paramétriques dans deux situations : une chaîne de Markov longue, plusieurs chaînes courtes indépendantes. Dans le cas non-paramétrique, les estimateurs du taux d'entropie sont obtenus à partir des estimateurs plug-in pour les probabilités de transition et pour la loi stationnaire de la chaîne. On montre que l'estimateur du taux d'entropie est fortement consistant pour un espace d'états dénombrable. Si l'espace des états est fini et la loi de transition n'est pas uniforme, on montre la normalité asymptotique. Dans le cas de plusieurs trajectoires, différents schémas sont proposés. L'estimateur du taux d'entropie construit à partir de l'observation de seulement deux pas de la chaîne de Markov, a les propriétés les plus intéressantes, parce qu'il utilise les résultats connus de l'estimation de l'entropie pour des suites i.i.d.

Dans le cas paramétrique, les estimateurs de l'entropie sont obtenus à partir de l'estimateur plug-in pour les paramètres. On montre leur convergence forte et la normalité asymptotique.

Ces résultats ont été généralisés dans **l'article 2** où l'on étudie le taux d'entropie des suites de v.a. dépendantes correspondant à des fonctionnelles d'entropie. On couvre ainsi les entropies : Shannon, Rényi, Tsallis, Sharma-Mittal, On montre que tous les taux d'entropie sont soit 0 soit infini ; exceptés les taux d'entropie de Shannon et de Rényi. Les entropies et les taux d'entropie marginales pour une chaîne de Markov paramétrique sont estimés en utilisant l'estimateur du MV pour le paramètre et ensuite la méthode de plug-in.

2.4 Modèles à longue mémoire

Une des suppositions standard dans un modèle avec ou sans points de rupture est l'indépendance des erreurs ou des régresseurs, ce qui constitue assez souvent une mauvaise approximation pour des phénomènes réels. Les processus à longue mémoire sont caractérisés par une décroissance lente des autocorrélations, ce qui change la méthodologie utilisée pour étudier le comportement asymptotique des estimateurs et des tests d'hypothèse. Une série temporelle à longue mémoire est le processus ARFIMA.

Les domaines d'application de ces processus sont très variés : climatologie, hydrologie, géophysique, chimie, économétrie, télécommunications, etc.

Le seul article qui aborde un modèle non-linéaire avec des régresseurs et des erreurs à longue mémoire est celui de Koul et Baillie (2003), mais dans lequel la vitesse des estimateurs est supposée connue. Dans l'article **Ciuperca(2011)** j'ai trouvé la vitesse de convergence et la loi asymptotique de l'estimateur des moindres carrés des paramètres pour un modèle non-linéaire. Les résultats obtenus dépendent des paramètres de longue mémoire et de l'espérance de la dérivée de la fonction de régression par rapport aux paramètres de régression. La distribution asymptotique est ensuite utilisée pour tester si à gauche et à droite d'un point de rupture on a le même modèle. Des simulations confirment les résultats théoriques.

Pour un modèle linéaire avec un seul point de rupture, les erreurs et les régresseurs Gaussiens (ou fonction de variables Gaussiennes) à longue mémoire, comme on l'a déjà vu, les outliers peuvent poser des problèmes. Dans ce cas, on peut utiliser la S-estimation : la convergence forte et la vitesse de convergence du S-estimateur sont obtenues (voir **article 1**). Ces résultats sont totalement différents des résultats obtenus précédemment pour les modèles à courte mémoire ou indépendants. Des simulations et une application au débit du fleuve Nil sont présentées. Une comparaison avec deux autres estimateurs proposés dans la littérature montre la supériorité du S-estimateur proposé.

2.5 Estimation semi- ou non-paramétrique

2.5.1 Maximum de vraisemblance(MV) empirique

Pour un modèle non-linéaire $Y_i = f(X_i; \beta) + \varepsilon_i$, $i = 1, \dots, n$ avec (ε_i, X_i) une suite de v.a. indépendantes, la variable réponse Y_i peut être manquante de manière aléatoire. Soit la v.a. $\delta_i = 1$ si Y_i observée et $\delta_i = 0$ si Y_i manquante. Soit la fonction de probabilité de sélection $\pi(x) = \mathbb{P}[\delta = 1 | X = x]$. Les paramètres β peuvent être estimés sur les données complètes, soit par moindres carrés $\hat{\beta}_{n,LS} = \arg \min_{\beta} \sum_{i=1}^n \delta_i [Y_i - f(X_i, \beta)]^2$ soit par moindres déviations $\hat{\beta}_{n,LAD} = \arg \min_{\beta} \sum_{i=1}^n \delta_i |Y_i - f(X_i, \beta)|$. Ces estimateurs permettent la construction de la statistique du MV empirique. On montre que cette statistique converge vers une loi de χ^2 , ce qui implique la construction des régions de confiance pour β . Les données manquantes sont reconstituées en utilisant une méthode semi-paramétrique et le théorème de Wilks est prouvé sur toutes les données reconstituées. Des simulations montrent, via la méthode Monte Carlo, que l'estimateur pro-

posé est meilleur, en terme de couvrement, que la méthode Normale. Les avantages de la méthode proposée sont exemplifiés aussi sur des données réelles (**article 21.**).

