



HAL
open science

Nonparametric estimation of regression level sets

Thomas Laloë, Rémi Servien

► **To cite this version:**

Thomas Laloë, Rémi Servien. Nonparametric estimation of regression level sets. 2011. hal-00674197v1

HAL Id: hal-00674197

<https://hal.science/hal-00674197v1>

Preprint submitted on 27 Feb 2012 (v1), last revised 8 Oct 2012 (v3)

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.

Nonparametric estimation of regression level sets

Thomas Laloë¹, Rémi Servien²

¹Université de Nice Sophia Antipolis, Laboratoire J-A Dieudonné,
Parc Valrose, 06108 Nice Cedex 2, France.
lalo@unice.fr

²Montpellier SupAgro-INRA, UMR MISTEA 729,
Bâtiment 29, 2 place Pierre Viala, 34060 Montpellier Cedex 2, France.
remi.servien@supagro.inra.fr

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 a 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 for the density function level sets. Finally we discuss the practical results obtained with a simulated data set.

Key-words and phrases: Regression function, Level set, Plug-in estimator, Kernel estimator.

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.) $\left((X_1, Y_1), \dots, (X_n, Y_n)\right)$ with the same distribution than (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 [3], Cuevas and Fraiman [7], Hartigan [11], Polonik [15], Tsybakov [19], Walther [20]. 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 and Stawitzki [13], Polonik [15]), or in clustering (Biau, Cadre and Pelletier [1], Cuevas, Febrero and Fraiman [6, 5]). In particular, Biau, Cadre and Pelletier [1] 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. It may be useful, for instance in order to automatically determine the location or the nature of a tumor. Note that, in these two 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 [12] mentioned it briefly in his survey. One can also cite the recent work of Cavalier [4], Polonik and Wang [16], Scott and Davenport [18], and Willett and Nowak [14]. The estimator proposed by Cavalier is based on the maximization of the excess mass. It is an adaptation of the estimator proposed by Tsybakov [19] for the density function. Scott and Davenport use a cost sensitive approach. The main advantage of our estimator is the simplicity of his calculation. Moreover, unlike the estimator proposed by Cavalier, our estimator does not require that the level sets are star shaped.

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

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

where $\mathcal{L}_n = \mathcal{L}_n(t)$ and $\mathcal{L} = \mathcal{L}(t)$.

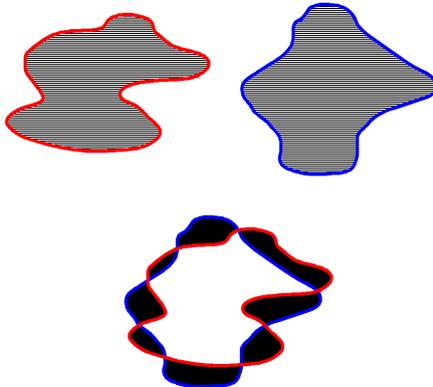


Figure 1: Symmetrical difference (in black) between two sets A (in red) and B (in blue).

Our goal is to establish some consistency results under reasonable assumptions on r and \hat{r}_n . We first use the results by Cuevas, González-Manteiga and Rodríguez-Casal [8] in order to state a consistency result for any consistent estimator \hat{r}_n of r . Then we particularize our approach by considering a kernel estimator of the regression function. For this estimator, we get a rate of convergence equivalent to the one obtained by Cadre [3] for the density function.

This paper is organized as follows. We give the main results in Section 2. Section 3 is devoted to the confrontation of our estimator to simulated data. Finally, proofs are collected in Section 4.

2 Main results

2.1 Consistency

From now on, $\|\cdot\|$ stands for the Euclidean norm on a finite dimensional space. Besides, for all integrable function $g : \mathbb{R}^d \rightarrow \mathbb{R}$, we denote by $\|g\|_p$ the norm defined by

$$\|g\|_p = \left(\int_{\mathbb{R}^d} |g(x)|^p dx \right)^{1/p}.$$

2.1.1 General estimator \hat{r}_n of r

In this paragraph, we assume that we know a consistent estimator \hat{r}_n of r (in a sense defined hereafter). Let us introduce the following assumption on the

regression function r :

A0 There exists t^- such that $\mathcal{L}(t^-)$ is compact. Besides, $\lambda(\{r = t\}) = 0$.

