A Graph-Based Estimator of the Number of Clusters

Gérard BIAU*, Benoît CADRE and Bruno PELLETIER

Institut de Mathématiques et de Modélisation de Montpellier
UMR CNRS 5149, Equipe de Probabilités et Statistique
Université Montpellier II, CC 051
Place Eugène Bataillon, 34095 Montpellier Cedex 5, France

biau,cadre,pelletier@math.univ-montp2.fr

Abstract

Assessing the number of clusters of a statistical population is one of the essential issues of unsupervised learning. Given n independent observations $X_1, \ldots, X_n$ drawn from an unknown multivariate probability density f, we propose a new approach to estimate the number of connected components, or clusters, of the t-level set $\mathcal{L}(t) = \{x : f(x) \geq t\}$. The basic idea is to form a rough skeleton of the set $\mathcal{L}(t)$ using any preliminary estimator of f, and to count the number of connected components of the resulting graph. Under mild analytic conditions on f, and using tools from differential geometry, we establish the consistency of our method.

Index Terms — Cluster analysis, Connected component, Level set, Graph, Tubular neighborhood.

AMS 2000 Classification: 62G05, 62G20.

1 Introduction

Clustering is the problem of identifying groupings of similar points that are relatively isolated from each other, or in other words to partition the data into dissimilar groups of similar items. This unsupervised learning paradigm is perhaps one of the most widely used statistical techniques for exploratory

*Corresponding author.

data analysis. Across all disciplines, from social sciences over biology to computer science, practitioners try to get a first intuition about their data by identifying meaningful groups of observations. We refer the reader to Duda, Hart and Stork [9], Chapter 10, and Hastie, Tibshirani and Friedman [12], Chapter 14, for a general background on the question.

A major challenge in cluster analysis is to assess the number of clusters, say k. Practically speaking, the identification of k is essential for effective and efficient data partitioning, and it should be seen as a step preliminary to any clustering algorithm. For instance, popular clustering algorithms such as k-means or Gaussian mixture modeling may generate bad results if initial partitions are not properly chosen.

Contrary to data analysis methods such as regression or classification, there are many ways to define clustering—even the question "What is clustering?" is difficult to answer in all generality (von Luxburg and Ben-David [14]). Thus, in order to make precise statements about k, a formal definition of cluster is needed. In the present paper, we will use the definition proposed by Hartigan [11]: Given a $\mathbb{R}^d$-valued random variable X with probability density f and a positive level t, a t-cluster is defined as a connected component (i.e., a maximal connected subset) of the t-level set

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

The advantage of this definition is that it is geometrically easy to understand. The level t should not be considered here as a smoothing parameter to be assigned in an optimum way: it just indicates the resolution level chosen for the practical clustering problem at hand. Thus, in this context, the number of connected components of $\mathcal{L}(t)$, say $k(t)$, is considered as the "true" number of clusters of the underlying distribution.

In the present paper, our purpose is to estimate the positive integer $k(t)$, given a random sample $X_1, \ldots, X_n$ drawn from f. A rough analysis suggests first to estimate the level sets of the probability density f (Polonik [16], Tsybakov [18], Cadre [3]), and then to evaluate the number of connected components of the resulting set estimate. However, this does not seem to be a promising strategy, especially because it requires assessing the level sets, which is, in the present context, a superfluous operation. (Note however that estimating the clusters can provide valuable information to group the data, see Cuevas, Febrero and Fraiman [7]). Therefore, we propose a different approach, which bypass the estimation of the level sets, and which is computationally simple. The basic idea is to form a rough skeleton of the level set $\mathcal{L}(t)$ using any preliminary estimator of f, and to count the number of connected components of the resulting graph. Practically speaking, the latter operation can be performed efficiently using for example a tree search algorithm such as Depth-First Search (Cormen, Leiserson and Rivest [5]). Our approach is close in spirit to that of Cuevas, Febrero and Fraiman [6], who analyse a simple algorithm to count the number of connected components of the Devroye-Wise [8] estimate of $\mathcal{L}(t)$. We also refer the reader to Duda, Hart and Stork [9], Section 10.12, for an account on related graph-theoretic methods for clustering purposes.

The paper is organized as follows. In Section 2, we introduce notation and define $k_n(t)$, our graph-based estimator of the number of clusters. The convergence of $k_n(t)$ towards $k(t)$ is studied in Section 3. Technical lemmas necessary to the proof of the results are postponed to the Appendix A.

