Low-Noise Density Clustering

Alessandro Rinaldo
Department of Statistics
Carnegie Mellon University

Larry Wasserman Department of Statistics
Carnegie Mellon University

Abstract

We study density-based clustering under low-noise conditions. Our framework allows for sharply defined clusters such as clusters on lower dimensional manifolds. We show that accurate clustering is possible even in high dimensions. We propose two data-based methods for choosing the bandwidth and we study the stability properties of density clusters. We show that a simple graph-based algorithm known as the "friends-of-friends" algorithm successfully approximates the high density clusters.

1 Introduction

It has been observed that classification methods can be very accurate in high dimensional problems, apparently contradicting the curse of dimensionality. A plausible explanation for this phenomenon is the "low-noise" condition due to Mammen and Tsybakov (1999). When the low noise condition holds, the probability mass near the decision boundary is low and fast rates of convergence of the classification error are possible in high dimensions.

Similarly, clustering methods can be very accurate in high dimensional problems. For example, clustering subjects based on gene profiles and clustering curves are both high dimensional problems where several methods have worked well despite the high dimensionality. This suggests that there should be a low noise condition that explains the success of clustering in high dimensional problems.

In this paper we focus on clusters that are defined as the connected components of high density regions (Cuevas and Fraiman, 1997; Hartigan, 1975). The advantage of density clustering over other methods is that there is a well-defined population quantity being estimated and density clustering allows the shape of the clusters to be very general. (A related but somewhat different approach for generally shaped clusters is spectral clustering; see (von Luxburg, 2007) and (Ng et al., 2002).) Of course, without some conditions, density estimation is subject to the usual curse of dimensionality. One would hope that an appropriate low noise condition would obviate the curse of dimensionality. Such assumptions have been proposed by Polonik (1995), Rigollet (2007), Rigollet and Vert (2006), and others. However, the assumptions used by these authors rule out the case where the clusters are very sharply defined, which should be the easiest cases, and, more generally, clusters defined on lower dimensional sets.

The purpose of this paper is to define a notion of low noise clusters that does not rule out the most favorable cases and is not limited to sets of full dimension. We study the risk properties of density-based clustering and its stability properties, and we provide data-based methods for choosing the smoothing parameters.

The following simple example helps to illustrate our motivation. We refer the reader to the next section for a more rigorous introduction. Suppose that a distribution P is a mixture of finitely many point masses at distinct points $x_1, \ldots, x_k$ where $x_j \in \mathbb{R}^d$. Specifically, suppose that $P = k^{-1} \sum_{j=1}^{k} \delta_j$ where δ_j is a point mass at x_j. The clusters are $C_1 = \{x_1\}, \ldots, C_k = \{x_k\}$. This is a trivial clustering problem even if the dimension d is very high. The clusters could not be more sharply defined yet the density does not even exist in the usual sense. This makes it clear that common assumptions about the density such as smoothness or even boundedness are not well-suited for density clustering.

Now let $p_h = dP_h/d\mu$ be the Lebesgue density of the measure P_h obtained by convolving P with the probability measure having Lebesgue density K_h, a kernel with bandwidth h. Unlike the original distribution

P, P_h has full-dimensional support for each positive h. The "mollified" density p_h contains all the information needed for clustering. Indeed, there exist constants $\overline{h} > 0$ and $\lambda \geq 0$ such that the following facts are true:

1. for all $0 < h < \overline{h}$, the level set $\{x : \ p_h(x) > \lambda\}$ has disjoint, connected components $C_1^h, \ldots, C_k^h$;
2. the components C_j^h contain the true clusters: $C_j \subset C_j^h$ for $j = 1, \ldots, k$;
3. although C_j^h overestimates the true cluster C_j, this overestimation is inconsequential since $P(C_j^h - C_j) = 0$ and hence a new observation will not be misclustered;
4. let $\widehat{p}_h$ denote the kernel density estimator using K_h with fixed bandwidth $0 < h < \overline{h}$ and based on a i.i.d. sample of size n from P. Then, $\sup_x |p_h(x) - \widehat{p}_h(x)| = O(\sqrt{\log n/n})$ almost everywhere P, which does not depend on the dimension d (see Section 3.1). The bias from using a fixed bandwidth h— which does not vanish as $n \to \infty$— does not adversely affect the clustering.

In summary, we can recover the true clusters using an estimator of the density p_h with a large bandwidth h. It is not necessary to assume that the true density is smooth or that it even exists.

Our contributions in this paper are the following:

1. We develop a notion of low noise clustering that applies to probability distribution that have non-smooth Lebesgue densities or do not even admit a density.
2. We find the rates of convergence for estimators of these clusters.
3. We study two data-driven methods for choosing the bandwidth.
4. We study the stability properties of density clusters.
5. We show that the depth-first search algorithm on the ρ-nearest neighborhood graph of $\{\widehat{p}_h > \lambda\}$ is effective at recovering the high-density clusters.

Section 2 contains notation and definitions. Section 3 contains results on rates of convergence. We give a data-driven method for choosing the bandwidth in Section 4. Section 4.2 contains results on cluster stability. The validity of the "friends-of-friends" algorithm for approximating the clusters is proved in Section 5. Section 6 contains some examples based on simulated data. Concluding remarks are in Section 7. All proofs are in the Section 8. Some technical details are in the appendix, Section 9.

Notation. For two sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$ and $a_n = \Omega(b_n)$ if there exists a constant $C > 0$ such that, for all n large enough, $|a_n|/b_n \leq C$ and $|a_n|/b_n \geq C$, respectively. If $a_n = \Omega(b_n)$ and $a_n = O(b_n)$, then we will write $a_n \asymp b_n$. We denote with $\mathbb{P}(E)$ the probability of a generic event E, whenever the underlying probability measure is implicitly understood from the context. Similarly, for a random quantity X, $\mathbb{P}(E|X)$ indicates the condition probability of the event E given X.

2 Background

We begin with some definitions and the low-noise assumptions.

2.1 Level Set Clusters

In this section we develop a probabilistic framework for the definition of clusters we have adopted. For ease of readability, the more technical measure-theoretic details are in Section 9.1.

Let P be a probability distribution on $\mathbb{R}^d$ whose support S is comprised of an unknown number m of disjoint compact sets $\{S_1, \ldots, S_m\}$ of different integral dimensions. These sets may consist, for example, of smooth submanifolds or even single points. We define the *geometric density* of P as the measurable function $p \colon \mathbb{R}^d \mapsto \mathbb{R}$ given by

$$p(x) = \lim_{h \downarrow 0} \frac{P(B(x,h))}{v_d h^d}, \tag{1}$$

where $B(x,\epsilon)$ is the Euclidean ball of radius h centered at x, μ is the d-dimensional Lebesgue measure and $v_d \equiv \mu(B(0,1))$. Note that, almost everywhere P, $p(x) = \infty$ if and only if x belongs to some set S_i having dimension strictly less than d and is positive and finite if and only if x belongs to some d-dimensional set S_i. In general, $\int_{\mathbb{R}^d} p(x)d\mu(x) \leq 1$ and, therefore, p is not necessarily a probability density. Nonetheless, p can be used to recover the support of P, since

$$S = \overline{\{x\colon p(x) > 0\}},$$

where for a set $A \subset \mathbb{R}^d$, $\overline{A}$ denotes its closure.

For $\lambda \geq 0$, define the λ-level set

$$L \equiv L(\lambda) = \overline{\Big\{x :\ p(x) > \lambda\Big\}}. \tag{2}$$

Throughout the paper, we will suppose that we are given a fixed value of $\lambda < \|p\|_\infty$, where $\|p\|_\infty \equiv \sup_{x\in\mathbb{R}^d} p(x)$. Often, λ is chosen so that $P(L(\lambda)) \approx 1-\alpha$ for some given α. In practice, it is advisable to present the results for a variety of values of λ as we discuss in Section 7.

Remark. Our definition of density clusters does not necessarily lead to a partition of the support of P, since $P(L^c)$ is not in general zero. The fact that there may a be a positive probability of observing a point outside L can be interpreted as a form of noise.

We assume that there are $k \geq 1$ disjoint, compact, connected sets $C_1, \ldots, C_k$ such that

$$L = C_1 \cup \cdots \cup C_k.$$

The value of k is not assumed to be known. The sets $C_1, \ldots, C_k$ are called the λ-*clusters* of p, or just *clusters*. In our setting, the C_j's need not be full dimensional. Indeed, C_j might be a lower-dimensional manifold or even a single point. Furthermore, if S_i has dimension smaller than d, then $C_j = S_i$, for some $j = 1, \ldots, k$. Thus, for any $\lambda \geq 0$, the λ-clusters of p will include all the lower-dimensional components of S. On the other hand, if S_i is full-dimensional, then there may be multiple clusters in it, depending on the value of λ.

We observe an i.i.d. sample $X = (X_1, \ldots, X_n)$ from P, from which we construct the kernel density estimator

$$\widehat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{c_d h^d} K\left(\frac{x - X_i}{h}\right), \quad \forall x \in \mathbb{R}^d. \tag{3}$$

We assume that the kernel $K\colon \mathbb{R}^d \mapsto \mathbb{R}$ is a smooth, bounded and nonnegative function with compact support. Further conditions on the kernel K are discussed in Section 3.1. We point out that, while the compactness assumption for the support of K simplifies our analysis, it is not essential and could be relaxed at the cost of additional technicalities.

Let $p_h\colon \mathbb{R}^d \mapsto \mathbb{R}$ be the measurable function given by

$$p_h(x) = \int_S K_h(x-y)dP(y) = \mathbb{E}(\widehat{p}_h(x)), \tag{4}$$

where $K_h(x) \equiv \frac{1}{c_d h^d} K\left(\frac{||x||}{h}\right)$, with $c_d \equiv \int K(x)d\mu(x)$. Also, let $K_h\mu$ be the probability measure given by $K_h\mu(A) = \int_A K_h(x)d\mu(x)$, for any Borel set $A \subseteq \mathbb{R}^d$. Then, p_h is the Lebesgue density of the probability measure P_h obtained by convolving P with $K_h\mu$. More precisely, for each measurable set A,

$$P_h(A) = \int_A \int_S K_h(x-y)dP(y)d\mu(x) = \int_A p_h(x)d\mu(x).$$

Borrowing some terminology from analysis, where the kernel K is referred to as a mollifier, we call the measure P_h and the density p_h as the mollified measure and mollified density, respectively. For each h, the mollification of P by K yields that

1. the mollified measure P_h has full-dimensional support $S + B(0,h)$ and is absolutely continuous with respect to μ; here, for two set A and B in $\mathbb{R}^d$, $A + B \equiv \{x + y, : x \in A, y \in B\}$ denotes its Minkowski sum;

2. the mollified density p_h is of class $\mathcal{C}^\alpha$ whenever K is of class $\mathcal{C}^\alpha$, with $\alpha \in \mathbb{N}_+ \cup \{\infty\}$.

Thus, mollifying P makes it better behaved. At the same time, P_h and p_h can be seen as approximations of the original measure P and the geometric density, respectively, in a sense made precise by the following result.

Lemma 2.1. *As $h \to 0$, P_h converges weakly to P and $\lim_{h\to 0} p_h(x) = p(x)$, almost everywhere P.*

To estimate the λ-clusters of p, we use the connected components of $\widehat{L}$, i.e. the λ-clusters of $\widehat{p}_h$. That is, we estimate L with

$$\widehat{L} = \widehat{L}_h = \Big\{x : \ \widehat{p}_h(x) > \lambda\Big\}. \tag{5}$$

In practice, finding the estimated clusters is computationally difficult. Indeed, to see if two points X_i and X_j are in the same cluster, we need to check every possible path connecting X_i and X_j. If the minimum of $\widehat{p}_h$ along at least one such path is larger than λ then X_i and X_j are in the same cluster. We discuss an algorithm for approximating the clusters in Section 5. Until then, we ignore the computational problems and assume that the λ-clusters of $\widehat{p}_h$ can be computed exactly.

2.2 Risk

We consider three different risk functions.

- The *level set risk* is defined to be $R^L(p, \widehat{p}_h) = \mathbb{E}(\rho_P^\dagger(p, \widehat{p}))$, where

$$\rho_P^\dagger(r, q) = \int_{\{r>\lambda\}\Delta\{q>\lambda\}} dP(x), \tag{6}$$

  and $A\Delta B = (A \cap B^c) \cup (A^c \cap B)$ is the symmetric set difference.

- Define the *excess mass functional* as

$$\mathcal{E}(A) = P(A) - \lambda\mu(A), \tag{7}$$

  for any measurable set $A \subset \mathbb{R}^d$. This functional is maximized by the true level set L; see Mueller and Sawitzki (1991) and Polonik (1995). We can use the excess mass functional as a risk function except, of course, that we maximize it rather than minimize it. Given an estimate $\widehat{L}$ of L we will then be interested in making the *excess mass risk*

$$R^M(p, \widehat{p}_h) = \mathcal{E}(L) - \mathbb{E}(\mathcal{E}(\widehat{L})) \tag{8}$$

  as small as possible. Furthermore, simple algebra reveals that maximizing $\mathcal{E}(A)$ is equivalent to minimizing,

$$\int_{A\Delta L} |p - \lambda| d\mu$$

  which is the loss function used by Willett and Nowak (2007).

- The *Modified Rand risk* is defined as follows. For an extended real valued non-negative function r (which may take on the value ∞), we write $x \overset{r}{\sim} y$ if there exists a path γ on the graph of r between $r(x)$ and $r(y)$ such that either $r^{-1}(\gamma) \subset L(\lambda)$ or $r^{-1}(\gamma) \subset L(\lambda)^c$. The *modified Rand risk* is defined as

$$R^R(p, \widehat{p}_h) = \mathbb{E}\left(\rho_P(p, \widehat{p}_h)\right), \tag{9}$$

where

$$\rho_P(r,q) = \mathbb{P}\Big(M_r(Z_1,Z_2) \neq M_q(Z_1,Z_2)\Big),$$

the pair (Z_1, Z_2) is an i.i.d. sample from P, and

$$M_r(x,y) = \begin{cases} 1 & \text{if } r(x) > \lambda \text{ and } r(y) > \lambda \text{ and } x \overset{r}{\sim} y \\ 0 & \text{if } r(x) < \lambda \text{ and } r(y) < \lambda \text{ and } x \overset{r}{\sim} y \\ * & \text{otherwise.} \end{cases}$$

Notice that in equation (9) the expectation is with respect to the joint distribution of the observed sample X and of the pair (Z_1, Z_2).

2.3 Low Noise Conditions

Throughout our analysis we assume the following low noise conditions.

(LN1) There exist positive constants γ, C_1 and $\overline{\epsilon}$ such that

$$\mathbb{P}\Big(|p(X) - \lambda| < \epsilon\Big) \leq C_1 \epsilon^\gamma, \quad \forall \epsilon \in [0, \overline{\epsilon}).$$

(LN2) There exist positive constants $\overline{h}$, C_2, C_3, $\xi \geq d$ and a permutation σ of $\{1, \ldots, k\}$ such that, for all $0 < h < \overline{h}$ and all $\lambda' \in (\lambda - \overline{\epsilon}, \lambda + \overline{\epsilon})$, $L_h(\lambda') = \bigcup_{j=1}^k C_j^h$ where

(a) $C_i^h \cap C_j^h = \emptyset$ for $1 \leq i < j \leq k$;

(b) $C_j \subseteq C_{\sigma(j)}^h$, for all $1 \leq j \leq k$;

(c) $\mathbb{P}(C_{\sigma(j)}^h - C_j) \leq C_2 h^\xi$;

(d) $\mu(L(\lambda - \epsilon) + B(0,h)) \leq C_3 h^d$ for all $0 \leq \epsilon \leq \overline{\epsilon}$.

2.4 Remarks on The Low Noise Conditions

Condition (LN1), first introduced in Polonik (1995), provides a way to relate the stochastic fluctuations of $\widehat{p}_h$ around its mean p_h to the clustering risk. Indeed, the larger γ, the smaller the effects of these fluctuations, and the easier it is to obtain good clusters from noisy estimates of p_h, for any $h < \overline{h}$. Conditions (LN2) (a) - (c) offer instead a way of controlling the approximation error (bias) we incur by estimating p_h instead of p, locally in a neighborhood of λ. Notice that we do not require p to satisfy any smoothness condition. See Section 3.4 for revised (LN2) conditions in case of smooth densities.

Conditions (LN2) (a) and (b), though quite mild, are particularly important, as they directly imply that the estimated density $\widehat{p}_h$ can be used quite effectively for clustering purposes, for a range of bandwidth values. This is is shown in the next, simple result. Let $N(\lambda)$, $N_h(\lambda)$ and $\widehat{N}_h(\lambda)$ denote the number of λ-clusters for p, p_h and $\widehat{p}_h$, respectively.

Lemma 2.2. *Under conditions (LN2) (a)-(b) and for all $\epsilon \in (0, \overline{\epsilon})$ and $h \in (0, \overline{h})$, on the event $\mathcal{E}_{h,\epsilon} \equiv \{\|\widehat{p}_h - p_h\|_\infty < \epsilon\}$,*

$$N(\lambda) = N_h(\lambda) = \widehat{N}_h(\lambda) = k.$$

Condition (LN2) (c) is quite weak. When P has full-dimensional support and the boundary of L has small curvature or $\overline{h}$ is small enough, ξ is typically d. If C_j has dimension smaller than d, then $\mathbb{P}(C_{\sigma(j)}^h - C_j) = 0$. In particular, $\xi = \infty$ occurs when $L = S$, which is the most favorable case.

Condition (LN2) (d) is technical and needed in Theorem 3.5 to obtain consistency rates for the excess mass risk. It is also very mild, as it holds, for example, if the boundary of L is smooth or even if L is a lower dimensional smooth manifold with bounded curvature.

Although the rates are not affected by the constants, in practice, they can have a significant effect on the results, since they may very well depend on d. This is especially true of C_1, as illustrated in Example 2.6 below.

We conclude this section with some comments on the parameter γ appearing in condition (LN1), whose value affects in a crucial way the consistency rates, with faster rates arising from larger values of γ. If S has dimension smaller than d, then, clearly, $\gamma = \infty$, thus throughout this subsection we assume that P is a probability measure on $\mathbb{R}^d$ having Lebesgue density p.

