Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
C. R. Acad. Sci. Paris, Ser. I 338 (2004) 311–316
COMPTES RENDUS MATHEMATIQUE

Statistique/Probabilités

Processus de Bickel–Rosenblatt pondéré et tests d'ajustement

Fateh Chebana

LSTA, boîte Courrier 158, 8A, Université Paris-6, 175, rue du Chevaleret, 75013 Paris, France

Reçu le 9 janvier 2003 ; accepté après révision le 12 novembre 2003

Présenté par Paul Deheuvels

Résumé

Le but de cette étude est de trouver la loi limite du processus

$$\left\{ I_n(W) := \int_A \big(f_n(x) - Ef_n(x)\big)^2 W(x)\,\mathrm{d}x,\ W \in \mathcal{W} \right\},$$

où $\mathcal{W}$ est une classe de fonctions de poids W, f_n est l'estimateur à noyau de la densité f et A est un sous-ensemble borélien de $\mathbb{R}$. On utilise ce résultat pour construire de nouveaux tests d'ajustement de la densité f, plus performants, sous certaines alternatives locales, que le test classique de Bickel–Rosenblatt. ***Pour citer cet article : F. Chebana, C. R. Acad. Sci. Paris, Ser. I 338 (2004).***

Abstract

Weighted Bickel–Rosenblatt process and goodness of fit tests. The goal of this work is to establish the limit distribution of the process

$$\left\{ I_n(W) := \int_A \big(f_n(x) - Ef_n(x)\big)^2 W(x)\,\mathrm{d}x,\ W \in \mathcal{W} \right\},$$

where $\mathcal{W}$ is a class of weight functions W, f_n is the kernel density estimator of the density f and A is a Borelian subset of $\mathbb{R}$. We apply this result to derive new statistics to test goodness-of-fit of the density function f. Under some local alternatives, these new tests are more powerful than the usual Bickel–Rosenblatt one. ***To cite this article: F. Chebana, C. R. Acad. Sci. Paris, Ser. I 338 (2004).***

Abridged English version

Let $X_1, X_2, \ldots,$ be a sequence of independent identically distributed real random variables with density f. Denote by $f_n(x) := (nh_n)^{-1} \sum_{i=1}^n K((x - X_i)/h_n)$ the kernel estimator of $f(x)$. We prove, under suitable

Adresse e-mail : chebana@ccr.jussieu.fr (F. Chebana).

1631-073X/$ – see front matter
doi:10.1016/j.crma.2003.11.031

hypotheses, that $\{nh_n^{1/2}(I_n(W) - EI_n(W)),\ W \in \mathcal{W}\}$ converges weakly to a Gaussian process, where $I_n(W) := \int_A (f_n(x) - Ef_n(x))^2 W(x)\,\mathrm{d}x$, $\mathcal{W}$ is a class of weight functions and A is a Borelian subset of $\mathbb{R}$.

We apply this result to construct new test statistics based on functionals of the process $\{I_n(W), W \in \mathcal{W}\}$. To test, at a given level α, the hypothesis H_0: $f = f_0$ versus the Pitman local alternatives H_n: $f(x) = f_0(x) + n^{-1/2}h_n^{-1/4}(\eta(x) + \mathrm{o}(1))$, where $f_0 = \frac{1}{2}I_{[-1,1]}$, we use an adapted version of our result; we obtain both critical region and asymptotic power of the new functional statistics.

Let $T_n(s) := h_n^{-1/2}(nh_n\tilde{I}_n(s) - \int_{-s}^{s} f_0(x)\,\mathrm{d}x \int K^2(x)\,\mathrm{d}x)$, where $\tilde{I}_n(s) = \int_{-s}^{s}(f_n(x) - f_0(x))^2\,\mathrm{d}x$. We propose here the statistic $S_n := \frac{1}{a}\int_0^a T_n(u)\,\mathrm{d}u$. We compare S_n and $T_n(a)$, $0 < a < 1$. The statistic S_n converges weakly, both under H_0 and H_n, to a Gaussian distribution.

We show that S_n is more powerful than $T_n(a)$, whenever the function η satisfies $a(\sqrt{3} - 1)\int_{-a}^{a}\eta^2(x)\,\mathrm{d}x \geqslant \sqrt{3}\int_{-a}^{a}|x|\eta^2(x)\,\mathrm{d}x$. A simple example is to take η as the identity function on the interval $[-a/2, a/2]$.

