Journal of Machine Learning Research 13 (2012) 905-948 Submitted 11/10; Revised 1/12; Published 4/12

# Stability of Density-Based Clustering

**Alessandro Rinaldo** ARINALDO@CMU.EDU
*Department of Statistics*
*Carnegie Mellon University*
*Pittsburgh, PA 15213, USA*

**Aarti Singh** AARTI@CS.CMU.EDU
*Machine Learning Department*
*Carnegie Mellon University*
*Pittsburgh, PA 15213, USA*

**Rebecca Nugent** RNUGENT@STAT.CMU.EDU
**Larry Wasserman**[*] LARRY@STAT.CMU.EDU
*Department of Statistics*
*Carnegie Mellon University*
*Pittsburgh, PA 15213, USA*

**Editor:** Sanjoy Dasgupta

## Abstract

High density clusters can be characterized by the connected components of a level set $L(\lambda) = \{x : \ p(x) > \lambda\}$ of the underlying probability density function $p$ generating the data, at some appropriate level $\lambda \geq 0$. The complete hierarchical clustering can be characterized by a cluster tree $\mathcal{T} = \bigcup_\lambda L(\lambda)$. In this paper, we study the behavior of a density level set estimate $\widehat{L}(\lambda)$ and cluster tree estimate $\widehat{\mathcal{T}}$ based on a kernel density estimator with kernel bandwidth $h$. We define two notions of instability to measure the variability of $\widehat{L}(\lambda)$ and $\widehat{\mathcal{T}}$ as a function of $h$, and investigate the theoretical properties of these instability measures.

**Keywords:** clustering, density estimation, level sets, stability, model selection

## 1. Introduction

A common approach to identifying high density clusters is based on using level sets of the density function (see, for instance, Hartigan, 1975; Rigollet and Vert, 2009). Let $X_1, \ldots, X_n$ be a random sample from a distribution $P$ on $\mathbb{R}^d$ with density $p$. For $\lambda > 0$ define the level set $L(\lambda) = \{x : p(x) > \lambda\}$. Assume that $L(\lambda)$ can be decomposed into disjoint, connected sets $L(\lambda) = \bigcup_{j=1}^{N(\lambda)} C_j$. Following Hartigan (1975), we refer to $\mathcal{C}_\lambda = \{C_1, \ldots, C_{N(\lambda)}\}$ as the *density clusters* at level $\lambda$. We call the collection of clusters

$$\mathcal{T} = \bigcup_{\lambda \geq 0} \mathcal{C}_\lambda$$

the *cluster tree* of the density $p$. Note that $\mathcal{T}$ does indeed have a tree structure: if $A, B \in \mathcal{T}$ then either, $A \subset B$, or $B \subset A$ or $A \cap B = \emptyset$. The cluster tree summarizes the cluster structure of the distribution; see Stuetzle and Nugent (2009).

∗. Also in the Machine Learning Department.

It is also possible to index the level sets by probability content. For $0 < \alpha < 1$, define the level set $M(\alpha) = L(\lambda_\alpha)$, where

$$\lambda_\alpha = \sup\{\lambda : \ P(L(\lambda)) \geq \alpha\}.$$

If the density does not contain any jumps or flat parts, then there is a one-to-one correspondence between the level sets indexed by the density level and the probability content. The cluster tree obtained from the clusters of $M(\alpha)$ for $0 \leq \alpha \leq 1$ is equivalent to $\mathcal{T}$. Relabeling the tree in terms of $\alpha$ may be convenient because $\alpha$ is more interpretable than $\lambda$, but the tree is the same. Figure 1 shows the cluster tree for a density estimate of a mixture of three normals (using a reference rule bandwidth). The cluster tree's two splits and subsequent three leaves correspond to the density estimate's modes. The tree is also indexed by $\lambda$, the density estimate's height, on the left and $\alpha$, the probability content, on the right. For example, the second split corresponds to $\lambda = 0.086$ and $\alpha = 0.257$. We note here that determining the true clusters for even this seemingly simple univariate distribution is not trivial for all $\lambda$; in particular, values of $\lambda$ near 0.04 and 0.09 will give ambiguous results.

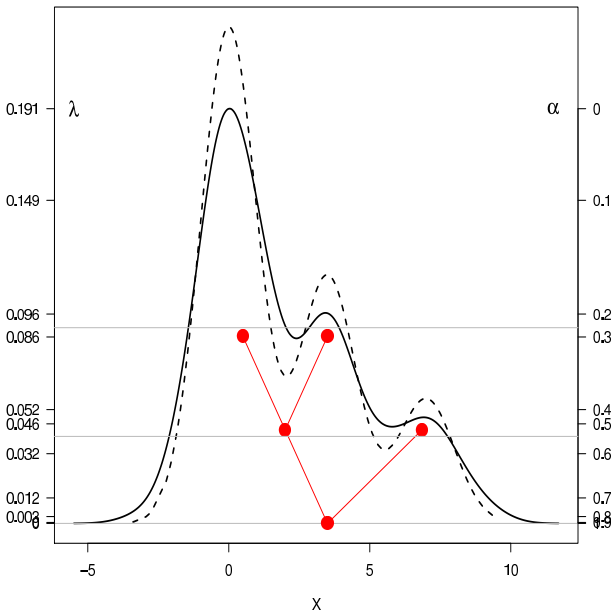


Figure 1: The cluster tree for a Gaussian kernel density estimate (normal reference rule bandwidth) of a sample from the mixture $(4/7)N(0,1) + (2/7)N(3.5,1) + (1/7)N(7,1)$; the tree is indexed by both $\lambda$ (left) and $\alpha$ (right). The dashed curve indicates the true underlying density. The gray lines indicate $L(0.04)$, $L(0.09)$.

In this paper we study some properties of clusters defined by density level sets and cluster trees. In particular, we consider their estimators based on a kernel density estimate and show how the bandwidth $h$ of the kernel affects the risk of these estimators. Then we investigate the notion of stability for density-based clustering. Specifically, we propose two measures of instability. The first, denoted by $\Xi_{\lambda,n}(h)$, measures the instability of a given level set. The second, denoted by $\Gamma_n(h)$, is a more global measure of instability.

Investigation of the stability properties of density clusters is the main focus of the paper. Stability has become an increasingly popular tool for choosing tuning parameters in clustering; see von Luxburg (2009), Lange et al. (2004), Ben-David et al. (2006), Ben-Hur et al. (2002), Carlsson and Memoli (2010), Meinshausen and Bühlmann (2010), Fischer and Buhmann (2003), and Rinaldo and Wasserman (2010). The basic idea is this: clustering procedures inevitably depend on one or more tuning parameters. If we choose a good value of the tuning parameter, then we expect that the clusters from different subsets of the data should be similar. While this idea sounds simple, the reality is rather complex. Figure 2 shows a plot of $\Xi_n$ and $\Gamma_n$ for our example. We see that $\Xi_{\lambda,n}(h)$ is a complicated function of $h$ while $\Gamma_n(h)$ is much simpler. Our results will explain this behavior.