Roughly speaking, the last part of this assumption means that the Lebesgue measure does not charge the set $\{x \in \mathbb{R}^d : r(x) = t\}$.

Theorem 3 by Cuevas, González-Manteiga and Rodríguez-Casal [8] states the almost sure consistency of $\lambda(\mathcal{L}_n \Delta \mathcal{L})$ for any consistent estimator \hat{r}_n of r . Theorem 2.1 below deals with the consistency of $\mathbb{E}\lambda(\mathcal{L}_n \Delta \mathcal{L})$.

Theorem 2.1 *Under Assumption A0, if*

$$\sup_{\mathbb{R}^d} |\hat{r}_n - r| \xrightarrow[n \rightarrow \infty]{} 0 \text{ a.s.},$$

then

$$\mathbb{E} \lambda(\mathcal{L}_n(t) \Delta \mathcal{L}(t)) \xrightarrow[n \rightarrow \infty]{} 0.$$

Then, provided that \hat{r}_n is consistent with respect to the L_p or supremum norm, the plug-in estimator of the level sets of the regression function is consistent with respect to the symmetrical difference. For example, the BSE estimator (Bosq and Lecoutre [2], Chapter 7), the k -nearest neighbor estimator ([2] Chapter 8), as well as the regressogram ([2], Chapter 6) satisfy this property. In the next paragraph, we focus on the case of the kernel estimator.

2.1.2 Kernel estimator r_n of r

Let us consider now the case of the kernel estimator of the regression function. 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 $\varphi(x) = r(x)f(x)$.

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

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

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

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

For a complete study of the kernel estimator of the regression function and its properties, we refer to the book of Prakasa Rao [17].

We derive the following result from Theorem 2.1 using the consistency properties of the kernel estimator of the regression function (Bosq and Lecoutre [2]).

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

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

This result gives concrete conditions to obtain the consistency of $\mathcal{L}_n(t) \Delta \mathcal{L}(t)$ for the kernel estimator. It then particularizes the previous result from Cuevas, González-Manteiga and Rodríguez-Casal [8]. In the next paragraph, we establish a rate of convergence for this estimator.

2.2 Rate of convergence

Remember that r_n is the kernel estimator of the regression function. 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 $\psi : \mathbb{R}^d \rightarrow \mathbb{R}$.

Let us mention that under Assumptions **A1** and **A2**, we have (Proposition A.2 in [3])

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

This property, first used by Polonik [15], is almost identical to the last part of Assumption **A0** in Paragraph 2.1.1. 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 $\mu : \mathbb{R}^+ \rightarrow \mathbb{R}$ such that $K(x) = \mu(\|x\|)$ for all $x \in \mathbb{R}^d$.

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 and Gariepy [9]). Recall that \mathcal{H} agrees with ordinary “ $(k-1)$ -dimensional surface area” on nice sets (Proposition A.1 in [3]). Finally, we set $\tilde{K} = \int K^2 d\lambda$.

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 2.2 *Under Assumptions A0 – A3, if $nh^d/(\log n)^7 \rightarrow \infty$ and $nh^{d+4} \log n \rightarrow 0$, then*

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

Note that our estimator is easy to calculate and requires reasonable assumptions. Cavalier [4] gets a better rate, but with an estimator more difficult to calculate and with more restrictive assumptions. For example, we do not need for the level sets to be star-shaped. However, the choice of the bandwidth h in our estimator is not an easy task, since an optimal bandwidth for estimating r is not necessarily optimal for estimating \mathcal{L} .

3 Study of finite sample behavior

In this section, we illustrate our method on a simple simulated data set. Consider the function $r : [-6.5, 4.5] \times [-6.5, 4.5] \rightarrow [-2, 2]$ defined by

$$r : (u, v) \mapsto \sin(u) + \sin(v).$$

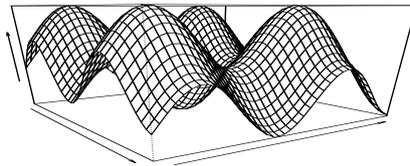


Figure 2: Representation of $r(u, v) = \sin(u) + \sin(v)$ on $[-6.5, 4.5] \times [-6.5, 4.5]$.

