



HAL
open science

Necessary and sufficient condition for the existence of a limit distribution of the nearest neighbour density estimator

Alain Berlinet, Rémi Servien

► **To cite this version:**

Alain Berlinet, Rémi Servien. Necessary and sufficient condition for the existence of a limit distribution of the nearest neighbour density estimator. *Journal of Nonparametric Statistics*, 2011, 23 (3), pp.633-643. 10.1080/10485252.2011.567334 . hal-00730477

HAL Id: hal-00730477

<https://hal.science/hal-00730477>

Submitted on 10 Sep 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

RESEARCH ARTICLE

Necessary and sufficient condition for the existence of a limit distribution of the nearest neighbour density estimator

Alain F. Berline^t^{a*} and Rémi Servien^b

^a*I3M UMR CNRS 5149*

Université Montpellier II, France

^b*UMR 729 MISTEA -*

Campus INRA SupAgro, Montpellier

(v1.0 October 2, 2010)

In density estimation, a number of convergence results can be stated with no restriction on the function to be estimated. Unlike these universal properties, the asymptotic normality of estimators often requires hypotheses on the derivatives of the underlying density together with additional conditions on the smoothing parameter. Yet, in spite of possible bad local behaviour of the density (it is not continuous or has infinite derivative), the convergence in law of the nearest neighbour estimator can still occur and provide confidence bands for the estimated density. Therefore a natural question arises: Is it possible to give a necessary and sufficient condition for the existence of a limit distribution of the nearest neighbour estimator? We answer this question by means of the regularity index recently introduced by Beirlant, Berline and Biau (1). As expected the limit distribution, when existing, is gaussian. Its mean and variance are given explicitly in terms of coefficients related to the regularity index. The second order exponent in the expansion of the small ball probability is shown to be the crucial parameter. In contrast to former results on sufficiency of conditions for asymptotic normality no hypothesis is needed on the continuity of the underlying density.

Keywords: Asymptotic normality; nearest neighbour estimator; regularity index; small ball probability.

AMS Subject Classification: 62G05; 60E05

*Corresponding author. Email: berline@math.univ-montp2.fr

1. Introduction

Sufficient conditions for convergence and asymptotic normality of nearest neighbour density estimators have been given a long time ago. A classical result (Loftsgaarden and Quesenberry (8)) states the pointwise convergence in probability of this estimator at points where the density is positive and continuous. It can be extended to any Lebesgue point (Bosq and Lecoutre (3)). Moore and Yackel (10) proved the asymptotic normality under additional conditions on the smoothing parameter and on the derivatives of the density. Their results were extended to multivariate nearest neighbour estimators with weight function by Mack (9). It seems that no author reconsidered this question until the paper by Berlinet and Levallois (2) who proved the asymptotic normality of the nearest neighbour density estimator in cases where the density has bad local behaviour. Indeed what is important is the local behaviour of the associated measure, more precisely the rate at which the derivative of the underlying measure is approximated by ratios of ball measures. We establish in Section 2 a necessary and sufficient condition on the number of neighbours for the existence of a limit distribution of the nearest neighbour estimator. As expected, the limit distribution when existing is gaussian. Its parameters are given explicitly in terms of coefficients appearing in the expansion of small ball probabilities. In Section 3, some simulations are presented in a case where the density has no derivative.

2. Definitions and results

We consider a sequence $(X_n)_{n \geq 1}$ of d -dimensional independent random vectors with probability distribution μ and density f with respect to the Lebesgue measure λ on \mathbb{R}^d equipped with the Euclidean norm $\|\cdot\|$. We denote by $B(x, \delta)$ (resp. $\bar{B}(x, \delta)$) the open (resp. closed) ball with center at x and radius δ .