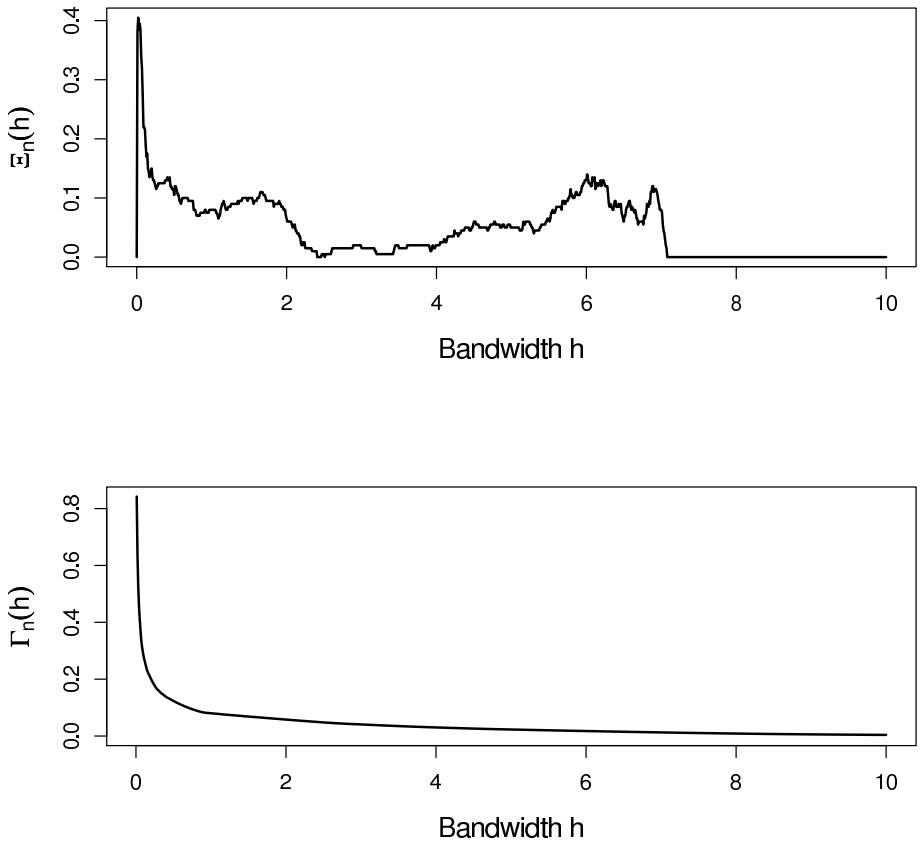


Figure 2: Plots of the fixed-$\lambda$ instability (top) $\Xi_{\lambda,n}(h)$ for $\lambda = 0.09$ and of the total variation instability $\Gamma_n(h)$ (bottom) for the mixture distribution in Figure 1 as functions of the bandwidth $h$.

Below we briefly describe our contributions.

- We consider plug-in estimates of the level sets $L(\lambda)$ corresponding to fixed density levels $\lambda$ and also to the level sets $L(\lambda_\alpha)$ corresponding to fixed probability contents $\alpha$ using kernel density estimators. We analyze the statistical properties of these plug-in estimates and formulate conditions on the density of the data generating distribution and on the kernel that guarantee accurate recovery of the level sets as $n$ becomes large.

- We formulate a notion of cluster stability of the level sets based on a splitting of the the data that quantifies the variability of the level set estimators we consider. We construct an estimator of the cluster instability and analyze its performance as $n$ become large, and argue that stability can provide a guidance on the optimal choice of the bandwidth parameter. As a result of our analysis, we are able to provide a rigorous characterization of the levels sets for which the the uncertainty is larger and, therefore, for which the cluster tree can be estimated with a smaller degree of accuracy. Our results suggest that the sample complexity for successful reconstruction of the cluster tree may vary significantly depending on whether we estimating a portion of the tree that is far removed from a branching region or not, and for those portion of the tree we provide some rates.

- We formulate and analyze a stronger notion of cluster stability that is based on the total variation distance between kernel density estimates computed over different data subsamples. This second kind of stability is more global and has natural and interesting connections with the problem of optimally estimating a density in $L_1$ norm.

After the writing of the first draft of this paper we learned of the interesting and relevant contributions by Chaudhuri and Dasgupta (2010), Kpotufe and von Luxburg (2011) and Steinwart (2011) who all consider the problem of estimating the cluster tree. Our results provide a different perspective on this issue as we concern ourselves with quantifying, based on stability criteria, the uncertainty of the cluster tree estimate. Furthermore, these papers only characterize the optimal scaling of parameters to guarantee cluster tree recovery and do not provide a data-driven way to choose these parameters. In this paper, we investigate stability as a means for data-adaptive choice of parameters such as the kernel bandwidth.

The paper is organized as follows. In Section 2 we describe the assumptions on the density and recall some facts about kernel density estimation. In Section 3 we construct plug-in estimates $\widehat{L}(\lambda)$ of the level set $L(\lambda)$, $\widehat{\mathcal{T}}$ of the cluster tree $\mathcal{T}$, and $\widehat{M}(\alpha)$ of the level set indexed by probability content $M(\alpha)$. In Section 4 we define and study a notion of the stability of $\widehat{L}(\lambda)$ and extend it to $\widehat{\mathcal{T}}$. We also consider an alternative version of our results when the level sets are indexed by probability content. We then describe another notion of stability of cluster trees based on total variation that leads to a constructive procedure for selecting the kernel bandwidth. In Section 5 we consider some numerical examples. Section 6 contains a discussion of the results and the proofs are in Section A. Throughout, we use symbols like $c, c_1, c_2, \ldots, C, C_1, C_2, \ldots$, to denote various positive constants whose value can change in different expressions.

## 2. Preliminaries

In this section we introduce some notation, state the assumptions on the density we will be using throughout and review some useful facts about kernel density estimation.

### 2.1 Notation

For $x \in \mathbb{R}^d$, let $\|x\|$ denote its Euclidean norm. Let $B(x,\varepsilon) = \{y : \ \|x-y\| \leq \varepsilon\} \subset \mathbb{R}^d$ denote a ball centered at $x$ with radius $\varepsilon$. For two sets $A$ and $B$ in $\mathbb{R}^d$, their Hausdorff distance is

$$d_\infty(A,B) = \inf\{\varepsilon : \ A \subset (B \oplus \varepsilon) \ \text{ and } \ B \subset (A \oplus \varepsilon)\},$$

where $A \oplus \varepsilon = \bigcup_{x \in A} B(x, \varepsilon)$, and

$$A \Delta B = (A \cap B^c) \bigcup (A^c \cap B)$$

denotes the symmetric set difference. Finally, we let $v_d = \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)}$ be the volume of the $d$-dimensional Euclidean unit ball.

For sequences of real numbers $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$ if there exists a $C > 0$ such that $|a_n| \leq C|b_n|$ for all $n$ large enough, and we will write $a_n = \omega(b_n)$ if there exists a constant $C > 0$ such that $|a_n| \geq C|b_n|$ for all $n$ large enough. When $\{a_n\}$ and $\{b_n\}$ are sequences of random variables described by a probability measure $P$, we will write $a_n = O_P(b_n)$ if, for any $\varepsilon > 0$, there exists a constant $C > 0$ such that $|a_n| \leq C|b_n|$ with $P$-probability at least $1 - \varepsilon$ for all $n$ large enough.

