Programme et activités de recherche

1 Travaux de thèse et développements récents

Nos travaux de thèse prennent place dans le vaste domaine de la Statistique Fonctionnelle. En premier lieu, nos efforts se sont concentrés sur l'étude du comportement asymptotique presque sûr des estimateurs à noyaux de la fonction de régression et ses dérivées. Notre but principal est d'établir des lois limites presque sûres, appelées **lois uniforme du logarithme** ([LUL]), concernant la déviation en norme uniforme de la régression et ses dérivées dans les cadres uni et multivariés. Les estimateurs concernés sont ceux de type Nadaraya-Watson, par lissage polynomial local (cf. Fan et Gijbels [13]) et par la méthode des ondelettes. Nous proposons également des extensions pour des M-estimateurs de la fonction de régression (méthode du maximum de vraisemblance [4] et régression robuste [6]). La méthode de démonstration est fondée sur la théorie moderne des processus empiriques.

Soient (X,Y), (X_1,Y_1), (X_2,Y_2),..., des couples de variables aléatoires à valeurs réelles indépendants et identiquement distribués. Le couple de variables aléatoires (X,Y) est supposé admettre une densité jointe $f_{X,Y}$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^2$ et une densité marginale f_X. Nous introduisons une fonction auxiliaire Borel mesurable $\phi(\cdot)$, supposée bornée sur chaque sous-ensemble compact de $\mathbb{R}$. Nous nous intéressons à l'estimation de la fonction de régression (ou espérance conditionnelle) de $\phi(Y)$ sachant $X=x$, définie par

$$m_\phi(x) = \mathbb{E}\big[\phi(Y)|X=x\big] = \frac{1}{f_X(x)}\int_{\mathbb{R}} \phi(y) f_{X,Y}(x,y)dy := \frac{r_\phi(x)}{f_X(x)}, \quad \text{lorsque } f_X(x)\neq 0.$$

Nos résultats seront établis uniformément en $x\in I$, avec I intervalle compact de $\mathbb{R}$ strictement contenu dans le support de f_X. Soit $k\geq 0$ un entier arbitraire, désignant le degré de dérivation. Dans un premier temps, nous considérons des estimateurs à noyau du type Nadaraya-Watson de $m_\phi^{(k)}(x)$, la dérivée d'ordre k de $m_\phi(x)$. L'estimateur de Nadaraya-Watson de la fonction de régression $m_\phi(x)$ est défini par,

$$\hat{m}_{\phi;n}(x) := \frac{(nh_n)^{-1}\sum_{i=1}^n \phi(Y_i)K\big((x-X_i)/h_n\big)}{(nh_n)^{-1}\sum_{i=1}^n K\big((x-X_i)/h_n\big)} =: \frac{\hat{r}_{\phi;n}(x)}{\hat{f}_{X;n}(x)}, \quad \text{lorsque } \hat{f}_{X;n}(x)\neq 0.$$

Les estimateurs des différentes dérivées de la régression sont obtenus en utilisant la formule de Leibniz concernant les dérivées de produit de fonctions. Par exemple, nous avons, lorsque $k=1$,

$$\hat{m}'_{\phi;n}(x) = \frac{\hat{r}'_{\phi;n}(x)}{\hat{f}_{X;n}(x)} - \frac{\hat{r}_{\phi;n}(x)\hat{f}'_{X;n}(x)}{\hat{f}^2_{X;n}(x)}, \quad \text{lorsque } \hat{f}_{X;n}(x)\neq 0.$$

En s'inspirant des travaux récents de Einmahl et Mason [11] et Deheuvels et Mason [10], sous certaines hypothèses classiques, nous avons obtenu le résultat suivant : lorsque $n\to\infty$,

$$\left|\left\{\frac{nh_n^{2k+1}}{2\log(1/h_n)}\right\}^{1/2} \sup_{x\in I} \pm\big\{\hat{m}^{(k)}_{\phi;n}(x) - m^{(k)}_{\phi;n}(x)\big\} - \sigma_\phi(I)\right| = o(1), \quad \text{presque sûrement,} \tag{1}$$

où

$$\sigma^2_\phi(I) = \sigma^2_{\phi,k}(I) := \sup_{x\in I}\left\{\frac{\mathrm{Var}[\phi(Y)|X=x]}{f_X(x)}\right\}\int_{\mathbb{R}}[K^{(k)}(t)]^2dt.$$