2 Notation and assumptions

Let f be a probability density function on $\mathbb{R}^d$. As explained earlier, for any $t > 0$ in the range of f, we let the t-level set be defined as $\mathcal{L}(t) = \{x \in \mathbb{R}^d : f(x) \geq t\}$, and denote by $k(t)$ the number of connected components of $\mathcal{L}(t)$. Recall that, in our framework, the integer $k(t)$ is considered as the "true" number of clusters of the statistical population associated with f, in the sense of Hartigan's definition [11]. In all of the following, $k(t)$ will be assumed finite.

Let $\mathcal{D}_n = \{X_1, \ldots, X_n\}$ be an i.i.d. sample drawn from f. In addition to the data set $\mathcal{D}_n$, it will also be supposed that at our disposal is an estimator f_n of f, based on $\mathcal{D}_n$, and obtained by an arbitrary method, e.g., a kernel density estimator, but many other choices are possible.

We now proceed to define the estimator of $k(t)$ by constructing a graph as follows. First, set $J_n(t) = \{i = 1, \ldots, n : f_n(X_i) \geq t\}$. Next, given a sequence (r_n) of (strictly) positive real numbers, consider the sample items falling in $J_n(t)$, and introduce the Card $J_n(t) \times$ Card $J_n(t)$ matrix $\mathbf{S}_n = [s_{ij}]$ with binary entries

$$s_{ij} = \begin{cases} 1 & \text{if } \|X_i - X_j\| \leq r_n \\ 0 & \text{otherwise,} \end{cases}$$

where $\|.\|$ is the Euclidean norm on $\mathbb{R}^d$. The matrix $\mathbf{S}_n$ induces a graph, $\mathcal{G}_n(t)$, the nodes of which are the points in $J_n(t)$, and where an edge joins node i and node j if and only if $s_{ij} = 1$, or, equivalently, $\|X_i - X_j\| \leq r_n$. This algorithm produces a skeleton of the set $J_n(t)$, where two elements x and x' of $J_n(t)$ are in the same cluster if and only if there exists a chain x, $x_1, \ldots, x_k$, x' in $J_n(t)$ such that x is connected to x_1, x_1 to x_2, and so on for the whole chain. Our proposal is to estimate $k(t)$ by $k_n(t)$, the number of connected components of the graph $\mathcal{G}_n(t)$, sometimes called ε-nearest neighbor graph. As explained in the Introduction, the evaluation of $k_n(t)$ does not require to estimate the whole set $\mathcal{L}(t)$. Moreover, its computation can be performed efficiently in $\mathrm{O}(V\mathcal{G}_n(t) + E\mathcal{G}_n(t))$ operations (e.g., via the Depth-First Search algorithm, see Cormen, Leiserson and Rivest [5]), where $V\mathcal{G}_n(t)$ (*resp.* $E\mathcal{G}_n(t)$) denotes the number of vertices (*resp.* edges) of the graph $\mathcal{G}_n(t)$.

Our main result states that $k_n(t)$ is a consistent estimator of $k(t)$. To prove this, and denoting by $\{f \in A\}$ the set $\{x \in \mathbb{R}^d : f(x) \in A\}$ for any Borel set $A \subset \mathbb{R}$, we shall need the following assumptions.

Assumption 1

(a) The probability density f is of class $\mathcal{C}^1$ on a neighborhood of $\{f = t\}$.

(b) For each $x \in \{f = t\}$, the gradient of f at x is non-zero.

Assumption 2

With probability 1, the estimator f_n is of class $\mathcal{C}^1$.

Note that Assumption 1 (b) is equivalent to the fact that the differential $\mathrm{d}f_x$ of f at x is surjective at every $x \in \{f = t\}$. Furthermore, Assumption 1 implies that $\{f = t\}$ has Lebesgue mass 0 and that each connected component of $\mathcal{L}(t)$ has positive Lebesgue mass, i.e., we have (i) $\lambda(\{f = t\}) = 0$, and (ii) $\lambda(\mathcal{C}_l(t)) > 0$, where λ is the Lebesgue measure on $\mathbb{R}^d$, and where the $\mathcal{C}_l$ are the connected components of $\mathcal{L}(t)$. At last, under Assumption 1, the set $\{f = t\}$ is a submanifold of $\mathbb{R}^d$ of codimension 1 by the Implicit Function Theorem. Finally, observe that Assumption 2 is not restrictive, and holds for example if f_n is of kernel type with a continuously differentiable kernel.

