Estimation de régularité locale

Alain Berlinet, Rémi Servien

Estimation de régularité locale[1]

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
rservien@math.univ-montp2.fr

Résumé

La dérivée symétrique d'une mesure de probabilité en un point de Lebesgue peut souvent être approximée à l'aide d'un développement faisant intervenir un indice de régularité. La connaissance de cet indice est d'un intérêt pratique. En effet, il permet par exemple de déterminer le comportement local de la mesure étudiée. Il intervient aussi dans l'évaluation du nombre de voisins à prendre en compte dans l'estimation de la densité ou dans l'estimation du mode. Cet indice de régularité est difficile à estimer à cause de son caractère fortement local comme nous pourrons le remarquer sur des exemples. Cependant, Beirlant, Berlinet et Biau (2008, *Annals of the Institute of Statistical Mathematics*, 60, 651-677) ont précédemment défini un estimateur de cet indice dans $\mathbb{R}^d$ en utilisant l'estimateur des plus proches voisins. Nous définissons et étudions de nouveaux estimateurs de l'indice de régularité basés sur différents estimateurs de la fonction de répartition.

Mots-clés **– Point de Lebesgue, Indice de régularité, Densité non lisse, Estimation du mode.**

Abstract

The symmetric derivative of a probability measure at a Lebesgue point can often be specified by a relationship involving a regularity index. Knowledge of this index is of practical interest. Indeed, it allows us to specify the local behavior of the measure under study. It is also useful in the evaluation of number of neighbors to take into account in density estimation or in mode estimation. This regularity index is hard to estimate due to his highly local nature as we can see in some examples. However, Beirlant, Berlinet and Biau (2008, *Annals of the Institute of Statistical Mathematics*, 60, 651-677) defined an estimator of the regularity index using the k-nearest neighbor d-dimensional density estimate. We define and study new estimators of this regularity index using different estimators of the cumulative distribution function.

[1]Travail réalisé dans le cadre d'une thèse co-financée par le CNRS et la région Languedoc-Roussillon.

Le sujet principal de cet exposé est lié au problème général de dérivation des mesures (Rudin (1987), Dudley (1989)). Il trouve ses motivations dans l'étude de problèmes d'estimation quand les conditions de régularité habituelles ne sont pas vérifiées. En effet, de nombreux théorèmes de convergence font intervenir des hypothèses de continuité qui ne sont en pratique pas toujours satisfaites. Nous utilisons donc des conditions moins contraignantes permettant d'étudier la régularité de la fonction de densité associée à la mesure considérée.

Un paramètre α_x appelé *indice de régularité* apparaît lorsqu'on essaie d'étudier localement le comportement d'une fonction de densité dérivée d'une mesure quelconque. Ce paramètre de régularité étant fortement local, son estimation est difficile.

Pour $d \geq 1$, notons $\mathcal{B}(\mathbb{R}^d)$ la tribu borélienne de $\mathbb{R}^d$. Nous considérons μ une mesure de probabilité sur $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Notons λ la mesure de Lebesgue sur $\mathbb{R}^d$ muni d'une norme notée $||.||$. Soit x un point de $\mathbb{R}^d$, δ un réel positif et $B_\delta(x)$ 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 2 mesures. Ainsi, si pour x fixé la limite suivante

$$f(x) = \lim_{\delta \to 0} \frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))} \tag{1}$$

existe, alors x est appelé un *point de Lebesgue* de la mesure μ. Si μ est absolument continue par rapport à λ, nous pouvons sélectionner parmi toutes les densités obtenues à partir de μ, une densité particulière f, qui satisfait (1) en tout point où cette limite existe. Notons que la notion de point de Lebesgue est plus large que la notion de continuité. Elle permet donc 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 tous les points de Lebesgue x de μ tels que

$$\left| \frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))} - f(x) \right| \leq \rho(\delta), \tag{2}$$

où ρ est une fonction mesurable telle que $\lim_{\delta \to 0} \rho(\delta) = 0$. 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 le voisinage de x, alors nous avons $\rho(\delta) = C_x \delta$ et x est ρ-régulier. Il est aussi clair que si f est une fonction localement hölderienne en x avec un exposant α_x, cela implique $\rho(\delta) = C_x/(\alpha_x + 1)\delta^{\alpha_x}$. De plus, il est possible de trouver des exemples de mesures ρ-régulières mais avec un mauvais comportement local de la densité, comme des discontinuités du second ordre. Pour des exemples, on peut se référer à Berlinet et Levallois (2000).