The problem of estimating the probability density has received considerable attention in the litterature. A lot of methods have been developped like histograms, kernel estimators and their variants. We shall consider in this paper the nearest neighbour estimator f_{k_n} of f based on $(X_n)_{n \geq 1}$. It is defined by

$$f_{k_n}(x) = \frac{k_n}{n\lambda(\bar{B}_n(x))}$$

where $\bar{B}_n(x) = \bar{B}(x, R_n(x))$ is the smallest closed ball with center x containing at least k_n sample points. The integer k_n plays the role of a smoothing parameter: when k_n is chosen too large the data are oversmoothed; they are undersmoothed in the opposite case. The estimate $f_{k_n}(x)$ is the ratio of the frequency of sample points falling into $\bar{B}_{k_n}(x)$ to the Lebesgue measure of $\bar{B}_{k_n}(x)$.

Berlinet and Levallois studied the asymptotic normality of this estimator and pointed out the key role of the local behaviour of the probability measure associated with the density, more precisely the role of the rate at which the derivative of the measure is approximated by the ratios of ball measures. If, for fixed x , the following limit

$$\ell(x) = \lim_{\delta \rightarrow 0} \frac{\mu(B(x, \delta))}{\lambda(B(x, \delta))} \quad (1)$$

exists, then x is called a Lebesgue point of the measure μ . This notion of Lebesgue

point is essential to state elegant results with few restrictions on the functions to be estimated. In Berlinet and Levallois (2), examples where the density has a bad local behaviour at Lebesgue points are examined. Yet, to evaluate rates of convergence or investigate asymptotic normality of estimators, we require not only the convergence of the ratio of ball measures but also information on higher order behaviour. In this context, Berlinet and Levallois define a ρ -regularity point of the measure μ as any Lebesgue point x satisfying

$$\left| \frac{\mu(B(x, \delta))}{\lambda(B(x, \delta))} - \ell(x) \right| \leq \rho(\delta), \quad (2)$$

where ρ is a measurable function such that $\lim_{\delta \downarrow 0} \rho(\delta) = 0$.

As it will be seen below, Berlinet and Levallois used this notion to extend previous asymptotic normality results of Moore and Yackel.

To specify an exact rate of convergence of the ratio of ball measures, Beirlant, Berlinet and Biau (1) assumed that a more precise relation than (2) holds at the Lebesgue point x , namely

$$\frac{\mu(B(x, \delta))}{\lambda(B(x, \delta))} = f(x) + C_x \delta^{\alpha_x} + o(\delta^{\alpha_x}) \text{ as } \delta \rightarrow 0, \quad (3)$$

where C_x is a non-zero constant and α_x is a positive real number. It is easy to show that relation (3) implies ρ -regularity at the point x with $\rho(\delta) = D_x \delta^{\alpha_x}$ and $D_x > C_x$. The constants C_x and α_x are uniquely determined (provided they exist). The index α_x is a regularity index which controls the degree of smoothness of the symmetric derivative of μ with respect to λ . The larger the value of α_x , the smoother the derivative of μ is at the point x . Beirlant, Berlinet and Biau used this regularity index to solve practical problems in nearest neighbour density estimation, such as removing bias or selecting the number of neighbours.

Equation (3) is clearly equivalent to the small ball probability expansion

$$P(\|X - x\| \leq \delta) = V_d \delta^d (f(x) + C_x \delta^{\alpha_x} + o(\delta^{\alpha_x}))$$

where X has density f and $V_d = \pi^{d/2} / \Gamma(1 + d/2)$ denotes the volume of the unit ball in \mathbb{R}^d . In other words, the second order term in the expansion of the small ball probability of radius δ at x is equal, up to a multiplicative constant, to $\delta^{d+\alpha_x}$.

2.1. Sufficient conditions of convergence

In discriminatory analysis, Fix and Hodges (6) introduced the classification rule based on nearest neighbour (see also Devroye, Györfi and Lugosi (4)). As to the nearest neighbour estimator, it was first studied by Loftsgaarden and Quesenberry (8) at points where the density is positive and continuous. These authors proved that the conditions

$$\lim_{n \rightarrow \infty} k_n = \infty \quad \text{and} \quad \lim_{n \rightarrow \infty} \frac{k_n}{n} = 0 \quad (4)$$

ensure the convergence in probability of $f_{k_n}(x)$ to $f(x)$. This result can be extended to any Lebesgue point (3).

Moore and Yackel proved the following asymptotic normality result.