First, a fairly general sufficient condition for assumption (LN1) to hold with $\gamma = 1$ at λ can be easily obtained using probabilistic arguments as follows. Let G denote the distribution of the random variable $Y = p(X)$ and suppose G has a Lebesgue density g which is bounded away from 0 and infinity on $(\lambda - \bar{\epsilon}, \lambda + \bar{\epsilon})$. Then, by the mean value theorem, for any $\epsilon < \bar{\epsilon}$,

$$\mathbb{P}(\lambda - \epsilon \leq p(X) \leq \lambda + \epsilon) = G(\{y \colon y \in (\lambda + \epsilon, \lambda - \epsilon)\}) = \epsilon g(\lambda + \eta),$$

for some $\eta \in (-\epsilon, \epsilon)$. Thus, (LN1) holds with $\gamma = 1$ at λ. See also Example 2.6 below. A more refined result based on analytic conditions is given next. Below $\mathcal{H}^{d-1}$ denotes the $(d-1)$-dimensional Hausdorff measure in $\mathbb{R}^d$. See Section 9.1 for the definition of Hausdorff measure.

Lemma 2.3. *Suppose that P is a probability measure on $\mathbb{R}^d$ having Lipschiz density p. Assume that, almost everywhere μ, $\|\nabla p(x)\| > 0$ and that $\mathcal{H}^{d-1}(\{x \colon p(x) = \lambda\}) < \infty$ for any $\lambda \in (0, \|p\|_\infty)$. Then, (LN1) holds with $\gamma = 1$ for each $\lambda \in (0, \|p\|_\infty)$ outside of a set of Lebesgue measure 0.*

A further point of interest is to characterize the set of λ values for which, given a class of densities, condition (LN1) holds with $\gamma \neq 1$. Clearly, if p has a jump discontinuity, then (LN1) is verified with $\gamma = \infty$, for all values of λ in some interval. On the other hand, on the account of the previous result, if $\|\nabla p\|$ is bounded away from 0 and ∞ in a neighborhood of $p^{-1}(\lambda)$, then $\gamma = 1$. Thus one could expect a value of γ different than 1 when ∇p does not exist or when $\|\nabla p\|$ is infinity or vanishes in $p^{-1}(\lambda)$. See the example on page 7 in Rigollet and Vert (2006), where (LN1) holds with $\gamma < 1$ if $q > d$ and $\gamma > 1$ if $q < d$, the former case corresponding to $\|\nabla p(x_0)\| = 0$ and the latter to $\lim_{x \to x_0} \|\nabla p(x)\| = \infty$. However, this would seem to indicate that, if p is sufficiently regular, the values of λ for which $\gamma \neq 1$ form a negligible set of $\mathbb{R}$. Lemma 2.3 above already shows that this set has Lebesgue measure zero if p is Lipschitz with non-vanishing gradient. Under stronger assumptions, it can be verified that this set is in fact finite.

Corollary 2.4. *Under the assumption of Lemma 2.3, if p is of class $\mathcal{C}^1$ and has compact support, then the set of λ such that (LN1) holds with $\gamma \neq 1$ is finite.*

Example 2.5. Sharp Clusters. Suppose that $p = \frac{dP}{d\mu} = \sum_{i=1}^m \pi_j p_j$ where p_i is a density with support on a compact, connected set S_i, $\sum_i \pi_i = 1$ and $\min_i \pi_i > 0$. Moreover suppose that

$$\min_{s \neq t} d(C_s, C_t) > 0$$

where $d(A, B) = \inf_{x \in A, y \in B} ||x - y||$. Finally suppose that

$$\min_j \inf_{x \in C_j} \pi_j p(x) > \lambda.$$

Sharp clusters of this type were considered by Singh et al. (2009), for example. It is easy to see that (LN1) and (LN2) hold with $\gamma = \xi = \infty$. A more general example in which one of the mixture component is supported on a lower dimensional set is shown in Figure 1. Here, the true distribution is $P = (1/3)\text{Unif}(-5.5, -4.5) + (1/3)\text{Unif}(4.5, 5.5) + (1/3)\delta_0$. The geometric density and the mollified density based on $h = .04$ are shown in the top plot. The point mass at 0 is indicated with a vertical bar. The bottom plot shows the true clusters and the mollified clusters based on p_h with $\lambda = .04$. The clusters based on p_h contain the true clusters and the difference between them is a set of zero probability.

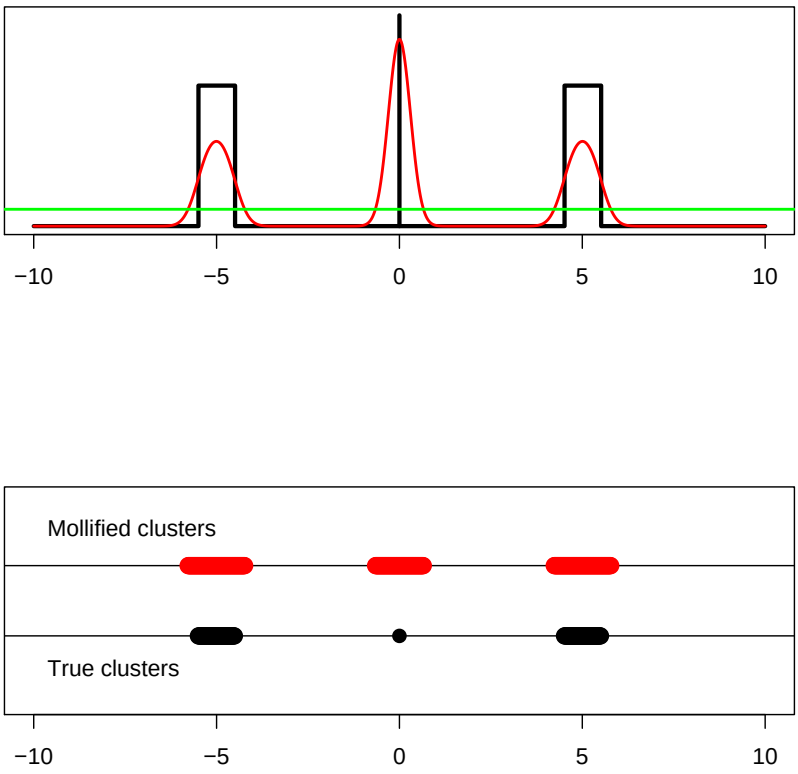

Figure 1: Sharp clusters. Top: the density of $P = (1/3)\text{Unif}(-5.5, -4.5) + (1/3)\text{Unif}(4.5, 5.5) + (1/3)\delta_0$ and the mollified density p_h for $h = .04$. The point mass at 0 is indicated with a vertical bar. Bottom: the true clusters and the mollified clusters of p_h with $\lambda = .04$.

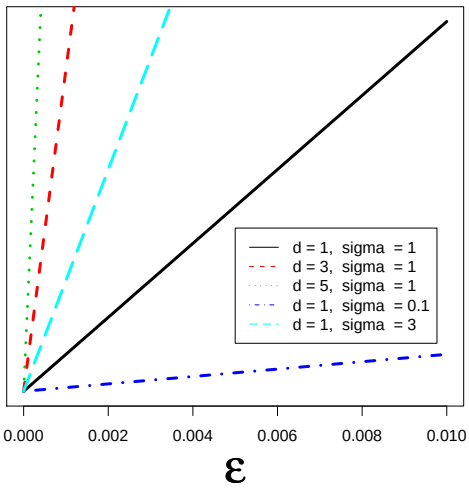

Figure 2: Noise exponent for Gaussians. Each curve shows $\mathbb{P}(|p(X)-\lambda|<\epsilon)$ versus ϵ for $\alpha = 1/2$. The plots are nearly linear since $\gamma = 1$ in this case.

Example 2.6. Normal Distributions. Suppose that $X \sim N_d(0,\Sigma)$, with Σ positive definite. Set $\sigma = |\Sigma|^{1/2}$. Then, (LN1) holds for any $0 \le \lambda \le \left(\sigma\left(\sqrt{2\pi}\right)^d\right)^{-1}$ with $\gamma = 1$ and $C_1 = C_d 2\sigma\left(\sqrt{2\pi}\right)^d$, where the constant C_d depends on d. For simplicity, we prove the claim only for $\lambda = \alpha\left(\sigma\left(\sqrt{2\pi}\right)^d\right)^{-1}$, where $\alpha \in (0,1)$. Cases in which $\alpha = 1$ or $\alpha = 0$ can be dealt with similarly. Let $W \sim \chi^2_d$ and notice that $X^\top \Sigma^{-1} X \stackrel{d}{=} W$. For all $\epsilon > 0$ smaller than

$$\min\left\{\frac{\alpha}{\sigma\left(\sqrt{2\pi}\right)^d}, \frac{(1-\alpha)}{\sigma\left(\sqrt{2\pi}\right)^d}\right\}, \tag{10}$$

simple algebra yields

$$\begin{aligned} P(|\phi_\sigma(X)-\lambda|<\epsilon) &= \mathbb{P}\left(2\log\tfrac{1}{\alpha-\epsilon\sigma\left(\sqrt{2\pi}\right)^d} \le W \le 2\log\tfrac{1}{\alpha+\epsilon\sigma\left(\sqrt{2\pi}\right)^d}\right) \\ &= 2\left(\log\tfrac{1}{\alpha-\epsilon\left(\sigma\sqrt{2\pi}\right)^d} - \log\tfrac{1}{\alpha+\eta\sigma\left(\sqrt{2\pi}\right)^d}\right) p_d\left(\log\tfrac{1}{\alpha+\eta\sigma\left(\sqrt{2\pi}\right)^d}\right), \end{aligned}$$

for some $\eta \in (-\epsilon,\epsilon)$ where p_d denotes the density of a χ^2_d distribution and the second equality holds in virtue of the mean value theorem. By a first order Taylor expansion, for $\epsilon \downarrow 0$, the first term on the right hand side of the previous display can be written as

$$2\epsilon\sigma\left(\sqrt{2\pi}\right)^d\left(\frac{1}{\alpha-\epsilon\sigma\left(\sqrt{2\pi}\right)^d}+\frac{1}{\alpha+\epsilon\sigma\left(\sqrt{2\pi}\right)^d}\right)+o(\epsilon^2).$$

Since $\left(\frac{1}{\alpha-\epsilon\sigma\left(\sqrt{2\pi}\right)^d}+\frac{1}{\alpha+\epsilon\sigma\left(\sqrt{2\pi}\right)^d}\right) p_d\left(\log\frac{1}{\alpha+\eta\sigma\left(\sqrt{2\pi}\right)^d}\right) \asymp 1$ for any $\epsilon \ge 0$ bounded by (10), the claim is proved. See Figure 2.

3 Rates of Convergence

In this section we study the rates of convergence in the three distances using deterministic bandwidths. We defer the discussion of random (data driven) bandwidths until Section 4.

3.1 Preliminaries

Before establishing consistency rates for the different risk measures described above, we discuss some necessary preliminaries.

In our analysis we require the event

$$\mathcal{E}_{h,\epsilon} \equiv \{||\widehat{p}_h - p_h||_\infty \leq \epsilon\}, \quad \epsilon > 0, h > 0, \tag{11}$$

to hold with high probability, for all n large enough. In fact, some control over $\mathcal{E}_{h,\epsilon}$ provides a means of bounding the clustering risks, as shown in the following result.

Lemma 3.1. *Let $\epsilon \in (0, \overline{\epsilon})$ and $h \in (0, \overline{h})$ be such that the conditions (LN1) and (LN2) (a)-(c) are satisfied. Then, on the event $\mathcal{E}_{h,\epsilon}$,*

$$L(\lambda + \epsilon) \subseteq \widehat{L}_h(\lambda) \subseteq L(\lambda + \epsilon) \cup A \cup B$$

where

$$A = L(\lambda - \epsilon) - L(\lambda + \epsilon)$$

and

$$B = L_h(\lambda - \epsilon) - L(\lambda - \epsilon).$$

Therefore, on $\mathcal{E}_{h,\epsilon}$,

$$P\left(\widehat{L}_h(\lambda) \Delta L(\lambda)\right) \leq C_1 \epsilon^\gamma + C_2 h^\xi. \tag{12}$$

In order to bound $\mathbb{P}(\mathcal{E}_{h,\epsilon}^c)$, we study the properties of the kernel estimator $\widehat{p}_h$. We will impose the following condition on the kernel K, due to Nolan and Pollard (1987).

(K) The kernel K is a bounded, squared integrable function on $\mathbb{R}^d$ in the linear span of nonnegative real-valued functions k on $\mathbb{R}^d$ such that the subgraph of k, $\{(s,u) \in \mathbb{R}^d \times \mathbb{R} \colon k(s) \geq u\}$, can be represented as a finite number of Boolean operations among sets of the form $\{(s,u) \in \mathbb{R}^d \times \mathbb{R} \colon p(s,u) \geq \psi(u)\}$, where p is a polynomial on $\mathbb{R}^d \times \mathbb{R}$ and ψ an arbitrary real-valued function. The number A and v are called the VC characteristics of K.

If property (K) holds, then, for any $h > 0$, the class of functions

$$\mathcal{F}_h = \left\{ K\left(\frac{x - \cdot}{h}\right), x \in \mathbb{R}^d \right\} \tag{13}$$

satisfies

$$\sup_P N(\mathcal{F}_h, L_2(P), \epsilon \|F_h\|_{L_2(P)}) \leq \left(\frac{A}{\epsilon}\right)^v,$$

where $N(T, d, \epsilon)$ denotes the ϵ-covering number of the metric space (T, d), F_h is the envelope function of $\mathcal{F}_h$ and the supremum is taken over the set of all probability measures on $\mathbb{R}^d$. See Giné and Guillou (2002).

Using condition (K), we can establish the following finite sample bound for $\mathbb{P}\left(\|\widehat{p}_h - p_h\|_{infty} > \epsilon\right)$, which is obtained as a direct application of results in Giné and Guillou (2002). Throughout the paper, we will assume that the conditions in the next Proposition are satisfied.

Proposition 3.2 (Gine and Guillon)**.** *Assume that the kernel satisfies the property (K) and that*

$$\sup_{t \in \mathbb{R}^d} \sup_{h > 0} \int_{\mathbb{R}^d} K_h^2(t - x) dP(x) < D < \infty. \tag{14}$$

1. *Let h be fixed. Then, there exist constants $L > 0$ and $C > 0$, which depend only on the VC characteristics of K, such that, for any $c_1 \geq C$ and $0 < \epsilon \leq \frac{c_1 D}{\|K\|_\infty}$, there exists an $n_0 > 0$, which depends on ϵ, D, $\|K\|_\infty$ and the VC characteristics of K, such that, for all $n \geq n_0$,*

$$\mathbb{P}\left\{\sup_{x\in\mathbb{R}^d} |\widehat{p}_h(x) - p_h(x)| > 2\epsilon\right\} \leq L \exp\left\{-\frac{1}{L}\frac{\log(1 + c_1/(4L))}{c_1}\frac{nh^d\epsilon^2}{D}\right\}. \tag{15}$$

2. *Let $h_n \to 0$ as $n \to \infty$ in such a way that $\frac{nh_n^d}{|\log h_n^d|} \to \infty$. If $\{\epsilon_n\}$ is a sequence such that*

$$\epsilon_n = \Omega\left(\sqrt{\frac{\log r_n}{nh_n^d}}\right), \tag{16}$$

where $r_n = \Omega\left(h_n^{-d/2}\right)$, then, for all n large enough, (15) holds with h and ϵ replaced by h_n and ϵ_n, respectively. In particular, the term on the right hand side of (15) vanishes at the rate $O\left(r_n^{-1}\right)$.

The above theorem imposes minimal assumptions on the kernel K and, more importantly, on the probability distribution P, whose density is not required to be bounded or smooth, and, in fact, may not even exist. Finally, we remark that, for fixed h, setting $\epsilon_n = \sqrt{\frac{2\log n}{h^d n C_K}}$ for an appropriate constant C_K (depending on K), an application of the Borel-Cantelli Lemma yields that, as $n \to \infty$, $\|p_h - \widehat{p}_h\|_\infty = O\left(\sqrt{\frac{\log n}{n}}\right)$ almost everywhere P. See also Einmahl and Mason (2005) for a uniform version of this claim, which also relies on assumption (K).

3.2 Rates of Convergence

We now derive the converge rates for the clustering risks defined in Section 2.2. Below we will write C_K for a constant whose value depends only on the VC characteristic of the kernel K and on the constant D appearing in (14).

Theorem 3.3. *(Level Set Risk.) Suppose that (LN1) and (LN2) (a)-(c) hold. For any $h \in (0, \overline{h})$ and $\epsilon \in (0, \overline{\epsilon})$,*

$$R^L(p, \widehat{p}_h) = O\left(\epsilon^\gamma + h^\xi + e^{-C_K n h^d \epsilon^2}\right). \tag{17}$$

In particular, setting

$$h_n = \left(\frac{\log n}{n}\right)^{\gamma/(2\xi+d\gamma)} \quad \text{and} \quad \epsilon_n = \sqrt{\frac{\log n}{C_K n h_n^d}}$$

we obtain

$$R^L(p, \widehat{p}_h) = O\left(\max\left\{\left(\frac{\log n}{n}\right)^{\frac{\gamma\xi}{2\xi+d\gamma}}, \frac{1}{n}\right\}\right). \tag{18}$$

According to the theorem, if $\xi \geq d\gamma/2$ then

$$\mathbb{E}(\rho^\dagger(p, \widehat{p}_h, P)) = O\left(\max\left\{\left(\frac{\log n}{n}\right)^{\gamma/4}, \frac{1}{n}\right\}\right).$$

In particular, fast rates (including a $\log n$ term) arise when

$$\frac{\gamma\xi}{2\xi + d\gamma} > \frac{1}{2},$$

which holds provided that $\gamma > \frac{\xi}{\xi - d/2}$. If $\xi = d$, this is satisfied when $\gamma > 2$. If $\xi = d$, the risk is of large order $(\log n/n)^{\frac{\gamma}{\gamma+2}}$, which for $\gamma = 1$, becomes of large order $(\log n/n)^{1/3}$. Note that these rates do not depend on the dimension d.

As a corollary, we can show that the expected proportion of sample points that are incorrectly assigned as clusters or noise vanishes at the same rate.

Corollary 3.4. *Let* $\widehat{f}_h = \frac{|\widehat{I}_h|}{n}$, *where*

$$\widehat{I}_h = \{i\colon \ \mathrm{sign}(\widehat{p}_h(X_i) - \lambda) \neq \mathrm{sign}(p(X_i) - \lambda)\}.$$

Then, $\mathbb{E}(\widehat{f}_h) \leq O\left(\epsilon^\gamma + h^\xi + e^{-C_K n h^d \epsilon^2}\right)$.

We now turn to the excess mass risk.