Ci-dessus, notons que $m^{(k)}_{\phi;n}(x)$ correspond à une approximation de l'espérance usuelle $\mathbb{E}[\hat{m}^{(k)}_{\phi;n}(x)]$. Cette approximation permet de linéariser la déviation stochastique et son expression comme une fonctionnelle du processus empirique, i.e.

$$\hat{m}^{(k)}_{\phi;n}(x) - m^{(k)}_{\phi;n}(x) \overset{p.s.}{=} \alpha_n(\nu_{x,k,n}) + o(1), \quad x \in I,$$

où $\alpha_n(\nu_{x,k,n})$ désigne le processus empirique bivarié basé sur les couples d'observations $(X_i, Y_i)_{i=1}^n$ et indexé par $\nu_{x,k,n} : \mathbb{R}^2 \to \mathbb{R}$ fonction choisie convenablement. Ainsi, l'étude de la convergence uniforme (sur un compact I) presque sûre de nos estimateurs se réduit à l'analyse du comportement asymptotique de la norme du supremum du processus empirique indicé par une certaine classe de fonctions $\nu_{x,k,n}$, elle-même indexée par les points de l'intervalle I. Précisons que la loi limite uniforme du logarithme (1) constitue un analogue à la loi du logarithme itéré classique pour le mode de convergence uniforme.
La méthodologie de démonstration repose sur la théorie moderne des processus empiriques et, plus particulièrement, sur l'étude du processus empirique local indicé par des classes de fonctions vérifiant certaines propriétés combinatoriales. De manière classique, la démonstration de notre loi limite (1) se scinde en deux parties, le traitement de la borne supérieure puis celui de la borne inférieure. Essentiellement, la démonstration de la borne supérieure s'appuie sur le principe du chaînage, qui consiste à ramener un supremum sur une classe infinie à un maximum, via une discrétisation préalable combinée à un contrôle des incréments. La première étape de la démonstration de la borne supérieure utilise une version maximale de l'inégalité de Bernstein afin d'estimer le supremum de la déviation sur une version discrétisée de l'intervalle I. Puis, le contrôle de l'oscillation maximale entre les points de la discrétisation fait appel à une remarquable inégalité exponentielle (démontrée par Talagrand, cf. [18]) combinée à une inégalité de moment appropriée. Cette inégalité exponentielle (de type Bernstein ou Borell), dite de concentration, est centrale dans nos travaux. La borne de moment, développée par Einmahl et Mason ([11] et [12]), permet d'appliquer efficacement l'inégalité de Talagrand et nécessite une condition d'entropie sur la classe de fonctions considérée. Cette condition d'entropie permet de contrôler la taille de la classe de fonctions. Typiquement, nous devons vérifier que le nombre de recouvrement de la classe de fonctions est uniformément polynomial, ceci étant satisfait pour certaines classes de fonctions particulières dénommées "*VC subgraph classes*" (cf. Van der Vaart et Wellner [19]) ou classes de graphes VC. Comme nous cherchons à borner des suprema sur des familles a priori non-dénombrables, une hypothèse spécifique de mesurabilité est également nécessaire pour formuler nos théorèmes sans mesures extérieures. Il est alors suffisant de vérifier que les classes de fonctions sont "mesurables ponctuellement" ("*pointwise measurable*"), i.e. pour chaque classe de fonctions $\mathcal{F}$, il existe une sous-classe $\mathcal{F}_0 \subseteq \mathcal{F}$ dénombrable et dense pour la topologie de la convergence simple. Si la classe $\mathcal{F}$ est mesurable ponctuellement, alors les mesures extérieures redeviennent des mesures et l'enveloppe mesurable de la classe $\mathcal{F}$ coïncide presque partout avec le supremum. Dans la plupart de nos applications, cette hypothèse spécifique de mesurabilité sera satisfaite via un argument de continuité (ou semi-continuité). La démonstration de la borne inférieure est fondée sur l'étude d'une version poissonisée du processus empirique indicé par des classes de fonctions (cette fois-ci, aux bords de l'ensemble limite) combiné à un lemme fondamental de poissonisation (cf. [9]).

