ARTICLE IN PRESS

Journal of the Korean Statistical Society ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at SciVerse ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

Nonparametric estimation of regression level sets using kernel plug-in estimator

T. Laloë[a,*], R. Servien[b]

[a] Université de Nice Sophia Antipolis, Laboratoire J-A Dieudonné, Parc Valrose, 06108 Nice Cedex 2, France

[b] Montpellier SupAgro-INRA, UMR MISTEA 729, 2 place Pierre Viala, 34060 Montpellier Cedex 2, France

ARTICLE INFO

Article history:
Received 12 March 2012
Accepted 5 October 2012
Available online xxxx

AMS 2000 subject classifications:
62G05
62G08

Keywords:
Regression function
Level set
Plug-in estimator
Kernel estimator

ABSTRACT

Let (X, Y) be a random pair taking values in $\mathbb{R}^d \times J$, where $J \subset \mathbb{R}$ is supposed to be bounded. We propose a plug-in estimator of the level sets of the regression function r of Y on X, using a kernel estimator of r. We consider an error criterion defined by the volume of the symmetrical difference between the real and estimated level sets. We state the consistency of our estimator, and we get a rate of convergence equivalent to the one obtained by Cadre (2006) for the density function level sets.

1. Introduction

In this paper, we consider the problem of estimating the level sets of a regression function. More precisely, consider a random pair (X, Y) taking values in $\mathbb{R}^d \times J$, where $J \subset \mathbb{R}$ is supposed to be bounded. The goal of this paper is then to build an estimator of the level sets of the regression function r of Y on X, defined for all $x \in \mathbb{R}^d$ by

$$r(x) = \mathbb{E}[Y|X = x].$$

For $t > 0$, a level set for r is defined by

$$\mathcal{L}(t) = \{x \in \mathbb{R}^d : r(x) > t\}.$$

Assume that we have an independent and identically distributed sample (i.i.d.) $((X_1, Y_1), \ldots, (X_n, Y_n))$ with the same distribution as (X, Y). We then consider a plug-in estimator of $\mathcal{L}(t)$. More precisely, we use a consistent estimator $\hat{r}_n$ of r, in order to estimate $\mathcal{L}(t)$ by

$$\mathcal{L}_n(t) = \{x \in \mathbb{R}^d : \hat{r}_n(x) > t\}.$$

Most of the research works on the estimation of level sets concern the density function. One can cite the works of Cadre (2006), Cuevas and Fraiman (1997), Hartigan (1987), Polonik (1995), Tsybakov (1997), Walther (1997). This large number of works on this subject is motivated by the high number of possible applications. Estimating these level sets can be useful in mode estimation (Müller & Sawitzki, 1991; Polonik, 1995), or in clustering (Biau, Cadre, & Pelletier, 2007; Cuevas, Febrero, &

* Corresponding author. Tel.: +33 492076020.
E-mail addresses: laloe@unice.fr (T. Laloë), remi.servien@supagro.inra.fr (R. Servien).

1226-3192/$ – see front matter
doi:10.1016/j.jkss.2012.10.001

Fraiman, 2000, 2001). In particular, Biau et al. (2007) use an estimator of the level sets of the density function to determine the number of clusters.

The same applications are possible with the regression function. Moreover, it is for instance possible to use an estimator of the level sets of the regression function to determine the path of water flow from a digital representation of an area. In the same vein, in medical imaging, people want to estimate the areas where some function of the image exceeds a fixed threshold. In medical decision making, we can also find a lot of applications. For instance, the severity of the cancer is characterized by a variable Y which directly impacts the choice of standard or aggressive chemotherapy. For osteosarcoma (Man et al., 2005), Y is the percent necrosis in the tumor after a first round of treatment. If $Y > 0.9$ (this threshold has been fixed by experts and is now the convention), the aggressive chemotherapy will be chosen. The problem is that Y is measured using an invasive biopsy. If we can collect from the patient a feature vector X (which acquisition is easier), such as gene expression levels, knowledge of the regression level sets would allow the choice of an efficient treatment planning without a biopsy. Note that, in these examples, the use of a compact set J is fully justified. This is generally the case in most practical situations, particularly in image analysis.

Despite the many potential applications, the estimation of the level sets of the regression function has not been widely studied. Müller (1993) mentioned it briefly in his survey. Nowak and Willett (2007) obtained minimax rates (for different smoothness classes) for estimators based on recursive dyadic partitions. Scott and Davenport (2007) use a cost sensitive approach and a different measure of risk. Cavalier (1997) and Polonik and Wang (2005) used estimators based on the maximization of the excess mass which was introduced by Müller and Sawitzki (1991) and Hartigan (1987). Cavalier demonstrated asymptotic minimax rate of convergence for piecewise polynomial estimators using smoothness assumptions on the boundary of the level sets. We used a different approach and construct a plug-in estimator using the kernel estimator of the regression. The main advantage of our estimator is the simplicity of his calculation, inherited from the plug-in approach. Moreover, our estimator does not require strong assumptions on the shape of level sets.

All our consistency results are in the sense of the symmetrical difference (Fig. 1), defined by

$$\mathcal{L}_n(t) \Delta \mathcal{L}(t) = (\mathcal{L}_n(t) \cap \mathcal{L}^C(t)) \cup (\mathcal{L}_n^C(t) \cap \mathcal{L}(t)).$$

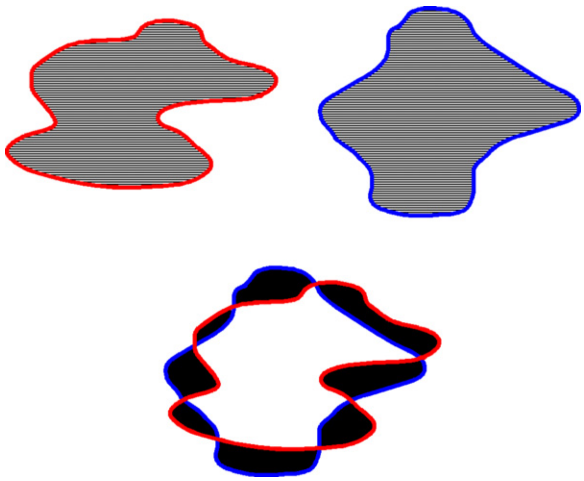

Fig. 1. Symmetrical difference (in black) between two sets A (in red) and B (in blue). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Our goal is to establish some consistency results under reasonable assumptions on r and $\hat{r}_n$. Using a kernel estimator for r, we get a rate of convergence equivalent to the one obtained by Cadre (2006) for the density function.

