The Annals of Statistics
2015, Vol. 43, No. 5, 2132–2167
DOI: 10.1214/15-AOS1331

FULLY ADAPTIVE DENSITY-BASED CLUSTERING

BY INGO STEINWART[1]

University of Stuttgart

The clusters of a distribution are often defined by the connected components of a density level set. However, this definition depends on the user-specified level. We address this issue by proposing a simple, generic algorithm, which uses an almost arbitrary level set estimator to estimate the smallest level at which there are more than one connected components. In the case where this algorithm is fed with histogram-based level set estimates, we provide a finite sample analysis, which is then used to show that the algorithm consistently estimates both the smallest level and the corresponding connected components. We further establish rates of convergence for the two estimation problems, and last but not least, we present a simple, yet adaptive strategy for determining the width-parameter of the involved density estimator in a data-depending way.

1. Introduction. One definition of density-based clusters, which was first proposed by Hartigan [10], assumes i.i.d. data $D = (x_1, \ldots, x_n)$ generated by some unknown distribution P that has a continuous density h. For a user-defined threshold $\rho \geq 0$, the clusters of P are then defined to be the connected components of the level set $\{h \geq \rho\}$. This so-called single level approach has been studied by several authors; see, for example, [6, 10, 14, 17, 20] and the references therein. Unfortunately, however, different values of ρ may lead to different (numbers of) clusters (see, e.g., the illustrations in [5, 19]), and there is no generally accepted rule for choosing ρ, either. In addition, using a couple of different candidate values creates the problem of deciding which of the resulting clusterings is best. For this reason, Rinaldo and Wasserman [20] note that research on data-dependent, automatic methods for choosing ρ (and the width parameter of the involved density estimator) "would be very useful."

A second, density-based definition for clustering, which is known as the cluster tree approach, avoids this issue by considering all levels and the corresponding connected components simultaneously. Its focus thus lies on the identification of the hierarchical tree structure of the connected components for different levels; see, for example, [5, 10, 13, 27, 28] for details. For example, Chaudhuri and Dasgupta [5] show, under some assumptions on h, that a modified single linkage algorithm recovers this tree in the sense of [11], and Kpotufe and von Luxburg [13] obtain similar results for an underlying k-NN density estimator. In addition,

Received March 2015; revised March 2015.
[1]Supported by DFG Grant STE 1074/2-1.
MSC2010 subject classifications. Primary 62H30, 91C20; secondary 62G07.
Key words and phrases. Cluster analysis, consistency, rates, adaptivity.

Kpotufe and von Luxburg [13] propose a simple pruning strategy that removes connected components that artificially occur because of finite sample variability. However, the notion of recovery taken from [11] only focuses on the correct estimation of the cluster tree structure and not on the estimation of the clusters itself; cf. the discussion in [24].

Defining clusters by the connected components of one or more level sets clearly requires us to estimate level sets in one form or the other. Level set estimation itself is a classical nonparametric problem, which has been considered by various authors; see, for example, [1–3, 7, 12, 15, 16, 18, 21, 22, 26, 29]. In these articles, two different performance measures are considered for assessing the quality of a density level set estimate, namely the mass of the symmetric difference between the estimate and the true level set, and the Hausdorff distance between these two sets. Estimators that are consistent with respect to the Hausdorff metric clearly capture all topological structures eventually, so that these estimators form an almost canonical choice for density-based clustering with fixed level ρ. In contrast, level set estimators that are only consistent with respect to the first performance measure are, in general, not suitable for the cluster problem, since even sets that are equal up to measure zero may have completely different topological properties.

Another, very recent density-based cluster definition (see [4]) uses Morse theory to define the clusters of P. The idea of this approach is best illustrated by water flowing on a terrain. Namely, for each mode x_0 of h, the corresponding modal cluster is the set of points from which water flows, on the steepest descent path, to x_0 on the terrain described by $-h$. Under suitable smoothness assumptions on h, it turns out that these modal clusters form a partition of the input space modulo a Lebesgue zero set. Unlike in the single level approach, essentially all points of the input domain are thus assigned to a cluster. However, the required smoothness assumptions are somewhat strong, and so far, a consistent estimator has only been found for the one-dimensional case; see [4], Theorem 1.

In this work, we consider none of these approaches. Instead, we follow the approach of [24]; that is, we are interested in estimating (a) the infimum of all ρ at which the level set has more than one component and (b) the corresponding components. In addition, the usual continuity assumption on h is avoided. Let us therefore briefly describe the approach of [24] here; more details can be found in Section 2.

Its first step consists of defining level sets M_ρ that are *independent of the actual choice of the density*; see (2.1). Here we note that this independence is crucial for avoiding ambiguities when dealing with discontinuous densities. So far, some approaches have been made to address these difficulties. For example, Cuevas and Fraiman [6] introduced a thickness assumption for sets C that rules out cases in which neighborhoods of $x \in C$ have not sufficient mass. This thickness assumption excludes some topological pathologies such as topologically connecting bridges of zero mass, while others, such as cuts of measure zero, are not addressed. These issues are avoided in [20] by considering level sets of convolutions $k * P$ of the

underlying distribution P with a continuous kernel k on $\mathbb{R}^d$ having a compact support. Since such convolutions are always continuous, these authors cannot only deal with discontinuous densities, but also with distributions that do not have a Lebesgue density at all. However, different kernels or kernel widths may lead to different level sets, and consequently, their approach introduces new parameters that are hard to control by the user. In this respect, recall that for some other functionals of densities, Donoho [8] could remove these ambiguities, but so far it is unclear whether this is also possible for cluster analysis.

In a second step, the infimum ρ^* over all levels ρ for which M_ρ contains more than one connected component is considered. To reliably estimate ρ^*, it is further assumed that there exists some $\rho^{**} > \rho^*$ such that the component structure of M_ρ remains persistent for all $\rho \in (\rho^*, \rho^{**}]$. Note that such persistence is assumed either explicitly or implicitly in basically all density-based clustering approaches (see, e.g., [5, 13]), as it seems intuitively necessary for dealing with vertically uncertainty caused by finite sample effects. Another assumption imposed on P, namely that M_ρ has exactly two components between ρ^* and ρ^{**}, seems to be more restrictive at first glance. However, the opposite is true: if, for example, $h : [0, 1] \to (0, \infty)$ is a continuous density with exactly two distinct, strict local minima at say x_1 and x_2, then we only have more than two connected components in a small range above ρ^* if $h(x_1) = h(x_2)$. Compared to the case $h(x_1) \neq h(x_2)$, the latter seems to be rather singular, in particular, if one considers higher-dimensional analogs. Finally note that we could look for further splits of components above the level ρ^{**} in a similar fashion. This way we would recover the cluster tree approach, and, at least for the one-dimensional case, also the Morse approach by some trivial modifications already discussed in [4]. However, such an iterative approach is clearly out of the scope of this paper.

The first main result of this paper is a generic algorithm, which is based on an arbitrary level set estimator, for estimating both ρ^* and the corresponding clusters. In the case in which the underlying level set estimator enjoys guarantees on its vertical and horizontal uncertainty, we further provide an error analysis for both estimation problems in terms of these guarantees. A detailed statistical analysis is then conducted for histogram-based level set estimators. Here, our first result is a finite sample bound, which is then used to derive (as in [24]) consistency. We further provide rates of convergence for estimating ρ^* under an assumption on P that describes how fast the connected components of M_ρ move apart for increasing $\rho \in (\rho^*, \rho^{**}]$. The next main result establishes rates of convergence for estimating the clusters. Here we additionally need the well-known flatness condition of Polonik (see [16]) and an assumption that describes the mass of δ-tubes around the boundaries of the M_ρ's. Unlike previous articles, however, we do not need to restrict our considerations to (essentially) rectifiable boundaries. All these rates can only be achieved if the histogram width is chosen in a suitable, distribution-dependent way, and therefore we finally propose a simple data-driven parameter

selection strategy. Our last main result shows that this strategy often achieves the above rates without knowing characteristics of P.

Since this work strongly builds upon [23, 24], let us briefly describe our main *new* contributions. First, in [24], only the consistency of the histogram-based algorithm is established; that is, no rate of convergence is presented. While in [23], such rates are established, the situation considered in [23] is different. Indeed, in [23], an algorithm that uses a Parzen window density estimator to estimate the level sets is considered. However, this algorithm requires the density to be α-Hölder continuous for known α. Second, neither of the papers considers a data-dependent way of choosing the width parameter of the involved density estimator. Besides these new contributions, this paper also adds a substantial amount of extra information regarding the imposed assumptions and, last but not least, polishes many of the results from [24].

The rest of this paper is organized as follows. In Section 2 we recall the cluster definition from [24] and generalize the clustering algorithm from [24]. In Section 3 we provide a finite-sample analysis for the case, in which the generic algorithm is fed with plug-in estimates of a histogram. In Section 4 we then establish consistency and the new learning rates. Section 5 contains the description and the analysis of the new data-driven width selection strategy. Proofs of some of our results that are new, compared to those in [23, 24], can be found in Section 6. The remaining proofs, auxiliary results and an example of a large class of distributions on $\mathbb{R}^2$ with continuous densities that satisfy all the assumptions made in this paper can be found in [25].

2. Preliminaries: Level sets, clusters and a generic algorithm. In this section we recall and refine several notions related to the definition of clusters in [24]. In addition, we present a generic clustering algorithm, which is based on the ideas developed in [24].

Let us begin by fixing some notation and assumptions used throughout this paper: (X, d) is always a compact metric space, and $\mathcal{B}(X)$ denotes its Borel σ-algebra. Moreover, μ is a *known* σ-finite measure on $\mathcal{B}(X)$, and P is an *unknown* μ-absolutely continuous distribution on $\mathcal{B}(X)$ from which the data $D = (x_1, \ldots, x_n) \in X^n$ will be drawn in an i.i.d. fashion. In the following, we always assume that μ has full support, that is, $\operatorname{supp} \mu = X$. Of course, the example we are most interested in is that of $X = [0, 1]^d$ and μ being the Lebesgue measure on X, but alternatives such as the surface measure on a sphere are possible, too.

Given an $A \subset X$, we write $\mathring{A}$ for its interior, $\overline{A}$ for its closure and $\partial A := \overline{A} \setminus \mathring{A}$ for its boundary. Finally, $\mathbf{1}_A$ denotes the indicator function of A and $A \triangle B$, the symmetric difference of two sets A and B.

2.1. *Density-independent density level sets.* Unlike most papers dealing with density-based clustering, we will not assume that the data-generating distribution

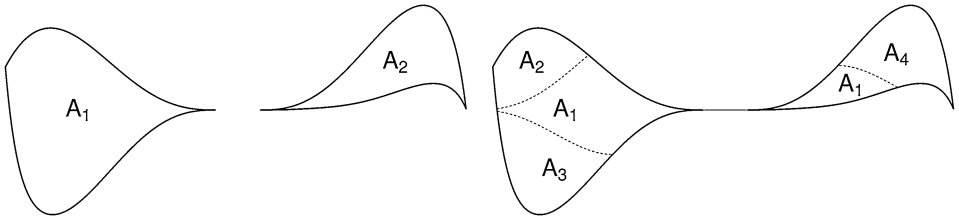

FIG. 1. *topologically relevant changes on sets of measure zero. Left*: *The thick solid lines indicate a set consisting of two connected components* A_1 *and* A_2. *If* $h = c\mathbf{1}_{A_1 \cup A_2}$ *is a density of* P *for a suitable constant* c, *then* A_1 *and* A_2 *are the connected components of* $\{h \geq \rho\}$ *for all* $\rho \in [0, c]$. *Right*: *This is a similar situation, but with topologically relevant changes on sets of measure zero. The straight horizontal thin line indicates a line of measure zero connecting the two components, and the dashed lines indicate cuts of measure zero. Clearly,* $h' := c\mathbf{1}_{A_1 \cup A_2 \cup A_3 \cup A_4}$ *is another density of* P, *but the connected components of* $\{h' \geq \rho\}$ *are the four sets* $A_1, \ldots, A_4$ *for all* $\rho \in [0, c]$.

P has a *continuous* density. Unfortunately, this generality makes it more challenging to define density-level-based clusters. Indeed, since the data is generated by P, we actually need to define clusters for distributions and not for densities. Consequently, a well-defined density-based notion of clusters either needs to be independent of the choice of the density, or pick, for each P, a somewhat canonical density. Now, if we assume that each considered P has a continuous density h, then these h's may serve as such canonical choices. In the absence of continuous densities, however, it is no longer clear how a "canonical" choice should look. In addition, the level sets of two different densities of the same P may have very distinct connected components (see, e.g., Figure 1) so that defining the clusters of P by the connected components of $\{h \geq \rho\}$ becomes inconsistent. In other words, neither of the two alternatives above is readily available for general P.

This issue is addressed in [24] by considering "density level sets" that are independent of the choice of the density. To recall this idea from [24], we fix an *arbitrary* μ-density h of P. Then, for every $\rho \geq 0$,

$$\mu_\rho(A) := \mu\big(A \cap \{h \geq \rho\}\big), \qquad A \in \mathcal{B}(X)$$

defines a σ-finite measure μ_ρ on $\mathcal{B}(X)$ that is actually *independent* of our choice of h. As a consequence, the set

$$M_\rho := \operatorname{supp} \mu_\rho, \tag{2.1}$$

which in [24] is called the density level set of P to the level ρ, is independent of this choice, too. It is shown in [24] (see also [25], Lemma A.1.1) that these sets are ordered in the usual way, that is, $M_{\rho_2} \subset M_{\rho_1}$ whenever $\rho_1 \leq \rho_2$. Furthermore, for any μ-density h of P, the definition immediately gives

$$\mu\big(\{h \geq \rho\} \setminus M_\rho\big) = \mu\big(\{h \geq \rho\} \cap (X \setminus M_\rho)\big) = \mu_\rho(X \setminus M_\rho) = 0; \tag{2.2}$$

that is, modulo μ-zero sets, the level sets $\{h \geq \rho\}$ are not larger than M_ρ. In fact, M_ρ turns out to be the smallest closed set satisfying (2.2), and it is shown in [24]

(see also [25], Lemma A.1.2) that we have both

$$\{h \overset{\circ}{\geq} \rho\} \subset M_\rho \subset \overline{\{h \geq \rho\}} \quad \text{and} \quad M_\rho \triangle \{h \geq \rho\} \subset \partial\{h \geq \rho\}. \tag{2.3}$$

For technical reasons we will not only need (2.2) but also the "converse" as well as a modification of (2.2). The exact requirements are introduced in the following definition, which slightly deviates from [24].

DEFINITION 2.1. We say that P is normal at level $\rho \geq 0$ if there exist two μ-densities h_1 and h_2 of P such that

$$\mu\big(M_\rho \setminus \{h_1 \geq \rho\}\big) = \mu\big(\{h_2 > \rho\} \setminus \overset{\circ}{M}_\rho\big) = 0.$$

Moreover, we say that P is normal if it is normal at every level.

It is shown in [25], Lemma A.1.3, that P is normal if it has both an upper semi-continuous μ-density h_1 and a lower semi-continuous μ-density h_2. Moreover, if P has a μ-density h such that $\mu(\partial\{h \geq \rho\}) = 0$, then P is normal at level ρ by (2.3). Finally, note that if the conditions of normality at level ρ are satisfied for some μ-densities h_1 and h_2 of P, then they are actually satisfied for all μ-densities h of P, and we have $\mu(M_\rho \triangle \{h \geq \rho\}) = 0$.

The remarks made above show that most distributions one would intuitively think of are normal. The next lemma demonstrates that there are also distributions that are not normal at a continuous range of levels.

LEMMA 2.2. *There exists a Lebesgue absolutely continuous distribution P on $[0, 1]$ and a $c > 0$ such that P is not normal at ρ for all $\rho \in (0, c]$.*

2.2. *Comparison of partitions and some notions of connectivity.* Following [24] we will define clusters with the help of connected components over a range of level sets. To prepare this definition, we recall some notions related to connectivity in this subsection. Moreover, we introduce a tool that makes it possible to compare the connected components of two level sets.

To motivate the following definition, which generalizes the ideas from [24], we note that the connected components of a set form a partition.

DEFINITION 2.3. Let $A \subset B$ be nonempty sets and $\mathcal{P}(A)$ and $\mathcal{P}(B)$ be partitions of A and B, respectively. Then $\mathcal{P}(A)$ is comparable to $\mathcal{P}(B)$, and we write $\mathcal{P}(A) \sqsubset \mathcal{P}(B)$ if, for all $A' \in \mathcal{P}(A)$, there is a $B' \in \mathcal{P}(B)$ with $A' \subset B'$.

Informally speaking, $\mathcal{P}(A)$ is comparable to $\mathcal{P}(B)$ if no cell $A' \in \mathcal{P}(A)$ is broken into pieces in $\mathcal{P}(B)$. In particular, if $\mathcal{P}_1$ and $\mathcal{P}_2$ are two partitions of A, then $\mathcal{P}_1 \sqsubset \mathcal{P}_2$ if and only if $\mathcal{P}_1$ is finer than $\mathcal{P}_2$.