In the following, ∇ stands for the gradient and $\|.\|_\infty$ denotes the supremum norm over $\mathbb{R}^d$.

3 Main result

Theorem 3.1 *Suppose that Assumption 1 and Assumption 2 hold. Let* (ε_n) *be a sequence of positive real numbers such that* $\varepsilon_n \to 0$ *and* $\varepsilon_n = \mathrm{o}(r_n)$*. Let* V *be a neighborhood of* $\{f = t\}$ *such that* $\inf_V \|\nabla f\| > 0$*. Then there exist two positive constants* C_1 *and* C_2 *such that:*

$$\mathbb{P}\big(k_n(t) \neq k(t)\big) \leq C_1 r_n^{-d} \exp(-C_2 n r_n^d) \\ + 2\mathbb{P}\big(\|f_n - f\|_\infty > \varepsilon_n\big) + \mathbb{P}\Big(\inf_V \|\nabla f_n\| < \frac{1}{2} \inf_V \|\nabla f\|\Big).$$

As an example, consider the case where f_n is a kernel estimator of f, i.e., for $x \in \mathbb{R}^d$,

$$f_n(x) = \frac{1}{n h_n^d} \sum_{i=1}^n K\Big(\frac{x - X_i}{h_n}\Big),$$

where the kernel K is a probability density on $\mathbb{R}^d$, and the smoothing parameter h_n vanishes as $n \to \infty$. For simplicity, assume that K is the Gaussian kernel and that f is a $\mathcal{C}^1$ probability density with bounded gradient. Let h_n be such that $h_n = \mathrm{o}(\varepsilon_n)$ and $n h_n^{d+1}/\log n \to \infty$. Using Bernstein inequality, one easily derives exponential bounds for the two terms above involving f_n and ∇f_n (see, e.g., Prakasa Rao [17]). Moreover, assuming that $h_n \leq \varepsilon_n^2$, together with the condition $n r_n^d/\log n \to \infty$, we obtain the result

$$\mathbb{P}\big(k_n(t) \neq k(t)\big) = \mathrm{O}\Big(\frac{1}{n^2}\Big).$$

Since $k_n(t)$ and $k(t)$ are integers, the Borel-Cantelli lemma shows that, with probability 1, $k_n(t) = k(t)$ for all n large enough.

Remark According to a referee, a challenging question is whether one can obtain similar results by using only the connected components of the standard ε-nearest and k-nearest neighbor graphs, for example by adapting methods of Brito, Chavez, Quiroz and Yukich [2] and Penrose [15].

Proof of Theorem 3.1 uses the following lemma.

Lemma 3.1 *Suppose that Assumption 1 holds. Then, for* $\varepsilon > 0$ *small enough, we have*

$$k(t - \varepsilon) = k(t) = k(t + \varepsilon).$$

Proof We only prove the equality $k(t) = k(t + \varepsilon)$, the other case being similar. On the one hand, for $\varepsilon > 0$ small enough, $k(t) \leq k(t + \varepsilon)$ since

$\lambda(\{f = t\}) = 0$. On the other hand, the inequality $k(t) \geq k(t+\varepsilon)$ for $\varepsilon > 0$ small enough is clear, since the gradient of f does not vanish on a neighborhood of $\{f = t\}$. $\square$

From now on, we denote by $\hat{k}_n(t)$ the number of connected components of the set

$$\mathcal{L}_n(t) = \{x \in \mathbb{R}^d : f_n(x) \geq t\}.$$

Lemma 3.2 *Suppose that Assumption 1 and Assumption 2 hold. Then, for $\varepsilon > 0$ small enough, the following inclusion between probability events holds for all $n \geq 1$:*

$$\Big[\|f_n - f\|_\infty \leq \varepsilon\Big] \cap \Big[\inf_V \|\nabla f_n\| \geq \frac{1}{2}\inf_V \|\nabla f\|\Big] \subset \Big[\hat{k}_n(t) = k(t)\Big],$$

where V is defined in Theorem 3.1.