We will be considering samples of $n$ independent and identically distributed random vectors from an unknown probability measure $P$ on $\mathbb{R}^d$ with Lebesgue density $p$. If $X$ and $Y$ are such samples, we will denote with $\mathbb{P}_{X,Y}$ the probability measures associated to them and with $\mathbb{E}_{X,Y}$ the corresponding expectation operator. Thus, if $\mathcal{A}$ is an event depending on $X$ and $Y$, we will write $\mathbb{P}_{X,Y}(\mathcal{A})$ for its probability. Finally, for a sample $X = (X_1, \ldots, X_n)$, we will denote with $\widehat{P}_X$ the empirical measure associated with it; explicitly, for any measurable set $A \subset \mathbb{R}^d$,

$$\widehat{P}_X(A) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \in A).$$

### 2.2 Assumptions

We will use the following assumptions on the density $p$ and its local behavior around a given density level $\lambda$.

(A0) *Compact Support* - The support $S$ of $p$ is compact.

(A1) *Lipschitz Density* - Assume that

$$p \in \Sigma(A) \equiv \left\{ p : \ |p(x) - p(y)| \leq A||x - y||, \ \text{ for all } x, y \in S \right\}$$

for some $A > 0$.

(A2) *Local density regularity at* $\lambda$- For a given density level of interest $\lambda$, there exist constants $0 < \kappa_1 \leq \kappa_2 < \infty$ and $0 < \varepsilon_0$ such that, for all $\varepsilon < \varepsilon_0$,

$$\kappa_1 \varepsilon \leq P(\{x : \ |p(x) - \lambda| \leq \varepsilon\}) \leq \kappa_2 \varepsilon.$$

It is possible to formulate condition (A2) more generally in terms of powers of $\varepsilon$, that is $\varepsilon^a$. However, as argued in Rinaldo and Wasserman (2010), the above statement typically holds with $a = 1$ for almost all $\lambda$.

Assumptions (A1) and (A2) impose some mild regularity conditions on the density: (A1) implies that the density cannot change drastically anywhere, while (A2) implies that the density cannot be too flat or steep locally around the level set. In particular, (A2) is necessary to ensure that small error in estimating the density level does not translate into a huge error in localizing the level set.

We remark that this assumption is an extension of the Tsybakov noise-margin condition for classification (see Mammen and Tsybakov, 1999; Tsybakov, 2004) to the density level set context and has been used in other work on density level-set estimation, such as Polonik (1995), Tsybakov (1997), Cuevas et al. (2006), Rigollet and Vert (2009), Singh et al. (2009) and Rinaldo and Wasserman (2010). Finally notice that (A0) and (A1) together imply that the density $p$ is bounded by some positive constant $p_{\max} < \infty$. These assumptions are stronger than necessary, but they simplify the proofs. Notice in particular, that assumptions (A1) and (A2) each rule out the case of sharp clusters, in which $S$ is the disjoint union of a finite number of compact sets over which $p$ is bounded from below by a positive constant. Finally, we remark that some of our results will only require a subset of these assumptions.

### 2.3 Estimating the Density

To estimate the density $p$ based on the i.i.d. sample $X = (X_1, \ldots, X_n)$, we use the kernel density estimator

$$\widehat{p}_{h,X}(u) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{u - X_i}{h}\right), \quad u \in \mathbb{R}^d,$$

where the kernel $K$ is a symmetric, non-negative function with compact support such that $\int_{R^d} K(z)dz = 1$ and $h > 0$ is the bandwidth. In some results we will consider specifically the *spherical kernel* $K(u) = \frac{I_{B(0,1)}(u)}{v_d}$, $u \in \mathbb{R}^d$, where $I_{B(0,1)}(\cdot)$ denotes the indicator function of the Euclidean ball $B(0,1)$.

For $h > 0$, let $p_h(u) = \mathbb{E}_X[\widehat{p}_{h,X}(u)]$. Note that $p_h$ is the Lebesgue density of the probability measure

$$P_h = P * \mathbb{K}_h,$$

where $*$ denotes convolution of probability measures and $\mathbb{K}_h$ denotes the probability measure of a random variable with density $K_h(z) = h^{-d}K(z/h)$, $z \in \mathbb{R}^d$.

We note that the compactness of $K$ and assumption (A0) on $p$ imply that the support of $P_h$ is compact, while assumption (A1) on $p$ further yields that $p_h \in \Sigma(A)$, both statements holding for all $h \geq 0$ (for a formal proof of the second claim, see the end of the proof of Lemma 5). Below, we will be concerned with given values of the density level $\lambda$ and of the probability parameter $\alpha \in (0,1)$ and will impose the following assumptions.

(B2) *Local density regularity at* $\lambda$- For a given density level $\lambda$, there exist positive constants $\kappa_1' \leq \kappa_2'$, $\varepsilon_0$ and $H$ bounded away from 0 and $\infty$, such that, for all $0 \leq \varepsilon < \varepsilon_0$,

$$\kappa_1'\varepsilon \leq \inf_{0\leq h\leq H} P(\{x\colon |p_h(x) - \lambda| \leq \varepsilon\}) \leq \sup_{0\leq h\leq H} P(\{x\colon |p_h(X) - \lambda| \leq \varepsilon\}) \leq \kappa_2'\varepsilon.$$

(B3) *Local density regularity at* $\alpha$- For a given probability value $\alpha$, there exist positive constants $\kappa_3$, $\eta_0$ and $H$ bounded away from 0 and $\infty$, such that, for all $0 \leq \eta < |\eta_0|$,

$$\sup_{0\leq h\leq H} d_\infty(M_h(\alpha), M_h(\alpha+\eta)) \leq \kappa_3|\eta|,$$

where $M_h(\alpha) = \{u : p_h(u) > \lambda_\alpha\}$.

Conditions (B2) and (B3) are used only for some specific results from Section 4.1 and Section 3.2, respectively. This will be explicitly mentioned in the statement of such results. In particular, condition (B2) is needed in order to explicitly state the behavior of the instability measure we define below. We conjecture that (B2) follows from condition (A2) on the true density $p$ and using kernels with compact support. This assumption holds for all density levels that are not too close to a local maxima or minima of the density. Assumption (B3) characterizes the regularity of the level sets of $p_h$ and essentially states that the boundary of these level sets is well-behaved and not space-filling (see Tsybakov, 1997; Singh et al., 2009, for analogous conditions). Both assumptions (B2) and (B3) could be stated more generally by assuming some uniformity over $\lambda$ and $\alpha$ respectively, but for the sake of readability we state them as point-wise conditions.

Our analysis depends crucially on the quantity $\|\widehat{p}_{h,X} - p_h\|_\infty = \sup_{u\in\mathbb{R}^d} |\widehat{p}_{h,X}(u) - p_h(u)|$, for which we use a probabilistic upper established by Giné and Guillou (2002), to which the reader is referred for details. To this end, we will make the following assumption on the kernel $K$:

(VC) The class of functions

$$\mathcal{F} = \left\{ K\left(\frac{x-\cdot}{h}\right), x \in \mathbb{R}^d, h > 0 \right\}$$

satisfies, for some positive numbers $V$ and $v$,

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

where $N(T;d;\varepsilon)$ denotes the $\varepsilon$-covering number of the metric space $(T,d)$, $F$ is the envelope function of $\mathcal{F}$ and the supremum is taken over the set of all probability measures $P$ on $\mathbb{R}^d$. The quantities $V$ and $v$ are called the VC characteristics of $\mathcal{F}$.