2.5.2 Estimation semi-paramétrique pour des lois des valeurs extrêmes

Dans le papier **Ciuperca et Mercadier(2010)** on propose un estimateur semi-paramétrique, qui généralise nombreux estimateurs existants, pour l'indice des valeurs extrêmes γ et pour le paramètre d'ordre 2 d'une loi des valeurs extrêmes. Pour une fonction poids g et une constante $\alpha > 0$, la statistique proposée a la forme :

$$\Gamma_{n,k}(g, \alpha) = \frac{\frac{1}{k} \sum_{i=1}^k g\left(\frac{i}{k+1}\right) \left[\log \frac{X_{n-i+1,n}}{X_{n-k,n}}\right]^\alpha}{\int_0^1 g(x)(-\log(x))^\alpha dx}$$

avec $X_{1,n} \leq \dots, \leq X_{n,n}$ sont les v.a. X_1, \dots, X_n ordonnées. On montre que $\Gamma_{n,k}(g, \alpha)$ est un estimateur consistant et asymptotiquement normal pour γ^α . Un estimateur (consistant et asymptotiquement normal) pour le paramètre de seconde ordre de la loi est proposé. Des simulations ont été réalisées.

2.6 Applications des statistiques

Après ma thèse, j'ai fait un post-doc d'un an sur la modélisation des pics de pollution dans la région parisienne dans le cadre d'un contrat de recherche entre le Laboratoire de Probabilités/Statistique de l'Université d'Orsay et AIRPARIF (organisme agréé par le ministère chargé de l'Environnement pour la surveillance de la qualité de l'air en Ile de France). Dans le cadre de ce contrat, j'ai travaillé sur la mise en place effective du système de prévision des pointes de pollution par l'ozone et les oxydes d'azote dans la région parisienne, à court et à moyen terme. A court terme, la concentration journalière maximale d'ozone (de l'après-midi) est prévue à 6h du matin, en utilisant des données de pollution et données météo mesurées dans la région parisienne. A moyen terme, la prévision est faite un, voire plusieurs jours à l'avance. Pour prévoir la pollution, on a utilisé plusieurs méthodes statistiques : analyse de données, méthode de régression par arbre CART et des méthodes non-paramétriques. La méthode retenue est la non-paramétrique, les pics de pollution étant prévus en proportion de 90%. On considère (Y_n) le processus à prévoir et (X_n) un processus de variables exogènes. Dans ce cas précis, Y_n est la concentration maximale journalière d'ozone pour le jour n et X_n contient des variables météo (température, direction et vitesse de vent, ...). Le modèle statistique considéré est :

$$\begin{aligned} Y_{n+1} &= F(Y_n, X_n) + \varepsilon_{1n} \\ X_{n+1} &= G(X_n) + \varepsilon_{2n} \end{aligned}$$

avec (ε_{1n}) et (ε_{2n}) des bruits blancs indépendants. Les fonctions F et G sont inconnues et elles sont estimées en utilisant une méthode à noyau, en prenant le noyau Nadaraya-Watson.

Le système de prévision a été implanté à partir de l'été 1997 à AIRPARIF et depuis il est opérationnel. Cette prévision de la pollution sert pour prévenir la population dans le cas d'un grand pic de pollution. Elle sert aussi aux décideurs politiques pour envisager des mesures qui vont défavoriser la formation de la pollution. (Articles **Bel et al.(1998)**, **Bel et al.(1999)**).

J'ai participé à un BQR basé sur la collaboration des mathématiciens de l'UMR 5208 et des médecins du Service d'Hématologie Clinique (Univ. Lyon 1). Ce projet a fait partie d'un programme prioritaire régional "Cancer" et il a été consacré au développement des outils de diagnostic des leucémies aiguës myéloïdes. Des méthodes statistiques ont permis de prévoir certains types de leucémies en connaissant l'immunophénotypage. (Article **Plesa et al.(2006)**).