Let us now assume that we have two partitions $\mathcal{P}(A)$ and $\mathcal{P}(B)$ such that $\mathcal{P}(A) \sqsubset \mathcal{P}(B)$. Then it is easy to see (cf. [25], Lemma A.2.1) that there exists a unique map $\zeta : \mathcal{P}(A) \to \mathcal{P}(B)$ such that, for all $A' \in \mathcal{P}(A)$, we have

$$A' \subset \zeta(A').$$

Following [24], we call ζ the cell relating map (CRM) between A and B. Moreover, we write $\zeta_{A,B} := \zeta$ when we want to emphasize the involved pair (A, B). Note that ζ is injective, if and only if no two distinct cells of $\mathcal{P}(A)$ are contained in the same cell of $\mathcal{P}(B)$. Conversely, ζ is surjective, if and only if every cell in $\mathcal{P}(B)$ contains a cell of $\mathcal{P}(A)$. Therefore, ζ is bijective, if and only if there is a structure preserving a one-to-one relation between the cells of the two partitions. In this case, we say that $\mathcal{P}(A)$ is *persistent* in $\mathcal{P}(B)$ and write $\mathcal{P}(A) \sqsubseteq \mathcal{P}(B)$.

The next lemma establishes a very useful composition formula for CRMs. For a proof, which is again inspired by [24], we refer to [25], Section A.2.

LEMMA 2.4. *Let $A \subset B \subset C$ be nonempty sets with partitions $\mathcal{P}(A)$, $\mathcal{P}(B)$ and $\mathcal{P}(C)$ such that $\mathcal{P}(A) \sqsubset \mathcal{P}(B)$ and $\mathcal{P}(B) \sqsubset \mathcal{P}(C)$. Then we have $\mathcal{P}(A) \sqsubset \mathcal{P}(C)$, and the corresponding CRMs satisfy*

$$\zeta_{A,C} = \zeta_{B,C} \circ \zeta_{A,B}.$$

The lemma above shows that the relations $\sqsubset$ and $\sqsubseteq$ are transitive. Moreover, if $\mathcal{P}(A) \sqsubseteq \mathcal{P}(C)$, then $\zeta_{A,B}$ must be injective, and $\zeta_{B,C}$ must be surjective, and we have $\mathcal{P}(A) \sqsubseteq \mathcal{P}(B)$ if and only if $\mathcal{P}(B) \sqsubseteq \mathcal{P}(C)$.

Now recall that an $A \subset X$ is (topologically) connected if, for every pair $A', A'' \subset A$ of relatively closed disjoint subsets of A with $A' \cup A'' = A$, we have $A' = \varnothing$ or $A'' = \varnothing$. The maximal connected subsets of A are called the connected components of A. It is well known that these components form a partition of A, which we denote by $\mathcal{C}(A)$. Moreover, for closed $A \subset B$ with $|\mathcal{C}(B)| < \infty$ we have $\mathcal{C}(A) \sqsubset \mathcal{C}(B)$; see [24] or [25], Lemma A.2.3.

Following [24], we will also consider a discrete version of path-connectivity. To recall the latter, we fix a $\tau > 0$ and an $A \subset X$. Then $x, x' \in A$ are τ-connected in A if there exist $x_1, \ldots, x_n \in A$ such that $x_1 = x$, $x_n = x'$ and $d(x_i, x_{i+1}) < \tau$ for all $i = 1, \ldots, n-1$. Clearly, being τ-connected gives an equivalence relation on A. We write $\mathcal{C}_\tau(A)$ for the resulting partition and call its cells the *τ-connected components of A*. It is shown in [24] (see also [25], Lemma A.2.7) that $\mathcal{C}_\tau(A) \sqsubset \mathcal{C}_\tau(B)$ for all $A \subset B$ and $\tau > 0$.

For a closed A and $\tau > 0$, we have $\mathcal{C}(A) \sqsubset \mathcal{C}_\tau(A)$ with a surjective CRM $\zeta : \mathcal{C}(A) \to \mathcal{C}_\tau(A)$; see [24] or [25], Proposition A.2.10. To characterize, when this CRM is even bijective, let us assume that $1 < |\mathcal{C}(A)| < \infty$. Then

$$\tau_A^* := \min\{d(A', A'') : A', A'' \in \mathcal{C}(A) \text{ with } A' \neq A''\} \tag{2.4}$$

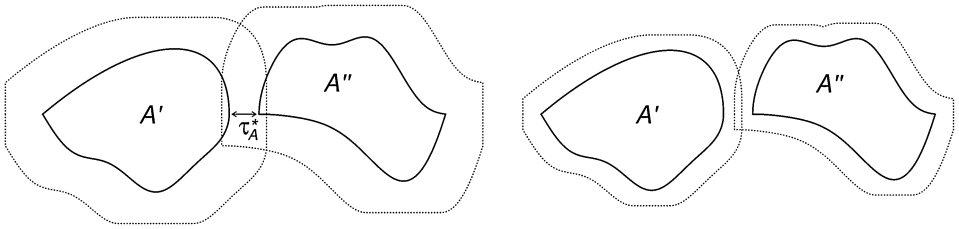

FIG. 2. *The role of τ_A^*. Left: A set A consisting of two connected components A' and A'' drawn in solid lines. The dotted lines indicate the contours of the set of all points that are within τ-distance of A', respectively A'', for some fixed $\tau > \tau_A^*$ and the sup-norm. Since there are some elements in A'' that are within τ-distance of A', there is only one τ-connected component, namely A. The CRM $\zeta : \mathcal{C}(A) \to \mathcal{C}_\tau(A)$ is thus surjective but not injective. Right: Here we have the same situation for some $\tau < \tau_A^*$. In this case, A' and A'' are also the τ-connected components of A, and the CRM $\zeta : \mathcal{C}(A) \to \mathcal{C}_\tau(A)$ is bijective.*

denotes the minimal distance between mutually different components of $\mathcal{C}(A)$. Now it is shown in [24] (or [25], Proposition A.2.10) that

$$\mathcal{C}(A) = \mathcal{C}_\tau(A) \quad \Longleftrightarrow \quad \tau \in (0, \tau_A^*];$$

see also Figure 2 for an illustration. In other words, τ_A^* is the largest (horizontal) granularity τ at which the connected components of A are not glued together. Finally, this threshold is ordered for closed $A \subset B$ in the sense that $\tau_A^* \geq \tau_B^*$ whenever $|\mathcal{C}(A)| < \infty$, $|\mathcal{C}(B)| < \infty$, and the CRM $\zeta : \mathcal{C}(A) \to \mathcal{C}(B)$ is injective. We refer to [24] or [25], Lemma A.2.11.

2.3. *Clusters.* Using the concepts developed in the previous subsections, we can now recall the definition of clusters from [24].

DEFINITION 2.5. The distribution P can be clustered between $\rho^* \geq 0$ and $\rho^{**} > \rho^*$ if P is normal and for all $\rho \in [0, \rho^{**}]$, the following three conditions are satisfied:

(i) we have either $|\mathcal{C}(M_\rho)| = 1$ or $|\mathcal{C}(M_\rho)| = 2$;
(ii) if we have $|\mathcal{C}(M_\rho)| = 1$, then $\rho \leq \rho^*$;
(iii) if we have $|\mathcal{C}(M_\rho)| = 2$, then $\rho \geq \rho^*$ and $\mathcal{C}(M_{\rho^{**}}) \sqsubseteq \mathcal{C}(M_\rho)$.

Using the CRMs $\zeta_\rho : \mathcal{C}(M_{\rho^{**}}) \to \mathcal{C}(M_\rho)$, we then define the clusters of P by

$$A_i^* := \bigcup_{\rho \in (\rho^*, \rho^{**}]} \zeta_\rho(A_i), \qquad i \in \{1, 2\},$$

where A_1 and A_2 are the two topologically connected components of $M_{\rho^{**}}$.

By conditions (iii) and (ii), we find $\rho < \rho^* \Rightarrow |\mathcal{C}(M_\rho)| = 1 \Rightarrow \rho \leq \rho^*$ as well as $\rho > \rho^* \Rightarrow |\mathcal{C}(M_\rho)| = 2 \Rightarrow \rho \geq \rho^*$ for all $\rho \in [0, \rho^{**}]$; see also Figure 3. At each

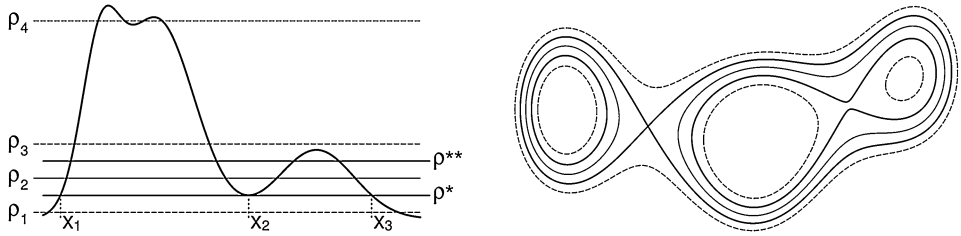

FIG. 3. *Definition of clusters. Left*: *A* 1*-dimensional mixture of three Gaussians together with the level* ρ^* *and a possible choice for* ρ^{**}*. The component structure at level* $\rho_2 \in (\rho^*, \rho^{**})$ *coincides with that at level* ρ^{**}*, while for* $\rho_1 < \rho^*$*, we only have one connected component. The levels* $\rho_3, \rho_4 > \rho^{**}$ *are not considered by Definition* 2.5*, and thus the component structure at these levels is arbitrary. Finally, the clusters of the distribution are the open intervals* (x_1, x_2) *and* (x_2, x_3)*. Right*: *Here we have a similar situation for a mixture of three* 2*-dimensional Gaussians drawn by contour lines. The thick solid lines again indicate the levels* ρ^* *and* ρ^{**}*, and the thin solid lines show a level* $\rho \in (\rho^*, \rho^{**})$*. The dashed lines correspond to a level* $\rho < \rho^*$ *and a level* $\rho > \rho^{**}$*. This time the clusters are the two connected components of the open set that is surrounded by the outer thick solid line.*

level below ρ^* there is thus only one component, while there are two components at all levels in between ρ^* and ρ^{**}. Moreover, in both cases the corresponding partitions are persistent.

Since all ζ_ρ's are bijective, we find $\zeta_\rho(A_1) \cap \zeta_\rho(A_2) = \varnothing$ for all $\rho \in (\rho^*, \rho^{**}]$, and using $\zeta_\rho(A_1) \nearrow A_i^*$ for $\rho \searrow \rho^*$, we conclude that $A_1^* \cap A_2^* = \varnothing$. In general, the sets A_i^* are neither open nor closed, and we may have $d(A_1^*, A_2^*) = 0$; that is, the clusters may touch each other; see again Figure 3.

2.4. *Cluster persistence under horizontal uncertainty.* In general, we can only expect nonparametric estimates of M_ρ that are both vertically and horizontally uncertain. To some extent the vertical uncertainty, which is caused by the estimation error, has already been addressed by the persistence assumed in our cluster definition. In this subsection, we complement this by recalling tools from [24] for dealing with horizontal uncertainty, which is usually caused by the approximation error.

To quantify horizontal uncertainty, we need for $A \subset X$, $\delta > 0$, the sets

$$A^{+\delta} := \{x \in X : d(x, A) \le \delta\},$$
$$A^{-\delta} := X \setminus (X \setminus A)^{+\delta},$$

where $d(x, A) := \inf_{x' \in A} d(x, x')$ denotes the distance between x and A. Simply speaking, adding a δ-tube to A gives $A^{+\delta}$, while removing a δ-tube gives $A^{-\delta}$. These operations, as well as closely related operations based on the Minkowski addition and difference have already been used in the literature on level set estimation; see, for example, [30]. Some simple properties of these operations can be found in [25], Lemma A.3.1.

Now let L_ρ be an estimate of M_ρ having vertical and horizontal uncertainty in the sense of

$$M_{\rho+\varepsilon}^{-\delta} \subset L_\rho \subset M_{\rho-\varepsilon}^{+\delta},$$

for some $\varepsilon, \delta > 0$. Ideally, we additionally have $\mathcal{C}(M_{\rho+\varepsilon}^{-\delta}) \sqsubseteq \mathcal{C}(L_\rho) \sqsubseteq \mathcal{C}(M_{\rho-\varepsilon}^{+\delta})$. To reliably use $\mathcal{C}(L_\rho)$ as an estimate of $\mathcal{C}(M_\rho)$, it then suffices to know $\mathcal{C}(M_{\rho+\varepsilon}^{-\delta}) \sqsubseteq \mathcal{C}(M_\rho) \sqsubseteq \mathcal{C}(M_{\rho-\varepsilon}^{+\delta})$. Unfortunately, however, the latter is typically not true. Indeed, even in the absence of horizontal uncertainty, we do not have $\mathcal{C}(M_{\rho+\varepsilon}) \sqsubseteq \mathcal{C}(M_{\rho-\varepsilon})$ if $\rho + \varepsilon > \rho^*$ and $\rho - \varepsilon < \rho^*$. Moreover, in the absence of vertical uncertainty, we usually do not have $\mathcal{C}(M_\rho^{-\delta}) \sqsubseteq \mathcal{C}(M_\rho) \sqsubseteq \mathcal{C}(M_\rho^{+\delta})$, either, as components of $\mathcal{C}(M_\rho)$ may be glued together in $\mathcal{C}(M_\rho^{+\delta})$ or cut apart in $\mathcal{C}(M_\rho^{-\delta})$; see Figure 5. To repair such cuts, our algorithm will consider τ-connected components instead of connected components. In the rest of this section we thus investigate under which conditions we do have $\mathcal{C}_\tau(M_{\rho+\varepsilon}^{-\delta}) \sqsubseteq \mathcal{C}(M_\rho) \sqsubseteq \mathcal{C}_\tau(M_{\rho-\varepsilon}^{+\delta})$. We begin with the following definition taken from [24] that excludes bridges and cusps that are too thin.

DEFINITION 2.6. We say that P has thick level sets of order $\gamma \in (0, 1]$ up to the level $\rho^{**} > 0$, if there exist constants $c_{\text{thick}} \geq 1$ and $\delta_{\text{thick}} \in (0, 1]$ such that, for all $\delta \in (0, \delta_{\text{thick}}]$ and $\rho \in [0, \rho^{**}]$, we have

$$\sup_{x \in M_\rho} d\big(x, M_\rho^{-\delta}\big) \leq c_{\text{thick}} \delta^\gamma. \tag{2.5}$$

In this case, we call $\psi(\delta) := 3c_{\text{thick}}\delta^\gamma$ the thickness function of P.

Thickness assumptions have been widely used in the literature on level set estimation (see, e.g., [22]), where the case $\gamma = 1$ is considered. To some extent, the latter is a natural choice, as is discussed in detail in [25], Section A.3. In particular, for $d = 1$ we always have $\gamma = 1$, and for $d = 2$ [25], Example B.2.1, provides a rich class of continuous densities with $\gamma = 1$. Figure 4 illustrates how different shapes of level sets lead to different γ's.

The following result, which summarizes some findings from [24] (see also [25], Theorems A.4.2 and A.4.4), provides an answer to our persistence question.

THEOREM 2.7. *Assume that P can be clustered between ρ^* and ρ^{**} and that it has thick level sets of order γ up to ρ^{**}. Let ψ be its thickness function. Using (2.4), we define the function $\tau^* : (0, \rho^{**} - \rho^*] \to (0, \infty)$ by*

$$\tau^*(\varepsilon) := \tfrac{1}{3}\tau^*_{M_{\rho^*+\varepsilon}}. \tag{2.6}$$

Then τ^* is increasing, and for all $\varepsilon^* \in (0, \rho^{**} - \rho^*]$, $\delta \in (0, \delta_{\text{thick}}]$, $\tau \in (\psi(\delta), \tau^*(\varepsilon^*)]$ and all $\rho \in [0, \rho^{**}]$, the following statements hold:

(i) *we have $1 \leq |\mathcal{C}_\tau(M_\rho^{+\delta})| \leq 2$ and $1 \leq |\mathcal{C}_\tau(M_\rho^{-\delta})| \leq 2$;*

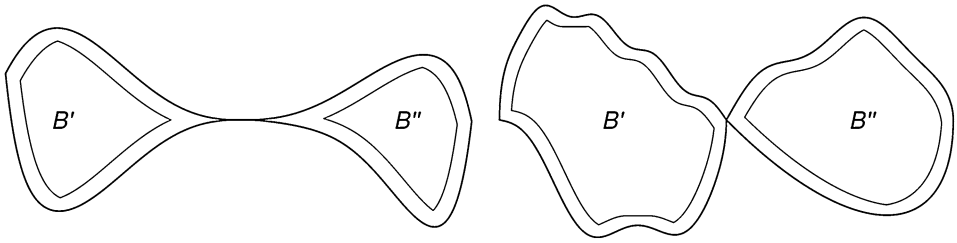

FIG. 4. *Thick level sets. Left: The thick solid line indicates a level set M_ρ below or at the level ρ^*, and the thin solid lines show the two components B' and B'' of $M_\rho^{-\delta}$. Because of the quadratic shape of M_ρ around the thin bridge, the set M_ρ has thickness of order $\gamma = 1/2$. Right: Here we have the same situation for a distribution that has thick level sets of order $\gamma = 1$. Note that the smaller γ on the left leads to a significantly wider separation of B' and B'' than on the right, which in turn requires larger τ to glue the parts together.*

(ii) *if $\rho < \rho^*$ or $\rho \geq \rho^* + \varepsilon^*$, then we have*

$$\mathcal{C}_\tau\bigl(M_\rho^{-\delta}\bigr) \sqsubseteq \mathcal{C}(M_\rho) = \mathcal{C}_\tau(M_\rho) \sqsubseteq \mathcal{C}_\tau\bigl(M_\rho^{+\delta}\bigr).$$

Theorem 2.7 in particular shows that for sufficiently small δ and τ, the component structure of M_ρ is not changed when δ-tubes are added or removed and τ-connected components are considered instead. Not surprisingly, however, the meaning of "sufficiently small," which is expressed by the functions τ^* and ψ, changes when we approach the level ρ^* from above. Moreover, note that even for sufficiently small δ and τ, Theorem 2.7 does not specify the structure of $\mathcal{C}_\tau(M_\rho^{-\delta})$ and $\mathcal{C}_\tau(M_\rho^{+\delta})$ at the levels $\rho \in [\rho^*, \rho^* + \varepsilon^*)$. In fact, for such ρ, the components of M_ρ may be accidentally glued together in $\mathcal{C}_\tau(M_\rho^{+\delta})$; see, for example, Figure 5. This effect complicates our analysis significantly.

