

Résumé

Depuis de longues années, l'estimation non-paramétrique des fonctions de la densité et de la régression est devenu un sujet de recherche intense, permettant l'élaboration de nombreuses méthodes. L'approche non-paramétrique est fondée sur l'idée de faire peu ou de ne faire aucune hypothèse concernant la distribution de l'échantillon d'observations. Ceci, contrairement à l'approche paramétrique qui suppose que la distribution suit un certain modèle décrit par un nombre fini de paramètres. Par conséquent, l'approche paramétrique réduit le problème à l'estimation des paramètres associés au modèle. Cependant, un modèle non-paramétrique offre une applicabilité bien plus large de par son nombre plus restreint d'hypothèses.

Une classe d'estimateurs bien connue et généralement utilisée est celle des *estimateurs à noyaux* qui permettent d'estimer, entre autres, les fonctions de la densité et de la régression. Typiquement, un estimateur à noyau construit à partir de variables aléatoires indépendentes $(X_1, Y_1), \dots, (X_n, Y_n)$ à valeurs dans $\mathbb{R}^d \times \mathbb{R}^r$, est défini par

$$\hat{\varphi}_{n,h}(t) = \frac{1}{nh^d} \sum_{i=1}^n \varphi(Y_i) K\left(\frac{t - X_i}{h}\right), \quad t \in \mathbb{R}^d,$$

où K est une fonction noyau, $0 < h < 1$ est une fenêtre, et $\varphi : \mathbb{R}^r \rightarrow \mathbb{R}$ est une fonction mesurable appropriée. Nous pouvons considérer ceci comme étant une moyenne pondérée des $\varphi(Y_i)$'s, où la fonction noyau détermine la forme des poids associés à un certain voisinage de X_i , et où la fenêtre contrôle la taille de ce voisinage. A part d'être mesurable et de vérifier $\int K(x)dx = 1$, aucune restriction supplémentaire n'est essentiellement imposée quant au choix de la fonction noyau. Par contre, comme l'estimateur à noyau hérite du degré de lissage de la fonction noyau, on choisira un noyau lisse ou moins lisse selon le degré de lissage désiré pour $\hat{\varphi}_{n,h}(t)$.

Le choix de la fenêtre est plus problématique car $\hat{\varphi}_{n,h}(t)$ est particulièrement sensible à ce choix. Dans le cas des estimateurs de type Nadaraya-Watson (probablement l'estimateur non-paramétrique le plus généralement utilisé), une fenêtre trop petite conduira à un estimateur proche d'une interpolation des données, tandis qu'une fenêtre trop large produira essentiellement une ligne horizontale. Le choix de la fenêtre a également un impact important sur le biais, défini comme étant la différence entre la quantité estimée et l'espérance de l'estimateur considéré. En général, une petite fenêtre constituera un estimateur ayant peu de biais mais une grande variance, tandis qu'une fenêtre plus grande aura plutôt tendance à produire un estimateur ayant une variance réduite, mais un biais plus important. Un objectif important dans le cadre de l'estimation par des estimateurs à noyaux est donc de déterminer une fenêtre qui permettra à l'estimateur d'avoir un bon équilibre entre son biais et sa variance. De manière générale, la fenêtre la mieux adaptée variera selon la situation et dépendra des données disponibles. Par conséquent, il n'est plus possible d'étudier le comportement de tels estimateurs avec les résultats "classiques", valables uniquement pour des estimateurs construits à partir d'une séquence de fenêtres déterministes. Au lieu de cela, de nouveaux résultats sont nécessaires afin de permettre l'étude d'estimateurs fondés sur des séquences de fenêtres aléatoires.

Comme point de départ nous considérons une solution présentée par Einmahl et Mason (2005) qui consiste à rajouter aux résultats classiques un supremum sur toute une gamme de fenêtres. De tels résultats sont maintenant généralement désignés sous le nom de "résultats uniformes en fenêtre" et prennent typiquement la forme suivante :

$$\limsup_{n \rightarrow \infty} \sup_{a_n \leq h \leq b_0} \sup_{\varphi \in \mathcal{F}} \sup_{t \in I} \frac{\sqrt{nh^d} |\hat{\varphi}_{n,h}(t) - \mathbb{E}\hat{\varphi}_{n,h}(t)|}{\sqrt{|\log h| \vee \log \log n}} < \infty, \quad \text{p.s.}, \quad (\star)$$

