Deviation in kernel density estimation: super-optimal case

Corinne Berzin-Joseph

C. R. Acad. Sci. Paris, t. 330, Série I, p. 825–830, 2000
Statistique/*Statistics*
(Probabilités/*Probability Theory*)

Deviation in kernel density estimation: super-optimal case

Corinne BERZIN-JOSEPH

Labsad, BSHM, Université Pierre-Mendès-France, 1251, avenue Centrale, B.P. 47, 38040 Grenoble cedex, France
E-mail: Corinne.Joseph@upmf-grenoble.fr

(Reçu le 20 avril 1999, accepté après révision le 16 mars 2000)

Abstract. As in a previous Note [3] we study the asymptotic behaviour of several non-linear functionals of the empirical bridge in the super-optimal case. We consider the asymptotic behaviour of the number of crossings for the perturbed process in case the window satisfies $\sqrt{n}\,h^2 \to +\infty$; applications of the asymptotics are found. We also obtain a central limit theorem for the integrated square error of density estimators and in general for a G-deviation in density estimation and for the Kullback deviation in the super-optimal case. © 2000 Académie des sciences/Éditions scientifiques et médicales Elsevier SAS

Déviation de l'estimateur à noyau de la densité : cas sur-optimal

Résumé. *Nous étudions comme dans une Note précédente* [3] *le comportement asymptotique de certaines fonctionnelles non linéaires du pont empirique dans le cas sur-optimal. Nous considérons le comportement asymptotique du nombre de passages par un niveau pour le processus lissé perturbé dans le cas où la fenêtre d'observation est telle que* $\sqrt{n}\,h^2 \to +\infty$ *; des applications concernant l'asymptotique sont proposées. Nous obtenons aussi un théorème central-limite pour l'erreur quadratique intégrée de l'estimateur à noyau de la densité et de manière générale pour une G-déviation quelconque ainsi qu' une application à la convergence pour la déviation Kullback, dans le cas sur-optimal.* © 2000 Académie des sciences/Éditions scientifiques et médicales Elsevier SAS

Version française abrégée

Dans cette Note nous étudions comme dans [3] le comportement asymptotique de fonctionnelles non linéaires de régularisations du pont empirique. Dans [3], en utilisant un résultat montré dans [2], pour ce genre de fonctionnelles du pont brownien, et le théorème de Komlós, Major et Tusnády (KMT) [5], nous avons étudié le comportement asymptotique du nombre de passages par un niveau pour le processus empirique lissé perturbé ([8] à [13]) dans le cas où la fenêtre d'observation est telle que $\sqrt{n}\,h^2 \to a$ (a pouvant être nul). Notre but est de regarder le même genre de problème pour une fenêtre telle que

Note présentée par Paul DEHEUVELS.

S0764-4442(00)00271-8/FLA

$\sqrt{n}\,h^2 \to +\infty$, ce qui suppose une renormalisation du pont empirique lissé perturbé, et nous obtenons dans un premier temps, dans le théorème 1, des limites non triviales en utilisant le théorème 1 de [2].

Dans un deuxième temps, utilisant le théorème 2 de [2], nous nous intéressons aux vitesses avec lesquelles ont lieu ces convergences (*cf.* théorème 2 pour l'énoncé).

Enfin, en nous plaçant dans le cas sur-optimal au lieu du cas optimal (resp. sous-optimal), nous obtenons dans les théorèmes 3 et 4, en utilisant les résultats généraux de [2], un TCL pour la G-déviation de l'estimateur à noyau de la densité et déduisons comme corollaire un TCL pour sa déviation de Kullback.

1. Introduction and notation

1. Let $X_1, \ldots, X_n$ i.i.d. r.v. with distribution function F on $[0,1]$ ($F(0) = 0$, $F(1) = 1$) and $\mathrm{d}F(u) = s(u)\,\mathrm{d}u$ with $s \in \mathrm{C}^2$ on $[0,1]$.