Let us now summarize the assumptions that will be used in the following.

ASSUMPTION C. We have a compact metric space (X, d), a finite Borel measure μ on X with $\operatorname{supp}\mu = X$ and a μ-absolutely continuous distribution P that can be clustered between ρ^* and ρ^{**}. In addition, P has thick level sets of order $\gamma \in (0, 1]$ up to the level ρ^{**}. We denote the corresponding thickness function by ψ and write τ^* for the function defined in (2.6).

2.5. *A generic clustering algorithm and its analysis.* In this section, we present and analyze a generic version of the clustering algorithm from [24]. The main difference between our algorithm and the algorithm of [24] is that our generic algorithm can use any level set estimator that has control over both its vertical and horizontal uncertainty.

Our first result, which is a generic version of [24], Theorem 24, relates the component structure of a family of level set estimates to the component structure of certain sets $M_{\rho+\varepsilon}^{-\delta}$. For a proof we refer to [25], Section A.6.

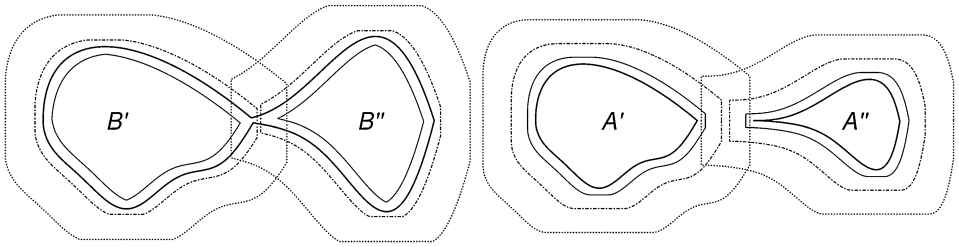

FIG. 5. *Difficulties around ρ^*. Left: The thick solid line indicates an M_ρ for $\rho < \rho^*$, and the thin solid lines show $M_\rho^{-\delta}$. While M_ρ consists of one connected component, $M_\rho^{-\delta}$ has two such components, B' and B'', and hence $\mathcal{C}(M_\rho^{-\delta})$ is not persistent in $\mathcal{C}(M_\rho)$. The two types of dotted lines indicate the set of all points that are within τ-distance of B', respectively B'' for two values of τ. Only for the larger τ we have $\mathcal{C}_\tau(M_\rho^{-\delta}) \sqsubseteq \mathcal{C}(M_\rho)$; that is, in this case τ-connectivity does glue the separated regions together. Right: The thick solid lines indicate an M_ρ for some $\rho \in (\rho^*, \rho^{**}]$ having two connected components, A' and A'', and thin solid lines show the two components of $M_\rho^{+\delta}$. The two types of dotted lines indicate the set of all points that are within τ-distance of $(A')^{+\delta}$, respectively $(A'')^{+\delta}$ for the two values of τ used left. This time, we have $\mathcal{C}(M_\rho) \sqsubseteq \mathcal{C}_\tau(M_\rho^{+\delta})$ only for the smaller value of τ. Together, these graphics thus illustrate that good values for δ and τ at one level may be bad at a different level. However, Theorem 2.7 shows that this undesired behavior can be excluded with the help of the functions τ^* and ψ for all levels $\rho \notin [\rho^*, \rho^* + \varepsilon^*)$.*

THEOREM 2.8. *Let Assumption C be satisfied. Furthermore, let $\varepsilon^* \in (0, \rho^{**} - \rho^*]$, $\delta \in (0, \delta_{\text{thick}}]$, $\tau \in (\psi(\delta), \tau^*(\varepsilon^*)]$ and $\varepsilon \in (0, \varepsilon^*]$. In addition, let $(L_\rho)_{\rho \geq 0}$ be a decreasing family of sets $L_\rho \subset X$ such that*

$$M_{\rho+\varepsilon}^{-\delta} \subset L_\rho \subset M_{\rho-\varepsilon}^{+\delta} \tag{2.7}$$

holds for all $\rho \geq 0$. Then, for all $\rho \in [0, \rho^{**} - 3\varepsilon]$ and the corresponding CRMs $\zeta : \mathcal{C}_\tau(M_{\rho+\varepsilon}^{-\delta}) \to \mathcal{C}_\tau(L_\rho)$, the following disjoint union holds:

$$\mathcal{C}_\tau(L_\rho) = \zeta\big(\mathcal{C}_\tau\big(M_{\rho+\varepsilon}^{-\delta}\big)\big) \cup \big\{B' \in \mathcal{C}_\tau(L_\rho) : B' \cap L_{\rho+2\varepsilon} = \varnothing\big\}. \tag{2.8}$$

Theorem 2.8 shows that for suitable δ, ε and τ, all τ-connected components B' of L_ρ are *either* contained in the image $\zeta(\mathcal{C}_\tau(M_{\rho+\varepsilon}^{-\delta}))$ *or* vanish at level $\rho + 2\varepsilon$, that is, $B' \cap L_{\rho+2\varepsilon} = \varnothing$. Now assume we can detect the latter components. By Theorem 2.8 we can then identify the τ-connected components B' that are contained in $\zeta(\mathcal{C}_\tau(M_{\rho+\varepsilon}^{-\delta}))$, and if, in addition, ζ is injective, these identified components have the same structure as $\mathcal{C}_\tau(M_{\rho+\varepsilon}^{-\delta})$. By Theorem 2.7 we can further hope that $\mathcal{C}_\tau(M_{\rho+\varepsilon}^{-\delta}) \sqsubseteq \mathcal{C}(M_{\rho+\varepsilon})$, so that we can relate the identified components to those of $\mathcal{C}(M_{\rho+\varepsilon})$. Assuming these steps can be carried out precisely, we obtain Algorithm 1; see also Figure 6, which scans through the values of ρ from small to large and stops as soon as it identifies either no component or at least two.

The following theorem provides bounds for the level ρ_D^* and the components $B_i(D)$ returned by Algorithm 1. It extends the analysis from [24].

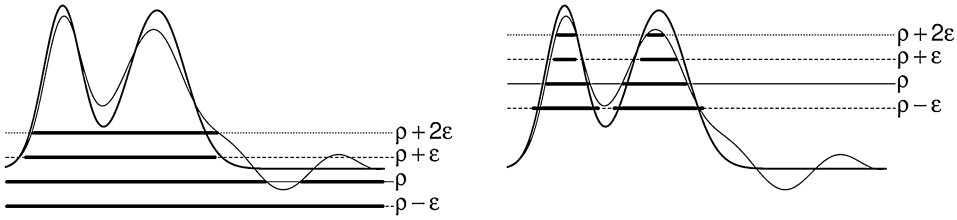

FIG. 6. *Illustration of Algorithm 1 Left: A density (thick solid line) having two modes on the left and a flat part on the right. A plug-in approach based on a density estimate (thin solid line with three modes) is used to provide the level set estimator L_ρ (bold horizontal line at level ρ), which satisfies $M^{-\delta}_{\rho+\varepsilon} \subset L_\rho \subset M^{+\delta}_{\rho-\varepsilon}$. Only the left component of L_ρ does not vanish at $\rho + 2\varepsilon$, and thus Algorithm 1 identifies only one component at its line 3. Right: Here we have the same situation at a higher level. This time both components of L_ρ do not vanish at $\rho + 2\varepsilon$, and hence Algorithm 1 identifies two components at its line 3.*

THEOREM 2.9. *Let Assumption C be satisfied. Furthermore, let $\varepsilon^* \le (\rho^{**} - \rho^*)/9$, $\delta \in (0, \delta_{\text{thick}}]$, $\tau \in (\psi(\delta), \tau^*(\varepsilon^*)]$ and $\varepsilon \in (0, \varepsilon^*]$. In addition, let D be a data set and $(L_{D,\rho})_{\rho\ge 0}$ be a decreasing family satisfying (2.7) for all $\rho \ge 0$. Then the following statements are true for Algorithm 1:*

(i) *the returned level ρ^*_D satisfies both $\rho^*_D \in [\rho^* + 2\varepsilon, \rho^* + \varepsilon^* + 5\varepsilon]$ and*

$$\tau - \psi(\delta) < 3\tau^*\big(\rho^*_D - \rho^* + \varepsilon\big); \tag{2.9}$$

Algorithm 1 Clustering with the help of a generic level set estimator

Require: Some $\tau > 0$ and $\varepsilon > 0$.
A decreasing family $(L_{D,\rho})_{\rho\ge 0}$ of subsets of X.
Ensure: An estimate of ρ^* and the clusters A^*_1 and A^*_2.
1: $\rho \leftarrow 0$
2: **repeat**
3: Identify the τ-connected components $B'_1, \ldots, B'_M$ of $L_{D,\rho}$ satisfying

$$B'_i \cap L_{D,\rho+2\varepsilon} \ne \varnothing.$$

4: $\rho \leftarrow \rho + \varepsilon$
5: **until** $M \ne 1$
6: $\rho \leftarrow \rho + 2\varepsilon$
7: Identify the τ-connected components $B'_1, \ldots, B'_M$ of $L_{D,\rho}$ satisfying

$$B'_i \cap L_{D,\rho+2\varepsilon} \ne \varnothing.$$

8: **return** $\rho^*_D := \rho$ and the sets $B_i(D) := B'_i$ for $i = 1, \ldots, M$.

(ii) *algorithm 1 returns two sets* $B_1(D)$ *and* $B_2(D)$, *and these sets can be ordered such that we have*

$$\sum_{i=1}^{2} \mu\bigl(B_i(D) \triangle A_i^*\bigr) \le 2 \sum_{i=1}^{2} \mu\bigl(A_i^* \setminus \bigl(A^i_{\rho_D^*+\varepsilon}\bigr)^{-\delta}\bigr) + \mu\bigl(M^{+\delta}_{\rho_D^*-\varepsilon} \setminus \{h > \rho^*\}\bigr). \tag{2.10}$$

Here, $A^i_{\rho_D^*+\varepsilon} \in \mathcal{C}(M_{\rho_D^*+\varepsilon})$ *are ordered in the sense of* $A^i_{\rho_D^*+\varepsilon} \subset A_i^*$.

3. Finite sample analysis of a histogram-based algorithm. In this section, we consider the case where the level set estimates $L_{D,\rho}$ fed into Algorithm 1 are produced by a histogram. The main result in this section shows that the error estimates of Theorem 2.9 hold with high probability.

To ensure (2.7), we will use, as in [24], partitions that are geometrically well behaved. To this end, recall that the diameter of an $A \subset X$ is

$$\operatorname{diam} A := \sup\bigl\{d(x, x') : x, x' \in A\bigr\}.$$

Now, the assumptions made on the used partitions are as follows:

ASSUMPTION A. For each $\delta \in (0, 1]$, $\mathcal{A}_\delta = (A_1, \ldots, A_{m_\delta})$ is a partition of X. Moreover, there exist constants $\mathrm{d} > 0$ and $c_{\mathrm{part}} \ge 1$ such that, for all $\delta \in (0, 1]$ and $i = 1, \ldots, m_\delta$, we have

$$\operatorname{diam} A_i \le \delta, \qquad m_\delta \le c_{\mathrm{part}} \delta^{-\mathrm{d}} \quad \text{and} \quad \mu(A_i) \ge c_{\mathrm{part}}^{-1} \delta^{\mathrm{d}}.$$

The most important examples of families of partitions satisfying Assumption A are hyper-cube partitions of $X \subset \mathbb{R}^d$ in combination with the Lebesgue measure; see [25], Example A.7.1, for details. Other situations in which partitions satisfying Assumption A can be found include spheres $X := \mathbb{S}^d \subset \mathbb{R}^{d+1}$ together with their surface measures and $\mathrm{d} = d - 1$, sufficiently compact metric groups in combination their Haar measure and *known*, sufficiently smooth d-dimensional sub-manifolds equipped their surface measure. For details we refer to [25], Lemma A7.2 and Corollary A.7.3.

Let us now assume that Assumption A is satisfied. Moreover, for a data set $D = (x_1, \ldots, x_n) \in X^n$ we denote, in a slight abuse of notation, the corresponding empirical measure by D, that is, $D := \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$, where δ_x is the Dirac measure at x. Then the resulting histogram is

$$h_{D,\delta}(x) = \sum_{j=1}^{m_\delta} \frac{D(A_j)}{\mu(A_j)} \cdot \mathbf{1}_{A_j}(x), \qquad x \in X. \tag{3.1}$$

The following theorem provides a finite sample analysis for using the plug-in estimates $L_{D,\rho} := \{h_{D,\delta} \ge \rho\}$ in Algorithm 1.

THEOREM 3.1. *Let Assumptions A and C be satisfied. For a fixed* $\delta \in (0, \delta_{\text{thick}}]$, $\varsigma \geq 1$, $n \geq 1$ *and* $\tau > \psi(\delta)$, *we fix an* $\varepsilon > 0$ *satisfying the bound*

$$\varepsilon \geq c_{\text{part}} \sqrt{\frac{E_{\varsigma,\delta}}{2\delta^{2\mathrm{d}} n}}, \tag{3.2}$$

where $E_{\varsigma,\delta} := \varsigma + \ln(2c_{\text{part}}) - \mathrm{d} \ln \delta$, *or if* P *has a bounded* μ*-density* h, *the bound*

$$\varepsilon \geq \sqrt{\frac{2c_{\text{part}}(1 + \|h\|_\infty) E_{\varsigma,\delta}}{\delta^{\mathrm{d}} n}} + \frac{2c_{\text{part}} E_{\varsigma,\delta}}{3\delta^{\mathrm{d}} n}. \tag{3.3}$$

We further pick an $\varepsilon^* > 0$ *satisfying*

$$\varepsilon^* \geq \varepsilon + \inf\{\varepsilon' \in (0, \rho^{**} - \rho^*] : \tau^*(\varepsilon') \geq \tau\}. \tag{3.4}$$

For each data set $D \in X^n$, *we now feed Algorithm 1 with the parameters* τ *and* ε, *and with the family* $(L_{D,\rho})_{\rho \geq 0}$ *given by*

$$L_{D,\rho} := \{h_{D,\delta} \geq \rho\}, \qquad \rho \geq 0.$$

If $\varepsilon^* \leq (\rho^{**} - \rho^*)/9$, *then with probability* P^n *not less than* $1 - e^{-\varsigma}$, *we have a* $D \in X^n$ *satisfying the assumptions and conclusions of Theorem 2.9.*

At this point we like to emphasize that a finite sample bound in the form of Theorem 3.1 can be derived from our analysis whenever Algorithm 1 uses a density level set estimator guaranteeing the inclusions $M^{-\delta}_{\rho+\varepsilon} \subset L_{D,\rho} \subset M^{+\delta}_{\rho-\varepsilon}$ with high probability. A possible example of such an alternative level set estimator is a plug-in approach based on a moving window density estimator, since for the latter it is possible to establish a uniform convergence result similar to [25], Theorem A.8.1; see, for example, [9, 23]. Unfortunately, the resulting level sets become *computationally unfeasible* when used naïvely, and hence we have not included this approach here. It is, however, an interesting open question, whether sets $L_{D,\rho}$ that are constructed differently from the moving window estimator can address this issue. So far, the only known result in this direction [23] constructs such sets for α-Hölder-continuous densities h with *known* α, but we conjecture that a similar construction may be possible for general h, too. In addition, strategies such as approximating the sets $L_{D,\rho}$ by fine grids may be feasible, at least for small dimensions, too.

4. Consistency and rates. The first goal of this section is to use the finite sample bound of Theorem 3.1 to show that Algorithm 1 estimates both ρ^* and the clusters A^*_i consistently. We then introduce some assumptions on P that lead to convergence rates for both estimation problems.

The following consistency result is a modification of [24], Theorem 26; see also [25], Section A.9, for a corresponding modification of its proof.