This paper is organized as follows. The definition of our estimator and consistency results are given in Section 2. In Section 3 we confront our estimator to simulated data. Finally, proofs are collected in Section 4.

2. Main results

2.1. Construction of the estimator

As announced, we use a plug-in approach. That is, given an estimator r_n of r we estimate $\{x \in \Lambda : r(x) > t\}$ by $\{x \in \Lambda : r_n(x) > t\}$. To estimate r, we choose to consider a kernel estimator.

Assume that we can write

$$r(x) = \frac{\varphi(x)}{f(x)},$$

where f is the density function of X, and φ is defined by $\varphi(x) = r(x)f(x)$.

Please cite this article in press as: Laloë, T., & Servien, R. Nonparametric estimation of regression level sets using kernel plug-in estimator. Journal of the Korean Statistical Society (2012), doi:10.1016/j.jkss.2012.10.001

Let K be a kernel on $\mathbb{R}^d$, that is a probability density on $\mathbb{R}^d$. We denote $h = h_n$ and $K_h(x) = K(x/h)$. From an i.i.d. sample $((X_1, Y_1), \ldots, (X_n, Y_n))$, we define, for all $x \in \mathbb{R}^d$,

$$\varphi_n(x) = \frac{1}{nh^d} \sum_{i=1}^{n} Y_i K_h(x - X_i) \quad \text{and} \quad f_n(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K_h(x - X_i).$$

For all $x \in \mathbb{R}^d$, the kernel estimator of r is then defined by

$$r_n(x) = \begin{cases} \varphi_n(x)/f_n(x) & \text{if } f_n(x) \neq 0 \\ 0 & \text{otherwise.} \end{cases}$$

The properties of this estimator are already well studied in the literature. For instance, the interesting reader can look at Bosq and Lecoutre (1987) or Gasser and Müller (1979).

Under the assumption

A0 There exists $t^- < t$ such that $\mathcal{L}(t^-)$ is compact. Besides, $\lambda(\{r = t\}) = 0$ (where λ stands for the Lebesgue measure),

a first consistency result can be trivially obtained from a slight modification of Theorem 3 by Cuevas, González-Manteiga, and Rodríguez-Casal (2006) and the consistency properties of the kernel estimator.

Proposition 2.1. *Under Assumption* A0, *if* K *is bounded, integrable, with compact support and Lipschitz, and if* $h \to 0$ *and* $nh^d/\log n \to \infty$, *then*

$$\mathbb{E}\,\lambda\left(\mathcal{L}_n(t) \Delta \mathcal{L}(t)\right) \underset{n\to\infty}{\to} 0.$$

Note that the last part of assumption A0 means that the regression function cannot have a null derivative at the estimated level set.

2.2. Rate of convergence

From now on, $\Theta \subset (0, \sup_{\mathbb{R}^d} r)$ is an open interval. Let us introduce the following assumptions:

A1 The functions r and f are twice continuously differentiable, and, $\forall t \in \Theta, \exists 0 < t^- < t : \inf_{\mathcal{L}(t^-)} f > 0$;

A2 For all $t \in \Theta$,

$$\inf_{r^{-1}(\{t\})} \|\nabla r\| > 0,$$

where, $\nabla \psi(x)$ stands for the gradient at $x \in \mathbb{R}^d$ of the differentiable function $\psi : \mathbb{R}^d \to \mathbb{R}$.

The assumptions A1 on the regularity are inherited from the classical assumptions in kernel estimation (Bosq & Lecoutre, 1987). Note that "harder" assumptions on the regularity of r and f will not improve the obtained rate of consistency. Moreover, let us mention that under Assumptions A1 and A2, we have (Proposition A.2 in Cadre (2006))

$$\forall t \in \Theta : \lambda(r^{-1}[t - \varepsilon, t + \varepsilon]) \to 0 \quad \text{as } \varepsilon \to 0.$$

Let us now introduce the assumptions on the kernel K.

A3 K is a continuously differentiable with a compact support. Moreover, there exists a decreasing function $\mu : \mathbb{R}^+ \to \mathbb{R}$ such that $K(x) = \mu(\|x\|)$ for all $x \in \mathbb{R}^d$.

We are now in a position to establish a rate of convergence for $\mathbb{E}\,\lambda(\mathcal{L}_n(t)\Delta\mathcal{L}(t))$.

Theorem 2.1. *Under Assumptions* A0–A3, *if* $nh^d/(\log n) \to \infty$ *and* $nh^{d+4}\log n \to 0$, *then for almost all* $t \in \Theta$

$$\mathbb{E}\,\lambda(\mathcal{L}_n(t)\Delta\mathcal{L}(t)) = O(1/\sqrt{nh^d}).$$

Remarks.

- Roughly speaking, the assumptions about the bandwidth impose to take h between $(\frac{\log n}{n})^{\frac{1}{d}}$ and $(n \log n)^{\frac{-1}{d+4}}$. Moreover, if we take $h = O((n\log n)^{\frac{-1}{d+4}})$, we get

$$\sqrt{nh^d} = O\left(\frac{n^{\frac{2}{d+4}}}{(\log n)^{\frac{d}{2(d+4)}}}\right) = O\left(\frac{n^{1/3}}{(\log n)^{1/6}}\right) \quad \text{with } d = 2,$$

that is a rate of the same order as Cadre (2001) in the density case.

- A remaining and crucial problem is the research of an optimal bandwidth h for our estimator. Indeed, if they are already results in the literature about an optimal bandwidth for the estimation of r, this bandwidth is not necessarily optimal for estimating $\mathcal{L}(t)$. However, in the simulations, we used a cross-validation procedure to choose a bandwidth.