Assumption (VC) holds for a large class of kernels, including, any compactly supported polynomial kernel and the Gaussian kernel. The lemma below follows from Giné and Guillou (2002) (see also Rinaldo and Wasserman, 2010).

**Lemma 1** *Assume that the kernel satisfies the VC property, and that*

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

*There exist positive constants $K_1$, $K_2$ and $C$, which depends on $B$ and the VC characteristic of $K$ such that the following hold:*

1. *For every $\varepsilon > 0$ and $h > 0$, there exists $n(\varepsilon, h)$ such that, for all $n \geq n(\varepsilon, h)$*

$$\mathbb{P}_X\left(\|\widehat{p}_{h,X} - p_h\|_\infty > \varepsilon\right) \leq K_1 e^{-K_2 n \varepsilon^2 h^d}. \tag{1}$$

2. *Let $h_n \to 0$ as $n \to \infty$ in such a way that*

$$\frac{n h_n^d}{\log n} \to \infty. \tag{2}$$

*Then, there exist a constant $K_3$ and a number $n_0 \equiv n_0(d, K_3)$ such that, setting $\varepsilon_n = \sqrt{\frac{K_3 \log n}{n h_n^d}}$,*

$$\mathbb{P}_X\left(\|\widehat{p}_{h_n,X} - p_{h_n}\|_\infty > \varepsilon_n\right) \leq \frac{1}{n}, \tag{3}$$

*for all $n \geq n_0(d, K_3)$.*

*The numbers $n(\varepsilon, h)$ and $n_0$ depend also on the VC characteristic of $K$ and on $B$. Furthermore, $n(\varepsilon, h)$ is decreasing in both $\varepsilon$ and $h$.*

This result requires virtually no assumptions on $p$ and only minimal assumptions on the kernel, which are satisfied by the most commonly used kernels.

The constraint in Equation (2), which in general cannot be dispensed with, has a subtle but important implication for our later results on instability. In fact, it implies that the bandwidth parameter $h_n$ is only allowed to vanish at a slower rate than $\left(\frac{\log n}{n}\right)^{1/d}$. As a result, our measures of instability defined in Sections 4.1 and 3.2 can be reliably estimated for values of the bandwidth $h \gg \left(\frac{\log n}{n}\right)^{1/d}$. Indeed, the threshold value $\left(\frac{\log n}{n}\right)^{1/d}$ is of the same order of magnitude of the maximal spacing among the points in a sample of size $n$ from $P$ (see, for instance, Penrose, 2003).

## 3. Estimating the Level Set and Cluster Tree

For a given density level $\lambda$ and kernel bandwidth $h$, the estimated level set is $\widehat{L}_{h,X}(\lambda) = \{x\colon\ \widehat{p}_{h,X}(x) > \lambda\}$. The clusters (connected components) of $\widehat{L}_{h,X}(\lambda)$ are denoted by $\widehat{C}_{h,\lambda}$ and the estimated cluster tree is

$$\widehat{\mathcal{T}}_h = \bigcup_{\lambda \geq 0} \widehat{C}_{h,\lambda}.$$

### 3.1 Fixed $\lambda$

We measure the quality of $\widehat{L}_{h,X}(\lambda)$ as an estimator of $L(\lambda)$ using the loss function

$$\mathcal{L}(h, X, \lambda) = \int_{L(\lambda) \Delta \widehat{L}_{h,X}(\lambda)} p(u)du,$$

where we recall that $\Delta$ denotes the symmetric set difference. The performance of plug-in estimators of density level sets has been studied earlier, but we state the results here in a form that provides insights into the performance of instability measures proposed in the next section.

**Theorem 2** *Assume that the density $p$ satisfies conditions (A0) and (A1) and let $D = \int \|z\| K(z)dz$ (which is finite by compactness of $K$). For any sequence $h_n = \omega((\log n/n)^{1/d})$, let*

$$\varepsilon_n = \sqrt{\frac{K_3 \log n}{n h_n^d}}$$

*and*

$$r_{h_n, \varepsilon_n, \lambda} = P\left(\{u\colon\ |p(u) - \lambda| < ADh_n + \varepsilon_n\}\right).$$

*Then, for all $n \geq n(n_0, \lambda, A, D, d)$,*

$$\mathbb{P}_X\left(\mathcal{L}(h_n, X, \lambda) \leq r_{h_n, \varepsilon_n, \lambda}\right) \geq 1 - \frac{1}{n}.$$

*If assumption (A2) holds for the density level $\lambda$, then for all $n \geq n(n_0, \lambda, A, D, \varepsilon_0, d)$,*

$$\mathbb{P}_X\Big(\mathcal{L}(h_n, X, \lambda) \leq \kappa_2(ADh_n + \varepsilon_n)\Big) \geq 1 - \frac{1}{n}.$$

The following corollary characterizes the optimal scaling of the bandwidth parameter $h_n$ that balances the approximation and estimation errors.

**Corollary 3** *The value of h that minimizes the bound on $\mathcal{L}$ is*

$$h_n^* = c\left(\frac{n}{\log n}\right)^{-\frac{1}{d+2}},$$

*where $c > 0$ is an appropriate constant.*

### 3.2 Fixed $\alpha$

Often it is more natural to index the density clusters by the probability mass contained in the corresponding high-density regions, instead of the associated density levels. The level set estimator indexed by the probability content $\alpha \in (0,1)$ is given as

$$\widehat{M}_{h,X}(\alpha) = \widehat{L}_{h,X}(\widehat{\lambda}_{h,\alpha,X}),$$

where

$$\widehat{\lambda}_{h,\alpha,X} = \sup\left\{\lambda : \widehat{P}_X(\{u : \widehat{p}_{h,X}(u) > \lambda\}) \geq \alpha\right\} \tag{4}$$

and $\widehat{p}_{h,X}$ is the kernel density estimate computed using the data $X$ with bandwidth $h$. This estimator was studied by Cadre et al. (2009), though using different techniques and in different settings than ours.

Let $\alpha \in (0,1)$ be fixed and define

$$\lambda_{h,\alpha} = \sup\{\lambda : P(p_h(X) > \lambda) \geq \alpha\}.$$

We first show that the deviation $|\lambda_{h,\alpha} - \lambda_\alpha|$ is of order $h$, uniformly over $\alpha$, under the very general assumption that the true density $p$ is Lipschitz.

**Lemma 4** *Assume the true density p satisfies the conditions (A0) and (A1). Then, for any $h > 0$,*

$$\sup_{\alpha\in(0,1)} |\lambda_{h,\alpha} - \lambda_\alpha| \leq ADh,$$

*where $D = \int_{\mathbb{R}^d} \|z\| K(z) dz$.*

**Remark:** More generally, if $p$ is assumed to be Hölder continuous with parameter $\beta$ then, under additional mild integrability conditions on $K$, it can be shown that $|\lambda_{h,\alpha} - \lambda_\alpha| = O(h^\beta)$, uniformly in $\alpha$.