Récemment j'ai collaboré avec un group de chercheurs/ médecins sur une étude de la schizophrénie (Article **Chambon et al.(2008)**).

2.7 Perspectives

1. Concernant les modèles non-linéaires de rupture, avec des variables aléatoires à longue mémoire, l'estimation des paramètres peut être envisagée.
 2. Pour les modèles paramétriques de rupture, trouver le nombre de points de rupture a été traité seulement de point de vue critère de choix. Le test d'hypothèse peut être une autre solution plus complète mais aussi plus complexe.
-

3 Activités d'enseignement

Les principales activités d'enseignement depuis mon arrivée à l'Université Lyon 1 :

1. **CM** : *DEA Mathématiques Université Lyon 1*, "Statistique Exploratoire" (2003/2004) :
 - (a) introduction aux Statistiques ;
 - (b) méthodes factorielles : Analyse en Composantes Principales(ACP), Analyse Factorielle Discriminante (AFD), Analyse Factorielle des Correspondances (AFC), Analyse des Corrélations Canoniques (ACC) ;
 - (c) méthodes de classification : classification par partition, classification hiérarchique.
2. **CM/TD/TP** : *Master Pro 1 et 2 de Mathématiques Appliquées*,
 - (a) "Statistique paramétrique"(M1)(2009/2011), "Statistique inférentielle"(M2)(2001-2010), "Techniques Probabilités et Statistiques"(2001/2009) :
 - i. statistique descriptive ;
 - ii. notions d'échantillonnage, Théorème de Cochran ;
 - iii. théorie de l'estimation : méthodes d'estimation, propriétés des estimateurs, familles exponentielles, estimateur par intervalle ;
 - iv. théorie des tests d'hypothèse : *tests paramétriques* : lemme de Neyman-Pearson, test du rapport de vraisemblance et de Wald, tests de Student et de Fisher pour une ou deux populations ; *test non-paramétriques* : Théorème de Pearson, test de χ^2 , test de Kolmogorov-Smirnov, de la médiane, test de Spearman, test de Wilcoxon.
 - v. modèles linéaires
 - A. régression linéaire multiple ;
 - B. analyse de variance ;
 - vi. En TD/TP, nombreuses applications concrètes sont traitées, en utilisant les logiciels SAS et R.
 - (b) "Analyse de données"(M2), (2001-2011) : méthodes factorielles : ACP, AFD, AFC, ACC. L'examen consiste d'un projet (individuel) sur des données réelles.
 - (c) "Outils statistiques avancés"(M2)(2006-2011) : le contenu peut être différent d'une année à l'autre :
 - i. séries temporelles ;
 - ii. modèles non-linéaires ;
 - iii. régression logistique ;
 - iv. modèles mixtes ;
 - v. modèles de censure ;
 - vi. méthodes de Jackknife et Bootstrap ;
 - vii. data mining : régression PLS, méthode CART

- (d) “Méthodes des Monte-Carlo”(M2)(2001-2005)
 - i. intégration par Monte-Carlo, Théorème de Rubinstein ;
 - ii. méthode de Bootstrap ;
 - iii. algorithme de Hastings-Metropolis.
- 3. Chaque année je propose (et j’encadre) 3-4 sujets de **TER** pour les étudiants en M1 PRO de Mathématiques Appliquées. J’encadre également des **stages en entreprise** des étudiants en M2 PRO de la même formation.
- 4. **CM/TD/TP** : *L1 et L2 MASS*,
 - (a) “Probabilités/Statistique”(L2)(2001-2007)
 - i. introduction à la théorie de l’estimation et des tests d’hypothèse ;
 - ii. modèles linéaires : régression et analyse de variance.
 - (b) “Statistique descriptive”(L1)(2002/2003) ;
- 5. **CM/TD/TP** : *Master Pro M2 Neurosciences*, “Statistique appliquée aux neurosciences” (2007/2008) :
 - (a) notions de probabilités ;
 - (b) statistique descriptive ;
 - (c) estimation et tests d’hypothèse ;
 - (d) ACP, AFC.
- 6. **TD** : *Ecole Ingénieurs Universitaire(ISTIL)*, “Introduction aux méthodes probabilistes et statistiques”(première année) (2008/2009) :
 - (a) modèle probabiliste, variables aléatoires discrètes, la loi Normale ;
 - (b) échantillons et lois d’échantillonnage ;
 - (c) estimation et test de la moyenne, tests de χ^2 .

Logiciels utilisés en TP : SAS, R et Matlab.

Fait à Villeurbanne, le 24 octobre 2011