où I est un rectangle compact dans \mathbb{R}^d , $b_0 < 1$ est une constante positive, et a_n est une suite de nombres non-aléatoires qui tend vers zéro à une vitesse appropriée. Ce supremum supplémentaire que l'on trouve dans le résultat mentionné ci-dessus nous permet entre autres de considérer les estimateurs à noyaux fondés sur des fenêtres aléatoires. En effet, lorsque $\hat{h}_n = H(X_1, \dots, X_n)$ est une suite de fenêtres telle que $\hat{h}_n \geq a_n$, le résultat asymptotique (\star) implique immédiatement qu'avec probabilité 1, $|\hat{\varphi}_{n,\hat{h}_n}(t) - \mathbb{E}\hat{\varphi}_{n,\hat{h}_n}(t)| \rightarrow 0$, uniformément sur un compact $I \subseteq \mathbb{R}^d$.

L'objectif principal de cette thèse est de démontrer comment un tel résultat d'uniformité peut également être obtenu pour de nombreuses et plus vastes classes d'estimateurs. Notre méthodologie s'inspire principalement des techniques développées dans Einmahl et Mason (2005), dont nous regroupons les principales étapes dans la Section 3.3. Par la suite, cette méthodologie sera dénommée par "Standard Methodology", ou plus simplement par [SM]. Cette méthode repose principalement sur la théorie des processus empiriques, et les outils principaux consistent essentiellement en quelques inégalités exponentielles de déviation et inégalités de moment appropriées. Tout au long des différents chapitres nous appliquerons ces techniques à plusieurs reprises afin d'établir la consistance uniforme en fenêtre de certaines classes d'estimateurs à noyaux.

Une première application de [SM] consiste à étendre le résultat (\star) vers un résultat où la norme uniforme est remplacée par une norme uniforme pondérée. Nous considérons le cas de l'estimateur à noyau de la densité $\hat{f}_{n,h}(t)$ et supposons que les poids soient déterminés par une fonction ψ appropriée. En partageant certaines idées développées par Giné, Koltchinskii et Zinn (2004), nous fournissons sous des hypothèses supplémentaires concernant les suites a_n, b_n et la fonction de densité, des conditions nécessaires et suffisantes afin que le processus

$$\sup_{a_n \leq h \leq b_n} \sup_{t \in \mathbb{R}^d} \frac{\sqrt{nh^d} |\psi(t)(\hat{f}_{n,h}(t) - \mathbb{E}\hat{f}_{n,h}(t))|}{\sqrt{|\log h|}}$$

soit stochastiquement et presque sûrement borné. Dans ce processus, ψ est une fonction de poids (pas nécessairement bornée), et a_n et b_n sont des suites de fonctions à variation régulière à l'infini d'indice négatif. Une preuve détaillée de ce résultat sera présentée dans le Chapitre 4.

Dans le cinquième chapitre, nous supposons que la fonction de régression $m(t) = \mathbb{E}[Y|X = t]$, $t \in \mathbb{R}$ soit $p + 1$ fois dérivable au point $x_0 \in \mathbb{R}$, et nous considérons une classe plus vaste d'estimateurs pour $m(x_0)$, la classe des estimateurs localement polynomiaux. Ceux-ci sont définis comme étant les solutions d'un problème des moindres carrés pondérés, où les poids sont fixés par $K((x_0 - X_i)/h)$. Nous notons que l'estimateur de Nadaraya–Watson appartient à cette classe, car il est la solution du problème des moindres carrés pondérés lorsque $p = 0$. Il s'avère que la méthode décrite dans [SM] et reposant sur la théorie des processus empiriques, est également applicable pour établir la consistance uniforme en fenêtre des estimateurs

localement polynomiaux. Si nous notons $\hat{m}_{n,h}^{(p)}(x_0)$ pour l'estimateur localement polynomial de degré $p \geq 0$ au point x_0 , nous démontrons dans la Section 5.2 que

$$\sup_{(\frac{c \log n}{n})^\gamma \leq h \leq b_n} \sup_{x_0 \in I} |\hat{m}_{n,h}^{(p)}(x_0) - m(x_0)| \longrightarrow 0, \quad \text{p.s.},$$

où b_n est une suite arbitraire qui tend vers zéro, et où $\gamma = 1$ ou $\gamma = 1 - 2/q$ selon si Y est bornée ou a un moment borné d'ordre $q > 2$.