Proof On the one hand, using Lemma 3.1, we know that, for $\varepsilon > 0$ small enough and all $n \geq 1$,

$$\Big[\|f_n - f\|_\infty \leq \varepsilon\Big] \subset \Big[k(t - \|f_n - f\|_\infty) = k(t + \|f_n - f\|_\infty)\Big]$$

between probability events. On the other hand, using the triangle inequality, we may write

$$\mathcal{L}(t + \|f_n - f\|_\infty) \subset \mathcal{L}_n(t) \subset \mathcal{L}(t - \|f_n - f\|_\infty). \tag{3.1}$$

For any $u > 0$, we denote by $\mathcal{C}_j(u)$, $j = 1, \ldots, k(u)$, the connected components of the set $\mathcal{L}(u)$. Then, for ε small enough, and after a possible rearrangement of the indices, we have $\mathcal{C}_j(t + \|f_n - f\|_\infty) \subset \mathcal{C}_j(t - \|f_n - f\|_\infty)$, for all $j = 1, \ldots, k(t)$, on the event $[\|f_n - f\|_\infty \leq \varepsilon]$. Consequently, on the event $[\|f_n - f\|_\infty \leq \varepsilon]$, $\hat{k}_n(t) \geq k(t)$.

Under Assumption 1, there exists a neighborhood U of $\{f = t\}$ on which $\mathrm{d}f$ is never zero. Without loss of generality, one can assume that $V \subset U$. Now ε can be chosen small enough for we have

$$\mathcal{L}(t - \|f_n - f\|_\infty) \setminus \mathcal{L}(t + \|f_n - f\|_\infty) \subset V$$

on the event $[\|f_n - f\|_\infty \leq \varepsilon]$. Also, from equation (3.1), it follows that $\partial\mathcal{L}_n(t) \subset \mathcal{L}(t - \|f_n - f\|_\infty) \setminus \mathcal{L}(t + \|f_n - f\|_\infty)$. Suppose that $\hat{k}_n(t) > k(t)$ on the event

$$\Big[\|f_n - f\|_\infty \leq \varepsilon\Big] \cap \Big[\inf_V \|\nabla f_n\| \geq \frac{1}{2}\inf_V \|\nabla f\|\Big].$$

Then f_n must assume a local minimum at some point, say x, in V with $f_n(x) < t$, which contradicts the fact that $\inf_V \|\nabla f_n\| > 0$. Hence, $\hat{k}_n(t) = k(t)$. $\square$

Proof of Theorem 3.1 Consider a covering $\mathcal{P}_n$ of $\mathcal{L}(t)$ composed of closed balls centered on points of $\mathcal{L}(t)$ and of radius $r_n/2$. Recall that by recursing to a metric entropy argument (see for example Györfi, Kohler, Krzyżak and Walk [10]), it may easily be shown that the minimal number of balls necessary to cover a given compact $\mathcal{D}$ of $\mathbb{R}^d$ by balls of radius r with centers in $\mathcal{D}$ is of order $\mathrm{O}(r^{-d})$. Thus, from now on, the covering $\mathcal{P}_n$ will be assumed to be constructed in such a way that

$$\mathrm{Card}\,(\mathcal{P}_n) \leq C_1 r_n^{-d} \tag{3.2}$$

for some positive constant C_1. Let us introduce the event

$$\Omega_n(t) = \Big[\forall A \in \mathcal{P}_n : \sum_{i \in J_n(t)} \mathbf{1}_A(X_i) \geq 1\Big].$$

Finally, we denote by δ the smallest distance between two connected components of $\mathcal{L}(t)$ when $k(t) \geq 2$, and let $\delta = +\infty$ otherwise. Note that $\delta > 0$ by assumption.

Observe that, on the event $\Omega_n(t)$, each element of $\mathcal{P}_n$, i.e., a ball of radius $r_n/2$, contains at least one data point X_i with $i \in J_n(t)$. Thus, as long as n is large enough such that (i) $r_n \leq \delta/2$, and (ii) ε_n is small enough for Lemma 3.1 to hold, we have

$$\Omega_n(t) \cap \Big[\|f_n - f\|_\infty \leq \varepsilon_n\Big] \subset \Big[k_n(t) = \hat{k}_n(t)\Big].$$