We selected this function to check that our estimator detects the different connected components of the level sets. Indeed, we see on Figure 2 that, for appropriate values of t , $\mathcal{L}(t) = \{(u, v) : r(u, v) > t\}$ admits four connected components. This is an interesting point since a possible application of the estimation of level sets is to use the connected components to do clustering (Biau, Cadre and Pelletier [1]).

Let X be a random pair with a binormal distribution on \mathbb{R}^2 (truncated on $[-6.5, 4.5] \times [-6.5, 4.5]$) centered on $(-1, -1)$ and with a covariance given by $6 * Id$. We set $Y_i = r(X_i) + \varepsilon_i$, where (X_1, \dots, X_n) is an i.i.d sample distributed as X , and $(\varepsilon_1, \dots, \varepsilon_n)$ is an i.i.d. sample with a normal distribution $\mathcal{N}(0, 0.01)$. For different values of n and t , we use the sample $\left((X_1, Y_1), \dots, (X_n, Y_n)\right)$ to estimate $\mathcal{L}(t)$. The problem of the selection of an optimal bandwidth h_n is not studied here. We simply choose a bandwidth fulfilling the conditions of Theorem 2.2. The choice of h_n is a full problem and it could be the subject of a future work.

We consider the levels $t = 0.5$, $t = 1$ and $t = 1.5$, and the sample sizes $n = 500$, $n = 2500$ and $n = 7500$. We evaluate our estimator on a regular grid of 3364 points on $[-6.5, 4.5] \times [-6.5, 4.5]$. Figure 3 represents the symmetrical difference between the real (**R**) and estimated (**E**) level sets, for the different values of t (in column) and n (in row). The points of $\mathbf{R} \setminus \mathbf{E}$ are represented in blue, and the one of $\mathbf{E} \setminus \mathbf{R}$ in red.

As expected, the quality of the estimation improves as n grows. However, we observe that a relatively large number of observations is needed to obtain a satisfactory estimation (based on a visual inspection).

We note that, in each situation, the connected components can be well distinguished (visually speaking). It comforts us in the idea that we could use level sets estimation to perform clustering. Besides, connecting clusters to level sets has already been proposed. We can cite for instance Hartigan [10] or Cuevas, Febrero and Freiman [6].