1. Introduction

Soit $X_1, X_2, \ldots,$ une suite de variables aléatoires réelles (v.a.r.) indépendantes et identiquement distribuées (iid.) de densité f. Pour tout $x \in \mathbb{R}$, désignons par

$$f_n(x) := \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)$$

l'estimateur à noyau de la densité $f(x)$, où $(h_n)_{n\in\mathbb{N}}$ est une suite de nombres réels positifs tendant vers zéro quand n tend vers l'infini et K est un noyau sur $\mathbb{R}$.

Une manière de mesurer la distance entre l'estimateur f_n et son espérance Ef_n consiste à prendre la norme L_2 pondérée donnée par

$$I_n(W) := \int_A \big(f_n(x) - Ef_n(x)\big)^2 W(x)\,\mathrm{d}x, \tag{1}$$

où W est une fonction de poids et A est un borélien de $\mathbb{R}$. Sous certaines conditions sur h_n, l'estimateur f_n est convergent dans L_2, voir [6]. Dans le cas iid., la normalité asymptotique de $I_n(W)$ a été étudiée par Bickel et Rosenblatt [1], et quand $W \equiv 1$ par Hall [5]. Tenreiro [8] a considéré des variables X_i dépendantes, sous certaines conditions de mélange (voir aussi les réferences ci-incluses). Tenreiro [8] a introduit une version du MISE (Mean Integrate Square Error) évaluée sur un sous-ensemble. Constatons que dans toutes les études précédentes la fonction de poids W est fixée. Le comportement fonctionnel du processus $J_n(K) := \int |f_{n,K}(x) - Ef_n(x)|\,\mathrm{d}x$, indexé par des noyaux K, a été envisagé récemment par Giné et al. [3]. Les statistiques $I_n(W)$, avec une fonction de poids fixée W, ont été proposées en premier par Bickel et Rosenblatt [1] comme des tests d'ajustement basés sur un estimateur de la densité (*BR* test). Partons de la même idée, on voudrait connaître le comportement asymptotique de l'ensemble de ces statistiques dans le but de proposer des tests fondés sur des fonctionnelles du processus $\{I_n(W),\ W \in \mathcal{W}\}$, où $\mathcal{W}$ est une classe de fonctions.

L'objectif de cette Note est donc d'étudier le comportement asymptotique du processus $I_n(\cdot)$ indexé par une classe de fonctions de poids sous une hypothèse d'entropie sur la classe $\mathcal{W}$. La preuve est divisée en deux parties. La première consiste à trouver la distribution limite des vecteurs fini-dimensionnels $I_n(W_k),\ k = 1, \ldots, m$, pour tout entier m positif. La démonstration est inspirée de [5]. La seconde est de prouver la tension du processus $\{I_n(W), \mathcal{W}\}$. Dans ce but, on adapte la démonstration de Giné et al. [3]. Le processus limite obtenu est un processus gaussien centré dont la structure de covariance est explicitement donnée. Enfin, on propose comme application de notre résultat une nouvelle statistique pour tester l'ajustement d'une densité. On présente ses performances sous des alternatives locales. Pour une certaine classe d'alternatives locales, on montre que notre test statistique est plus puissant que celui de Bickel–Rosenblatt.

2. Résultats

Dans cette section nous présentons nos résultats. Le premier porte sur les lois fini-dimensionnelles tandis que le second prouve la tension du processus $I_n(\cdot)$. Finalement, on démontre la convergence faible du processus $I_n(\cdot)$. Pour des raisons statistiques, on établit également la loi limite du processus

$$\left\{\tilde{I}_n(W) := \int_A \big(f_n(x) - f(x)\big)^2 W(x)\,\mathrm{d}x,\ W \in \mathcal{W}\right\},$$

en imposant quelques hypothèses de régularité.

Avant de présenter nos résultats, nous introduisons les hypothèses et les notations suivantes.

2.1. Hypothèses et notations

(K1) Le noyau K est positif et borné, (K2) $\int K(z)\,\mathrm{d}z = 1$, (K3) K est à support borné.
(D) La densité f est bornée et uniformément continue sur l'ensemble A^ϵ, l'ensemble dilaté de A, pour ϵ positif.
(W1) Les fonctions de poids W sont bornées sur A^ϵ et continues presque partout (p.p.) sur l'intérieur de A.
(W2) La classe $\mathcal{W}$ est uniformément bornée.
(H) La suite h_n est telle que : $h_n \to 0$ et $nh_n \to \infty$ quand $n \to \infty$.