Consequently, using Lemma 3.2, we deduce that

$$\begin{aligned}
&\Omega_n(t) \cap \Big[\|f_n - f\|_\infty \leq \varepsilon_n\Big] \cap \Big[\inf_V \|\nabla f_n\| \geq \frac{1}{2}\inf_V \|\nabla f\|\Big] \\
&\subset \Big[k_n(t) = \hat{k}_n(t)\Big] \cap \Big[\hat{k}_n(t) = k(t)\Big] \\
&\subset \big[k_n(t) = k(t)\big].
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\big(k_n(t) = k(t)\big) \\
&\quad \geq \mathbb{P}\Big(\Omega_n(t) \cap \big[\|f_n - f\|_\infty \leq \varepsilon_n\big] \cap \big[\inf_V \|\nabla f_n\| \geq \frac{1}{2} \inf_V \|\nabla f\|\big]\Big) \\
&\quad = \mathbb{P}\big(\Omega_n(t)\big) - \mathbb{P}\Big(\Omega_n(t) \cap \Big(\big[\|f_n - f\|_\infty > \varepsilon_n\big] \cup \big[\inf_V \|\nabla f_n\| < \frac{1}{2} \inf_V \|\nabla f\|\big]\Big)\Big) \\
&\quad \geq \mathbb{P}\big(\Omega_n(t)\big) - \mathbb{P}\big(\|f_n - f\|_\infty > \varepsilon_n\big) - \mathbb{P}\big(\inf_V \|\nabla f_n\| < \frac{1}{2} \inf_V \|\nabla f\|\big). \quad (3.3)
\end{aligned}$$

Now we proceed to bound from below the term $\mathbb{P}(\Omega_n(t))$. We have:

$$\begin{aligned}
\mathbb{P}\big(\Omega_n^c(t)\big) &\leq \mathbb{P}(\exists A \in \mathcal{P}_n : \sum_{i \in J_n(t)} \mathbf{1}_A(X_i) = 0 \text{ and } \|f_n - f\|_\infty \leq \varepsilon_n) \\
&\quad + \mathbb{P}\big(\|f_n - f\|_\infty > \varepsilon_n\big) \\
&\leq \operatorname{Card}(\mathcal{P}_n) \sup_{A \in \mathcal{P}_n} \mathbb{P}(\forall i \in J_n(t) : X_i \in A^c \text{ and } \|f_n - f\|_\infty \leq \varepsilon_n) \\
&\quad + \mathbb{P}(\|f_n - f\|_\infty > \varepsilon_n). \quad (3.4)
\end{aligned}$$

Set $\bar{J}_n(t) = \{i = 1, \ldots, n : f(X_i) \geq t + \varepsilon_n\}$. On the event $[\|f_n - f\|_\infty \leq \varepsilon_n]$, we have $\bar{J}_n(t) \subset J_n(t)$. Consequently, for all $A \in \mathcal{P}_n$,

$$\mathbb{P}(\forall i \in J_n(t) : X_i \in A^c \text{ and } \|f_n - f\|_\infty \leq \varepsilon_n) \leq \mathbb{P}(\forall i \in \bar{J}_n(t) : X_i \in A^c). \quad (3.5)$$

But, by definition of $\bar{J}_n(t)$,

$$\begin{aligned}
&\mathbb{P}(\forall i \in \bar{J}_n(t) : X_i \in A^c) \\
&\quad = \mathbb{P}\big(\forall i = 1, \ldots, n : (f(X_i) \geq t + \varepsilon_n \text{ and } X_i \in A^c) \text{ or } (f(X_i) < t + \varepsilon_n)\big) \\
&\quad = \Big[\mu\big(\{f \geq t + \varepsilon_n\} \cap A^c\big) + \mu\big(\{f < t + \varepsilon_n\}\big)\Big]^n \\
&\quad = \Big[1 - \mu\big(A \cap \{f \geq t + \varepsilon_n\}\big)\Big]^n, \quad (3.6)
\end{aligned}$$

where μ denotes the probability distribution associated with f.

Since $\varepsilon_n = \mathrm{o}(r_n)$, it follows from Proposition A.2 that there exists a positive constant C_2, independent of n and A, such that

$$\mu\big(A \cap \{f \geq t + \varepsilon_n\}\big) \geq C_2 r_n^d. \quad (3.7)$$

Thus, we deduce from (3.2) and (3.4)–(3.7) that

$$\mathbb{P}\big(\Omega_n^c(t)\big) \leq C_1 r_n^{-d} (1 - C_2 r_n^d)^n + \mathbb{P}\big(\|f_n - f\|_\infty > \varepsilon_n\big).$$