Theorem 3.5. *(Excess Mass.) Suppose that (LN1) and (LN2) hold. Then, for any* $h \in (0, \overline{h})$, $\epsilon \in (0, \overline{\epsilon})$ *and* $\lambda > \epsilon$,

$$R^M(p, \widehat{p}_h) = O\left(\epsilon^{\gamma+1} + h^d + (1 + \lambda C_S) e^{-n C_K \epsilon^2 h^d}\right),$$

where C_S *is a constant independent of* ϵ *and* h. *Thus, setting*

$$h_n = \left(\frac{\log n}{n}\right)^{\frac{\gamma+1}{d(\gamma+3)}} \quad \text{and} \quad \epsilon_n = \sqrt{\frac{\log(n/(1+\lambda C_S))}{C_K n h_n^d}},$$

we obtain

$$R^M(p, \widehat{p}_h) = O\left(\max\left\{\left(\frac{\log n}{n}\right)^{\frac{\gamma+1}{\gamma+3}}, \frac{1}{n}\right\}\right) \tag{19}$$

We remark that the rates we obtain are all faster then $(\log /n)^{1/3}$ and dimension independent.

In our last result, we show that the modified rand risk satisfies the same upper bound (17) established for the level set risk, and, therefore, vanishes at the same rate.

Theorem 3.6. *(Rand Distance) Suppose that (LN1) and (LN2) (a) - (c) hold and let* $h \in (0, \overline{h})$ *and* $\epsilon \in (0, \epsilon_0)$. *Then,* $R^R(p, \widehat{p}_h) = O\left(\epsilon^\gamma + h^\xi + e^{-C_K n h^d \epsilon^2}\right)$, *and the same rates of Theorem 3.3 hold.*

Remark. Our proofs rely in a fundamental way on Proposition 3.2, which holds with virtually almost no conditions on the probability P. On the other hand, this yields the additional $\log n$ term in our results. Such term can be eliminated by assuming smoothness conditions on p. See Section 3.4 below.

3.3 Biased Clusters

In some cases, we might be content with estimating the level set $L_h(\lambda)$, which is a biased version of $L(\lambda)$. That is, the fringe $L_h(\lambda) - L(\lambda)$ may not be of great practical concern. In that case we have the following result, which gives faster, dimension independent rates. The proof is similar to the proofs of the previous results and is omitted.

Theorem 3.7. *Let* $h \in (0, \overline{h})$ *be fixed. Under (LN1) and (LN2) (a)-(b),* $R^L(p, \widehat{p}_h)$ *and* $R^R(p, \widehat{p}_h)$ *are* $O\left(\max\left\{\left(\frac{\log n}{n}\right)^{\gamma/2}, \frac{1}{n}\right\}\right)$. *If, in addition, (LN2) (d) is satisfied, then*

$$R^M(p, \widehat{p}_h) = O\left(\max\left\{\left(\frac{\log n}{n}\right)^{\frac{1+\gamma}{2}}, \frac{1}{n}\right\}\right). \tag{20}$$

3.4 Some Special Cases

The case $\gamma = \infty$. Then, the sequence $\{\epsilon_n\}$ does not need to vanish. Thus, setting $h_n = \left(\frac{\log n}{n}\right)^{1/d}$, since $\xi \geq d$, it is immediate to see that the three clustering risks vanish at a rate that is at least $O\left(\frac{\log n}{n}\right)$.

The case $\gamma = \xi = \infty$. This is the most favorable case for clustering, corresponding for example, to sharp clusters. The rate $O(1/n)$ is achieved with any fixed positive bandwidth smaller than $\overline{h}$.

The smooth case. If P has full-dimensional support and the Lebesgue density p is smooth, different results are possible. For example, using the same settings of Rigollet and Vert (2006), if p is β-times Hölder differentiable, then the bias condition (LN2) is superfluous, as

$$\|p_h - p\|_\infty \leq Ch^\beta, \tag{21}$$

for some constant C which depends only on the kernel K. Choosing h such that $Ch^\beta < \epsilon$, on the event $\mathcal{E}_{h,\epsilon}$, the triangle inequality yields $\|\widehat{p}_h - p\|_\infty < 2\epsilon$. Thus, for each $\epsilon < \frac{\overline{\epsilon}}{2}$ and each h such that $Ch^\beta < \epsilon$, on $\mathcal{E}_{h,\epsilon}$, instead of (12), one obtains

$$P\left(\widehat{L}_h(\lambda)\Delta L(\lambda)\right) \leq C_1 2^\gamma \epsilon^\gamma.$$

Then, setting $h_n = (\log n/n)^{\frac{1}{2\beta+d}}$ and $\epsilon_n = \Omega((\log n/n))^{\frac{\beta}{2\beta+d}})$, we see that both $R^L(p, \widehat{p}_h)$ and $R^R(p, \widehat{p}_h)$ are of order $O((\log n/n)^{\frac{\gamma\beta}{2d+\beta}})$, while $R^M(p, \widehat{p}_h)$ is of order $O((\log n/n)^{\frac{(\gamma+1)\beta}{2d+\beta}})$. These, are, up to an extra logarithmic factor, the minimax rates established by Rigollet and Vert (2006). In fact, under these smoothness assumptions, and since the bias can be uniformly controlled as in (21), then, by a combination of Fubini's theorem and of a peeling argument as in Audibert and Tsybakov (2007) and Rigollet and Vert (2006), the exponential term $O\left(e^{-C_K n h^d \epsilon^2}\right)$ becomes redundant and one obtains rates without the extra logarithmic term.

4 Choosing the Bandwidth

In this section we discuss two data-driven method for choosing the bandwidth. Before we explain the details, we point out that L_2 cross-validation is not appropriate for this problem. In fact, we are allowing for the case where P may have atoms, in which case it is well known that cross-validation chooses $h = 0$.

4.1 Excess Mass

We propose choosing h by splitting the data and maximizing an empirical estimate of the excess mass functional. Polonik (1995) used this approach to choose a level set from among a fixed class $\mathcal{L}$ of level sets. Here, we are choosing a bandwidth, or, in other words, we are choosing a level set from a random class of level sets $\mathcal{L} = \{\{\widehat{p}_h > \lambda\} : \ h > 0\}$ depending on the observed sample X. The steps are in Table 1.

Remarks.

1. To implement the method, we need to compute $\mu(L_h)$. In practice $\mu(L_h)$ can be approximated by

$$\frac{1}{M}\sum_{i=1}^{M} \frac{I(\widehat{p}_h(U_i) > \lambda)}{g(U_i)}$$

   where $U_1, \ldots, U_M$ is a sample from a convenient density g. In particular, one can choose $g = \widehat{p}_H$ for some large bandwidth H. Choosing $M \approx n^2$ ensures that the extra error this importance sampling estimator is $O(1/n)$ which is negligible. We ignore this error in what follows.

2. Technically, the method only applies for $\lambda > 0$, at least in terms of the theory that we derive. In practice, it can be used for $\lambda = 0$. In this case, $\widehat{\mathcal{E}}(h)$ becomes 1 when h is large. We then take $\widehat{h}$ to be the smallest h for which $\widehat{\mathcal{E}}(h) = 1$.

1. Split the data into two halves which we denote by $X = (X_1, \ldots, X_n)$ and $Z = (Z_1, \ldots, Z_n)$.

2. Let $\mathcal{H}$ be a finite set of bandwidths. Using X, construct kernel density estimators $\{\widehat{p}_h : \ h \in \mathcal{H}\}$. Let $L_h = \{x : \ \widehat{p}_h(x) > \lambda\}$.

3. Using Z, estimate the excess mass functional

$$\widehat{\mathcal{E}}(h) = \frac{1}{n}\sum_{i=1}^{n} I(Z_i \in L_h) - \lambda\mu(L_h).$$

4. Let $\widehat{h}$ be the maximizer of $\widehat{\mathcal{E}}(h)$ and set $\widehat{L} = L_{\widehat{h}}$.

Table 1: Selecting the bandwidth using the excess mass risk.

Below we use the notation $\mathcal{E}_X(\cdot)$ instead of $\mathcal{E}(\cdot)$ to indicate that the excess mass functional (7) is evaluated at a random set depending on the training set X. Accordingly, with some abuse of notation, for a $h > 0$, we will write $\mathcal{E}_X(h) = \mathcal{E}(L_h)$, with L_h the λ-level set of $\widehat{p}_h$. Below $\mathcal{H}$ is a countable subset of $[0, \overline{h}]$. The next result is closely related to Theorem 7.1 of Györfi et al. (2002).

Theorem 4.1. *Let $h_* = \text{armax}_{h\in\mathcal{H}}\mathcal{E}_X(h)$. For any $\delta > 0$,*

$$\mathbb{E}(\mathcal{E}_X(h_*)) - \mathbb{E}(\mathcal{E}_X(\widehat{h})) \leq d(\delta, \kappa)\frac{1 + \log 2}{n} \tag{22}$$

where the expectation is with respect to the joint distribution of the training and test set, $d(\delta, \kappa) = \frac{2}{\kappa}\delta(1 + \delta)\left(16\gamma^2 + \delta(7 + 16\gamma^2)\right)$, with $\kappa = 2 + \lambda\mu(S + B(0, \overline{h}))$ and $\gamma^2 = \frac{7}{4}\left(e^{4/7} - 1\right)$.

Now we construct a grid $\mathcal{H}_n$ of size depending on n that is guaranteed to ensure that optimizing over $\mathcal{H}_n$ implies we are adapting over γ.

Theorem 4.2. *Suppose (LN1) and (LN2) hold. Let*

$$\delta_n = \frac{9}{4L_n}\left(\frac{\log n}{n}\right)^{2/3}$$

where $L_n = \log n - \log\log n$. Let $G_n = \{\gamma_1, \ldots, \gamma_N\}$ where $\gamma_j = (j-1)\delta_n$ and N is the smallest integer greater than or equal to

$$\frac{2L_n(L_n - \log 2)}{9\log 2}\left(\frac{n}{\log n}\right)^{2/3}.$$

Let $\mathcal{H}_n = \{h_n(\gamma)\ \gamma \in G_n\}$ where $h_n(\gamma) = (\log n/n)^{(\gamma+1)/(d(3+\gamma))}$. Then

$$\mathcal{E}(L) - \mathbb{E}(\mathcal{E}(\widehat{L})) \leq O\left(\frac{\log n}{n}\right)^{\frac{\gamma+1}{\gamma+3}}.$$

Remarks.

1. We are choosing the bandwidth from a single split of the data. An alternative is to split the data many times and combines the estimates over multiple splits.

2. When $\mu(L) = 0$, we have that $h_* = 0$. The above theorems are still valid in this case. Thus the case where P is atomic is included while it is ruled out for L_2 cross-validation.

4.2 Stability

Another method for selecting the bandwidth is to choose the value for h that produces stable clusters, in a sense defined below. The use of stability has gained much popularity in clustering; see Ben-Hur et al. (2002) and Lange et al. (2004) for example. In the context of k-means clustering and related methods, Ben-David et al. (2006) showed that minimizing instability leads to poor clustering. Here we investigate the use of stability for density clustering.

Given three independent samples $X = (X_1, \ldots, X_n)$, $Y = (Y_1, \ldots, Y_n)$ and $Z = (Z_1, \ldots, Z_n)$, we define the instability be

$$\Xi(h) = \rho(\widehat{p}_h, \widehat{q}_h, \widehat{P}_Z) \tag{23}$$

where $\widehat{p}_h$ is constructed from X, $\widehat{q}_h$ is constructed from Y and $\widehat{P}_Z$ is the empirical distribution based on Z. (In practice, we split the data into three subsets.)

Rather than study stability in generality, we consider a special case involving the following extra conditions:

1. **Sharp Clusters.** Assume that $P = \sum_{j=1}^m \pi_j P_j$ where $\sum_j \pi_j = 1$, and P_j is uniform on the compact set C_j of full dimension d. Thus, $p(z) = \sum_j \Delta_j I(z \in C_j)$ where $\Delta_j = \pi_j / \mu(C_j)$. Let $\underline{\Delta} = \min_j \Delta_j > 0$ and let $\overline{\Delta} = \max_j \Delta_j$.

2. **Spherical Kernel.** We use a spherical kernel so that

$$\widehat{p}_h(z) = \frac{1}{nh^d} \sum_{i=1}^n \frac{I(||z - X_i|| \leq h)}{v_d} = \frac{\widehat{P}(B(x,h))}{h^d v_d}$$

   where $v_d = \pi^{d/2}/\Gamma(d/2+1)$ denotes the volume of the unit ball and $\widehat{P}$ is the empirical measure.

3. S **is a standard set.** Assume that there exists a $\delta \in (0,1)$ such that

$$\mu(B(z,h) \cap L) \geq \delta \mu(B(z,h)) \quad \text{for all } z \in S, \text{ and all } h < \text{diam}(S),$$

   where $\text{diam}(S) = \sup_{(x,y) \subset S} \|x - y\|$ indicates the diameter of the set S. The notion of a standard set was originally introduced by Cuevas and Fraiman (1997).

4. **Choice of** λ**.** We take $\lambda = 0$.

We will focus on the level set distance $\rho^\dagger(r, q, P) = \int_{\{r > \lambda\} \Delta \{q > \lambda\}} dP(x)$. The level set instability is

$$\Xi(h) = \rho^\dagger(\widehat{p}, \widehat{q}, \widehat{P}_Z). \tag{24}$$

Under these settings, the graph $\Xi(h)$ is typically unimodal with $\Xi(0) = \Xi(\infty) = 0$. Hence, it makes no sense to minimize Ξ. Instead, we will fix a constant α and choose

$$\widehat{h} = \inf\left\{ h : \sup_{t > h} \Xi(t) \leq \alpha \right\}. \tag{25}$$

Theorem 4.3. *Let* $h_* = \text{diam}(L)$. *Under conditions 1-4,*

1. $\Xi(0) = 0$ *and* $\Xi(h) = 0$, *for all* $h \geq h_*$;
2. $\sup_{0 < h < h_*} \mathbb{E}(\Xi(h)) \leq 1/2$;
3. *As* $h \to 0$, $\mathbb{E}(\Xi(h)) \asymp h^d$;

4. for each $h \in (0, h_*)$,

$$C_3(h_* - h)^{d(n+1)} C_4^n \leq \mathbb{E}(\Xi(h)) \leq 2C_1(h_* - h)^{n+1} C_2^n$$

where

$$\begin{aligned} C_1 &= \frac{\pi^{d/2} h_*^{d-1}}{2^d \Gamma((d/2)+1)}, \quad C_2 = \frac{\pi^{d/2} h_*^{d-1}}{\Gamma((d/2)+1)} \\ C_3 &= \frac{\delta \underline{\Delta} \pi^{d/2}}{\Gamma((d/2)+1)}, \quad C_4 = \frac{\underline{\Delta} \delta \pi^{d/2}}{\Gamma(d/2+1)}. \end{aligned}$$

To see the implication of Theorem 4.3, we proceed as follows. Consider a grid of values $\mathcal{H} \subset (0, h_*)$ of cardinality n^β, for some $0 < \beta < 1$. By Hoeffding's inequality, with probability at least $1 - \frac{1}{n}$, we have that

$$\sup_{h \in \mathcal{H}} |\Xi(h) - \mathbb{E}(\Xi(h))| \leq w_n \equiv \sqrt{\frac{2\log(2n)(1-\beta)}{n}}.$$

Replacing $\mathbb{E}(\Xi(h))$ by $\Xi(h) + w_n$ and $\Xi(h) - w_n$ in the upper and lower bounds of part *4.* of Theorem 4.3, respectively, setting them both equal to α and then finally solving for h, we conclude that the selected $\widehat{h}$ is upper bounded by

$$h_* - \left(\frac{\alpha - w_n}{2C_1}\right)^{1/(n+1)} C_2^{-\frac{n}{n+1}}$$

and lower bounded by

$$h_* - \left(\frac{\alpha + w_n}{C_3}\right)^{1/(d(n+1))} C_4^{-\frac{n}{d(n+1)}}$$

with probability larger than $1 - \frac{1}{n}$. Thus, as $n \to \infty$, the resulting bandwidth does not tend to 0. Hence, the stability based method leads to bandwidths that are quite different than the method in the previous section. Our explanation for this finding is that the stability criterion is essentially aimed at reducing the variability of the clustering solution, but it is virtually unaffected by the bias caused by large bandwidths.

In the analysis above we assumed for simplicity that $\lambda = 0$. When $\lambda > 0$, the instability $\Xi(h)$ can have some large peaks for very large h. This occurs when h is large enough so that some mode of $p_h(x)$ is close to λ. Choosing h according to (25) will then lead to serious oversmoothing. Instead, we can choose $\widehat{h}$ as follows. Let $h_0 = \text{argmax}_h \Xi(h)$ and define

$$\widehat{h} = \inf\left\{h : \ h \geq h_0, \ \ \Xi(h) \leq \alpha\right\}. \tag{26}$$

We will revisit this issue in Section 6. A theoretical analysis of this modified procedure is tedious and, in the interest of space, we shall not pursue it here.

5 Approximating the Clusters

In this section we study a graph-based algorithm, which is often called the "friends-of-friends" algorithm, for finding the connected components of $\widehat{L}_h$ and for estimating the number of λ clusters $N(\lambda)$ that is based on the ρ-nearest neighborhood graph of $\{X_i : \widehat{p}_h(X_i) > \lambda\}$. The "friends-of-friends" algorithm and its variants have been used by Devroy and Wise (1980) and Tsybakov and Korostelev (1993) for support estimation and, more recently, by Cuevas et al. (2000) and Biau et al. (2007) for estimating the number of λ-clusters. Our results offer similar guarantees but hold under more general settings.

Lemma 2.2 shows that, under mild, conditions and when the sample size is large enough, $N(\lambda) = \widehat{N}_h(\lambda)$ uniformly over $h \in (0, \overline{h})$ with high probability. However, computing the number of connected components of $\widehat{L}_h(\lambda)$ exactly is computationally difficult, especially if d is large. The "friends-of-friends" algorithm is instead fast and easy to implement. The algorithm proceeds as follows. For some $h \in (0, \overline{h}$ and given $\lambda \geq 0$,

1. Compute the kernel density estimate $\widehat{p}_h$;
2. compute the ρ-nearest neighborhood graph of $\{X_i : \widehat{p}_h(X_i) > \lambda\}$, that is the graph $\mathcal{G}_{h,n}$ on $\{X_i : \widehat{p}_h(X_i) > \lambda\}$ where there is an edge between any two nodes if and only if they both belong to a ball of radius ρ;
3. compute the connected components of $\mathcal{G}_{h,n}$ using a depth-first search.

