HAL open science

A note on extreme values and kernel estimators of sample boundaries

Stéphane Girard, Pierre Jacob

To cite this version:

Stéphane Girard, Pierre Jacob. A note on extreme values and kernel estimators of sample boundaries. Statistics and Probability Letters, 2008, 78 (12), pp.1634-1638. 10.1016/j.spl.2008.01.046 . hal-00724899

HAL Id: hal-00724899

https://inria.hal.science/hal-00724899

Submitted on 23 Aug 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.

A note on extreme values and kernel estimators of sample boundaries

Stéphane Girard[1,*] and Pierre Jacob[2]

INRIA Rhône-Alpes, team Mistis,
655, avenue de l'Europe, Montbonnot,
38334 Saint-Ismier Cedex, France.
Stephane.Girard@inrialpes.fr

[2] EPS/I3M, Université Montpellier 2,
Place Eugène Bataillon, 34095 Montpellier Cedex 5, France.
jacob@math.univ-montp2.fr

Abstract

In a previous paper [3], we studied a kernel estimate of the upper edge of a two-dimensional bounded set, based upon the extreme values of a Poisson point process. The initial paper [1] on the subject treats the frontier as the boundary of the support set for a density and the points as a random sample. We claimed in [3] that we are able to deduce the random sample case from the point process case. The present note gives some essential indications to this end, including a method which can be of general interest.

Keywords and phrases: support estimation, asymptotic normality, kernel estimator, extreme values.

1 Introduction and main results

As in the early paper of Geffroy [1], we address the problem of estimating a subset D of $\mathbb{R}^2$ given a sequence of random points $\Sigma_n = \{Z_1, ..., Z_n\}$ where the $Z_i = (X_i, Y_i)$ are independent and uniformly distributed on D. The problem is reduced to functional estimation by defining

$$D = \left\{(x, y) \in \mathbb{R}^2 / 0 \leq x \leq 1; 0 \leq y \leq f(x)\right\},$$

where f is a strictly positive function. Given an increasing sequence of integers $0 < k_n < n$, $k_n \uparrow \infty$, for $r = 1, ..., k_n$, let $I_{n,r} = [(r-1)/k_n, r/k_n[$ and

$$U_{n,r} = \max\left\{Y_i / (X_i, Y_i) \in \Sigma_n; X_i \in I_{n,r}\right\}, \tag{1}$$

*Corresponding author

where it is conveniently understood that $\max \varnothing = 0$. Now, let K be a bounded density, which has a support within a compact interval $[-A, A]$, a bounded first derivative and which is piecewise C^2, and $h_n \downarrow 0$ a sequence of positive numbers. Following [3], Section 6, we define the estimate

$$\hat{f}_n(x) = \frac{1}{k_n} \sum_{r=1}^{k_n} K_n(x - x_r) \left(U_{n,r} + \frac{1}{n - k_n} \sum_{s=1}^{k_n} U_{n,s} \right), x \in \mathbb{R}, \tag{2}$$

where x_r is the center of $I_{n,r}$ and, as usually,

$$K_n(t) = \frac{1}{h_n} K\left(\frac{t}{h_n}\right), t \in \mathbb{R}.$$

The perhaps curious second term in brackets in formula (2) is designed for reducing the bias (see [3], Lemma 8). Note that $\hat{f}_n$ can be rewritten as a linear combination of extreme values

$$\hat{f}_n(x) = \frac{1}{k_n} \sum_{r=1}^{k_n} \beta_{n,r}(x) U_{n,r},$$

where

$$\beta_{n,r}(x) = \frac{1}{k_n} K_n(x - x_r) + \frac{1}{k_n(n - k_n)} \sum_{s=1}^{k_n} K_n(x - x_s).$$

In the sequel, we suppose that f is α-Lipschitzian, $0 < \alpha \leq 1$, and strictly positive. Our result is the following:

Theorem 1 *If $h_n k_n \to \infty$, $n = o(k_n^{1/2} h_n^{-1/2-\alpha})$, $n = o(k_n^{5/2} h_n^{3/2})$ and $k_n = o(n/\ln n)$, then for every $x \in]0, 1[$,*

$$\left(n h_n^{1/2} / k_n^{1/2}\right) \left(\hat{f}_n(x) - f(x)\right) \Rightarrow \mathcal{N}\left(0, \sigma^2\right),$$

with $\sigma = \|K\|_2 / c$.