Using the inequality $1 - u \le \exp(-u)$ for $u \in \mathbb{R}$, we can now conclude from (3.3) that

$$\mathbb{P}\big(k_n(t) = k(t)\big) \geq 1 - C_1 r_n^{-d} \exp(-C_2 n r_n^d) \\ - 2\mathbb{P}\big(\|f_n - f\|_\infty > \varepsilon_n\big) - \mathbb{P}\Big(\inf_V \|\nabla f_n\| < \frac{1}{2} \inf_V \|\nabla f\|\Big),$$

as desired. $\square$

A Geometrical results

Let us start with some definitions. For general references, we refer the reader to Bredon [1], Chavel [4], and Kobayashi and Nomizu [13]. Let (M, σ) be a smooth and closed (i.e., compact and without boundary) submanifold of $\mathbb{R}^d$. Let T_pM be the tangent space to M at p, and let TM be the tangent bundle of M. For all $p \in M$, T_pM may be considered as a subspace of $\mathbb{R}^d$ via the canonical identification of $T_p\mathbb{R}^d$ with $\mathbb{R}^d$ itself. Via this identification, the normal space $T_pM^\perp$ to M at p is the orthogonal complement of T_pM in $\mathbb{R}^d$. The normal bundle of M in $\mathbb{R}^d$ is defined by $TM^\perp = \cup_{p \in M} T_pM^\perp$, with bundle projection map $\pi : TM^\perp \to M$ defined by $\pi\langle p, v\rangle = p$, i.e., each element $< p, v >$ of $TM^\perp$ is mapped on p by π.

Now let $\theta : TM^\perp \to \mathbb{R}^d$ be given by $\theta\langle p, v\rangle = p + v$. Also let $TM^\perp_\varepsilon = \{\langle p, v\rangle \in TM^\perp : \|v\| < \varepsilon\}$. Then the Tubular Neighborhood Theorem (see e.g., Bredon [1], page 93) states that there exists an $\varepsilon > 0$ such that $\theta : TM^\perp_\varepsilon \to \mathbb{R}^d$ is a diffeomorphism onto the neighborhood $\mathcal{V}(M, \varepsilon) = \{x \in \mathbb{R}^d : \mathrm{dist}(x, M) < \varepsilon\}$ of M in $\mathbb{R}^d$, which is called a tubular neighborhood of radius ε of M in $\mathbb{R}^d$.

Proposition A.1 *Let $\mathcal{D}$ be a connected domain of $\mathbb{R}^d$ with smooth boundary $\partial\mathcal{D}$. Then there exists $\rho > 0$ such that, for all $r \le \rho$ and all $x \in \mathcal{D}$, there exists a point $y \in \mathcal{D}$ such that*

$$B\Big(y, \frac{r}{2}\Big) \subset B(x, r) \cap \mathcal{D}.$$

Proof By the Tubular Neighborhood Theorem, there exists a $r_0 > 0$ such that the set

$$\mathcal{V}(\partial\mathcal{D}, r_0) = \{x \in \mathbb{R}^d : \mathrm{dist}(x, \partial\mathcal{D}) \le r_0\}$$

is diffeomorphic to the subset

$$T\partial\mathcal{D}^\perp_\varepsilon = \{< p, v > \in T\partial\mathcal{D}^\perp : \|v\| \le \varepsilon\}$$

of the normal bundle $T\partial\mathcal{D}^\perp$ of $\partial\mathcal{D}$. Thus each $x \in \mathcal{V}(\partial\mathcal{D}, r_0)$ projects uniquely onto $\partial\mathcal{D}$, and may be expressed as

$$x = p_x + v_x e_{p_x},$$

where e_p denotes the unit-norm section of $T\partial\mathcal{D}^\perp$ pointing inwards $\mathcal{D}$, i.e., e_p is the unit normal vector field to $\partial\mathcal{D}$ directed towards the interior of $\mathcal{D}$.

Set $r \leq r_0/2$. Clearly, for all $x \in \mathcal{D}$ such that $B(x, r) \subset \mathcal{D}$, we have

$$B\Big(x, \frac{r}{2}\Big) \subset B(x, r).$$

Now we examine those cases for which $B(x, r) \cap \mathcal{D}^c \neq \emptyset$. In this configuration, we have

$$B(x, r) \subset \mathcal{V}(\partial\mathcal{D}, r_0),$$

since, for all $y \in B(x, r)$,

$$\mathrm{dist}(y, \partial\mathcal{D}) \leq \mathrm{dist}(y, x) + \mathrm{dist}(x, \partial\mathcal{D}) \leq r_0.$$

