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

Rémi Servien

Laboratoire Jean Kuntzmann
INP Grenoble

Neuvième Colloque Jeunes Probabilistes et Statisticiens
Mai 2010
Le Mont Dore

Plan de la présentation

Estimateur classique des k_n plus proches voisins

- $S_n = \{X_1, ..., X_n\}$ de v.a. iid sur $\mathbb{R}^d$ tiré à partir d'une mesure de probabilité μ
- Estimateur de la densité des k_n-plus proches voisins :

$$f_{k_n}(x) = \frac{k_n}{n\lambda(B_{k_n}(x))}$$

où $B_{k_n}(x)$ est la plus petite boule de centre x contenant k_n points de l'échantillon.

- Estimateur convergent

Théorème (Moore et Yackel, 1977)

Si f est continue et à dérivées bornées dans un voisinage de x avec $f(x) > 0$ et sous les conditions de convergence de f_{k_n}

$$\lim_{n\to\infty} k_n = \infty \text{ et } \lim_{n\to\infty} \frac{k_n}{n} = 0$$

et sous la condition supplémentaire

$$\lim_{n\to\infty} \frac{k_n}{n^{2/3}} = 0$$

alors la variable aléatoire

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

converge vers une loi $\mathcal{N}(0,1)$.

Définitions

- x est un point de Lebesgue de la mesure μ si

$$\lim_{\delta \to 0^+} \frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))} = f(x) \text{ existe.}$$

- x est un point ρ-régulier de la mesure μ si

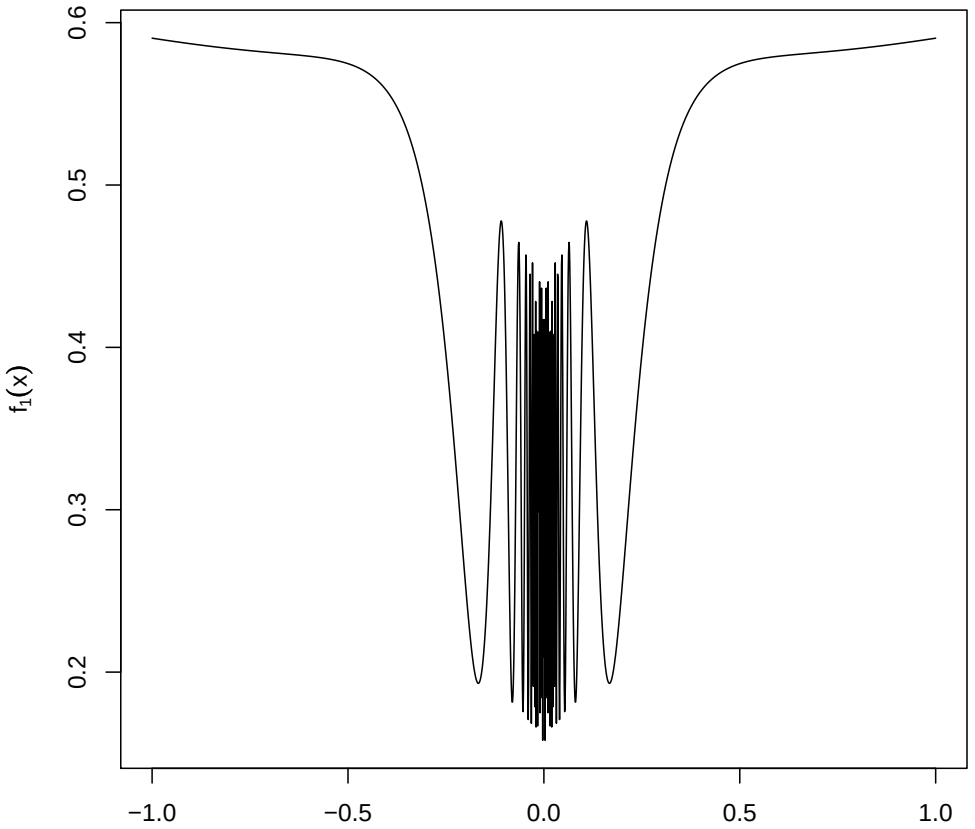
$$\left| \frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))} - f(x) \right| \leq \rho(\delta), \tag{1}$$

où ρ est une fonction mesurable telle que $\lim_{\delta \to 0^+} \rho(\delta) = 0$ (Berlinet et Levallois, 2000).