2 Proofs

If formally the definition of $\hat{f}_n$ is identical here and in [3] the fundamental difference lies in the fact that in [3] the sample is replaced by a homogeneous Poisson point process with a mean measure $\mu_n = nc\lambda 1_D$ where λ is the Lebesgue measure of $\mathbb{R}^2$ and $c^{-1} = \lambda(D)$. Here we denote by $\Sigma_{0,n}$ this point process and we need, for the sake of approximation, two further Poisson point processes $\Sigma_{1,n}$ and $\Sigma_{2,n}$. The point processes $\Sigma_{j,n}$ are constructed as in [2], extending an original idea of J. Geffroy. Given a sequence $\gamma_n \downarrow 0$, consider independent Poisson random variables $N_{1,n}$, $M_{1,n}$, $M_{2,n}$, independent of the sequence (Z_n), with parameters $\mathbb{E}(N_{1,n}) = n(1 - \gamma_n)$ and $\mathbb{E}(M_{1,n}) = \mathbb{E}(M_{2,n}) = n\gamma_n$. Define $N_{0,n} = N_{1,n} + M_{1,n}$, $N_{2,n} = N_{0,n} + M_{2,n}$ and take $\Sigma_{j,n} = \left\{Z_1, ..., Z_{N_{j,n}}\right\}$, $j = 0, 1, 2$. For $j = 0, 1, 2$ we define $U_{j,n,r}$ and $\hat{f}_{j,n}$ by imitating (1) and (2). Finally, let us introduce the event $E_n = \{\Sigma_{1,n} \subseteq \Sigma_n \subseteq \Sigma_{2,n}\}$. The following lemma is the starting point of our "random sandwiching" technique.

Lemma 1 *One always has $\hat{f}_{1,n} \leq \hat{f}_{0,n} \leq \hat{f}_{2,n}$. Moreover, if E_n holds, $\hat{f}_{1,n} \leq \hat{f}_n \leq \hat{f}_{2,n}$.*

Proof : The definition of the random sets $\Sigma_{j,n}$, $j = 0, 1, 2$ implies that $\Sigma_{1,n} \subseteq \Sigma_{0,n} \subseteq \Sigma_{2,n}$. Thus, since $\beta_{n,r}(x) \geq 0$ for all $r = 1, \ldots, k_n$, we have $\hat{f}_{1,n} \leq \hat{f}_{0,n} \leq \hat{f}_{2,n}$. Similarly, E_n implies that $\hat{f}_{1,n} \leq \hat{f}_n \leq \hat{f}_{2,n}$. ■

The success of the approximation between $\hat{f}_n$ and $\hat{f}_{0,n}$ is based upon two lemmas. The first one shows how large is the probability of the event E_n.

Lemma 2 *For n large enough,*

$$\mathbb{P}\left(\Omega \setminus E_n\right) \leq 2\exp\left(-\frac{1}{8}n\gamma_n^2\right).$$

Proof : Using the Laplace transform of a Poisson random variable X with parameter $\lambda > 0$, we get for $\varepsilon/2\lambda$ small enough,

$$\mathbb{P}\left(|X-\lambda| > \varepsilon\right) < \exp\left(-\varepsilon^2/4\lambda\right),$$

see for instance Lemma 1 in [2]. Clearly, $\Omega \setminus E_n = \{N_{1,n} > n\} \cup \{N_{2,n} < n\}$ and thus

$$\mathbb{P}\left(\Omega \setminus E_n\right) \leq \exp\left(-\frac{n\gamma_n^2}{4(1-\gamma_n)}\right) + \exp\left(-\frac{n\gamma_n^2}{4(1+\gamma_n)}\right).$$

The lemma follows. ■

The second lemma is essential to control the approximation obtained when the event E_n holds.