Let φ be an even estimation kernel, F_n the empirical distribution of the sample and, $\widehat{s}_n$ the kernel density estimator defined by:

$$\widehat{s}_n(u) = \frac{1}{h}\int_{-\infty}^{\infty} \varphi\left(\frac{u-y}{h}\right) \mathrm{d}F_n(y);$$

$\widehat{F}_n$ will be the corresponding distribution function.

Define as in [8] to [13] the perturbed empirical process $\widetilde{\beta}_n^h(u) = \sqrt{n}\,(\widehat{F}_n(u) - F(u))$ (the bridge is said to be perturbed because we cut off $F(u)$ from $\widehat{F}_n(u)$ in place of $\mathrm{E}(\widehat{F}_n(u))$, $\widetilde{\beta}_n^h(u)$ having so a bias).

Yukich [13] proved that if $\sqrt{n}\,h^2 \to a$ (a can take the value 0), then $\widetilde{\beta}_n^h(u)$ converges in law to $b^F(u) + \frac{a}{2}\,K_2 s'(u)$, where $b^F(\cdot)$ is an F-Brownian motion bridge and $K_2 = \int_{-\infty}^{\infty} \varphi(v)\,v^2\,\mathrm{d}v$.

We consider the case $\sqrt{n}\,h^2 \to +\infty$. $\widetilde{\beta}_n^h(u)$ can be decomposed as

$$\widetilde{\beta}_n^h(u) = \beta_n^h(u) + \sqrt{n}\,\big(\widetilde{F}_n(u) - F(u)\big),$$

where $\beta_n^h(u)$ is the φ-regularization of size h of the empirical bridge $\beta_n^F(u) = \sqrt{n}\,(F_n(u) - F(u))$ and

$$\widetilde{F}_n(u) = \int_{-\infty}^{\infty} \varphi(v) F(u - hv)\,\mathrm{d}v.$$

With these notations $\frac{2}{K_2 h^2}(\widetilde{F}_n(u) - F(u)) \to s'(u)$.

So it is natural to consider $c_n\widetilde{\beta}_n^h(\cdot)$ with $c_n = 2/(K_2\sqrt{n}\,h^2)$ and to consider as in [3], the asymptotic behaviour of $\widetilde{N}^{c_n\widetilde{\beta}_n^h}_{[h,1-h]}(x)$ the number of times that the process $c_n\widetilde{\beta}_n^h(\cdot)$ crosses level x in the time interval $[h, 1-h]$.

On one hand, we propose non trivial limits. On the other hand we propose the speed at which the convergence takes place.

2. As in [3] we define $\sqrt{n}\,[\widehat{s}_n(u) - \mathrm{E}(\widehat{s}_n(u))] = (\beta_n^h)'(u)$.

It is well known that the optimal window (resp. sub-optimal) for the quadratic risk is $h(n) = \mathrm{O}(n^{-1/5})$ (resp. $h(n) = \mathrm{o}(n^{-1/5})$).

We are interested here in the super-optimal case (i.e., when $nh^5 \to +\infty$) and we show a central limit theorem for

$$\frac{\sqrt{n}}{K_2 h^2}\int_h^{1-h} \left\{\big(\widehat{s}_n(u) - s(u)\big)^2 - \mathrm{E}\big[\big(\widehat{s}_n(u) - s(u)\big)^2\big]\right\} \mathrm{d}u$$

getting thus a part of the results of P. Hall [6].

Finally we are interested in the "modified" Kullback deviation, i.e.:

$$\mathcal{K}(\widehat{s}_n, s) = \int_h^{1-h} \widehat{s}_n(u) \ln\left(\frac{\widehat{s}_n(u)}{s(u)}\right) \mathrm{d}u$$

(we are integrating between h and $1-h$ and not between 0 et 1 hence the expression "modified" Kullback deviation).

A TCL can also be obtained for

$$\frac{2\sqrt{n}}{K_2 h^2}\left\{\mathcal{K}(\widehat{s}_n, s) - \frac{1}{2}\int_h^{1-h} \frac{1}{s(u)} \mathrm{E}\big[\widehat{s}_n(u) - s(u)\big]^2 \mathrm{d}u - \int_h^{1-h} \big(\widehat{s}_n(u) - s(u)\big)\,\mathrm{d}u\right\}.$$

To finish, using our techniques we also consider the same type of problem for a G deviation and obtain a central limit theorem.