Theorem 2.1: *Under the convergence conditions (4), if*

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

if f is continuous at x and has bounded derivatives in a neighbourhood of x and if $f(x) > 0$, then

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

tends in distribution to the standard normal $\mathcal{N}(0, 1)$ as n tends to infinity.

In practice, we have no information on the existence of derivatives. It is less restrictive to suppose ρ -regularity. Assuming ρ -regularity and a suitable rate of convergence for $\rho(R_n(x))$, Berlinet and Levallois obtained the following theorem.

Theorem 2.2: *Suppose that the convergence conditions (4) hold, that x is a ρ -regularity point of the measure μ and that $f(x) > 0$. If, as n tends to infinity,*

$$\sqrt{k_n} \rho(R_n(x)) \tag{5}$$

tends to 0 in probability then

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

tends in distribution to $\mathcal{N}(0, 1)$.

When the function f satisfies a Lipschitz condition of order $\alpha > 0$ at x , i.e.

$$|f(x) - f(y)| \leq C_x \|x - y\|^\alpha \tag{6}$$

for any y in some neighbourhood of x , it is possible to give a condition on the sequence (k_n) implying condition (5). This corollary of Theorem 2.2 was given in the univariate case by Berlinet and Levallois but a quick inspection of their proof easily yields to the d -dimensional version given in Corollary 2.3 below. This proof essentially relies on the fact that condition (6) implies ρ -regularity at the point x with

$$\rho(\delta) = D_x \delta^\alpha.$$

Hence Theorem 2.2 gives the result.

Corollary 2.3: *Suppose that the convergence conditions (4) hold, that f satisfies condition (6) and that $f(x) > 0$. If*

$$\lim_{n \rightarrow \infty} \frac{k_n^{1+d/(2\alpha)}}{n} = 0$$

then

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

tends in distribution to $\mathcal{N}(0, 1)$.

2.2. Main results

Now we consider the question of existence of a limit distribution for the nearest neighbour density estimator.

Theorem 2.4: *Suppose that the convergence conditions (4) hold, that x is a Lebesgue point in \mathbb{R}^d where (3) is satisfied with $f(x) > 0$. Then, as n tends to infinity,*

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

converges in distribution if and only if the sequence

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

has a finite limit κ . When this last condition is satisfied, the asymptotic law of $T_n(x)$ is

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

Theorem 2.4 provides a necessary and sufficient condition for the existence of a limit distribution and gives explicitly this distribution when it does exist.

2.3. Proofs

Proof of Theorem 2.4

Theorem 2.4 is a straightforward consequence of Lemma 2.7 and 2.8 below. □

The proof of the lemmas relies on the following decomposition:

$$T_n(x) = a_n(b_n + c_n)$$

where

$$\begin{aligned}
 a_n &= \frac{k_n}{n\mu(B_n(x))}, \\
 b_n &= \sqrt{k_n} \frac{1}{f(x)} \left(\frac{\mu(B_n(x))}{\lambda(B_n(x))} - f(x) \right) \\
 \text{and } c_n &= \frac{n}{\sqrt{k_n}} \left(\frac{k_n}{n} - \mu(B_n(x)) \right).
 \end{aligned}$$

Moore and Yackel (10) proved the following result.

Lemma 2.5: *For the sequences of random variables (a_n) and (c_n) defined above in the decomposition of $T_n(x)$, we have, under the convergence conditions (4),*

$$a_n \xrightarrow{P} 1 \quad \text{and} \quad c_n \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

as n tends to infinity.

Let us now turn to the asymptotic behaviour of the third quantity b_n introduced in the decomposition of $T_n(x)$.