Notons par Σ la $m \times m$-matrice dont les éléments sont

$$\sigma(k,l) = 2\int_A W_k(x)W_l(x)f^2(x)\,\mathrm{d}x \int\left[\int K(s)K(z+s)\,\mathrm{d}s\right]^2 \mathrm{d}z, \quad k,l = 1,\ldots,m.$$

Soit $G_p(\cdot)$ un processus gaussien centré de matrice de covariance Σ. Dans la proposition suivante, nous donnons les lois de dimensions finies du processus $I_n(\cdot)$.

Proposition 2.1. *Sous les hypothèses* (K1), (K2), (D), (W1) *et* (H), *on a*

$$nh_n^{1/2}\big\{I_n(W_k) - EI_n(W_k),\ k = 1,\ldots,m\big\} \xrightarrow{\mathcal{D}} \mathcal{N}_m(0,\Sigma).$$

Le résultat suivant implique l'équicontinuité du processus $I_n(\cdot)$, et par conséquent sa tension (voir [10]). Pour éviter des problèmes de mesurabilité, on utilise la mesure extérieure P^* quand cela est nécessaire (voir [10], p. 37). La norme d'Orlicz $\|\cdot\|_\psi^*$ et les nombres de recouvrement $\mathcal{N}(\mathcal{W}, \|\cdot\|_2, \epsilon)$ sont définis dans [10], p. 83. Dans la proposition suivante, on considère pour tout x positif, la fonction $\psi_{1/2}$ donnée par

$$\psi_{1/2}(x) := \frac{\exp(\sqrt{x+1}) - \mathrm{e}}{2\mathrm{e}}.$$

Proposition 2.2. *Supposons vérifiées les hypothèses* (K1)–(K3), (D), (H) *et* (W2). *De plus, l'ensemble A défini dans* (1) *est supposé borné. Supposons également que*

$$\int_0^\infty \big(\log\mathcal{N}\big(\mathcal{W}, \|\cdot\|_2, \epsilon\big)\big)^2\,\mathrm{d}\epsilon < \infty. \tag{2}$$

On obtient alors

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\Big\|\sup_{\|W\|_2\leqslant\delta,\ W\in\mathcal{W}} nh_n^{1/2}\big|I_n(W) - EI_n(W)\big|\Big\|_{\psi_{1/2}}^* = 0.$$

Notre résultat principal est donné dans le théorème qui suit. Il est fondé sur les deux propositions précédentes. Il donne la loi limite du processus $I_n(\cdot)$.

Théorème 2.3. *Sous les hypothèses des Propositions* 2.1 *et* 2.2, *on a*

$$\{nh_n^{1/2}(I_n(W) - EI_n(W)),\ W \in \mathcal{W}\} \xrightarrow{\mathcal{D}} \{G_P(W),\ W \in \mathcal{W}\},$$

quand $n \to \infty$, *où* $\xrightarrow{\mathcal{D}}$ *désigne la convergence faible dans l'espace* $\ell^\infty(\mathcal{W})$ *de toutes les fonctions bornées à valeurs réelles définies sur la classe* $\mathcal{W}$. *De plus le processus limite ainsi obtenu est tendu et ses trajectoires sont* $\|\cdot\|_2$*-uniformément continues.*

Pour des raisons statistiques, il est préférable d'avoir le même résultat pour le processus MISE $\tilde{I}_n(\cdot)$. Pour cela, on impose en plus les hypothèses de régularité suivantes :

(K′) $\int zK(z)\,dz = 0$ et (K″) $\int z^2K(z)\,dz < \infty$.
(D′) La densité f est deux fois dérivable et sa dérivée seconde f'' est bornée et uniformément continue sur A^ϵ.

Théorème 2.4. *Sous les hypothèses du Théorème* 2.3, (K′), (K″) *et* (D′), *et l'hypothèse additive* $nh_n^5 \to 0$, *on obtient les mêmes conclusions que celles du Théorème* 2.3 *pour le processus*