Le sixième chapitre sera consacré à la consistance ponctuelle uniforme en fenêtre du processus $\hat{\varphi}_{n,h}(t)$, c.à.d. uniformément pour toute une rangée de fenêtres, mais pour un point $t \in \mathbb{R}^d$ fixé. Bien que le résultat uniforme décrit dans (\star) implique la convergence ponctuelle, plusieurs améliorations peuvent être réalisées à plusieurs niveaux. Nous montrons que

$$\limsup_{n \rightarrow \infty} \sup_{a_n \leq h \leq b_0} \sup_{\varphi \in \mathcal{F}} \frac{\sqrt{nh^d} |\hat{\varphi}_{n,h}(t) - \mathbb{E}\hat{\varphi}_{n,h}(t)|}{\sqrt{\log \log n}} < \infty, \quad \text{p.s.},$$

où $a_n^d = c \log \log n/n$ ou $a_n^d \geq n^{-1}(\log n)^{2/(p-2)}$ selon si la fonction enveloppe de la classe \mathcal{F} admet une fonction génératrice bornée, ou a un moment borné d'ordre $p > 2$. Clairement, une première amélioration concerne la vitesse de convergence, qui est sensiblement meilleure dans ce cas-ci. Une deuxième amélioration est réalisée sur la gamme des fenêtres autorisées, qui est plus large que celle dans le cas uniforme. Enfin, le résultat obtenu ci-dessus est applicable plus généralement que le cas uniforme (\star) , qui n'est valable que pour des classes bornées ou pour lesquelles un moment d'ordre $p > 2$ de la fonction enveloppe existe. D'un point de vue de consistance, ceci implique qu'en terme de vitesses de convergence et de suites de fenêtres autorisées, les classes de fonctions bornées sont équivalentes à celles ayant une fonction enveloppe admettant une fonction génératrice bornée. Ce résultat surprenant semblait ne pas encore être connu, ni même dans le cas classique traitant les fenêtres déterministes. Jusqu'à présent, la question de savoir si une telle conclusion peut également être obtenue dans le cas où la convergence est uniforme sur des ensembles compacts (donc pour $t \in I$), est toujours ouverte.

Une application importante du résultat mentionné ci-dessus est la consistance uniforme en fenêtre d'une version à noyau de l'estimateur de Hill pour l'index d'une distribution de type Pareto. Si $\hat{\tau}_{n,h}$ représente cet estimateur fondé sur n variables indépendentes et suivant une loi Pareto

d'index $\tau > 0$, il est démontré dans la Section 6.4 que

$$\sup_{a_n \leq h \leq b_n} |\hat{\tau}_{n,h} - 1/\tau| = o_{\mathbb{P}}(1),$$

où a_n et b_n sont des suites non-aléatoires vérifiant entre autres $b_n \rightarrow 0$ et $na_n \rightarrow \infty$.

Comme dernière application de [SM], nous considérons les “ U -statistiques conditionnelles” qui constituent une classe d’estimateurs à noyaux beaucoup plus large que celle des estimateurs de type Nadaraya–Watson. Ces estimateurs sont des estimateurs consistants de la fonction de régression “multivariée”, représentée par $m_{\varphi}(\mathbf{t}) = \mathbb{E}[\varphi(Y_1, \dots, Y_m) | (X_1, \dots, X_m) = \mathbf{t}]$, $\mathbf{t} \in \mathbb{R}^m$. Dans le Chapitre 8, nous représentons une U -statistique conditionnelle par $\hat{m}_{n,h,\varphi}(\mathbf{t})$ et nous établissons un résultat de consistance uniforme en fenêtre pour $\hat{m}_{n,h,\varphi}(\mathbf{t})$, qui est une conséquence du résultat asymptotique suivant :

$$\limsup_{n \rightarrow \infty} \sup_{a_n^{\gamma} \leq h < b_n} \sup_{\varphi \in \mathcal{F}} \sup_{\mathbf{t} \in I^m} \frac{\sqrt{nh^m} |\hat{m}_{n,h,\varphi}(\mathbf{t}) - \mathbb{E} \hat{m}_{n,h,\varphi}(\mathbf{t})|}{\sqrt{|\log h| \vee \log \log n}} < \infty, \quad \text{p.s.},$$

où $a_n = c(\log n/n)^{1/m}$, $I^m = I \times \dots \times I$, et où $\gamma = 1$ ou $\gamma = 1 - 2/p$ selon si la fonction enveloppe de \mathcal{F} est bornée ou a un moment borné d’ordre $p > 2$. Notons que ce résultat implique entre autres la consistance uniforme en fenêtre des estimateurs de type Nadaraya–Watson qui sont des cas particuliers de $\hat{m}_{n,h,\varphi}(\mathbf{t})$ lorsque $m = 1$.