





C. R. Acad. Sci. Paris, Ser. I 343 (2006) 595-600

http://france.elsevier.com/direct/CRASS1/

Statistique/Probabilités

Principe de grandes déviations uniforme pour l'estimateur de la densité par la méthode des delta-suites

Noureddine Berrahou

L.S.T.A., université de Paris 6. 175, rue du Chevaleret, 75013 Paris, France
Reçu le 22 juillet 2005 ; accepté après révision le 26 septembre 2006
Disponible sur Internet le 24 octobre 2006
Présenté par Paul Deheuvels

Résumé

L'objet de cette Note est d'établir un principe de grandes déviations pour la déviation uniforme de l'estimateur de la densité par la méthode des delta-suites. Un résultat général est obtenu pour une delta-suite régulière quelconque et une discussion des hypothèses est donnée pour toutes les delta-suites associées à des méthodes d'estimation usuelles. Une application de nos résultats à l'évaluation des taux d'erreur associés aux différentes méthodes d'estimation est considérée. L'estimateur est construit ici à partir d'une suite de variables aléatoires i.i.d. *Pour citer cet article : N. Berrahou, C. R. Acad. Sci. Paris, Ser. I 343 (2006).* © 2006 Académie des sciences. Publié par Elsevier Masson SAS. Tous droits réservés.

Abstract

Uniform large deviations principle for the delta-sequence method density estimator. In this Note, we obtain a large deviations principle for the uniform deviation of the delta-sequence density estimator. A general result is stated for any regular delta-sequence and a discussion of hypotheses for the most usual methods is given. Implications of our results in the study of the inaccuracy rate of these estimates with the comparisons of performances of various methods are given. The estimation is based upon sequences of i.i.d. random variables. *To cite this article: N. Berrahou, C. R. Acad. Sci. Paris, Ser. I 343 (2006).*© 2006 Académie des sciences. Publié par Elsevier Masson SAS. Tous droits réservés.

Abridged English version

Let \mathcal{X} be an open interval of the real line \mathbb{R} . A sequence $\{\delta_m(x,u)\}$ of bounded measurable functions on $\mathcal{X} \times \mathcal{X}$ is a delta-sequence on \mathcal{X} if, for each $x \in \mathcal{X}$ and each \mathcal{C}^{∞} function φ with support in \mathcal{X} we have

$$\lim_{m\to\infty}\int\limits_{\mathcal{V}}\delta_m(x,u)\varphi(u)\,\mathrm{d}u=\varphi(x).$$

Let $X_1, X_2, ..., X_n$ be a sequence of i.i.d real random variables defined on a probability space (Ω, \mathcal{F}, P) and taking values in a set $\mathcal{X} \subseteq \mathbb{R}$. Denote by F the distribution function of X and by f its probability density function with

Adresse e-mail: berrahou@ccr.jussieu.fr (N. Berrahou).

respect to the Lebesgue measure over \mathcal{X} . We define the delta-sequence estimator of f associated to the sequence $\{\delta_m(x,u)\}$ by

$$f_n(x) = \frac{1}{n} \sum_{i=1}^n \delta_m(x, X_i),$$

where $m = m_n$ is a sequence of positive real numbers that tends to infinity with n.

In this Note, we obtain the large deviations principle for the uniform deviation of the delta-sequence density estimator f_n with respect to the underling density f. A general result pertaining with any regular delta-sequence is stated and a discussion of hypotheses for the most usual methods is given. Our results allow us to deduce some implications upon the inaccuracy rates pertaining with the usual estimates. More precisely, solving the minimization problem, under some constraints, of the function $Q(K) = \int K^2(x) \, dx$, we obtain that the kernel method associated to the kernel

$$K_b(x) = \begin{cases} \frac{1}{6b} \left(1 - \frac{1}{6b} x \right) & \text{if } x \in [0, 6b[, \\ \frac{1}{6b} \left(1 + \frac{1}{6b} x \right) & \text{if } x \in] - 6b, 0], \\ 0 & \text{elsewhere,} \end{cases}$$

offer the best performances with respect to the whole methods discussed in this Note. This work follows up several results of large deviations in nonparametric function estimation establish by Bitouzé and Louani [3] and Louani [5–7] and treating the kernel and the orthogonal series methods.

