Clustering by estimation of density level sets at a fixed probability

Benoît CADRE[a], Bruno PELLETIER[b,*], and Pierre PUDLO[c]

[a] IRMAR, ENS Cachan Bretagne, CNRS, UEB
Campus de Ker Lann
Avenue Robert Schuman, 35170 Bruz, France
cadre@bretagne.ens-cachan.fr

[b] IRMAR, Université Rennes 2, CNRS, UEB
Campus Villejean
Place du Recteur Henri Le Moal
35043 Rennes Cedex, France
bruno.pelletier@univ-rennes2.fr

[c] I3M, UMR CNRS 5149
Université Montpellier II, CC 051
Place Eugène Bataillon, 34095 Montpellier Cedex 5, France
pudlo@math.univ-montp2.fr

Abstract

In density-based clustering methods, the clusters are defined as the connected components of the upper level sets of the underlying density f. In this setting, the practitioner fixes a probability p, and associates with it a threshold $t^{(p)}$ such that the level set $\{f \geq t^{(p)}\}$ has a probability p with respect to the distribution induced by f. This paper is devoted to the estimation of the threshold $t^{(p)}$, of the level set $\{f \geq t^{(p)}\}$, as well as of the number $k(t^{(p)})$ of connected components of this level set. Given a nonparametric density estimate $\hat{f}_n$ of f based on an i.i.d. n-sample drawn from f, we first propose a computationally simple estimate $t_n^{(p)}$ of $t^{(p)}$, and we establish a concentration inequality for this estimate. Next, we consider the plug-in level set estimate $\{\hat{f}_n \geq t_n^{(p)}\}$, and we establish the exact convergence rate of the Lebesgue measure of the symmetric difference between $\{f \geq t^{(p)}\}$ and $\{\hat{f}_n \geq t_n^{(p)}\}$. Finally, we propose a computationally simple graph-based estimate of $k(t^{(p)})$, which is shown to be consistent. Thus, the methodology yields a complete procedure for analyzing the grouping structure of the data, as p varies over $(0;1)$.

[*] Corresponding author.

Index Terms — Kernel estimate; Density level sets; Nonparametric statistics; Clustering.
AMS 2000 Classification — 62H30, 62H12.

1 Introduction

Cluster analysis encompasses a number of popular statistical techniques aiming at classifying the observations into different groups, called clusters, of similar items; see, e.g., Chapter 10 in Duda et al. (2000), and Chapter 14 in Hastie et al. (2009), for a general exposition on the subject. In general, no prior knowledge on the groups and their number is available, in which case clustering is an unsupervised learning problem. According to Hastie et al. (2009), clustering methods may be categorized in three ensembles, namely combinatorial algorithms, mixture modeling, and mode seekers. The methods proposed and studied in this paper pertain to the third class and rely on the ideas of density-based clustering; see Hartigan (1975).

Let us recall the nonparametric definition of a cluster given by Hartigan (1975). Let X be a $\mathbb{R}^d$-valued random variable with density f. For any $t \geq 0$, denote by $\mathscr{L}(t)$ the t-upper-level set of f, i.e.,

$$\mathscr{L}(t) = \{f \geq t\} = \{x \in \mathbb{R}^d : f(x) \geq t\}. \tag{1.1}$$

According to Hartigan (1975), the clusters are the connected components of $\mathscr{L}(t)$, whence relating population clusters with domains of mass concentration.

Density level sets are therefore the basic objects of Hartigan's approach to the clustering problem. They also play a prominent role in various scientific applications, including anomaly or novelty detection, medical imaging, and computer vision. The theory behind their estimation has developed significantly in the recent years. Excess-mass level set estimates are studied in Hartigan (1987), Muller and Sawitzki (1991), Nolan (1991), Polonik (1995, 1997), Tsybakov (1997). Other popular level set estimates are the plug-in level set estimates, formed by replacing the density f with a density estimate $\hat{f}_n$ in (1.1). Under some assumptions, consistency and rates of convergence (for the volume of the symmetric difference) have been established in Baillo et al. (2000, 2001), Baillo (2003), and an exact convergence rate is obtained in Cadre (2006). Recently, Mason and Polonik (2009) derive the asymptotic normality of the volume of the symmetric difference for kernel plug-in level set estimates; see also related works in Molchanov (1998), Cuevas et al. (2006).

In the context of clustering, algorithms relying on the definition of Hartigan (1975) are typically composed of two main operations. First, observations falling into an estimation of $\mathscr{L}(t)$ are extracted, and next, these extracted observations are partitioned into groups; see, e.g., Cuevas et al. (2000), Biau et al. (2007), and the references therein. However, to interpret the cluster analysis, the extracted set of observations must be related to a probability instead of a threshold of the level set. Such an objective may be reached as follows: given a probability $p \in (0;1)$, define $t^{(p)}$ as the largest threshold such that the probability of $\mathscr{L}(t^{(p)})$ is greater than p, i.e.,

$$t^{(p)} = \sup\big\{t \geq 0 : \mathbb{P}\big(X \in \mathscr{L}(t)\big) \geq p\big\}. \tag{1.2}$$

Note that $\mathbb{P}(X \in \mathscr{L}(t^{(p)})) = p$ whenever $\mathbb{P}(f(X) = t^{(p)}) = 0$. The parameter p has to be understood as a resolution level fixed by the practitioner: if p is close to 1, almost all the sample is in the level set, while if p is small, $\mathscr{L}(t^{(p)})$ is a small domain concentrated around the largest mode of f.

Hence, in a cluster analysis, the practitioner fixes a probability p, depending on the objectives of his study. For a complete study, he needs to estimate, from a set of observations, the threshold $t^{(p)}$, the level set $\mathscr{L}(t^{(p)})$, as well as the number of clusters, i.e. the number of connected components of $\mathscr{L}(t^{(p)})$. Assessing the number of clusters is also a major challenge in cluster analysis, due to its interpretation in terms of population diversity. When a hierarchical cluster analysis is needed, a dendrogram (see, e.g., Hastie et al., 2009, p. 521) may be produced by varying the value of p over $(0,1)$. The aim of this paper is to address these estimation problems, given a set of i.i.d. observations $X_1, \cdots, X_n$ drawn from f.

Estimation of $t^{(p)}$ and $\mathscr{L}(t^{(p)})$. In Cadre (2006), a consistent estimate of $t^{(p)}$ is defined as a solution in t of the equation

$$\int_{\{\hat{f}_n \geq t\}} \hat{f}_n(x)\mathrm{d}x = p, \tag{1.3}$$

where $\hat{f}_n$ is a nonparametric density estimate of f based on the observations $X_1, \cdots, X_n$. In practice, though, computing such an estimate would require multiple evaluations of integrals, yielding a time-consuming procedure. Following an idea that goes back to Hyndman (1996), we propose to consider the estimate $t_n^{(p)}$

defined as the $(1-p)$-quantile of the empirical distribution of $\hat{f}_n(X_1),\ldots,\hat{f}_n(X_n)$. Such an estimate may be easily computed using an order statistic. We first establish a concentration inequality for $t_n^{(p)}$, depending on the supremum norm of $\hat{f}_n - f$ (Theorem 2.1). Next we specialize to the case where $\hat{f}_n$ is a nonparametric kernel density estimate, and we consider the plug-in level set estimate $\mathscr{L}_n(t_n^{(p)})$ defined by

$$\mathscr{L}_n(t_n^{(p)}) = \{\hat{f}_n \geq t_n^{(p)}\}.$$

The distance between two Borel sets in $\mathbb{R}^d$ is defined as the Lebesgue measure λ of the symmetric difference denoted Δ (i.e., $A\Delta B = (A\cap B^c)\cup(A^c\cap B)$ for all sets A,B). Our second result (Theorem 2.3) states that, under suitable conditions, $\mathscr{L}_n(t_n^{(p)})$ is consistent in the sense that

$$\sqrt{nh_n^d}\,\lambda\left(\mathscr{L}_n(t_n^{(p)})\Delta\mathscr{L}(t^{(p)})\right) \xrightarrow{\mathbb{P}} C_f^{(p)},$$

where $C_f^{(p)}$ is an explicit constant depending on f and p, and which can be consistently estimated. An analogous result is obtained in Corollary 2.1 in Cadre (2006) by using the threshold estimate (1.3). Note that the two threshold estimates are defined differently: the first one is defined as a solution of the integral equation (1.3), and our estimate is defined as an empirical quantile of a non-independent sequence of random variables. As a consequence, establishing Theorem 2.3 requires different arguments developed in the proof exposed in Section 6.

Estimation of the number of clusters. Then, we consider the estimation of the number of clusters of $\mathscr{L}(t^{(p)})$. A theoretical estimator could be defined as the number of connected components of the plug-in level set estimate $\mathscr{L}_n(t_n^{(p)})$, for any estimate $t_n^{(p)}$ of $t^{(p)}$. However, heavy numerical computations are required to evaluate this number in practice, especially when the dimension d is large. For this reason, stability criterions with respect to resampling, or small perturbations of the data set, are frequently employed in practice, despite the negative results of Ben-David et al. (2006) and Ben-David et al. (2007). The approach developed in Biau et al. (2007) and summarized below is based on a graph and leads to a dramatic decrease of the computational burden; see also Ben-David et al. (2006). In Biau et al. (2007), the threshold $t>0$ is fixed. Set $(r_n)_n$ a sequence of positive numbers, and define the graph $\mathscr{G}_n(t)$ whose vertices are the observations X_i for which $\hat{f}_n(X_i)\geq t$, and where two vertices are connected by an edge whenever they are at a distance no more than r_n. Biau et al. (2007) prove that, with probability one, the graph $\mathscr{G}_n(t)$ and the set $\mathscr{L}(t)$ have the same number of connected components, provided n is large enough. Hence the number of connected components, say $k_n(t)$, of $\mathscr{G}_n(t)$ is a strongly consistent estimate of the number of connected components $k(t)$ of $\mathscr{L}(t)$.

In practice, however, only the probability p is fixed, hence the threshold defined by (1.2) is unknown. Moreover, in the above-mentioned paper, the behavior of $k_n(t)$ depends on the behavior of the gradient of $\hat{f}_n$; when $\hat{f}_n$ is a kernel density estimate for instance, this leads to restrictive conditions on the bandwidth sequence. In comparison with Biau et al. (2007), one can sum up our contribution (Theorem 3.1) as follows: only the probability p is fixed, and the associated threshold is estimated, leading to an efficient and tractable method for clustering. Moreover, the concentration inequality for the estimator is obtained whatever the behavior of the gradient of $\hat{f}_n$, hence a better inequality.

The paper is organized as follows. Section 2 is devoted to the estimation of the threshold $t^{(p)}$ and the level set $\mathscr{L}(t^{(p)})$. In Section 3, we study the estimator of the number of clusters of $\mathscr{L}(t^{(p)})$. The methodology is illustrated in Section 4 on a simulated data set and applied next to a real data set. Section 5, Section 6, and Section 7 are devoted to the proofs. Finally, several auxiliary results for the proofs are postponed in the Appendices, at the end of the paper.