We will show that, if ρ is chosen appropriately,y then, with high probability as $n \to \infty$,

1. the number of connected components of $\mathcal{G}_{h,n}$, $\widehat{N}_h^G(\lambda)$, matches the number of true clusters, $N(\lambda) = k$;
2. there exists a permutation of $\{1, \dots, k\}$ such that, for each j and j',

$$C_j^h \subseteq \bigcup_{x \in \mathcal{C}_{\sigma(j)}} B(x, \rho) \quad \text{and} \quad \left(\bigcup_{x \in \mathcal{C}_{\sigma(j)}} B(x, \rho) \right) \cap \left(\bigcup_{x \in \mathcal{C}_{\sigma(j')}} B(x, \rho) \right) = \emptyset, \tag{27}$$

   where $\mathcal{C}_1, \dots, \mathcal{C}_k$ are the connected components of $\mathcal{G}_{h,n}$.

We will assume the following regularity condition on the densities p_h, which is satisfied if the kernel K is of class $\mathcal{C}^1$ and P is not flat in a neighborhood of λ:

(G) There exist constants $\epsilon_1 > 0$ and $C_g > 0$ such that for each $h \in (0, \overline{h})$, p_h is of class $\mathcal{C}^1$ on $\{x : |p_h(x) - \lambda| < \epsilon_1\}$ and

$$\inf_{h \in (0,\overline{h})} \inf_{x \in \{|p_h(x) - \lambda| < \epsilon_1\}} \|\nabla p_h(x)\| > C_g. \tag{28}$$

Let $\delta_h = \min_{i \neq j} \operatorname{dist}(C_i^h, C_j^h)$ and set $\delta = \inf_{h \in (0,\overline{h})} \delta_h$. Notice that, under (LN2) (b), $\delta > 0$. Finally, let $\mathcal{O}_{h,n}$ denote the event in equation (27), which clearly implies the event $\{\widehat{N}_h^G(\lambda) = k\}$.

Theorem 5.1. *Assume conditions (G) and (LN2) (a)-(b) and let $d^* = \dim(L)$. Assume further that there exists a constant $\overline{C}$ such that, for every $r \leq \delta/2$ and for P-almost all $x \in S \cap L$,*

$$P(B(x, r)) > \overline{C} r^{d_i}, \tag{29}$$

where $d_i = \dim(S_i)$, with $x \in S_i$. Then, there exists positive constants $\overline{\rho}$ and $\overline{M}$, depending on d^* and L such that, for every $\rho < \min\{\delta/2, \overline{\rho}\}$, there exists a number $\epsilon(\rho)$ such that, for any $\epsilon < \eta(\rho)$,

$$\mathbb{P}\left(\mathcal{O}_{h,n}^c\right) \leq \mathbb{P}(\mathcal{E}_{h,\epsilon}^c) + \overline{M} \rho^{-d^*} e^{-\overline{C} n \rho^{d^*}},$$

uniformly in $h \in (0, \overline{h})$.

The previous result deserves few comments. First, the constants $\overline{\rho}$, $\overline{M}$ and $\overline{C}$ depend on d^*. Secondly, assumption (29) is a generalization to lower dimensional sets of the standardness assumption introduced in Cuevas and Fraiman (1997). It is clearly true for components P_i of full-dimensional support that are absolutely continuous with respect to the Lebesgue measure. Finally, in view of Lemma 8.1 (and, specifically, of the way $\epsilon(\rho, \tau)$ is defined), Theorem 5.1 holds for sequences $\{\epsilon_n\}$, $\{h_n\}$ and $\{\rho_n\}$ such that

1. $\epsilon_n = o(1)$,
2. $\sup_n h_n \leq \overline{h}$;
3. $\sup_n \rho_n < \min\{\delta/2, \overline{\rho}_d\}$ and $\epsilon_n = o(\rho_n)$.

In particular, if $h_n = o(1)$, then, following Proposition 3.2, the term $\mathbb{P}(\mathcal{E}_{h_n,\epsilon_n}^c)$ vanishes if $\frac{n h_n^d}{|\log h_n^d|} \to \infty$ Interestingly enough, condition (LN1) plays no role in Theorem 5.1.

We now consider a bootstrap extension of the previous algorithm, as suggested in Cuevas et al. (2000). For any h, let $X^* = (X_1^*, \dots, X_N^*)$ denote a bootstrap sample from $\widehat{p}_h$ conditionally on $\{\widehat{p}_h > \lambda\}$ and let $\mathcal{G}_{n,h}^*$ denote the ρ-neighborhood graph with node set X^*. Finally, let $\mathcal{O}_{h,n}^*$ be the event given in equation (27), except that $\mathcal{C}_1, \dots, \mathcal{C}_k$ are now the connected components of $\mathcal{G}_{h,n}^*$.

Theorem 5.2. *Assume conditions (LN2) (a)-(b) and (G). Suppose that there exist positive constants* $\overline{C}$ *and* $\overline{\rho}$ *such that*

$$\inf_{h\in(0,\overline{h})}\int_{A_h\cap L_h(\lambda)} p_h d\mu > \overline{C}\rho^d. \tag{30}$$

for any ball A_h *of radius* $\rho < \overline{\rho}$ *and center in* $L_h(\lambda)$*. Then, for any* $\rho \le \min\{\delta/2, \overline{\rho}\}$*, there exists a positive number* $\epsilon(\rho)$ *such that, for each* $\epsilon < \epsilon(\rho)$*,*

$$\mathbb{P}\left((\mathcal{O}^*_{h,n})^c\right) \le \mathbb{P}(\mathcal{E}^c_{h,\epsilon}) + \overline{M}\rho^{-d}e^{-NC\rho^d},$$

2 uniformly in $h \in (0, \overline{h})$*, where* $\overline{M}$ *and* C *are positive constants independent of* h *and* ρ*.*

The constants C, $\overline{C}$, $\overline{\rho}$ and $\overline{M}$ depend on both d and $S + B(0, \overline{h})$. In our settings, condition (30) clearly holds if P has full-dimensional support. In the Appendix 9.2 we derive sufficient conditions for (30) to hold with lower dimensional level sets, based (29). They are verified if, for example, the boundary of L_h has bounded condition number (see Niyogi et al., 2008), uniformly in $h \in (0, \overline{h})$. Just like with Theorem 5.1, using Lemma 8.1, it can be verified that the theorem holds if one consider sequences of parameters depending on the sample size such that $\epsilon_n = o(1)$, $\epsilon_n = o(\rho_n)$, $\sup_n \rho_n < \max\{\delta/2, \overline{\rho}\}$ and $\sup_n h_n < \overline{h}$, provided that the conditions of Proposition 3.2 are met.

Despite the similar form for the error bounds of Theorems 5.1 and 5.2, there are some marked differences. In fact, in Theorem 5.1 the performance of the algorithm depends directly on the sample size n and, in particular, on the actual dimension $d^* \le d$ of the support of P, with smaller values of d^* yielding better guarantees. In contrast, besides n, the performance of the algorithm based on the bootstrap sample depends on the ambient dimension d, regardless of d^*, and on the bootstrap sample size N. By choosing N very large, the expression $\mathbb{P}(\mathcal{E}^c_{h,\epsilon})$ becomes the leading term in the upper bound of the probability of the event $(\mathcal{O}^*_{h,n})^c$.

6 Examples

In this section we consider a few examples to illustrate the methods.

6.1 A One Dimensional Example

In Section 4.2 we pointed out that when $\lambda > 0$ and large, it is safer to use the modified rule $\widehat{h} = \inf\{h : \ h \ge h_0, \ \ \Xi(h) \le \alpha\}$ where $h_0 = \text{argmax}_h \Xi(h)$, in place of the original rule $\widehat{h} = \inf\{h : \ \sup_{t>h} \Xi(t) \le \alpha\}$. We illustrate this with a simple one dimensional example.

Figure 3 shows an example based on $n = 200$ points from the density p that is uniform on $[0, 1] \cup [5, 6]$. When $\lambda = 0$ (top) the original rule works fine. (We use $\alpha = 0.05$.) The selected bandwidth is small leading to the very wiggly density estimator in the top right plot. However, this estimator correctly estimates the level set and the clusters. In the bottom we have $\lambda = .3$. When h is large, there is a blip in the instability curve corresponding to the fact that the modes of $p_h(x)$ are close to λ. The original rule corresponds to the second vertical line in the bottom left plot. The resulting density estimator shown in the bottom right plot is oversmoothed and leads to no points being in the set $\widehat{p}_h > \lambda$. The modified rule corresponds to the first vertical line in the bottom left plot. This bandwidth works fine.

Figure 4 compares the instability method (top) with the excess mass method (bottom). Both methods recover the level set and the clusters. We took $\lambda = .3$ in both cases. Because λ is very large, the excess mass becomes undefined for large h since $p_h(x) < \lambda$ for all x, which we denoted by setting the risk to 0 in the bottom left plot.

6.2 Fuzzy Stick With Spiral

Figure 5 shows data from a fuzzy stick with a spiral. The stick has noise while the spiral is supported on a lower dimensional curve. Figure 6 shows the clusterings from the instability method and the excess risk

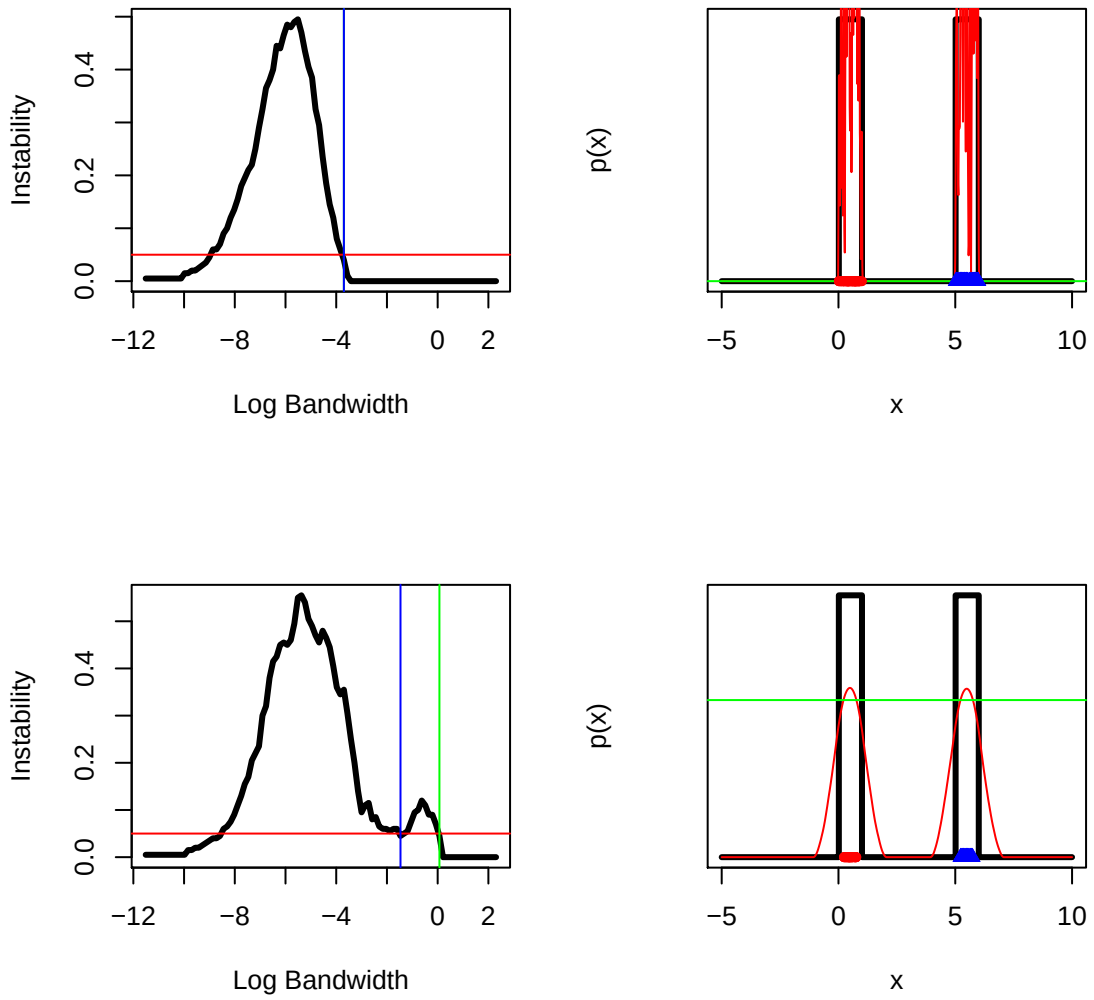

Figure 3: The left plots show the instability as a function of log bandwidth. The horizontal line shows $\alpha = 0.05$. The right plots show the true density and the kernel density estimator based on the selected bandwidth h. In the top plots, $\lambda = 0$. In the bottom plots, $\lambda = .3$.

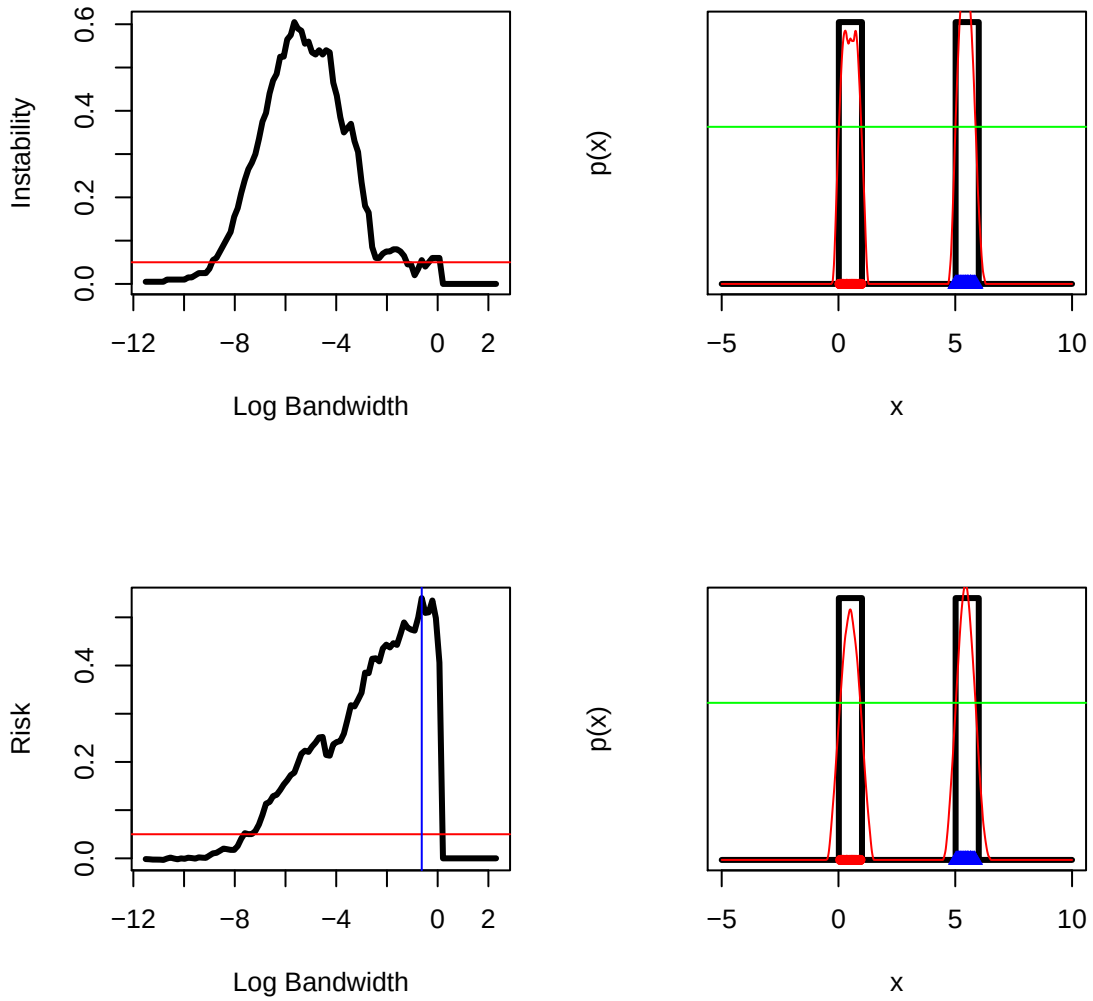

Figure 4: The top left plot shows the instability as a function of log bandwidth. The top right plot shows the true density and the kernel density estimator based on the selected bandwidth h using the modified rule. The bottom left plot shows the estimated excess mass risk as a function of log bandwidth. The top right plot shows the true density and the kernel density estimator based on the selected bandwidth h obtained by maximizing the excess mass. In both bottom plots, $\lambda = .3$. Both methods recover the level set and the clusters.

method with $\lambda = 0$. Both recover the clusters perfectly. Note that the excess risk is necessarily equal to 1 for large h. In this case we take $\widehat{h}$ to be the smallest h of all bandwidths that maximize the excess mass. We see that both methods recover the clusters.

6.3 Two Moons

This is a 20 dimensional example. The data lie on two half-moons embedded in $\mathbb{R}^{20}$. The results are shown in Figure 7. Only the first two coordinates of the data are plotted. Again we see that both methods recover the clusters.

7 Discussion

As is common in density clustering, we have assumed a fixed, given value of λ. In practice, we recommend that the results should be computed for a range of values of λ (see, e.g., Stuetzle and Nugent, 2009, and references therein). It is important to choose a different bandwidth for each λ. Indeed, inspection of the proof of Theorem 3.5 shows that the optimal bandwidth, as a function of λ, has the form,

$$h(\lambda) \asymp \left(\frac{\log n}{n\lambda^{\frac{2}{\gamma+1}}} \right)^{\frac{\gamma+1}{(d(\gamma+3))}}$$

for λ large enough. Hence, $h(\lambda) \to 0$ as λ increases. Further research on data-dependent methods to choose λ and ρ (the parameter in the friends-of-friends algorithm) would be very useful.

We discussed the idea of using stability to choose a bandwidth. We saw that the behavior of the selected bandwidth is quite different than with the excess mass method. This method seems to work well for density clustering unlike what happens for k-means clustering (Ben-David et al., 2006). We believe that the stability method deserves more scrutiny. In particular, it would be helpful to understand the behavior of the stability measure under more general conditions. Also, the detailed theoretical properties of the modified method for selecting h based on stability should be explored.

Finally, we note that there is growing interest in spectral clustering methods (von Luxburg (2007)). We believe there are connections between the work reported here and spectral methods. We will report on this connection in the future.