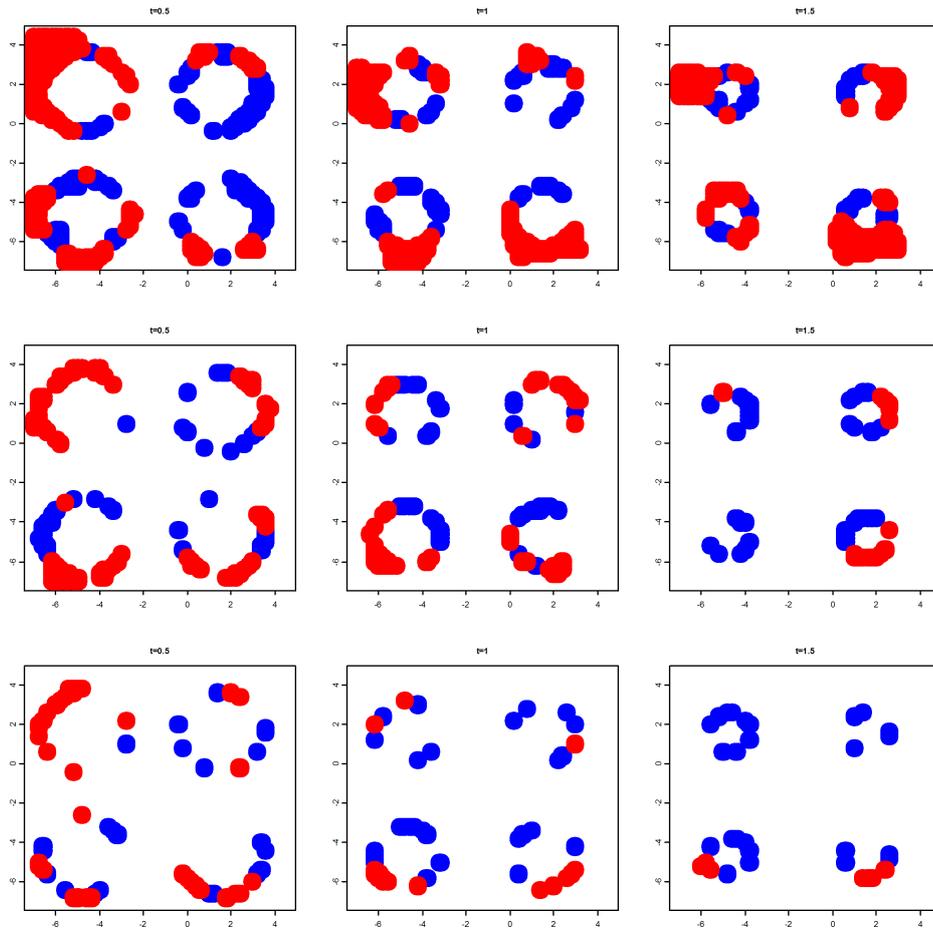


Figure 3: Graphical representation of the symmetrical difference between the real and estimated level sets.

4 Proofs

This section is dedicated to the proofs of Theorem 2.1 and Theorem 2.2. From now on, c , C_1 , and C_2 are non-negative constants, whose values may change from line to line.

4.1 Proof of Theorem 2.1

Let $\varepsilon > 0$, for all $n \in \mathbb{N}$, we define $E_{n,\varepsilon} = \{x \in \mathcal{L}_n(t) \cup \mathcal{L}(t) : |\hat{r}_n(x) - r(x)| \leq \varepsilon\}$. Since $\sup_{\mathbb{R}^d} |\hat{r}_n - r| \rightarrow 0$ a.s., we have that $\mathcal{L}_n(t) \subset \mathcal{L}(t^-)$ for n large enough which gives us $E_{n,\varepsilon} = \{x \in \mathcal{L}(t^-) : |\hat{r}_n(x) - r(x)| \leq \varepsilon\}$.

Moreover we have

$$\lambda(\mathcal{L}_n(t)\Delta\mathcal{L}(t)) = \lambda((\mathcal{L}_n(t)\Delta\mathcal{L}(t)) \cap E_{n,\varepsilon}) + \lambda((\mathcal{L}_n(t)\Delta\mathcal{L}(t)) \cap E_{n,\varepsilon}^c).$$

Since $(\mathcal{L}_n(t)\Delta\mathcal{L}(t)) \cap E_{n,\varepsilon} \subset \{t - \varepsilon \leq r \leq t + \varepsilon\}$, we obtain

$$\lambda(\mathcal{L}_n(t)\Delta\mathcal{L}(t)) \leq \lambda(\{t - \varepsilon \leq r \leq t + \varepsilon\}) + \lambda(E_{n,\varepsilon}^c),$$

and

$$\mathbb{E}\lambda(\mathcal{L}_n(t)\Delta\mathcal{L}(t)) \leq \lambda(\{t - \varepsilon \leq r \leq t + \varepsilon\}) + \mathbb{E}\lambda(E_{n,\varepsilon}^c). \quad (1)$$

Finally, as $\sup_{\mathbb{R}^d} |\hat{r}_n - r| \rightarrow 0$ a.s., then $\lambda(E_{n,\varepsilon}^c) \xrightarrow{n \rightarrow \infty} 0$ and, according to assumption **A0**, $\mathbb{E}\lambda(E_{n,\varepsilon}^c) \xrightarrow{n \rightarrow \infty} 0$ by the Lebesgue's dominated convergence theorem. \square

4.2 Proof of Theorem 2.2

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 ([2]), chapter 4 and 5.

4.2.1 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 $nh^{d+4}/\log n \rightarrow 0$, then there exists $\Gamma > 0$ such that*

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

Note that the condition $nh^{d+4}/\log n \rightarrow 0$ is satisfied under the assumptions of Theorem 2.2.