Exemple

On définit, pour $x \in [-1, 1]$ et $x \neq 0$, la densité

$$f_1(x) = \frac{\sqrt{|x|} + 2 - \cos(1/x) + 2x \sin(1/x)}{c}$$



Exemple

- Discontinuité du 2nd ordre en 0.
- Mais

$$\frac{\mu_1(B_h(0))}{2h} = \frac{2}{c} + \frac{2}{3c}h^{1/2} + o(h^{1/2})$$

- 0 est un point de Lebesgue de la mesure μ_1 de densité f_1

Théorème (Berlinet et Levallois, 2000)

Sous les conditions de convergence de f_{k_n}, si x est un point ρ-régulier de la mesure μ avec $f(x) > 0$, alors la condition

$$\sqrt{k_n}\rho(R_n(x)) \xrightarrow{P} 0$$

lorsque n tend vers l'infini implique la convergence en distribution de la variable aléatoire

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

vers une loi $\mathcal{N}(0,1)$.

Indice de régularité

Si en x, point de Lebesgue de la mesure μ, nous avons

$$\frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))} = f(x) + C_x\delta^{\alpha_x} + o(\delta^{\alpha_x}) \text{ quand } \delta \to 0^+, \qquad (2)$$

où $C_x \neq 0$ et $\alpha_x > 0$.

L'indice α_x est appelé indice de régularité de la mesure μ au point x.

Exemple 1

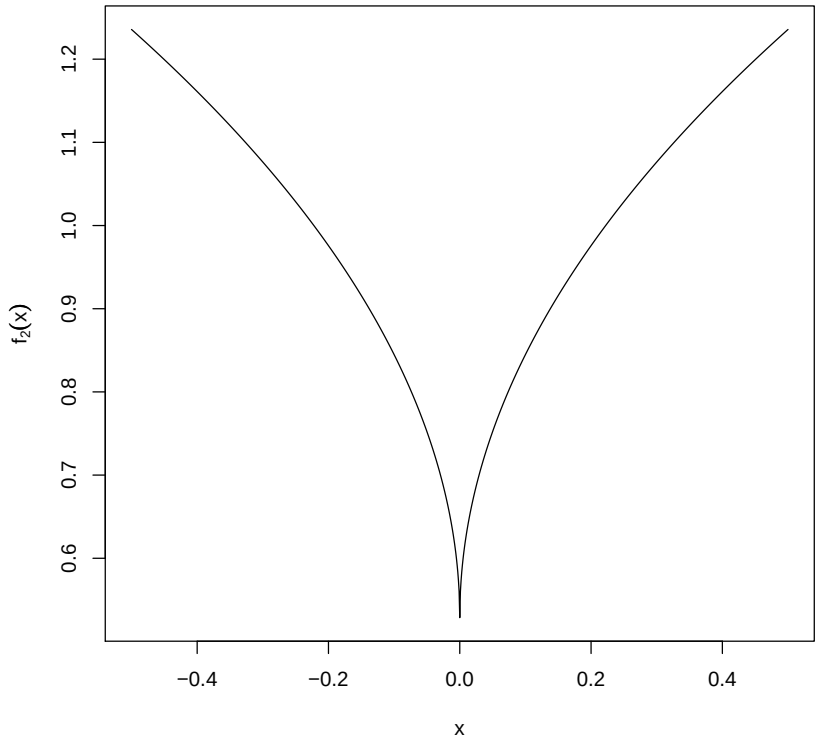
- $$f_1(x) = \frac{\sqrt{|x|} + 2 - \cos(1/x) + 2x\sin(1/x)}{c}$$

- $$\frac{\mu_1(B_h(0))}{2h} = \frac{2}{c} + \frac{2}{3c}h^{1/2} + o(h^{1/2})$$

- 0 est un point de Lebesgue de la mesure μ_1 de densité f_1 avec un indice de régularité $\alpha_0 = 1/2$.