8 Proofs

Proof of Lemma 2.1. The weak convergence follows from the fact that P is a Radon measure (see, e.g., Leoni and Fonseca, 2007, Theorem 2.79). As for the second part, if $x \in S_i$, where S_i has Hausdorff dimension d, then $p(x) = \pi_i p_i(x)$, with p_i a Lebesgue density, and the result follows directly from Leoni and Fonseca (2007, Theorem 2.73, part *(ii)*). On the other hand if $d_i < d$, then it is necessary to modify the arguments as follows. Since K is smooth and supported on $B(0,1)$, there exists a η such that $K\left(\frac{x-y}{h}\right) > \eta$, for each h, if $\|x-y\| < \eta h$. Set $C = \frac{\eta^{d_i+1} c_{d_i}}{c_d}$. Then,

$$\begin{aligned} p_h(x) &= \frac{1}{c_d h^d} \int_{S_i \cap B(x,h)} K\left(\frac{x-y}{h}\right) dP(y) \\ &\geq \frac{1}{c_d h^d} \eta \int_{S_i \cap B(x,\eta h)} dP(y) \\ &= \frac{\eta^{d_i+1} c_{d_i}}{c_d h^{d-d_i}} \frac{1}{c_{d_i} (\eta h)^{d_i}} P_i(B(x,\eta,h)) \\ &= \frac{C}{h^{(d-d_i)}} \frac{P_i(B(x,\eta,h))}{\mu(B(x,\eta,h))}. \end{aligned}$$

By (38), $\frac{P_i(B(x,\eta,h))}{\mu(B(x,\eta,h))} \to p_i(x) < \infty$, almost everywhere $\mathcal{H}^{d_i}$, while $\frac{C}{h^{(d-d_i)}} \to \infty$ as $h \to 0$, thus showing that $\lim_{h\to 0} p_h(x) = \infty$. ∎

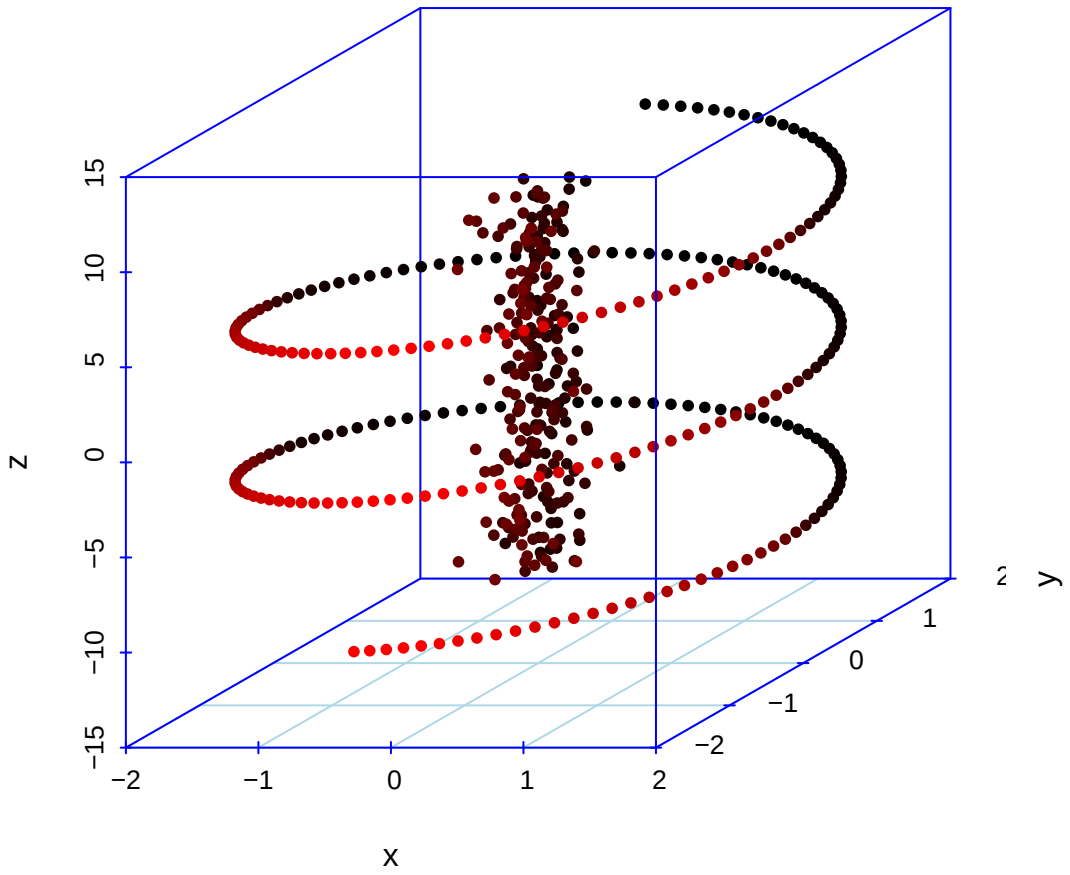

Figure 5: 500 data points from a fuzzy stick plus a spiral.

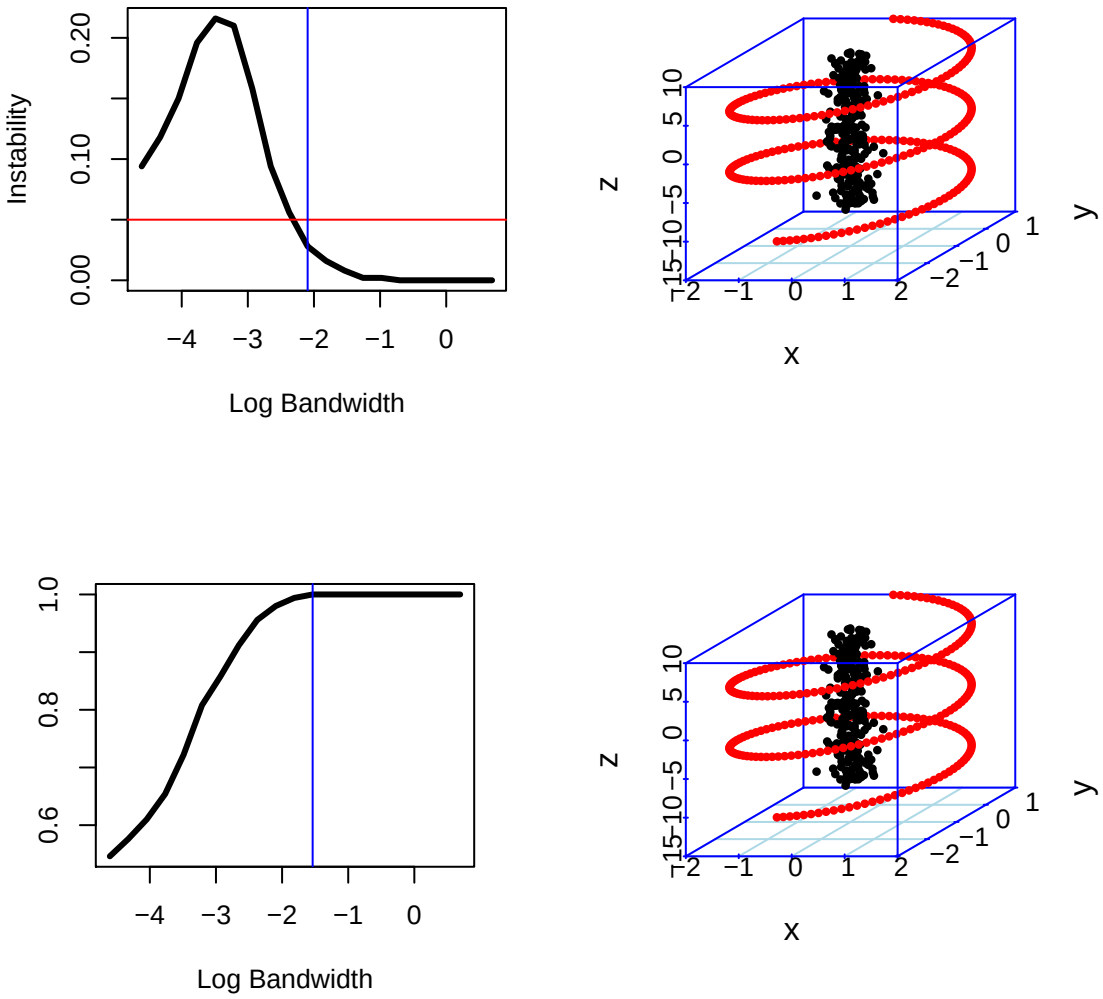

Figure 6: Clusters obtained from instability (top) and excess mass (bottom).

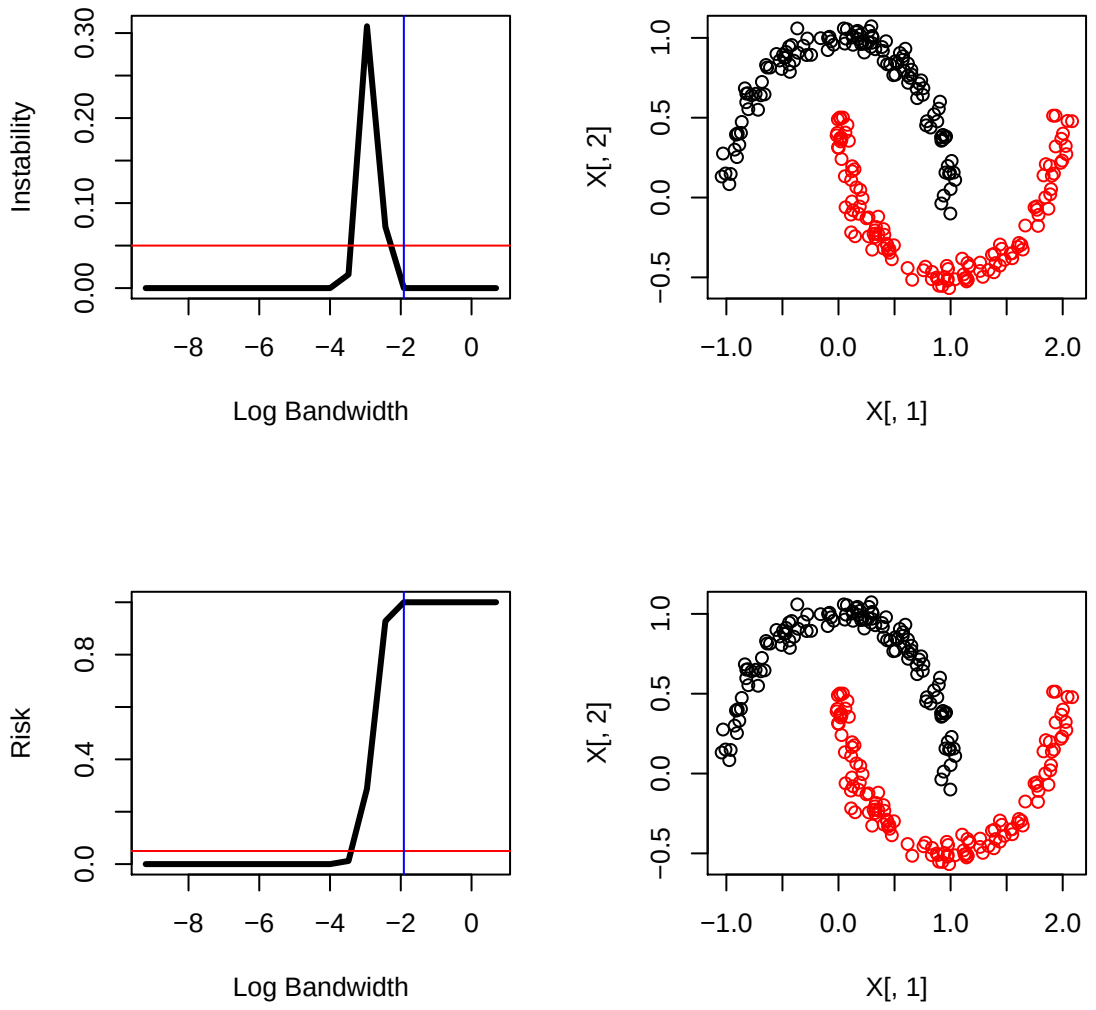

Figure 7: Clusters obtained from instability (top) and excess mass (bottom). The data are in $\mathbb{R}^{20}$ but only the first two components are plotted.

Proof of Lemma 2.2. By assumptions (LN2) (a) and (b), for any $0 \leq \epsilon < \bar{\epsilon}$ and $0 < h < \overline{h}$,

$$N_h(\lambda - \epsilon) = N_h(\lambda) = N_h(\lambda + \epsilon) = N(\lambda) = k.$$

On the event $\mathcal{E}_{h,\epsilon}$ it holds that

$$L_h(\lambda + \|p_h - \widehat{p}_h\|_\infty) \subseteq \widehat{L}_h(\lambda) \subseteq L_h(\lambda - \|p_h - \widehat{p}_h\|_\infty),$$

which implies that, on the same event,

$$k = N_h(\lambda + \|p_h - \widehat{p}_h\|_\infty) \leq \widehat{N}_h(\lambda) \leq N_h(\lambda - \|p_h - \widehat{p}_h\|_\infty) = k.$$

■

Proof of Lemma 2.3. Since p is Lipschitz and integrable, $p^{-1}(\lambda)$ is $\mathcal{H}^{d-1}$-measurable, so the integral $\mathcal{H}^{d-1}(\{x\colon p(x) = \lambda\})$ is well defined for $\lambda \in (0, \|p\|_\infty)$, where $\mathcal{H}^{d-1}$ denote the $(d-1)$-dimensional Hausdorff measure in $\mathbb{R}^d$. Furthermore, we can use the coarea formula. See Evans and Gariepy (1992) and Ambrosio et al. (2000) for backgrounds on Hausdorff measures and the coarea formula. By the Rademacher Theorem, the set E_1 of points where p is not differentiable has Lebesgue measure zero. By Lemma 2.96 in Ambrosio et al. (2000), the set $E_2 = \{x\colon \|\nabla p(x)\| = 0\}$ is such that $\mathcal{H}^{d-1}\{p^{-1}(\lambda) \cap E_2\} = 0$, for all $\lambda \in (0, \|p\|_\infty)$ outside of a set $E_3 \subset \mathbb{R}$ of Lebesgue measure 0. Without loss of generality, below we may assume that E_1 and E_2 are empty. Thus, we can assume that, for any $\lambda \in (0, \|p\|_\infty) \cap E_3^c$, there exists positive numbers $\bar{\epsilon}$, C and M such that

(i) $\inf_{x \in \{x\colon |p(x)-\lambda| < \bar{\epsilon}\}} \|\nabla p(x)\| > C$, almost everywhere-$\mu$;

(ii) $\sup_{\eta \in (-\bar{\epsilon}, \bar{\epsilon})} \mathcal{H}^{d-1}(\{x\colon p(x) = \lambda + \eta\}) < M$.

Then, for each $\epsilon \in (0, \bar{\epsilon})$,

$$\begin{aligned}
P(\{x\colon |p(x) - \lambda| < \epsilon\}) &= \int p(x) 1_{\{|p(x)-\lambda|<\epsilon\}} d\mu(x) \\
&= \int \tfrac{p(x)}{\|\nabla p(x)\|} 1_{\{|p(x)-\lambda|<\epsilon\}} \|\nabla p(x)\| d\mu(x) \\
&= \int_{-\epsilon}^{+\epsilon} \int_{\{p^{-1}(\lambda+u)\}} \tfrac{p(x)}{\|\nabla p(x)\|} d\mathcal{H}^{n-1}(x) du \\
&= \int_{-\epsilon}^{+\epsilon} (\lambda + u) \int_{\{p^{-1}(\lambda+u)\}} (\|\nabla p(x)\|)^{-1} d\mathcal{H}^{n-1}(x) du \\
&\leq \tfrac{2\lambda M}{C} \epsilon,
\end{aligned}$$

where the second equality holds because $\|\nabla p(x)\|$ is bounded away from 0 on $\{x\colon |p(x) - \lambda| < \epsilon\}$ by (i), the third equality is a direct application of the coarea formula (see, e.g., Proposition 3 page 118 in Evans and Gariepy, 1992) and the last inequality follows from (i) and (ii). ■

Proof of Corollary 2.4. Following the proof of Lemma 2.3 and using our additional assumption that p is of class $\mathcal{C}^1$, without any loss of generality, below we can assume that the set E_1 and E_2 are empty and we recall that E_3 has Lebesgue measure 0. Let $\lambda \notin E_3$ be such that

$$\inf_{x \in p^{-1}(\lambda)} \|\nabla p(x)\| > 0.$$

We now claim that there exists a non-empty neighborhood U of λ for which

$$\inf_{\lambda \in U} \inf_{x \in p^{-1}(\lambda)} \|\nabla p(x)\| > 0.$$

Indeed, arguing by contradiction, suppose that the previous display were not verified for any neighborhood U of λ. Then, there exist sequences $\{\lambda_n\} \subset \mathbb{R}$ and $\{x_n\} \subset S$ such that $\lim_n \lambda_n = \lambda$, and $x_n \in p^{-1}(\lambda_n)$ and $\nabla p(x_n) = 0$ for each n. By compactness, it is possible to extract a subsequence $\{x_{n_k}\}$ of $\{x_n\}$ such that $x_{n_k} \to x$, for some $x \in p^{-1}(\lambda)$. Since p is of class $\mathcal{C}^1$, this implies that $\nabla p(x_{n_k}) \to \nabla p(x)$ as well. However, $\nabla p(x_{n_k}) = 0$ for each k by construction, while $\nabla p(x) \neq 0$. This produces a contradiction. Thus, for each λ that is not a critical point, one can find a neighborhood of positive length containing it and, by Lemma 2.3, LN1 holds at λ with $\gamma = 1$. Since, using compactness again, $\|p\|_\infty < \infty$, this implies that there can only be a finite number of critical points for which γ may differ from 1. ■

Proof of Lemma 3.1. Since $\epsilon < \bar{\epsilon}$ and $h < \overline{h}$, in virtue of (LN2) (b) it holds that, on $\mathcal{E}_{h,\epsilon}$,

$$\widehat{L}_h(\lambda) \supseteq L_h(\lambda + \epsilon) \supseteq L(\lambda + \epsilon),$$

and

$$\widehat{L}_h(\lambda) \subseteq L_h(\lambda - \epsilon) = L(\lambda - \epsilon) \cup (L_h(\lambda - \epsilon) - L(\lambda - \epsilon)).$$