$$\{nh_n^{1/2}(\tilde{I}_n(W) - E\tilde{I}_n(W)),\ W \in \mathcal{W}\}.$$

2.2. Remarques

(i) Les trajectoires du processus I_n appartiennent à $\ell^\infty(\mathcal{W})$.
(ii) Les hypothèses (K1), (K2), (K′) et (K″) sont standards dans la littérature de l'estimation fonctionnelle, voir par exemple [5] et [1]. Pour (K3) Ghosh et Huang [4] ont montré que le choix optimal du noyau pour les tests d'ajustement du type BR est la densité uniforme sur un intervalle borné. Les hypothèses (D), (D′) et (W1) ont été introduites dans [8].
(iii) Si f est une densité uniforme sur l'intervalle $[0, 1]$, le processus G_P donné dans le Théorème 2.3 est le processus P-mouvement Brownien sur $\mathcal{W}$, à une constante près qui dépend de K ; voir Définition 12 dans [7], p. 147.
(iv) Dans cette Note, on a considéré uniquement le cas $nh_n^5 \to 0$. Pour les autres cas (a) $nh_n^5 \to \lambda,\ 0 < \lambda < \infty$, et (b) $nh_n^5 \to \infty$, la Proposition 2.2 est valable avec des modifications au niveau des suites normalisantes, $n^{9/10}$ dans le cas (a), et $\sqrt{n}h_n^{-2}$ dans le cas (b). Toutefois le cas $nh_n^5 \to 0$ est statistiquement recommandé dans certains cas ; voir [2], p. 329.
(v) Les Théorèmes 2.3 et 2.4 pourraient être adaptés au cas multi-dimensionnel, sous des modifications semblables à celles que l'on peut trouver dans [5] qui traite la loi limite de $I_n(1)$.

2.3. Exemples

(i) La classe $\mathcal{W} = \{I_{[-t,t]}, t \in [0, T]\}$ satisfait les hypothèses du Théorème 2.3, puisque son nombre de recouvrement $\mathcal{N}(\mathcal{W}, \|\cdot\|_2, \varepsilon)$ est plus petit que $2\varepsilon^{-2}$ (voir [7], p. 146).
(ii) La classe $\mathcal{W} = \{W(\cdot\lambda)\colon \lambda \geqslant 1\}$, où W est p.p. continue, bornée et à variation bornée et de support $[-1/2, 1/2]$. Le nombre de recouvrement de $\mathcal{W}$ est tel que : $\mathcal{N}(\mathcal{W}, \|\cdot\|_2, \varepsilon) \leqslant C\varepsilon^{-2} + 1$ pour tout $\varepsilon \leqslant C$. Cette classe vérifie donc la condition (2) ; c'est l'Exemple 1.2 dans [3].

3. Application

La loi limite du processus I_n (respectivement de $\tilde{I}_n$) vient du Théorème 2.3 (respectivement du Théorème 2.4). On peut utiliser la version fonctionnelle de la méthode delta (voir [9], Chapitres 3 et 20) ou plus généralement le lemme de Slutsky, dans le but d'obtenir des propriétés asymptotiques des statistiques basées sur I_n (respectivement sur $\tilde{I}_n$). Cependant, il n'est pas toujours facile d'obtenir des formules explicites des fonctionnelles d'un processus gaussien. Dans le cas des transformations linéaires, on trouve une distribution gaussienne, dont on sait calculer les paramètres (voir Lemme 3.9.8 dans [10], p. 377). En utilisant ce lemme, on donne l'application explicite suivante :

Cherchons à tester au niveau α donné l'hypothèse $H_0 : f = f_0$ contre la suite des alternatives locales $H_n : f = g_n$, où $g_n(x) = f_0(x) + \alpha_n(\eta(x) + o(1))$. La fonction η est p.p. continue, $\|\eta\|_\infty < \infty$, intégrable et $\alpha_n := n^{-1/2}h_n^{-1/4}$. Le terme $o(1)$ est supposé uniforme en x.

Cette famille d'alternatives a été presentée dans [1] ; ce sont des alternatives de Pitman locales à f_0.

Pour obtenir la région critique ainsi que la puissance asymptotique, on donne une version modifiée du Théorème 2.4.

Lemme 3.1.

(i) *Sous H_0 et si les hypothèses du Théorème* 2.4 *sont satisfaites et que f_0' et f_0'' sont intégrables sur A^ϵ, alors quand $nh_n^{9/2} \to 0$, on a*

$$h_n^{-1/2}\left\{nh_n\tilde{I}_n(W) - \int_A f_0(x)W(x)\,\mathrm{d}x \int K^2(t)\,\mathrm{d}t, W \in \mathcal{W}\right\} \xrightarrow{\mathcal{D}} \{G_P(W),\ W \in \mathcal{W}\},$$

où $G_P(\cdot)$ est défini dans le Théorème 2.3.

(ii) *Sous H_n et les hypothèses de* (i), *si $nh_n^{3/2} \to \infty$, on a*

$$h_n^{-1/2}\left\{nh_n\tilde{I}_n(W) - \int_A f_0(x)W(x)\,\mathrm{d}x \int K^2(t)\,\mathrm{d}t, W \in \mathcal{W}\right\} \xrightarrow{\mathcal{D}} \{G'_P(W),\ W \in \mathcal{W}\},$$

où $G'_P(\cdot) = G_P(W) + \int_A \eta^2(x)W(x)\,\mathrm{d}x$.