2 Level set and threshold estimation

2.1 Notations

Let $\hat{f}_n$ be an arbitrary nonparametric density estimate of f. For $t \geq 0$, the t-upper level sets of f and $\hat{f}_n$ will be denoted by $\mathscr{L}(t)$ and $\mathscr{L}_n(t)$ respectively, i.e.,

$$\mathscr{L}(t) = \{f \geq t\}, \quad \text{and} \quad \mathscr{L}_n(t) = \{\hat{f}_n \geq t\}.$$

Given a real number p in $(0;1)$, our first objective is to estimate a level $t^{(p)} \in \mathbb{R}$ such that $\mathscr{L}(t^{(p)})$ has μ-coverage equal to p, where μ is the law of X. To this aim, let H and H_n be the functions defined for all $t \geq 0$ respectively by

$$H(t) = \mathbb{P}(f(X) \leq t), \quad H_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\hat{f}_n(X_i) \leq t\}.$$

Next, for all $p \in (0;1)$, we define the $(1-p)$-quantile of the law of $f(X)$, i.e.

$$t^{(p)} = \inf\{t \in \mathbb{R} : H(t) \geq 1-p\}, \tag{2.1}$$

and its estimate based on the sample $\hat{f}_n(X_1), \cdots, \hat{f}(X_n)$:

$$t_n^{(p)} = \inf\{t \in \mathbb{R} : H_n(t) \geq 1 - p\}. \tag{2.2}$$

In comparison with the estimator of $t^{(p)}$ defined as a solution of (1.3), the estimate $t_n^{(p)}$ is easily computed, by considering the order statistic induced by the sample $\hat{f}_n(X_1)$,..., $\hat{f}_n(X_n)$. Moreover, note that the set of discontinuities for H is at most countable, and that whenever H is continuous at $t^{(p)}$, the two definitions (1.2) and (2.1) coincide. In this case, we have $\mu(\mathscr{L}(t^{(p)})) = p$. We shall consider $\mathscr{L}_n(t_n^{(p)})$ as an estimate of $\mathscr{L}(t^{(p)})$.

Whenever f is of class C^1, we let $\mathscr{T}_0$ be the subset of the range of f defined by

$$\mathscr{T}_0 = \left\{ t \in (0; \sup_{\mathbb{R}^d} f) : \inf_{\{f=t\}} \|\nabla f\| = 0 \right\}.$$

This set naturally arises when considering the distribution of $f(X)$. Indeed, the Implicit Function Theorem implies that $\mathscr{T}_0$ contains the set of points in $(0; \sup_{\mathbb{R}^d} f)$ which charges the distribution of $f(X)$. We shall assume throughout that the density f satisfies the following conditions.

Assumption 1 [on f]

(i) The density f is of class C^2 with a bounded hessian matrix, and $f(x) \to 0$ as $\|x\| \to \infty$.

(ii) $\mathscr{T}_0$ has Lebesgue content 0.

(iii) $\lambda(\{f = t\}) = 0$ for all $t > 0$.

Assumptions 1-*(ii)* and 1-*(iii)* are essentially imposed for the sake of the simplicity of the exposition, allowing the main results to be stated for almost all $p \in (0; 1)$.

By Assumption 1-*(i)*, the upper t-level set $\mathscr{L}(t)$ is compact for all $t > 0$, as well as its boundary $\{f = t\}$. Assumption 1-*(iii)*, which ensures the continuity of H, roughly means that each flat part of f has a null volume. Moreover, it is proved in Lemma A.1 that under Assumption 1-*(i)*, we have $\mathscr{T}_0 = f(\mathscr{X}) \setminus \{0; \sup_{\mathbb{R}^d} f\}$, where $\mathscr{X} = \{\nabla f = 0\}$ is the set of critical points of f. Suppose in addition that f

is of class C^k, with $k \geq d$. Then, Sard's Theorem (see, e.g., Aubin, 2000) ensures that the Lebesgue measure of $f(\mathscr{X})$ is 0, hence implying Assumption 1-*(ii)*.

Let us introduce some additional notations. We let $\|.\|_2$ and $\|.\|_\infty$ be the $L^2(\lambda)$- and $L^\infty(\lambda)$-norms on functions respectively, and $\|.\|$ be the the usual Euclidean norm. At last, $\mathscr{H}$ stands for the $(d-1)$-dimensional Hausdorff measure (see, e.g., Evans and Gariepy, 1992). Recall that $\mathscr{H}$ agrees with ordinary $(d-1)$-dimensional surface area on nice sets.

The next subsection is devoted to the study of the asymptotic behavior of $t_n^{(p)}$ and $\mathscr{L}_n(t_n^{(p)})$, when $t_n^{(p)}$ is defined by (2.2). The case of an arbitrary density estimate $\hat{f}_n$ is considered first. Next, we specialize the result in the case where $\hat{f}_n$ is a kernel density estimator.

2.2 Asymptotic behavior of $t_n^{(p)}$

Our first result provides a concentration inequality for $t_n^{(p)}$ defined by (2.2) when $\hat{f}_n$ is an arbitrary density estimate.

Theorem 2.1. *Suppose that f satisfies Assumption 1. Then, for almost all $p \in (0;1)$ and for all $\eta > 0$, we have*

$$\mathbb{P}\left(|t_n^{(p)} - t^{(p)}| \geq \eta\right) \leq \mathbb{P}\left(\|\hat{f}_n - f\|_\infty \geq C_1\eta\right) + C_2 n^2 \exp\left(-nC_1\eta^2\right),$$

where C_1 and C_2 are positive constants.

We now specialize the above result in the case where $\hat{f}_n$ is a nonparametric kernel density estimate of f with kernel K and bandwidth sequence $(h_n)_n$, namely

$$\hat{f}_n(x) = \frac{1}{nh_n^d}\sum_{i=1}^n K\left(\frac{x - X_i}{h_n}\right). \tag{2.3}$$

The following assumptions on h_n and K will be needed in the sequel.

Assumption 2a [on h_n]

$$\frac{nh_n^d}{\log n} \to \infty, \quad \text{and} \quad nh_n^{d+4} \to 0.$$

Assumption 3 [on K]
The kernel K is a density on $\mathbb{R}^d$ with radial symmetry:

$$K(x) = \Phi\left(\|x\|\right),$$

where $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ is a decreasing function with compact support.

Under Assumption 3 the class of functions

$$\left\{K\left(\frac{x-.}{h}\right) \,:\, h > 0\,;\, x \in \mathbb{R}^d\right\}$$

has a polynomial discrimination (see, e.g., Pollard, 1984, Problem II.28, p. 42). Then, sharp almost-sure convergence rates on $\hat{f}_n - f$ can be established (see, e.g., Giné and Guillou, 2002, Einmahl and Mason, 2005). More precisely, since $\|\mathbb{E}\hat{f}_n - f\|_\infty = O(h_n^2)$ under Assumption 1-*(i)*, one deduces from the above-mentioned papers that, if Assumptions 2a and 3 also hold, then for all $\eta > 0$,

$$\sum_n \mathbb{P}\big(v_n \|\hat{f}_n - f\|_\infty \geq \eta\big) < \infty, \tag{2.4}$$

where $(v_n)_n$ is any sequence satisfying $v_n = o(\sqrt{nh_n^d/\log n})$. Combined with the concentration inequality in Theorem 2.1, we obtain the following corollary.

Corollary 2.2. *Suppose that f satisfies Assumption 1. Let $\hat{f}_n$ be the nonparametric kernel density estimate (2.3) satisfying Assumptions 2a and 3. Then, for almost all $p \in (0;1)$, we have*

$$\frac{\sqrt{nh_n^d}}{\log n}\left|t_n^{(p)} - t^{(p)}\right| \xrightarrow{\text{a.s}} 0.$$

Even if the above result is non-optimal, it turns out to be enough for a cluster analysis, as showed in the next section.

2.3 Asymptotic behavior of $\mathscr{L}_n(t_n^{(p)})$

We shall need a slightly stronger assumption than Assumption 2a on the bandwidth sequence $(h_n)_n$.

Assumption 2b [on h_n]

$$\frac{nh_n^d}{(\log n)^{16}} \to \infty, \quad \text{and} \quad nh_n^{d+4}(\log n)^2 \to 0.$$

Under this set of conditions on the bandwidth sequence, one may apply the main result in Cadre (2006).

The next result is an equivalent to Corollary 2.1 of Cadre (2006), in which the estimate of $t^{(p)}$ is defined as a solution of (1.3). It shows that $\mathscr{L}_n(t_n^{(p)})$ is consistent for the volume of the symmetric difference. Hence, this estimate can be used as a reliable basis for performing a cluster analysis in practice.

Theorem 2.3. *Suppose that f satisfies Assumption 1 and that $d \geq 2$. Let $\hat{f}_n$ be the nonparametric kernel density estimate (2.3) satisfying Assumptions 2b and 3. Then, for almost all $p \in (0;1)$, we have*

$$\sqrt{nh_n^d}\,\lambda\left(\mathscr{L}_n(t_n^{(p)})\Delta\mathscr{L}(t^{(p)})\right) \xrightarrow{\mathbb{P}} \sqrt{\frac{2}{\pi}}\,\|K\|_2\,t^{(p)}\int_{\{f=t^{(p)}\}}\frac{1}{\|\nabla f\|}\mathrm{d}\mathscr{H}.$$

The deterministic limit in the above theorem depends on the unknown density f. However, one can prove that if $(\alpha_n)_n$ is a sequence of positive numbers tending to 0 and such that $\alpha_n^2 nh_n^d/(\log n)^2 \to \infty$, then, for almost all $p \in (0;1)$,

$$\frac{t_n^{(p)}}{\alpha_n}\lambda\left(\mathscr{L}_n(t_n^{(p)})\setminus\mathscr{L}_n(t_n^{(p)}+\alpha_n)\right) \xrightarrow{\mathbb{P}} t^{(p)}\int_{\{f=t^{(p)}\}}\frac{1}{\|\nabla f\|}\mathrm{d}\mathscr{H}.$$

The proof of the above result is similar to the one of Lemma 4.6 in Cadre (2006), using our Corollary 2.2. Combined with Theorem 2.3, we then have, for almost all $p \in (0;1)$,

$$\frac{\alpha_n\sqrt{nh_n^d}}{t_n^{(p)}\lambda\left(\mathscr{L}_n(t_n^{(p)})\setminus\mathscr{L}_n(t_n^{(p)}+\alpha_n)\right)}\lambda\left(\mathscr{L}_n(t_n^{(p)})\Delta\mathscr{L}(t^{(p)})\right) \xrightarrow{\mathbb{P}} \sqrt{\frac{2}{\pi}}\|K\|_2,$$

which yields a feasible way to estimate $\lambda(\mathscr{L}_n(t_n^{(p)})\Delta\mathscr{L}(t^{(p)}))$.

Remark 2.4. *According to Proposition A.2 in Appendix A, on any interval $I \subset (0;\sup_{\mathbb{R}^d} f)$ with $I \cap \mathscr{T}_0 = \emptyset$, the random variable $f(X)$ has a density on I, which is given by*

$$g(t) = t\int_{\{f=t\}}\frac{1}{\|\nabla f\|}\mathrm{d}\mathscr{H}, \quad t \in I.$$

Thus the normalized distance between $\mathscr{L}_n(t_n^{(p)})$ and $\mathscr{L}(t^{(p)})$ in Theorem 2.3 corresponds to the density g at point $t^{(p)}$, up to a multiplicative constant.