Because $L(\lambda + \epsilon) \subseteq L(\lambda) \subseteq L(\lambda - \epsilon)$, the above inclusions imply, still on $\mathcal{E}_{h,\epsilon}$, that

$$\begin{aligned} \widehat{L}_h(\lambda) \Delta L(\lambda) &\subseteq (L(\lambda - \epsilon) - L(\lambda + \epsilon)) \cup (L_h(\lambda - \epsilon) - L(\lambda - \epsilon)) \\ &= A \cup B, \end{aligned}$$

where it is clear that the sets A and B are disjoint. Taking expectation with respect to P of the indicators of the sets $\widehat{L}_h(\lambda)\Delta L(\lambda)$, A and B and using conditions (LN1) and (LN2) (c) yield (12). ■

Proof of Proposition 3.2. The claimed results are a direct consequence of Corollary 2.2 in Giné and Guillou (2002). We outline the details below. We use the function class (13) to rewrite the left hand side of (14) as

$$\mathbb{P}\left\{ \left\| \sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X_1)] \right\|_{\mathcal{F}_h} > 2\epsilon n h^d \right\}$$

and then proceed to apply Giné and Guillou (2002, Corollary 2.2). Following their notation, we set $t = nh^d\epsilon$ and, since,

$$\sup_{f \in \mathcal{F}_h} \mathrm{Var}[f] \leq \sup_{z} \int_{\mathbb{R}^d} K^2\left(\frac{z - x}{h}\right) dP(x) \leq h^d D,$$

we can further take $\sigma^2 = h^d D$ and $U = C\|K\|_\infty$, where C is a positive constant, depending on h, such that $\sigma < U/2$. Then, conditions (2.4) (2.5) and (2.6) of Giné and Guillou (2002) are satisfied for all n bigger than some finite n_0, which depends on the VC characteristics of K, D, $\|K\|_\infty$, C and ϵ. Part 2 is proved in a very similar way. In this case, we set $\sigma_n^2 = h_n^d D$ and $U = \|K\|_\infty$. For all n large enough, condition (2.5) is trivially satisfied because $h_n = o(1)$, while equations (2.4) and (2.6) hold true in virtue of (16). The unspecified constants now depend on the specific sequences $\{h_n\}$ and $\{\epsilon_n\}$, as well as on the VC characteristics of K, D, and $\|K\|_\infty$. ■

Proof of Theorem 3.3. We can write

$$\begin{aligned} \mathbb{E}(\rho^\dagger(p, \widehat{p}_h, P)) &= \mathbb{E}\left(\textstyle\int_{\widehat{L}_h(\lambda)\Delta L(\lambda)} dP; \mathcal{E}_{h,\epsilon}\right) + \mathbb{E}\left(\textstyle\int_{\widehat{L}_h(\lambda)\Delta L(\lambda)} dP; \mathcal{E}_{h,\epsilon}^c\right) \\ &\leq C_1\epsilon^\gamma + C_2 h^\xi + \mathbb{E}\left(\textstyle\int_{\widehat{L}_h(\lambda)\Delta L(\lambda)} dP; \mathcal{E}_{h,\epsilon}^c\right) \\ &\leq C_1\epsilon^\gamma + C_2 h^\xi + \mathbb{P}\left(\mathcal{E}_{h,\epsilon}^c\right) \\ &\leq C_1\epsilon^\gamma + C_2 h^\xi + Le^{-nC_K h^d \epsilon^2}, \end{aligned}$$

where the first inequality stems from (12), the second from the fact that $\int_{\widehat{L}_h(\lambda)\Delta L(\lambda)} dP \leq 1$ with probability one and the third from Proposition 3.2. The claimed rates are established using simple algebra. Notice that the choice of the sequence $\{\epsilon_n\}$ does not violate condition (16). ■

Proof of Corollary 3.4. For each $i \in \{1, \ldots, n\}$,

$$\mathbb{P}\left(i \in \widehat{I}_h | \mathcal{E}_h\right) \leq \mathbb{P}\left(X_i \in \widehat{L}_h \Delta L | \mathcal{E}_{h,\epsilon}\right) \leq C(\epsilon^\gamma + h^\xi) \frac{1}{\mathbb{P}(\mathcal{E}_{h,\epsilon})}$$

where $C = C_1 + C_2$ and the last inequality is due to Lemma 3.1. Thus,

$$\begin{aligned} \mathbb{E}(|\widehat{I}_h|) &\leq \textstyle\sum_{i=1}^n \mathbb{P}\left(i \in \widehat{I} | \mathcal{E}_{h,\epsilon}\right) \mathbb{P}(\mathcal{E}_{h,\epsilon}) + n\mathbb{P}(\mathcal{E}_h^c) \\ &\leq n\left(C(\epsilon^\gamma + h^\xi) + \mathbb{P}(\mathcal{E}_h^c)\right). \end{aligned}$$

■

Proof of Theorem 3.5. On the event $\mathcal{E}_{h,\epsilon}$, we obtain, in virtue of Lemma 3.1,

$$\int_{\{\widehat{L}_h(\lambda)\Delta L(\lambda)\}} |p-\lambda| d\mu \leq \int_{L(\lambda-\epsilon)-L(\lambda+\epsilon)} |p-\lambda| d\mu + \int_{L_h(\lambda-\epsilon)-L(\lambda-\epsilon)} |p-\lambda| d\mu. \tag{31}$$

The first term on the right hand side of (31) can be bounded as follows.

$$\begin{array}{rcl} \int_{L(\lambda-\epsilon)-L(\lambda+\epsilon)} |p-\lambda| d\mu(x) & = & \int_{\{x\colon |p(x)-\lambda|<\epsilon\}} |p-\lambda| d\mu(x) \\ & \leq & \epsilon \int_{\{x\colon |p(x)-\lambda|<\epsilon\}} d\mu(x) \\ & = & \frac{\epsilon}{\lambda-\epsilon} \int_{\{x\colon |p(x)-\lambda|<\epsilon\}} (\lambda-\epsilon) d\mu \\ & \leq & \frac{\epsilon}{\lambda-\epsilon} \int_{\{x\colon |p(x)-\lambda|<\epsilon\}} p(x) d\mu(x) \\ & \leq & \frac{C_1}{\lambda-\epsilon} \epsilon^{\gamma+1}. \end{array}$$

where in the last inequality is due to condition LN1. As for the second term of the right hand side of (31), we obtain

$$\int_{L_h(\lambda-\epsilon)-L(\lambda-\epsilon)} |p-\lambda| d\mu \leq \lambda \int_{L_h(\lambda-\epsilon)-L(\lambda-\epsilon)} d\mu,$$

since, for $x \in L_h(\lambda-\epsilon)-L(\lambda-\epsilon)$, $p(x) < \lambda-\epsilon$ We now claim that $L_h(\lambda-\epsilon)-L(\lambda-\epsilon) \subseteq L(\lambda-\epsilon)+B(0,h)$. In fact, if $w \notin L(\lambda-\epsilon)+B(0,h)$, then either $p(w) > \lambda-\epsilon$ or, for every $z \in B(w,h)$, $p(z) < \lambda-\epsilon$. Since the kernel K has compact support, the former case implies that $p_h(w) < \lambda-\epsilon$ as well. Therefore,

$$\begin{array}{rcl} w & \in & \{x\colon p(x) > \lambda-\epsilon\} \cup \{x\colon p_h(x) < \lambda-\epsilon\} \\ & = & \{x\colon p(x) < \lambda-\epsilon, p_h(x) > \lambda-\epsilon\}^c \\ & = & (L_h(\lambda-\epsilon) - L(\lambda-\epsilon))^c. \end{array}$$

Thus, invoking (LN2) (d),

$$\int_{L_h(\lambda-\epsilon)-L(\lambda-\epsilon)} |p-\lambda| d\mu \leq \lambda\mu(L(\lambda-\epsilon)+B(0,h)) = O(h^d).$$

We conclude that

$$\mathbb{E}\left(\int_{\{\widehat{L}_h(\lambda)\Delta L(\lambda)\}} |p-\lambda| d\mu; \mathcal{E}_{h,\epsilon}\right) = O\left(\epsilon^{\gamma+1} + h^d\right).$$

Next, by compactness of S,

$$\mathbb{E}\left(\int_{\{\widehat{L}_h(\lambda)\Delta L(\lambda)\}} |p-\lambda| d\mu; \mathcal{E}_h^c\right) \leq 1 + \lambda\mu(S+B(0,\overline{h})) \leq (1+\lambda)C_S,$$

for some positive constant C_S, uniformly in $h < \overline{h}$. Thus, recalling that $\mathcal{E}(L) - \mathbb{E}(\mathcal{E}(\widehat{L})) = \mathbb{E}\left(\int_{\{\widehat{L}_h(\lambda)\Delta L(\lambda)\}} |p-\lambda| d\mu\right)$, the result follows. ∎

Proof of Theorem 3.6. By Lemma 2.2, on the event $\mathcal{E}_{h,\epsilon}$, $\widehat{C}_h(\lambda) = K$. For each $i = 1\ldots, K$, denote with $C_i(\lambda)$, $C_i^h(\lambda)$ and $\widehat{C}_i(\lambda)$ the λ-clusters of p, p_{h_n} and $\widehat{p}_{h_n}$, respectively, ordered in such a way that $C_i(\lambda) \subseteq C_i^h(\lambda)$ for all i. Assumptions (LN2) and the arguments used in the proof of Lemma 2.2 show that, for all i, $C_i(\lambda) \subseteq C_i^h(\lambda-\epsilon)$ and $\widehat{C}_{\sigma(i)}(\lambda) \subseteq C_i^h(\lambda-\epsilon)$, for some permutation σ of $\{1,\ldots,K\}$. Since, by assumption (LN2), $C_i^h(\lambda-\epsilon) \cap C_j^h(\lambda-\epsilon) = \emptyset$ for all $i \neq j$, we conclude that

$$\left(C_i(\lambda) \cup \widehat{C}_{\sigma(i)}(\lambda)\right) \bigcap \left(\bigcup_{j\neq i} (C_j(\lambda) \cup \widehat{C}_{\sigma(j)}(\lambda))\right) = \emptyset, \tag{32}$$

for every $i = 1, \ldots, K$. Next, set $A_{\epsilon,h} = \{x \in \mathbb{R}^d \colon \|p_h - \widehat{p}_h\|_\infty < \epsilon\}$ and

$$B_{h,\epsilon} = \{z \in \mathbb{R} \colon z \notin (L(\lambda - \epsilon) - L(\lambda + \epsilon)) \cup (L_h(\lambda - \epsilon) - L(\lambda - \epsilon))\}.$$

We claim that, for any $(x, z_1, z_2) \in A_{\epsilon,h} \times B_{h,\epsilon} \times B_{h,\epsilon}$, $M_{\widehat{p}_h}(z_1, z_2) = M_p(z_1, z_2)$. First observe that equation (32) implies that, for any $x \in A_{\epsilon,h}$ and for each $z_1, z_2 \in L$, the set of all pairs (z_1, z_2) such that $M_{\widehat{p}_h}(z_1, z_2) = 1$ and $M_p(z_1, z_2) = 0$ or $M_{\widehat{p}_h}(z_1, z_2) = 0$ and $M_p(z_1, z_2) = 1$ is empty. Thus, it is sufficient to show that, for each such triple (x, z_1, z_2), it is not possible for $M_{\widehat{p}_h}(z_1, z_2), M_p(z_1, z_2)$ to be equal to any of the remaining four patterns $(0, *)$, $(1, *)$, $(*, 0)$, $(*, 1)$. Indeed, for any $x \in A_{\epsilon,h}$, should any such pattern occurs, then because of (32), z_1 or z_2 (possibly both) would have to belong to $L(\lambda)\Delta\widehat{L}_h(L)$, which by Lemma 3.1 is a subset of $(L(\lambda - \epsilon) - L(\lambda + \epsilon)) \cup (L_h(\lambda - \epsilon) - L(\lambda - \epsilon)) = B^c_{\epsilon,h}$. This gives a contradiction, so the claim is proved true. By independence, the probability that $(x, z_1, z_2) \notin A_{\epsilon,h} \times B_{h,\epsilon} \times B_{h,\epsilon}$ is bounded by

$$\mathbb{P}\left(\mathcal{E}^c_{h,\epsilon}\right) + 2\left(\int_{L(\lambda-\epsilon)-L(\lambda+\epsilon)} dP + \int_{L_h(\lambda-\epsilon)-L(\lambda-\epsilon)} dP\right) = O(\epsilon^\gamma + h^\xi + e^{-Cnh^d\epsilon^2}),$$

with the last identity stemming from Proposition 3.2 and assumptions (LN1) and (LN2) (a) - (c). ■

Proof of Theorem 4.1. This follows by combining the version of Talagrand's inequality for empirical processes as given in Massart (2007) with a straightforward adaptation of the proof of Theorem 7.1 of Györfi et al. (2002). For completeness, we provide the details.

Define $\widehat{h} = \text{argsup}_{h \in \mathcal{H}} \widehat{\mathcal{E}}(L_h)$, where

$$\widehat{\mathcal{E}}(L_h) = \frac{1}{n}\sum_{i=1}^{n} I(Z_i \in L_h) - \lambda\mu(L_h).$$

and $h_* = \text{argsup}_{h \in \mathcal{H}} \mathcal{E}_X(L_h)$. Set $\Gamma(h) = \mathcal{E}_X(L_{h_*}) - \mathcal{E}_X(L_h)$, where $h \in \mathcal{H}$. Recall that both L_{h^*} and $L_h = \{x \colon \widehat{p}_h > \lambda\}$, are random sets depending on the training set X. We will bound $\mathbb{E}(\Gamma(\widehat{h}))$, where the expectation is over the joint distribution of X and Y.

We can write

$$\mathbb{E}(\Gamma(\widehat{h})|X) = \underbrace{\mathbb{E}(\Gamma(\widehat{h})|X) - (1+\delta)\widehat{\Gamma}(\widehat{h})}_{T_1} + \underbrace{(1+\delta)\widehat{\Gamma}(\widehat{h})}_{T_2}$$

where $\widehat{\Gamma}(h) = \widehat{\mathcal{E}}(L_{h_*}) - \widehat{\mathcal{E}}(L_h)$. Note that

$$\widehat{\Gamma}(\widehat{h}) = \widehat{\mathcal{E}}(L_{\widehat{h}}) - \widehat{\mathcal{E}}(L_{h_*}) \leq \widehat{\mathcal{E}}(L_{h_*}) - \widehat{\mathcal{E}}(L_{h_*}) = 0.$$

Thus, $\mathbb{E}(T_2|X) \leq 0$. We conclude that

$$\mathbb{E}(\Gamma(\widehat{h})) = \mathbb{E}(\mathbb{E}(\Gamma(\widehat{h})|X)) = \mathbb{E}(\mathbb{E}(T_1|X)) + \mathbb{E}(\mathbb{E}(T_2|X)) \leq \mathbb{E}(\mathbb{E}(T_1|X)). \tag{33}$$

Now we bound $\mathbb{E}(T_1|X)$. Consider the empirical process

$$Z = \sup_{h \in \mathcal{H}} \widehat{\Gamma}(h),$$

so that $Z = \widehat{\Gamma}(\widehat{h})$ and $\mathbb{E}(\Gamma(\widehat{h})|X) = \mathbb{E}(Z|X)$. We have

$$\begin{aligned}
\mathbb{P}(T_1 \geq s|X) &= \mathbb{P}\left(\mathbb{E}(Z|X) - (1+\delta)Z) \geq s \,\middle|\, X\right) \\
&= \mathbb{P}\left(\mathbb{E}(Z|X) - Z \geq \frac{s + \delta\mathbb{E}(Z|X)}{1+\delta} \,\middle|\, X\right).
\end{aligned}$$

Notice that, conditionally on X, $Z = \frac{1}{n}\sup_{h\in\mathcal{H}}\sum_{i=1}^n f_h(Y_i)$, where, for each $h \in \mathcal{H}$, $f_h\colon \mathbb{R}^d \mapsto \mathbb{R}$ is the function given by

$$f_h(x) = I(x \in L_{h_*}) - \lambda\mu(L_{h_*}) - (I(x \in L_h) - \lambda\mu(L_h))\,.$$

with $\|f_h\|_\infty < \kappa$. Let $\sigma^2 \equiv \mathbb{E}(\frac{1}{n}\sup_{h\in\mathcal{H}}\sum_{i=1}^n f_h^2(Y_i)|X)$ and notice that $\sigma^2 \le \kappa\mathbb{E}(\sup_h \widehat{\Gamma}(h)|X) = \kappa\mathbb{E}(Z|X)$. Thus,

$$\mathbb{P}(T_1 \ge s|X) \le \mathbb{P}\left(\mathbb{E}(Z|X) - Z \ge \frac{s + \delta\sigma^2/\kappa}{1+\delta}\;\middle|\; X\right),$$

which, by Corollary 13 in Massart (2007), is upper bounded by

$$2\exp\left\{-\frac{n\left(\frac{s+\delta\sigma^2/\kappa}{1+\delta}\right)^2}{4(4\gamma^2\sigma^2 + \frac{7}{4}\kappa\epsilon)}\right\}.$$

Then, some algebra (see Problem 7.1 in Györfi et al., 2002) yields the final bound

$$\mathbb{P}(T_1 \ge s|X) \le 2\exp\left\{\frac{-ns}{d(\delta,\kappa)}\right\},$$

where $d(\delta,\kappa)$ is given the in the statement of the theorem.

Set $u = \frac{d(\delta,\kappa)}{n}\log 2$. Then,

$$\begin{aligned}
\mathbb{E}(T_1|X) &= \int_0^\infty \mathbb{P}(T_1 > s|X)ds \le u + \int_u^\infty \mathbb{P}(T_1 > s|X)ds \\
&= u + \frac{2d(\delta,\kappa)}{n}\exp\left\{-\frac{nu}{d(\delta,\kappa)}\right\} \\
&= d(\delta,\kappa)\frac{1+\log 2}{n}.
\end{aligned}$$

From (33) we conclude that

$$\mathbb{E}(\Gamma(\widehat{h})) \le d(\delta,\kappa)\frac{1+\log 2}{n}.$$

and so

$$\mathbb{E}(M(\widehat{h})) \le \mathbb{E}(M(h_*)) + d(\delta,\kappa)\frac{1+\log 2}{n}$$

which implies that

$$\mathbb{E}(\mathcal{E}(\widehat{h})) \ge \mathbb{E}(\mathcal{E}(h_*)) - d(\delta,\kappa)\frac{1+\log 2}{n}.$$

This shows (22).