Par la suite, nous avons étendu le résultat (1) au cadre strictement multidimensionnel, pour $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^d$ et $\phi(Y) \in \mathbb{R}^q$. Le passage au cas multivarié lorsque $X \in \mathbb{R}^p$ ne présente pas de difficultés supplémentaires. Il suffit d'adapter convenablement les hypothèses sur le noyau K et la fenêtre h_n. Par contre, une normalisation préliminaire est nécessaire pour caractériser l'ensemble limite lorsque $Y \in \mathbb{R}^d$ ou $\phi(Y) \in \mathbb{R}^q$. Alors, nous utilisons efficacement un argument de Finkelstein afin d'obtenir une loi uniforme du logarithme à partir de (1). Comme exemple d'application directe, pour le choix particulier de $\phi(\cdot) = \mathbb{1}\{\cdot \leq \mathbf{t}\}$, $\mathbf{t} \in \mathbb{R}^d$, nous obtenons une loi limite pour la déviation en norme uniforme de

la fonction de répartition conditionnelle. Le choix de la fonction ϕ peut amener à considérer d'autres problèmatiques liées à la régression (cf. [15] et section 2). A ce propos, il est intéressant de pouvoir étendre le résultat (1) en considérant l'uniformité sur le paramètre fonctionnel ϕ, tel que, lorsque $n \to \infty$,

$$\left| \left\{ \frac{nh_n^{2k+1}}{2\log(1/h_n)} \right\}^{1/2} \sup_{\phi\in\Phi} \sup_{x\in I} \pm\left\{ \hat{m}_{\phi;n}^{(k)}(x) - m_{\phi;n}^{(k)}(x) \right\} - \sup_{\phi\in\Phi}\{\sigma_\phi(I)\} \right| = o(1), \quad \text{presque sûrement.} \tag{2}$$

Cette égalité est vérifiée par les classes de fonctions Φ possédant la propriété d'entropie présentée ci-dessus (cf. [11]), par exemple la classe des fonctions monotones et ses translatés.

En pratique, nos lois limites permettent de construire des bornes de confiance asymptotiques pour les différentes dérivées de la régression. Dans cette optique, signalons que la méthode que nous utilisons permet également de traiter la convergence uniforme presque sûre d'estimateurs de la régression plus sophistiqués tels les estimateurs obtenus par lissage polynomial local. Les estimateurs par polynômes locaux possèdent de meilleures propriétés théoriques et pratiques que les estimateurs de type Nadaraya-Watson (notamment, un meilleur biais). En reprenant les arguments développés précédemment, nous établissons une loi limite uniforme du logarithme concernant **les estimateurs par polynômes locaux** de la régression et ses dérivées (cf. [5]). En collaboration avec Anne Massiani et Pierre Ribereau, nous obtenons également des lois limites similaires pour **des estimateurs de la régression par ondelettes** (estimateurs linéaires et à seuil, [3]). Notons que le résultat présenté dans le cadre des ondelettes peut se généraliser au cadre des estimateurs par projection, i.e. en utilisant un noyau généralisé à la place du noyau habituel. En effet, la grande majorité des estimateurs par projection de la régression peuvent s'écrire sous la forme suivante :

$$\hat{m}_n(x) = \frac{\sum_{i=1}^n Y_i K_n\big(x, X_i\big)}{\sum_{i=1}^n K_n\big(x, X_i\big)}.$$

Pour conclure l'aspect statistique de notre étude, nous avons introduit une approche simple permettant de déterminer le paramètre de lissage optimal dans le cadre de la convergence presque sûre. Cette fenêtre optimale est obtenue de manière classique, en équilibrant un terme de biais au carré et un terme de variance. Notons que, pour construire nos intervalles de confiance, il faut remplacer cette fenêtre théorique optimale par une version aléatoire de type plug-in obtenue par mimimisation de l'erreur quadratique ou MSE (procédure automatique). En combinant cette approche pour le choix optimal de la fenêtre avec une procédure automatique et en injectant la formule de cette fenêtre aléatoire dans nos estimateurs de la régression, nous obtenons des intervalles de confiance uniformes asymptotiquement optimaux.