3 Estimation of the number of clusters

3.1 Notations

As in the previous section, let us start with an arbitrary nonparametric density estimate $\hat{f}_n$ of f. We first recall the methodology developed by Biau et al. (2007) to estimate the number of clusters $k(t)$ of $\mathscr{L}(t)$. Set

$$J_n(t) = \left\{ j \leq n : \hat{f}_n(X_j) \geq t \right\},$$

i.e., $\{X_j \, : \, j \in J_n(t)\}$ is the part of the n-sample lying in the t-level set of $\hat{f}_n$. Let $(r_n)_n$ be a sequence of positive numbers vanishing as $n \to \infty$. Define the graph $\mathscr{G}_n(t)$ with vertices $\{X_j \, : \, j \in J_n(t)\}$ and where, for $i, j \in J_n(t)$, X_i and X_j are joined by an edge if and only if $\|X_i - X_j\| \leq r_n$. Then we set $k_n(t)$ as the number of connected components of the graph $\mathscr{G}_n(t)$.

Under suitable assumptions, Biau et al. (2007) prove that, with probability one, $k_n(t) = k(t)$ provided n is large enough. In our setting however, the threshold $t^{(p)}$ is unknown and has to be estimated. Hence, the main result in Biau et al. (2007) may not be applied in order to estimate the number of clusters $k(t^{(p)})$.

Let $t_n^{(p)}$ be an arbitrary estimator of $t^{(p)}$. In the next subsection, we state a concentration inequality for $k_n(t_n^{(p)})$. Then, we specialize this result to the case where $\hat{f}_n$ is the kernel estimate (2.3) and $t_n^{(p)}$ is given by (2.2).

3.2 Asymptotic behavior of $k_n(t_n^{(p)})$

In what follows, ω_d denotes the volume of the Euclidean unit ball in $\mathbb{R}^d$.

Theorem 3.1. *Suppose that f satisfies Assumption 1. Let $(\varepsilon_n)_n$ and $(\varepsilon'_n)_n$ be two sequences of positive numbers such that $\varepsilon_n + \varepsilon'_n = o(r_n)$. For almost all $p \in (0;1)$, there exists a positive constant C, depending only on f and p, such that, if n is large enough,*

$$\begin{aligned} \mathbb{P}\big(k_n(t_n^{(p)}) \neq k(t^{(p)})\big) \quad \leq \quad & 2\,\mathbb{P}\big(\|\hat{f}_n - f\|_\infty > \varepsilon_n\big) + 2\,\mathbb{P}\big(|t_n^{(p)} - t^{(p)}| > \varepsilon'_n\big) \\ & + C r_n^{-d} \exp\left(-t^{(p)} \frac{\omega_d}{4^{d+1}} n r_n^d\right). \end{aligned}$$

In comparison with the result in Biau et al. (2007) for a fixed threshold, the above concentration inequality does not require any assumption on the gradient of $\hat{f}_n$. As a consequence, when $\hat{f}_n$ is a nonparametric kernel estimate for instance, the conditions imposed on the bandwidth are less restrictive.

Now consider the particular case where $\hat{f}_n$ is defined by (2.3) and $t_n^{(p)}$ is defined by (2.2). Letting $(v_n)_n$ be a sequence such that $v_n = o(\sqrt{nh_n^d/\log n})$, and choosing the sequences $(r_n)_n$, $(\varepsilon'_n)_n$ and $(\varepsilon_n)_n$ in Theorem 3.1 so that $\varepsilon_n = \varepsilon'_n = 1/v_n$ and $v_n r_n \to \infty$, we deduce the following from Theorem 3.1, Theorem 2.1 and (2.4).

Corollary 3.2. *Suppose that f satisfies Assumption 1. Let $\hat{f}_n$ be the kernel density estimate* (2.3) *satisfying Assumptions 2a and 3, and let $t_n^{(p)}$ be the estimate of $t^{(p)}$ defined by* (2.2)*. Then, for almost all $p \in (0;1)$, we have almost surely*

$$k_n(t_n^{(p)}) = k(t^{(p)}),$$

provided n is large enough.

4 Numerical illustrations

4.1 A simulated example: two-dimensional Gaussian mixture

To illustrate the practical implementation of our proposed estimators, we consider a Gaussian mixture model inspired from Baudry et al. (2010) with six components on $\mathbb{R}^2$. The underlying density f is expressed as

$$f(x) = \sum_{\ell=1}^{6} p_\ell \varphi(x, \mu_\ell, \Sigma_\ell), \tag{4.1}$$

where $\varphi(\cdot, \mu, \Sigma)$ denotes the Gaussian density with mean μ and covariance matrix Σ. The mixture coefficients p_ℓ are taken as $p_1 = p_2 = p_3 = p_4 = 1/5$, and $p_5 = p_6 = 1/10$, and the parameters of the Gaussian components are specified as follows: $\mu_1 = (-0.3; -0.3)$, $\mu_2 = (3.0; 3.0)$, $\mu_3 = \mu_4 = (0; 3.0)$, $\mu_5 = \mu_6 = (3.0; 0)$,

$$\Sigma_1 = \begin{pmatrix} 0.39 & -0.28 \\ -0.28 & 0.39 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 0.36 & 0.30 \\ 0.30 & 0.36 \end{pmatrix},$$

$$\Sigma_3 = \Sigma_5 = \begin{pmatrix} 0.33 & 0 \\ 0 & 0.01 \end{pmatrix}, \quad \Sigma_4 = \Sigma_6 = \begin{pmatrix} 0.01 & 0 \\ 0 & 0.33 \end{pmatrix}.$$

A data set of size $n = 600$ has been simulated according to model (4.1). A scatterplot of the simulated data is represented in Figure 1 together with level curves (contours) of the underlying density f. The data points aggregate into 4 groups for a large range of threshold values: two ellipsoids originating from the 1^{st} and 2^{nd} components, and two crosses which result from the combination of the 3^{th} and 4^{th} components (upper left cross) and from the combination of the 5^{th} and 6^{th} components (lower right cross). More precisely, using Monte Carlo simulations, we evaluated numerically the theoretical values of the threshold $t^{(p)}$ and of the number of connected components $k(t^{(p)})$ of the level set, for various probability numbers p. The results of these computations are reported in the column labeled "Asymptotics" of Table 1. For probability levels greater than about 50.3%, the theoretical level set has 4 connected components.

Parameter calibration. Our proposed clustering algorithm depends on three parameters: the probability p of the level set, the bandwidth h_n of the kernel density estimate, and the connectivity radius r_n of the neighborhood graph.

The parameter p must be chosen in advance by the practitioner. Recall that p represents the resolution level of the clustering analysis, and that by definition of the threshold estimate $t_n^{(p)}$, p also represents the proportion of the data points that will be extracted for further partitioning. For instance, values of p as large as 90% will lead to a clustering of most of the data set, while values of p on the order of 50% may be selected for identifying regions of high density in the data set.

Values for the bandwidth h_n of the kernel density estimate may be determined using any standard bandwidth selection procedure. There exists an important literature on this subject which we do not review in the present paper. In these experiments, we employed a classical cross-validation bandwidth selection method, as implemented in nonparametric codes for the `R` statistical software.

For the selection of the connectivity radius r_n, note that for Theorem 3.1 and Corollary 3.2 to hold, it is not necessary that r_n goes to 0 as $n \to \infty$. Indeed, it is sufficient that r_n be smaller than the minimum distance between any two distinct connected components of the level set $\mathscr{L}(t^{(p)})$ for all n large enough. That said, if r_n is taken as a sequence decreasing to 0, then r_n must not decrease faster than the oscillations of $\|\hat{f}_n - f\|_\infty$; see the conditions and the inequality in Theorem 3.1. Hence in practice, appropriate values of r_n must not be too large to prevent undesired connections between connected components, and not too small to gurantee the connectivity of the graph, i.e., that its topology coincides with the one of the level set. Based on these observations, we propose the following heuristic to automatically select r_n. For each observation X_j in the extracted data set, we first compute the distance to its M^{th} nearest neighbor among the extracted data, for some integer M. Then we select r_n as the maximal value of these distances, which implies that each extracted point is connected to at least M other points in the graph. Alternatively, one may also consider for r_n an empirical quantile of these distances corresponding to an order π close to 1. In these simulations, the choice of $M = 10$ proved satisfactory and a correct estimation of the number of clusters has been obtained for any quantile order $\pi \geq 0.96$.

Results. The methodology has been applied on 40 independent simulated data sets, each of size $n = 600$ and drawn from the model (4.1). Two types of kernels have been considered for density estimation: an Epanechnikov kernel, and a Gaussian kernel. The Epanechnikov kernel satisfies Assumption 3 [on K] which requires the kernel to have compact support, but a Gaussian kernel is considered for the purpose of comparison. The results are exposed in Table 1. As discussed above, the column "Asymptotics" provide the numerical evaluation of the theoretical threshold $t^{(p)}$ and of the number of connected components $k(t^{(p)})$ for various probability numbers p. The last four columns of Table 1 give statistics computed from the 40 data sets: the mean and standard deviation of the threshold estimate $t_n^{(p)}$, and the most frequent value of the estimate of the number of clusters $k_n(t_n^{(p)})$, with the frequency of the most frequent value.

For values of p greater than about 50%, the estimations of $t^{(p)}$ with the Epanechnikov and Gaussian kernels are comparable and accurate. For smaller values of p, $t^{(p)}$ tends to be under-estimated, especially when using the Epanechnikov kernel. This is explained by a failure of the automatic cross-validation bandwidth selection procedure with an Epanechnikov kernel, which has led to an over-smoothed density estimate. On the other hand, the data points extracted using these two density estimates are almost identical. Recall that an observation X_i is extracted whenever $\hat{f}_n(X_i)$ is greater than $t_n^{(p)}$, and that $t_n^{(p)}$ is obtained from the empirical distribution of the $\hat{f}_n(X_i)$'s. Hence two density estimates may lead to somewhat different estimations of $t^{(p)}$ while still yielding similar extracted observations. In both cases (Epanechnikov and Gaussian kernels) the estimations of the number of clusters are comparable and accurate. One example of the resulting partition of a data set of size $n = 600$ is represented in Figure 2, using the Epanechnikov kernel (left panel) and the Gaussian kernel (right panel). In both cases, the probability level has been taken as $p = 90\%$, and the automatic selection of r_n with $M = 10$ neighbors led to the formation of 4 groups.

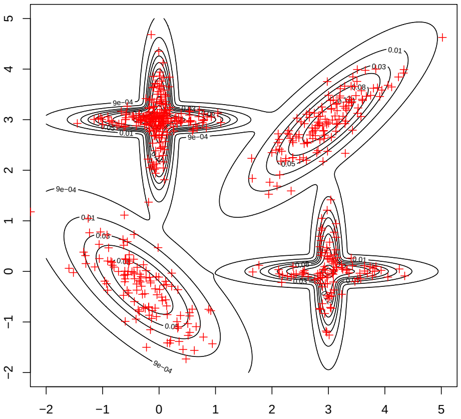

Figure 1: Scatterplot of a data set of size $n = 600$ drawn from the density f defined in (4.1) and contour curves of f. The observations aggregate into 4 groups for a large range of threshold values of the level set.