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 $\sup_{\mathcal{L}(t^-)} |f_n - f| \rightarrow 0$ a.s. under the assumptions of Lemma 4.1 (Bosq and Lecoutre [2]), there exists $\theta > 0$ such that $\inf_{\mathcal{L}(t^-)} f_n > \theta$ a.s. for n large enough. So we can write

$$\begin{aligned} \sup_{\mathcal{L}(t^-)} |r_n - r| &= \sup_{\mathcal{L}(t^-)} \left| \frac{\varphi_n - \varphi}{f_n} + r \frac{f_n - f}{f_n} \right| \\ &\leq c \left(\sup_{\mathcal{L}(t^-)} |\varphi_n - \varphi| + \sup_{\mathcal{L}(t^-)} |f_n - f| \right). \end{aligned} \quad (2)$$

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 $B_k = B(x_k, \rho_n)$ ($k = 1, \dots, \ell_n$) of radius ρ_n .

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

$$\begin{aligned} A_n(x, x') &= \frac{1}{n} \sum_{i=1}^n Y_i [K_h(x - X_i) - K_h(x' - X_i)] \\ &\quad - \mathbb{E} \frac{1}{n} \sum_{i=1}^n Y_i [K_h(x - X_i) - K_h(x' - X_i)], \end{aligned}$$

which leads us to

$$\sup_{\mathcal{L}(t^-)} |\varphi_n - \varphi| \leq \sup_{1 \leq k \leq \ell_n} |\varphi_n(x_k) - \mathbb{E}\varphi_n(x_k)| + \sup_{x \in \mathcal{L}(t^-)} |A_n(x, x_k)| + \sup_{\mathcal{L}(t^-)} |\mathbb{E}\varphi_n - \varphi|. \quad (3)$$

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

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

As a consequence, we have

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

One can choose

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

such that

$$\mathbb{P} \left(\sup_{x \in \mathcal{L}(t^-)} |A_n(x, x_k)| > \log n / \sqrt{nh^d} \right) = 0. \quad (4)$$

Then, using the arguments of the proof of Theorem 5.II.3 in [2], we obtain

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

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

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

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

$$\sqrt{nh^d} \mathbb{P} \left(\sup_{1 \leq k \leq l_n} |\varphi_n(x_k) - \mathbb{E}\varphi_n(x_k)| > \varepsilon_0 \sqrt{\frac{\log n}{nh^d}} \right) \leq c n^{1/2+ad-2\varepsilon_0/c} \sqrt{h^d} \quad (5)$$

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 $nh^{d+4}/\log n \rightarrow 0$ we obtain

$$\sqrt{nh^d} \mathbb{P} \left(\sup_{\mathcal{L}(t^-)} |\mathbb{E}\varphi_n - \varphi| \geq \frac{c}{2} \sqrt{\frac{\log n}{nh^d}} \right) \rightarrow 0. \quad (6)$$

From (3) and using (4), (5) and (6) we obtain

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

From (2) and such as $\sup_{\mathcal{L}(t^-)} |f_n - f| \rightarrow 0$ a.s., we conclude the proof. \square

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

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

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

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

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 $\bar{\Phi}(x) = 1 - \Phi(x)$.

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 $i = 1, \dots, n$,

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

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-Essen inequality gives us

$$|\mathbb{P}(r_n(x) < t) - \Phi(t_n(x))| \leq \frac{c}{\sqrt{nV_n(x, t)^3}} \mathbb{E} |(Y - t)K_h(x - X) - \mathbb{E}(Y - t)K_h(x - X)|^3. \quad (7)$$

Finally, under Assumptions **A1** and **A3**, we have (see for example Bosq and Lecoutre [2])

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

and

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

The lemma can then be deduced from (7). \square

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

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

The following result is proven in Cadre [3] (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 [3],

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

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

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

We are now in a position to prove Lemma 4.4 below.

Lemma 4.4 *If $nh^d/(\log n)^7 \rightarrow \infty$ and $nh^{d+4} \log n \rightarrow 0$, then for all $t \in \Theta_0$,*

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

and

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

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

$$E_n \leq c \sqrt{nh^d} \lambda(\mathcal{V}_n^t) \sup_{\mathcal{V}_n^t} |t_n - \bar{t}_n|. \quad (8)$$