1. Introduction

Soit $\mathcal{X} \subseteq \mathbb{R}$ un intervalle ouvert. Une suite $\{\delta_m(x,u)\}$ de fonctions mesurables bornées sur $\mathcal{X} \times \mathcal{X}$ est une deltasuite sur \mathcal{X} si pour tout $x \in \mathcal{X}$ et pour toute fonction φ de classe \mathcal{C}^{∞} à support dans \mathcal{X} on a

$$\lim_{m\to\infty}\int\limits_{\mathcal{X}}\delta_m(x,u)\varphi(u)\,\mathrm{d}u=\varphi(x).$$

Soit X une variable aléatoire définie sur un espace de probabilité (Ω, \mathcal{F}, P) à valeurs dans \mathcal{X} . On note F sa fonction de répartition et f sa densité de probabilité. On considère X_1, X_2, \ldots, X_n une suite d'observations indépendantes de la variable X. L'estimateur de f par la méthode des delta-suites, en un point $x \in \mathbb{R}$, est donné par

$$f_n(x) = \frac{1}{n} \sum_{i=1}^n \delta_m(x, X_i),$$

où $m = m_n$ est une suite de réels positifs tendant vers l'infini avec n et satisfaisant des conditions qui seront spécifiées plus loin.

La méthode des delta-suites pour l'estimation de la densité englobe plusieurs méthodes classiques d'estimation parmi lesquelles on peut citer la méthode du noyau, la méthode des séries orthogonales, l'approche utilisant la transformée de Fourier et aussi la méthode des histogrammes (voir Walter et Blum [8]).

Dans cette Note, nous établissons un principe de grandes déviations pour la déviation uniforme de l'estimateur de la densité de probabilité par la méthode des delta-suites. Un résultat général est obtenu pour une delta-suite régulière quelconque et une discussion des hypothèses est donnée pour toutes les delta-suites associées à des méthodes d'estimation usuelles. Nous proposons une application de nos résultats à l'évaluation des taux d'erreur associés aux différentes méthodes d'estimation permettant ainsi de comparer leurs performances. Ce travail fait suite à plusieurs résultats de grandes déviations en estimation fonctionnelle non paramétrique établis par Bitouzé et Louani [3] et Louani [5–7] et traitant de la méthode du noyau et de la méthode des systèmes orthogonaux.

2. Résultats

Des hypothèses de régularité de la fonction f et de la suite $\{\delta_m(x,u)\}$ sont requises pour établir nos résultats. Elles se présentent sous la forme suivante :

- (H₁) Pour tout x dans \mathcal{X} , $\int_{\mathcal{X}} \delta_m(x, u) du = 1$.
- (H₂) Pour t > 0 et tout x dans \mathcal{X} , la quantité

$$L(t,x) := \lim_{n \to \infty} m_n \int_{\mathcal{X}} \left[\exp \left\{ \frac{t}{m_n} \delta_{m_n}(x,u) \right\} - 1 \right] f(u) \, \mathrm{d}u,$$

existe, et la convergence est uniforme par rapport à x. En outre, la fonction L(t,x) est dérivable par rapport à la variable t.

Introduisons maintenant des fonctions qui vont nous permettre de formuler nos résultats. Pour tout $x \in \mathcal{X}$ et tout

$$\Gamma_{f(x)}(\lambda) = \min \left\{ \Gamma_{f(x)}^{-}(\lambda), \Gamma_{f(x)}^{+}(\lambda) \right\}, \quad \text{où } \Gamma_{f(x)}^{\pm}(\lambda) = \sup_{t>0} \left\{ t \left(\lambda \pm f(x) \right) - L(\pm t, x) \right\},$$

et, pour tout $\lambda > 0$, on note $g(\lambda) = \inf\{\Gamma_{f(x)}(\lambda): x \in \mathcal{X}\}.$

- (H₃) La fonction g est continue croissante.
- (H_4) Il existe une suite $(H_n)_{n\geqslant 1}$ de réels positifs tendant vers l'infini et une suite de fonctions positives g_n , telles que,
 - (i) pour tout $\epsilon > 0$, il existe $n_0 > 0$ tel que, pour tout $n \ge n_0$, $\sup_{|x| > H_n} |\delta_{m_n}(x, u)| \le g_n(u) + \epsilon$,
 - (ii) pour tout $\zeta > 0$, $\lim_{n \to \infty} \frac{m_n}{n} E(g_n(X_1)) \exp(\zeta n/m_n) = 0$.
- (H₅) Il existe une partition $-H_n = \stackrel{\circ}{a_0} < a_1 < \cdots < a_{d_n} = H_n$ de l'intervalle $[-H_n, H_n]$, telle que, (i) pour tout $\epsilon > 0$, il existe $\eta_n > 0$ tel que pour tout $1 \leqslant j \leqslant d_n$ et tout $z \in \mathcal{X}$,

$$\sup_{\{(x,u)\in(a_{j-1},a_j)^2:\,|x-u|\leqslant\eta_n\}}\left|\delta_{m_n}(x,z)-\delta_{m_n}(u,z)\right|\leqslant\epsilon.$$