Exemple 2

$$f_2(x) = 1 - \sqrt{2}/3 + \sqrt{|x|} \text{ si } x \in [-0.5, 0.5]$$



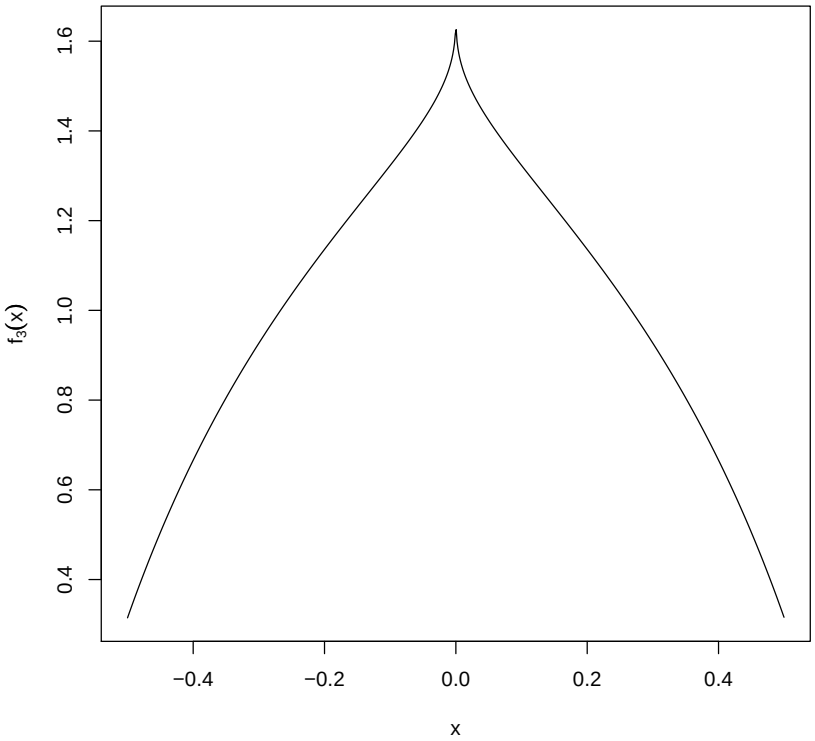
Pour $\mu_2(x)$ on a :

- $\alpha_x = 1$ si $x \neq 0$
- $\alpha_x = 0.5$ si $x = 0$

Caractère très fortement local de l'indice de régularité

Exemple 3

$$\begin{aligned} f_3(x) &= \frac{1}{\log|x|} + 1 - a \text{ si } x \in [-0.5, 0.5]\backslash 0 \\ &= 1 - a \qquad\qquad \text{si } x = 0 \end{aligned}$$



- 0 est un point de Lebesgue de la mesure μ_3
- On a ρ-régularité avec $\rho(\delta) = -1/\log\delta$
- Il n'y a pas d'indice de régularité en 0.

Théorème

Si x est un point de Lebesgue où (2) est vérifié avec $f(x) > 0$, et sous les conditions de convergence de f_{k_n}, 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