La fonction ρ de (2) n'est clairement pas unique et dépend de la norme choisie sur $\mathbb{R}^d$. Il est possible d'aller plus loin que la relation (2) et d'envisager qu'en x, point de

Lebesgue de la mesure μ, nous ayons

$$\frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))} = f(x) + C_x\delta^{\alpha_x} + o(\delta^{\alpha_x}) \text{ quand } \delta \downarrow 0, \tag{3}$$

où C_x est une constante différente de 0 et α_x un nombre réel strictement positif. Ces constantes sont alors uniques et, trivialement, cette relation implique la ρ-régularité en x avec $\rho(\delta) \sim \delta^{\alpha_x}$.

L'indice α_x est appelé *indice de régularité* et reflète le degré de régularité de la mesure μ par rapport à la mesure de Lebesgue λ. Concrètement, plus α_x sera grand, plus la dérivée de μ sera lisse autour du point x.

La connaissance de cet indice est d'un intérêt pratique concernant le comportement local de la mesure. Il intervient également dans différents problèmes d'estimation : dans l'estimation du nombre optimal de voisins pour l'estimateur des plus proches voisins de la densité ou encore dans l'estimation du mode d'une densité inconnue.

Un premier estimateur de l'indice de régularité a été défini par Beirlant, Berlinet et Biau (2008). Ils démontrent tout d'abord que, si (3) est vérifiée, on obtient pour $\tau > 1$,

$$\lim_{\delta\to 0} \frac{\Phi_{\tau^2\delta}(x) - \Phi_{\tau\delta}(x)}{\Phi_{\tau\delta}(x) - \Phi_\delta(x)} = \tau^{\alpha_x}, \tag{4}$$

où

$$\Phi_\delta(x) = \frac{\mu(B_\delta(x))}{\lambda(B_\delta(x)}.$$

Ils utilisent ensuite l'estimateur des k_n-plus proches voisins usuel défini 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. En le combinant avec la relation (4) ils définissent l'estimateur suivant pour $\tau > 1$,

$$\hat{\alpha}_{n,x} = \frac{d}{\log\tau} \log \frac{f_{\lfloor\tau^2 k_n\rfloor}(x) - f_{\lfloor\tau k_n\rfloor}(x)}{f_{\lfloor\tau k_n\rfloor}(x) - f_{\lfloor k_n\rfloor}(x)}$$

si $[f_{\lfloor\tau^2 k_n\rfloor}(x) - f_{\lfloor\tau k_n\rfloor}(x)]/[f_{\lfloor\tau k_n\rfloor}(x) - f_{\lfloor k_n\rfloor}(x)] > 0$ et $\hat{\alpha}_{n,x} = 0$ sinon. Cet estimateur converge en probabilité vers α et sa normalité asymptotique est démontrée. Il est également possible, à l'aide d'un terme correctif dépendant de α, d'améliorer les performances de l'estimateur f_{k_n}.

En remplaçant dans l'équation (4) la fonction de répartition par ses estimateurs empiriques ou à noyaux, nous déterminons de nouveaux estimateurs convergents pour l'indice de régularité.