Please cite this article in press as: Laloë, T., & Servien, R. Nonparametric estimation of regression level sets using kernel plug-in estimator. Journal of the Korean Statistical Society (2012), doi:10.1016/j.jkss.2012.10.001

2.3. Discussion about the rate

In this section, we provide a short comparison with the estimator proposed by Cavalier. Indeed, we choose this estimator because it is proven to be optimal (Cavalier, 1997).

The main idea of this estimator is that the level set $\mathcal{L}(t)$ minimizes the excess mass

$$M(G) = \int_G f(x)dx - t * \lambda(G).$$

Starting from this, Cavalier proposes to introduce estimators with piecewise-polynomial structure based on the maximization of local empirical excess mass. Assuming that $\mathcal{L}(t)$ can be expressed as

$$\mathcal{L}(t) = \{x = (r, \varphi), 0 \leq r < 2\pi\},$$

with g a 2π-periodic continuous function on $\mathbb{R}$, one starts by computing a piecewise-polynomial estimator $\hat{g}$ of g. Then, the estimate of $\mathcal{L}(t)$ is given by the closure of

$$\{(r, \varphi) : 0 \leq r < \hat{g}(\varphi), 0 \leq \varphi < 2\pi\}.$$

Note that this estimate is always star-shaped.

Depending on the used kind of design points, Cavalier obtains optimal rates of consistency.

If our estimator fails to get an optimal rate, its main advantage is its simplicity. Indeed, where getting the estimator $\hat{g}$ of g could be a little difficult, our estimator is really easy to implement. One only needs to do is compute a kernel estimation of the regression function (with one of the various existing R packages) and use the results to estimate the level set. Moreover, despite the regularity assumptions for f and r inherited from the use of a kernel estimator, our rate of consistency is obtained for general shapes of level sets. For example, we do not require that the level sets are star-shaped.

3. Study of finite sample behavior

In this section, we illustrate our method on a simple simulated data set. Consider the function r defined on $\mathbb{R}^2$ (Fig. 2) by

$$r : (u, v) \mapsto \frac{\sin(u) + \sin(v)}{\log\left(\sqrt{u^2 + v^2} + 1\right)}.$$

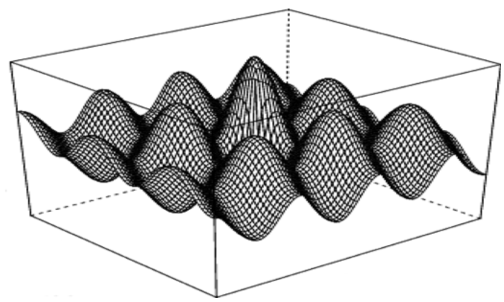

Fig. 2. Representation of r.

Theoretically, our results are established for a function r defined on $\mathbb{R}^2$. However, in order to compute easily the volume of the symmetrical difference, we will restrict ourselves to a bounded square. Let X be a random pair with a uniform distribution on the square $[-20, 20] \times [-20, 20]$. The size of the square is large enough to contain the level sets we will consider. We set $Y_i = r(X_i) + \varepsilon_i$, where $(X_1, \ldots, X_n)$ is an i.i.d. sample distributed as X, and $(\varepsilon_1, \ldots, \varepsilon_n)$ is an i.i.d. sample with a normal distribution centered on 0 and with standard deviation $\mathbf{0.1}$ ($X \equiv \mathcal{N}(0, 0.1)$).

3.1. Illustration of the rate

In this section we illustrate our theoretical rate of convergence obtained in Theorem 2.1. We use the function *npreg* of the R package "*np*" to perform the kernel estimation function, and the bandwidth is given by $h = (n \log n)^{-1.1/6}$. Then, we use a Monte Carlo approach to estimate the volume of the symmetrical difference (on the square $[-20, 20] \times [-20, 20]$). The error is then expressed in percents of the volume of the square.

For a level $t = 1$ and different values of n, we give the error multiplied by the rate of Theorem 2.1 $\sqrt{nh^2}$ in Fig. 3.

This figure seems to confirm the rate obtained in Theorem 2.1. Note that we consider a very large square, what can decrease artificially the error. However, it does not really matter since it does not change the conclusion.

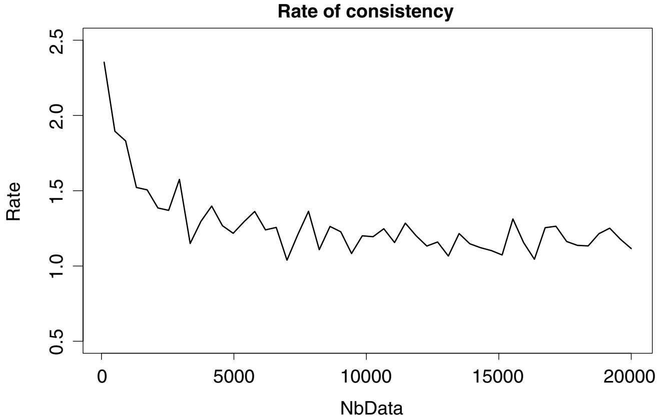

Fig. 3. Illustration of the rate: the curve represents the estimated error multiplied by $\sqrt{nh^2}$.

3.2. Selection of h by cross-validation

Now we use a simple cross-validation to select the bandwidth: for each value of n we use half the dataset to compute the kernel estimator with h (and the level set estimator) on a grid of 20 values between the limits allowed by the assumptions of Theorem 2.1 (see the first remark below Theorem 2.1). Then, we use the remaining part of the dataset to evaluate the volume of the symmetrical difference and select the optimal h. We compare the error obtained to the previous ones in Fig. 4.