Proof of Theorem 4.2. Let γ be as defined in (LN1). Define $r_n(s) = \left(\frac{\log n}{n}\right)^{\frac{s+1}{s+3}}$. Let $\Upsilon_n = 2(L_n - \log 2)/\log 2$ where $L_n = \log n - \log\log n$. Note that $r_n(s)$ is decreasing in s and that

$$r_n(\Upsilon_n) \le 2r_n(\infty).$$

Hence, $\inf_{s\in[0,\Upsilon_n]} r(s) \le 2\inf_{s\ge 0} r(s)$. Now $|r'(s)| \le 2L_n r(0)/9$ for all s. Therefore, for each j, $r(\gamma_j) = r(j\delta_n + \delta_n) \ge r(j\delta_n) - \delta_n\sup_\gamma |r'(\gamma)| \ge r(\gamma_j) - 2\delta_n L_n r(0)/9 \ge r(\gamma_j)/2$. Let $h_n \equiv (\log n/n)^{(\gamma+1)/(d(3+\gamma))}$. By Theorem 3.5, $R^M(p,\widehat{p}_{h_n}) = O((\log n/n)^{(\gamma+1)/(\gamma+3)})$. Let $h_* \in \mathcal{H}_n$ minimize $R^M(p,\widehat{p}_h)$ for $h \in \mathcal{H}_n$. Thus,

$R^M(p, \widehat{p}_{h_*}) \leq 2R^M(p, \widehat{p}_{h_n})$. So,

$$\begin{aligned} R^M(p, \widehat{p}_{\widehat{h}}) &\leq d(\delta, \kappa)\frac{1+\log 2}{n} + R^M(p, \widehat{p}_{h_*}) \\ &\leq d(\delta, \kappa)\frac{1+\log 2}{n} + 2R^M(p, \widehat{p}_{h_n}) \\ &= d(\delta, \kappa)\frac{1+\log 2}{n} + 2r(\gamma) \\ &= O\left(\frac{\log n}{n}\right)^{\frac{\gamma+1}{\gamma+3}}. \end{aligned}$$

■

Proof of Theorem 4.3. (1) When $h = 0$, $\{\widehat{p}_h > \lambda\} = X$ and $\{\widehat{q}_h > \lambda\} = Y$ so that $\{\widehat{p}_h > \lambda\} \Delta \{\widehat{q}_h > \lambda\} = (X, Y)$. Since P has a Lebesgue density, with probability one, $d\widehat{P}_Z$ puts no mass on (X, Y) and, therefore, $\Xi(0) = 0$. By compactness of S, if $h \geq \text{diam}(S)$, then $\|\widehat{p}_h\|_\infty = \|\widehat{q}_h\|_\infty = \frac{1}{h^d v_d}$, with the supremum attained by any $z \in S$. Thus, as $h \to \infty$, $\|\widehat{p}_h - \widehat{q}_h\|_\infty \to 0$ and consequently, $\Xi(\infty) \to 0$.

(2) Note that

$$\begin{aligned} \Xi(h) &= \rho^\dagger(\widehat{p}_h, \widehat{q}_h, \widehat{P}_Z) = \int_{\{\widehat{p}_h > \lambda\} \Delta \{\widehat{q}_h > \lambda\}} d\widehat{P}_Z(z) \\ &= \int I(\widehat{p}_h(z) > \lambda, \widehat{q}_h(z) \leq \lambda) d\widehat{P}_Z(z) + \int I(\widehat{p}_h(z) \leq \lambda, \widehat{q}_h(z) > \lambda) d\widehat{P}_Z(z). \end{aligned}$$

Define $\xi(h) = \mathbb{E}(\Xi(h)|X, Y)$. Then,

$$\begin{aligned} \xi(h) &= \rho^\dagger(\widehat{p}_h, \widehat{q}_h, P) \\ &= \int I(\widehat{p}_h(z) > \lambda, \widehat{q}_h(z) \leq \lambda) dP(z) + \int I(\widehat{p}_h(z) \leq \lambda, \widehat{q}_h(z) > \lambda) dP(z) \\ &\stackrel{d}{=} 2\int I(\widehat{p}_h(z) > \lambda, \widehat{q}_h(z) \leq \lambda) dP(z), \end{aligned}$$

where $\stackrel{d}{=}$ denotes identity in distribution. Let $\pi_h(z) = \mathbb{P}(\widehat{p}_h(z) \leq \lambda) = \mathbb{P}(\widehat{q}_h(z) \leq \lambda)$. By Fubini's theorem and independence,

$$\begin{aligned} \mathbb{E}(\Xi(h)) &= \mathbb{E}(\xi(h)) \\ &= 2\int_{\mathbb{R}^d} \mathbb{P}(\widehat{p}_h(z) > \lambda, \widehat{q}_h(z) \leq \lambda) dP(z) \\ &= 2\int_{\mathbb{R}^d} \mathbb{P}(\widehat{p}_h(z) > \lambda)\mathbb{P}(\widehat{q}_h(z) \leq \lambda) dP(z) \\ &= 2\int_{\mathbb{R}^d} \pi_h(z)(1 - \pi_h(z)) dP(z). \end{aligned} \tag{34}$$

Since $\pi_h(z)(1 - \pi_h(z)) \leq 1/4$ for all n, h and z, (2) follows.

(3) Let $W = (X, Y)$ be the $2n$-dimensional vector obtained by concatenating X and Y and define the event

$$\mathcal{A}_h = \{B(W_i, h) \cap B(W_j, h) = \emptyset, \forall i \neq j\}.$$

Let h be small enough such that $\lambda n h^d v_d < 1$ (trivially satisfied if $\lambda = 0$). Then, for any realization w of the vector W for which the event $\mathcal{A}_h$ occurs,

$$\int I(\widehat{p}_h(z) > \lambda, \widehat{q}_h(z) \leq \lambda) dP(z) = \sum_{i=1}^{2n} P(B(w_i, h)).$$

By our assumptions,

$$2n\delta h^d v_d \leq \sum_{i=1}^{2n} P(B(w_i,h)) \leq 2n\overline{\Delta} h^d v_d.$$

Using the union bound, we also have

$$\mathbb{P}\left(\mathcal{A}_h^c\right) \leq \binom{2n}{2}(2h)^d v_d \overline{\Delta}.$$

Thus it follows that, for fixed n, $\mathbb{E}\left(\xi(h)\right) \to 0$ as $h \to 0$ according to

$$2n\delta h^d v_d \leq \mathbb{E}\left(\xi(h)\right) \leq h^d v_d 2\overline{\Delta} \max\left\{2^d n(2n-1), 2n\right\}.$$

(4) By the same arguments used in the proof of point (1), for all $h \geq h_*$, $\xi(h) = 0$ almost everywhere with respect to the joint distribution of X and Y, and, therefore, $\mathbb{E}(\xi(h)) = 0$. Thus, we need only to consider the case $0 < h \leq h_*$.

Set $p_{z,h} = P(B(z,h))$ and denote with $X_{z,h}$ a random variable with distribution $\text{Bin}(n, p_{z,h})$. Then,

$$\mathbb{P}(\widehat{p}_h(z) = 0) = \mathbb{P}(X_{z,h} = 0) = (1 - p_{z,h})^n.$$

For each $z \in S$, set $D(z,h) = \{z' \in S\colon \|z - z'\| < h\}$ and $S_h = \{z\colon D(z,h) \neq S\}$. Furthermore, set $p_{h,\max} = \sup_{z \in S_h}\{p_{z,h}\}$ and $p_{h,\min} = \inf_{z \in S_h}\{p_{z,h}\}$. Then, the expected instability cam be written as

$$\mathbb{E}\left(\Xi(h)\right) = 2\int_{S_h} \pi_h(z)(1 - \pi_h(z))dP(z)$$

so that $A_{h,n} \leq \mathbb{E}\left(\Xi(h)\right) \leq B_{h,n}$, where

$$\begin{aligned} A_{h,n} &\equiv 2P(S_h)(1 - p_{h,\max})^n \left(1 - (1 - p_{h,\min})^n\right), \\ B_{h,n} &\equiv 2P(S_h)(1 - p_{h,\min})^n \left(1 - (1 - p_{h,\max})^n\right). \end{aligned}$$

We will now upper bound $B_{h,n}/2$. For the first term we proceed as follows. There exists a sphere $E = B(z_0, h_*/2)$ such that $S \subset E$. (For example, choose any two points z, z' such that $||z - z'|| = h^*$. Set $z_0 = (z + z')/2$.) Let $A = B(z_0, h_*/2) - B(z_0, (h_* - h)/2)$. We claim that $S_h \subset A$. This follows since if $z \in A^c \cap S$ then $z \in B(z_0, h/2)$ and then $\sup_{z' \in S} ||z - z'|| =\leq \sup_{z \in B(z_0, h/2), z' \in B(z_0, h_*/2)} ||z - z'|| = h$. Thus if $z \in S_h$ then $z \in A \cap S \subset A$. Hence

$$P(S_h) \leq P(A) \leq \overline{\Delta}\mu(A) = \overline{\Delta}\frac{((h_*/2)^d - (h/2)^d)\pi^{d/2}}{\Gamma((d/2)+1)} \leq C_1(h_* - h)$$

where

$$C_1 = \frac{\pi^{d/2} h_*^{d-1}}{2^d \Gamma((d/2)+1)}.$$

For the second term, let $z_0 = \text{argmin}_z p_{z,h}$. Then,

$$\begin{aligned} 1 - p_{h,\min} &= 1 - P(B(z_0,h)) = P(B(z,h_*)) - P(B(z_0,h)) \\ &= P(B(z,h_*) - B(z_0,h)) \leq \overline{\Delta}\mu(B(z,h_*) - B(z_0,h)) \\ &\leq \overline{\Delta}\frac{(h_*^d - h^d)\pi^{d/2}}{\Gamma((d/2)+1)} = C_2(h_* - h) \end{aligned}$$

where $C_2 = \frac{\pi^{d/2} h_*^{d-1}}{\Gamma((d/2)+1)}$. The third term is bounded above by 1. Hence, $B_n \leq C_1 C_2^n (h_* - h)^{n+1}$.

Now we lower bound $A_{h,n}/2$. First we claim that S_h contains the intersection of a sphere of radius $r/2$ where $r = h_* - h$, with S. Indeed, let $z \in S_h$. Then there exists $z' \in S$ such that $||z - z'|| \leq h_* = h + r$. Let $w \in B(z', r/2)$. By the triangle inequality, $||w - z|| \geq h + r/2$. So $B(z', r/2) \cap S \subset S_h$. Therefore,

$$\begin{aligned} P(S_h) &\geq P(B(z', r/2) \cap S) \geq \underline{\Delta}\mu(B(z', r/2) \cap S) \\ &\geq \delta\underline{\Delta}\mu(B(z', r/2)) = C_3(h_* - h)^d \end{aligned}$$

where $C_3 = \frac{\delta\underline{\Delta}\pi^{d/2}}{\Gamma((d/2)+1)}$.

To lower bound the second term, Let $z_0 = \text{argmax}_z p_{z,h}$. Then,

$$\begin{aligned} 1 - p_{h,\max} &= 1 - P(B(z_0, h)) = P(B(z, h_*)) - P(B(z_0, h)) \\ &= P(B(z, h_*) - B(z_0, h)) \geq \underline{\Delta}\mu((B(z, h_*) - B(z_0, h)) \cap S) \\ &\geq \underline{\Delta}\delta\mu((B(z, h_*) - B(z_0, h))) = \underline{\Delta}\delta\frac{(h_* - h)^d\pi^{d/2}}{\Gamma(d/2 + 1)} \\ &= C_4(h_* - h)^d \end{aligned}$$

where $C_4 = \frac{\underline{\Delta}\delta\pi^{d/2}}{\Gamma(d/2+1)}$. Thus, $(1 - p_{h\max})^n \geq C_4^n(h_* - h)^{nd}$. For the third term, argue as above that $1 - p_{h,\min} \leq C_2(h_* - h)$ so the third term is larger than $1/2$ when h is close enough to h_*. Hence, $A_n \geq \frac{C_3}{2}C_4^n(h_* - h)^{d(n+1)}$. ■

Proof of Theorem 5.1. By our assumptions (see Section 2.1),

$$0 < \lim_{r \to 0} \frac{P(B(x, r))}{r^{d_i}} < \infty$$

where $d_i = \dim(S_i)$, for any x outside of a set of P_i measure zero. By Theorem 5.7 in Mattila (1999), d_i is also the box-counting dimension of S_i. Thus, $d^* = \max_i d_i$. Combined with (29) this implies that, without loss of generality, we can assume that there exist constants $\overline{C} > 0$ and $\overline{\rho} > 0$ such that for every ball B of radius $\rho < \overline{\rho}$ and center in $L(\lambda)$, $P(B) > \overline{C}\rho^{d^*}$.

Let $\mathcal{A}$ be a covering of $L(\lambda)$ with balls of radius $\rho/2$ and centers in $L(\lambda)$, with $\rho < \overline{\rho}$. By compactness of L, $|\mathcal{A}| \leq \overline{M}\rho^{-d^*}$, where $\overline{M}$ depends on d^* and $L(\lambda)$ but not on ρ.

Next, by Lemma 2.2, on the event $\mathcal{E}_{h,\epsilon} = \{\|p_h - \widehat{p}_h\|_\infty < \epsilon\}$, the set $\widehat{L}_h$ consists of k disjoint connected sets. Since $\rho < \delta/2$, this implies, on the same event, that $\widehat{N}_h^G(\lambda) \geq k$. Thus, on the event $\mathcal{E}_{h,\epsilon}$, for some $\epsilon < \epsilon_1$ to be specified below, a sufficient condition for the event $\mathcal{O}_{h,n}$ to be verified is that every $A \in \mathcal{A}$ contains at least one point from the set $\widehat{J}_h \equiv \{i \colon \widehat{p}_h(X_i) > \lambda\}$ (similar arguments are used also in Cuevas et al., 2000; Biau et al., 2007). We conclude that the probability of $\mathcal{O}_{h,n}^c$ is bounded from above by

$$\mathbb{P}(\mathcal{E}_{h,\epsilon}^c) + \overline{M}\rho^{-d^*} \sup_{A \in \mathcal{A}_h} \mathbb{P}\left(\left\{X_i \notin A, \forall i \in \widehat{J}_h\right\} \cap \mathcal{E}_{h,\epsilon}\right).$$

Since, on the event $\mathcal{E}_{h,\epsilon}$ the set $J_{h_n} = \{i \colon p_{h_n}(X_i) > \lambda + \epsilon\}$ is contained in $\widehat{J}_h$, we further have that, for each $A \in \mathcal{A}_h$,

$$\mathbb{P}\left(\left\{X_i \notin A, \forall i \in \widehat{J}_h\right\} \cap \mathcal{E}_{h,\epsilon}\right) \leq (1 - P(A \cap \{p_h > \lambda + \epsilon\}))^n, \tag{35}$$

where the inequality stems from the identity among events

$$\{X_i \notin A, \forall i \in J_h\} = \bigcap_i \{\{p_{h_n}(X_i) > \lambda + \epsilon\} \cap A^c\} \cup \{p_{h_n}(X_i) < \lambda + \epsilon\}\},$$

and the independence of the X_i's. By Lemma 8.1, for any fixed $0 < \tau < 1/2$, there exists a point $y \in L(\lambda) \cap L_h(\lambda + \epsilon)$ such that $B\left(y, \frac{\tau\rho}{2}\right) \subset A \cap L_h(\lambda + \epsilon)$, for all $\epsilon < \epsilon(\rho, \tau)$. Thus,

$$P(A \cap L_h(\lambda + \epsilon)) \geq P\left(B\left(y, \frac{\tau\rho}{2}\right)\right) \geq \overline{C}\left(\frac{\tau\rho}{2}\right)^{d^*},$$

for all $\epsilon < \epsilon(\rho,\tau)$, where the second inequality is verified since $\frac{\rho\tau}{2} < \overline{\rho}$. Set $\epsilon(\rho) = \min\{\epsilon_1, \epsilon(\rho,\tau)\}$. The result now follows from collecting all the terms and the inequality $(1-x)^n \leq e^{-nx}$, valid for all $0 \leq x \leq 1$. ■

Proof of Theorem 5.2. Let $\mathcal{A}_h$ be a covering of $L_h(\lambda)$ by balls of radius $\rho/2$ and centers in $L_h(\lambda)$. By the same arguments used in the proof of the theorem 5.1, the probability of the event $(\mathcal{O}^*_{h,n})^c$ is bounded by

$$\mathbb{P}(\mathcal{E}^c_{h,\epsilon}) + \overline{M}\rho^{-d} \sup_{A\in\mathcal{A}_h} \mathbb{P}\left(\left\{X^*_j \notin A, \forall j\right\} \cap \mathcal{E}_{h,\epsilon}\right),$$

where the probability is over the original sample $X = (X_1,\ldots,X_n)$ and the bootstrap sample $X^* = (X^*_1,\ldots,X^*_N)$. Here the value of $\epsilon < \epsilon_1$ used in the definition of the event $\mathcal{E}_{h,\epsilon}$ is to be specified below. Because of compactness of the support of P, $\overline{M}$ is a constant depending only d and $S + B(0,\overline{h})$.