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

$$\begin{aligned}
& \frac{1}{\sqrt{nh^d}} |t_n(x) - \bar{t}_n(x)| \\
& \leq |t - r(x)| \left| \frac{f(x)}{\sqrt{\tilde{K}f(x)(v(x) + t^2)}} - \frac{\mathbb{E} f_n(x)}{\sqrt{V_n(x, t)h^{-d}}} \right| \\
& \quad + \sqrt{\frac{h^d}{V_n(x, t)}} \mathbb{E} f_n(x) |r(x) - \tilde{\mathbb{E}} r_n(x)| \\
& \leq \sqrt{\frac{\log n}{nh^d}} \left| \sqrt{\frac{|f(x)V_n(x, t)h^{-d} - (\mathbb{E} f_n(x))^2 \tilde{K}(v(x) + t^2)|}{\tilde{K}(v(x) + t^2)V_n(x, t)h^{-d}}} \right| \\
& \quad + \sqrt{\frac{h^d}{V_n(x, t)}} \mathbb{E} f_n(x) |r(x) - \tilde{\mathbb{E}} r_n(x)| \tag{9}
\end{aligned}$$

Remember that

$$|\tilde{\mathbb{E}} r_n - r| \leq \frac{1}{f_n} |\mathbb{E} \phi_n - \phi| + |r| |\mathbb{E} f_n - f| \tag{10}$$

Since \mathcal{V}_n^t is included in $\mathcal{L}(t^-)$, we can deduce (Bosq and Lecoutre [2]) from **A1**, **A3** and (10) that

$$\sup_{\mathcal{V}_n^t} |\tilde{\mathbb{E}} r_n - r| \leq ch^2. \tag{11}$$

Moreover, if we set

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

we can write

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

Again, since $\mathcal{V}_n^t \subset \mathcal{L}(t^-)$, we can deduce (Bosq and Lecoutre [2]) from **A1** and **A3** that

$$\sup_{x \in \mathcal{V}_n^t} |f(x)V_n(x, t)h^{-d} - (\mathbb{E} f_n(x))^2 \tilde{K}(v(x) + t^2)| \leq ch. \tag{12}$$

We deduce from (9), (11) and (12) that

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

Then, thanks to (8) and since $t \in \Theta_0$, we have for n large enough

$$E_n \leq c \sqrt{\log n} \left(\sqrt{h \log n} + \sqrt{nh^{k+4}} \right). \quad (13)$$

Finally, Lemma 4.2 leads us to

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

which tends to 0 since $\lambda(r^{-1}[t - \varepsilon, t + \varepsilon]) \rightarrow 0$. This and (13) ends the proof. \square

4.2.2 Proof of Theorem 2.2

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

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

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

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

We have

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

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 such that

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

which leads us to

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

and

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

Using the arguments of the proof of Proposition 3.1 in [3], we obtain

$$C_1 \frac{\sqrt{t\tilde{K}}}{\sqrt{2\pi}} \int_{\partial\mathcal{L}(t)} \frac{1}{\|\nabla r\|} \partial\mathcal{H} \leq \underline{\lim}_{n \rightarrow \infty} \sqrt{nh^d} I_n \leq \overline{\lim}_{n \rightarrow \infty} \sqrt{nh^d} I_n \leq C_2 \frac{\sqrt{t\tilde{K}}}{\sqrt{2\pi}} \int_{\partial\mathcal{L}(t)} \frac{1}{\|\nabla r\|} \partial\mathcal{H}.$$

With similar arguments, we have

$$C_1 \sqrt{\frac{t}{2\pi}} \tilde{K} \int_{\partial\mathcal{L}(t)} \frac{d\mathcal{H}}{\|\nabla r\|} \leq \underline{\lim}_{n \rightarrow \infty} \sqrt{nh^d} \bar{I}_n \leq \overline{\lim}_{n \rightarrow \infty} \sqrt{nh^d} \bar{I}_n \leq C_2 \sqrt{\frac{t}{2\pi}} \tilde{K} \int_{\partial\mathcal{L}(t)} \frac{d\mathcal{H}}{\|\nabla r\|}.$$