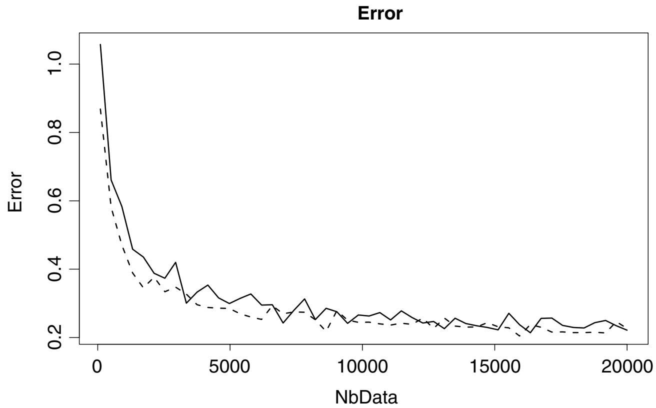

Fig. 4. Comparison of the error: the plain line stands for $h = (n \log n)^{-1.1/6}$, and the dotted line for h selected by a cross-validation.

We see that our choice process of h does not improve the estimation of the level sets. If we compare the error multiplied by $\sqrt{nh^2}$ (Fig. 5), we see that we select a lower bandwidth.

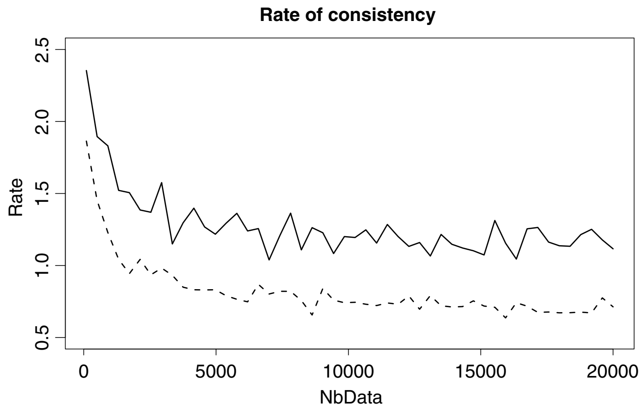

Fig. 5. Comparison of the rates: the plain line stands for $h = (n \log n)^{-1.1/6}$, and the dotted line for h selected by a cross-validation.

Please cite this article in press as: Laloë, T., & Servien, R. Nonparametric estimation of regression level sets using kernel plug-in estimator. Journal of the Korean Statistical Society (2012), doi:10.1016/j.jkss.2012.10.001

However, we cannot generalize about it since we are here in a very simple case. Moreover, we use a naive cross-validation method. Looking for more efficient methods to derive an optimal bandwidth for the level-set estimation is still an interesting and opened question. For this, we could first think of the adaptation of method used for density level set estimation like Rinaldo, Singh, Nugent, and Wasserman (2012) or Samworth and Wand (2010) for example.

4. Proofs

This section is dedicated to the proof of Theorem 2.1. From now on, c is a non-negative constant, which value may change from line to line.

4.1. Proof of Theorem 2.1

In this proof, some arguments are classical result from the kernel density (or regression) estimation theory. For more details, we refer the reader to the book by Bosq and Lecoutre (1987, Chapters 4 and 5).

From now on, we denote by ∂A the boundary of any subset $A \subset \mathbb{R}^d$. Besides, we introduce $\mathcal{H}$ the $(d-1)$-dimensional Hausdorff measure (Evans & Gariepy, 2000). Recall that $\mathcal{H}$ agrees with ordinary "$(k-1)$-dimensional surface area" on nice sets (Proposition A.1 in Cadre (2006)). Finally, we set $\widetilde{K} = \int K^2 d\lambda$.

4.1.1. Preliminary results

All the results in this sections are stated under Assumptions A0–A3. The proof of the theorem relies on the four following lemmas.

Let us define

$$\Omega_{n,c} = \left\{ \sqrt{nh^d} \sup_{\mathcal{L}_n(t) \cup \mathcal{L}(t)} |r_n - r| \geq c\sqrt{\log n} \right\}.$$

Lemma 4.1. *If $nh^{d+4}/\log n \to 0$, then there exists $\Gamma > 0$ such that*

$$\sqrt{nh^d}\,\mathbb{P}(\Omega_{n,\Gamma}) \to 0.$$

Note that the condition $nh^{d+4}/\log n \to 0$ is satisfied under the assumptions of Theorem 2.1.

Proof of Lemma 4.1. As r is continuous, we have $\sup_{\mathcal{L}(t^-)} |r| < c$. Assuming that $\inf_{\mathcal{L}(t^-)} f > 0$, then, since $\sup_{\mathcal{L}(t^-)} |f_n - f| \to 0$ a.s. under the assumptions of Lemma 4.1 (Bosq & Lecoutre, 1987), there exists $\theta > 0$ such that $\inf_{\mathcal{L}(t^-)} f_n > \theta$ a.s. for n large enough. So we can write

$$\begin{aligned} \sup_{\mathcal{L}(t^-)} |r_n - r| &= \sup_{\mathcal{L}(t^-)} \left| \frac{\varphi_n - \varphi}{f_n} + r\frac{f_n - f}{f_n} \right| \\ &\leq c \left(\sup_{\mathcal{L}(t^-)} |\varphi_n - \varphi| + \sup_{\mathcal{L}(t^-)} |f_n - f| \right). \end{aligned} \tag{1}$$

We have

$$\sup_{\mathcal{L}(t^-)} |\varphi_n - \varphi| \leq \sup_{\mathcal{L}(t^-)} |\varphi_n - \mathbb{E}\,\varphi_n| + \sup_{\mathcal{L}(t^-)} |\mathbb{E}\,\varphi_n - \varphi|.$$

We cover $\mathcal{L}(t^-)$ with ℓ_n balls $B_k = B(x_k, \rho_n)$ $(k = 1, \ldots, \ell_n)$ of radius ρ_n.