Set $x = p_x + v_x e_{p_x}$, and consider the ball $B(y, r/2)$ centered at $y = p_x + (v_x + r/2)e_{p_x}$ and of radius $r/2$. This ball is clearly contained in $B(x, r)$. Now, suppose that $B(y, r/2)$ is not included in $\mathcal{D}$. Then, there exists a point $q \in \partial\mathcal{D}$ such that

$$\mathrm{dist}(q, y) < \frac{r}{2}.$$

But

$$\begin{aligned}
\mathrm{dist}(q, y) &\geq \mathrm{dist}(y, \partial\mathcal{D}) \\
&= \frac{r}{2} + v_x \\
&\geq \frac{r}{2},
\end{aligned}$$

hence a contradiction. $\square$

Proposition A.2 *Suppose that the probability density f satisfies Assumption 1. Let $r_n \to 0$ and let $\varepsilon_n = \mathrm{o}(r_n)$. Then there exists a constant $C > 0$ such that, for all n large enough, and for all $x \in \{f \geq t\}$,*

$$\mu\big(B(x, r_n) \cap \{f \geq t + \varepsilon_n\}\big) \geq C r_n^d,$$

where μ denotes the probability distribution associated with f.

Proof Observe first that, since f satisfies Assumption 1, there exists an open neighborhood of $\{f = t\}$ on which $\mathrm{d}f$ is surjective. Consequently, from the Implicit Function Theorem, there exists $\varepsilon_0 > 0$ such that, for all $\varepsilon \le \varepsilon_0$, $\{f = t + \varepsilon\}$ is a submanifold of $\mathbb{R}^d$ of codimension 1.

Consequently, for all n large enough such that $\varepsilon_n \le \varepsilon_0$, and for all $x \in \{f \ge t + \varepsilon_n\}$, the result follows from Proposition A.1. Thus there remains to examine those cases for which $x \in \{t \le f < t + \varepsilon_n\}$. For this purpose, we first prove that, for all n large enough, $B(x, r_n)$ has a non-empty intersection with $\{f \ge t + \varepsilon_n\}$ for all $x \in \{t \le f < t + \varepsilon_n\}$.

For all $\varepsilon \le \varepsilon_0$, denote by $r(\varepsilon) > 0$ the maximal radius of a tubular neighborhood of $\{f = t + \varepsilon\}$, the existence of which follows from the Tubular Neighborhood Theorem, i.e., $r(\varepsilon)$ is the largest number such that $\{x \in \mathbb{R}^d : \mathrm{dist}(x, \{f = t + \varepsilon\})\}$ is a tubular neighborhood of $\{f = t + \varepsilon\}$. Set $\rho = \inf_{0 \le \varepsilon \le \varepsilon_0} r(\varepsilon)$. Note that $\rho > 0$. Since $\varepsilon_n \to 0$, for all n large enough, we have

$$\{f = t\} \subset \mathcal{V}\big(\{f = t + \varepsilon_n\}, \rho\big).$$

Also, observe that in this case, $\{f = t + \varepsilon_n\} \subset \mathcal{V}(\{f = t\}, \rho)$. Thus, each $x \in \{f = t + \varepsilon_n\}$ may be expressed as $x = p_x + v_x e_{p_x}$, where $p_x \in \{f = t\}$ and where $v_x = \mathrm{dist}(x, \{f = t\})$. Expanding f at p_x yields

$$f(p_x + v_x e_{p_x}) = f(p_x) + D_{e_{p_x}} f(p_x + \xi e_{p_x}) v_x$$

i.e.,

$$t + \varepsilon_n = t + D_{e_{p_x}} f(p_x + \xi e_{p_x}) v_x$$

for some $\xi > 0$, and where $D_u f(y)$ denotes the directional derivative of f at y in the direction u. Since $\mathrm{d}f$ is surjective for all x in $\{t \le f < \varepsilon_0\}$, it follows that there exists a constant $C > 0$ such that

$$\sup_{q \in \{f = t + \varepsilon_n\}} \mathrm{dist}\big(q, \{f = t\}\big) \le C\varepsilon_n. \tag{A.1}$$

Consequently, since $\varepsilon_n = \mathrm{o}(r_n)$, for all n large enough, the ball $B(x, r_n)$ has a non-empty intersection with $\{f \ge t + \varepsilon_n\}$ for all $x \in \{t \le f < t + \varepsilon_n\}$.

Now, for all n large enough, each $x \in \{t \le f < t + \varepsilon_n\}$ may be expressed as $x = p_x - v_x e_{p_x}$, where $p_x \in \{f = t + \varepsilon_n\}$, and where $v_x > 0$. Also, for all n large enough, the two following assertions hold:

(*i*) $B(x, r_n) \subset \mathcal{V}\big(\{f = t + \varepsilon_n\}, \rho\big)$.