2. Hypotheses and notation

(H1) *Assume the kernel φ satisfies*: $\int_{-1}^{1} \varphi(t)\,\mathrm{d}t = 1$, $\varphi \in \mathrm{C}^1$, $\varphi \geqslant 0$ and even and has support a subset of $[-1,1]$. Define $\psi(u) = \varphi * \varphi(u)$ and $\theta(u) = \psi(u)\,\|\varphi\|_2^{-2}$, $u \in \mathbb{R}$. Let $K_2 = \int_{-\infty}^{\infty} \varphi(v)\,v^2\,\mathrm{d}v$.

(H2) *For the function f*: $f \in \mathrm{C}^2$ and f'' is bounded.

(H3) *For the function s*: $s \in \mathrm{C}^3$ on $[0,1]$, s is bounded below by $m > 0$ and above by M on $[0,1]$.

(H4) *For the function G*: $G \in \mathrm{C}^3$ and $\left|G^{(3)}(x)\right| \leqslant |P(x)|$, for $x \in \mathbb{R}$, where P is a polynomial.

Note that in the following hypotheses, (H2) and (H3) are not necessarily assumed.

Let $\{H_n,\ n \in \mathbb{N}\}$ be the Hermite polynomials, orthogonal with respect to the standard Gaussian measure ϕ and with leading coefficient equal to 1. We shall note, for $u \in [0,1]$,

$$\sigma_a^2(u) = \sum_{k=2}^{\infty} k!\, d_{k,a}^2(u) \int_{-2}^{2} \big[\theta(\omega)\big]^k \,\mathrm{d}\omega,$$

where $(d_{k,a}(u))_{k=0}^{\infty}$ are the Hermite coefficients of

$$\ell_a(x) = \sqrt{\frac{\pi}{2}}\left|x + \frac{aK_2 s''(u)}{2\sqrt{s(u)}\,\|\varphi\|_2}\right|$$

and $\theta(\omega)$ has been defined before in hypothesis (H1).

Also

$$\widetilde{\Sigma}_n^h(f) = \int_{-\infty}^{\infty} f(x)\,\widetilde{N}_{[h,1-h]}^{c_n\widetilde{\beta}_n}(x)\,\mathrm{d}x \quad \text{with } c_n = \frac{2}{K_2\sqrt{n}\,h^2},$$

where $\widetilde{N}_{[h,1-h]}^{c_n\widetilde{\beta}_n}(x)$ is the number of times that the process $c_n\widetilde{\beta}_n(\cdot)$ crosses level x in $[h,1-h]$ and f is a function, not necessarily a density.

We shall write $\sigma_2^2 = \int_0^1\int_0^1 F(u\wedge v)\,[1-F(u\vee v)]\,s^{(3)}(u)\,s^{(3)}(v)\,\mathrm{d}u\,\mathrm{d}v$, where $u\wedge v = \min(u,v)$ and $u\vee v = \max(u,v)$.

We also note

$$\sigma_G^2 = \int_0^1\int_0^1 F(u\wedge v)\big[1-F(u\vee v)\big]\,G''\big(s''(u)\big)\,G''\big(s''(v)\big)\,s^{(3)}(u)\,s^{(3)}(v)\,\mathrm{d}u\,\mathrm{d}v$$

and

$$\sigma_3^2 = \int_0^1 \int_0^1 F(u \wedge v)\big[1 - F(u \vee v)\big] \left(\frac{s''(u)}{s(u)}\right)' \left(\frac{s''(v)}{s(v)}\right)' \mathrm{d}u\,\mathrm{d}v;$$

N shall stand for a standard Gaussian r.v., $\equiv$ will denote equality in distribution and $\mathcal{D}$ (resp. $\mathcal{P}$) convergence in distribution (resp. in probability), E will take place for the expectation and $|\cdot|$ for the absolute value.