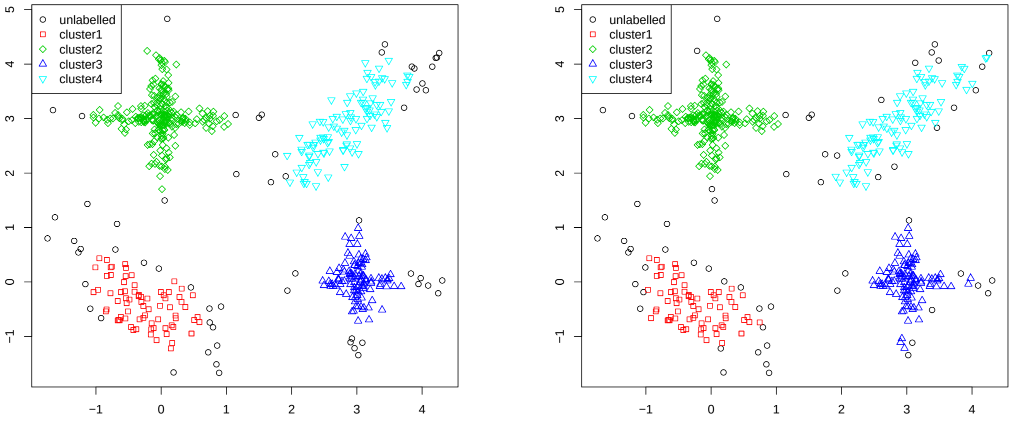

Figure 2: Clustering result using the Epanechnikov (left) and Gaussian (right) kernel density estimates at a probability level $p = 90\%$.

Table 1: Summary statistics for the estimation of the threshold $t^{(p)}$ and of the number of connected components $k(t^{(p)})$ using a kernel density estimate with an Epanechnikov and a Gaussian kernel. The theoretical values of $t^{(p)}$ and $k(t^{(p)})$, evaluated numerically using Monte Carlo simulations, are reported in the column "Asymptotics" for various values of p. The stars *** indicate the values of p where the function $p \mapsto k(t^{(p)})$ presents a jump. The following statistics are formed based on 40 independent data sets, each of size $n = 600$: the mean and standard deviation (SD) of the threshold estimate $t_n^{(p)}$, the most frequent value of $k_n(t_n^{(p)})$ (majority) and its corresponding frequency (vote).

	Asymptotics		**Epanechnikov kernel**			
p	$t^{(p)}$	$k(t^{(p)})$	$t_n^{(p)}$		$k_n(t_n^{(p)})$	
			mean	SD	majority	(vote)
0.900	0.0236	4	0.0250	0.00426	4	(90.0%)
0.800	0.0475	4	0.0407	0.00690	4	(87.5%)
0.503	0.1180	***	0.0784	0.01470	4	(47.5%)
0.450	0.1388	3	0.0868	0.01712	3	(57.5%)
0.396	0.1611	***	0.1006	0.01985	2	(47.5%)
0.200	0.3616	2	0.1891	0.04552	1	(92.5%)
0.085	0.5574	***	0.2678	0.08164	1	(100.0%)
0.040	0.8013	1	0.2835	0.08951	1	(100.0%)
0.000	1.1082		0.2929	0.09463		

	Asymptotics		**Gaussian kernel**			
p	$t^{(p)}$	$k(t^{(p)})$	$t_n^{(p)}$		$k_n(t_n^{(p)})$	
			mean	SD	majority	(vote)
0.900	0.0236	4	0.0366	0.00319	4	(82.5%)
0.800	0.0475	4	0.0576	0.00443	4	(80.0%)
0.503	0.1180	***	0.1192	0.00894	4	(67.5%)
0.450	0.1388	3	0.1354	0.01110	3	(35.0%)
0.396	0.1611	***	0.1565	0.01378	2	(57.5%)
0.200	0.3616	2	0.2865	0.03033	2	(62.5%)
0.085	0.5574	***	0.4676	0.05290	1	(100.0%)
0.040	0.8013	1	0.5527	0.05433	1	(100.0%)
0.000	1.1082		0.6169	0.05762		

4.2 Real dataset

In this section, we consider the classical Old Faithful Geyser dataset available from the `R` software which contains 272 observations of the waiting time between eruptions and the duration of the eruption; see Azzalini and Bowman (1990). For the analysis, the two variables have been first standardized to null mean and unit variance. The joint density has been estimated by a nonparametric kernel density estimate with a Gaussian kernel and automatic bandwidth selection, as described above.

The probability level of the analysis has been fixed at $p = 99\%$, which yields a threshold estimate of $t_n^p = 0.029$. The data points together with the contour curves of level t_n^p are represented in Figure 3. For this resolution level of 99%, all but 3 data points belong to the estimated level set.

Next, we constructed a neighborhood graph on the extracted data points. The connectivity radius of the graph has been selected as $r_n = 0.35$ through the distribution of the ten nearest neighbors of each point in the extracted dataset.

The graph on the extracted points is displayed in the left panel of Figure 4. The graph has two connected components, indicating two groups at the resolution level of $p = 99\%$. The results of the clustering are represented on the right panel of Figure 4, revealing the two distinct groups.

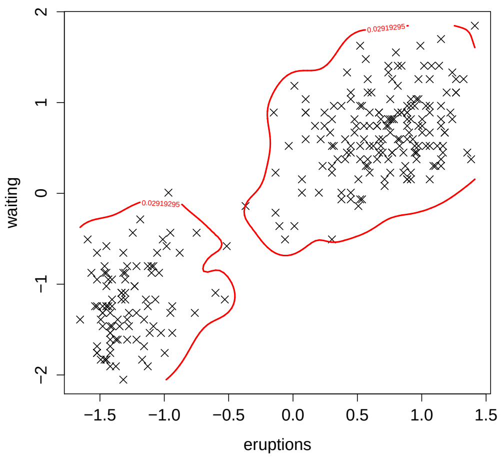

Figure 3: Data points and contour curves of level $t^p = 0.029$ corresponding to a probability level of 99%.

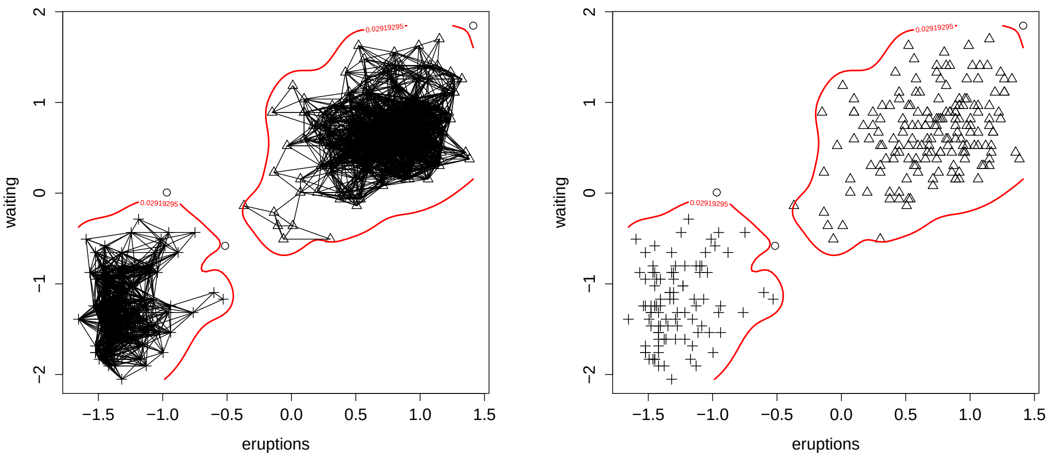

Figure 4: Neighborhood graph on the extracted points (left) and result of the clustering (right).

5 Proof of Theorem 2.1: convergence of $t_n^{(p)}$

5.1 Auxiliary results

We shall assume throughout this subsection that Assumptions 1, 2a and 3 hold.

First note that under Assumption 1, H is a bijection from $(0;\sup_{\mathbb{R}^d} f)$ to $(0;1)$. Indeed, Assumption 1-*(iii)* implies that H is a continuous function. Moreover, under Assumption 1-*(i)*, H is increasing: for suppose it were not, then for some $t \geq 0$ and some $\varepsilon > 0$,

$$0 = H(t+\varepsilon) - H(t) = \int_{\{t<f\leq t+\varepsilon\}} f \mathrm{d}\lambda,$$

which is impossible, because $\lambda(\{t < f < t+\varepsilon\}) > 0$. Then we denote by G the inverse of H restricted to $(0;\sup_{\mathbb{R}^d} f)$.

Lemma 5.1. *The function G is almost everywhere differentiable.*

Proof. As stated above, H is increasing. Hence, by the Lebesgue derivation Theorem, for almost all t, H is differentiable with derivative $H'(t) > 0$. Thus, G is almost everywhere differentiable. □

The Levy metric $d_{\mathscr{L}}$ between any real-valued functions φ_1, φ_2 on $\mathbb{R}$ is defined by

$$d_{\mathscr{L}}(\varphi_1,\varphi_2) = \inf\left\{\theta > 0 : \forall x \in \mathbb{R}, \varphi_1(x-\theta) - \theta \leq \varphi_2(x) \leq \varphi_1(x+\theta) + \theta\right\},$$

(see, e.g., Billingsley, 1995, 14.5). Recall that convergence in distribution is equivalent to convergence of the underlying distribution functions for the metric $d_{\mathscr{L}}$.

Lemma 5.2. *Let x_0 be a real number, and let φ_1 be an increasing function with a derivative at point x_0. There exists $C > 0$ such that, for any increasing function φ_2 with $d_{\mathscr{L}}(\varphi_1,\varphi_2) \leq 1$,*

$$|\varphi_1(x_0) - \varphi_2(x_0)| \leq C d_{\mathscr{L}}(\varphi_1,\varphi_2).$$

Proof. Let θ be any positive number such that, for all $x \in \mathbb{R}$,

$$\varphi_1(x-\theta) - \theta \leq \varphi_2(x) \leq \varphi_1(x+\theta) + \theta. \tag{5.1}$$

Since φ_1 is differentiable at x_0,

$$\varphi_1(x_0 \pm \theta) = \varphi_1(x_0) \pm \theta \varphi_1'(x_0) + \theta \psi_{\pm}(\theta) \tag{5.2}$$

where each function $\psi_{\pm}$ satisfies $\psi_{\pm}(\theta) \to 0$ when $\theta \to 0^+$. Using (5.1) and (5.2), we obtain

$$-\theta(\varphi_1'(x_0)+1) + \theta\psi_-(\theta) \leq \varphi_2(x_0) - \varphi_1(x_0) \leq \theta(\varphi_1'(x_0)+1) + \theta\psi_+(\theta).$$

Taking the infimum over θ satisfying (5.1) gives the announced result with any C such that, for all $\delta \leq 1$,

$$\left|\varphi_1'(x_0)+1\right| + \max\left(|\psi_-(\delta)|, |\psi_+(\delta)|\right) \leq C. \qquad \square$$

Let $\mathscr{L}^\ell(t)$ denote the lower t-level set of the unknown density f, i.e., $\mathscr{L}^\ell(t) = \{x \in \mathbb{R}^d : f(x) \leq t\}$. Moreover, we set

$$V_n = \sup_{t \geq 0} \left| \mu_n\left(\mathscr{L}^\ell(t)\right) - \mu\left(\mathscr{L}^\ell(t)\right) \right|, \tag{5.3}$$

where $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ is the empirical measure indexed by the sample, δ_x denoting the Dirac measure at point x. The next lemma borrows elements from the Vapnik-Chervonenkis theory; we refer the reader to Devroye et al. (1996) for materials on the subject.

