Condition nécessaire et suffisante de convergence en loi de l'estimateur des plus proches voisins

Alain Berlinet, Rémi Servien

Condition nécessaire et suffisante de convergence en loi de l'estimateur des plus proches voisins

Alain Berlinet & Rémi Servien

Institut de Mathématiques et de Modélisation de Montpellier
UMR CNRS 5149
Equipe de Probabilités et Statistique
Université Montpellier II
CC 051, Place Eugène Bataillon
34095 Montpellier Cedex 5, France

Résumé

L'estimateur des plus proches voisins de la densité est un estimateur simple et facile à mettre en oeuvre. Sa normalité asymptotique a été établie par Moore et Yackel (1977) sous des hypothèses faisant intervenir les dérivées de la densité. Sans faire d'hypothèse de continuité sur la densité, nous donnons une condition nécessaire et suffisante de convergence en loi de cet estimateur. Nous utilisons pour cela l'indice de régularité d'une mesure de probabilité (Beirlant, Berlinet et Biau (2008)) qui intervient de fait dans la loi limite.

Mots-clés **– Normalité asymptotique, Estimateur des plus proches voisins, Indice de régularité.**

Abstract

The nearest neighbour estimator is a well-known estimator of the density. Its asymptotic normality was obtained by Moore and Yackel (1977) under conditions on the derivatives of the density. We establish a necessary and sufficient condition for the existence of a limit distribution without any continuity hypothesis on the density. We use for this the regularity index of a probability measure (Beirlant, Berlinet and Biau (2008)) which plays a part in the asymptotic distribution.

Keywords **– Asymptotic normality, Nearest neighbour estimator, Regularity index.**

Un estimateur de la densité simple et facile à mettre en oeuvre est l'estimateur des plus proches voisins défini de la façon suivante. Soit $(k_n)_{n\geq 1}$ une suite d'entiers positifs tels que pour tout $n \geq 1$ on ait $1 \leq k_n \leq n$ et $(X_i)_{i\geq 1}$ une suite de variables aléatoires indépendantes de même loi μ à densité f par rapport à la mesure de Lebesgue λ sur $\mathbb{R}^d$ muni de sa topologie usuelle. L'estimateur des plus proches voisins de f au point x est défini (presque sûrement) par

$$f_{k_n}(x) = \frac{k_n}{n\lambda(\overline{B}_{k_n}(x))},$$

où $\overline{B}_{k_n}(x)$ est la plus petite boule fermée de centre x contenant au moins k_n points de l'échantillon. L'entier k_n joue donc le rôle d'un paramètre de lissage.

En analyse discriminante, Fix et Hodges (1951) ont introduit la règle de classification basée sur les plus proches voisins (voir également Devroye, Györfi et Lugosi (1996) sur ce sujet). L'application de cette règle à l'estimation de la densité est due à Loftsgaarden et Quesenberry (1965) qui ont démontré la convergence en probabilité de l'estimateur sous les hypothèses suivantes

$$\lim_{n\to\infty} k_n = \infty \qquad \text{et} \qquad \lim_{n\to\infty} \frac{k_n}{n} = 0. \tag{1}$$

Par la suite, Moore et Yackel (1977) ont obtenu le résultat asymptotique suivant

$$\sqrt{k_n}\frac{f_{k_n}(x) - f(x)}{f(x)} \xrightarrow{L} \mathcal{N}(0,1)$$

si f est différentiable en x et à dérivées partielles bornées dans un voisinage de x avec $f(x) > 0$ et en ajoutant la condition

$$\lim_{n\to\infty} \frac{k_n}{n^{2/(d+2)}} = 0$$

aux conditions (1) (Bosq et Lecoutre (1987)).
Nous montrons qu'il est possible d'obtenir une condition nécessaire et suffisante pour la convergence en loi de l'estimateur f_{k_n} sous des hypothèses beaucoup moins restrictives sur f. Soit x un point de $\mathbb{R}^d$. Pour δ réel positif notons $B_\delta(x)$ la boule ouverte de centre x et de rayon δ. Afin de mesurer le comportement local de $\mu(B_\delta(x))$ par rapport à $\lambda(B_\delta(x))$ nous pouvons considérer le quotient de ces deux mesures. Ainsi, si pour x fixé la limite suivante

$$\lim_{\delta\to 0} \frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))} \tag{2}$$

existe, alors x est appelé *point de Lebesgue* de la mesure μ. Il est important de noter que la notion de point de Lebesgue permet d'élargir certains résultats en diminuant les