THEOREM 4.1. *Let Assumptions A and C be satisfied, and let* (ε_n), (δ_n) *and* (τ_n) *be strictly positive sequences converging to zero such that* $\psi(\delta_n) < \tau_n$ *for all sufficiently large* n, *and*

$$\lim_{n\to\infty} \frac{\ln \delta_n^{-1}}{n\delta_n^{2\mathrm{d}}\varepsilon_n^2} = 0. \tag{4.1}$$

For $n \geq 1$, *consider Algorithm 1 with the input parameters* ε_n, τ_n *and the family* $(L_{D,\rho})_{\rho\geq 0}$ *given by* $L_{D,\rho} := \{h_{D,\delta_n} \geq \rho\}$. *Then, for all* $\epsilon > 0$, *we have*

$$\lim_{n\to\infty} P^n\big(\big\{D \in X^n : 0 < \rho_D^* - \rho^* \leq \epsilon\big\}\big) = 1,$$

and if $\mu(\overline{A_i^* \cup A_2^*} \setminus (A_1^* \cup A_2^*)) = 0$, *we also have*

$$\lim_{n\to\infty} P^n\big(\big\{D \in X^n : \mu\big(B_1(D) \triangle A_1^*\big) + \mu\big(B_2(D) \triangle A_2^*\big) \leq \epsilon\big\}\big) = 1,$$

where, for $B_1(D)$ *and* $B_2(D)$, *we use the same numbering as in* (2.10).

Note that the assumption $\mu(\overline{A_i^* \cup A_2^*} \setminus (A_1^* \cup A_2^*)) = 0$ is satisfied if there exists a μ-density h of P such that $\mu(\partial\{h \leq \rho^*\}) = 0$; see [25], Section A.9.

Theorem 4.1 shows that for suitably chosen parameters and histogram-based level set estimates Algorithm 1 asymptotically recovers both ρ^* and the clusters A_1^* and A_2^*, if the distribution P has level sets that are thicker than a user-specified order γ. To illustrate this, suppose that we choose $\delta_n \sim n^{-\alpha}$ and $\varepsilon_n \sim n^{-\beta}$ for some $\alpha, \beta > 0$. Then it is easy to check that (4.1) is satisfied if and only if $2(\alpha\mathrm{d} + \beta) < 1$. For $\tau_n \sim n^{-\alpha\gamma} \ln n$, we then have $\psi(\delta_n) < \tau_n$ for all sufficiently large n, and therefore, Algorithm 1 recovers the clusters for all distributions P that have thick levels of order γ. Similarly, the choice $\tau_n \sim (\ln n)^{-1}$ leads to consistency for all distributions P that have thick levels of some order $\gamma > 0$. Finally note that (4.1) can be replaced by

$$\frac{\ln \delta_n^{-1}}{n\delta_n^{\mathrm{d}}\varepsilon_n^2} \to 0$$

if we restrict our consideration to distributions with bounded μ-densities. The proof of this is a straightforward modification of the proof of Theorem 4.1.

To give two examples, recall from the discussion in [25], Section A.5, that for the one-dimensional case $X = [a, b]$, we always have $\gamma = 1$. In two dimensions this is, however, no longer true as, for example, Figure 4 illustrates. Nonetheless, there do exist many examples of both discontinuous and continuous densities for which we have thickness $\gamma = 1$; see [25], Section B.2. Finally note that the construction used there can be easily generalized to higher dimensions.

For our next goal, which is establishing rates for both $\mu(B_i(D) \triangle A_i^*) \to 0$ and $\rho_D^* \to \rho^*$, we need, as usual, some assumptions on P. Let us begin by introducing an assumption that leads to rates for the estimation of ρ^*.

DEFINITION 4.2. Let Assumption C be satisfied. Then the clusters of P have separation exponent $\kappa \in (0, \infty]$ if there is a constant $\underline{c}_{\text{sep}} > 0$ such that

$$\tau^*(\varepsilon) \geq \underline{c}_{\text{sep}} \varepsilon^{1/\kappa}$$

for all $\varepsilon \in (0, \rho^{**} - \rho^*]$. Moreover, the separation exponent κ is exact if there exists another constant $\overline{c}_{\text{sep}} > 0$ such that, for all $\varepsilon \in (0, \rho^{**} - \rho^*]$, we have

$$\tau^*(\varepsilon) \leq \overline{c}_{\text{sep}} \varepsilon^{1/\kappa}.$$

The separation exponent describes how fast the connected components of the M_ρ approach each other for $\rho \searrow \rho^*$. Note that the separation exponent is monotone, that is, a distribution having separation exponent κ also has separation exponent κ' for all $\kappa' < \kappa$. In particular, the "best" separation exponent is $\kappa = \infty$, and this exponent describes distributions, for which we have $d(A_1^*, A_2^*) \geq \underline{c}_{\text{sep}}$; that is, the clusters A_1^* and A_2^* do not touch each other.

To illustrate the separation exponent, let us consider $X := -[3, 3]$ and, for $\theta, \beta \in (0, \infty]$ and $\rho^* \in [0, 1/6)$, the distribution $P_{\theta,\beta}$ that has the density

$$h_{\theta,\beta}(x) := \rho^* + c_{\theta,\beta}\bigl(\mathbf{1}_{[0,1]}(|x|)|x|^\theta + \mathbf{1}_{[1,2]}(|x|) + \mathbf{1}_{[2,3]}(|x|)(3 - |x|)^\beta\bigr), \tag{4.2}$$

where $c_{\theta,\beta}$ is a constant ensuring that $h_{\theta,\beta}$ is a probability density; see also Figure 7 for two examples. Note that $P_{\theta,\beta}$ can be clustered between ρ^* and $\rho^{**} := \rho^* + c_{\theta,\beta}$. Moreover, $P_{\theta,\beta}$ always has exact separation exponent θ.

The polynomial behavior in the upper vicinity of ρ^* of the distributions (4.2) is somewhat archetypal for smooth densities on $\mathbb{R}$. For example, for C^2-densities h whose first derivative h' has exactly one zero x_0 in the set $\{h = \rho^*\}$ and whose second derivative satisfies $h''(x_0) > 0$, one can easily show with the help of Taylor's theorem that their behavior in the upper vicinity of ρ^* is asymptotically identical to that of (4.2) for $\kappa = \theta = 2$ and $\beta = 1$. Moreover, larger values for $\kappa = \theta$ can be achieved by assuming that higher derivatives of h vanish at x_0. Analogously, the class of continuous densities on $\mathbb{R}^2$ from [25], Section B.2, have separation exponent $\kappa = 2$ (see [25], Example B.2.1), as these densities, similar to Morse functions, behave like $x_1^2 - x_2^2$ in the vicinity of the saddle point. Again, the construction can be modified to achieve other exponents.

In the following we show how the separation exponent influences the rate for estimating ρ^*. We begin with a finite sample bound.

THEOREM 4.3. *Let Assumptions A and C be satisfied, and assume additionally that P has a bounded μ-density h and that its clusters have separation exponent $\kappa \in (0, \infty]$. For some fixed $\delta \in (0, \delta_{\text{thick}}]$, $\varsigma \geq 1$, $n \geq 1$ and $\tau \geq 2\psi(\delta)$, we pick an $\varepsilon > 0$ satisfying (3.3), that is,*

$$\varepsilon \geq \sqrt{\frac{2c_{\text{part}}(1 + \|h\|_\infty)(\varsigma + \ln(2c_{\text{part}}) - d\ln\delta)}{\delta^d n}} + \frac{2c_{\text{part}}(\varsigma + \ln(2c_{\text{part}}) - d\ln\delta)}{3\delta^d n}.$$

Let us assume that $\varepsilon^* := \varepsilon + (\tau/\underline{c}_{\mathrm{sep}})^\kappa$ *satisfies* $\varepsilon^* \le (\rho^{**} - \rho^*)/9$. *Then if Algorithm 1 receives the input parameters* ε, τ *and the family* $(L_{D,\rho})_{\rho\ge 0}$ *given by* $L_{D,\rho} := \{h_{D,\delta} \ge \rho\}$, *the probability* P^n *of a* $D \in X^n$ *that satisfies*

$$\varepsilon < \rho_D^* - \rho^*, \tag{4.3}$$

$$\rho_D^* - \rho^* \le (\tau/\underline{c}_{\mathrm{sep}})^\kappa + 6\varepsilon \tag{4.4}$$

is not less than $1 - e^{-\varsigma}$. *Moreover, if the separation exponent* κ *is exact and* $\kappa < \infty$, *then we can replace* (4.3) *by*

$$\frac{1}{4}\left(\frac{\tau}{6\overline{c}_{\mathrm{sep}}}\right)^\kappa + \varepsilon < \rho_D^* - \rho^*. \tag{4.5}$$

The finite sample guarantees of Theorem 4.3 can be easily used to derive (exact) rates for $\rho_D^* \to \rho^*$. The following corollary presents, modulo (double) logarithmic factors, the best rates we can derive by this approach.

COROLLARY 4.4. *Let Assumptions A and C be satisfied, and assume that* P *has bounded* μ*-density and that its clusters have separation exponent* $\kappa \in (0, \infty)$. *Furthermore, let* (ε_n), (δ_n) *and* (τ_n) *be sequences with*

$$\varepsilon_n \sim \left(\frac{\ln n \cdot \ln\ln n}{n}\right)^{\gamma\kappa/(2\gamma\kappa+\mathrm{d})}, \qquad \delta_n \sim \left(\frac{\ln n}{n}\right)^{1/(2\gamma\kappa+\mathrm{d})} \quad \text{and} \quad \tau_n \sim \varepsilon_n^{1/\kappa},$$

and assume that, for $n \ge 1$, *Algorithm 1 receives the input parameters* ε_n, τ_n *and the family* $(L_{D,\rho})_{\rho\ge 0}$ *given by* $L_{D,\rho} := \{h_{D,\delta_n} \ge \rho\}$. *Then there exists a constant* $\overline{K} \ge 1$ *such that for all sufficiently large* n, *we have*

$$P^n\big(\{D \in X^n : \rho_D^* - \rho^* \le \overline{K}\varepsilon_n\}\big) \ge 1 - \frac{1}{n}. \tag{4.6}$$

Moreover, if the separation exponent κ *is exact, there exists another constant* $\underline{K} \ge 1$ *such that for all sufficiently large* n, *we have*

$$P^n\big(\{D \in X^n : \underline{K}\varepsilon_n \le \rho_D^* - \rho^* \le \overline{K}\varepsilon_n\}\big) \ge 1 - \frac{1}{n}. \tag{4.7}$$

Finally, if $\kappa = \infty$, *then* (4.7) *holds for all sufficiently large* n *if*

$$\varepsilon_n \sim \left(\frac{\ln n \cdot \ln\ln n}{n}\right)^{1/2}, \qquad \delta_n \sim (\ln\ln n)^{-1/(2\mathrm{d})} \quad \text{and} \quad \tau_n \sim (\ln\ln n)^{-\gamma/(3\mathrm{d})}.$$

Recall that for the one-dimensional distributions (4.2) we have $\gamma = 1$ and $\kappa = \theta$, so that the exponent in the rates above becomes $\frac{\theta}{2\theta+1}$. In particular, for the C^2-case discussed there, we have $\theta = 2$, and thus we get a rate with exponent $2/5$, while for $\theta \to \infty$ the exponent converges to $1/2$. Similarly, for the typical, two-dimensional distributions considered in [25], Section B.2, we have $\gamma = 1$, $\kappa = 2$ and $\mathrm{d} = 2$, and hence the exponent in the rate is $1/3$.

Our next goal is to establish rates for $\mu(B_i(D) \triangle A_i^*) \to 0$. Since this is a modified level set estimation problem, let us recall some assumptions on P, which have been used in this context. The first assumption in this direction is a one-sided variant of a well-known condition introduced by Polonik [16].

DEFINITION 4.5. Let μ be a finite measure on X and P be a distribution on X that has a μ-density h. For a given level $\rho \geq 0$, we say that P has flatness exponent $\vartheta \in (0, \infty]$ if there exists a constant $c_{\text{flat}} > 0$ such that

$$\mu\bigl(\{0 < h - \rho < s\}\bigr) \leq (c_{\text{flat}} s)^{\vartheta}, \qquad s > 0. \tag{4.8}$$

Clearly, the larger the ϑ, the more steeply h must approach ρ from above. In particular, for $\vartheta = \infty$, the density h is allowed to take the value ρ but is otherwise bounded away from ρ. For example, the densities in (4.2) have a flatness exponent $\vartheta = \min\{1/\theta, 1/\beta\}$ if $\theta < \infty$ and $\beta < \infty$ and a flatness exponent $\vartheta = \infty$ if $\theta = \beta = \infty$. Finally, for the two-dimensional distributions of [25], Section B.2, the flatness exponent is not fully determined by their definition, but some calculations show that we have $\vartheta \in (0, 1]$.

Next, we describe the roughness of the boundary of the clusters.

DEFINITION 4.6. Let Assumption C be satisfied. Given some $\alpha \in (0, 1]$, the clusters have an α-smooth boundary if there exists a constant $c_{\text{bound}} > 0$ such that, for all $\rho \in (\rho^*, \rho^{**}]$, $\delta \in (0, \delta_{\text{thick}}]$ and $i = 1, 2$, we have

$$\mu\bigl((A_\rho^i)^{+\delta} \setminus (A_\rho^i)^{-\delta}\bigr) \leq c_{\text{bound}} \delta^{\alpha}, \tag{4.9}$$

where A_ρ^1 and A_ρ^2 denote the two connected components of the level set M_ρ.

In $\mathbb{R}^d$, considering $\alpha > 1$ does not make sense, and for an $A \subset \mathbb{R}^d$ with rectifiable boundary, we always have $\alpha = 1$; see [25], Lemma A.10.4. The α-smoothness of the boundary thus enforces a uniform version of this, which, however, is not very restrictive; see, for example, the densities of (4.2), for which we have $\alpha = 1$ and $c_{\text{bound}} = 4$, and [25], Example B.2.2, for which we also have $\alpha = 1$.

The following assumption collects all conditions we need to impose on P to get rates for estimating the clusters.

ASSUMPTION R. Assumptions A and C are satisfied, and P has a bounded μ-density h. Moreover, P has a flatness exponent $\vartheta \in (0, \infty]$ at level ρ^*, its clusters have an α-smooth boundary for some $\alpha \in (0, 1]$ and its clusters have a separation exponent $\kappa \in (0, \infty]$.

Let us now investigate how well our algorithm estimates the clusters A_1^* and A_2^*. As usual, we begin with a finite-sample estimate.

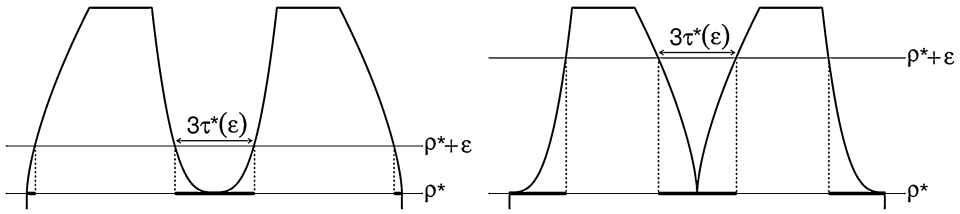

FIG. 7. *Separation and flatness. Left: The density $h_{\theta,\beta}$ described in (4.2) for $\theta = 3$ and $\beta = 2/3$. The bold horizontal line indicates the set $\{\rho^* < h < \rho^* + \varepsilon\}$, and $3\tau^*(\varepsilon)$ describes the width of the valley at level $\rho^* + \varepsilon$. Right: Here we have the same situation for $\theta = 2/3$ and $\beta = 3$. The value of ε is chosen such that $3\tau^*(\varepsilon)$ equals the value on the left. The smaller value of θ narrows the valley, and hence ε needs to be chosen larger. As a result, it becomes more difficult to estimate ρ^* and the clusters. Indeed, ignoring logarithmic factors, Corollary 4.4 gives a rate of $n^{-3/7}$ on the left and a rate of $n^{-2/7}$ on the right, while Corollary 4.8 gives a rate of $n^{-1/7}$ on the left and a rate of $n^{-2/21}$ on the right. Finally, in the most typical case $\theta = 2$ and $\beta = 1$ not illustrated here, we obtain the rates $n^{-1/3}$ and $n^{-1/5}$.*

THEOREM 4.7. *Let Assumption R be satisfied, and assume that δ, ε, τ, ε^*, ς, n and $(L_{D,\rho})_{\rho\geq 0}$ are as in Theorem 4.3. Then the probability P^n of having a data set $D \in X^n$ satisfying (4.3), (4.4) and*

$$\mu\big(B_1(D) \triangle A_1^*\big) + \mu\big(B_2(D) \triangle A_2^*\big) \leq 6c_{\text{bound}}\delta^\alpha + \big(c_{\text{flat}}(\tau/\underline{c}_{\text{sep}})^\kappa + 7c_{\text{flat}}\varepsilon\big)^\vartheta$$

is not less than $1 - e^{-\varsigma}$, where the sets $B_1(D)$ and $B_2(D)$ are ordered as in (2.10). Moreover, if the separation exponent κ is exact and satisfies $\kappa < \infty$, then (4.5) also holds for these data sets D.

Note that for *finite* values of ϑ and κ, the bound in Theorem 4.7 behaves like $\delta^\alpha + \tau^{\vartheta\kappa} + \varepsilon^\vartheta$, and in this case it is thus easy to derive the best convergence rates our analysis yields. The following corollary presents corresponding results and also provides rates for the cases $\vartheta = \infty$ or $\kappa = \infty$.