Lemma 5.3. *There exists a constant C such that, for all $\eta > 0$, we have*

$$\mathbb{P}(V_n \geq \eta) \leq Cn^2 \exp\left(-n\eta^2/32\right).$$

Proof. Let $\mathscr{A}$ be the collection of lower level sets, namely

$$\mathscr{A} = \{\mathscr{L}^\ell(t), t \geq 0\}.$$

Observe that the Vapnik-Chervonenkis dimension of $\mathscr{A}$ is 2. Then, by the Vapnik-Chervonenkis inequality (see, e.g., Devroye et al., 1996, Theorem 12.5), we obtain the stated result. □

5.2 Proof of Theorem 2.1

We first proceed to bound $d_{\mathscr{L}}(H,H_n)$. We have $H_n(t) = \mu_n\big(\mathscr{L}_n^\ell(t)\big)$, and $H(t) = \mu\big(\mathscr{L}^\ell(t)\big)$ where $\mathscr{L}_n^\ell(t) = \{x \in \mathbb{R}^d : \hat{f}_n(x) \leq t\}$ and $\mathscr{L}^\ell(t) = \{x \in \mathbb{R}^d : f(x) \leq t\}$. The triangular inequality gives

$$\mathscr{L}^\ell\left(t - \|\hat{f}_n - f\|_\infty\right) \subset \mathscr{L}_n^\ell(t) \subset \mathscr{L}^\ell\left(t + \|\hat{f}_n - f\|_\infty\right),$$

which, applying μ_n, yields

$$\mu_n\left(\mathscr{L}^\ell\left(t - \|\hat{f}_n - f\|_\infty\right)\right) \leq H_n(t) \leq \mu_n\left(\mathscr{L}^\ell\left(t + \|\hat{f}_n - f\|_\infty\right)\right).$$

Moreover, by definition of V_n in (5.3), we have

$$H(s) - V_n \leq \mu_n\big(\mathscr{L}^\ell(s)\big) \leq H(s) + V_n,$$

for all real number s. The two last inequalities give

$$H(t - \|\hat{f}_n - f\|_\infty) - V_n \leq H_n(t) \leq H(t + \|\hat{f}_n - f\|_\infty) + V_n.$$

Using the fact that H is non-decreasing, we obtain

$$d_{\mathscr{L}}(H,H_n) \leq \max\left(\|\hat{f}_n - f\|_\infty, V_n\right). \tag{5.4}$$

By Lemma 5.1, G is almost everywhere differentiable. Let us fix $p \in (0;1)$ such that G is differentiable at $1-p$, and observe that $G(1-p) = t^{(p)}$. Denote by G_n the pseudo-inverse of H_n, i.e.

$$G_n(s) = \inf\{t \geq 0 : H_n(t) \geq s\},$$

and remark that $G_n(1-p) = t_n^{(p)}$. Moreover, we always have $d_{\mathscr{L}}(H,H_n) \leq 1$ because $0 \leq H(t) \leq 1$ and $0 \leq H_n(t) \leq 1$ for all $t \in \mathbb{R}$. Hence, since $d_{\mathscr{L}}(H,H_n) = d_{\mathscr{L}}(G,G_n)$, we obtain from Lemma 5.2 that for some constant C,

$$\left|t_n^{(p)} - t^{(p)}\right| = |G_n(1-p) - G(1-p)| \leq C d_{\mathscr{L}}(H,H_n).$$

Theorem 2.1 is now a straightforward consequence of (5.4) and Lemma 5.3. $\square$

6 Proof of Theorem 2.3: convergence of $\mathscr{L}_n(t_n^{(p)})$

6.1 Auxiliary results

We shall assume throughout this subsection that Assumptions 1, 2b and 3 hold.

Lemma 6.1. *For almost all $p \in (0;1)$, we have*

(i) $(\log n) \times \lambda\left(\mathscr{L}_n(t^{(p)}) \Delta \mathscr{L}(t^{(p)})\right) \xrightarrow{\mathbb{P}} 0$ *and*

(ii) $(\log n) \times \lambda\left(\mathscr{L}_n(t_n^{(p)}) \Delta \mathscr{L}(t^{(p)})\right) \xrightarrow{\mathbb{P}} 0.$

Proof. We only prove (ii). Set $\varepsilon_n = \log n / \sqrt{nh_n^d}$, which vanishes under Assumption 2b. Moreover, let $\mathscr{N}_1$, $\mathscr{N}_2$ be defined as

$$\begin{aligned}\mathscr{N}_1^c &= \left\{p \in (0;1) \,:\, \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \lambda\left(\left\{t^{(p)} - \varepsilon \leq f \leq t^{(p)} + \varepsilon\right\}\right) \text{ exists}\right\};\\ \mathscr{N}_2^c &= \left\{p \in (0;1) \,:\, \frac{1}{\varepsilon_n} |t_n^{(p)} - t^{(p)}| \xrightarrow{\text{a.s.}} 0\right\}.\end{aligned}$$

Both $\mathscr{N}_1$ and $\mathscr{N}_2$ have a null Lebesgue measure: the first property is a consequence of the Lebesgue derivation Theorem and the fact that H is a bijection from $(0; \sup_{\mathbb{R}^d} f)$ onto $(0;1)$. The second one is a direct consequence of Theorem 2.1.

Hence, one only needs to prove the lemma for all $p \in \mathscr{N}_1^c \cap \mathscr{N}_2^c$. We now fix p in this set, and we denote by Ω_n the event

$$\Omega_n = \{\|\hat{f}_n - f\|_\infty \leq \varepsilon_n\} \cap \{|t_n^{(p)} - t^{(p)}| \leq \varepsilon_n\}.$$

Since $\mathbb{P}(\Omega_n) \to 1$ by (2.4), it suffices to show that the stated convergence holds on the event Ω_n. Simple calculations yields

$$\begin{aligned}&\lambda\left(\mathscr{L}_n(t_n^{(p)}) \Delta \mathscr{L}(t^{(p)})\right)\\ &\qquad = \lambda\left(\left\{\hat{f}_n \geq t_n^{(p)}; f < t^{(p)}\right\}\right) + \lambda\left(\left\{\hat{f}_n < t_n^{(p)}; f \geq t^{(p)}\right\}\right).\end{aligned}$$

But, on the event Ω_n, we have $\hat{f}_n + \varepsilon_n \geq f \geq \hat{f}_n - \varepsilon_n$ and $t_n^{(p)} - \varepsilon_n \leq t^{(p)} \leq t_n^{(p)} + \varepsilon_n$. Consequently, if n is large enough,

$$\begin{aligned}
&\lambda\left(\mathscr{L}_n(t_n^{(p)}) \Delta \mathscr{L}(t^{(p)})\right) \\
&\quad \leq \lambda\left(\left\{t^{(p)} - 2\varepsilon_n \leq f < t^{(p)}\right\}\right) + \lambda\left(\left\{t^{(p)} \leq f \leq t^{(p)} + 2\varepsilon_n\right\}\right) \\
&\quad = \lambda\left(\left\{t^{(p)} - 2\varepsilon_n \leq f \leq t^{(p)} + 2\varepsilon_n\right\}\right) \\
&\quad \leq C\varepsilon_n,
\end{aligned}$$

for some constant C, because $p \in \mathscr{N}_1^c$ and $(\varepsilon_n)_n$ vanishes. The last inequality proves the lemma, since by Assumption 2b, $\varepsilon_n \log n \to 0$. □

In the sequel, $\tilde{\mu}_n$ denotes the smoothed empirical measure, which is the random measure with density $\hat{f}_n$, defined for all Borel set $A \subset \mathbb{R}^d$ by

$$\tilde{\mu}_n(A) = \int_A \hat{f}_n \mathrm{d}\lambda.$$

Lemma 6.2. *For almost all $p \in (0;1)$,*

(i) $\sqrt{nh_n^d}\left\{\tilde{\mu}_n(\mathscr{L}_n(t^{(p)})) - \mu(\mathscr{L}_n(t^{(p)}))\right\} \xrightarrow{\mathbb{P}} 0$ *and*

(ii) $\sqrt{nh_n^d}\left\{\tilde{\mu}_n(\mathscr{L}_n(t_n^{(p)})) - \mu(\mathscr{L}_n(t_n^{(p)}))\right\} \xrightarrow{\mathbb{P}} 0.$

Proof. We only prove (ii). Fix $p \in (0;1)$ such that the result in Lemma 6.1 holds. Observe that

$$\begin{aligned}
&\left|\tilde{\mu}_n(\mathscr{L}_n(t_n^{(p)})) - \mu(\mathscr{L}_n(t_n^{(p)}))\right| \\
&\quad \leq \int_{\mathscr{L}_n(t_n^{(p)}) \Delta \mathscr{L}(t^{(p)})} |\hat{f}_n - f| \mathrm{d}\lambda + \left|\int_{\mathscr{L}(t^{(p)})} (\hat{f}_n - f) \mathrm{d}\lambda\right| \\
&\quad \leq \lambda\left(\mathscr{L}_n(t_n^{(p)}) \Delta \mathscr{L}(t^{(p)})\right) \|\hat{f}_n - f\|_\infty + \left|\int_{\mathscr{L}(t^{(p)})} (\hat{f}_n - f) \mathrm{d}\lambda\right|. \quad (6.1)
\end{aligned}$$

Recall that K is a radial function with compact support. Since $nh_n^{d+4} \to 0$ and $\mathscr{L}(t^{(p)})$ is compact for all $p \in (0;1)$, it is a classical exercise to prove that for all $p \in (0;1)$,

$$\sqrt{nh_n^d} \int_{\mathscr{L}(t^{(p)})} (\hat{f}_n - f) \mathrm{d}\lambda \xrightarrow{\mathbb{P}} 0. \quad (6.2)$$

(see, e.g., Cadre, 2006, Lemma 4.2). Moreover, by (2.4) and Lemma 6.1,

$$\sqrt{nh_n^d}\,\lambda\left(\mathscr{L}_n(t_n^{(p)})\Delta\mathscr{L}(t^{(p)})\right)\|\hat{f}_n - f\|_\infty \xrightarrow{\mathbb{P}} 0. \tag{6.3}$$

The inequalities (6.1), (6.2) and (6.3) prove the assertion of the lemma. □

Lemma 6.3. *For almost all $p \in (0;1)$,*

$$\sqrt{nh_n^d}\Big\{\mu\left(\mathscr{L}_n(t_n^{(p)})\right) - \mu\left(\mathscr{L}(t^{(p)})\right)\Big\} \xrightarrow{\mathbb{P}} 0.$$

Proof. Let $\varepsilon_n = \log n/\sqrt{nh_n^d}$ and $\mathscr{N}$ be the set defined by

$$\mathscr{N}^c = \left\{p \in (0;1) : \frac{1}{\varepsilon_n}|t_n^{(p)} - t^{(p)}| \xrightarrow{\text{a.s.}} 0\right\}.$$

By Corollary 2.2, $\mathscr{N}$ has a null Lebesgue measure. If $p \in \mathscr{N}^c$, then almost surely, we have $t^{(p)} - \varepsilon_n \leq t_n^{(p)} \leq t^{(p)} + \varepsilon_n$ for large enough n. Hence,

$$\mathscr{L}_n(t^{(p)} + \varepsilon_n) \subset \mathscr{L}_n(t_n^{(p)}) \subset \mathscr{L}_n(t^{(p)} - \varepsilon_n).$$

Consequently, one only needs to prove that for almost all $p \in \mathscr{N}^c$, the two results above hold:

$$\sqrt{nh_n^d}\left\{\mu\left(\mathscr{L}_n(t^{(p)} \pm \varepsilon_n)\right) - \mu\left(\mathscr{L}(t^{(p)})\right)\right\} \xrightarrow{\mathbb{P}} 0. \tag{6.4}$$

For the sake of simplicity, we only prove the "+" part of (6.4).