Consider $x \in \mathcal{L}(t^-)$, we denote by B_k the ball containing x. Then we set, for $x, x' \in \mathcal{L}(t^-)$,

$$A_n(x, x') = \frac{1}{n}\sum_{i=1}^{n} Y_i \left[K_h(x - X_i) - K_h(x' - X_i)\right] - \mathbb{E}\,\frac{1}{n}\sum_{i=1}^{n} Y_i \left[K_h(x - X_i) - K_h(x' - X_i)\right],$$

which leads us to

$$\sup_{\mathcal{L}(t^-)} |\varphi_n - \varphi| \leq \sup_{1 \leq k \leq \ell_n} |\varphi_n(x_k) - \mathbb{E}\varphi_n(x_k)| + \sup_{x \in \mathcal{L}(t^-)} |A_n(x, x_k)| + \sup_{\mathcal{L}(t^-)} |\mathbb{E}\,\varphi_n - \varphi|. \tag{2}$$

Then, since K is Lipschitz, there exists $\gamma > 0$ such that

$$\begin{aligned} |A_n(x, x_k)| &\leq ch^{-d-\gamma}\rho_n^{\gamma} \left(\frac{1}{n}\sum_{i=1}^{n} |Y_i| + \mathbb{E}|Y| \right) \\ &\leq ch^{-d-\gamma}\rho_n^{\gamma} \quad \text{since } Y \text{ is bounded.} \end{aligned}$$

Please cite this article in press as: Laloë, T., & Servien, R. Nonparametric estimation of regression level sets using kernel plug-in estimator. Journal of the Korean Statistical Society (2012), doi:10.1016/j.jkss.2012.10.001

As a consequence, we have

$$\mathbb{P}\left(\sup_{x\in\mathcal{L}(t^-)} |A_n(x, x_k)| > \frac{c}{4}\sqrt{\frac{\log n}{nh^d}}\right) \le \mathbb{P}\left(ch^{-d-\gamma}\rho_n^{\gamma} > \frac{c}{4}\sqrt{\frac{\log n}{nh^d}}\right).$$

One can choose

$$\rho_n = n^{-a}, \quad a > 0 \quad \text{and} \quad \rho_n^{\gamma} = o\left(h^{d+\gamma}\sqrt{\frac{\log n}{nh^d}}\right),$$

such that

$$\mathbb{P}\left(\sup_{x\in\mathcal{L}(t^-)} |A_n(x, x_k)| > \log n/\sqrt{nh^d}\right) = 0. \tag{3}$$

Then, using the arguments of the proof of Theorem 5.II.3 in Bosq and Lecoutre (1987), we obtain

$$\forall\varepsilon > 0, \quad \mathbb{P}\left(\sup_{1\le k\le l_n} |\varphi_n(x_k) - \mathbb{E}\varphi_n(x_k)| > \varepsilon\right) < 2\ell_n e^{-\frac{nh^d\varepsilon^2}{c}}.$$

If we set $\varepsilon = \varepsilon_0\sqrt{\log n/nh^d}$, we have

$$\mathbb{P}\left(\sup_{1\le k\le l_n} |\varphi_n(x_k) - \mathbb{E}\varphi_n(x_k)| > \varepsilon_0\sqrt{\frac{\log n}{nh^d}}\right) \le c\ell_n n^{-2\varepsilon_0/c}$$
$$\le cn^{-2\varepsilon_0/c}\rho_n^{-d}.$$

Remember that $\rho_n = n^{-a}$, with $a > 0$, one gets

$$\sqrt{nh^d}\mathbb{P}\left(\sup_{1\le k\le l_n} |\varphi_n(x_k) - \mathbb{E}\varphi_n(x_k)| > \varepsilon_0\sqrt{\frac{\log n}{nh^d}}\right) \le cn^{1/2+ad-2\varepsilon_0/c}\sqrt{h^d} \tag{4}$$

which tends to 0 choosing $\varepsilon_0 > \frac{(1/2+ad)c}{2}$.

Moreover, under A3, K is even which gives us

$$\sup_{\mathcal{L}(t^-)} |\mathbb{E}\,\varphi_n - \varphi| = O\left(\sqrt{\frac{\log n}{nh^d}}\right),$$

and, using that $nh^{d+4}/\log n \to 0$ we obtain

$$\sqrt{nh^d}\mathbb{P}\left(\sup_{\mathcal{L}(t^-)} |\mathbb{E}\,\varphi_n - \varphi| \ge \frac{c}{2}\sqrt{\frac{\log n}{nh^d}}\right) \to 0. \tag{5}$$

From (2) and using (3)–(5) we obtain

$$\sqrt{nh^d}\mathbb{P}\left(\sup_{\mathcal{L}(t^-)} |\varphi_n - \varphi| \ge c\sqrt{\frac{\log n}{nh^d}}\right) \to 0.$$

From (1) and such as $\sup_{\mathcal{L}(t^-)} |f_n - f| \to 0$ a.s., we conclude the proof. □

Consider $t \in \Theta$. For all $x \in \mathcal{L}(t^-)$, we define

$$V_n(x, t) = \mathrm{Var}((Y - t)K_h(x - X)) \quad \text{and} \quad \widetilde{\mathbb{E}}\,r_n(x) = \mathbb{E}\,\varphi_n(x)/\mathbb{E}\,f_n(x).$$

For all $x \in \mathcal{L}(t^-)$ such that $V_n(x, t) \ne 0$, we set

$$t_n(x) = \mathbb{E}f_n(x)\sqrt{\frac{nh^{2d}}{V_n(x, t)}}(t - \widetilde{\mathbb{E}}r_n(x)).$$

Besides, we consider the sets

$$\mathcal{V}_n^t = r^{-1}[t, t + \Gamma\sqrt{\log n/nh^d}] \cap \mathcal{L}(t^-) \quad \text{and} \quad \bar{\mathcal{V}}_n^t = r^{-1}[t - \Gamma\sqrt{\log n/nh^d}, t] \cap \mathcal{L}(t^-).$$

Finally, we denote by Φ the distribution function of the standard normal $\mathcal{N}(0, 1)$, and we define $\overline{\Phi}(x) = 1 - \Phi(x)$.