(ii) $\lim_{n\to\infty} m_n \log(H_n/\eta_n)/n = 0$

Le principe de grandes déviations uniforme est donné dans le théorème suivant :

Théorème 2.1. Supposons que les hypothèses (H_1) – (H_5) soient satisfaites, et que la densité f soit continue. Si $m_n \rightarrow$ ∞ et $m_n/n \to 0$ quand $n \to \infty$, alors

Pour tout fermé U de \mathbb{R}^+ ,

$$\limsup_{n\to\infty} \frac{m_n}{n} \log P(\|f_n - f\|_{\infty} \in U) \leqslant -\inf_{\lambda \in U} g(\lambda).$$

Pour tout ouvert V de \mathbb{R}^+ .

$$\liminf_{n\to\infty} \frac{m_n}{n} \log P(\|f_n - f\|_{\infty} \in V) \geqslant -\inf_{\lambda \in V} g(\lambda).$$

Remarque 1. La fonction L(t,x) peut être décomposée sous la forme L(t,x) = f(x)I(t), pour plusieurs types d'estimateurs couverts par la méthode des delta-suites. Une discussion à ce propos est donnée dans Berrahou [2]. Cette décomposition nous permet d'obtenir une forme plus explicite des fonctions de taux obtenues dans nos résultats. En effet, si I est deux fois dérivable et de dérivée ψ inversible, en prenant son inverse défini sur $D=(t_0,t_1)$, avec $t_0 = \inf\{\psi(t)\}\$ et $t_1 = \sup\{\psi(t)\}\$, par $\psi^{-1}(t) = \inf\{s: \psi(s) \ge t\}$, il suit alors que

$$\Gamma_{f(x)}^{\pm}(\lambda) = \begin{cases} f(x) \left[\left(1 \pm \frac{\lambda}{f(x)} \right) \psi^{-1} \left(1 \pm \frac{\lambda}{f(x)} \right) - I \circ \psi^{-1} \left(1 \pm \frac{\lambda}{f(x)} \right) \right] & \text{si } \left(1 \pm \frac{\lambda}{f(x)} \right) \in D_{\pm}, \\ \infty & \text{sinon,} \end{cases}$$

avec $D_- = [t_0, 1)$ et $D_+ = (0, t_1]$. Et pour tout $\lambda > 0$, on a $g(\lambda) = \min\{\Gamma_M^-(\lambda), \Gamma_M^+(\lambda)\}$, avec $M := \sup_{x \in \mathcal{X}} f(x) < 0$ ∞ . En outre, quand $\lambda \to 0$, on a $g(\lambda) = \frac{\lambda^2}{2M\psi'(0)}(1 + o(1))$.

Remarque 2. Dans la cas où l'on suppose que la fonction ψ est dérivable et de dérivée ψ' , on peut voir facilement que la fonction g est continue croissante. De ce fait, la condition (H₃) est verifiée pour toutes les méthodes d'estimations discutées dans ce travail.

2.1. Discussion des conditions

2.1.1. Estimation par la méthode du noyau

Soit K une fonction positive bornée et d'intégrale 1 telle que $\lim_{x\to\infty} |x|K(x) = 0$. Pour tout $(x,u) \in \mathbb{R}^2$, on considère la fonction

$$\delta_m(x, u) = mK(m(x - u)).$$

Nous introduisons maintenant quelques conditions suffisantes permettant aux hypothèses (H₁)-(H₅) d'être satisfaites.

- (A_1) K est une fonction lipschitzienne,
- (A₂) Il existe une suite (H_n) de réels positifs tendant vers l'infini telle que $P(|X| > H_n) \le \epsilon_n$ et pour $\zeta > 0$, $\lim_{n\to\infty} \frac{\epsilon_n m_n^2 \exp\{\xi n/m_n\}}{n} = 0.$ (A₃) $\lim_{n\to\infty} m_n \log(H_n m_n^2)/n = 0.$

Lemme 2.2. On suppose les hypothèses (A_1) – (A_3) satisfaites et la densité f uniformément continue. Si $m_n \to \infty$ quand $n \to \infty$, alors les conditions (H₁)-(H₅) sont vérifiées.