Lemma 3 *If $k_n = o(n/\log n)$ and $n = O\left(k_n^{1+\alpha}\right)$, then uniformly on $r = 1, \ldots, k_n$,*

$$\mathbb{E}\left(U_{2,n,r} - U_{1,n,r}\right) = O\left(\frac{k_n\gamma_n}{n}\right).$$

Proof : Let us define $m_{n,r} = \min_{x \in I_{n,r}} f(x)$ and $M_{n,r} = \max_{x \in I_{n,r}} f(x)$. Then,

$$\begin{aligned}
& \mathbb{E}\left(U_{2,n,r} - U_{1,n,r}\right) \\
= & \int_0^{M_{n,r}} \left(\mathbb{P}\left(U_{2,n,r} > y\right) - \mathbb{P}\left(U_{1,n,r} > y\right)\right) dy \\
= & \int_0^{m_{n,r}} \left(\mathbb{P}\left(U_{2,n,r} > y\right) - \mathbb{P}\left(U_{1,n,r} > y\right)\right) dy + \int_{m_{n,r}}^{M_{n,r}} \left(\mathbb{P}\left(U_{2,n,r} > y\right) - \mathbb{P}\left(U_{1,n,r} > y\right)\right) dy \\
\stackrel{def}{=} & A_{n,r} + B_{n,r}.
\end{aligned}$$

Introducing $\lambda_{n,r} = \int_{I_{n,r}} f(x)dx$, we can write $A_{n,r}$ as

$$A_{n,r} = \int_0^{m_{n,r}} \exp\left(\frac{n(1-\gamma_n)}{k_n}\left(y - k_n\lambda_{n,r}\right)\right) dy - \exp\left(\frac{n(1+\gamma_n)}{k_n}\left(y - k_n\lambda_{n,r}\right)\right) dy.$$

Now, $A_{n,r}$ is expanded as a sum $A_{1,n,r} + A_{2,n,r}$ with

$$\begin{aligned}
A_{1,n,r} &= \frac{k_n}{n(1-\gamma_n)} \exp\left(\frac{n(1-\gamma_n)}{k_n}\left(m_{n,r} - k_n\lambda_{n,r}\right)\right) - \frac{k_n}{n(1+\gamma_n)} \exp\left(\frac{n(1+\gamma_n)}{k_n}\left(m_{n,r} - k_n\lambda_{n,r}\right)\right), \\
A_{2,n,r} &= \frac{k_n}{n(1+\gamma_n)} \exp\left(-n(1+\gamma_n)\lambda_{n,r}\right) - \frac{k_n}{n(1-\gamma_n)} \exp\left(-n(1-\gamma_n)\lambda_{n,r}\right).
\end{aligned}$$

The part $A_{2,n,r}$ is easily seen to be a $o\left(n^{-s}\right)$ where s is a arbitrarily large exponent under the condition $k_n = o(n/\log n)$. Now, If a, b, x, y are real numbers such that $x < y < 0 < b < a$, we have $0 < ae^y - be^x < (a-b) + b\left(y - x\right)$. Applying to $A_{1,n,r}$ this inequality yields

$$A_{1,n,r} \leq \frac{k_n}{n}\frac{2\gamma_n}{(1-\gamma_n^2)} + \left(M_{n,r} - m_{n,r}\right)\frac{2\gamma_n}{(1+\gamma_n)}.$$

Under the hypothesis that f is $\alpha-$Lipschitzian, and the condition $n = O\left(k_n^{1+\alpha}\right)$, we have $(M_{n,r} - m_{n,r}) = O(k_n/n)$, so that $A_{n,r} = A_{1,n,r}+A_{2,n,r} = O\left(k_n\gamma_n/n\right)$. Now, for $m_{n,r} \leq y \leq M_{n,r}$, it is easily seen that

$$\mathbb{P}\left(U_{2,n,r} > y\right) - \mathbb{P}\left(U_{1,n,r} > y\right) \leq 2\gamma_n \frac{n}{k_n}(M_{n,r} - m_{n,r}),$$

and thus

$$B_{n,r} \leq 2\gamma_n \frac{n}{k_n}\left(M_{n,r} - m_{n,r}\right)^2 = O\left(\frac{k_n}{n}\gamma_n\right).$$

Clearly, the bounds on $A_{n,r}$ and $B_{n,r}$ are uniform in $r = 1, ..., k_n$, and thus we obtain the result. ■

We quote a technical lemma.

Lemma 4 *If $k_n = o\left(n\right)$ and $h_n k_n \to \infty$ when $n \to \infty$,*

$$\lim_{n\to\infty} \sum_{r=1}^{k_n} \beta_{n,r}(x) = 1.$$

Proof : Remarking that

$$\sum_{r=1}^{k_n} \beta_{n,r}(x) = \frac{n}{n-k_n}\frac{1}{k_n}\sum_{r=1}^{k_n} K_n(x - x_r),$$

the result follows from the well-known property

$$\lim_{n\to\infty} \frac{1}{k_n}\sum_{r=1}^{k_n} K_n(x - x_r) = 1,$$

see for instance [3], Corollary 2. ■

The next proposition is the key tool to extend the results obtained on Poisson processes to samples.