Lemma 4.2. *There exists $c > 0$ such that for all $n \geq 1$, $t \in \mathbb{R}$ and $x \in \mathcal{L}(t^-)$:*

$$|\mathbb{P}(r_n(x) \leq t) - \Phi(t_n(x))| \leq \frac{c}{\sqrt{nh^d}}.$$

Proof of Lemma 4.2. Set, for $i = 1, \dots, n$,

$$Z_i(x, t) = (Y_i - t)K_h(x - X_i), \qquad Z(x, t) = (Y - t)K_h(x - X).$$

By definition, we have $V_n(x, t) = \text{Var}(Z(x, t))$, and

$$\begin{aligned}\mathbb{P}(r_n(x) < t) &= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i(x, t) < 0\right) \\ &= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} (Z_i(x, t) - \mathbb{E}\, Z(x, t)) < -\mathbb{E}\, Z(x, t)\right) \\ &= \mathbb{P}\left(\sqrt{\frac{n}{V_n(x, t)}}\frac{1}{n}\sum_{i=1}^{n} (Z_i(x, t) - \mathbb{E}\, Z(x, t)) < t_n(x)\right).\end{aligned}$$

Then, the Berry–Esseen inequality (Berry, 1941) gives us

$$|\mathbb{P}(r_n(x) < t) - \Phi(t_n(x))| \leq \frac{c}{\sqrt{nV_n(x, t)^3}}\mathbb{E}\,|(Y - t)K_h(x - X) - \mathbb{E}(Y - t)K_h(x - X)|^3 . \tag{6}$$

Finally, under Assumptions A1 and A3, we have (see for example Bosq and Lecoutre (1987))

$$\sup_{x \in \mathcal{L}(t^-)} |(Y - t)K_h(x - X) - \mathbb{E}(Y - t)K_h(x - X)|^3 \leq ch^d$$

and

$$\inf_{x \in \mathcal{L}(t^-)} V_n(x, t) \geq ch^d.$$

The lemma can then be deduced from (6). □

Define now Θ_0 the set of all t in Θ such that

$$\lim_{\varepsilon \searrow 0} \frac{1}{\varepsilon}\lambda\left(r^{-1}[t - \varepsilon, t]\right) = \lim_{\varepsilon \searrow 0} \frac{1}{\varepsilon}\lambda\left(r^{-1}[t, t + \varepsilon]\right) = \int_{\partial\mathcal{L}(t)} \|\nabla r\|^{-1} d\mathcal{H}.$$

The following result is proven in Cadre (2006, Lemma 3.2).

Lemma 4.3. *$\Theta_0 = \Theta$ almost everywhere.*

Note that under Assumptions A1 and A2, we obtain, thanks to Proposition A.2 in Cadre (2006),

$$\lambda\left(r^{-1}[t - \varepsilon, t + \varepsilon]\right) = \lambda\left(r^{-1}(t - \varepsilon, t + \varepsilon)\right),$$

for all $t \in \Theta$ and $\varepsilon > 0$ small enough.

Finally, we set

$$v(x) = \text{Var}(Y|X = x) + r^2(x),$$

and, for $t \in \Theta$ and $x \in \mathcal{L}(t^-)$,

$$\bar{t}_n(x) = f(x)\sqrt{\frac{nh^d}{\widetilde{K}f(x)(v(x) + t^2)}}(t - r(x)).$$

We are now in a position to prove Lemma 4.4.

Lemma 4.4. *If $nh^d/(\log n) \to \infty$ and $nh^{d+4}\log n \to 0$, then for all $t \in \Theta_0$,*

$$\lim_{n\to\infty} \sqrt{nh^d}\left[\int_{\mathcal{V}_n^t} \mathbb{P}(r_n(x) < t)dx - \int_{\mathcal{V}_n^t} \Phi(\bar{t}_n(x))dx\right] = 0,$$

and

$$\lim_{n\to\infty} \sqrt{nh^d}\left[\int_{\overline{\mathcal{V}}_n^t} \mathbb{P}(r_n(x) > t)dx - \int_{\overline{\mathcal{V}}_n^t} \overline{\Phi}(\bar{t}_n(x))dx\right] = 0.$$

Proof of Lemma 4.4. We only prove the first equation, the second one can be obtained with similar arguments.

Define E_n by

$$E_n = \sqrt{nh^d}\int_{\mathcal{V}_n^t} |\Phi(t_n(x))dx - \Phi(\bar{t}_n(x))dx|.$$

As Φ is Lipschitz we have

$$E_n \le c\sqrt{nh^d}\lambda(\mathcal{V}_n^t)\sup_{\mathcal{V}_n^t}|t_n - \bar{t}_n|. \tag{7}$$

By definition of $t_n(x)$ and $\bar{t}_n(x)$, we have, for all $x \in \mathcal{V}_n^t$,

$$\begin{aligned}\frac{1}{\sqrt{nh^d}}|t_n(x) - \bar{t}_n(x)| &\le |t - r(x)|\left|\frac{f(x)}{\sqrt{\widetilde{K}f(x)(v(x)+t^2)}} - \frac{\mathbb{E}f_n(x)}{\sqrt{V_n(x,t)h^{-d}}}\right| + \sqrt{\frac{h^d}{V_n(x,t)}}\mathbb{E}f_n(x)|r(x) - \widetilde{\mathbb{E}}\,r_n(x)| \\ &\le \sqrt{\frac{\log n}{nh^d}}\left|\sqrt{\frac{|f(x)V_n(x,t)h^{-d} - (\mathbb{E}f_n(x))^2\widetilde{K}(v(x)+t^2)|}{\widetilde{K}(v(x)+t^2)V_n(x,t)h^{-d}}}\right| \\ &\quad + \sqrt{\frac{h^d}{V_n(x,t)}}\mathbb{E}f_n(x)|r(x) - \widetilde{\mathbb{E}}\,r_n(x)|.\end{aligned} \tag{8}$$

Remember that

$$|\widetilde{\mathbb{E}}\,r_n(x) - r(x)| \le \frac{1}{f_n(x)}|\mathbb{E}\phi_n(x) - \phi(x)| + |r(x)|\mathbb{E}|f_n(x) - f(x)|. \tag{9}$$