2.1.2. Estimation par la méthode des systèmes orthogonaux

Soit $\mathcal{X} = (a, b)$ un intervalle de \mathbb{R} et soit $\{e_i(x)\}_{i \in \mathbb{N}}$ un système orthonormal complet défini sur \mathcal{X} représentant les fonctions propres d'un opérateur compact sur $L^2(\mathcal{X})$. Pour $(x, u) \in \mathcal{X}^2$, on considère la suite de fonctions

$$\delta_m(x, u) = \sum_{j=1}^m e_j(x)e_j(u).$$

Dans le Lemme 2.3, ci-dessous, des conditions suffisantes permettant aux hypothèses (H₄)–(H₅) d'être satisfaites sont données. Pour cela, introduisons d'abord ces conditions.

(B₁) Pour tout $j \in \mathbb{N}$, e_j est une fonction lipschitzienne, avec une constante v_j et pour tout $n \ge 1$,

$$M_n := \sup_{j \leq m_n} \sup_{x \in \mathcal{X}} |e_j(x)| < \infty.$$

- (B₂) $\lim_{n\to\infty} m_n \log(M_n \sum_{j\leqslant m_n} v_j)/n = 0.$
- (B_3) Il existe une suite (H_n) de réels positifs tendant vers l'infini et une suite $(u_{i,n})$ de réels positifs tendant vers zéro, quand $n \to \infty$, telles que

 - (i) pour tout $|x| > H_n$, $|e_j(x)| \le u_{j,n}$, (ii) $\lim_{n \to \infty} \frac{m_n}{n} \log(H_n) = 0$, (iii) pour tout $\zeta > g(\lambda)$, $\lim_{n \to \infty} \frac{m_n}{n} \exp\{\zeta n/m_n\} \sum_{j \le m_n} u_{j,n} E((|e_j(X_1)|) = 0$, où E(X) désigne l'espérance mathématique de X.

Lemme 2.3. On suppose les hypothèses (B₁)–(B₃) satisfaites. Alors les conditions (H₄)–(H₅) sont réalisées.

Une discussion des hypothèses (H₁)-(H₂) et (B₁)-(B₃) est faite dans Louani [5] dans le cas de l'estimation utilisant la base trigonométrique et le système de Haar.

2.1.3. Estimation par la méthode des histogrammes

Dans ce cas, la delta-suite est donnée pour tout $(x, u) \in (0, 1)^2$ par

$$\delta_m(x, u) = m \sum_{j=1}^m \chi_j(x) \chi_j(u),$$

avec χ_1 la fonction indicatrice de (0, 1/m), et χ_j la fonction indicatrice de [(j-1)/m, j/m) pour $j=2,\ldots,m$. On peut montrer facilement que si $m_n \to \infty$ et $m_n \log(m_n)/n \to 0$ lorsque $n \to \infty$, et si f est continue, alors les conditions (H₁)–(H₅) sont satisfaites.

2.1.4. Estimation par la méthode du noyau de Fejér

Dans ce cas la delta-suite est définie, pour tout $(x, u) \in [-\pi, \pi]^2$ par

$$\delta_m(x,u) = \frac{\sin^2((m+1)(x-u)/2)}{2\pi(m+1)\sin^2((x-u)/2)}.$$

Lemme 2.4. On suppose les hypothèses (A_2) – (A_3) satisfaites, la densité f continue et m_n pair. Si $m_n \to \infty$ lorsque $n \to \infty$, alors les conditions (H₁)-(H₅) sont réalisées.

2.1.5. Estimation par la méthode de la transformée de Fourier

Pour $x, u \in \mathbb{R}$, on considère la suite de fonctions,

$$\delta_m(x,u) = \frac{1}{2\pi} \int_{-m}^m e^{is(x-u)} \, \mathrm{d}s.$$

Lemme 2.5. On suppose les hypothèses (A_2) – (A_3) satisfaites, la densité f dérivable et $\sup_x |f'(x)| < \infty$. Si $m_n \to \infty$ lorsque $n \to \infty$, alors les conditions (H₁)-(H₅) sont réalisées.