The following lemma bounds the deviation of $|\widehat{\lambda}_{h,\alpha,X} - \lambda_{h,\alpha}|$.

**Lemma 5** *Assume that the true density satisfies (A0)-(A1) and the density level sets of $p_h$ corresponding to probability content $\alpha$ satisfy (B3). Then, for any $0 < h \leq H$, any $\varepsilon < \eta_0 - 1/n$, and all $n \geq n(\varepsilon, h)$,*

$$\mathbb{P}_X\left(|\widehat{\lambda}_{h,\alpha,X} - \lambda_{h,\alpha}| \geq \varepsilon(A\kappa_3 + 1) + A\kappa_3/n\right) \leq K_1 e^{-K_2 n h^d \varepsilon^2} + 8ne^{-n\varepsilon^2/32}, \tag{5}$$

*where A is the Lipschitz constant and $\kappa_3$ is the constant in (B3).*

Using Lemma 4 and Lemma 5, we immediately obtain the following bound on the deviation of the estimated level $\widehat{\lambda}_{h,\alpha,X}$ from the true density level $\lambda_\alpha$ corresponding to probability content $\alpha$.

**Corollary 6** *Under the same conditions of Lemma 5,*

$$\mathbb{P}_X\left(|\widehat{\lambda}_{h,\alpha,X} - \lambda_\alpha| \geq ADh + \varepsilon(A\kappa_3 + 1) + A\kappa_3/n\right) \leq K_1 e^{-K_2 nh^d \varepsilon^2} + 8ne^{-n\varepsilon^2/32}.$$

We now study the performance of the level set estimator indexed by probability content using the following loss function

$$\mathcal{L}^*(h, X, \alpha) = P(M(\alpha)\Delta\widehat{M}_{h,X}(\alpha)) = \int_{M(\alpha)\Delta\widehat{M}_{h,X}(\alpha)} p(u)du.$$

**Theorem 7** *Assume that the density $p$ satisfies conditions (A0) and (A1) and the level set of $p_h$ indexed by probability content $\alpha$ satisfies (B3). For any sequence $h_n = \omega((\log n/n)^{1/d})$, let*

$$\varepsilon_n = \sqrt{\frac{K_3 \log n}{nh_n^d}}$$

*and set*

$$C_{1,n} = ADh_n + \varepsilon_n, \quad C_{2,n} = ADh_n + (A\kappa_3 + 1)\varepsilon_n + A\kappa_3/n$$

*and*

$$r_{h_n,\varepsilon_n,\alpha} = P\left(\{u\colon |p(u) - \lambda_\alpha| \leq C_{1,n} + C_{2,n}\}\right).$$

*Then, for $h_n = \omega((\log n/n)^{1/d})$ and $h_n \leq H$, we have for all $n \geq n(n_0, \eta_0, K_3, d)$,*

$$\mathbb{P}_X(\mathcal{L}^*(h_n, X, \alpha) \leq r_{h_n,\varepsilon_n,\alpha}) \geq 1 - \frac{2}{n}.$$

*In particular, if assumption (A2) also holds for density level $\lambda_\alpha$, then, for all $n \geq n(n_0, \eta_0, K_3)$,*

$$\mathbb{P}_X\left(\mathcal{L}^*(h_n, X, \alpha) \leq \kappa_2(C_{1,n} + C_{2,n})\right) \geq 1 - \frac{2}{n}.$$

**Corollary 8** *The value of $h$ that minimizes the upper bound on $\mathcal{L}$ is*

$$h^*_{n,\alpha} = c\left(\frac{n}{\log n}\right)^{-\frac{1}{d+2}}$$

*where $c > 0$ is a constant.*

## 4. Stability

The loss $\mathcal{L}$ is a useful theoretical measure of clustering accuracy. Balancing the terms in the upper bound on the loss gives an indication of the optimal scaling behavior of $h$. But estimating the loss is difficult and the value of the constant $c$ in the expression for $h^*_n$ is unknown. Thus, in practice, we need an alternative method to determine $h$. Instead of minimizing the loss, we consider using the stability of $\widehat{L}_{h,X}(\lambda)$ and $\widehat{\mathcal{T}}_h$ to choose $h$. As we discussed in the introduction, stability ideas have been used for clustering before. But the behavior of stability measures can be quite complicated. 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. On the other hand, Rinaldo and Wasserman (2010) showed that, for density-based clustering, stability-based methods can sometimes lead to good results. This motivates us to take a deeper look at stability for density clustering. In this section, we investigate two measures of cluster stability.

The first measure of cluster stability we analyze is the *level set stability*, which we denote, for a fixed density level $\lambda$ and a varying bandwidth value $h$, with $\Xi_{\lambda,n}(h)$. Assuming for convenience that the sample size is $3n$, we randomly split the data into three pieces $(X,Y,Z)$ each of size $n$. Let $\widehat{p}_{h,X}$ be the density estimator constructed from $X$ and $\widehat{p}_{h,Y}$ be the density estimator constructed from $Y$. The sample instability statistic is

$$\Xi_{\lambda,n}(h) = \widehat{P}_Z(\widehat{L}_{h,X}(\lambda)\Delta\widehat{L}_{h,Y}(\lambda)), \tag{6}$$

where $\widehat{P}_Z$ denote the empirical measure induced by $Z$. The measure $\Xi_{\lambda,n}(h)$ is the stability of a fixed level set, as a function of $h$. We will see that $\Xi_n$ has surprisingly complex behavior. See Figure 2. First of all, $\Xi_n(0) = 0$. This is an artifact and is due to the fact that the level sets get small as $h \to 0$. As $h$ increases, $\Xi_{\lambda,n}(h)$ first increases and then gets smaller. Once it gets small enough, the level sets have become stable and we have reached a good value of $h$. However, after this point, $\Xi_{\lambda,n}(h)$ continues to rise and fall. The reason is that, as $h$ gets larger, $p_h(x)$ decreases. Every time we reach a value of $h$ such that a mode of $p_h$ has height $\lambda$, $\Xi_{\lambda,n}(h)$ will increase. $\Xi_{\lambda,n}(h)$ is thus a non-monotonic function whose mean and variance become large at particular values of $h$. This behavior will be described explicitly in the theory and simulations that follow. As a practical matter, since $\Xi_{\lambda,n}(h)$ vanishes for very small values of $h$, we recommend to exclude all values of $h$ before the first local maximum of $\Xi_{\lambda,n}(h)$. Then, a reasonable choice of $h$ is the smallest value $h^*$ for which $\Xi_{\lambda,n}(h)$ remains less than some maximal pre-specified probability value $\beta$ for the empirical instability, such as 5% or 10%, for all $h \geq h^*$. The parameter $\beta$ is an entirely subjective quantity to be chosen by the practitioner, akin to the type-I-error parameter in standard hypothesis testing, and quantifies the maximal amount of empirical instability that one is willing to accept.