COROLLARY 4.8. *Assume that Assumption R is satisfied, and write $\varrho := \min\{\alpha, \vartheta\gamma\kappa\}$. Furthermore, let (ε_n), (δ_n) and (τ_n) be sequences with*

$$\varepsilon_n \sim \left(\frac{\ln n}{n}\right)^{\varrho/(2\varrho+\vartheta\mathrm{d})} (\ln\ln n)^{-\vartheta\mathrm{d}/(8\varrho+4\vartheta\mathrm{d})},$$

$$\delta_n \sim \left(\frac{\ln n \cdot \ln\ln n}{n}\right)^{\vartheta/(2\varrho+\vartheta\mathrm{d})} \quad \text{and}$$

$$\tau_n \sim \left(\frac{\ln n \cdot (\ln\ln n)^2}{n}\right)^{\vartheta\gamma/(2\varrho+\vartheta\mathrm{d})}.$$

Assume that, for $n \geq 1$, Algorithm 1 receives the parameters ε_n, τ_n and the family $(L_{D,\rho})_{\rho\geq 0}$ given by $L_{D,\rho} := \{h_{D,\delta_n} \geq \rho\}$. Then there is a constant $\overline{K} \geq 1$ such

that, for all $n \geq 1$ *and the ordering as in* (2.10), *we have*

$$P^n\left(D : \sum_{i=1}^{2} \mu\bigl(B_i(D) \triangle A_i^*\bigr) \leq K\left(\frac{\ln n \cdot (\ln\ln n)^2}{n}\right)^{\vartheta\varrho/(2\varrho+\vartheta\mathrm{d})}\right) \geq 1 - \frac{1}{n}.$$

Let us now compare the established rates for estimating ρ^* and the clusters in the most important case, that is, $\alpha = 1$. If $\vartheta\gamma\kappa \leq 1$, we obtain $\varrho = \vartheta\gamma\kappa$ in Corollary 4.8, and the exponent in the asymptotic behavior of the optimal (δ_n) becomes $\frac{1}{2\gamma\kappa+\mathrm{d}}$. Since this equals the exponent in Corollary 4.4, and, modulo the extra $\ln\ln n$ terms, we also have the same behavior for (ε_n) and (τ_n) in both corollaries, we conclude that we obtain the rates in Corollaries 4.4 and 4.8 with (essentially) the *same* controlling sequences (ε_n), (δ_n) and (τ_n) of Algorithm 1. If $\vartheta\gamma\kappa \leq 1$, we can thus achieve the best rates for estimating ρ^* and the clusters *simultaneously*. Unfortunately, this changes if $\vartheta\gamma\kappa > 1$. Indeed, while the exponent for (δ_n) in Corollary 4.4 remains the same, it changes from $\frac{1}{2\gamma\kappa+\mathrm{d}}$ to $\frac{\vartheta}{2+\vartheta\mathrm{d}}$ in Corollary 4.8, and a similar effect takes place for the sequences (ε_n) and (τ_n). The reason for this difference is that in the case $\vartheta\gamma\kappa > 1$ the estimation of ρ^* is easier than the estimation of the level set M_{ρ^*}, and since for estimating the clusters we need to do both, the level set estimation rate determines the rate for estimating the clusters.

To illustrate this difference between the estimation of ρ^* and the clusters in more detail, let us consider the toy model (4.2) in the case $\theta = \beta = \infty$, that is, $\kappa = \infty$. Then the clusters are stumps, and the sets M_ρ do not change between ρ^* and ρ^{**}. Intuitively, the best choice for estimating ρ^* are then sufficiently small but fixed values for δ_n and τ_n, so that ε_n converges to 0 as fast as possible. In Corollary 4.4 this is mimicked by choosing very slowly decaying sequences (δ_n) and (τ_n). On the other hand, to find A_1^* and A_2^* it suffices to identify one $\rho \in (\rho^*, \rho^{**}]$ and to estimate the connected components of M_ρ. The best way to achieve this is to use a sufficiently small but fixed value for ε_n and sequences (δ_n) and (τ_n) that converge to zero as fast as possible. In Corollary 4.8 this is mimicked by choosing a very slowly decaying sequence (ε_n) and quickly decaying sequences (δ_n) and (τ_n).

As for estimating the critical level ρ^*, we do not know so far, whether our rates for estimating the clusters are minmax optimal, but our conjecture is that they are optimal modulo the logarithmic terms. To motivate our conjecture, let us consider the case $\alpha = \gamma = 1$. Moreover, assume that two-sided versions of [25], (A.10.4) and (A.10.6), hold for all $\rho \in (\rho^*, \rho^{**}]$, respectively, $\rho = \rho^*$. Then we have $\kappa = \theta$ and $\vartheta = 1/\theta$ by [25], Lemmas A.10.1 and A.10.5, and thus we find $\varrho = 1$. Consequently, the rates in Corollary 4.8 have the exponent $\frac{1}{2\theta+d}$. This is exactly the same exponent as the one obtained in [22] for minmax optimal and adaptive Hausdorff estimation of a fixed level set. In addition, it seems that their lower bound, which is based on [29], is, modulo logarithmic factors, the same for assessing the estimator in the way we have done it in Corollary 4.8. While this

coincidence indicates that our rates may be (essentially) optimal, it is, of course, not a rigorous argument. A detailed analysis is, however, out of the scope of this paper. Another interesting question, which is also out of the scope, is whether the estimates $B_i(D)$ approximate the true clusters A_i^* in the Hausdorff metric, too, and if so, whether we can achieve the rates reported in [22].

5. Data-dependent parameter selection. In the last section we derived rates of convergence for both the estimation of ρ^* and the clusters. In both cases, our best rates required sequences (ε_n), (δ_n) and (τ_n) that did depend on some properties of P, namely α, κ, ϑ. Of course, these parameters are not available to us in practice, and therefore the obtained rates are of little practical value. The goal of this final section is to address this issue by proposing a simple data-dependent parameter selection strategy that is able to recover the rates of Corollary 4.4 without knowing anything about P. We further show that this selection strategy recovers the rates of Corollary 4.8 in the case of $\vartheta\gamma\kappa \le \alpha$.

We begin by presenting the parameter selection strategy. To this end, let $\Delta \subset (0, 1]$ be finite and $n \ge 1$, $\varsigma \ge 1$. For $\delta \in \Delta$, we fix a $\tau_{\delta,n} > 0$ and define

$$\varepsilon_{\delta,n} := C\sqrt{\frac{c_{\mathrm{part}}(\varsigma + \ln(2c_{\mathrm{part}}|\Delta|) - \mathrm{d}\ln\delta)\ln\ln n}{\delta^{\mathrm{d}} n}} + \frac{2c_{\mathrm{part}}(\varsigma + \ln(2c_{\mathrm{part}}|\Delta|) - \mathrm{d}\ln\delta)}{3\delta^{\mathrm{d}} n}, \tag{5.1}$$

where $C \ge 1$ is some user-specified constant. Now assume that, for each $\delta \in \Delta$, we run Algorithm 1 with the parameters $\varepsilon_{\delta,n}$ and $\tau_{\delta,n}$, and the family $(L_{D,\rho})_{\rho\ge 0}$ given by $L_{D,\rho} := \{h_{D,\delta} \ge \rho\}$. We write $\rho^*_{D,\delta}$ for the corresponding level returned by Algorithm 1. Let us consider a width $\delta^*_{D,\Delta} \in \Delta$ that achieves the smallest returned level, that is,

$$\delta^*_{D,\Delta} \in \arg\min_{\delta\in\Delta} \rho^*_{D,\delta}. \tag{5.2}$$

Note that in general, this width may not be uniquely determined, so that in the following we need to additionally assume that we have a well-defined choice, for example, the smallest $\delta \in \Delta$ satisfying (5.2). Moreover, we write

$$\rho^*_{D,\Delta} := \rho^*_{D,\delta^*_{D,\Delta}} = \min_{\delta\in\Delta} \rho^*_{D,\delta} \tag{5.3}$$

for the smallest returned level. Note that unlike $\delta^*_{D,\Delta}$, the level $\rho^*_{D,\Delta}$ is always unique. Finally, we define $\varepsilon_{D,\Delta} := \varepsilon_{\delta^*_{D,\Delta},n}$ and $\tau_{D,\Delta} := \tau_{\delta^*_{D,\Delta},n}$.

Our first goal is to show that $\rho^*_{D,\Delta}$ achieves the rates of Corollary 4.4 for suitably chosen Δ and $\tau_{\delta,n}$. We begin with a finite sample guarantee.

THEOREM 5.1. *Let Assumptions A and C be satisfied, and assume that P has a bounded μ-density h, and that the two clusters of P have separation exponent*

$\kappa \in (0,\infty]$. *For a fixed finite* $\Delta \subset (0, \delta_{\text{thick}}]$, *and* $n \geq 1$, $\varsigma \geq 1$ *and* $C \geq 1$, *we define* $\varepsilon_{\delta,n}$ *by* (5.1) *and choose* $\tau_{\delta,n}$ *such that* $\tau_{\delta,n} \geq 2\psi(\delta)$ *for all* $\delta \in \Delta$. *Furthermore, assume that* $C^2 \ln\ln n \geq 2(1+\|h\|_\infty)$ *and* $\varepsilon^*_\delta := \varepsilon_{\delta,n} + (\tau_{\delta,n}/\underline{c}_{\text{sep}})^\kappa \leq (\rho^{**}-\rho^*)/9$ *for all* $\delta \in \Delta$. *Then we have*

$$P^n\Big(\Big\{D \in X^n : \varepsilon_{D,\Delta} < \rho^*_{D,\Delta} - \rho^* \leq \min_{\delta\in\Delta}\big((\tau_{\delta,n}/\underline{c}_{\text{sep}})^\kappa + 6\varepsilon_{\delta,n}\big)\Big\}\Big) \geq 1 - e^{-\varsigma}.$$

Moreover, if the separation exponent κ *is exact and* $\kappa < \infty$, *then the assumptions above actually guarantee*

$$P^n\Big(D : \min_{\delta\in\Delta}\big(c_1\tau^\kappa_{\delta,n} + \varepsilon_{\delta,n}\big) < \rho^*_{D,\Delta} - \rho^* \leq \min_{\delta\in\Delta}\big(c_2\tau^\kappa_{\delta,n} + 6\varepsilon_{\delta,n}\big)\Big) \geq 1 - e^{-\varsigma},$$

where $c_1 := \frac{1}{4}(6\overline{c}_{\text{sep}})^{-\kappa}$ *and* $c_2 := \underline{c}_{\text{sep}}^{-\kappa}$, *and similarly*

$$P^n\big(\big\{D \in X^n : c_1\tau^\kappa_{D,\Delta} + \varepsilon_{D,\Delta} < \rho^*_{D,\Delta} - \rho^* \leq c_2\tau^\kappa_{D,\Delta} + 6\varepsilon_{D,\Delta}\big\}\big) \geq 1 - e^{-\varsigma}.$$

Theorem 5.1 establishes the same finite sample guarantees for the estimator $\rho^*_{D,\Delta}$ as Theorem 4.3 did for the simpler estimator ρ^*_D. Therefore, it is not surprising that for suitable choices of Δ, the rates of Corollary 4.4 can be recovered, too. The next corollary shows that this can actually be achieved for candidate sets Δ that are completely independent of P.

COROLLARY 5.2. *Assume that Assumptions* A *and* C *are satisfied, that* P *has a bounded* μ*-density* h *and that the two clusters of* P *have separation exponent* $\kappa \in (0,\infty]$. *For* $n \geq 16$, *we consider the interval*

$$I_n := \left[\left(\frac{\ln n \cdot (\ln\ln n)^2}{n}\right)^{1/\mathrm{d}}, \left(\frac{1}{\ln\ln n}\right)^{1/\mathrm{d}}\right]$$

and fix some $n^{-1/\mathrm{d}}$*-net* $\Delta_n \subset I_n$ *of* I_n *with* $|\Delta_n| \leq n$. *Furthermore, for some fixed* $C \geq 1$ *and* $n \geq 16$, *we write* $\tau_{\delta,n} := \delta^\gamma \ln\ln\ln n$ *and define* $\varepsilon_{\delta,n}$ *by* (5.1) *for all* $\delta \in \Delta_n$ *and* $\varsigma = \ln n$. *Then there exists a constant* $\overline{K}$ *such that, for all sufficiently large* n, *we have*

$$P^n\left(D : \varepsilon_{D,\Delta_n} < \rho^*_{D,\Delta_n} - \rho^* \leq \overline{K}\left(\frac{\ln n \cdot (\ln\ln n)^2}{n}\right)^{\gamma\kappa/(2\gamma\kappa+\mathrm{d})}\right) \geq 1 - \frac{1}{n}. \tag{5.4}$$

If, in addition, the separation exponent κ *is exact and* $\kappa < \infty$, *then there is another constant* $\underline{K}$ *such that for all sufficiently large* n, *we have*

$$\begin{aligned} P^n\Bigg(D : \underline{K}\left(\frac{\ln n \cdot \ln\ln n}{n}\right)^{\gamma\kappa/(2\gamma\kappa+\mathrm{d})} &\leq \rho^*_{D,\Delta_n} - \rho^* \\ &\leq \overline{K}\left(\frac{\ln n \cdot (\ln\ln n)^2}{n}\right)^{\gamma\kappa/(2\gamma\kappa+\mathrm{d})}\Bigg) \\ &\geq 1 - \frac{1}{n}. \end{aligned}$$

Finally, we show that our parameter selection strategy partially recovers the rates for estimating the clusters A_i^* obtained in Corollary 4.8.

COROLLARY 5.3. *Assume that Assumption* R *is satisfied with* $\alpha \geq \vartheta\gamma\kappa$ *and exact separation exponent* κ. *Then, for the procedure of Corollary* 5.2, *there is a* $K \geq 1$ *such that for* $n \geq 1$ *and the ordering as in* (2.10), *we have*

$$P^n\left(D : \sum_{i=1}^{2} \mu\big(B_i(D) \triangle A_i^*\big) \leq K\left(\frac{\ln n \cdot (\ln\ln n)^2}{n}\right)^{\vartheta\gamma\kappa/(2\gamma\kappa+\vartheta\mathrm{d})}\right) \geq 1 - \frac{1}{n}.$$

Unfortunately, the simple parameter selection strategy (5.2) is not adaptive in the case $\alpha < \vartheta\gamma\kappa$, that is, in the case in which the estimation of ρ^* is easier than the estimation of the corresponding clusters. It is unclear to us whether in this case a two-stage procedure that first estimates ρ^* by ρ^*_{D,Δ_n} as above, and then uses a different strategy to estimate the connected components at the level ρ^*_{D,Δ_n} can be made adaptive.

6. Selected proofs. In this section we present some selected proofs. All remaining proofs can be found in [25].

PROOF OF LEMMA 2.2. Let (x_n) be an enumeration of $\mathbb{Q} \cap [0,1]$ and $I_n := [x_n - 2^{-n-2}, x_n + 2^{-n-2}] \cap [0,1]$ for $n \geq 1$. For $x \in [0,1]$ and $I_0 := [0,1]$, we further define

$$f(x) := \sup_{n\geq 0} n\mathbf{1}_{I_n}(x),$$

that is, $f(x)$ equals the largest integer $n \geq 0$ (including infinity) such that $x \in I_n$. For $c > 0$ specified below, we now define

$$h(x) := \begin{cases} 2c - \dfrac{c}{f(x)}, & \text{if } f(x) > 0, \\ 0, & \text{else.} \end{cases}$$

Then h is measurable, nonnegative and Lebesgue-integrable, and hence we can choose c such that $\int_0^1 h(x)\,dx = 1$. Then h is a density of a Lebesgue-absolutely continuous distribution P. Moreover, note that $h(x) \geq 2c - c/n$ for all $x \in I_n$ and $n \geq 1$. For a fixed $\rho \in (0, 2c)$ we now write $n_\rho := c/(2c-\rho)$. Then we have $2c - c/n \geq \rho$ if and only if $n \geq n_\rho$. Consequently, the set

$$A_\rho := \bigcup_{n\geq n_\rho}^{\infty} \mathring{I}_n$$

satisfies $A_\rho \subset \{h \geq \rho\}$. Moreover, since A_ρ is open, we find $A_\rho \subset \{h \mathring{\geq} \rho\}$, and thus

$$\overline{A_\rho} \subset \overline{\{h \mathring{\geq} \rho\}} \subset M_\rho$$

by [25], Lemma A.1.2. In addition, we have $\{x_n : n \geq n_\rho\} \subset A_\rho$, and since the former set is dense in $[0,1]$, we conclude that $M_\rho = [0,1]$. On the other hand, the Lebesgue measure λ of $\{h \geq \rho\}$ can be estimated by

$$\lambda\bigl(\{h \geq \rho\}\bigr) \leq \lambda\bigl(\{h > 0\}\bigr) = \lambda\left(\bigcup_{n=1}^{\infty} I_n\right) \leq \sum_{n=1}^{\infty} \lambda(I_n) \leq \sum_{n=1}^{\infty} 2^{-n-1} = \frac{1}{2},$$

and hence we conclude that $\lambda(M_\rho \setminus \{h \geq \rho\}) \geq 1/2$. In other words, P is not normal at level ρ. $\square$

PROOF OF THEOREM 2.7. The monotonicity of τ^* is shown in [25], Theorem A.4.2, and (i) follows from parts (i) of [25], Theorems A.4.2 and A.4.4.