3. Perturbed empirical process

A function f is said to be locally Lipschitz, if

$$\big|f(x) - f(y)\big| \leqslant \big|Q(x,y)\big|\,|x - y|, \tag{1}$$

where Q is a polynomial.

THEOREM 1. – *Suppose that f satisfies* (1), $s \in \mathrm{C}^2$ *on* $[0,1]$ (*instead of* (H3)), φ *satisfies* (H1) *and* $h \to 0$, $nh^4 \to +\infty$. *Then*:
(a) *if* $nh^5 \to +\infty$,

$$\widetilde{\Sigma}_n^h(f) \xrightarrow{\mathcal{P}} \int_0^1 f\big(s'(u)\big)\,|s''(u)|\,\mathrm{d}u;$$

(b) *if* $nh^5 \to a^2$ (*a can take the value* 0 *and if it is the case we will suppose furthermore that* (H3) *holds*), *then*

$$\sqrt{nh^5}\,\widetilde{\Sigma}_n^h(f) \xrightarrow{\mathcal{D}} \frac{2}{K_2} \int_0^1 f\big(s'(u)\big)\,\mathrm{E}\left|N + \frac{aK_2 s''(u)}{2\sqrt{s(u)}\|\varphi\|_2}\right| \sqrt{s(u)}\,\|\varphi\|_2\,\mathrm{d}u.$$

Remark 1. – The motivation of the study of $\widetilde{\Sigma}_n^h(f)$ comes from the fact that, on one hand, if f tends to the Dirac delta distribution we then obtain the asymptotic behaviour of the number of crossings of the perturbed empirical process.

It is important to note that there exist other ways to apprehend the asymptotic behaviour of the perturbed empirical process and there is a quite substantial literature on the subject (*see* for example [8] to [12]).

On the other hand it is interesting to note the form of the limit in (a).

Indeed, $\int_0^1 f(s'(u))\,|s''(u)|\,\mathrm{d}u$ can be expressed thanks to the Banach–Kac formula ([1] and [7]) as $\int_{-\infty}^{+\infty} N_{[0,1]}^{s'}(x)\,f(x)\,\mathrm{d}x$ and if the function f is near from the Dirac delta distribution in zero then we obtain an estimator for the number of crossings of 0 by the function s' on $[0,1]$ and then for the number of local extrema of the function s on $[0,1]$.

Furthermore, if we work with the up-crossings (resp. the down-crossings) of $c_n\,\widetilde{\beta}_n(\cdot)$ we obtain an estimator for the number of local maxima (resp. minima) of the function s.

Proof. – The proof follows from Theorem 1 of [2].

Moreover, we have:

THEOREM 2. – *Assume that f verifies* (1), $s \in \mathrm{C}^3$ *on* $[0,1]$, (H1), $h \to 0$ *and* $nh^4 \to +\infty$, *then*:
(a) *if* $nh^6 \to +\infty$ *and if* $|s''|$ *is bounded below on* $[0,1]$, *then*

$$c_n^{-1}\left[\widetilde{\Sigma}_n^h(f) - \int_h^{1-h} f\big(c_n\widetilde{\beta}_n^h(u)\big)\left|c_n\sqrt{n}\,\big(\widetilde{F}_n'(u) - F'(u)\big)\right|\mathrm{d}u\right] \xrightarrow{\mathcal{D}} \int_0^1 f\big(s'(u)\big)\,\mathrm{sign}\big(s''(u)\big)\,\mathrm{d}b^F(u);$$

(b) *if* $(nh^5 - a^2)/\sqrt{h} \to 0$ (*a can take the value zero*), *under* (H2) *and* (H3),