The second measure of cluster stability we consider is the *total variation* stability, denoted, for a varying value of the bandwidth $h$, as $\Gamma_n(h)$. Assuming again for simplicity that the sample is of size $2n$, we randomly split the data into two parts $(X,Y)$ of equal sizes $n$. Then, for a given bandwidth $h$, we compute separately on each of the two samples $X$ and $Y$ the kernel density estimates $\widehat{p}_{h,X}$ and $\widehat{p}_{h,Y}$, respectively. The total variation stability is defined to be the quantity

$$\Gamma_n(h) \equiv \sup_{B\in\mathcal{B}} \left| \int_B \widehat{p}_{h,X}(u)du - \int_B \widehat{p}_{h,Y}(u)du \right| = \frac{1}{2}\int |\widehat{p}_{h,X}(u) - \widehat{p}_{h,Y}(u)|\, du, \tag{7}$$

where the supremum is over all Borel sets $B$. Note that the total variation stability is a function of $h$. Unlike the level set stability, the total variation stability is a global measure of cluster stability in the sense that it takes into account the difference between $\widehat{p}_{h,X}$ and $\widehat{p}_{h,Y}$ overall all measurable sets, not just over the level sets. Thus, total variation stability is a much stronger notion of cluster stability. In fact, when $\Gamma_n(h)$ is small, the whole cluster tree is stable. It turns out that the behavior of $\Gamma_n(h)$ is much simpler. It is monotonically decreasing as a function of $h$. In this case we recommend choosing $h$ to be the smallest bandwidth value $h^*$ for which the instability is no larger than a pre-specified probability values $\beta \in (0,1)$, that is $\Gamma_n(h^*) \leq \beta$.

The motivation for choosing the bandwidth parameter $h$ in the way described above is as follows. We cannot estimate loss exactly. But we can use the instability to estimate variability. Our choice of

$h$ corresponds to making the bias as small as possible while maintaining control over the variability. This is very much in the spirit of the Neyman-Pearson approach to hypothesis testing where one tries to make the power of a test as large as possible while controlling the probability of false positives. Put another way, $P_h = P * \mathbb{K}_h$ has a blurred version of the shape information in $P$. *We are choosing the smallest h such that the shape information in $P_h$ can be reliably recovered.*

Before getting into the details, which turn out to be somewhat technical, here is a very loose description of the results. For large $h$, $\Gamma_n(h) \approx 1/\sqrt{nh^d}$. On the other hand, $\Xi_{\lambda,n}(h)$ tends to oscillate up and down corresponding to the presence of modes of the density. In regions where it is small, it also behaves like $1/\sqrt{nh^d}$.

### 4.1 Level Set Stability

For the analysis of the level set stability we focus on a single level set indexed by some density level value $\lambda \geq 0$. Consider two independent samples $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$ and set

$$\xi_{\lambda,n}(h) \quad = \quad \mathbb{E}_{XY}\left(P\left(\widehat{L}_{h,X}(\lambda)\Delta\widehat{L}_{h,Y}(\lambda)\right)\right).$$

The quantity $\xi_{\lambda,n}(h)$ measures the expected disagreement between level sets based on two samples as a function of the bandwidth $h$.

The definition of $\xi_{\lambda,n}$ depends on $P$ which, of course, we do not know. To estimate $\xi_{\lambda,n}(h)$ we use the sample instability statistic defined above in Equation (6), where it was assumed for simplicity that the sample size is $3n$ and the data were randomly split into three pieces $(X,Y,Z)$ each of size $n$. It is immediate to see that the expectation of the sample instability statistic is precisely $\xi_{\lambda,n}(h)$, that is

$$\xi_{\lambda,n}(h) = \mathbb{E}_{X,Y,Z}[\Xi_{\lambda,n}(h)].$$

Note that since we are using the empirical distribution $\widehat{P}_Z$, the sample instability can be rewritten as

$$\begin{aligned}\Xi_{\lambda,n}(h) &= \frac{1}{n}\sum_{i=1}^{n} I(Z_i \in (\widehat{L}_{h,X}(\lambda)\Delta\widehat{L}_{h,Y}(\lambda))) \\ &= \frac{1}{n}\sum_{i=1}^{n} I(\text{sign}(\widehat{p}_{h,X}(Z_i) - \lambda) \neq \text{sign}(\widehat{p}_{h,Y}(Z_i) - \lambda)).\end{aligned}$$

The above equation show that, for a fixed $\lambda$, $\Xi_{\lambda,n}(h)$ is obtained as the fraction of the observations in $Z$ where $\widehat{p}_{h,X}(Z_i) < \lambda < \widehat{p}_{h,Y}(Z_i)$ or $\widehat{p}_{h,X}(Z_i) > \lambda > \widehat{p}_{h,Y}(Z_i)$. This representation is closely tied to the use of the *sample level sets* to construct the cluster tree (Stuetzle and Nugent, 2009) where each level set is represented only by the observations associated with its connected components rather than the feature space. Using the empirical distribution $\widehat{P}_Z$ also removes the need to determine the exact shape of the level sets of the density estimate. The top graph of Figure 2 shows the sample instability as a function of $h$ for $\lambda = 0.09$ for our example distribution depicted in Figure 1. Note that the instability initially drops and then oscillates before dropping to zero at $h = 7.08$, indicating the multi-modality seen in Figure 1. More discussion of this example is in Section 5.

As mentioned at the end of section 2.3, for values of $h \ll \left(\frac{\log n}{n}\right)^{1/d}$, the kernel density estimate $\widehat{p}_h$ is no longer a reliable estimate of $p_h$. The following simple but important boundary properties of $\Xi_n$ and $\xi$ describes the behavior of the empirical and expected instability when $h$ is either too small or too large.

**Lemma 9** *For fixed n and $\lambda > 0$,*

$$\lim_{h\to 0} \xi_{\lambda,n}(h) = \lim_{h\to\infty} \xi_{\lambda,n}(h) = \lim_{h\to 0} \Xi_{\lambda,n}(h) = \lim_{h\to\infty} \Xi_{\lambda,n}(h) = 0,$$

*where the last two limits occurs almost surely. In particular, $\xi_{\lambda,n}(h) = O(h^d)$, as $h \to 0$.*

We now study the behavior of the mean function $\xi_{\lambda,n}(h)$. Let $u \in \mathbb{R}^d$, $h > 0$ and $\varepsilon > 0$, and define

$$\pi_h(u) = \mathbb{P}_X(\widehat{p}_{h,X}(u) > \lambda) \quad \text{and} \quad U_{h,\varepsilon} = \{u\colon |p_h(u) - \lambda| < \varepsilon\}. \tag{8}$$

**Theorem 10** *Let $u \in \mathbb{R}^d$, $h > 0$ and $\varepsilon > 0$.*

1. *The following identity holds:*

$$\xi_{\lambda,n}(h) = 2\int_{\mathbb{R}^d} \pi_h(u)(1-\pi_h(u))dP(u).$$

2. *Also, for all $n \geq n(\varepsilon, h)$,*

$$r_{h,\varepsilon}\, \underline{A}_{h,\varepsilon} \leq \xi_{\lambda,n}(h) \leq r_{h,\varepsilon}\, \overline{A}_{h,\varepsilon} + 2K_1 e^{-K_2 n h^d \varepsilon^2},$$