Lemma 2.6: *Under the convergence conditions (4), and if (3) is satisfied, we have*

$$\lim_{n \rightarrow \infty} \left(\frac{n}{k_n} \right)^{\alpha_x/d} \frac{1}{\sqrt{k_n}} b_n = \frac{C_x}{V_d^{\alpha_x/d}} \left(\frac{1}{f(x)} \right)^{1+\alpha_x/d}.$$

Proof:

From (3) we obtain that the function

$$\Phi(\delta) = \left(\frac{\mu(B_\delta(x))}{\lambda(B_\delta(x))} - f(x) \right) \frac{1}{\delta^{\alpha_x}}$$

is such that

$$\lim_{\delta \rightarrow 0^+} \Phi(\delta) = C_x.$$

Using this function Φ in the expression of b_n , we have

$$b_n = \Phi(R_n) \left(\frac{nV_d R_n^d}{k_n} \right)^{\alpha_x/d} \frac{1}{V_d^{\alpha_x/d} f(x)} \frac{k_n^{\alpha_x/d+1/2}}{n^{\alpha_x/d}}$$

and

$$\left(\frac{n}{k_n} \right)^{\alpha_x/d} \frac{1}{\sqrt{k_n}} b_n = \Phi(R_n) \left(\frac{nV_d R_n^d}{k_n} \right)^{\alpha_x/d} \frac{1}{V_d^{\alpha_x/d} f(x)}.$$

Then, under the convergence conditions (4), we observe that

$$R_n \xrightarrow{P} 0 \quad \text{and} \quad \frac{k_n}{nV_d R_n^d} = f_{k_n}(x) \xrightarrow{P} f(x).$$

The convergence in probability of (R_n) to zero implies that

$$\Phi(R_n) \xrightarrow{P} C_x.$$

This, together with the convergence in probability of $f_{k_n}(x)$ to $f(x)$, gives the result. \square

Lemma 2.7: Sufficient condition.

Under the conditions of Theorem 2.4, if

$$\lim_{n \rightarrow \infty} \frac{k_n^{1+d/2\alpha_x}}{n} = \kappa$$

then

$$T_n(x) \xrightarrow{\mathcal{D}} \mathcal{N} \left(\frac{C_x \kappa^{\alpha_x/d}}{V_d^{\alpha_x/d}} \left(\frac{1}{f(x)} \right)^{1+\alpha_x/d}, 1 \right).$$

Proof: Conditions of Theorem 2.4 and Lemma 2.6 lead us to

$$\lim_{n \rightarrow \infty} b_n = \frac{C_x \kappa^{\alpha_x/d}}{V_d^{\alpha_x/d}} \left(\frac{1}{f(x)} \right)^{1+\alpha_x/d} \quad \text{in probability}$$

and Lemma 2.5 allows to conclude. \square

Lemma 2.8: Necessary condition.

Under the conditions of Theorem 2.4, if the sequence $\left(\frac{k_n^{1+d/(2\alpha_x)}}{n} \right)$ does not converge then the sequence of random variables

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

does not converge in distribution.

Proof: If the sequence $\left(\frac{k_n^{1+d/2\alpha_x}}{n} \right)$ is not bounded, then there exists a subsequence $\left(\frac{k_{q(n)}^{1+d/2\alpha_x}}{q(n)} \right)$ which tends to infinity. So

$$\forall M > 0, P(|b_{q(n)}| > M) \xrightarrow{n \rightarrow \infty} 1$$

and (b_n) , and consequently $(T_n(x))$, do not converge in distribution.

If the sequence $\left(\frac{k_n^{1+d/2\alpha_x}}{n} \right)$ is bounded and do not converge, then we can extract two subsequences which converge to two different limits l_1 and l_2 . Using Lemma 2.5, the

corresponding subsequences of $(T_n(x))$ converge in distribution respectively to $\mathcal{N}(l_1, 1)$ and $\mathcal{N}(l_2, 1)$. Therefore, (Z_n) does not converge in distribution. \square

3. An example

We illustrate the above considerations with the probability density f defined on $(-1, 1)$ by

$$f(x) = \frac{5}{2} \left(1 - |x|^{1/4}\right).$$

The associated distribution function satisfies

$$F(x) = \begin{cases} (1/2) + (5/2) x + 2 (-x)^{5/4} & \text{if } -1 \leq x \leq 0 \\ (1/2) + (5/2) x - 2 x^{5/4} & \text{if } 0 \leq x \leq 1. \end{cases}$$