(ii) Let us first consider the case $\rho < \rho^*$. Since P can be clustered, we have $|\mathcal{C}(M_\rho)| = 1$, and [25], Proposition A.2.10, gives both $\tau^*_{M_\rho} = \infty$ and $\mathcal{C}(M_\rho) = \mathcal{C}_\tau(M_\rho)$. By [25], Lemma A.4.1, we further find $\mathcal{C}_\tau(M_\rho) \sqsubseteq \mathcal{C}_\tau(M_\rho^{+\delta})$. Finally, part (ii) of [25], Lemma A.4.3, yields $1 \leq |\mathcal{C}_\tau(M_\rho^{-\delta})| \leq |\mathcal{C}(M_\rho)| = 1$, and hence its part (iii) gives the persistence $\mathcal{C}_\tau(M_\rho^{-\delta}) \sqsubseteq \mathcal{C}(M_\rho)$.

In the case $\rho \geq \rho^* + \varepsilon^*$, $\mathcal{C}_\tau(M_\rho) \sqsubseteq \mathcal{C}_\tau(M_\rho^{+\delta})$ follows from part (ii) of [25], Theorem A.4.2, and the equality $\mathcal{C}(M_\rho) = \mathcal{C}_\tau(M_\rho)$ follows from [25], Proposition A.2.10, in combination with $\tau \leq \tau^*(\varepsilon^*) \leq \tau^*(\rho - \rho^*)$. By part (ii) of [25], Theorem A.4.4, we further know $\mathcal{C}_\tau(M_{\rho^{**}}^{-\delta}) \sqsubseteq \mathcal{C}_\tau(M_{\rho^{**}}^{+\delta})$. Using $\rho \geq \rho^* + \varepsilon^*$ and part (iv) of [25], Theorem A.4.2, we find $|\mathcal{C}_\tau(M_\rho^{-\delta})| = 2$, and hence part (iii) of [25], Theorem A.4.4, gives $\mathcal{C}_\tau(M_\rho^{-\delta}) \sqsubseteq \mathcal{C}(M_\rho)$. $\square$

PROOF OF THEOREM 2.9. (i) The first bound on ρ_D^* directly follows from part (i) of [25], Theorem A.6.2.

To show (2.9), we observe that parts (iii) and (iv) of [25], Theorem A.6.2, imply $2 = |\mathcal{C}_\tau(M^{-\delta}_{\rho_D^*+\varepsilon})| = |\mathcal{C}(M_{\rho_D^*+\varepsilon})|$. Since we further have $\rho_D^* + \varepsilon \leq \rho^* + \varepsilon^* + 6\varepsilon \leq \rho^{**}$ by the first bound on ρ_D^*, part (iii) of [25], Lemma A.4.3, thus shows

$$d(B_1, B_2) \geq \tau - 2\psi^*_{M_{\rho_D^*+\varepsilon}}(\delta) \geq \tau - 2c_{\text{thick}}\delta^\gamma > \tau - \psi(\delta),$$

where B_1 and B_2 are the two connected components of $M_{\rho_D^*+\varepsilon}$. On the other hand, the definition of $\tau^*_{M_{\rho_D^*+\varepsilon}}$ in [25], Proposition A.2.10, together with the definition of τ^* in (2.6) gives

$$3\tau^*\bigl(\rho_D^* - \rho^* + \varepsilon\bigr) = \tau^*_{M_{\rho_D^*+\varepsilon}} = d(B_1, B_2).$$

Combining both we find (2.9).

(ii) Part (iii) of [25], Theorem A.6.2, shows that Algorithm 1 returns two sets. Our next goal is to find a suitable ordering of these sets. To this end, we adopt

the notation of [25], Theorem A.6.2. Moreover, we denote the two topologically connected components of $M_{\rho^{**}}$ by A_1 and A_2. We further write

$$V^i_{\rho^*_D+\varepsilon} := \zeta_{\rho^{**},\rho^*_D+\varepsilon}\big(\zeta^{-1}_{\rho^{**}}(A_i)\big), \qquad i = 1, 2,$$

for the two τ-connected components of $M^{-\delta}_{\rho^*_D+\varepsilon}$. Note that part (iv) of [25], Theorem A.6.2, ensures that we can actually make this definition, and, in addition, it shows $V^1_{\rho^*_D+\varepsilon} \neq V^2_{\rho^*_D+\varepsilon}$. Moreover, by parts (ii) and (iii) of [25], Theorem A.6.2, we may assume that the sets returned by Algorithm 1 are ordered in the sense of $B_i(D) = \zeta(V^i_{\rho^*_D+\varepsilon})$, that is,

$$B_i(D) = \zeta \circ \zeta_{\rho^{**},\rho^*_D+\varepsilon}\big(\zeta^{-1}_{\rho^{**}}(A_i)\big), \qquad i = 1, 2. \tag{6.1}$$

To simplify notation in the following calculations, we write $B_i := B_i(D)$ for $i \in \{1, 2\}$ and $\rho := \rho^*_D$. Consequently, $A^1_{\rho+\varepsilon}$ and $A^2_{\rho+\varepsilon}$ are the two connected components of $M_{\rho+\varepsilon} = M_{\rho^*_D+\varepsilon}$, which by Definition 2.5 can be ordered in the sense of $A^i_{\rho+\varepsilon} \subset A^*_i$. Moreover, $V^1_{\rho+\varepsilon}$ and $V^2_{\rho+\varepsilon}$ become the two τ-connected components of $M^{-\delta}_{\rho+\varepsilon}$. For $i \in \{1, 2\}$, we further write $W^i_{\rho+\varepsilon} := (A^i_{\rho+\varepsilon})^{-\delta}$. Our first goal is to show that

$$W^i_{\rho+\varepsilon} \subset V^i_{\rho+\varepsilon}, \qquad i \in \{1, 2\}. \tag{6.2}$$

To this end, we fix an $x \in W^1_{\rho+\varepsilon}$. Since $W^1_{\rho+\varepsilon} \subset A^1_{\rho+\varepsilon}$ and $W^1_{\rho+\varepsilon} \subset M^{-\delta}_{\rho+\varepsilon}$, where the latter follows from $(A^1_{\rho+\varepsilon})^{-\delta} \subset M^{-\delta}_{\rho+\varepsilon}$, we then have $x \in A^1_{\rho+\varepsilon}$ and $x \in V^1_{\rho+\varepsilon} \cup V^2_{\rho+\varepsilon}$. Let us assume that $x \in V^2_{\rho+\varepsilon}$. Then we have $V^2_{\rho+\varepsilon} \cap A^1_{\rho+\varepsilon} \neq \varnothing$. Now, the diagram of [25], Theorem A.6.2, shows that $\zeta_{\rho+\varepsilon} : \mathcal{C}_\tau(M^{-\delta}_{\rho+\varepsilon}) \to \mathcal{C}(M_{\rho+\varepsilon})$ satisfies $\zeta_{\rho+\varepsilon_n}(V^2_{\rho+\varepsilon}) = A^2_{\rho+\varepsilon}$, and hence we have $V^2_{\rho+\varepsilon} \subset A^2_{\rho+\varepsilon}$. Consequently, $V^2_{\rho+\varepsilon} \cap A^1_{\rho+\varepsilon} \neq \varnothing$ implies $A^2_{\rho+\varepsilon} \cap A^1_{\rho+\varepsilon} \neq \varnothing$, which is a contradiction. Therefore, we have $x \in V^1_{\rho+\varepsilon}$; that is, we have shown (6.2) for $i = 1$. The case $i = 2$ can be shown analogously.

By (6.2) we find $W^i_{\rho+\varepsilon} \subset V^i_{\rho+\varepsilon} \subset B_i$, and thus $\mu(A^*_i \setminus B_i) \leq \mu(A^*_i \setminus W^i_{\rho+\varepsilon})$ for $i = 1, 2$. Conversely, using $\mu(B \setminus A) = \mu(B) - \mu(A \cap B)$ twice, we obtain

$$\begin{aligned}
\mu\big(B_1 \setminus (A^*_1 \cup A^*_2)\big) &= \mu(B_1) - \mu\big(B_1 \cap (A^*_1 \cup A^*_2)\big) \\
&\geq \mu(B_1) - \mu\big(B_1 \cap A^*_1\big) - \mu\big(B_1 \cap A^*_2\big) \\
&= \mu\big(B_1 \setminus A^*_1\big) - \mu\big(B_1 \cap A^*_2\big).
\end{aligned}$$

Since $B_1 \cap B_2 = \varnothing$ implies $B_1 \cap A^*_2 \subset A^*_2 \setminus B_2$, we thus find

$$\begin{aligned}
\mu\big(B_1 \triangle A^*_1\big) &= \mu\big(B_1 \setminus A^*_1\big) + \mu\big(A^*_1 \setminus B_1\big) \\
&\leq \mu\big(B_1 \setminus (A^*_1 \cup A^*_2)\big) + \mu\big(A^*_2 \setminus B_2\big) + \mu\big(A^*_1 \setminus B_1\big) \\
&\leq \mu\big(B_1 \setminus \{h > \rho^*\}\big) + \mu\big(A^*_1 \setminus W^1_{\rho+\varepsilon}\big) + \mu\big(A^*_2 \setminus W^2_{\rho+\varepsilon}\big),
\end{aligned}$$

where in the last estimate we also used [25], (A.1.3). Repeating this estimate for $\mu(B_2 \triangle A_2^*)$ and using $B_1 \cup B_2 \subset L_{D,\rho} \subset M_{\rho-\varepsilon}^{+\delta}$ yields the assertion. $\square$

PROOF OF THEOREM 3.1. Let us fix a $D \in X^n$ with $\|h_{D,\delta} - h_{P,\delta}\|_\infty < \varepsilon$. By the first estimate of [25], Theorem A.8.1, we see that the probability P^n of such a D is not smaller than $1 - e^{-\varsigma}$. In the case of a bounded density and (3.3), the same holds by the second estimate of [25], Theorem A.8.1, and

$$\begin{aligned}
&\sqrt{\frac{6c_{\text{part}}\|h\|_\infty\varsigma + \ln(2c_{\text{part}}) - \mathrm{d}\ln\delta}{3\delta^{\mathrm{d}}n} + \left(\frac{2c_{\text{part}}\varsigma}{3\delta^{\mathrm{d}}n}\right)^2} + \frac{c_{\text{part}}\varsigma}{3\delta^{\mathrm{d}}n} \\
&\qquad \le \sqrt{\frac{6c_{\text{part}}\|h\|_\infty\varsigma + \ln(2c_{\text{part}}) - \mathrm{d}\ln\delta}{3\delta^{\mathrm{d}}n}} + \frac{2c_{\text{part}}\varsigma}{3\delta^{\mathrm{d}}n} \\
&\qquad \le \sqrt{\frac{2c_{\text{part}}(1+\|h\|_\infty)(\varsigma + \ln(2c_{\text{part}}) - \mathrm{d}\ln\delta)}{\delta^{\mathrm{d}}n}} \\
&\qquad\quad + \frac{2c_{\text{part}}(\varsigma + \ln(2c_{\text{part}}) - \mathrm{d}\ln\delta)}{3\delta^{\mathrm{d}}n},
\end{aligned}$$

where we use $\ln(2c_{\text{part}}) \ge \mathrm{d}\ln\delta$. Now, [25], Lemma A.8.2, shows (2.7) for all $\rho \ge 0$. Let us check that the remaining assumptions of Theorem 2.9 are also satisfied if $\varepsilon^* \le (\rho^{**} - \rho^*)/9$. Clearly, we have $\delta \in (0, \delta_{\text{thick}}]$, $\varepsilon \in (0, \varepsilon^*]$ and $\psi(\delta) < \tau$. To show $\tau \le \tau^*(\varepsilon^*)$ we write

$$E := \{\varepsilon' \in (0, \rho^{**} - \rho^*] : \tau^*(\varepsilon') \ge \tau\}.$$

Since we assume $\varepsilon^* < \infty$, we obtain $E \ne \varnothing$ by the definition of ε^*. There thus exists an $\varepsilon' \in E$ with $\varepsilon' \le \inf E + \varepsilon \le \varepsilon^*$. Using the monotonicity of τ^* established in [25], Theorem A.4.2, we then conclude that $\tau \le \tau^*(\varepsilon') \le \tau^*(\varepsilon^*)$, and hence all assumptions of Theorem 2.9 are indeed satisfied. $\square$

PROOF OF THEOREM 4.3. Let us begin by checking the conditions of Theorem 3.1. Obviously, ε is chosen this way, and the definition of ε^* together with the assumption $\varepsilon^* \le (\rho^{**} - \rho^*)/9$ yields

$$(\tau/\underline{c}_{\text{sep}})^\kappa \le \varepsilon^* < \rho^{**} - \rho^*. \tag{6.3}$$

By the assumed separation exponent κ, we thus find in the case $\kappa < \infty$ that

$$\begin{aligned}
\inf\{\tilde\varepsilon \in (0, \rho^{**} - \rho^*] : \tau^*(\tilde\varepsilon) \ge \tau\} &\le \inf\{\tilde\varepsilon \in (0, \rho^{**} - \rho^*] : \underline{c}_{\text{sep}}\tilde\varepsilon^{1/\kappa} \ge \tau\} \\
&= (\tau/\underline{c}_{\text{sep}})^\kappa.
\end{aligned}$$

Consequently, (3.4) holds in the case $\kappa < \infty$. Moreover, in the case $\kappa = \infty$, (6.3) together with $\rho^{**} < \infty$ implies $\tau \le \underline{c}_{\text{sep}}$. In addition, the separation exponent $\kappa = \infty$ ensures $\tau^*(\tilde\varepsilon) \ge \underline{c}_{\text{sep}}$ for all $\tilde\varepsilon > 0$, and hence we obtain

$$\varepsilon + \inf\{\tilde\varepsilon \in (0, \rho^{**} - \rho^*] : \tau^*(\tilde\varepsilon) \ge \tau\} = \varepsilon \le \varepsilon^*;$$

that is, (3.4) is also established in the case $\kappa = \infty$. Now, applying Theorem 3.1, we see that $\rho_D^* \in [\rho^* + 2\varepsilon, \rho^* + \varepsilon^* + 5\varepsilon]$ with probability P^n not less than $1 - e^{-\varsigma}$; that is, (4.3) is proved. In addition, the definition of ε^* yields

$$\rho_D^* - \rho^* \le \varepsilon^* + 5\varepsilon \le (\tau/\underline{c}_{\text{sep}})^\kappa + 6\varepsilon,$$

and hence we obtain (4.4). Let us finally show (4.5). To this end, we first observe that Theorem 3.1 ensures

$$\begin{aligned}\tau/2 \le \tau - \psi(\delta) < 3\tau^*\big(\rho_D^* - \rho^* + \varepsilon\big) &\le 3\overline{c}_{\text{sep}}\big(\rho_D^* - \rho^* + \varepsilon\big)^{1/\kappa} \\ &< 3\overline{c}_{\text{sep}} 2^{1/\kappa}\big(\rho_D^* - \rho^*\big)^{1/\kappa},\end{aligned}$$

where in the last step, we use the already established (4.3). By some elementary transformations we conclude that

$$\frac{1}{2}\left(\frac{\tau}{6\overline{c}_{\text{sep}}}\right)^\kappa < \rho_D^* - \rho^*,$$

and combining this with $2\varepsilon \le \rho_D^* - \rho^*$, we obtain the assertion. □

PROOF OF COROLLARY 4.4. We first show (4.7) for $\kappa < \infty$ and sufficiently large n with the help of Theorem 4.3. To this end, we define $\varepsilon_n^* := \varepsilon_n + (\tau_n/\underline{c}_{\text{sep}})^\kappa$ for $n \ge 1$. Since (ε_n), (δ_n) and (τ_n) converge to 0, we then have $\delta_n \in (0, \delta_{\text{thick}}]$ and $\varepsilon_n^* \le (\rho^{**} - \rho^*)/9$ for all sufficiently large n. Furthermore, our definitions ensure $\tau_n/\delta_n^\gamma \to \infty$, and hence we have $\tau_n \ge 6c_{\text{thick}}\delta_n^\gamma = 2\psi(\delta_n)$ for all sufficiently large n, too. Before we can apply Theorem 4.3, it thus remains to show (3.3) for sufficiently large n. To this end, we observe that for $\varsigma_n := \ln n$ and $\xi_n := 2c_{\text{part}}(\varsigma_n + \ln(2c_{\text{part}}) - \mathrm{d}\ln\delta_n)$, we have

$$\varepsilon_n' := \sqrt{\frac{(1 + \|h\|_\infty)\xi_n}{\delta_n^{\mathrm{d}} n}} + \frac{\xi_n}{3\delta_n^{\mathrm{d}} n} \preceq \left(\frac{\ln n}{n}\right)^{\gamma\kappa/(2\gamma\kappa+d)}.$$

Using $\varepsilon_n \cdot (\frac{\ln n}{n})^{-\gamma\kappa/(2\gamma\kappa+d)} \to \infty$, we then see that $\varepsilon_n \ge \varepsilon_n'$ for all sufficiently large n. Now, applying Theorem 4.3, namely (4.4), we obtain an $n_0 \ge 1$ and a constant $\overline{K}$ such that (4.6) holds for all $n \ge n_0$. Moreover, if κ is exact, (4.5) yields a constant $\underline{K}$ such that (4.7) holds for all $n \ge n_0$.