*where $r_{h,\varepsilon} = P(U_{h,\varepsilon})$,*

$$\overline{A}_{h,\varepsilon} = \sup_{u\in U_{h,\varepsilon}} 2\pi_h(u)(1-\pi_h(u))$$

*and*

$$\underline{A}_{h,\varepsilon} = \inf_{u\in U_{h,\varepsilon}} 2\pi_h(u)(1-\pi_h(u)).$$

Part 2 of the previous theorem implies that the behavior of $\xi$ is essentially captured by the behavior of the probability content $r_{h,\varepsilon}$. This quantity is, in general, a complicated function of both $h$ and $\varepsilon$. While it is easy to see that, for fixed $h$ and a sufficiently well-behaved density $p$, $r_{h,\varepsilon} \to 0$ as $\varepsilon \to 0$, for fixed $\varepsilon$, $r_{h,\varepsilon}$ can instead be a non-monotonic function of $h$. See, for example, the bottom right plot in Figure 3, which displays the values $r_{h,\varepsilon}$ as a function of $h \in [0, 4.5]$ and for $\varepsilon$ equal to 0.02, 0.05 and 0.1 for the mixture density of Figure 1. In particular, the fluctuations of $r_{h,\varepsilon}$ as a function of $h$ are related to the values of $h$ for which the critical points of $p_h$ are in the interval $[\lambda - \varepsilon, \lambda + \varepsilon]$. The main point to notice is that $r_{h,\varepsilon}$ is a complicated, non-monotonic function of $h$. This explains why $\Xi_n(h)$ is non-monotonic in $h$.

We now provide an upper and lower bound on the values of $\overline{A}_{h,\varepsilon}$ and $\underline{A}_{h,\varepsilon}$, respectively, under the simplifying assumption that $K$ is the spherical kernel. Notice that, while $\overline{A}_{h,\varepsilon}$ remains bounded away from $\infty$ for any sequence $\varepsilon_n \to 0$ and $h_n = \omega(n^{-1/d})$, the same is not true for $\underline{A}_{h,\varepsilon}$, which remains bounded away from 0 as long as $\varepsilon_n = \Theta(\frac{1}{nh_n^d})$ and $h_n = \omega(n^{-1/d})$.

**Lemma 11** *Assume that $K$ is the spherical kernel and let $0 < \varepsilon \leq \lambda/2$. For a given $\delta \in (0,1)$, let*

$$h(\delta, \varepsilon) = \sup\Big\{h\colon \sup_{u\in U_{h,\varepsilon}} P(B(u,h)) \leq 1-\delta\Big\}.$$

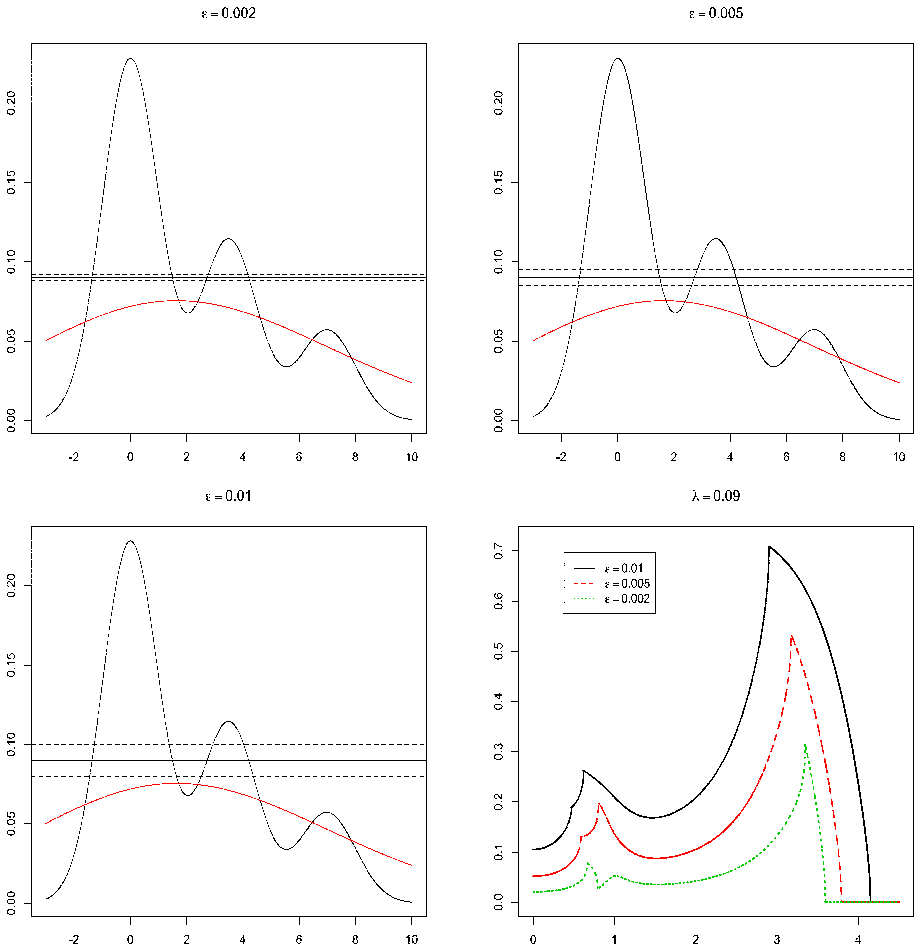


Figure 3: Top plots and left bottom plot: two densities $p_h$ corresponding to the mixture distribution of Figure 1 for $h = 0$, the true density (in black) and $h = 4.5$ (in red); the horizontal lines indicate the level set value of $\lambda = 0.09$, $\lambda + \varepsilon$ and $\lambda - \varepsilon$, for $\varepsilon$ equal to 0.02, 0.05 and 0.1. Right bottom plot: probability content values $r_{h,\varepsilon}$ as a function of $h \in [0, 4.5]$ for the three values of $\varepsilon$.

*Then, for all* $h \leq h(\delta, \varepsilon)$,

$$\overline{A}_{h,\varepsilon} \leq 2\left(1 - \Phi\left(-\sqrt{nh^d}\varepsilon\frac{2v_d}{3\lambda}\right) + \frac{C(\delta,\lambda)}{\sqrt{nh^d}}\right)^2,$$

*and*

$$\underline{A}_{h,\varepsilon} \geq 2\left(1 - \Phi\left(\sqrt{nh^d}\varepsilon\frac{2v_d}{\delta\lambda}\right) - \frac{C(\delta,\lambda)}{\sqrt{nh^d}}\right)^2,$$

*where* $\Phi$ *denote the cumulative distribution function of a standard normal random variable and*

$$C(\delta,\lambda) = \frac{33}{4}\sqrt{\frac{2}{\delta v_d \lambda}}.$$

The dips in Figure 2 correspond to values for which $p_h$ does not have a mode at height $\lambda$. In this case, (B2) holds and we have $r_{h,\varepsilon} = \Theta(\varepsilon)$. Now choosing $\varepsilon \approx \sqrt{\log n/(nh^d)}$ for the upper bound and $\varepsilon \approx \sqrt{1/(nh^d)}$ for the lower bound, we have that $\overline{A}_{h,\varepsilon}$ and $\underline{A}_{h,\varepsilon}$ are bounded, and the theorem yields

$$\sqrt{\frac{C_1}{nh^d}} \leq \xi_{\lambda,n}(h) \leq \sqrt{\frac{C_2 \log n}{nh^d}}.$$

Next we investigate the extent to which $\Xi_{\lambda,n}(h)$ is concentrated around its mean $\xi_{\lambda,n}(h) = \mathbb{E}[\Xi_{\lambda,n}(h)]$. We first point out that, for any fixed $h$, the variance of the instability can be bounded by $\xi_{\lambda,n}(h)(1/2 - \xi_{\lambda,n}(h))$.