$$\frac{1}{\sqrt{h}}\left[\sqrt{\frac{\pi}{2}}\|\varphi\|_2^{-1}\sqrt{nh^5}\,\widetilde{\Sigma}_n^h(f)-\frac{2}{K_2}\sqrt{\frac{\pi}{2}}\int_h^{1-h} f\big(c_n\widetilde{\beta}_n^h(u)\big)\,\mathrm{E}\left|N+\frac{aK_2s''(u)}{2\sqrt{s(u)}\|\varphi\|_2}\right|\sqrt{s(u)}\,\mathrm{d}u\right]$$
$$\xrightarrow{\mathcal{D}}-\frac{2}{K_2}\sqrt{\frac{\pi}{2}}\|\varphi\|_2^{-1}\int_0^1 b^F(u)\left(f\big(s'(u)\big)\,\mathrm{E}\left[N\left|N+\frac{aK_2s''(u)}{2\sqrt{s(u)}\|\varphi\|_2}\right|\right]\right)'\mathrm{d}u$$
$$+\frac{2}{K_2}\int_0^1\sigma_a(u)f\big(s'(u)\big)\sqrt{s(u)}\,\mathrm{d}\widetilde{W}(u), \tag{2}$$

where $\widetilde{W}(\cdot)$ is a Brownian motion independent of $b^F(\cdot)$.

Remark 2. – In the special case where $a=0$, note that $\sigma_0^2(u)$ does not depend on u.

Proof. – Theorem 2 is a consequence of the KMT theorem (replacing $(\beta_n^h)'(\cdot)$ by the derivative of the F-Browian bridge's φ-regularization of size h) and Theorem 2 of [2].

4. The integrated square-error of density estimators

Let

$$D_n=\frac{\sqrt{n}}{K_2h^2}\int_h^{1-h}\Big\{\big(\widehat{s}_n(u)-s(u)\big)^2-\mathrm{E}\left[\big(\widehat{s}_n(u)-s(u)\big)^2\right]\Big\}\,\mathrm{d}u.$$

We are interested in the asymptotic behaviour of the second order deviation D_n in the super-optimal case (i.e., when $nh^5\to+\infty$) studied by P. Hall [6] for a multivariate density. Later Csörgő & Horváth [4] have studied the case of a p-deviation in the optimal (resp. sub-optimal) case, i.e., $nh^5\to a>0$ (resp. $nh^5\to 0$). Our approach includes the case of a G-deviation in the super-optimal case (see the following section).

THEOREM 3. – *Under* (H1) *and* (H3), *if* $h\to 0$, $nh^5\to+\infty$, *then*

$$D_n\xrightarrow{\mathcal{D}}-\int_0^1 b^F(u)\,s^{(3)}(u)\,\mathrm{d}u\equiv N\big(0,\sigma_2^2\big).$$

Remark 3. – The normalization is the same as in P. Hall [6] where $d(n)=\sqrt{n}/h^2$.

Proof. – We prove (thanks to Theorem 2 of [2] and the KMT theorem) that D_n is asymptotically equivalent to $\int_0^1 s''(u)\,\mathrm{d}b^F(u)$ and then converges to a Normal.

4.1. Kullback deviation

Consider the "modified" Kullback deviation between the kernel estimator of the density $\widehat{s}_n$ and the true density s verifying (H3). Let

$$\mathcal{K}\big(\widehat{s}_n,s\big)=\int_h^{1-h}\widehat{s}_n(u)\ln\left(\frac{\widehat{s}_n(u)}{s(u)}\right)\mathrm{d}u.$$

As in [3] we can show that this definition makes sense under the condition $h\to 0$, $nh^{1+a}\to+\infty$ for some $a>0$ and then when $nh^5\to+\infty$.

COROLLARY 1. – *Under* (H1), (H3) *and the asumptions,* $h\to 0$, $nh^5\to+\infty$ *and* $nh^8\to 0$,

$$K_n=\frac{2\sqrt{n}}{K_2h^2}\left\{\mathcal{K}\big(\widehat{s}_n,s\big)-\frac{1}{2}\int_h^{1-h}\frac{1}{s(u)}\,\mathrm{E}\big[\widehat{s}_n(u)-s(u)\big]^2\,\mathrm{d}u-\int_h^{1-h}\big(\widehat{s}_n(u)-s(u)\big)\,\mathrm{d}u\right\}$$
$$\xrightarrow{\mathcal{D}}-\int_0^1 b^F(u)\left[\frac{s''(u)}{s(u)}\right]'\mathrm{d}u\equiv N\big(0,\sigma_3^2\big).$$

Proof. – We prove that K_n is asymptotically equivalent to

$$\frac{\sqrt{n}}{K_2 h^2}\int_h^{1-h}\frac{1}{s(u)}\left\{\left(\widehat{s}_n(u)-s(u)\right)^2-\mathrm{E}\left[\left(\widehat{s}_n(u)-s(u)\right)^2\right]\right\}\mathrm{d}u$$

and then by Theorem 3 converges to a Normal.