Les méthodes issues de la théorie des processus empiriques ont de nombreuses applications en statistique. Elles permettent de déterminer des lois limites pour une classe importante d'estimateurs, les M-estimateurs (cf. chapitre 3, [19]). Le dernier chapitre de la thèse est consacré à l'estimation de $\theta(\cdot)$, paramètre fonctionnel inconnu de la distribution conditionnelle de $Y|X = x$. Nous supposons que la densité conditionnelle est telle que $f_{Y|X}(y|x) := g(y;\theta(x))$ où g désigne une fonction de forme connue. En utilisant le principe d'estimation du maximum de vraisemblance local, nous pouvons construire un estimateur à noyau de $\theta(x)$, solution d'un certaine équation. Pour la déviation de cet estimateur, nous démontrons une nouvelle loi limite uniforme du logarithme en s'appuyant sur l'heuristique de la démonstration de la normalité asymptotique de l'estimateur du maximum de vraisemblance classique combinée à nos résultats sur la régression nonparamétrique. Nous avons, lorsque $n \to \infty$,

$$\left| \left\{ \frac{nh_n}{2\log(1/h_n)} \right\}^{1/2} \sup_{x\in I} \pm\left\{ \hat{\theta}_n(x) - \theta(x) \right\} - \sigma_\theta(I) \right| \overset{p.s.}{=} o(1), \tag{3}$$

où

$$\sigma_\theta(I) = \sup_{x\in I} \left\{ \frac{1}{f_X(x) I_\theta(x)} \int_{\mathbb{R}} [K^2(u)] du \right\}^{1/2}.$$

Ci-dessus, $I_\theta(x)$ désigne l'information de Fisher locale et la borne limite obtenue est du type Cramer-Rao (i.e. le problème de l'efficacité est traité). La démonstration procède en trois étapes principales, démontrer la consistance (forte) de notre estimateur, déterminer la vitesse de convergence optimale puis, via un argument du type Slutsky, établir la loi limite.

Par la suite, nous avons étendu nos travaux concernant l'estimation de la fonction de régression en considérant l'approche M-fonctionnelle. La courbe de régression est alors la quantité $m(x) = m(x;\psi)$, solution du problème suivant,

$$m(x) = \arg\min_\theta \mathbb{E}\Big[\rho(Y-\theta)\big|X=x\Big] \Leftrightarrow \mathbb{E}\Big[\psi(Y-m(x))\big|X=x\Big] = 0,$$

lorsque ρ désigne une certaine fonction de perte supposée convexe, différentiable et de dérivée $\dot\rho = \psi$. Le nouvel estimateur de la fonction de régression, noté $\hat m_n(x;\psi)$, est alors défini comme la solution ou le zéro d'une équation, i.e.

$$\sum_{i=1}^n \psi\big(Y_i - \hat m_n(x;\psi)\big) K\Big(\frac{x-X_i}{h_n}\Big) = 0.$$

Par exemple, lorsque la distribution conditionnelle admet un centre de symétrie, le choix de fonction suivant est approprié (cf. [16]),

$$\psi_k(x) = [x]_{-k}^k := \begin{cases} -k & \text{si } x \le -k, \\ x & \text{si } |x| \ge k, \\ k & \text{si } x \ge k. \end{cases}$$

Suivant le choix de k, l'estimateur $\hat m_n(x)$ oscille alors entre la moyenne conditionnelle (lorsque $k \to \infty$) et la médiane conditionnelle (lorsque $k \to 0$). Les résultats obtenus concernent de nombreux estimateurs de la régression, ils prennent la forme suivante (cf. [6])

$$\Big|\Big\{\frac{nh_n}{2\log(1/h_n)}\Big\}^{1/2} \sup_{\psi\in\Psi}\sup_{x\in I} \pm\big\{\hat m_{\psi;n}(x) - m_\psi(x)\big\} - \sigma_\Psi(I)\Big| \overset{p.s.}{=} o(1), \tag{4}$$

où

$$\sigma_\Psi(I) = \sup_{\psi\in\Psi}\sup_{x\in I}\left\{\frac{\mathrm{Var}\Big[\psi\big(Y-m_\psi(x)\big)\big|X=x\Big]}{g(x)\mathbb{E}^2\Big[\psi'\big(Y-m_\psi(x)\big)\Big]}\int_{\mathbb{R}} K^2(u)du\right\}^{1/2}.$$

Ci-dessus, Ψ désigne une classe de fonctions de type VC.

