Available online at www.sciencedirect.com
SCIENCE DIRECT®

C. R. Acad. Sci. Paris, Ser. I 337 (2003) 347–352

COMPTES RENDUS MATHEMATIQUE

Statistique/Probabilités

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

Noureddine Berrahou

LSTA, Université de Paris 6, 175, rue du Chevaleret, 8ème étage, batiment A, 75013 Paris, France

Reçu le 21 février 2003 ; accepté après révision le17 juillet 2003

Présenté par Paul Deheuvels

Résumé

L'objet de cette Note est d'établir un principe de grandes déviations ponctuel pour 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 des corollaires avec des fonctions de taux explicites sont déduits pour des delta-suites associées à des méthodes d'estimation usuelles. L'estimation est ici faite à partir de suites de variables aléatoires indépendantes et identiquement distribuées. ***Pour citer cet article : N. Berrahou, C. R. Acad. Sci. Paris, Ser. I 337 (2003).***

Abstract

Large deviations principle for the delta-sequence method density estimator. In this Note we obtain pointwise large deviations principle for the delta-sequence method density estimator. A general result is stated for any regular delta-sequence and corollaries with explicit rate functions are derived for delta-sequences, associated to usual estimation methods. The estimation is based upon sequences of independent and identically distributed random variables. ***To cite this article: N. Berrahou, C. R. Acad. Sci. Paris, Ser. I 337 (2003).***

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 $\mathcal{C}^\infty$ function φ with support in $\mathcal{X}$ we have

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

Let $X_1,X_2,\ldots,X_n$ be a sequence of i.i.d. real random variable defined on a probability space $(\Omega,\mathcal{F},P)$ 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

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

1631-073X/$ – see front matter
doi:10.1016/S1631-073X(03)00361-3

with respect to the Lebesgue measure over $\mathcal{X}$. 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 are concerned with pointwise large deviation principle for the delta-sequence method density estimator. In the main theorem, we establish a general result for any regular delta-sequence $\{\delta_m(x,u)\}$. This result is obtained under the hypotheses (H_1), (H_2), given below, which involve the regularity of both the density function f and the delta-sequence $\{\delta_m(x,u)\}$. The proof uses the usual arguments of the Cramér theorem proof; namely, Tchebycheff inequality is applied for derive the upper bound and an exponential change of measure is used to obtain the lower bound. For several estimation methods, the function $L(t,x)$ that appears in the hypothesis (H_2) is split up with respect to the form

$$L(t,x) = f(x)I(t).$$

This decomposition enables us to obtain a more explicit form of the rate function relative to our results. Investigations are carried out to determine the form of the function $I(t)$ associated to several usual estimation methods. In case of the kernel method, we obtain

$$I(t) = \int_{-\infty}^{\infty}\left[\exp\left\{tK(u)\right\} - 1\right]\mathrm{d}u.$$

For the density type estimation method, we have

$$I(t) = \lim_{n\to\infty} m_n \int_{-\infty}^{\infty}\left[\exp\left\{t\frac{g_{m_n}(u)}{m_n}\right\} - 1\right]\mathrm{d}u,$$

see below for the notation details. The study of the delta-sequence associated to Fejér kernel estimation method, gives

$$I(t) = \sum_{p=1}^{\infty}\frac{t^p}{(2\pi)^{p-1}p!(2p-1)!}\sum_{0\leqslant j\leqslant p}(-1)^j C_{2p}^j (p-j)^{2p-1}.$$

In the case of delta-sequence associated to Fourier transform estimation method, we obtain

$$I(t) = \int_{-\infty}^{\infty}\left[\exp\left\{t\frac{\sin u}{\pi u}\right\} - 1\right]\mathrm{d}u.$$

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 delta-suite sur $\mathcal{X}$ si pour tout $x \in \mathcal{X}$ et pour toute fonction φ de classe $\mathcal{C}^\infty$ à support dans $\mathcal{X}$ on a

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

Soit X une variable aléatoire définie sur un espace de probabilité $(\Omega, \mathcal{F}, P)$ 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 d'autres conditions qui seront spécifiées plus loin.

La méthode des delta-suites pour l'estimation de la densité englobe plusieures méthodes classiques d'estimation parmi lesquelles on peut citer la méthode du noyau introduite par Rosenblatt [8], Parzen [7], la méthode des séries orthogonales developpée en premier lieu par Chencov [3], l'approche utilisant la transformée de Fourier developpée par Blum et Susarla [2] et aussi la méthode de type densité introduite par Walter et Blum [10].

Dans cette Note, nous établissons un principe de grandes déviations ponctuel pour l'estimateur de la densité de probabilité par la méthode des delta-suites. Un résultat général est établi pour une delta-suite régulière quelconque et des corollaires avec des fonctions de taux explicites sont déduits pour des delta-suites associées à des méthodes d'estimation usuelles. Ce travail fait suite à plusieurs résultats de grandes déviations en estimation fonctionnelle non paramétrique établis par Bitouzé et Louani [1] et Louani [5,6] et traitant de la méthode du noyau et de la méthode des systèmes orthogonaux. Nos résultats en constituent une généralisation. Notons que nos résultats peuvent être étendus aisément au cas de l'estimation de la densité dans $\mathbb{R}^d$ ou dans un espace métrique abstrait.