Simple calculations show that for $0 < x - h < x + h < 1$, the small ball probability with radius h at x is given by

$$\begin{aligned} F(x+h) - F(x-h) &= 5h - 5hx^{1/4} + \frac{5}{32} \frac{h^3}{x^{7/4}} + O(h^5) \\ &= 2h (f(x) + C_x h^{\alpha_x} + o(h^{\alpha_x})) \end{aligned}$$

with

$$\begin{cases} C_x = (5/64) x^{-7/4} \\ \alpha_x = 2 \end{cases}$$

and that for $0 < h < 1$, one has

$$\begin{aligned} F(h) - F(-h) &= 5h - 4h^{5/4} \\ &= 2h (f(0) + C_0 h^{\alpha_0}) \end{aligned}$$

with

$$\begin{cases} C_0 = -2 \\ \alpha_0 = 1/4. \end{cases}$$

If k_n is chosen as $c n^\beta$ where c and β are two positive constants ($\beta < 1$), we know from Theorem 2.4 that for $x \neq 0$, $(T_n(x))$ converges in distribution if and only if $\beta \leq 4/5$. The asymptotic distribution is $\mathcal{N}\left(\frac{5^{-2} 2^{-1/4} c^{5/2}}{(2^{-2} 3^{7/4})^3}, 1\right)$ if $\beta = 4/5$ and $\mathcal{N}(0, 1)$ if $0 < \beta < 4/5$.

From the same theorem we know that $(T_n(0))$ converges in distribution if and only if $\beta \leq 1/3$. The asymptotic distribution is $\mathcal{N}(-4c^{3/4} 5^{-5/4}, 1)$ if $\beta = 1/3$ and $\mathcal{N}(0, 1)$ if $0 < \beta < 1/3$.

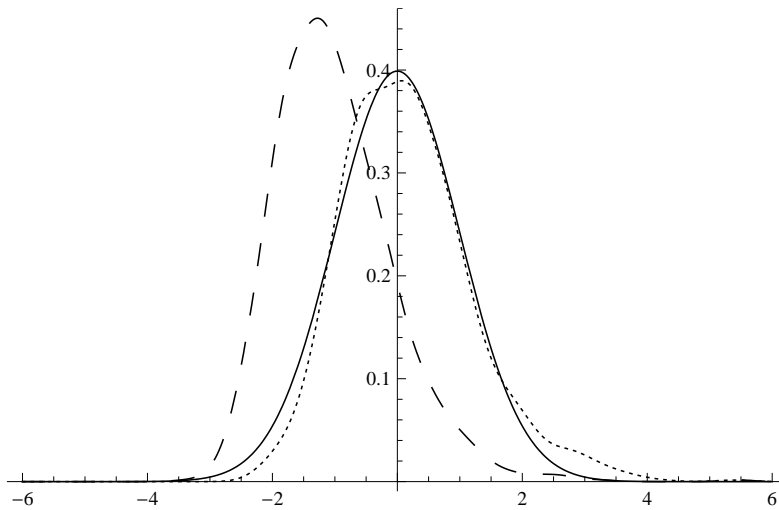


Figure 1. k_n equal to the integer part of $n^{1/2}$.

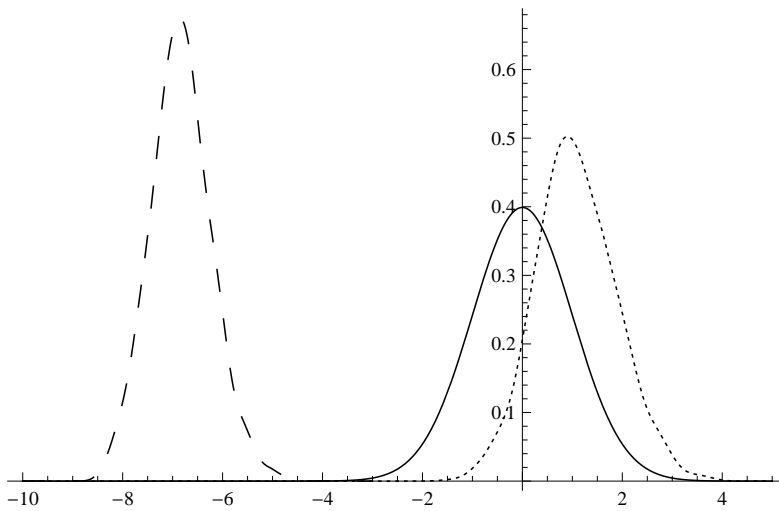


Figure 2. k_n equal to the integer part of $n^{4/5}$.