One can follow the arguments of the proofs of Propositions 3.1 and 3.2 in Cadre (2006), to obtain that for almost all $p \in \mathscr{N}^c$, there exists $J = J(p)$ with

$$\sqrt{nh_n^d}\,\mu\left(\mathscr{L}_n(t^{(p)} + \varepsilon_n) \cap \mathscr{V}_n\right) \xrightarrow{\mathbb{P}} J \quad \text{and}$$
$$\sqrt{nh_n^d}\,\mu\left(\mathscr{L}_n(t^{(p)} + \varepsilon_n)^c \cap \bar{\mathscr{V}}_n\right) \xrightarrow{\mathbb{P}} J,$$

where we set

$$\mathscr{V}_n = \left\{t^{(p)} - \varepsilon_n \leq f < t^{(p)}\right\} \quad \text{and} \quad \bar{\mathscr{V}}_n = \left\{t^{(p)} \leq f < t^{(p)} + 3\varepsilon_n\right\}.$$

Thus, for almost all $p \in \mathscr{N}^c$

$$\sqrt{nh_n^d}\Big\{\mu\left(\mathscr{L}_n(t^{(p)} + \varepsilon_n) \cap \mathscr{V}_n\right) - \mu\left(\mathscr{L}_n(t^{(p)} + \varepsilon_n)^c \cap \bar{\mathscr{V}}_n\right)\Big\} \xrightarrow{\mathbb{P}} 0. \tag{6.5}$$

Now let $p \in \mathscr{N}^c$ satisfying the above result, and set $\Omega_n = \{\|\hat{f}_n - f\|_\infty \leq 2\varepsilon_n\}$. By (2.4), $\mathbb{P}(\Omega_n) \to 1$ hence one only needs to prove that the result holds on the event Ω_n. But, on Ω_n,

$$\begin{aligned}
&\mu\left(\mathscr{L}_n(t^{(p)}+\varepsilon_n)\right) - \mu\left(\mathscr{L}(t^{(p)})\right) \\
&\quad = \mu\left(\left\{\hat{f}_n \geq t^{(p)}+\varepsilon_n; f < t^{(p)}\right\}\right) - \mu\left(\left\{\hat{f}_n < t^{(p)}+\varepsilon_n; f \geq t^{(p)}\right\}\right) \\
&\quad = \mu\left(\mathscr{L}_n(t^{(p)}+\varepsilon_n) \cap \mathscr{V}_n\right) - \mu\left(\mathscr{L}_n(t^{(p)}+\varepsilon_n)^c \cap \bar{\mathscr{V}}_n\right).
\end{aligned}$$

Consequently, by (6.5), we have on Ω_n

$$\sqrt{nh_n^d}\left\{\mu\left(\mathscr{L}_n(t^{(p)}+\varepsilon_n)\right) - \mu\left(\mathscr{L}(t^{(p)})\right)\right\} \xrightarrow{\mathbb{P}} 0.$$

This proves the "+" part of (6.4). The "−" part is obtained with similar arguments. □

6.2 Proof of Theorem 2.3

Let $t_0 \in \mathscr{T}_0^c$. Since f is of class C^2, there exists an open set $I(t_0)$ containing t_0 such that

$$\inf_{\{f \in I(t_0)\}} \|\nabla f\| > 0.$$

Thus, by Theorem 2.1 in Cadre (2006), we have, for almost all $t \in I(t_0)$,

$$\sqrt{nh_n^d}\,\lambda\left(\mathscr{L}_n(t)\Delta\mathscr{L}(t)\right) \xrightarrow{\mathbb{P}} \sqrt{\frac{2}{\pi}}\,\|K\|_2\, t\int_{\{f=t\}} \frac{1}{\|\nabla f\|}\mathrm{d}\mathscr{H}.$$

Recalling now that the Lebesgue measure of $\mathscr{T}_0$ is 0, and that H is a bijection from $(0;\sup_{\mathbb{R}^d} f)$ onto $(0;1)$, it follows that the above result remains true for almost all $p \in (0;1)$, with $t^{(p)}$ instead of t. As a consequence, one only needs to prove that for almost all $p \in (0;1)$, $\sqrt{nh_n^d}\,D_n(p) \to 0$ in probability, where

$$D_n(p) = \lambda\left(\mathscr{L}_n(t_n^{(p)})\Delta\mathscr{L}(t^{(p)})\right) - \lambda\left(\mathscr{L}_n(t^{(p)})\Delta\mathscr{L}(t^{(p)})\right).$$

After some calculations, $D_n(p)$ may be expressed as

$$D_n(p) = \int_{\mathbb{R}^d} \mathbf{1}\{t_n^{(p)} \leq \hat{f}_n < t^{(p)}\} g\,\mathrm{d}\lambda - \int_{\mathbb{R}^d} \mathbf{1}\{t^{(p)} \leq \hat{f}_n < t_n^{(p)}\} g\,\mathrm{d}\lambda,$$

where $g = 1 - 2\mathbf{1}\{f \geq t^{(p)}\}$. For simplicity, we assume that $0 < t_n^{(p)} \leq t^{(p)}$. Recall that $\tilde{\mu}_n$ is the random measure with density $\hat{f}_n$. Thus,

$$D_n(p) \leq \lambda\left(\left\{t_n^{(p)} \leq \hat{f}_n < t^{(p)}\right\}\right) \leq \frac{1}{t_n^{(p)}}\tilde{\mu}_n\left(\left\{t_n^{(p)} \leq \hat{f}_n < t^{(p)}\right\}\right).$$

The factor $1/t_n^{(p)}$ in the right-hand side of the last inequality might be asymptotically bounded by some constant C, using Corollary 2.2. Hence, for all n large enough, and for almost all $p \in (0;1)$,

$$D_n(p) \leq C\left|\tilde{\mu}_n\left(\mathscr{L}_n(t_n^{(p)})\right) - \tilde{\mu}_n\left(\mathscr{L}_n(t^{(p)})\right)\right|. \tag{6.6}$$

The right-hand term in (6.6) may be bounded from above by

$$\begin{aligned}\left|\tilde{\mu}_n\left(\mathscr{L}_n(t_n^{(p)})\right) - \tilde{\mu}_n\left(\mathscr{L}_n(t^{(p)})\right)\right| \leq\ & \left|\tilde{\mu}_n\left(\mathscr{L}_n(t_n^{(p)})\right) - \mu\left(\mathscr{L}_n(t_n^{(p)})\right)\right| \\ & + \left|\mu\left(\mathscr{L}_n(t_n^{(p)})\right) - \mu\left(\mathscr{L}(t^{(p)})\right)\right| \\ & + \left|\mu\left(\mathscr{L}(t^{(p)})\right) - \tilde{\mu}_n\left(\mathscr{L}_n(t^{(p)})\right)\right|.\end{aligned}$$

By Lemma 6.2 and Lemma 6.3, we obtain, for almost all $p \in (0;1)$,

$$\sqrt{nh_n^d}\left\{\tilde{\mu}_n\left(\mathscr{L}_n(t_n^{(p)})\right) - \tilde{\mu}_n\left(\mathscr{L}_n(t^{(p)})\right)\right\} \xrightarrow{\mathbb{P}} 0,$$

which, according to (6.6), gives the stated result. □

7 Proof of Theorem 3.1: convergence of $k_n(t_n^{(p)})$

7.1 Preliminaries

We assume in the whole section that Assumption 1 holds. Since H is a bijection from $(0;\sup_{\mathbb{R}^d} f)$ onto $(0;1)$ and since the Lebesgue measure of $\mathscr{T}_0$ is 0, one only needs to prove Theorem 3.1 for each probability $p \in (0;1)$ such that $t^{(p)} \notin \mathscr{T}_0$. Now fix such a probability p. Because f is of class C^2, there exists a closed interval $I \subset (0;+\infty)$ such that $t^{(p)}$ is in the interior of I, and $\inf_{\{f \in I\}} \|\nabla f\| > 0$. For ease of notation, we now set

$$k_n^{(p)} = k_n(t_n^{(p)}),\ k^{(p)} = k(t^{(p)}),\ J_n = J_n(t_n^{(p)}),\ \text{and } \mathscr{G}_n = \mathscr{G}_n(t_n^{(p)}).$$

In what follows, $B(x,r)$ stands for the Euclidean closed ball centered at $x \in \mathbb{R}^d$ with radius r.

Let $\mathscr{P}_n$ be a finite covering of $\mathscr{L}(t^{(p)} + \varepsilon_n + \varepsilon'_n)$ by closed balls $B(x, r_n/4)$ with centers at $x \in \mathscr{L}(t^{(p)} + \varepsilon_n + \varepsilon'_n)$, constructed in such a way that

$$\text{Card}\,(\mathscr{P}_n) \leq C_1 r_n^{-d}, \tag{7.1}$$

for some positive constant C_1. Let J'_n be the subset of J_n defined by

$$J'_n = \{j \in J_n : f(X_j) \geq t^{(p)} + \varepsilon_n + \varepsilon'_n\}.$$

Define the event $\mathit{\Gamma}_n$ on which every ball of the covering $\mathscr{P}_n$ contains at least one point X_j with $j \in J'_n$, i.e.,

$$\mathit{\Gamma}_n = \left\{\forall A \in \mathscr{P}_n, \exists j \in J'_n \text{with} X_j \in A\right\}.$$

Finally, we set

$$\mathit{\Gamma}'_n = \mathit{\Gamma}_n \cap \left\{\|\hat{f}_n - f\|_\infty \leq \varepsilon_n\right\} \cap \left\{|t_n^{(p)} - t^{(p)}| \leq \varepsilon'_n\right\}.$$

In the sequel, the statement "n is large enough" means that n satisfies the three following conditions:

(i) $(r_n/4)^2 + \Big(4(\varepsilon_n + \varepsilon'_n)/r_n\Big)^2 < \min(\alpha^2, \beta^2)$, where α and β are given by Proposition B.3 and Proposition B.4 respectively,

(ii) $\left[t^{(p)} - \varepsilon_n - \varepsilon'_n; t^{(p)} + \varepsilon_n + \varepsilon'_n\right] \subset I$ and,

(iii) $r_n < D_{\min}$.

In condition *(iii)* above, $D_{\min}$ denotes the smallest distance between two different connected components of $\mathscr{L}(\min I)$. By Lemma B.1, each level set $\mathscr{L}(t)$ has exactly $k^{(p)}$ connected components, provided $t \in I$. Hence,

$$D_{\min} = \min_{1 \leq \ell < \ell' \leq k^{(p)}} \text{dist}\big(\mathscr{C}_\ell(\min I), \mathscr{C}_{\ell'}(\min I)\big),$$

where for all t, the $\mathscr{C}_\ell(t)$'s denote the connected components of $\mathscr{L}(t)$.

Lemma 7.1. *Assume that n is large enough. Then, on $\mathit{\Gamma}'_n$, $k_n^{(p)} = k^{(p)}$.*

Proof. In the proof, a graph is denoted as (set of vertices, set of edges).