For a set $S \subseteq \mathbb{R}^d$, we denote with $S^{\otimes n}$ the $n-$fold Cartesian product of S and with $P^h_{X^*|X=x}$ the conditional distribution of the bootstrap sample X^* given $X = x$, with $x = (x_1,\ldots,x_n)$. Let $\mathcal{E}_n = \{x \in S^{\otimes n} \colon \|p_h - \widehat{p}_h\|_\infty \leq \epsilon\}$, where $\widehat{p}_h$ is the kernel density estimate based on x. Then, for each $A \in \mathcal{A}_h$,

$$\mathbb{P}\left(\left\{X^*_j \notin A, \forall j\right\} \cap \mathcal{E}_{h,\epsilon}\right) = E_X\left(P_{X^*|X}\left((A^c)^{\otimes n}\right); \mathcal{E}_n\right),$$

where, if $X \sim P$, $E_X(f(X);\mathcal{E}) \equiv \int_{\{x\in\mathcal{E}\}} f(x)dP(x)$. For every $x \in \mathcal{E}_n$, by the conditional independence of X^* given $X = x$,

$$\begin{aligned} P_{X^*|X=x}\left((A^c)^{\otimes n}\right) &= \left(1 - \frac{\int_{A\cap L_h(\lambda)} \widehat{p}_h(v)dv}{\int_{\{\widehat{p}_h > \lambda\}} \widehat{p}_h(v)dv}\right)^N \\ &\leq \left(1 - \frac{\int_{A\cap L_h(\lambda+\epsilon)} (p_h - \epsilon)d\mu}{V(h,\epsilon)}\right)^N. \end{aligned}$$

where the inequality is due to the fact that $x \in \mathcal{E}_n$ and with

$$V(h,\epsilon) = \int_{L_h(\max\{\lambda-\epsilon,0\})} (p_h + \epsilon)d\mu.$$

By Lemma 8.1, for any fixed $\tau < 1/2$ and each h, there exists a point $y \in L_h(\lambda) \cap L_h(\lambda+\epsilon)$ such that $B\left(y, \frac{\tau\rho}{2}\right) \subset A \cap L_h(\lambda+\epsilon)$, for all $\epsilon < \epsilon(\rho,\tau)$. Thus,

$$\int_{A\cap L_h(\lambda+\epsilon)} (p_h - \epsilon)d\mu \geq \int_{B\left(y,\frac{\tau\rho}{2}\right)} (p_h - \epsilon)d\mu = \int_{B\left(y,\frac{\tau\rho}{2}\right)} p_h d\mu - \epsilon\mu\left(B\left(y, \frac{\tau\rho}{2}\right)\right).$$

Next,

$$V(h,\epsilon) = \int_{L_h(\lambda)} p_h d\mu + \epsilon\mu\left(L_h(\max\{\lambda-\epsilon,0\})\right) + \int_{L_h(\lambda)-L_h(\max\{\lambda-\epsilon,0\})} p_h d\mu.$$

Following the proof of Lemma 8.1, one can verify that, because of assumption (G), $\inf_{h\in(0,\overline{h})} \mu\left(L_h(\lambda) - L_h(\max\{\lambda-\epsilon,0\})\right) \to 0$, as $\epsilon \to 0$. Thus,

$$\frac{\int_{A\cap L_h(\lambda+\epsilon)}(p_h-\epsilon)d\mu}{V(h,\epsilon)} \geq \frac{\int_{B(y,\frac{\tau\rho}{2})} p_h d\mu}{\int_{L_h(\lambda)} p_h d\mu}(1+o(1)),$$

as $\epsilon \to 0$. Then, using (30) and the facts $\tau < 1/2$ and $\int_{L_h(\lambda)} p_h d\mu \leq 1$ for each h, we conclude that there exists a $\epsilon(\rho,\tau)$ such that

$$\frac{\int_{A\cap L_h(\lambda+\epsilon)}(p_h-\epsilon)d\mu}{V(h,\epsilon)} \geq C\rho^d$$

for all $0 < \epsilon < \epsilon(\rho,\tau)$ and for some appropriate constant C, independent of ρ and h. Thus,

$$P_{X^*|X=x}\left((A^c)^{\otimes n}\right) \leq e^{-NC\rho^d}$$

and the results now follows by setting $\epsilon(\rho) = \min\{\epsilon_1, \epsilon(\rho,\tau)\}$. ■

Lemma 8.1. *Assume conditions (LN2) (a) and (b) and condition (G). Then, for any $0 < \tau < 1$ and $\rho > 0$, there exists a positive number $\epsilon(\rho, \tau)$ such that, for all $\epsilon < \epsilon(\rho, \tau)$,*

$$\sup_{h \in (0, \overline{h})} \sup_{x \in L(\lambda)} \operatorname{dist}(x, L_h(\lambda + \epsilon)) < \tau\rho. \tag{36}$$

Proof of Lemma 8.1. The claim follows by minor modifications of the arguments used in the Appendix of Biau et al. (2007). We provide some details for completeness and refer to Lee (2003) for background. Because of assumption (G) and in virtue of the regular level set theorem (see, e.g. Lee, 2003, Corollary 8.10), for any $\epsilon \in (0, \epsilon_1)$ and $h \in (0, \overline{h})$, the set $\{x \colon p_h(x) = \lambda + \epsilon\}$ is a closed embedded submanifold of $\mathbb{R}^d$. Let $r(\epsilon, h)$ be the maximal radius of the tubular neighborhood around $\{x \colon p_h(x) = \lambda + \epsilon\}$. Set $\overline{r}_h = \inf_{\epsilon < \epsilon_1} r(\epsilon, h)$ and notice that $\overline{r}_h > 0$ is positive for each $h \in (0, \overline{h})$. Then, following the proof of Biau et al. (2007, Proposition A.2), if $\epsilon < \epsilon_1$, for any $h \in (0, \overline{h})$,

$$\sup_{x \in \partial L_h(\lambda)} \operatorname{dist}(x, L_h(\lambda + \epsilon)) \leq C_g^{-1} \epsilon, \tag{37}$$

where C_g in the same constant appearing in (28) (see Equation (A.1) in Biau et al., 2007). In fact, since C_g does not depend on h, (37) holds uniformly over $h \in (0, \overline{h})$. Set $\epsilon(\rho, \tau) = \sup\{\epsilon \in (0, \epsilon_1) \colon C\epsilon < \tau\rho\}$. Then, as $L(\lambda) \subseteq L_h(\lambda)$ by (LN2) (b), (36) is verified for each $\epsilon < \epsilon(\rho, \tau)$. ■

9 Appendix

9.1 The Geometric Density

In this section we describe in detail our assumptions on the unknown distribution P. For the sake of completeness, we provide the basic definitions of Hausdorff measure, Hausdorff dimension and rectifiability. We refer the reader to Evans and Gariepy (1992), Mattila (1999), Ambrosio et al. (2000) and Federer (1969) for all the relevant geometric and measure theoretic background.

Let $k \in [0, \infty)$. The k-dimensional Hausdorff measure of a set E in $\mathbb{R}^d$ is defined as $\mathcal{H}^k(E) \equiv \lim_{\delta \downarrow 0} \mathcal{H}^k_\delta(E)$, where, for $\delta \in (0, \infty]$,

$$\mathcal{H}^k_\delta(E) = \frac{v_k}{2^k} \inf \left\{ \sum_{i \in I} (\operatorname{diam}(E_i))^k \colon \operatorname{diam}(E_i) < \delta \right\}$$

where the infimum is over all the countable covers $\{E_i\}_{i \in I}$ of E, with the convention $\operatorname{diam}(\emptyset) = 0$. The Hausdorff dimension of a set $E \subset \mathbb{R}^d$ is

$$\inf \left\{ k \geq 0 \colon \mathcal{H}^k(E) = 0 \right\}.$$

Note that $\mathcal{H}^0$ is the counting measure, while $\mathcal{H}^d$ coincides with the (outer) Lebesgue measure. When $1 \leq k < d$ is an integer, $\mathcal{H}^k(E)$ coincides with the k-dimensional area of E, if E is contained in a $\mathcal{C}^1$ k-dimensional manifold embedded in $\mathbb{R}^d$.

The set E is said to be $\mathcal{H}^k$-rectifiable if $\mathcal{H}^k(E) < \infty$ and there exists countably many Lipschitz functions $f_i \colon \mathbb{R}^k \mapsto \mathbb{R}^d$ such that

$$\mathcal{H}^k \left(E - \bigcup_{i=0}^{\infty} f_i(\mathbb{R}^k) \right).$$

A Radon measure ν in $\mathbb{R}^d$ is said to be k-rectifiable if there exists a $\mathcal{H}^k$-rectifiable set S and a Borel function $f \colon S \mapsto \mathbb{R}^d$ such that

$$\nu(A) = \int_{A \cap S} f(x) d\mathcal{H}^k(x),$$

for each measurable set $A \subseteq \mathbb{R}^d$.

Throughout this article, we assume that P is a finite mixture of probability measures supported on disjoint compact sets of possibly different integral dimensions. Formally, for each Borel set $A \subseteq \mathbb{R}^d$ and for some integer m,

$$P(A) = \sum_{i=1}^{m} \pi_i P_i(A),$$

where π is a point in the interior of the $(m-1)$-dimensional standard simplex and each P_i is a d_i-rectifiable Radon measure with compact and connected support S_i, where $d_i \in \{0, 1, \ldots, d\}$ and $S_i \cap S_j = \emptyset$, for each $i \neq j$. By Theorem 3.2.18 in Federer (1969), each of the lower dimensional rectifiable sets comprising the support of P, can be represented as the union of $\mathcal{C}^1$ embedded submanifolds, almost every everywhere P. Thus, we are essentially allowing P to be a mixture of distributions supported on disjoint submanifolds of different dimensions and finite sets.

Our assumptions imply that, for every mixture component P_i, there exists a measurable real valued function p_i vanishing outside S_i such that

$$p_i(x) = \lim_{h \to 0} \frac{P_i\left(B(x,h)\right)}{v_{d_i} h^{d_i}}, \quad \forall x \in S_i, \tag{38}$$

where v_{d_i} is the volume of the unit Euclidean ball in $\mathbb{R}^{d_i}$. Furthermore, for any measurable set A,

$$P_i(A) = \int_{A \cap S_i} p_i(x) d\mathcal{H}^{d_i}(x),$$

where $\mathcal{H}^{d_i}$ denotes the d_i-dimensional Hausdorff measure on $\mathbb{R}^d$. Notice that we also have $\max_i \mathcal{H}^{d_i}(S_i) < \infty$.

We do not assume any knowledge of the probability measures comprising the mixture P, of their number, supports and dimensions, nor of the vector of mixing probabilities π.

In virtue of (38), $p(x) = \infty$ if and only if $p_i(x) > 0$ for some i such that $d_i < d$, which implies that $p(x) = \infty$ if and only if $x \in S_i$ with $d_i < d$, almost everywhere P. Similarly, S_i has Hausdorff dimension d if and only if $p(x) = \pi_i p_i(x)$ for each $x \in S_i$, almost everywhere μ. Furthermore, if $x \notin S$, then $p(x) = 0$. Notice that, since $\mu\left(\{x \colon p(x) = \infty\}\right) = 0$, the geometric density p needs not be a probability density because, in general, $\int_{\mathbb{R}^d} p(x) d\mu(x) \leq 1$. Nonetheless,

$$S = \overline{\{x \colon p(x) > 0\}}.$$

As a final remark, even though the geometric density p is very different from the mixture densities p_i, for our clustering purposes, we need only to concern ourselves with estimating the level sets of p.

9.2 On condition (30)

In this section we show that condition (G), (29) and condition (T), given below, imply (30). We focus only on the case in which L has dimensional smaller than d. If L is full-dimensional, then it is easy to see that (30) holds.

For each h, let $d(h) \equiv \inf_{x \in L, y \in \partial L_h} \|x - y\|$. Notice that $0 < d_h \leq h$. Let $1 < D < 2$. For any $c \in (0,1)$ define $h_1(c)$ to be the infimum of all $h < \overline{h}$ such that for every ball of radius ρ no smaller than $Dd(h)$ and center in $L_h(\lambda)$ there exists a ball $B \subset A \cap L_h$ of radius $c\rho$. Set $h_1(c) = \infty$ when the infimum does not exist. Let $c_* = \sup\{c \in (0,1) \colon h_1(c) < \infty\} > 0$. It can be seen that, for any $c < c_*$, $h_1(c) < \infty$ and that $h_1(c) \to 0$ as $c \downarrow 0$.

Next, let A_r be a ball of radius r and center in L. Then, $P(\partial A_r) = 0$ for all $r \in (0, \delta/2) - R$, where $R \subset (0, \delta/2)$ is finite (possible empty). Thus, by Theorem 4.2 in Rao (1962), for any $0 < U < 1$ there exists a $h_2(U)$ such that, for every $h < h_2(U)$,

$$\frac{P_h(A_r)}{P(A_r)} > U,$$

uniformly over the set of balls A_r with center in L and radius $r \in (0, \delta/2) - R$. Furthermore, since $\sup_{x\in L, y\in L_h} \|x - y\| \to 0$ as $h \to 0$, we can choose U and $h_1(U)$ such that $\frac{P_h(A_r \cap L_h)}{P(A_r)} > U$, for all $h < h_1(U)$, uniformly over the balls A_r. Therefore, by choosing an appropriate c and U, we may then assume that $h_1(c) \le h_2(U)$. Set $h^* = (h_2(U) - h_1(c))/2$.

Let $\rho < \overline{\rho}$ be given. We consider 3 cases

1. $\rho \ge Dd(h^*)$ and $h \ge h^*$.
   In this simple case, we immediately obtain
   $$P_h(A_h \cap L_h) \ge \lambda\mu(B) = \lambda v_d c^d \rho^d. \tag{39}$$

2. $\rho \ge Dd(h^*)$ and $h < h^*$.
   By the definition of $d(h^*)$, there exists a ball B of radius $(D-1)\rho$ and center in L contained in $A \cap L_h$. Thus,
   $$P_h(A_h \cap L_h) \ge P_h(B) \ge UP(B) > U\overline{C}((D-1)\rho)^{d^*}, \tag{40}$$
   where the last inequality stems from (29).

3. $\rho < Dd(h^*)$.
   First suppose that A_h is centered in L. If $h \ge h^*$, then the ball B having the same center as A_h and radius $\min\{\rho, d(h^*)\}$ is entirely contained in L_h. Thus, $P_h(A_h \cap L_h) \ge P_h(B)$, and $P_h(B)$ at least $\lambda v_d \rho^d$ if $\rho < d(h^*)$ and at least $\frac{\lambda}{D} v_d \rho^d$ if $d(h^*) \le \rho \le Dd(h^*)$. If $h < h^*$, then
   $$P_h(A_h \cap L_h) \ge UP(A_h) > U\overline{C}\rho^{d^*}, \tag{41}$$
   We now consider the other case of A_h centered in $L_h - L$. If $\rho \ge Dd(h)$, then the same arguments as in case 2. above applies. Thus we only need to consider the case $\rho < Dd(h)$. Actually, since $D > 1$, analyzing the case $\rho < d(h)$ will be enough. In fact, we will show that there exists a ball B of radius $c\tau\rho$ such that $B \subset A_h \cap L_h$, where τ is the constant in assumption (T); this will imply that
   $$P_h(A_h \cap L_h) \ge P_h(B) \ge \lambda v_d (\tau c)^d \rho^d. \tag{42}$$
   To prove the claim, first observe that, for each ball A_h of any radius and center in L_h, there exists a ball A_1 with the same radius and center in ∂L_h such that
   $$\mu(A) \ge \mu(A_1).$$
   Thus, without loss of generality, we can assume that A_h is centered in ∂L_h. Next, by assumption (G), for each h, ∂L_h is a $(d-1)$-dimensional closed embedded submanifold of $\mathbb{R}^d$. Thus, by a straightforward adaptation of Proposition A.1 in Biau et al. (2007), there exists a $\rho(h)$ such that for each $r < \rho(h)$ and each ball A_h of radius r and center in ∂L_h, there exists a ball B of radius cr such that $B \subset A_h \cap L_h$. Suppose that the following condition holds:

   (T) there exists a constant $0 < \tau < 1$ such that
   $$\inf_{0<h<\overline{h}} \frac{d(h)}{\rho(h)} > \tau.$$
   Thus, there exists a ball $B \subset A_h \cap L_h$ of radius $\tau c\rho$ if if $\rho(h) < \rho < d(h)$ and of radius $c\rho$ if $\rho < \rho(h)$. Since $\tau < 1$, (42) is verified.

The claim follows from taking the minimal value of the constants in (39), (40) , (41) and (42).

References

Ambrosio, L., N. Fusco, and D. Pallara (2000). *Functions of Bounded Variation and Free Discontinuity Problems*. Oxford University Press.

Audibert, J. and A. Tsybakov (2007). Fast learning rates for plug-in classifiers. *Annals of Statistics 2*, 608–633.

Ben-David, S., U. von Luxburg, and D. Pall (2006). A sober look at clustering stability. *Learning Theory*, 5–19.

Ben-Hur, A., A. Elisseef, and I. Guyon (2002). A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing*.

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

Cuevas, A., M. Febrero, and R. Fraiman (2000). Estimating the number of clusters. *The Canadian Journal of Statistics 28*.

Cuevas, A. and R. Fraiman (1997). A plug-in approach to support estimation. *The Annals of Statistics 25*(6), 2300–2312.

Devroy, L. and L. Wise (1980). Detection of abnormal behavior via nonparametric estimation of the support. *SIAM, Journal of Applied Mathematics 38*, 3.

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

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

Federer, H. (1969). *Geometric Measure Theory*. Springer.

Giné, E. and A. Guillou (2002). Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'institut Henri Poincaré (B),* Probabilités et Statistiques *38*(6), 907–921.

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

Hartigan, J. (1975). *Clustering Algorithms*. Wiley.

Lange, T., V. Roth, M. Braun, and J.Buhmann (2004). Stability-based validation of clustering solutions. *Neural Computation*.

Lee, J. (2003). *Introduction to Smooth Manifolds*. Springer.

Leoni, G. and I. Fonseca (2007). *Modern Methods in the Calculus of Variations: L^p Spaces*. Springer.

Mammen, E. and A. B. Tsybakov (1999). Smooth discrimination analysis. *The Annals of Statistics 27*, 1808–1829.

Massart, P. (2007). About the constants in talagrand's concentration inequalities for empirical processes. *Annals of Probability 28*, 863–884.

Mattila, P. (1999). *Geometry of Sets and Measures in Euclidean Spaces: Fractals and Rectifiability*. Cambridge University Press.

Mueller, D. and G. Sawitzki (1991). Excess mass estimates and test for multimodality. *Journal of the American Statistical Association 86*, 738–746.

Ng, A., M. Jordan, and Y. Weiss (2002). On spectral clustering: Analysis and an algorithm. *NIPS*.

Niyogi, P., S. Smale, and S. Weinberger (2008). Finding the homology of submanifolds with high confidence. *Discrete and Compuational Geometry 38*(1-3), 419–441.

Nolan, D. and D. Pollard (1987). U-processes: Rates of convergence. *Annals of Statistics 15*(2), 780–799.

Polonik, W. (1995). Measuring mass concentrations and estimating density contour clusters-an excess mass approach. *The Annals of Statistics 23*, 855–881.

Rao, R. (1962). Relations between weak and uniform convergence of measures with applications. *Annals of Mathematical Statistics 33*(2), 659–680.

Rigollet, P. (2007). Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Leanring Research 8*, 1369–1392.

Rigollet, P. and R. Vert (2006). Fast rates for plug-in estimators of density level sets. *arXiv:math/0611473*.

Singh, A., R. Nowak, and X. Zhu (2009). Unlabeled data: Now it helps, now it doesn't. *NIPS*.

Stuetzle, W. and R. Nugent (2009). A generalized single linkage method for estimating the cluster tree of a density. *Journal of Computational and Graphical Statistics*.

Tsybakov, A. B. and A. Korostelev (1993). *Minimax Theory of Image Reconstruction*. Springer-Verlag.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing 17*, 395–416.

Willett, R. and R. Nowak (2007). Minimax optimal level-set estimation. *IEEE Transactions on Image Processing 12*, 2965–2979.