In the case $\kappa = \infty$, we first observe that $\varepsilon_n^* := \varepsilon_n + (\tau_n/\underline{c}_{\text{sep}})^\kappa$ satisfies $\varepsilon_n^* = \varepsilon_n$ for all n with $\tau_n < \underline{c}_{\text{sep}}$, that is, for all sufficiently large n. Moreover, we have $\tau_n/\delta_n^\gamma \to \infty$, and, like the case $\kappa < \infty$, it thus suffices to show (3.3) for sufficiently large n. To this end, we observe that for $\varsigma_n := \ln n$ and ε_n' as above, we find that, for all sufficiently large n,

$$\varepsilon_n' \le c_2\left(\frac{\ln n \cdot \sqrt{\ln\ln n}}{n}\right)^{1/2} \le \varepsilon_n,$$

where c_2 is a suitable constant independent of n. Consequently, (4.3) and (4.4) yield (4.7) for all sufficiently large n. □

LEMMA 6.1. *Under the assumptions of Theorem 2.9 we have*

$$\sum_{i=1}^{2} \mu\bigl(B_i(D) \triangle A_i^*\bigr) \le 2 \sum_{i=1}^{2} \mu\bigl(A^i_{\rho_D^*+\varepsilon} \setminus \bigl(A^i_{\rho_D^*+\varepsilon}\bigr)^{-\delta}\bigr) + \mu\bigl(M^{+\delta}_{\rho_D^*-\varepsilon} \setminus M_{\rho_D^*-\varepsilon}\bigr) + \mu\bigl(\{\rho^* < h < \rho_D^* + \varepsilon\}\bigr).$$

PROOF OF LEMMA 6.1. We will use inequality (2.10) established in Theorem 2.9. To this end, we first observe that [25], (A.1.3), implies

$$\mu\bigl(M^{+\delta}_{\rho-\varepsilon} \setminus \{h > \rho^*\}\bigr) = \mu\Biggl(M^{+\delta}_{\rho-\varepsilon} \setminus \bigcup_{\rho'>\rho^*} M_{\rho'}\Biggr) \le \mu\bigl(M^{+\delta}_{\rho-\varepsilon} \setminus M_{\rho-\varepsilon}\bigr).$$

To bound the remaining terms on the right-hand side of (2.10), we further observe that the disjoint relation $A \cap B^{+\delta} = (A \cap (B^{+\delta} \setminus B)) \cup (A \cap B)$ applied to $B := X \setminus A^i_{\rho+\varepsilon}$ yields

$$\begin{aligned}
\mu\bigl(A_i^* \setminus \bigl(A^i_{\rho+\varepsilon}\bigr)^{-\delta}\bigr) &= \mu\bigl(A_i^* \cap \bigl(X \setminus A^i_{\rho+\varepsilon}\bigr)^{+\delta}\bigr) \\
&= \mu\bigl(A_i^* \cap \bigl(X \setminus A^i_{\rho+\varepsilon}\bigr)^{+\delta} \cap A^i_{\rho+\varepsilon}\bigr) + \mu\bigl(A_i^* \setminus A^i_{\rho+\varepsilon}\bigr) \\
&= \mu\bigl(A^i_{\rho+\varepsilon} \setminus \bigl(A^i_{\rho+\varepsilon}\bigr)^{-\delta}\bigr) + \mu\bigl(A_i^* \setminus A^i_{\rho+\varepsilon}\bigr).
\end{aligned}$$

Moreover, $A^i_{\rho+\varepsilon} \subset A_i^*$, $A_1^* \cap A_2^* = \varnothing$ together with [25], (A.1.2) and (A.1.3), imply

$$\begin{aligned}
\mu\bigl(A_1^* \setminus A^1_{\rho+\varepsilon}\bigr) + \mu\bigl(A_2^* \setminus A^2_{\rho+\varepsilon}\bigr) &= \mu\bigl(\bigl(A_1^* \cup A_2^*\bigr) \setminus \bigl(A^1_{\rho+\varepsilon} \cup A^2_{\rho+\varepsilon}\bigr)\bigr) \\
&= \mu\bigl(\{\rho^* < h < \rho + \varepsilon\}\bigr).
\end{aligned}$$

Combining all estimates with (2.10), we obtain the assertion. □

PROOF OF THEOREM 4.7. Since Assumption R includes the assumptions made in Theorem 4.3, we obtain (4.3) and (4.4). Furthermore, recall that the proofs of Theorems 4.3 and 3.1 show that the probability P^n of having a dataset $D \in X^n$ satisfying the assumptions of Theorem 2.9 is not less than $1 - e^{-\varsigma}$. For such D, Lemma 6.1 is applicable, and hence we obtain

$$\begin{aligned}
&\mu\bigl(B_1(D) \triangle A_1^*\bigr) + \mu\bigl(B_2(D) \triangle A_2^*\bigr) \\
&\quad\le \mu\bigl(M^{+\delta}_{\rho_D^*-\varepsilon} \setminus M_{\rho_D^*-\varepsilon}\bigr) + \mu\bigl(\{\rho^* < h < \rho_D^* + \varepsilon\}\bigr) \\
&\qquad + 2\mu\bigl(A^1_{\rho_D^*+\varepsilon} \setminus \bigl(A^1_{\rho_D^*+\varepsilon}\bigr)^{-\delta}\bigr) + 2\mu\bigl(A^2_{\rho_D^*+\varepsilon} \setminus \bigl(A^2_{\rho_D^*+\varepsilon}\bigr)^{-\delta}\bigr) \\
&\quad\le \mu\bigl(M^{+\delta}_{\rho_D^*-\varepsilon} \setminus M_{\rho_D^*-\varepsilon}\bigr) + \mu\bigl(\{0 < h - \rho^* < \rho_D^* - \rho^* + \varepsilon\}\bigr) + 4c_{\text{bound}}\delta^{\alpha},
\end{aligned}$$

where in the second estimate we use that the clusters have an α-smooth boundary by Assumption R. Moreover, the α-smooth boundaries also yield

$$\begin{aligned}\mu\bigl(M^{+\delta}_{\rho^*_D-\varepsilon} \setminus M_{\rho^*_D-\varepsilon}\bigr) &\le \mu\bigl(\bigl(A^1_{\rho^*_D-\varepsilon}\bigr)^{+\delta} \setminus M_{\rho^*_D-\varepsilon}\bigr) + \mu\bigl(\bigl(A^2_{\rho^*_D-\varepsilon}\bigr)^{+\delta} \setminus M_{\rho^*_D-\varepsilon}\bigr)\\ &\le \mu\bigl(\bigl(A^1_{\rho^*_D-\varepsilon}\bigr)^{+\delta} \setminus A^1_{\rho^*_D-\varepsilon}\bigr) + \mu\bigl(\bigl(A^2_{\rho^*_D-\varepsilon}\bigr)^{+\delta} \setminus A^2_{\rho^*_D-\varepsilon}\bigr)\\ &\le 2c_{\mathrm{bound}}\delta^\alpha.\end{aligned}$$

Finally, by (4.4) and the flatness exponent ϑ from Assumption R, we find

$$\mu\bigl(\{0 < h - \rho^* < \rho^*_D - \rho^* + \varepsilon\}\bigr) \le \bigl(c_{\mathrm{flat}}\bigl(\rho^*_D - \rho^* + \varepsilon\bigr)\bigr)^\vartheta \le \bigl((\tau/\underline{c}_{\mathrm{sep}})^\kappa + 7\varepsilon\bigr)^\vartheta.$$

Combining these three estimates, we then obtain the assertion. □

PROOF OF COROLLARY 4.8. To apply Theorem 4.7 we check that ε_n, δ_n and τ_n satisfy the assumptions of Theorem 4.3 for $\varsigma_n := \ln n$ and all sufficiently large n. To this end, we observe that for $\varsigma_n := \ln n$ and $\xi_n := 2c_{\mathrm{part}}(\varsigma_n + \ln(2c_{\mathrm{part}}) - \mathrm{d}\ln\delta_n)$, we have

$$\varepsilon'_n := \sqrt{\frac{(1+\|h\|_\infty)\xi_n}{\delta_n^{\mathrm{d}} n}} + \frac{\xi_n}{3\delta_n^{\mathrm{d}} n} \preceq \left(\frac{\ln n}{n}\right)^{\varrho/(2\varrho+\vartheta d)} (\ln\ln n)^{-\vartheta d/(4\varrho+2\vartheta d)}.$$

Using $\varepsilon_n \cdot (\frac{\ln n}{n})^{-\varrho/(2\varrho+\vartheta d)} (\ln\ln n)^{\vartheta d/(4\varrho+2\vartheta d)} \to \infty$, we then see that $\varepsilon_n \ge \varepsilon'_n$ for all sufficiently large n. Moreover, the remaining conditions on ε_n, δ_n and τ_n from Theorem 4.3 are clearly satisfied for all sufficiently large n, and hence we can apply Theorem 4.7 for such n. This yields

$$\mu\bigl(B_1(D) \vartriangle A^*_1\bigr) + \mu\bigl(B_2(D) \vartriangle A^*_2\bigr) \le 6c_{\mathrm{bound}}\delta_n^\alpha + \bigl(c_{\mathrm{flat}}(\tau_n/\underline{c}_{\mathrm{sep}})^\kappa + 7c_{\mathrm{flat}}\varepsilon_n\bigr)^\vartheta$$

with probability P^n not smaller than $1 - 1/n$ for all sufficiently large n. Some elementary calculations then show that there is a K with

$$\begin{aligned}P^n\Biggl(D : \mu\bigl(B_1(D) \vartriangle A^*_1\bigr) + \mu\bigl(B_2(D) \vartriangle A^*_2\bigr) \le K\left(\frac{\ln n \cdot (\ln\ln n)^2}{n}\right)^{\vartheta\varrho/(2\varrho+\vartheta d)}\Biggr)\\ \ge 1 - \frac{1}{n}\end{aligned}$$

for all sufficiently large n. Moreover, since we always have

$$\mu\bigl(B_1(D) \vartriangle A^*_1\bigr) + \mu\bigl(B_2(D) \vartriangle A^*_2\bigr) \le 2\mu(X) < \infty,$$

it is an easy exercise to suitably increase K such that the desired inequality actually holds for all $n \ge 1$. □

PROOF OF THEOREM 5.1. First observe that $C^2\ln(\ln n) \ge 2(1+\|h\|_\infty)$ guarantees that all $\varepsilon_{\delta,n}$ satisfy (3.3) for $\varsigma' := \varsigma + \ln|\Delta|$. Consequently, Theorem 4.3, namely (4.3) and (4.4), yields

$$P^n\bigl(\{D \in X^n : \varepsilon_{\delta,n} < \rho^*_{D,\delta} - \rho^* \le \bigl(\tau^\gamma_{\delta,n}/\underline{c}_{\mathrm{sep}}\bigr)^\kappa + 6\varepsilon_{\delta,n}\}\bigr) \ge 1 - e^{-\varsigma - \ln|\Delta|}$$

for all $\delta \in \Delta$. Applying the union bound, we thus find

$$P^n\bigl(D \in X^n : \varepsilon_{\delta,n} < \rho^*_{D,\delta} - \rho^* \leq \bigl(\tau^\gamma_{\delta,n}/\underline{c}_{\text{sep}}\bigr)^\kappa + 6\varepsilon_{\delta,n} \text{ for all } \delta \in \Delta\bigr) \geq 1 - e^{-\varsigma}.$$

Let us now consider a $D \in X^n$ such that $\varepsilon_{\delta,n} < \rho^*_{D,\delta} - \rho^* \leq (\tau^\gamma_{\delta,n}/\underline{c}_{\text{sep}})^\kappa + 6\varepsilon_{\delta,n}$ for all $\delta \in \Delta$. Then the definitions of $\rho^*_{D,\Delta}$ and $\varepsilon_{D,\Delta}$ [see (5.3)] imply

$$\rho^*_{D,\Delta} - \rho^* = \min_{\delta\in\Delta} \rho^*_{D,\delta} - \rho^* \in \Bigl(\min_{\delta\in\Delta} \varepsilon_{\delta,n}, \min_{\delta\in\Delta}\bigl((\tau^\gamma_{\delta,n}/\underline{c}_{\text{sep}})^\kappa + 6\varepsilon_{\delta,n}\bigr)\Bigr]$$

and $\varepsilon_{D,\Delta} = \varepsilon_{\delta^*_{D,\Delta},n} < \rho^*_{D,\delta^*_{D,\Delta}} - \rho^* = \rho^*_{D,\Delta} - \rho^*$; that is, we have shown the first assertion. To show the remaining assertions, we first observe that a literal repetition of the argument above, in which we only replace the use of (4.3) by that of (4.5), yields

$$P^n\bigl(D \in X^n : c_1\tau^\kappa_{\delta,n} + \varepsilon_{\delta,n} < \rho^*_{D,\delta} - \rho^* \leq c_2\tau^\kappa_{\delta,n} + 6\varepsilon_{\delta,n} \text{ for all } \delta \in \Delta\bigr) \geq 1 - e^{-\varsigma}.$$

Using (5.3) we then immediately obtain the second assertion, while considering $\delta = \delta^*_{D,\Delta}$ gives the third assertion. □

PROOF OF COROLLARY 5.2. Let us fix an $n \geq 16$. For later use we note that this choice implies $I_n \subset (0, 1]$. Our first goal is to show that we can apply Theorem 5.1 for sufficiently large n. To this end, we first observe that $\max \Delta_n = (\ln\ln n)^{-\mathrm{d}} \to 0$ for $n \to \infty$, and hence we obtain $\Delta_n \subset (0, \delta_{\text{thick}}]$ for all sufficiently large n. Analogously, $\max \Delta_n \ln\ln\ln n \to 0$ implies $\max_{\delta\in\Delta_n}(\tau_{\delta,n}/\underline{c}_{\text{sep}})^\kappa \leq (\rho^{**} - \rho^*)/18$ for all sufficiently large n, and the definition of $\tau_{\delta,n}$ ensures $\min_{\delta\in\Delta_n} \tau_{\delta,n} \geq 2\psi(\delta)$ for all sufficiently large n. Let us now show that eventually we also have $\max_{\delta\in\Delta_n} \varepsilon_{\delta,n} \leq (\rho^{**} - \rho^*)/18$. To this end, note that the derivative of $g_n : (0, \infty) \to \mathbb{R}$ defined by

$$g_n(\delta) := \frac{\ln(2c_{\text{part}}|\Delta_n|n) - \mathrm{d}\ln\delta}{\delta^d n}$$

is given by

$$g'_n(\delta) = -\frac{\mathrm{d}(1 + \ln(2c_{\text{part}}|\Delta_n|n) - \mathrm{d}\ln\delta)}{\delta^{1+\mathrm{d}} n},$$

and using $c_{\text{part}} \geq 1$, we thus find that g_n is monotonically decreasing on $(0, 1]$ for all $n \geq 1$. In addition, using $|\Delta_n| \leq n$ we obtain

$$\begin{aligned} g_n(\min I_n) &= g_n\left(\left(\frac{\ln n \cdot (\ln\ln n)^2}{n}\right)^{1/d}\right) \\ &= \frac{\ln(2c_{\text{part}}|\Delta_n|n) + \ln n - \ln\ln n - 2\ln\ln\ln n}{\ln n \cdot (\ln\ln n)^2} \\ &\leq \frac{4\ln n - \ln\ln n - 2\ln\ln\ln n}{\ln n \cdot (\ln\ln n)^2} \\ &\leq \frac{4}{(\ln\ln n)^2} \end{aligned}$$

for all $n \geq \max\{16, 2c_{\text{part}}\}$, and hence $g_n(\min I_n)\ln\ln n \to 0$ for $n \to \infty$. Since the definition of $\varepsilon_{\delta,n}$ gives $\varepsilon_{\delta,n} = C\sqrt{c_{\text{part}} g_n(\delta)\ln\ln n} + \frac{2}{3}c_{\text{part}} g_n(\delta)$, we can thus conclude that

$$\begin{aligned}\max_{\delta\in\Delta_n} \varepsilon_{\delta,n} &\leq \max_{\delta\in\Delta_n} C\sqrt{c_{\text{part}} g_n(\delta)\ln\ln n} + \max_{\delta\in\Delta_n} c_{\text{part}} g_n(\delta)\\ &\leq C\sqrt{c_{\text{part}} g_n(\min I_n)\ln\ln n} + c_{\text{part}} g_n(\min I_n) \to 0\end{aligned}$$

for $n \to \infty$. This ensures the desired $\max_{\delta\in\Delta_n} \varepsilon_{\delta,n} \leq (\rho^{**} - \rho^*)/18$ for all sufficiently large n. Combining this with our previous estimate, we find

$$\max_{\delta\in\Delta_n} \bigl((\tau_{\delta,n}/\underline{c}_{\text{sep}})^\kappa + \varepsilon_{\delta,n}\bigr) \leq (\rho^{**} - \rho^*)/9$$

for all sufficiently large n, and thus we can apply Theorem 5.1 for such n.