Proposition 1 *If $k_n = o(n/\log n)$, $h_n k_n \to \infty$, and $n = O\left(k_n^{1+\alpha}\right)$, then, for every $x \in]0,1[$,*

$$(nh_n^{1/2}/k_n^{1/2})\mathbb{E}\left(\left|\hat{f}_n(x) - \hat{f}_{0,n}(x)\right|\right) \to 0.$$

Proof : From Lemma 1, we have

$$\begin{aligned}
\mathbb{E}\left(\left|\hat{f}_n(x) - \hat{f}_{0,n}(x)\right| \mathbf{1}_{E_n}\right) &\leq \mathbb{E}\left(\hat{f}_{2,n}(x) - \hat{f}_{1,n}(x)\right) \\
&= \sum_{r=1}^{k_n} \beta_{n,r}(x)\mathbb{E}(U_{2,n,r} - U_{1,n,r}) \\
&\leq \sum_{r=1}^{k_n} \beta_{n,r}(x) \max_{1\leq s\leq k_n} \mathbb{E}(U_{2,n,s} - U_{1,n,s}) \\
&= O\left(\frac{k_n\gamma_n}{n}\right),
\end{aligned}$$

in view of Lemma 3 and Lemma 4. As a consequence,

$$(nh_n^{1/2}/k_n^{1/2})\mathbb{E}\left(\left|\hat{f}_n(x)-\hat{f}_{0,n}(x)\right|\mathbf{1}_{E_n}\right)=O\left(k_n^{1/2}h_n^{1/2}\gamma_n\right). \tag{3}$$

Now, let $M=\sup\{f(x), x\in[0,1]\}$. Then, applying Lemma 4 again,

$$\max\left\{\hat{f}_n(x),\hat{f}_{0,n}(x)\right\}\leq M\sum_{r=1}^{k_n}\beta_{n,r}(x)=O(1),$$

and therefore, from Lemma 2,

$$\begin{aligned}(nh_n^{1/2}/k_n^{1/2})\mathbb{E}\left(\left|\hat{f}_n(x)-\hat{f}_{0,n}(x)\right|\mathbf{1}_{\Omega\setminus E_n}\right) &= (nh_n^{1/2}/k_n^{1/2})O(1)\mathbb{P}(\Omega\setminus E_n)\\ &= o(n)\exp\left(-\frac{1}{8}n\gamma_n^2\right)\\ &= o(1)\exp\left(-\frac{n}{k_n}\left(\frac{1}{8}k_n\gamma_n^2-\frac{k_n}{n}\log n\right)\right). \end{aligned}\tag{4}$$

From (3) and (4) it suffices to take $\gamma_n=k_n^{-1/2}$ to obtain the desired result. ■

The main theorem is now obtained without difficulty.

Proof of Theorem 1. Under the conditions $h_nk_n\to\infty$, $n=o(k_n^{1/2}h_n^{-1/2-\alpha})$, $n=o(k_n^{5/2}h_n^{3/2})$ and $k_n=o(n/\ln n)$, Theorem 5 of [3] asserts that

$$\left(nh_n^{1/2}/k_n^{1/2}\right)\left(\hat{f}_{0,n}(x)-f(x)\right)\Rightarrow\mathcal{N}\left(0,\sigma^2\right),$$

while from Proposition 1,

$$\left(nh_n^{1/2}/k_n^{1/2}\right)\left(\hat{f}_{0,n}(x)-\hat{f}_n(x)\right)\xrightarrow{\mathbb{P}}0.$$

Thus, the result is an immediate application of Slutsky's theorem. ■

References

[1] Geffroy J. (1964) Sur un problème d'estimation géométrique. *Publications de l'Institut de Statistique de l'Université de Paris*, XIII, 191-200.

[2] Geffroy, J., Girard, S. and Jacob, P. (2006) Asymptotic normality of the L_1- error of a boundary estimate, *Nonparametric Statistics*, **18**(1), 21-31.

[3] Girard, S. and Jacob, P. (2004) Extreme values and kernel estimates of point processes boundaries. *ESAIM: Probability and Statistics*, **8**, 150-168.