Since $\mathcal{V}_n^t$ is included in $\mathcal{L}(t^-)$, we can deduce (Bosq & Lecoutre, 1987) from A1, A3 and (9) that

$$\sup_{x\in\mathcal{V}_n^t}|\widetilde{\mathbb{E}}\,r_n(x) - r(x)| \le ch^2. \tag{10}$$

Moreover, if we set

$$V_n^1(x) = \text{Var } K_h(x - X), \qquad V_n^2 = \text{Var } Y\,K_h(x - X),$$

we can write

$$\begin{aligned}&\left|f(x)V_n(x,t)h^{-d} - (\mathbb{E}f_n(x))^2\,\widetilde{K}(v(x)+t^2)\right| \\ &\quad\le |f(x)|\left|V_n(x,t)h^{-d} - \widetilde{K}\mathbb{E}f_n(x)(v(x)+t^2)\right| + c|f(x) - \mathbb{E}f_n(x)| \\ &\quad\le |f(x)|\left|V_n(x,t)h^{-d} - \widetilde{K}f(x)(v(x)+t^2)\right| + c|f(x) - \mathbb{E}f_n(x)| \\ &\quad\le |f(x)|\left(t^2|V_n^1(x)h^{-d} - \widetilde{K}f(x)| + |V_n^2(x)h^{-d} - \widetilde{K}f(x)v(x)|\right. \\ &\qquad\left. +\ 2t\,|\text{Cov}\,(YK_h(x-X), K_h(x-X))|\right) + c|f(x) - \mathbb{E}f_n(x)| \\ &\quad\le c\left(|V_n^1(x)h^{-d} - \widetilde{K}f(x)| + |V_n^2(x)h^{-d} - \widetilde{K}f(x)v(x)|\right. \\ &\qquad\left. +\ |\text{Cov}\,(YK_h(x-X), K_h(x-X))| + |f(x) - \mathbb{E}f_n(x)|\right).\end{aligned}$$

Again, since $\mathcal{V}_n^t \subset \mathcal{L}(t^-)$, we can deduce (Bosq & Lecoutre, 1987) from A1 and A3 that

$$\sup_{x\in\mathcal{V}_n^t}|f(x)V_n(x,t)h^{-d} - (\mathbb{E}f_n(x))^2\widetilde{K}(v(x)+t^2)| \le ch. \tag{11}$$

We deduce from (8), (10) and (11) that

$$\sup_{x\in\mathcal{V}_n^t}|t_n(x) - \bar{t}_n(x)| \le c\left(\sqrt{h\log n} + \sqrt{nh^{k+4}}\right).$$

Then, thanks to (7) and since $t \in \Theta_0$, we have for n large enough

$$E_n \le c\sqrt{\log n}\left(\sqrt{h\log n} + \sqrt{nh^{k+4}}\right), \tag{12}$$

which tends to 0 under the assumptions on h of Theorem 2.1. Finally, Lemma 4.2 leads us to

$$\sqrt{nh^d}\left[\int_{\mathcal{V}_n^t}\mathbb{P}(r_n(x) < t)dx - \int_{\mathcal{V}_n^t}\Phi(t_n(x))dx\right] \le c\lambda(\mathcal{V}_n^t)$$

which tends to 0 since $\lambda(r^{-1}[t-\varepsilon, t+\varepsilon]) \to 0$. This and (12) ends the proof. □

Please cite this article in press as: Laloë, T., & Servien, R. Nonparametric estimation of regression level sets using kernel plug-in estimator. Journal of the Korean Statistical Society (2012), doi:10.1016/j.jkss.2012.10.001

4.1.2. Proof of Theorem 2.1

We first note that

$$\mathbb{E}\,\lambda\left(\mathcal{L}_n(t)\Delta\mathcal{L}(t)\right) = \int_{\mathcal{L}(t^-)\cap\{r\geq t\}} \mathbb{P}(r_n(x) < t)dx + \int_{\mathcal{L}(t^-)\cap\{r<t\}} \mathbb{P}(r_n(x) \geq t)dx.$$

Set

$$\mathbb{P}_{n,t}(x) = \mathbb{P}(r_n(x) < t), \qquad \overline{\mathbb{P}}_{n,t}(x) = \mathbb{P}(r_n(x) \geq t)$$

and remember that

$$\mathcal{V}_n^t = r^{-1}[t, t + \Gamma\sqrt{\log n/nh^d}] \cap \mathcal{L}(t^-) \quad \text{and} \quad \bar{\mathcal{V}}_n^t = r^{-1}[t - \Gamma\sqrt{\log n/nh^d}, t] \cap \mathcal{L}(t^-).$$

Consider $t \in \Theta_0$ and define

$$I_n = \int_{\mathcal{V}_n^t} \Phi(\bar{t}_n(x))dx, \qquad \bar{I}_n = \int_{\overline{\mathcal{V}}_n^t} \overline{\Phi}(\bar{t}_n(x))dx.$$

We have

$$I_n = \frac{1}{\sqrt{2\pi\widetilde{K}}} \int_{\mathcal{V}_n^t} \int_{-\infty}^{b_n(x)} \exp\left(\frac{-u^2}{2\widetilde{K}}\right) du\, dx$$

where $b_n(x) = \sqrt{f(x)nh^d}(t - r(x))/\sqrt{v(x) + t^2}$.