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 :

(H_1) Pour tout x dans $\mathcal{X}$

$$\int_{\mathcal{X}} \delta_m(x, u)\, \mathrm{d}u = 1.$$

(H_2) 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.

Théorème 2.1. *On suppose que les hypothèses* (H_1), (H_2) *sont satisfaites. Si* $m_n \to \infty$ *et* $m_n/n \to 0$ *quand* $n \to \infty$*, alors pour tout* $x \in \mathcal{X}$*, on a*

(1) *pour tout fermé* U *de* $\mathbb{R}^+$,

$$\limsup_{n\to\infty} \frac{m_n}{n} \log P\left(\left|f_n(x) - f(x)\right| \in U\right) \leqslant -\inf_{\lambda\in U} \Gamma_x(\lambda),$$

(2) *pour tout ouvert* V *de* $\mathbb{R}^+$,

$$\liminf_{n\to\infty} \frac{m_n}{n} \log P\left(\left|f_n(x) - f(x)\right| \in V\right) \geqslant -\inf_{\lambda\in V} \Gamma_x(\lambda),$$

où $\Gamma_x(\lambda) = \min\{\Gamma_x^-(\lambda), \Gamma_x^+(\lambda)\}$*, avec* $\Gamma_x^{\pm}(\lambda) = \sup_{t>0}\{t(\lambda \pm f(x)) - L(\pm t, x)\}$.

La démonstration de ce théorème s'appuie sur les techniques développées par Cramér, à savoir le passage par des inégalités de type Tchebycheff pour établir les bornes supérieures et le changement exponentiel de mesure pour obtenir les bornes inférieures (voir, par exemple, le Théorème 2.2.3 dans Dembo et Zeitouni [4]).

Remarque 1. Pour plusieurs méthodes d'estimation qui seront discutées plus loin, la fonction $L(t,x)$ peut être décomposée sous la forme

$$L(t,x)=f(x)I(t).$$

Cette décomposition nous permet d'obtenir une forme plus explicite des fonctions de taux obtenues dans nos résultats. En effet, si I est dérivable et de dérivée ψ inversible, en prenant son inverse définie sur $D=(t_0,t_1)$, avec $t_0=\inf_t\{\psi(t)\}$ et $t_1=\sup_t\{\psi(t)\}$, par

$$\psi^{-1}(t)=\inf\{s:\psi(s)\geqslant t\},$$

il suit alors que

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

avec $D_-=[t_0,1)$ et $D_+=(0,t_1]$.

Remarque 2. Il a été établi dans Louani ([5], Lemme 1), que $\Gamma_x(\lambda)=\Gamma_{x,+}(\lambda)$ quand ψ est convexe et $t_1-1\geqslant 1-t_0$.

Dans ce qui suit, nous allons spécifier la forme de la fonction $I(t)$ relative à la décomposition de la fonction $L(t,x)$ dans le cas de choix de delta-suites associées à des méthodes d'estimation classiques.

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

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

Parzen [7] a montré que $\{\delta_m\}$ constitue une delta-suite. Comme $\int_{-\infty}^{\infty}K(u)\,\mathrm{d}u=1$, il est clair que la condition (H_1) est satisfaite.

Corollaire 2.2. *Supposons que $m_n\to\infty$ lorsque $n\to\infty$. Si f est continue, alors $I(t)=\int_{-\infty}^{\infty}[\exp\{tK(u)\}-1]\,\mathrm{d}u$.*

Remarque 3. Nos résultats imposent des conditions moins contraignantes que celles imposées dans Louani [5,6] puisque la condition d'existence et de bornitude de la dérivée est relaxée.

Estimation par la méthode de type densité. Soit $Y_1,Y_2,\ldots$ une suite de variables aléatoires i.i.d., centrées et de variance finie ayant une densité g bornée. Soit g_m la densité associée à la moyenne $\overline{Y}_m=m^{-1}\sum_{i=1}^m Y_i$. En utilisant l'inégalité de Tchebycheff, on peut montrer que la fonction

$$\delta_m(x,u)=g_m(x-u),\quad x,u\in\mathbb{R},$$

est une delta-suite. Comme g_m est une densité, il est clair que l'hypothèse (H_1) est verifiée. Le corollaire suivant identifie la fonction I relative à cette méthode d'estimation.

Corollaire 2.3. *Supposons que $m_n\to\infty$ lorsque $n\to\infty$ et f continue. Si pour tout $t>0$, la quantité*

$$I(t):=\lim_{n\to\infty}m_n\int_{-\infty}^{\infty}\left[\exp\left\{t\frac{g_{m_n}(u)}{m_n}\right\}-1\right]\mathrm{d}u$$

existe et si pour tout $\alpha > 0$, $\sup_{|u|\geqslant\alpha} g_{m_n}(u) \to 0$ *quand* $n \to \infty$, *alors* $L(x,t) = f(x)I(t)$.