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

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

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

Exemple

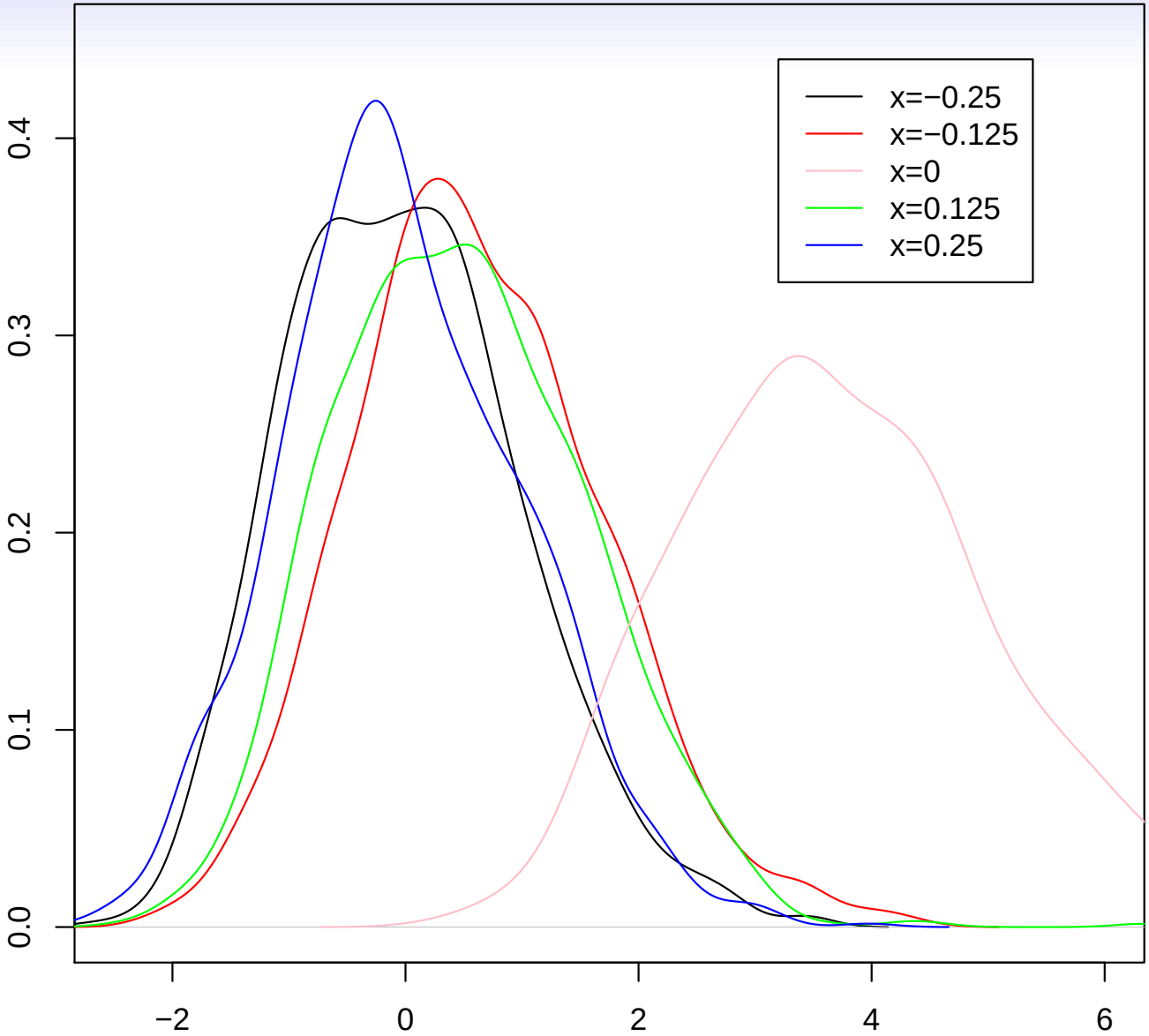
$f_2(x) = 1 - \sqrt{2}/3 + \sqrt{|x|}$ si $x \in [-0.5, 0.5]$

Pour $\mu_2(x)$ on a :

- $\alpha_x = 1$ si $x \neq 0$
- $\alpha_x = 0.5$ si $x = 0$

Sans prendre en compte la spécificité de f_2 en 0, on choisit $k_n = \sqrt{n}$ pour $n = 10000$.