contraintes sur les fonctions à estimer. Dans ce contexte, Berlinet et Levallois (2000) définissent un point ρ-régulier de la mesure μ comme un point de Lebesgue x de μ tel que

$$\left|\frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))} - \lim_{\eta\to 0}\frac{\mu(B_\eta(x))}{\lambda(B_\eta(x))}\right| \leq \rho(\delta), \tag{3}$$

où ρ est une fonction mesurable telle que $\lim_{\delta\downarrow 0}\rho(\delta) = 0$. Par exemple, si $d = 1$ et si la mesure μ a une densité f avec une dérivée f' bornée par une constante C_x dans un voisinage de x, alors nous avons la ρ-régularité en x avec $\rho(\delta) = C_x\delta$. Si f est une fonction localement höldérienne en x avec un exposant α_x, cela implique la ρ-régularité avec $\rho(\delta) = C_x\delta^{\alpha_x}/(\alpha_x + 1)$. Il est également possible de trouver des exemples de mesures ρ-régulières mais avec un mauvais comportement local de la densité, comme des discontinuités du second ordre.

Nous supposerons ici qu'une relation plus précise que la ρ-régularité a lieu. Nous considérons qu'en x, point de Lebesgue de la mesure μ de densité f, nous avons

$$\frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))} = f(x) + C_x\delta^{\alpha_x} + o(\delta^{\alpha_x}) \text{ quand } \delta \downarrow 0, \tag{4}$$

où C_x est une constante différente de 0 et α_x un nombre réel strictement positif que nous appelons *indice de régularité*. Ces constantes sont alors uniques et il est clair que cette relation implique la ρ-régularité avec $\rho(\delta) = C_x\delta^{\alpha_x}$. Beirlant, Berlinet et Biau (2008) ont utilisé cet indice pour résoudre certains problèmes liés à f_{k_n} (dépendance au nombre de voisins k_n, calcul d'un k_n optimal). Cet indice joue également un rôle primordial dans le résultat qui suit.

A l'aide de ces définitions, nous prouvons finalement le résultat suivant

Théorème 1 *Si x est un point de Lebesgue où (4) est vérifié avec $f(x) > 0$, et sous les conditions (1), la variable aléatoire*

$$T_n(x) = \sqrt{k_n}\frac{f_{k_n}(x) - f(x)}{f(x)}$$

converge en loi si et seulement si la suite

$$\left(\frac{k_n^{1+1/2\alpha_x}}{n}\right)_{n\geq 1}$$

a une limite finie κ. Lorsque cette condition est vérifiée, la loi asymptotique de $T_n(x)$ est

$$\mathcal{N}\left(\frac{C_x\kappa^{\alpha_x}}{2^{\alpha_x}}\left(\frac{1}{f(x)}\right)^{\alpha_x+1}, 1\right).$$

Nous avons obtenu une condition nécessaire et suffisante pour la convergence en loi de $T_n(x)$ ainsi que l'expression de la loi asymptotique, qui ne peut être que normale. On observera que l'on retrouve la condition suffisante de Moore et Yackel (1977) lorsque $\kappa = 0$ et que les conditions de différentiabilité sont satisfaites. Les hypothèses faites font intervenir l'indice de régularité de la mesure étudiée mais ne demandent aucune condition de continuité autour du point d'estimation.

Bibliographie

[1] Beirlant, J., Berlinet, A. et Biau, G. (2008) Higher order estimation at Lebesgue points, *Annals of the Institute of Statistical Mathematics*, 60, 651–677.
[2] Berlinet, A. et Levallois, S. (2000) Higher order analysis at Lebesgue points, *G. G. Roussas Festschrift - Asymptotics in Statistics and Probability*, M.L. Puri, 1–16.
[3] Bosq, D. et Lecoutre, J.-P. (1987) *Théorie de l'Estimation Fonctionnelle*, Economica.
[4] Devroye, L., Györfi, L. et Lugosi, G. (1996) *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag, New-York.
[5] Fix, E. et Hodges, J.L. Jr. (1951) Discriminatory analysis, nonparametric discrimination: Consistency properties, *USAF School of Aviation Medicine*, Randolph Field, Texas.
[6] Loftsgaarden, D. et Quesenberry, C.P (1965) A nonparametric estimate of a multivariate density function, *The Annals of Mathematical Statistics*, 1049–1051.
[7] Moore, D.S. et Yackel, J.W. (1977) Large sample properties of nearest neighbour density function estimates, *Statistical Decision Theory and Related Topics II*, Gupta, S.S. et Moore, D.S., Academic Press, New-York.