Before we proceed, let us now fix an $n \geq 16$ and assume that without loss of generality that Δ_n is of the form $\Delta = \{\delta_1, \ldots, \delta_m\}$ with $\delta_{i-1} < \delta_i$ for all $i = 2, \ldots, m$. We write $\delta_0 := \min I_n$ and $\delta_{m+1} := \max I_n$. Our intermediate goal is to show that

$$\delta_i - \delta_{i-1} \leq 2n^{-1/\mathrm{d}}, \qquad i = 1, \ldots, m+1. \tag{6.4}$$

To this end, we fix an $i \in \{1, \ldots, m\}$ and write $\bar{\delta} := (\delta_i + \delta_{i-1})/2 \in I_n$. Since Δ_n is an $n^{-1/\mathrm{d}}$-net of I_n, we then have $\delta_i - \bar{\delta} \leq n^{-1/\mathrm{d}}$ or $\bar{\delta} - \delta_{i-1} \leq n^{-1/\mathrm{d}}$, and from both, (6.4) follows. Moreover, to show (6.4) in the case $i = m+1$, we first observe that there exists an $\delta_i \in \Delta_n$ with $\delta_i - \delta_m \leq n^{-1/\mathrm{d}}$ since Δ_n is an $n^{-1/\mathrm{d}}$-net of I_n. Using our ordering of Δ_n, we can assume without loss of generality that $i = m$, which immediately implies (6.4).

We now prove the first assertion in the case $\kappa < \infty$. To this end, we write

$$\delta_n^* := \left(\frac{\ln n \cdot \ln\ln n}{n}\right)^{1/(2\gamma\kappa+\mathrm{d})},$$

where we note that for sufficiently large n we have $\delta_n^* \in I_n$. In the following we thus restrict our considerations to such n. Then there exists an index $i \in \{1, \ldots, m+1\}$ such that $\delta_{i-1} \leq \delta_n^* \leq \delta_i$, and by (6.4) we conclude that $\delta_n^* \leq \delta_i \leq \delta_n^* + 2n^{-1/\mathrm{d}}$. Clearly, this yields

$$\begin{aligned}\min_{\delta\in\Delta_n} \bigl(c_2\tau_{\delta,n}^\kappa + 6\varepsilon_{\delta,n}\bigr) &= \min_{\delta\in\Delta_n} \bigl(c_2\delta^{\gamma\kappa}(\ln\ln\ln n)^\kappa + 6\varepsilon_{\delta,n}\bigr)\\ &\leq c_2\delta_i^{\gamma\kappa}(\ln\ln\ln n)^\kappa + 6\varepsilon_{\delta_i,n}\\ &\leq c_2\bigl(\delta_n^* + 2n^{-1/\mathrm{d}}\bigr)^{\gamma\kappa}(\ln\ln\ln n)^\kappa + 6\varepsilon_{\delta_i,n}\\ &\leq 6c_2\left(\frac{\ln n \cdot (\ln\ln n)^2}{n}\right)^{1/(2\gamma\kappa+\mathrm{d})} + 6\varepsilon_{\delta_i,n}\end{aligned} \tag{6.5}$$

for all sufficiently large n, where $c_2 := \underline{c}_{\text{sep}}^{-\kappa}$ is the constant from Theorem 5.1. Moreover, using $|\Delta_n| \leq n$ and the monotonicity of g_n, we further obtain

$$g_n(\delta_i) \leq g_n(\delta_n^*) = \frac{\ln(2c_{\text{part}}|\Delta_n|n) - \mathrm{d}\ln\delta_n^*}{(\delta_n^*)^{\mathrm{d}}n} \leq \frac{\ln(2c_{\text{part}}) + 2\ln n - \mathrm{d}\ln\delta_n^*}{(\delta_n^*)^{\mathrm{d}}n}$$

$$\leq \frac{4\ln n}{(\delta_n^*)^{\mathrm{d}}n} \tag{6.6}$$

$$\leq \frac{4}{(\ln\ln n)^{\mathrm{d}/(2\gamma\kappa+\mathrm{d})}} \cdot \left(\frac{\ln n}{n}\right)^{2\gamma\kappa/(2\gamma\kappa+\mathrm{d})}$$

for all sufficiently large n. By the relation between $\varepsilon_{\delta,n}$ and $g_n(\delta)$, we then find

$$\varepsilon_{\delta_i,n} \leq 2C\sqrt{c_{\text{part}}}\left(\frac{\ln n \cdot \ln\ln n}{n}\right)^{\gamma\kappa/(2\gamma\kappa+\mathrm{d})} + 3c_{\text{part}}\left(\frac{\ln n}{n}\right)^{2\gamma\kappa/(2\gamma\kappa+\mathrm{d})},$$

and combining this estimate with (6.5) and Theorem 5.1, we obtain the first assertion in the case $\kappa < \infty$.

Let us now consider the case $\kappa = \infty$. To this end, we fix an n such that

$$\delta_n^* := \left(\frac{1}{\ln\ln n}\right)^{1/\mathrm{d}}$$

satisfies $(\delta_n^* + 2n^{-1/\mathrm{d}})^\gamma \ln\ln\ln n < \underline{c}_{\text{sep}}$, and thus

$$\left((\delta_n^* + 2n^{-1/\mathrm{d}})^\gamma \ln\ln\ln n/\underline{c}_{\text{sep}}\right)^\kappa = 0.$$

Since $\delta_n^* \in I_n$, there also exists an index $i \in \{1,\ldots,m+1\}$ such that $\delta_{i-1} \leq \delta^* \leq \delta_i$, and by (6.4) we again conclude $\delta^* \leq \delta_i \leq \delta^* + 2n^{-1/\mathrm{d}}$. Clearly, the latter implies

$$\begin{aligned}\min_{\delta\in\Delta_n}\left((\tau_{\delta,n}/\underline{c}_{\text{sep}})^\kappa + 6\varepsilon_{\delta,n}\right) &\leq \left(\delta_i^\gamma \ln\ln\ln n/\underline{c}_{\text{sep}}\right)^\kappa + 6\varepsilon_{\delta_i,n}\\ &\leq \left((\delta_n^* + 2n^{-1/\mathrm{d}})^\gamma \ln\ln\ln n/\underline{c}_{\text{sep}}\right)^\kappa + 6\varepsilon_{\delta_i,n}\\ &= 6\varepsilon_{\delta_i,n}\end{aligned}$$

by our assumptions on n. Analogously to (6.6) we further find, for sufficiently large n, that

$$g_n(\delta_i) \leq g_n(\delta_n^*) \leq \frac{3\ln n - \mathrm{d}\ln\delta_n^*}{(\delta_n^*)^{\mathrm{d}}n} = \frac{3\ln n + \ln\ln\ln n}{n(\ln\ln n)^{-1}} \leq 4\frac{\ln n \cdot \ln\ln n}{n},$$

and by the relation between $\varepsilon_{\delta,n}$ and $g(\delta)$, we then find the assertion with the help of Theorem 5.1.

Let us finally prove the second assertion. To this end we first recall that we have already seen that for sufficiently large n, we can apply Theorem 5.1. Thus it suffices to find a lower bound for the right-hand side of

$$\min_{\delta\in\Delta_n}\left(c_1\tau_{\delta,n}^\kappa + \varepsilon_{\delta,n}\right) \geq \min\{1, c_1\} \cdot \min_{\delta\in\Delta_n}\left(\tau_{\delta,n}^\kappa + \varepsilon_{\delta,n}\right), \tag{6.7}$$

where c_1 is the constant appearing in Theorem 5.1. Now, for $n \geq 16$, we have $I_n \subset (0, 1]$, and thus we find $\delta \in (0, 1]$ for all $\delta \in \Delta_n$. For sufficiently large n this yields

$$\begin{aligned}
&\min_{\delta\in\Delta_n}\bigl(\tau_{\delta,n}^{\kappa} + \varepsilon_{\delta,n}\bigr)\\
&\quad= \min_{\delta\in\Delta_n}\left(\delta^{\gamma\kappa}(\ln\ln\ln n)^{\kappa} + C\sqrt{c_{\mathrm{part}} g_n(\delta)\ln\ln n} + \frac{2}{3}c_{\mathrm{part}} g_n(\delta)\right)\\
&\quad\geq \min_{\delta\in\Delta_n}\bigl(\delta^{\gamma\kappa} + C\sqrt{c_{\mathrm{part}} g_n(\delta)\ln\ln n}\bigr)\\
&\quad\geq \min_{\delta\in\Delta_n}\left(\delta^{\gamma\kappa} + C\sqrt{\frac{c_{\mathrm{part}}\ln n\cdot\ln\ln n}{\delta^d n}}\right)\\
&\quad\geq \min_{\delta\in(0,1]}\left(\delta^{\gamma\kappa} + C\sqrt{\frac{c_{\mathrm{part}}\ln n\cdot\ln\ln n}{\delta^d n}}\right).
\end{aligned}$$

An elementary application of calculus then yields the assertion. □

PROOF OF COROLLARY 5.3. As in the proof of Corollary 4.8 it suffices to show the assertion for sufficiently large n. Now, we have seen in the proof of Corollary 5.2 that for sufficiently large n, Inequality (5.4) follows from the fact that the procedure satisfies the assumptions of Theorem 5.1 for such n and $\varsigma := \ln n$. Consequently, for sufficiently large n, the probability P^n of having a data set $D \in X^n$ satisfying both (5.4) and the third inequality of Theorem 5.1 is not less than $1 - 1/n$. Let us fix such a D. Then we have

$$c_1\tau_{D,\Delta}^{\kappa} + \varepsilon_{D,\Delta} \leq \rho_{D,\Delta_n}^{*} - \rho^{*} \leq \overline{K}\left(\frac{\ln n\cdot(\ln\ln n)^2}{n}\right)^{\gamma\kappa/(2\gamma\kappa+\mathrm{d})}. \tag{6.8}$$

Moreover, an elementary estimate yields

$$c_1\tau_{D,\Delta}^{\kappa} + \varepsilon_{D,\Delta} \geq \min\{1/7, c_1\underline{c}_{\mathrm{sep}}^{\kappa}\}\cdot\bigl((\tau_{D,\Delta}/\underline{c}_{\mathrm{sep}})^{\kappa} + 7\varepsilon_{D,\Delta}\bigr),$$

and setting $c := \min\{1/7, c_1\underline{c}_{\mathrm{sep}}^{\kappa}\}$, we hence obtain

$$(\tau_{D,\Delta}/\underline{c}_{\mathrm{sep}})^{\kappa} + 7\varepsilon_{D,\Delta} \leq c^{-1}\overline{K}\left(\frac{\ln n\cdot(\ln\ln n)^2}{n}\right)^{\gamma\kappa/(2\gamma\kappa+\mathrm{d})}. \tag{6.9}$$

In addition, for sufficiently large n, inequality (6.8) implies

$$\delta_{D,\Delta}^{*} \leq \tau_{D,\Delta}^{1/\gamma} \leq (4\overline{K})^{1/\gamma\kappa}(6\overline{c}_{\mathrm{sep}})^{1/\gamma}\left(\frac{\ln n\cdot(\ln\ln n)^2}{n}\right)^{1/(2\gamma\kappa+\mathrm{d})}. \tag{6.10}$$

Now we have already seen in the proofs of Theorem 5.1 and Corollary 5.2 that for sufficiently large n, the assumptions on δ, $\varepsilon_{\delta,n}$, $\varepsilon_{\delta,n}^{*}$, τ_n, $\varsigma_n := \ln n$ and n of Theorem 4.3 are satisfied for all $\delta \in \Delta_n$ simultaneously. We can thus combine (6.9) and (6.10) with Theorem 4.7 to obtain the assertion. □

SUPPLEMENTARY MATERIAL

Supplement to "Fully adaptive density-based clustering" (DOI: 10.1214/15-AOS1331SUPP; .pdf). We provide two appendices A and B. In Appendix A, several auxiliary results, which are partially taken from [24], are presented, and the assumptions made in the paper are discussed in more detail. In Appendix B, we present a couple of two-dimensional examples that show that the assumptions imposed in the paper are not only met by many discontinuous densities, but also by many continuous densities.

REFERENCES

[1] BAÍLLO, A., CUESTA-ALBERTOS, J. A. and CUEVAS, A. (2001). Convergence rates in nonparametric estimation of level sets. *Statist. Probab. Lett.* **53** 27–35. MR1843338

[2] BAÍLLO, A., CUEVAS, A. and JUSTEL, A. (2000). Set estimation and nonparametric detection. *Canad. J. Statist.* **28** 765–782. MR1821433

[3] BEN-DAVID, S. and LINDENBAUM, M. (1997). Learning distributions by their density levels: A paradigm for learning without a teacher. *J. Comput. System Sci.* **55** 171–182. MR1473058

[4] CHAÓN, J. C. (2014). A population background for nonparametric density-based clustering. Technical report. Available at arXiv:1408.1381.

[5] CHAUDHURI, K. and DASGUPTA, S. (2010). Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems* **23** (J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel and A. Culotta, eds.) 343–351.

[6] CUEVAS, A. and FRAIMAN, R. (1997). A plug-in approach to support estimation. *Ann. Statist.* **25** 2300–2312. MR1604449

[7] DEVROYE, L. and WISE, G. L. (1980). Detection of abnormal behavior via nonparametric estimation of the support. *SIAM J. Appl. Math.* **38** 480–488. MR0579432

[8] DONOHO, D. L. (1988). One-sided inference about functionals of a density. *Ann. Statist.* **16** 1390–1420. MR0964930

[9] GINÉ, E. and GUILLOU, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. Henri Poincaré Probab. Stat.* **38** 907–921. MR1955344

[10] HARTIGAN, J. A. (1975). *Clustering Algorithms*. Wiley, New York. MR0405726

[11] HARTIGAN, J. A. (1981). Consistency of single linkage for high-density clusters. *J. Amer. Statist. Assoc.* **76** 388–394. MR0624340

[12] HARTIGAN, J. A. (1987). Estimation of a convex density contour in two dimensions. *J. Amer. Statist. Assoc.* **82** 267–270. MR0883354

[13] KPOTUFE, S. and VON LUXBURG, U. (2011). Pruning nearest neighbor cluster trees. In *Proceedings of the* 28*th International Conference on Machine Learning* (L. Getoor and T. Scheffer, eds.) 225–232. ACM, New York.

[14] MAIER, M., HEIN, M. and VON LUXBURG, U. (2009). Optimal construction of k-nearest-neighbor graphs for identifying noisy clusters. *Theoret. Comput. Sci.* **410** 1749–1764. MR2514706

[15] MÜLLER, D. W. and SAWITZKI, G. (1991). Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.* **86** 738–746. MR1147099

[16] POLONIK, W. (1995). Measuring mass concentrations and estimating density contour clusters—An excess mass approach. *Ann. Statist.* **23** 855–881. MR1345204

[17] RIGOLLET, P. (2007). Generalized error bounds in semi-supervised classification under the cluster assumption. *J. Mach. Learn. Res.* **8** 1369–1392. MR2332435

[18] RIGOLLET, P. and VERT, R. (2009). Optimal rates for plug-in estimators of density level sets. *Bernoulli* **15** 1154–1178. MR2597587

[19] RINALDO, A., SINGH, A., NUGENT, R. and WASSERMAN, L. (2012). Stability of density-based clustering. *J. Mach. Learn. Res.* **13** 905–948. MR2930628

[20] RINALDO, A. and WASSERMAN, L. (2010). Generalized density clustering. *Ann. Statist.* **38** 2678–2722. MR2722453

[21] SCOVEL, C., HUSH, D. and STEINWART, I. (2005). Learning rates for density level detection. *Anal. Appl. (Singap.)* **3** 357–371. MR2181253

[22] SINGH, A., SCOTT, C. and NOWAK, R. (2009). Adaptive Hausdorff estimation of density level sets. *Ann. Statist.* **37** 2760–2782. MR2541446

[23] SRIPERUMBUDUR, B. K. and STEINWART, I. (2012). Consistency and rates for clustering with DBSCAN. In *Proceedings of the* 15*th International Conference on Artificial Intelligence and Statistics* 2012 (N. Lawrence and M. Girolami, eds.). *JMLR Workshop and Conference Proceedings* **22** 1090–1098.

[24] STEINWART, I. (2011). Adaptive density level set clustering. In *Proceedings of the* 24*th Conference on Learning Theory* 2011 (S. Kakade and U. von Luxburg, eds.). *JMLR Workshop and Conference Proceedings* **19** 703–738. JMRL.

[25] STEINWART, I. (2015). Supplement to "Fully adaptive density-based clustering." DOI:10.1214/15-AOS1331SUPP.

[26] STEINWART, I., HUSH, D. and SCOVEL, C. (2005). A classification framework for anomaly detection. *J. Mach. Learn. Res.* **6** 211–232. MR2249820

[27] STUETZLE, W. (2003). Estimating the cluster type of a density by analyzing the minimal spanning tree of a sample. *J. Classification* **20** 25–47. MR1983120

[28] STUETZLE, W. and NUGENT, R. (2010). A generalized single linkage method for estimating the cluster tree of a density. *J. Comput. Graph. Statist.* **19** 397–418. MR2675094

[29] TSYBAKOV, A. B. (1997). On nonparametric estimation of density level sets. *Ann. Statist.* **25** 948–969. MR1447735

[30] WALTHER, G. (1997). Granulometric smoothing. *Ann. Statist.* **25** 2273–2299. MR1604445

INSTITUTE FOR STOCHASTICS AND APPLICATIONS
FACULTY 8: MATHEMATICS AND PHYSICS
UNIVERSITY OF STUTTGART
D-70569 STUTTGART
GERMANY
E-MAIL: ingo.steinwart@mathematik.uni-stuttgart.de