Exemple



Simulations

$f_2(x) = 1 - \frac{\sqrt{2}}{3} + \sqrt{|x|}, x \in [-0.5, 0.5]$

Hypothèses du théorème : en 0 : $k_n << n^{0.5}$ et $k_n << n^{2/3}$ sinon.

$x =$	-0.25	-0.125	0	0.125
$k_n = n^{1/3}$	[0.74 ;1.25]	[0.72 ;1.22]	[0.34 ;0.55]	[0.63,1.06]
$k_n = n^{2/5}$	[0.90 ;1.32]	[0.66 ;0.96]	[0.37 ;0.53]	[0.65,0.94]
$k_n = n^{0.6}$	[1.02 ;1.05]	[0.85 ;0.87]	[0.55 ;0.56]	[0.85 ;0.88]
$f_2(x)$	1.03	0.88	0.53	0.88

TABLEAU: Estimations de f_2 en différents points x pour $n = 500000$ points

Autre utilisation de l'indice de régularité

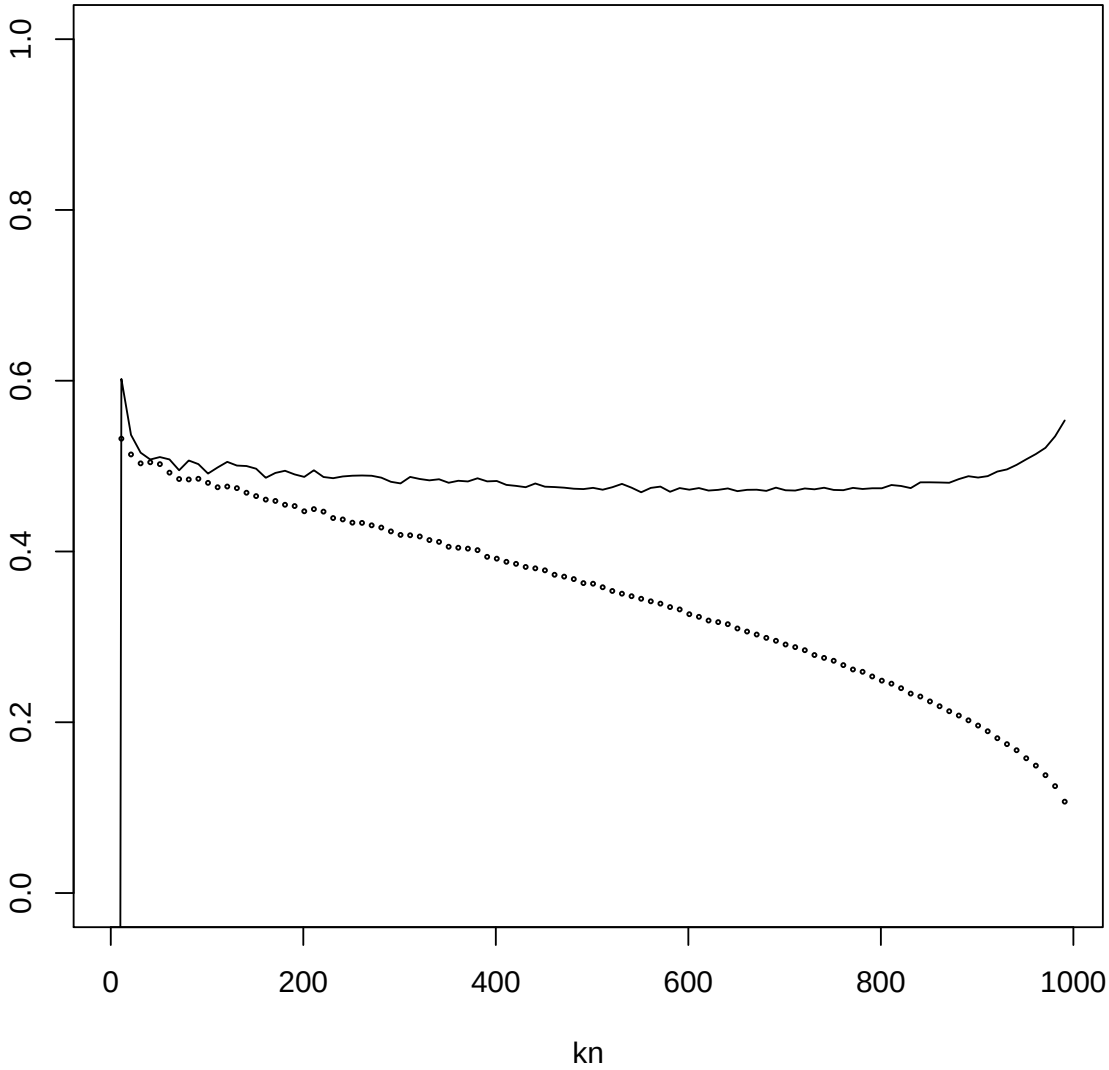
- Estimateur de la densité des k_n-plus proches voisins :