These inequalities, Lemma 4.4 and Lemma 4.3 give us, for all $t \in \Theta$,

$$\min \left(\underline{\lim}_{n \rightarrow \infty} \sqrt{nh^d} \int_{\mathcal{V}_n^t} \mathbb{P}_{n,t}(x) dx, \underline{\lim}_{n \rightarrow \infty} \sqrt{nh^d} \int_{\bar{\mathcal{V}}_n^t} \bar{\mathbb{P}}_{n,t}(x) dx \right) \geq C_1 \sqrt{\frac{2t}{\pi}} \tilde{K} \int_{\partial\mathcal{L}(t)} \frac{d\mathcal{H}}{\|\nabla r\|},$$

and

$$\max \left(\overline{\lim}_{n \rightarrow \infty} \sqrt{nh^d} \int_{\mathcal{V}_n^t} \mathbb{P}_{n,t}(x) dx, \overline{\lim}_{n \rightarrow \infty} \sqrt{nh^d} \int_{\bar{\mathcal{V}}_n^t} \bar{\mathbb{P}}_{n,t}(x) dx \right) \leq C_2 \sqrt{\frac{2t}{\pi}} \tilde{K} \int_{\partial\mathcal{L}(t)} \frac{d\mathcal{H}}{\|\nabla r\|},$$

and Lemma 4.1 concludes the proof. \square

References

- [1] G. Biau, B. Cadre, and B. Pelletier. A graph-based estimator of the number of clusters. *ESAIM. Probability and Statistics*, 11:272–280, 2007.
- [2] D. Bosq and J. P. Lecoutre. *Théorie de l'Estimation Fonctionnelle*. Ecole Nationale de la Statistique et de l'Administration Economique et Centre d'Etudes des Programmes Economiques. Economica, 1987.
- [3] B. Cadre. Kernel estimation of density level sets. *Journal of Multivariate Analysis*, 97(4):999–1023, 2006.
- [4] L. Cavalier. Nonparametric estimation of regression level sets. *Statistics*, 29(2):131–160, 1997.
- [5] A. Cuevas, M. Febrero, and R. Fraiman. Estimating the number of clusters. *The Canadian Journal of Statistics*, 28(1):367–382, 2000.
- [6] A. Cuevas, M. Febrero, and R. Fraiman. Cluster analysis: a further approach based on density estimation. *Computational Statistics and Data Analysis*, 36(4):441–459, 2001.
- [7] A. Cuevas and R. Fraiman. A plug-in approach to support estimation. *The Annals of Statistics*, 25(6):2300–2312, 1997.
- [8] A. Cuevas, W. González-Manteiga, and A. Rodríguez-Casal. Plug-in estimation of general level sets. *Australian and New Zealand Journal of Statistics*, 48(1):7–19, 2006.
- [9] L. C. Evans and R. F. Gariepy. *Measure Theory and Fine Properties of Functions*. Studies in Advanced Mathematics. CRC Press, 2000.
- [10] J. A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, New York-London-Sydney, 1975. Wiley Series in Probability and Mathematical Statistics.
- [11] J. A. Hartigan. Estimation of a convex density contour in two dimensions. *Journal of the American Statistical Association*, 82(397):267–270, 1987.
- [12] D. Müller. The excess mass approach in statistics. *Beiträge zur Statistik, Universität Heidelberg*, 3, 1993.
- [13] D. W. Müller and G. Sawitzki. Excess mass estimates and tests for multimodality. *Journal of the American Statistical Association*, 86(415):738–746, 1991.
- [14] R. D. Nowak and R. M. Willett. Minimax optimal level-set estimation. *IEEE Transactions on Image Processing*, 16(12):2965–2979, 2007.
- [15] W. Polonik. Measuring mass concentrations and estimating density contour clusters—an excess mass approach. *The Annals of Statistics*, 23(3):855–881, 1995.
- [16] W. Polonik and Z. Wang. Estimation of regression contour clusters: an application of the excess mass approach to regression. *Journal of multivariate analysis*, 94:227–249, 2005.

- [17] B. L. S. Prakasa Rao. *Nonparametric Functional Estimation*. Academic Press, Orlando, 1983.
- [18] C. Scott and M. Davenport. Regression level set estimation via costsensitive classification. *IEEE Transaction on Signal Processing*, 55:2752–2757, 2007.
- [19] A. B. Tsybakov. On nonparametric estimation of density level sets. *The Annals of Statistics*, 25(3):948–969, 1997.
- [20] G. Walther. Granulometric smoothing. *The Annals of Statistics*, 25(6):2273–2299, 1997.