5. G-deviation for density estimation

Let

$$D_{n,G}=\sqrt{n}\,h^2\frac{K_2}{2}\int_h^{1-h}\left\{G\big(c_n\sqrt{n}\big(\widehat{s}_n(u)-s(u)\big)\big)-\mathrm{E}\big[G\big(c_n\sqrt{n}\big(\widehat{s}_n(u)-s(u)\big)\big)\big]\right\}\mathrm{d}u.$$

THEOREM 4. – *Under* (H1), (H3) *and* (H4), *if* $h\to 0$ *and* $\sqrt{h}\,nh^5\to+\infty$, *then*

$$D_{n,G}\xrightarrow{\mathcal{D}}-\int_0^1 b^F(u)\,G''\big(s''(u)\big)\,s^{(3)}(u)\,\mathrm{d}u\;\equiv\;N\big(0,\sigma_G^2\big).$$

Remark 4. – In the particular case where $G(x)=x^2$, Theorem 4 is true under the less restrictive condition $nh^5\to+\infty$ and more generally for G a polynomial. Indeed, if G is polynomial with degree m it is enough to use a Taylor's development of order m for G in a neighbourhood of $c_n\sqrt{n}\left[\mathrm{E}(\widehat{s}_n(u))-s(u)\right]$, obtaining thus exactly a development we prove that its second term gives the required limit and that the other terms tend to zero by the KMT theorem and Theorem 2 in [2].

References

[1] Banach S., Sur les lignes rectifiables et les surfaces dont l'aire est finie, Fund. Math. 7 (1925) 225–237.
[2] Berzin-Joseph C., León J., Ortega J., Increments and crossings for the Brownian bridge: weak convergence, C. R. Acad. Sci. Paris., Série I 327 (1998) 587–592.
[3] Berzin-Joseph C., León J., Ortega J., Study of the asymptotic behaviour of non-linear functionals for the empirical bridge via strong approximations, C. R. Acad. Sci. Paris., Série I 327 (1998) 671–676.
[4] Csörgő M., Horváth L., Central limit theorems for L_p-norms of density estimators, Probab. Th. Rel. Fields. (1988) 269–291.
[5] Csörgő M., Révész P., Strong Approximations in Probability and Statistics, Academic Press, New York.
[6] Hall P., Central limit theorem for integrated square error of multivariate nonparametric density estimators, J. Multiv. Anal. 14 (1984) 1–16.
[7] Kac M., On the average number of real roots of a random algebraic equation, Bull. Amer. Math. Soc. 49 (1943) 314–320.
[8] Ralescu S.S., Asymptotic deviations between perturbed empirical and quantile processes, J. Statis. Plan. Inf. 32 (1992) 243–258.
[9] Ralescu S.S., Sun S., Necessary and sufficient conditions for the asymptotic normality of perturbed sample quantiles, J. Statis. Plan. Inf. 35 (1993) 55–64.
[10] Ralescu S.S., The asymptotic law of the local oscillation modulus of the empirical process, J. Statis. Plan. Inf. 35 (1993) 139–156.
[11] Ralescu S.S., A law of the iterated logarithm for perturbed kernel quantiles near the origin, in: Res. Dev. Probab. Statis., Brunner E., Denker M. (Eds.), 1996, pp. 45–57.
[12] Ralescu S.S., A Bahadur–Kiefer law for the Nadaraya empiric-quantile processes, Th. Probab. Appl. 41 (1997) 296–306.
[13] Yukich J.E., A note on limit theorems for perturbed empirical processes, Stoch. Proc. Appl. 33 (1989) 163–173.