$$f_{k_n}(x) = \frac{k_n}{n\lambda(B_{k_n}(x))}.$$

- Estimateur récursif (Beirlant, Berlinet et Biau (2008)) :

$$f_{k_n}^{(r)}(x) = f_{k_n}(x) - a(d, k_n, \alpha_x).$$

Estimation de $f_4(x) = 0.5\exp(|x|)$ en $x = 0$



Conclusion

- Condition nécessaire et suffisante pour la convergence en loi de f_{k_n}
- Problèmes :
  - Pas d'indice de régularité exact pour certaines fonctions ρ-régulières
  - Indice de régularité difficile à calculer
  - Définition inutilisable pour un rapport de mesures d'ensembles non centrés au point d'estimation x et n'étant pas forcément des boules

  $\rightarrow$ Nouvelle définition de l'indice de régularité

Nouvelle définition

On définit l'ensemble E_x par

$$E_x = \Big\{ \alpha \geq 0 \text{ tel que } \exists C > 0, \exists \lambda_0 > 0, \text{ tels que } \forall I \in \mathcal{I}_x \text{ vérifiant } \lambda(I) < \lambda_0 \text{ on ait } \left| \frac{\mu(I)}{\lambda(I)} - f(x) \right| \leq C\lambda(I)^\alpha \Big\}$$

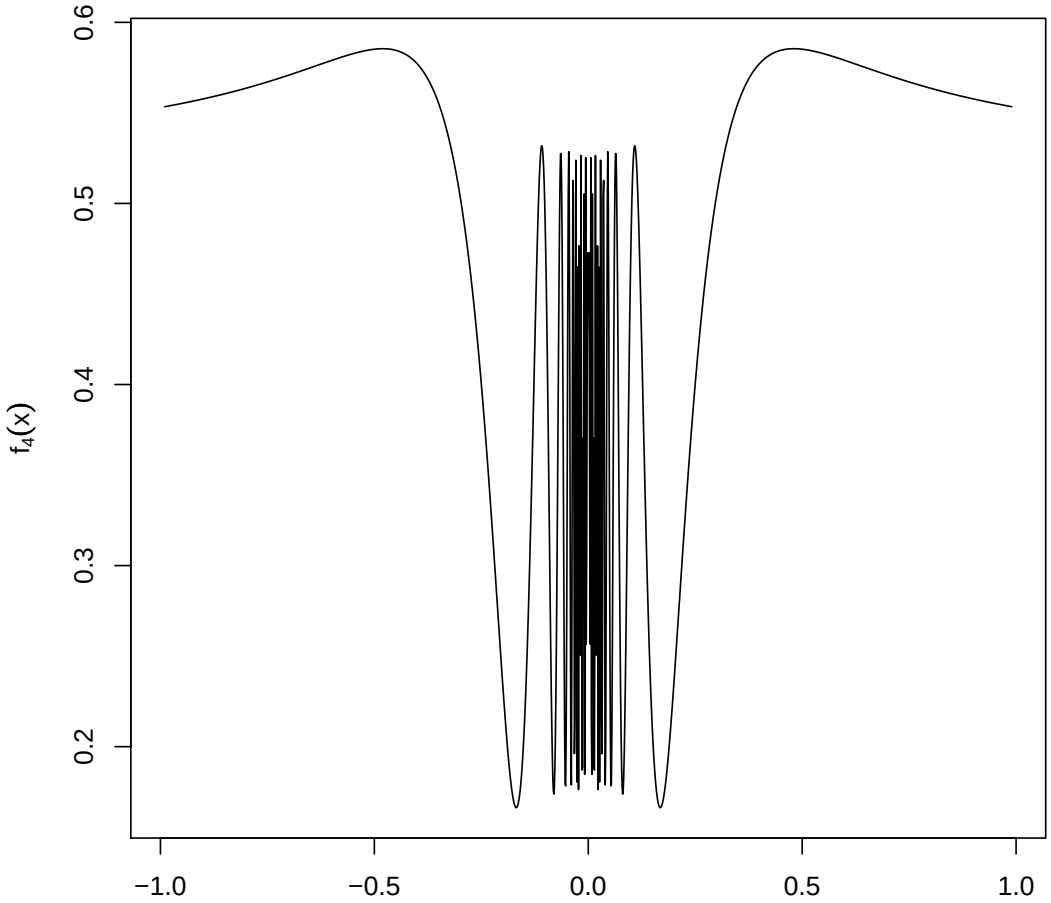
où $\mathcal{I}_x$ est l'ensemble des intervalles contenant x.
S'il existe un réel α_x vérifiant