Besides,

$$b_n(x) = \sqrt{\frac{|\varphi(x)|}{v(x) + t^2}}\, b'_n(x),$$

with $b'_n(x) = \sqrt{nh^d}(t - r(x))/\sqrt{|r(x)|}$. Then we can find two positive constants C_1 and C_2 (whose values will then change from line to line) such that

$$C_1 b'_n(x) \leq b_n(x) \leq C_2 b'_n(x),$$

which leads us to

$$I_n \geq \frac{C_1}{\sqrt{2\pi\widetilde{K}}} \int_{\mathcal{V}_n^t} \int_{-\infty}^{b'_n(x)} \exp\left(\frac{-C_1^2 u^2}{2\widetilde{K}}\right) du\, dx,$$

and

$$I_n \leq \frac{C_2}{\sqrt{2\pi\widetilde{K}}} \int_{\mathcal{V}_n^t} \int_{-\infty}^{b'_n(x)} \exp\left(\frac{-C_2^2 u^2}{2\widetilde{K}}\right) du\, dx.$$

Using the arguments of the proof of Proposition 3.1 in Cadre (2006), we obtain

$$C_1 \frac{\sqrt{t\widetilde{K}}}{\sqrt{2\pi}} \int_{\partial\mathcal{L}(t)} \frac{1}{\|\nabla r\|} \partial\mathcal{H} \leq \varliminf_{n\to\infty} \sqrt{nh^d} I_n \leq \varlimsup_{n\to\infty} \sqrt{nh^d} I_n \leq C_2 \frac{\sqrt{t\widetilde{K}}}{\sqrt{2\pi}} \int_{\partial\mathcal{L}(t)} \frac{1}{\|\nabla r\|} \partial\mathcal{H}.$$

With similar arguments, we have

$$C_1 \sqrt{\frac{t}{2\pi}\widetilde{K}} \int_{\partial\mathcal{L}(t)} \frac{d\mathcal{H}}{\|\nabla r\|} \leq \varlimsup_{n\to\infty} \sqrt{nh^d} \bar{I}_n \leq \varliminf_{n\to\infty} \sqrt{nh^d} \bar{I}_n \leq C_2 \sqrt{\frac{t}{2\pi}\widetilde{K}} \int_{\partial\mathcal{L}(t)} \frac{d\mathcal{H}}{\|\nabla r\|}.$$

These inequalities, Lemmas 4.4 and 4.3 concludes the proof. □

Acknowledgments

The authors thank the associate editor and two anonymous referees for relevant remarks and constructive comments on a previous version of the paper.

References

Berry, A. (1941). The accuracy of the Gaussian approximation to the sum of independent variables. *Transactions of the American Mathematical Society*, *49*, 122–136.
Biau, G., Cadre, B., & Pelletier, B. (2007). A graph-based estimator of the number of clusters. *ESAIM Probability and Statistics*, *11*, 272–280.
Bosq, D., & Lecoutre, J. P. (1987). *Théorie de l'Estimation Fonctionnelle, Ecole Nationale de la Statistique et de l'Administration Economique et Centre d'Etudes des Programmes Economiques*. Economica.
Cadre, B. (2001). Convergent estimators for the L_1-median of a Banach valued random variable. *Statistics*, *35*(4), 509–521.
Cadre, B. (2006). Kernel estimation of density level sets. *Journal of Multivariate Analysis*, *97*(4), 999–1023.
Cavalier, L. (1997). Nonparametric estimation of regression level sets. *Statistics*, *29*(2), 131–160.
Cuevas, A., Febrero, M., & Fraiman, R. (2000). Estimating the number of clusters. *The Canadian Journal of Statistics*, *28*(1), 367–382.
Cuevas, A., Febrero, M., & Fraiman, R. (2001). Cluster analysis: a further approach based on density estimation. *Computational Statistics and Data Analysis*, *36*(4), 441–459.
Cuevas, A., & Fraiman, R. (1997). A plug-in approach to support estimation. *The Annals of Statistics*, *25*(6), 2300–2312.
Cuevas, A., González-Manteiga, W., & Rodríguez-Casal, A. (2006). Plug-in estimation of general level sets. *Australian and New Zealand Journal of Statistics*, *48*(1), 7–19.
Evans, L. C., & Gariepy, R. F. (2000). *Studies in advanced mathematics*, *Measure theory and fine properties of functions*. CRC Press.
Gasser, T., & Müller, H. (1979). Kernel estimation of regression functions. In Th. Gasser et M. Rosenblatt (Ed.), *Lecture notes in mathematics*, *Smoothing techniques for curve estimation* (pp. 23–68). Berlin: Springer Verlag.
Hartigan, J. A. (1987). Estimation of a convex density contour in two dimensions. *Journal of the American Statistical Association*, *82*(397), 267–270.
Man, T., Chintagumpala, M., Visvanathan, J., Shen, J., Perlaky, L., Johnson, J., Davino, N., Murray, J., Helman, L., Meyer, W., Triche, T., Wong, K., & Laus, C. (2005). Expression profiles of osteosarcoma that can predict response to chemotherapy. *Cancer Research*, *65*, 8142–8150.
Müller, D. (1993). The excess mass approach in statistics. *Beiträge zur Statistik Nr. 3*. Universität Heidelberg.
Müller, D., & Sawitzki, G. (1991). Excess mass estimates and tests for multimodality. *Journal of American Statistical Society*, *86*, 738–746.
Nowak, R. D., & Willett, R. M. (2007). Minimax optimal level-set estimation. *IEEE Transactions on Image Processing*, *16*(12), 2965–2979.
Polonik, W. (1995). Measuring mass concentrations and estimating density contour clusters-an excess mass approach. *The Annals of Statistics*, *23*(3), 855–881.
Polonik, W., & Wang, Z. (2005). Estimation of regression contour clusters: an application of the excess mass approach to regression. *Journal of Multivariate Analysis*, *94*, 227–249.
Rinaldo, A., Singh, A., Nugent, R., & Wasserman, L. (2012). Stability of density-based clustering. *Journal of Machine Learning*, *13*, 905–948.
Samworth, R., & Wand, M. (2010). Asymptotics and optimal bandwidth selection for highest density region estimation. *Annals of Statistics*, *38*(3), 1767–1792.
Scott, C., & Davenport, M. (2007). Regression level set estimation via costsensitive classification. *IEEE Transaction on Signal Processing*, *55*, 2752–2757.
Tsybakov, A. B. (1997). On nonparametric estimation of density level sets. *The Annals of Statistics*, *25*(3), 948–969.
Walther, G. (1997). Granulometric smoothing. *The Annals of Statistics*, *25*(6), 2273–2299.

Please cite this article in press as: Laloë, T., & Servien, R. Nonparametric estimation of regression level sets using kernel plug-in estimator. Journal of the Korean Statistical Society (2012), doi:10.1016/j.jkss.2012.10.001