On $\mathscr{V}_n' = \{X_j \,:\, j \in J_n'\}$, the graph $\mathscr{G}_n = (\mathscr{V}_n, \mathscr{E}_n)$ induces the subgraph $\mathscr{G}_n' = (\mathscr{V}_n', \mathscr{E}_n')$. We first proceed to prove that $\mathscr{G}_n'$ has exactly $k^{(p)}$ connected components on Γ_n'. To this aim, observe first that

$$J_n' \subset \{j \leq n : f(X_j) \geq t^{(p)} + \varepsilon_n + \varepsilon_n'\}, \tag{7.2}$$

by definition of J_n', provided $\|\hat{f}_n - f\|_\infty \leq \varepsilon_n$ and $|t_n^{(p)} - t^{(p)}| \leq \varepsilon_n'$. Conversely, if $j \leq n$ is such that $f(X_j) \geq t^{(p)} + \varepsilon_n + \varepsilon_n'$ and if $|t_n^{(p)} - t^{(p)}| \leq \varepsilon_n'$, then

$$\hat{f}_n(X_j) \geq f(X_j) - \|\hat{f}_n - f\|_\infty \geq f(X_j) - \varepsilon_n \geq t^{(p)} + \varepsilon_n' \geq t_n^{(p)}.$$

Hence, if $|t_n^{(p)} - t^{(p)}| \leq \varepsilon_n'$

$$J_n' \supset \{j \leq n : f(X_j) \geq t^{(p)} + \varepsilon_n + \varepsilon_n'\}. \tag{7.3}$$

This shows that the two sets in (7.2) and (7.3) are in fact equal as soon as $\|\hat{f}_n - f\|_\infty \leq \varepsilon_n$ and $|t_n^{(p)} - t^{(p)}| \leq \varepsilon_n'$, i.e.,

$$J_n' = \{j \leq n : f(X_j) \geq t^{(p)} + \varepsilon_n + \varepsilon_n'\}. \tag{7.4}$$

In particular, on Γ_n', we have

$$\mathscr{V}_n' = \mathscr{V}_n \cap \mathscr{L}(t^{(p)} + \varepsilon_n + \varepsilon_n'). \tag{7.5}$$

We are now ready to prove that $k_n' = k^{(p)}$. Since n is large enough, $\mathscr{L}(t^{(p)} + \varepsilon_n + \varepsilon_n')$ has exactly $k^{(p)}$ connected components by Lemma B.1. Hence, one only needs to prove that any pair of vertices X_i and X_j of $\mathscr{G}_n'$ is linked by an edge of $\mathscr{E}_n'$ if and only if both vertices lie in the same connected components of $\mathscr{L}(t^{(p)} + \varepsilon_n + \varepsilon_n')$. First, if X_i and X_j belong to different connected components of $\mathscr{L}(t^{(p)} + \varepsilon_n + \varepsilon_n')$, then necessarily, $\|X_i - X_j\| \geq D_{\min}$. Since n is large enough, we have $r_n < D_{min}$, and so no edge of $\mathscr{G}_n'$ connects X_i to X_j. Second, if X_i and X_j belong to the same connected component of $\mathscr{L}(t^{(p)} + \varepsilon_n + \varepsilon_n')$, then on Γ_n', they are contained in some balls of $\mathscr{P}_n$. If they happen to lie in the same ball, then $\|X_i - X_j\| \leq r_n/2$ and so they are connected by an edge in $\mathscr{G}_n'$. Otherwise, there exists a path of edges in $\mathscr{G}_n'$ connecting X_i to X_j, and so they belong to the same connected component of $\mathscr{G}_n'$. This follows from the fact that, whenever n is large enough, the union of the balls of $\mathscr{P}_n$ has the same topology as $\mathscr{L}(t^{(p)} + \varepsilon_n + \varepsilon_n')$; in particular, their number of connected components are equal. As a consequence, since on Γ_n, each ball of $\mathscr{P}_n$ contains at least one vertex of $\mathscr{G}'_n$, it follows that $\mathscr{G}'_n$ has exactly $k^{(p)}$ connected components, i.e.,

$$k'_n = k^{(p)}. \tag{7.6}$$

And, if we decompose $\mathscr{G}'_n$ into its connected components

$$\mathscr{G}'_n = (\mathscr{V}'_{n,1}, \mathscr{E}'_{n,1}) \cup \cdots \cup (\mathscr{V}'_{n,k^{(p)}}, \mathscr{E}'_{n,k^{(p)}}),$$

we have also obtained that

$$\mathscr{V}'_{n,\ell} = \mathscr{V}'_n \cap \mathscr{C}_\ell(\min I), \quad \ell = 1, \ldots, k^{(p)}. \tag{7.7}$$

Now let

$$\mathscr{V}''_n = \mathscr{V}_n \setminus \mathscr{V}'_n.$$

A moment's thought reveals that

$$\mathscr{V}''_n \subset \mathscr{L}(t^{(p)} - \varepsilon_n - \varepsilon'_n) \setminus \mathscr{L}(t^{(p)} + \varepsilon_n + \varepsilon'_n). \tag{7.8}$$

For all vertices X_j in $\mathscr{V}''_n$, we have $B(X_j, r_n/4) \cap \mathscr{L}(t^{(p)} + \varepsilon_n + \varepsilon'_n) \neq \emptyset$ by Proposition B.4, so that $B(X_j, r_n/4)$ intersects some ball of the covering $\mathscr{P}_n$. This proves that any vertex of $\mathscr{V}''_n$ is connected by an edge of $\mathscr{G}_n$ to at least one vertex in $\mathscr{V}'_n$. Consequently, $k_n^{(p)}$ is smaller than the number of connected components of $\mathscr{G}'_n$, which is equal to $k^{(p)}$ by (7.6). But since n is large enough, so that $r_n < D_{min}$, each vertex in $\mathscr{V}''_n$ cannot be connected simultaneously to different components of $\mathscr{G}'_n$ by (7.7) and (7.8). Therefore $k_n^{(p)} \geq k^{(p)}$ and so $k_n^{(p)} = k^{(p)}$. □

7.2 Proof of Theorem 3.1

By Lemma 7.1, provided n is large enough,

$$\Gamma'_n \subset \left\{ k_n^{(p)} = k^{(p)} \right\}.$$

We assume in this section that n is large enough, so that the set of assumptions on n of the preliminaries holds. If we set

$$\Gamma''_n = \left\{ \|\hat{f}_n - f\|_\infty \leq \varepsilon_n \right\} \cap \left\{ |t_n^{(p)} - t^{(p)}| \leq \varepsilon'_n \right\},$$

we then have

$$\mathbb{P}\big(k_n^{(p)} \neq k^{(p)}\big) \leq \mathbb{P}(\Gamma_n'^c) \leq \mathbb{P}(\Gamma_n^c) + \mathbb{P}\big(\Gamma_n''^c\big). \tag{7.9}$$

We now proceed to bound $\mathbb{P}(\Gamma_n^c)$. First observe that

$$\begin{aligned}\mathbb{P}(\Gamma_n^c) &\leq \mathbb{P}\left(\Gamma_n'' \,;\, \exists A \in \mathscr{P}_n : \sum_{j \in J_n'} \mathbf{1}_A(X_j) = 0\right) + \mathbb{P}\big(\Gamma_n''^c\big) \\ &\leq \operatorname{Card}(\mathscr{P}_n) \sup_{A \in \mathscr{P}_n} \mathbb{P}\left(\Gamma_n'' \,;\, \forall i \in J_n' : X_i \in A^c\right) + \mathbb{P}\big(\Gamma_n''^c\big). \qquad (7.10)\end{aligned}$$

Denote by $\bar{J}_n$ the set $\bar{J}_n = \{j \leq n : f(X_j) \geq t^{(p)} + \varepsilon_n + \varepsilon_n'\}$, and recall that by (7.4), J_n' and $\bar{J}_n$ coincide on Γ_n''. Then, for all $A \in \mathscr{P}_n$,

$$\begin{aligned}\mathbb{P}\big(\Gamma_n'' \,;\, \forall i \in J_n' : X_i \in A^c\big) &\leq \mathbb{P}\big(\forall i \in \bar{J}_n, X_i \in A^c\big) \\ &= \left(1 - \mu\left(A \cap \mathscr{L}(t^{(p)} + \varepsilon_n + \varepsilon_n')\right)\right)^n. \qquad (7.11)\end{aligned}$$

By Proposition B.3, for any closed ball A centered at x in $\mathscr{L}(t^{(p)})$ with radius $r_n/4$, we have

$$\begin{aligned}\mu\left(A \cap \mathscr{L}(t^{(p)} + \varepsilon_n + \varepsilon_n')\right) &\geq t^{(p)} \lambda\left(A \cap \mathscr{L}(t^{(p)} + \varepsilon_n + \varepsilon_n')\right) \\ &\geq t^{(p)} \frac{\omega_d}{4^{d+1}} r_n^d. \qquad (7.12)\end{aligned}$$

With (7.1), (7.10), (7.11) and (7.12), we deduce that

$$\mathbb{P}(\Gamma_n^c) \leq C_1 r_n^{-d} \left(1 - t^{(p)} \frac{\omega_d}{4^{d+1}} r_n^d\right)^n + \mathbb{P}(\Gamma_n''^c).$$

According to (7.9) and the inequality $1 - u \leq \exp(-u)$ for all $u \in \mathbb{R}$, we obtain

$$\mathbb{P}\left(k_n^{(p)} \neq k^{(p)}\right) \leq C_1 r_n^{-d} \exp\left(-t^{(p)} \frac{\omega_d}{4^{d+1}} n r_n^d\right) + 2\mathbb{P}(\Gamma_n''^c).$$

This concludes the proof. □

A Auxiliary results on f and H

In this Appendix, we only presume that Assumption 1-*(i)* holds. Recall that $\mathscr{X}$ is the subset of $\mathbb{R}^d$ composed of the critical points of f, i.e.,

$$\mathscr{X} = \{\nabla f = 0\}.$$

The following lemma characterizes the set $\mathscr{T}_0$.

Lemma A.1. *We have* $f(\mathscr{X}) \setminus \{0; \sup_{\mathbb{R}^d} f\} = \mathscr{T}_0$.

Proof. Let $x \in \mathscr{X}$. If $f(x) \neq 0$ or $f(x) \neq \sup_{\mathbb{R}^d} f$, then obviously $f(x) \in \mathscr{T}_0$ and hence, $f(\mathscr{X}) \setminus \{0; \sup_{\mathbb{R}^d} f\} \subset \mathscr{T}_0$. Conversely, $\mathscr{T}_0 \subset f(\mathscr{X})$ by continuity of ∇f and because the set $\{f = t\}$ is compact whenever $t \neq 0$. □

The next proposition describes the absolutely continuous part of the random variable $f(X)$.