$$\alpha_x = \sup E_x \tag{1}$$

alors α_x sera l'indice de régularité de la mesure μ au point x.
Si $\sup E_x = +\infty$ alors nous aurons $\alpha_x = +\infty$.

Exemple

$f_4(x) = \frac{2-\cos(1/x)+2x\sin(1/x)}{c}$ définie sur $\mathbb{R}$ pour $x \in [-1,1]\backslash 0$ avec $c = 4 + 2\sin 1$ et μ_1 sa mesure de probabilité associée.



En 0 :

- Point de Lebesgue
- ρ-régularité avec $\rho(\delta) = \delta/c$
- Pas d'indice de régularité exact car

$$\frac{\mu_4([-h,h])}{2h} - f_4(0) = \frac{1}{c} h \sin\left(\frac{1}{h}\right)$$

$\rightarrow$ Utilisation de la nouvelle définition : $\alpha_0 = 1$

Théorème

Si x est un point de Lebesgue où (1) est vérifié avec $f(x) > 0$ et $\alpha_x \in E_x$, et sous les conditions de convergence de f_{k_n}, la condition supplémentaire

$$\lim_{n\to\infty} \frac{k_n^{1+1/2\alpha_x}}{n} = 0$$

implique la convergence asymptotique de la variable aléatoire

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

vers une loi $\mathcal{N}(0,1)$.

Application à l'histogramme

L'histogramme s'écrit

$$f_h(x) = \frac{\nu_{nq}}{nh_n}$$

avec $x \in \Pi_{nq}$, ν_{nq} étant le nombre de points dans la $q$ème cellule, $q \in \mathbb{Z}$ et h_n un nombre réel positif dépendant de n.
Si

$$\lim_{n\to\infty} h_n = 0 \quad \text{et} \quad \lim_{n\to\infty} nh_n = +\infty. \tag{2}$$

alors

$$f_h(x) \xrightarrow{L^2} f(x).$$

Théorème

Si x est un point de Lebesgue où (1) est vérifié avec $f(x) > 0$ et $\alpha_x \in E_x$, et sous les conditions (2), la condition supplémentaire

$$\lim_{n\to\infty} nh_n^{2\alpha_x+1} = 0$$

implique la convergence asymptotique de la variable aléatoire

$$H_n(x) = \sqrt{nh_n}\frac{f_h(x) - f(x)}{\sqrt{f(x)}}$$

vers une loi $\mathcal{N}(0,1)$.