Remarque 4. On peut voir facilement qu'on peut écrire $g_m(x) = mh_m(x)$, où h_m est la densité de $\sum_{i=1}^m Y_i$. La différence de cette méthode avec la méthode du noyau réside dans le fait que de façon générale h_m est une fonction dépendant de m alors que le noyau K est fixe.

Estimation par la méthode des systèmes orthogonaux. Soit $\mathcal{X} = (a,b)$ un intervalle de $\mathbb{R}$ et soit $\{\psi_m(x)\}$ 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 fonction

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

Walter ([9], p. 500) a montré que $(\delta_m(x,u))$ constitue une delta-suite. La fonction $(\delta_m(x,u))$ peut être associée à de nombreux systèmes orthogonaux incluant, entre autres, les fonctions trigonométriques, les polynômes de Legendre et les fonctions d'Hermite. Dans le cas des fonctions trigonométriques, $(\delta_m(x,u))$ est le noyau de Dirichlet.

En s'appuyant sur les résultats établis par Bitouzé et Louani [1] il ressort que si les hypothèses (H_1), (H_2) sont verifieés, alors le théorème reste vrai. En prenant le système de Haar pour construire la delta-suites (δ_m), il ressort que si la densité est continue, les résultats sont vrais avec

$$I(t) = \mathrm{e}^t - 1.$$

Par ailleurs, pour une delta-suite (δ_m) construite avec le système trigonométrique, il ressort que si

$$\sum_{r\in\mathbb{Z}} \int f(u)\mathrm{e}^{-\mathrm{i}ru}\,\mathrm{d}u < \infty,$$

la fonction I associée est donnée par

$$I(t) = 2\pi \sum_{p=1}^{\infty} \frac{t^p}{p!(p-1)!} \sum_{0\leqslant j\leqslant p/2} (-1)^j C_p^j \left(\frac{p}{2} - j\right)^{p-1}.$$

Estimation par la méthode du noyau de Fejér. On considère le noyau de Fejér défini, pour tout $u \in [-\pi,\pi]$ par

$$F_m(u) = \frac{\sin^2((m+1)u/2)}{2\pi(m+1)\sin^2(u/2)},$$

Winter [11] a démontré que la suite de fonction

$$\delta_m(x,u) = F_m(u-u),$$

constitue une delta-suite. Le corollaire suivant nous permet d'identifier la fonction $I(t)$ associée à ce cas.

Corollaire 2.4. *Supposons que* m_n *est pair et* $m_n \to \infty$ *lorsque* $n \to \infty$. *Alors*

$$I(t) = \sum_{p=1}^{\infty} \frac{t^p}{(2\pi)^{p-1}p!(2p-1)!} \sum_{0\leqslant j\leqslant p} (-1)^j C_{2p}^j (p-j)^{2p-1}.$$

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 $m_n \to \infty$ lorsque $n \to \infty$ et si f est continue, alors la fonction I associée à ce cas est donnée par $I(t) = (\mathrm{e}^t - 1)$.

Estimation par la méthode de la transformée de Fourier. Pour $x, u \in \mathbb{R}$, on considère la fonction

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

Blum et Susarla [2] ont montré que $(\delta_m(x, u))$ constitue une delta-suite. La fonction I associée à ce cas est donnée dans le corollaire ci-dessous.

Corollaire 2.5. *Supposons que $m_n \to \infty$ lorsque $n \to \infty$. Si f est dérivable et* $\sup_x |f'(x)| < \infty$*, alors* $I(t) = \int_{-\infty}^{\infty} [\exp\{t \frac{\sin u}{\pi u}\} - 1]\,\mathrm{d}u$.

Références

[1] 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 329 (1999) 441–444.
[2] J. Blum, V. Susarla, A Fourier inversion method for the estimation of a density and its derivatives, J. Austral. Math. Soc. 22 (1977) 166–171.
[3] N.N. Chencov, Evaluation of an unknown distribution density from observations, Soviet Math. 3 (1962) 1559–1562.
[4] A. Dembo, O. Zeitouni, Large Deviations Techniques and Applications, Jones and Bartelett, 1993.
[5] D. Louani, Principe de grandes déviations pour l'estimateur à noyau de la densité de probabilité, C. R. Acad. Sci. Paris, Sér. I 324 (1997) 569–572.
[6] D. Louani, Large deviations limit theorems for the kernel density estimator, Scand. J. Statist. 25 (1998) 243–253.
[7] E. Parzen, On the estimation of a probability density function and mode, Ann. Math. Statist. 33 (1962) 1065–1076.
[8] M. Rosenblatt, Remarks on some nonparametric estimates of a density function, Ann. Math. Statist. 27 (1956) 832–837.
[9] G. Walter, Expansions of distributions, Trans. Amer. Math. Soc. 116 (1965) 492–510.
[10] G. Walter, J. Blum, Probability density estimation using delta sequences, Ann. Statist. 7 (1979) 328–340.
[11] B.B. Winter, Rate of strong consistency of two nonparametric density estimators, Ann. Statist. 3 (1975) 759–766.