Since f is uniformly continuous the choice $k_n = c n^\beta, 0 < \beta < 1$, implies the strong uniform consistency conditions of Devroye and Wagner (5)

$$\lim_{n \rightarrow \infty} \frac{k_n}{\log n} = \infty \quad \text{and} \quad \lim_{n \rightarrow \infty} \frac{k_n}{n} = 0.$$

Note also that the optimal k_n in the sense of pointwise asymptotic mean square error at any point x of $(-1, 1)$ different from 0 is of the form $c n^{4/5}$ (see Fukunaga and Hostetler (7)) and that this imply the convergence of $T_n(x)$ to a non centered gaussian distribution.

A rule of thumb for k_n was early proposed by Loftsgaarden and Quesenberry (8). It consists in taking k_n as an integer close to \sqrt{n} . This rule ensures that the conditions of

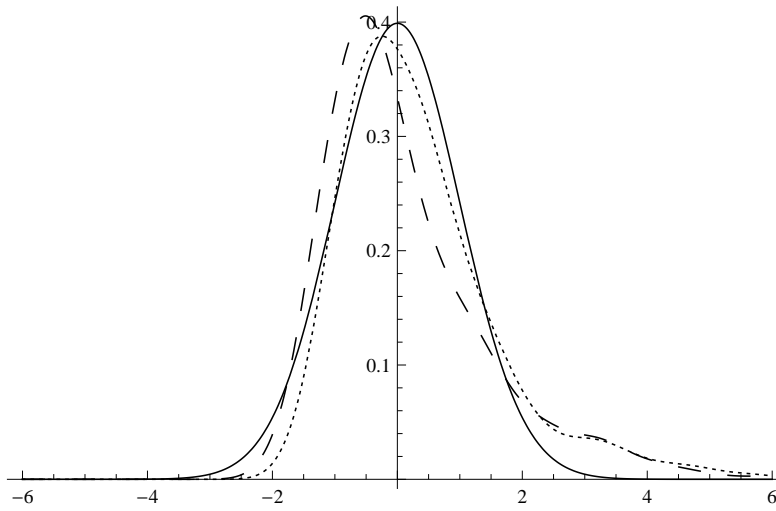


Figure 3. k_n equal to the integer part of $n^{1/3}$.

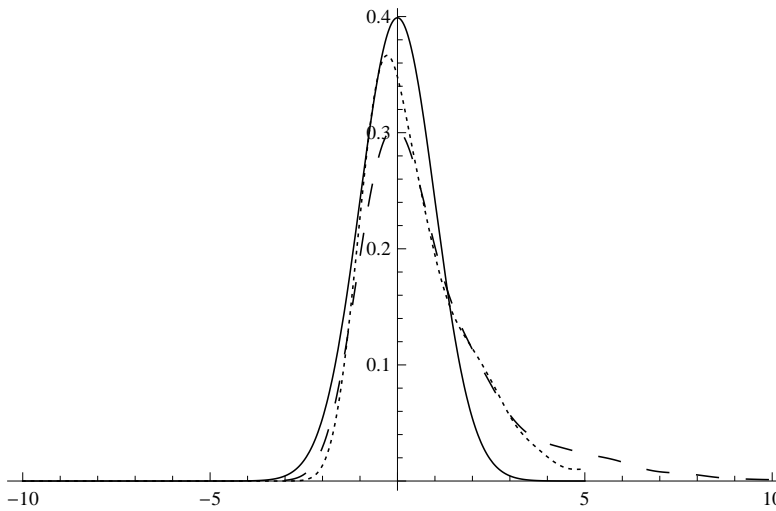


Figure 4. k_n equal to the integer part of $n^{1/4}$.

convergence

$$\lim_{n \rightarrow \infty} k_n = \infty \quad \text{and} \quad \lim_{n \rightarrow \infty} \frac{k_n}{n} = 0$$

are satisfied. It also ensures that the sufficient condition of Theorem 2.1

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

is fulfilled. Therefore the distribution of the statistic

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

tends to $\mathcal{N}(0, 1)$ as n tends to infinity whenever the conditions of Theorem 2.1 on f and its derivatives are satisfied. This is why the rule “ k_n close to \sqrt{n} ” is popular in nearest neighbour density estimation (Figure 1 described below was built with such k_n).