(*ii*) $B(x, r_n) \cap \{f \geq t + \varepsilon_n\} \neq \emptyset$.

Let $y = p_x + [(r_n - v_x)/2]e_{p_x}$, and consider the ball $B(y, (r_n - v_x)/2)$. Clearly,

$$B\Big(y, \frac{r_n - v_x}{2}\Big) \subset B(x, r_n).$$

Suppose that $B(y, (r_n - v_x)/2)$ is not included in $\{f \geq t + \varepsilon_n\}$. Then, there exists some point $q \in \{f = t + \varepsilon_n\}$ such that

$$\mathrm{dist}(q, y) < \frac{r_n - v_x}{2}.$$

But

$$\begin{aligned} \mathrm{dist}(q, y) &\geq \mathrm{dist}\big(y, \{f = t + \varepsilon_n\}\big) \\ &= \frac{r_n - v_x}{2}, \end{aligned}$$

hence a contradiction. Consequently,

$$B\Big(y, \frac{r_n - v_x}{2}\Big) \subset B(x, r_n) \cap \{f \geq t + \varepsilon_n\}. \tag{A.2}$$

From (A.2) and (A.1), it follows that

$$\mu\big(B(x, r_n) \cap \{f \geq t + \varepsilon_n\}\big) \geq \omega_d \left(r_n - C\varepsilon_n\right)^d,$$

where $\omega_d = \lambda(B(0, 1))$. Finally, the result follows from the fact that $\varepsilon_n = \mathrm{o}(r_n)$. □

Acknowledgments The authors are indebted to two referees for a very careful reading of the paper and stimulating questions and remarks.

References

[1] Bredon, G.E. (1993). *Topology and Geometry*, Vol. 139 of *Graduate Texts in Mathematics*, Springer-Verlag, New York.

[2] Brito, M.R., Chavez, E.L., Quiroz, A.J. and Yukich, J.E. (1997). Connectivity of the mutual k-nearest neighbor graph in clustering and outlier detection, *Statist. Probab. Lett.*, Vol. 35, pp. 33–42.

[3] Cadre, B. (2006). Kernel estimation of density level sets, *J. Multivariate Anal.*, Vol. 97, pp. 999–1023.

[4] Chavel, I. (1993). *Riemannian Geometry: A Modern Introduction*, Cambridge University Press, Cambridge.

[5] Cormen, T.H., Leiserson, C.E. and Rivest, R.L. (1990). *Introduction to Algorithms*, The MIT Press, Cambridge.

[6] Cuevas, A., Febrero, M. and Fraiman, R. (2000). Estimating the number of clusters, *Canad. J. Statist.*, Vol. 28, pp. 367–382.

[7] Cuevas, A., Febrero, M. and Fraiman, R. (2001). Cluster analysis: a further approach based on density estimation, *Comput. Statist. Data Anal.*, Vol. 36, pp. 441–459.

[8] Devroye, L. and Wise, G. (1980). Detection of abnormal behavior via nonparametric estimation of the support, *SIAM J. Appl. Math.*, Vol. 38, pp. 480–488.

[9] Duda, R.O., Hart, P.E. and Stork, D.G. (2000). *Pattern Classification*, Second Edition, Wiley-Interscience, New York.

[10] Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*, Springer-Verlag, New York.

[11] Hartigan, J.A. (1975). *Clustering Algorithms*, John Wiley, New York.

[12] Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York.

[13] Kobayashi, S. and Nomizu, K. (1996). *Foundations of Differential Geometry*, Vol. I & II, Second Edition, Wiley, New York.

[14] von Luxburg, U. and Ben-David, S. (2005). Towards a statistical theory of clustering, *PASCAL Workshop on Statistics and Optimization of Clustering*.

[15] Penrose, M.D. (1999). A strong law for the longest edge of the minimal spanning tree, *Ann. Probab.*, Vol. 27, pp. 246–260.

[16] Polonik, W. (1995). Measuring mass concentrations and estimating density contour clusters–an excess mass approach, *Ann. Statist.*, Vol. 23, pp. 855–881.

[17] Prakasa Rao, B.L.S. (1983). *Nonparametric Functional Estimation*, Academic Press, Orlando.

[18] Tsybakov, A.B. (1997). On nonparametric estimation of density level sets, *Ann. Statist.*, Vol. 25, pp. 948–969.