L'indice de régularité intervient aussi dans l'estimation du mode θ d'une densité de probabilité, c'est à dire l'argument maximum de cette densité. En partant de l'échantillon $S_n = \{X_1, ..., X_n\}$ tiré de la densité inconnue f, Abraham, Biau et Cadre (2003) utilisent tout d'abord l'estimateur à noyau de la densité pour estimer la densité f. Pour tout $x \in \mathbb{R}^d$, on définit alors l'estimateur f_{h_n} par

$$f_{h_n}(x) = \frac{1}{nh_n^d} \sum_{i=1}^{n} k\left(\frac{x - X_i}{h_n}\right),$$

où k est un noyau et h_n la fenêtre de lissage strictement supérieure à 0 et telle que $h_n \to 0$ quand $n \to +\infty$. Ils définissent ensuite leur estimateur θ_n du mode comme un élément de l'ensemble

$$\left\{ x \in S_n : f_{h_n}(x) = \max_{1 \leq i \leq n} f_{h_n}(X_i) \right\}.$$

Cet estimateur converge alors presque sûrement vers le mode θ. Il est à noter que ce résultat a été obtenu dans le cas d'une densité f continue autour du mode. Nous avons élargi ce résultat avec les deux seules hypothèses suivantes

$$\text{diam } A(\varepsilon) \to 0 \text{ lorsque } \varepsilon \to 0 \ \ (H1),$$

où $A(\varepsilon) = \left\{ x \in \mathbb{R}^d : f(x) > f(\theta) - \varepsilon \right\}$ et

$$\forall \varepsilon > 0, \ \mathrm{P}\{X \in A(\varepsilon)\} > 0 \ \ (H2).$$

Il ne subsiste donc plus d'hypothèses sur l'éventuelle continuité de la fonction f qui est remplacée par l'hypothèse $(H2)$. Cette dernière sera notamment vérifée dans le cas où θ est un point de Lebesgue.

Ils obtiennent ensuite une bonne vitesse de convergence et un intervalle de confiance asymptotique dépendant de certaines constantes dont κ, appelé *indice de pic*, défini comme un réel strictement supérieur à 0 et vérifiant l'inégalité suivante

$$0 < \liminf_{\varepsilon \to 0} \frac{\text{diam } A(\varepsilon)}{\varepsilon^\kappa} \leq \limsup_{\varepsilon \to 0} \frac{\text{diam } A(\varepsilon)}{\varepsilon^\kappa} < \infty. \tag{5}$$

Nous élargirons également ces résultats en utilisant nos hypothèses moins restrictives. L'indice de pic mesure donc la diminution plus ou moins rapide des valeurs de la densité autour du mode θ. Plus la densité diminuera brutalement autour de θ, plus κ sera élevé et inversement. Si la mesure associée à la densité f a un indice de régularité α en θ, alors le niveau $f(\theta + \delta) - f(\theta)$ est d'ordre δ^α. Par le même raisonnement nous avons diam $A(\varepsilon)$ d'ordre $\varepsilon^{1/\alpha}$ et f a donc un indice de pic qui vaut $1/\alpha$ en son mode θ. Nous avons donc

$$\kappa = \frac{1}{\alpha}.$$

Trouver un bon estimateur de l'indice de pic revient donc à en trouver un pour l'indice de régularité. Ce dernier est en effet nécessaire dans le but de générer des intervalles de confiance asymptotiques pour l'estimateur θ_n du mode.

En résumé, l'indice de régularité α_x nous donne d'importantes indications sur le comportement d'une mesure autour du point x. Il est important de noter que α_x existe dans le cas de densité non nécessairement continue, ceci nous garantissant un large cadre de travail. Il intervient également dans différents calculs de paramètres intimement liés au caractère lisse ou non lisse de la mesure. Nous avons en effet pu voir son importance dans le calcul d'un estimateur du mode ou dans celui du nombre de voisins en estimation de la densité.

Bibliographie

[1] Abraham, C., Biau, G. et Cadre, B. (2003) Simple estimation of the mode of a multivariate density. *The Canadian Journal of Statistics*, 31, 23–34.
[2] Beirlant, J., Berlinet, A. et Biau, G. (2008) Higher order estimation at Lebesgue points. *Annals of the Institute of Statistical Mathematics*, 60, 651–677.
[3] Berlinet, A. et Levallois, S. (2000) Higher order analysis at Lebesgue points. In M. Puri, editor, *G.G Roussas Festschrift - Asymptotics in Statistics and Probability*, 1–16.
[4] Dudley, R. (1989) *Real Analysis and Probability*, Chapman and Hall, New-York.
[5] Rudin, W. (1987) *Real and Complex Analysis*, McGraw-Hill, New-York.