2.2. Applications

Une des applications offertes par les résultats de grandes déviations est la possibilité de pouvoir comparer les performances des méthodes d'estimation. Les résultats obtenus ici nous permettent de comparer les différentes méthodes d'estimation usuelles en utilisant les taux d'erreur associés. A cet effet, considérons la probabilité de dispersion d'un estimateur g_n de f uniformément en dehors d'un λ -voisinage de f définie par $\gamma(\lambda, f, g_n)$ $P(\|g_n - f\|_{\infty} > \lambda)$. Le taux d'erreur associé à un estimateur f_n par la méthode des delta-suites est défini par $e(\lambda, f, (f_n)) = -\lim_{n \to \infty} \frac{m_n}{n} \log \gamma(\lambda, f, f_n)$. Le Théorème 2.1 établi plus haut nous permet d'identifier $e(\lambda, f, (f_n))$ par $g(\lambda)$.

La comparaison des méthodes d'estimation s'effectue alors en comparant les fonctions de taux g associées. Il ressort en conclusion qu'une méthode est préférable à une autre lorsque pour tout $\lambda > 0$ la fonction g associée à la première méthode est supérieure à celle associé à la deuxième et inversement. Localement en $\lambda = 0$, la forme locale de g permet, de se limiter à la comparaison de la quantité $\psi'(0)$ associée aux différentes méthodes.

Les valeurs de $\psi'(0)$ associées aux méthodes d'estimation discutées dans ce travail sont présentées dans le Tableau 1.

La recherche du noyau optimal minimisant la quantité $\int K^2(x) dx$ a été effectuée dans le passé. Nous renvoyons au livre de Devroye et Györfi [4] pour le cas univarié et à Berlinet [1] dans le cas multivarié. Nous considérons ici l'optimalité sous l'ensemble des contraintes C suivantes :

- (i) K est une fonction continue positive et bornée,
- (ii) $\forall x \in \mathcal{X}, K(x) = K(-x),$
- (iii) $\int_{-\infty}^{\infty} K(x) dx = 1,$ (iv) $\int_{0}^{\infty} x K(x) dx = b > 0.$

Tableau 1 Les valeurs de $\psi'(0)$ associées aux méthodes d'estimation

Méthode d'estimation	$\psi'(0)$	Méthode d'estimation	$\psi'(0)$
méthode du noyau	$\int K^2(z) dz$	la base trigonométrique	2π
la base de Haar	1	méthode des histogrammes	1
noyau de Fejér	$1/3\pi$	transformée de Fourier	$1/\pi$

En s'appuyant sur le Théorème 2 établi par Berlinet [1], on obtient que le minimum 1/(9b) atteint par la quantité $\int K^2(x) dx$ sous l'ensemble des contraintes C est réalisé par le noyau

$$K_b(x) = \begin{cases} \frac{1}{6b} \left(1 - \frac{1}{6b} x \right) & \text{si } x \in [0, 6b[, \\ \frac{1}{6b} \left(1 + \frac{1}{6b} x \right) & \text{si } x \in] - 6b, 0], \\ 0 & \text{sinon.} \end{cases}$$

Remarque 3. La méthode du noyau utilisant le noyau K_b avec $b > \pi/3$ offre les meilleures performances relativement à l'ensemble des méthodes discutées dans ce travail.

Références

- [1] A. Berlinet, An optimality theorem and its application to estimation, Cahiers Centre Études Rech. Opér. 28 (1-3) (1986) 79-88.
- [2] N. Berrahou, Principe de grandes déviations pour l'estimateur de la densité par la méthode des delta-suites, C. R. Acad. Sci. Paris, Ser. I 337 (2003) 347–352.
- [3] D. Bitouzé, D. Louani, Un théorème de grandes déviations pour l'estimateur de la densité par la méthode de projection, C. R. Acad. Sci. Paris, Sér. I Math. 329 (1999) 441–444.
- [4] L. Devroye, L. Györfi, Nonparametric Density Estimation. The L₁-View, John Wiley, New York, 1985.
- [5] D. Louani, Large deviations results for orthogonal series density estimators and some applications, Math. Methods Statist. 12 (2) (2003) 177–196.
- [6] D. Louani, Large deviations limit theorems for the kernel density estimator, Scand. J. Statist. 25 (1998) 243–253.
- [7] D. Louani, Principe de grandes déviations pour l'estimateur à noyau de la densité de probabilité, C. R. Acad. Sci. Paris, Sér. I Math. 324 (1997) 569–572.
- [8] J. Walter, G. Blum, Probability density estimation using delta sequences, Ann. Statist. 7 (1979) 328–340.