2 Perspectives de recherche

2.1 Tests d'adéquation

Une application évidente de nos lois limites (1) est la construction de tests asymptotiques concernant la norme uniforme de la courbe de régression. Pour ce faire, il est possible d'utiliser l'équivalence asymptotique entre la régression non-paramétrique à échantillonnage aléatoire et le modèle de bruit blanc gaussien (cf. [7]) combiné aux résultats de Ingster [17] sur les tests asymptotiques minimax. (Essentiellement, nous utilisons les travaux de Ingster pour obtenir une famille appropriée d'hypothèses alternatives.) En parallèle, nous étudions des méthodes récentes de **tests** minimax et adaptatifs de **spécification de modèles** (cf. [14]) de régression. Dans cette même optique, nous travaillons également sur un test lié à l'estimation optimale de l'indice des valeurs extrêmes (cf. ci-dessous).

2.2 Estimation non-paramétrique, classification de courbes

D'une manière plus générale, les outils de démonstration que nous utilisons ont un vaste potentiel d'application en estimation non-paramétrique. Par exemple, nos résultats sont applicables au cadre des estimateurs par ondelettes et permettent de retrouver des résultats minimax (cf. [3]). Il est également possible d'obtenir des résultats du type [LUL] pour une large classe d'estimateurs par projection et dans le cadre de données dépendantes, via des inégalités de découplage. Parallèlement, je souhaite aussi développer mes connaissances concernant les classes VC et leurs nombreuses applications en statistique. En **classification**, le cadre de travail est celui de la régression bornée ($X \in \mathbb{R}^d$ et $Y \in \{0,1\}$) et les méthodes de démonstration de consistence forte utilisent également la théorie des processus empiriques indicés par des classes de fonctions. En particulier, nos résultats de convergence p.s. impliquent directement la consistance forte des classifieurs fondés sur les estimateurs à noyaux de la régression présentés ci-dessus (voir également [1]).

2.3 Théorie des valeurs extrêmes

Soit $\{X_i\}_{i=1}^n$, une suite de v.a. réelles de fonction de répartition commune F appartenant au domaine d'attraction des extrêmes, i.e.

$$G_\gamma(x) = F^n(a_n x + b_n) = \begin{cases} \exp\Big\{ -(1+\gamma x)^{-1/\gamma} \Big\} & \gamma \neq 0 \\ \exp\Big\{ -\exp(-x) \Big\} & \gamma \neq 0. \end{cases}$$

Pour estimer γ, on utilise une fraction k des dernières statistiques d'ordre. L'estimateur de γ le plus célèbre est l'estimateur de Hill $\hat{H}_{n;k}$, défini par

$$\hat{H}_{n;k} := \frac{1}{k} \sum_{i=1}^{k} \log \frac{X_{n-j+1,n}}{X_{n-k+1,n}}, \quad k = 1, \ldots, n-1.$$

Pour une fonction de répartition de type Pareto, on a

$$1 - F(x) \approx x^{-1/\gamma} \Leftrightarrow U(x) = F^{-1}(1 - x^{-1}) \approx x^\gamma.$$

Nous avons donc $\log U(x) = \gamma \log x$ et le Pareto Quantile Plot $\{-\log(j/(n+1)), \log x_{n-j+1,n}\}$ est donc asymptotiquement de pente γ, ce qui justifie la formule de $\hat{H}_{n;k_n}$. Après transformation sur les statistiques d'ordres (cf. Théorème de représentation de Renyi), nous pouvons exprimer notre modèle sous une forme exponentielle et appliquer les résultats précédents, i.e. tester l'adéquation à un modèle exponentiel homogène. Ensuite, nous réalisons notre test à partir des k dernières données en augmentant k jusqu'à rejeter le modèle exponentiel. Cette procédure nous permet alors de trouver le k optimal pour construire l'estimateur de Hill. La méthodologie de démonstration repose essentiellement sur une adaptation des méthodes développées dans [14]. Notre but est ensuite d'appliquer ces résultats à des données climatologiques, illustrés de simulations.