To illustrate the behaviour of the statistic $T_n(x)$ ($x = 0$ and $x = 0.5$) we simulated m times a set S_i ($1 \leq i \leq m$) of n pseudo-random variables with density f by means of the rejection method. Each set S_i of n variables was used to compute $f_{k_n}(x)$, k_n being equal to the integer part of n^γ , $\gamma \in \{\frac{1}{4}, \frac{1}{3}, \frac{1}{2}, \frac{4}{5}\}$. Then $T_n(x)$ was computed by using the exact value of $f(x)$. In this way we got m values of $T_n(x)$ and used them to estimate its distribution by a kernel estimate with gaussian kernel

$$K(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right)$$

and window-width

$$h_m = \left(\frac{4}{3m}\right)^{1/5} s_m$$

where s_m is the standard deviation of the sample (see Bosq and Lecoutre (3)).

The resulting eight estimators (with $n = m = 1000$) are drawn in Figures 1 to 4 together with the density of $\mathcal{N}(0, 1)$. The estimated density of $T_n(1/2)$ (resp. $T_n(0)$) is in dotted (resp. dashed) line and the density of $\mathcal{N}(0, 1)$ is in solid line. The choice of k_n implies the following situations.

Figure 1.

The distribution of $T_n(1/2)$ converges to $\mathcal{N}(0, 1)$.

The distribution of $T_n(0)$ does not converge.

Figure 2.

The distribution of $T_n(1/2)$ converges to $\mathcal{N}(1.043\dots, 1)$.

The distribution of $T_n(0)$ does not converge.

Figure 3.

The distribution of $T_n(1/2)$ converges to $\mathcal{N}(0, 1)$.

The distribution of $T_n(0)$ converges to $\mathcal{N}(-0.534\dots, 1)$.

Figure 4.

The distribution of $T_n(1/2)$ and the distribution of $T_n(0)$ converge to $\mathcal{N}(0, 1)$.

The lesson to draw for the practitioner is that the choice of k_n should be made locally with great care (this is known but deserves reminder) and that whenever the set of data is large enough a preliminary estimation of α_x , as proposed by Beirlant, Berlinet and Biau (1), may act as a safeguard against hasty conclusions.

References

- [1] J. Beirlant, A. Berlinet, and G. Biau. Higher order estimation at Lebesgue points. *Annals of the Institute of Statistical Mathematics*, 60:651–677, 2008.
- [2] A. Berlinet and S. Levallois. Higher order analysis at Lebesgue points. In M. Puri, editor, *G. G. Roussas Festschrift - Asymptotics in Statistics and Probability*, pages 1–16. 2000.
- [3] D. Bosq and J.-P. Lecoutre. *Théorie de l'Estimation Fonctionnelle*. Economica, 1987.
- [4] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New-York, 1996.
- [5] L. Devroye and T. J. Wagner. The strong uniform consistency of nearest neighbor density estimates. *The Annals of Statistics*, 5:536–540, 1977.
- [6] E. Fix and J. J. Hodges. Discriminatory analysis, nonparametric discrimination: consistency properties. *USAF School of Aviation Medicine*, 1951.
- [7] K. Fukunaga and L. Hostetler. Optimization of k -nearest-neighbor density estimates. *IEEE Trans. Information Theory*, IT-19:320–326, 1973.
- [8] D. Loftsgaarden and C. Quesenberry. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, pages 1049–1051, 1965.
- [9] Y. P. Mack. Asymptotic normality of multivariate k -NN density estimates. *Sankhyā*, 42:53–63, 1980.
- [10] D. Moore and J. Yackel. Large sample properties of nearest neighbor density function estimates. In S. Gupta and D. Moore, editors, *Statistical Decision Theory and Related Topics II*. Academic Press, New-York, 1977.