Proposition A.2. *Let I be a compact interval of $\mathbb{R}_+^\star$ such that $I \cap \mathscr{T}_0 = \emptyset$. Then, the random variable $f(X)$ has a density g on I, which is given by*

$$g(t) = t \int_{\{f=t\}} \frac{1}{\|\nabla f\|} \mathrm{d}\mathscr{H}, \quad t \in I.$$

Proof. Since $\{f \in I\}$ is compact and $\{f \in I\} \cap \{\nabla f = 0\} = \emptyset$, we have

$$\inf_{\{f \in I\}} \|\nabla f\| > 0.$$

Now, let J be any interval included in I. Observe that f is a locally Lipschitz function and that $\mathbf{1}\{f \in J\}$ is integrable. According to Theorem 2, Chapter 3 in Evans and Gariepy (1992),

$$\mathbb{P}\left(f(X) \in J\right) = \int_{\{f \in J\}} f \mathrm{d}\lambda = \int_J \left(\int_{\{f=s\}} \frac{f}{\|\nabla f\|} \mathrm{d}\mathscr{H} \right) \mathrm{d}s,$$

hence the lemma. □

B Auxiliary results on $\mathscr{L}(t)$

In this Appendix, we only presume that Assumption 1-*(i)* holds. We denote by I a closed and non-empty interval such that

$$\inf_{\{f \in I\}} \|\nabla f\| > 0.$$

The following lemma, stated without proof, is a consequence of Theorem 3.1 in Milnor (1963) p.12 and Theorem 5.2.1 in Jost (1995) p.176; see also Lemma A.1 in Pelletier and Pudlo (2008). Recall that for all $t \geq 0$, $\mathscr{L}(t) = \mathscr{C}_1(t) \cup \cdots \cup \mathscr{C}_{k(t)}(t)$, where the $\mathscr{C}_\ell(t)$'s denote the connected components of $\mathscr{L}(t)$.

Lemma B.1. *There exists a one-parameter group of diffeomorphisms* $(\varphi_u)_{u\in\mathbb{R}}$ *such that for all* s,t *in* I, φ_{t-s} *is a diffeomorphism from* $\mathscr{L}(s)$ *onto* $\mathscr{L}(t)$. *Consequently, for all* s,t *in* I,

(i) $\mathscr{L}(s)$ *and* $\mathscr{L}(t)$ *have the same number of connected components;*

(ii) $\mathscr{C}_\ell(t) \subset \mathscr{C}_\ell(s)$ *whenever* $s < t$ *and* $1 \le \ell \le k(t)$*;*

(iii) $\mathscr{C}_\ell(t) = \varphi_{t-s}(\mathscr{C}_\ell(s))$ *whenever* $1 \le \ell \le k(t)$.

Lemma B.2. *Let* $t \in I$, *and fix* $x \in \mathscr{L}(t)$.

(i) If x *is in the interior of* $\mathscr{L}(t)$, *then*

$$\lim_{(\delta,r)\to(0,0)} r^{-d}\lambda\left(B(x,r)\cap\mathscr{L}(t+r\delta)\right) = \omega_d.$$

(ii) If x *is on the boundary of* $\mathscr{L}(t)$, *then*

$$\lim_{(\delta,r)\to(0,0)} r^{-d}\lambda\left(B(x,r)\cap\mathscr{L}(t+r\delta)\right) = \frac{\omega_d}{2}.$$

Proof. *(i)* If x is in the interior of $\mathscr{L}(t)$, then x is in the interior of $\mathscr{L}(t+\delta_0)$ for some δ_0. Thus, for some $r_0 > 0$, $B(x,r_0) \subset \mathscr{L}(t+\delta_0)$. We can assume that $r_0 < 1$. Then, if $\delta < \delta_0$ and $r < r_0$,

$$B(x,r) \subset B(x,r_0) \subset \mathscr{L}(t+\delta_0) \subset \mathscr{L}(t+r\delta).$$

Hence, for such a pair (r,δ), $B(x,r)\cap\mathscr{L}(t+r\delta) = B(x,r)$, which gives the result.

(ii) Let x be an element of the boundary of $\mathscr{L}(t)$, and denote by $\mathscr{H}_{1/r}$ the homothety with center x and similitude ratio $1/r$. For $r,\delta > 0$, we have

$$\frac{1}{r^d}\lambda\left(B(x,r)\cap\mathscr{L}(t+r\delta)\right) = \lambda\left(A_{r,\delta}\right) \tag{B.1}$$

where $A_{r,\delta} = \mathscr{H}_{1/r}\big(B(x,r)\cap\mathscr{L}(t+r\delta)\big)$. We claim that as $(r,\delta)\to(0,0)$, the indicator function of the set $A_{r,\delta}$ converges toward the indicator function of the set

$$A_{0,0} = \big\{\xi\in\mathbb{R}^d : \|\xi - x\| \le 1, \nabla f(x)\cdot(\xi - x) > 0\big\}.$$

Observe that $\mathscr{H}_{1/r}\big(B(x,r)\big) = B(x,1)$, and fix $\xi \in B(x,1)$. Then, ξ is in $\mathscr{H}_{1/r}\big(\mathscr{L}(t+r\delta)\big)$ if and only if $f\big(x+r(\xi-x)\big) \ge t+r\delta$. Moreover, $f\big(x+r(\xi-x)\big) = t+$

$r\nabla f(x)\cdot(\xi - x) + o(r)$ when $r \to 0$. Recalling that $\nabla f(x) \neq 0$, this gives, for any $\xi \in \mathbb{R}^d$,

$$\lim_{(\delta,r)\to(0,0)} \mathbf{1}\{\xi \in A_{r,\delta}\} = \mathbf{1}\{\xi \in A_{0,0}\}.$$

But, $\lambda(A_{0,0}) = \omega_d/2$, and the indicator functions of $A_{\delta,r}$ are bounded by the indicator function of $B(x,1)$. Therefore it follows that $\lambda(A_{r,\delta}) \to \omega_d/2$ as $(r,\delta) \to (0,0)$ by Lebesgue dominated convergence Theorem. Reporting this fact in equation (B.1) leads to the assertion *(ii)* of the lemma. □

Proposition B.3. *Let $t \in I$. There exists $\alpha > 0$ such that, if $r^2 + \delta^2 < \alpha^2$, then, for all $x \in \mathscr{L}(t)$,*

$$\lambda\left(B(x,r) \cap \mathscr{L}(t + r\delta)\right) \geq Cr^d,$$

where C is any positive constant such that $C < \omega_d/2$.

Proof. Let $U = \{(x,r,\delta) : f(x) \geq t,\ r \geq 0,\ \delta \geq 0\}$, and consider the map $\psi : U \to \mathbb{R}_+$ given by

$$\psi(x,\delta,r) = \begin{cases} r^{-d}\lambda\left(B(x,r)\cap\mathscr{L}(t+r\delta)\right) & \text{if } r > 0, \\ \omega_d & \text{if } r = 0,\ f(x) > t, \\ \omega_d/2 & \text{if } r = 0,\ f(x) = t. \end{cases}$$

By Lemma B.2, ψ is bounded from below by some constant $C < \omega_d/2$ on $V \cap U$, where V is some open neighborhood of $\mathscr{L}(t) \times (0,0)$. Since $\mathrm{dist}(\mathscr{L}(t) \times (0,0), \mathbb{R}^{d+3} \setminus V) > 0$, as a distance between two disjoint closed sets, one of them being compact, the result is proved. □

The proof of the next result is left to the reader, since it can be obtained by adapting the proofs of Lemma B.2 and Proposition B.3.

Proposition B.4. *There exists $\beta > 0$ such that, if $r^2 + \delta^2 < \beta^2$, then, for all (t,x) such that $t \in I$ and $x \in \mathscr{L}(t - r\delta)$, the closed ball $B(x, r/4)$ intersects $\mathscr{L}(t + r\delta)$.*

References

Aubin, T. (2000). *A course in Differential Geometry.* Graduate Studies in Mathematics, American Mathematical Society, Providence, Rhode Island.

Azzalini, A. and Bowman, A.W. (1990). A look at some data on the Old Faithful geyser. *Applied Statistics*, 39: 357365.

Baillo, A. (2003). Total error in a plug-in estimator of level sets. *Statist. and Probab. Lett.*, 65(4):411–417.

Baillo, A., Cuestas-Alberto, J., and Cuevas, A. (2001). Convergence rates in nonparametric estimation of level sets. *Statist. and Probab. Lett.*, 53(1):27–35.

Baillo, A., Cuevas, A., and Justel, A. (2000). Set estimation and nonparametric detection. *Canad. J. of Statist.*, 28(4):765–782.

Baudry, J.P., Raftery, A.E., Celeux, G., Lo, K. and Gottardo, R. (2010) Combining Mixture Components for Clustering. *J. Comput. Graph. Statist.*, 19(2):332–353.

Ben-David, S., Pál, D., and Simon, H. (2007). Stability of *k*-means clustering. In Bshouty, N. and Gentile, C., editors, *Learning theory*, Lecture Notes in Comput. Sci. 4539, pages 20–34. Springer.

Ben-David, S., von Luxburg, U., and Pál, D. (2006). A sober look at clustering stablity. In Lugosi, G. and Simon, H., editors, *Learning theory*, Lecture Notes in Comput. Sci. 4005, pages 5–19. Springer.

Biau, G., Cadre, B., and Pelletier, B. (2007). A graph-based estimator of the number of clusters. *ESAIM Probab. Stat.*, 11:272–280.

Billingsley, P. (1995). *Probability and Measure*. Wiley-Intersciences.

Cadre, B. (2006). Kernel estimation of density level sets. *J. of Multivariate Anal.*, 97(4):999–1023.

Cuevas, A., Febrero, M., and Fraiman, R. (2000). Estimating the number of clusters. *Canad. J. of Statist.*, 28(2):367–382.

Cuevas, A., Gonzalez-Manteiga, W., and Rodriguez-Casal, A. (2006). Plug-in estimation of general level sets. *Aust. N. Z. J. of Stat.*, 48(1):7–19.

Devroye, L., Gyorfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New-York.

Duda, R., Hart, P., and Stork, D. (2000). *Pattern Classification*. Second Edition, Wiley-Interscience, New York.

Einmahl, U. and Mason, D. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.*, 33(3):1380–1403.

Evans, L. and Gariepy, R. (1992). *Measure Theory and Fine Properties of Functions*. CRC Press, Boca Raton, FL.

Giné, E. and Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. H. Poincaré Prob. Statist.*, 38(6):907–921.

Hartigan, J. (1975). *Clustering Algorithms*. John Wiley, New-York.

Hartigan, J. (1987). Estimation of a convex density contour in two dimensions. *J. Amer. Statist. Assoc.*, 82(397):267–270.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New-York, 2nd edition.

Hyndman, R. (1996). Computing and graphing highest density regions. *The American Statistician*, 50:120–126.

Jost, J. (1995). *Riemannian Geometry and Geometric Analysis*. Universitext. Springer-Verlag, Berlin.

Mason, D. and Polonik, W. (2009). Asymptotic normality of plug-in level set estimates. *Annals of Applied Probability*, 19:1108-1142.

Milnor, J. (1963). *Morse Theory*. Annals of Mathematical Studies. Princeton University Press, Princeton.

Molchanov, I. (1998). A limit theorem for solutions of inequalities. *Scand. J. of Statist.*, 25(1):235–242.

Muller, D. and Sawitzki, G. (1991). Excess mass estimates and tests of multimodality. *J. Amer. Assoc.*, 86(415):738–746.

Nolan, D. (1991). The excess-mass elipsoid. *J. Multivariate Anal.*, 39(2):348–371.

Pelletier, B. and Pudlo, P. (2008). Strong consistency of spectral clustering on level sets. *Submitted*.

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer Series in Statistics, Springer, New-York, Berlin, Heidelberg, Tokyo.

Polonik, W. (1995). Measuring mass concentration and estimating density contour clusters—an excess mass approach. *Ann. Statist.*, 23(3):855–881.

Polonik, W. (1997). Minimum volume sets and generalized quantile processes. *Stochastic Process. Appl.*, 69(1):1–24.

Tsybakov, A. (1997). On nonparametric estimation of density level sets. *Ann. Statist.*, 25(3):948–969.