Pour la simplicité, on prend $f_0 = \frac{1}{2}I_{[-1,1]}$. Soit $T_n(s) := h_n^{-1/2}(nh_n\tilde{I}_n(s) - \int_{-s}^{s} f_0(x)\,\mathrm{d}x \int K^2(x)\,\mathrm{d}x)$, où $\tilde{I}_n(s) = \int_{-s}^{s}(f_n(x) - f_0(x))^2\,\mathrm{d}x$. On compare les statistiques $S_n := \frac{1}{a}\int_0^a T_n(u)\,\mathrm{d}u$ et $T_n(a)$ (la statistique $T_n(a)$ est proposée par Bickel et Rosenblatt [1] pour $s = a$) pour $0 < a < 1$. Pour cela, on prend la classe $\mathcal{W} = \{W_s = I_{[-s,s]}, s \in [0, a]\}$ et $A = [-a, a]$. La variance dans le Théorème 2.3 devient $\sigma(W_s, W_t) = C(K)\min(s, t)$; où $C(K) := \int[\int K(s)K(z+s)\,\mathrm{d}s]^2\,\mathrm{d}z$. Le Lemme 3.1 et Lemme 3.9.8 dans [10], donnent sous $H_0 : S_n \xrightarrow{\mathcal{D}} \mathcal{N}(0, \tau^2)$, et sous $H_n : S_n \xrightarrow{\mathcal{D}} \mathcal{N}(m, \tau^2)$, où

$$\tau^2 := \frac{2C(K)}{a^2}\int_{-a}^{a}\big(a - |x|\big)^2\,\mathrm{d}x \quad \text{et} \quad m := \frac{1}{a}\int_{-a}^{a}\big(a - |x|\big)\eta^2(x)\,\mathrm{d}x.$$

La puissance asymptotique de S_n par rapport aux alternatives H_n est donc

$$\beta_{S_n} = 1 - \phi\left(\phi^{-1}(1-\alpha) - \frac{m}{\tau}\right),$$

tandis que celle de la statistique $T_n(a)$ est (voir [1])

$$\beta_{T_n} = 1 - \phi\left(\phi^{-1}(1-\alpha) - \frac{\int_{-a}^{a} \eta^2(x)\,\mathrm{d}x}{\sqrt{2C(K)\int_{-a}^{a} f_0^2(x)\,\mathrm{d}x}}\right),$$

où ϕ est la fonction de répartition de $\mathcal{N}(0,1)$.

Vue la croissance de ϕ, on constate que le test statistique S_n est meilleur que $T_n(a)$, si on prend la fonction η dans la classe

$$a\left(\sqrt{3}-1\right)\int_{-a}^{a} \eta^2(x)\,\mathrm{d}x \geqslant \sqrt{3}\int_{-a}^{a} |x|\eta^2(x)\,\mathrm{d}x.$$

Un exemple simple d'eléments de cette classe est la fonction identité sur l'intervalle $[-a/2, a/2]$.

Remerciement

Je tiens à remercier le Professeur Michel Broniatowski pour ses conseils et remarques.

Références

[1] P.J. Bickel, M. Rosenblatt, On some global measures of the deviations of density function estimates, Ann. Statist. 1 (1973) 1071–1095.
[2] Y. Fan, Testing the goodness of fit of a parametric density function by kernel method, Econometric Theory 10 (1994) 316–356.
[3] E. Giné, D.M. Mason, A.Y. Zaitsev, The L_1-norm density estimator process, Ann. Probab. 3 (2003) 719–768.
[4] B.K. Ghosh, G.W. Huang, The power and optimal kernel of the Bickel–Rosenblatt test for goodness of fit, Ann. Statist. 19 (1991) 999–1009.
[5] P. Hall, Central limit theorem for integrated square error of multivariate nonparametric density estimators, J. Multivariate Anal. 14 (1984) 1–16.
[6] E. Nadaraya, Nonparametric Estimate of Probability Densities and Regression Curves, Kluwer Academic, 1989.
[7] D. Pollard, Convergence of Stochastic Processes, Springer, New York, 1984.
[8] C. Tenreiro, Loi asymptotique des erreurs quadratiques intégrées des estimateurs à noyau de la densité et de la régression sous des conditions de dépendence, Portugal. Math. 54 (1997) 187–213.
[9] A.W. van der Vaart, Asymptotic in Statistics, Cambridge University Press, New York, 1998.
[10] A.W. van der Vaart, J.A. Wellner, Weak Convergence and Empirical Processes, Springer-Verlag, Berlin, 1996.