2.4 Données Censurées

En cours d'avancement, j'explore également quelques pistes concernant l'estimation en présence de **données censurées** et plus particulièrement, l'obtention de lois limites pour l'analyse de durées de vie conditionnelles en présence de censures. Pour faire le lien entre mes travaux sur la régression et les applications biomédicales, on peut se referer à l'article de Deheuvels et Derzko [8] qui traite le cas particulier de la régression dichotomique. En épidémiologie et dans le cadre des **système multi-états**, l'utilisation

de méthodes non-paramétriques d'estimation permet de donner de meilleurs résultats pratiques que lorsqu'on suppose une relation de dépendance spécifique entre les données (modèle de régression logistique ou modèle logit). En collaboration avec M. Derzko (Sanofi-Synthelabo) et Pierre Ribereau, nous cherchons à établir les propriétés théoriques d'un estimateur de type Nadaraya-Watson dans le cadre où le régresseur est censuré à droite.

Références

[1] Abraham, C., Biau, G. et Cadre, B. (2005). On the kernel rule for function classification. *Annals of the Institute of Statistical Mathematics, to appear.*

[2] Blondin, D. (2004). Estimation nonparamétrique multidimensionnelle des dérivées de la régression. *C.R.A.S, Ser I*, **339** 713-716.

[3] Blondin, D., Massiani, A. et Ribereau, P. (2005). Vitesses de convergence uniforme presque sûre d'estimateurs non-paramétriques de la régression. *C.R.A.S, Ser I*, **340**, 525-528.

[4] Blondin, D. (2006). Vitesse de convergence presque sûre de l'estimateur à noyau du maximum de vraisemblance local. *C.R.A.S, Ser I*, **342.3**, 207-210.

[5] Blondin, D. (2006). Rates of strong uniform consistency for local least squares kernel regression estimators. Soumis à *Stat. & Proba. letters*.

[6] Blondin, D. (2006). Rates of strong uniform consistency for robust kernel-type regression M-estimators. Soumis à *Journal of Nonpar. Statist.*.

[7] Brown, L. D., Cai, T. C., Low, M. G. et Zhang, C.-H. (2002). Asymptotic equivalence theory for nonparametric regression with random design. *Ann. Statist.*, **30.3**, 688-707.

[8] Deheuvels, P. et Derzko G. (2002). Nonparametric estimation of dichotomic regression with biomedical applications. *C.R.A.S, Ser I* **334.1** 59-63.

[9] Deheuvels, P. et Mason, D. M. (1992). Functional laws of the iterated logarithm for the increments of empirical and quantile processes. *Ann. Probab.*, **20**, 1248-1287.

[10] Deheuvels, P. et Mason, D. M. (2004). General asymptotic confidence bands based on kernel-type function estimators. *Statistical Inference for Stochastic Processes*, **7.3**, 225-277.

[11] Einmahl, U., Mason, D.M. (2000). An empirical process approach to the uniform consistency of kernel-type function estimators. *Journal of Theoritical Probability*, **13.1**, 1-37.

[12] Einmahl, U. et Mason, D.M. (2005). Uniform in bandwidth consistency of kernel-type functions estimators. *Ann. Statist.*, **33.3**, 1380-1403.

[13] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Monographs on Statistics and Applied Probability, 66. Chapman & Hall, London.

[14] Guerre, E. et P. Lavergne (2005). Data-driven Rate-optimal Specification Testing in Regression Models. *Annals of Statistics* **33**, 840-870.

[15] Härdle, W., Janssen, P. and Serfing, R. (1988). Strong uniform consistency rates of estimators of conditional functionals. *Ann. Statist.*, **16.4**, 1428-1449.

[16] Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, **35**, 73-101.

[17] Ingster, Y. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives, I. *Math. Methods of Statist.*, **2.2**, 85-114.

[18] Talagrand, M. (1994). Sharper bounds for Gaussian and emprical processes. *Ann. Probab.*, **22.1**, 28-76.

[19] Van der Vaart, A. W. et Wellner, J. A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer-Verlag, New York.