**Lemma 12** *For any $h > 0$,*

$$\mathrm{Var}[\Xi_{\lambda,n}(h)] \leq \xi_{\lambda,n}(h)\left(\frac{n+1}{2n} - \xi_{\lambda,n}(h)\right) \approx \xi_{\lambda,n}(h)\left(\frac{1}{2} - \xi_{\lambda,n}(h)\right).$$

The previous results highlight the interesting feature that the empirical instability will be less variable around the values of $h$ for which the expected instability is very small (close to 0) or very large (close to $1/2$).

**Lemma 13** *Suppose that $h > 0$, $\varepsilon > 0$, $\eta \in (0,1)$ and $t > 0$ are such that*

$$t(1-\eta) \geq r_{h,\varepsilon} + 2K_1 e^{-K_2 n \varepsilon^2 h^d}, \tag{9}$$

*where $r_{h,\varepsilon} = P(U_{h,\varepsilon})$. Then, for all $n \geq n(\varepsilon, h)$,*

$$\mathbb{P}_{X,Y,Z}\left(\left|\Xi_{\lambda,n}(h) - \xi_{\lambda,n}(h)\right| > t\right) \leq e^{-ntC_\eta} + 2K_1 e^{-nK_2 h^d \varepsilon^2}$$

*where*

$$C_\eta = 9(1-\eta)\left(\frac{3-2\eta}{3(1-\eta)} - \sqrt{\frac{3-\eta}{3(1-\eta)}}\right).$$

### 4.2 Stability of Level Sets Indexed by Probability Content

As in the fixed-$\lambda$ case, we assume for simplicity that the sample has size $3n$ and split it equally in three parts: $X$, $Y$ and $Z$. We now define the fixed-$\alpha$ instability as

$$\Xi_{\alpha,n}(h) = \widehat{P}_Z(\widehat{M}_{h,X}(\alpha) \Delta \widehat{M}_{h,Y}(\alpha)),$$

where

$$\widehat{M}_{h,X}(\alpha) = \{x\colon \widehat{p}_{h,X}(x) > \widehat{\lambda}_{h,\alpha,X}\},$$

with $\widehat{\lambda}_{h,\alpha,X}$ estimated as in (4) using the points in $X$; we similarly estimate $\widehat{M}_{h,Y}(\alpha)$. As before, $\widehat{P}_Z$ denote the empirical measure arising from $Z$. Again, we use the observations to represent $\widehat{M}_{h,X}$, $\widehat{M}_{h,Y}$ as done for $\Xi_{\lambda,n}(h)$ for a fixed $\lambda$. Examples of $\Xi_{\alpha,n}(h)$ as a function of $h, \alpha$ can be seen in Section 5.

The expected instability is

$$\xi_{\alpha,n}(h) = \mathbb{E}_{X,Y,Z}[\Xi_{\alpha,n}(h)].$$

We begin by studying the behavior of the expected instability.

**Theorem 14** *Let $u \in \mathbb{R}^d$, $h > 0$ and $\varepsilon > 0$, and set*

$$\pi_{h,\alpha}(u) = \mathbb{P}_X(\widehat{p}_{h,X}(u) > \widehat{\lambda}_{h,\alpha,X}) \quad \text{and } U_{h,2\varepsilon,\alpha} = \{u\colon |p_h(u) - \lambda_{\alpha,h}| \le 2\varepsilon\}.$$

1. *The expected instability can be expressed as*

$$\xi_{\alpha,n}(h) = \mathbb{E}_{X,Y,Z}[\Xi_{\alpha,n}(h)] = 2\int_{\mathbb{R}^d} \pi_{h,\alpha}(u)(1-\pi_{h,\alpha}(u))dP(u).$$

2. *Let $\varepsilon < \eta_0 - 1/n$ and $\widetilde{\varepsilon} = \varepsilon(A\kappa_3 + 1) + A\kappa_3/n$. Then, for all $n \ge n(\varepsilon, h)$,*

$$P(U_{h,2\widetilde{\varepsilon},\alpha})\underline{A}_{h,\varepsilon,\alpha} \le \xi_{\lambda,n}(h) \le P(U_{h,2\widetilde{\varepsilon},\alpha})\overline{A}_{h,\varepsilon,\alpha} + 4K_1 e^{-K_2 nh^d\varepsilon^2} + 16ne^{-n\varepsilon^2/32},$$

*where*

$$\overline{A}_{h,\varepsilon,\alpha} = \sup_{u \in U_{h,2\widetilde{\varepsilon},\alpha}} 2\pi_{h,\alpha}(u)(1-\pi_{h,\alpha}(u))$$

*and*

$$\underline{A}_{h,\varepsilon,\alpha} = \inf_{u \in U_{h,2\widetilde{\varepsilon},\alpha}} 2\pi_{h,\alpha}(u)(1-\pi_{h,\alpha}(u)).$$

3. *Assume in addition that $K$ is the spherical kernel and that $\widetilde{\varepsilon} \le \inf_h \frac{\lambda_{h,\alpha}}{4}$. For a given $\delta \in (0,1)$, let*

$$h(\delta,\varepsilon,\alpha) = \sup\Big\{h\colon \sup_{u \in U_{h,\widetilde{\varepsilon},\alpha}} P(B(u,h)) \le 1-\delta\Big\}.$$

*Then, for all $h \le h(\delta,\varepsilon,\alpha)$,*

$$\overline{A}_{h,\varepsilon,\alpha} \le 2\left(1 - \Phi\left(-3\sqrt{nh^d}\widetilde{\varepsilon}\frac{2\nu_d}{3\lambda_{h,\alpha}}\right) + \frac{C(\delta,\lambda_{h,\alpha})}{\sqrt{nh^d}} + 4K_1 e^{-K_2 nh^d\varepsilon^2} + 16ne^{-n\varepsilon^2/32}\right)^2,$$

*and*

$$\underline{A}_{h,\varepsilon,\alpha} \ge 2\left(1 - \Phi\left(3\sqrt{nh^d}\widetilde{\varepsilon}\frac{2\nu_d}{\delta\lambda_{h,\alpha}}\right) - \frac{C(\delta,\lambda_{h,\alpha})}{\sqrt{nh^d}} - 4K_1 e^{-K_2 nh^d\varepsilon^2} - 16ne^{-n\varepsilon^2/32}\right)^2,$$

*where $\Phi$ denote the cumulative distribution function of a standard normal random variable and*

$$C(\delta,\lambda_{h,\alpha}) = \frac{33}{4}\sqrt{\frac{2}{\delta\nu_d\lambda_{h,\alpha}}}.$$

As for the fluctuations of $\Xi_{\alpha,n}(h)$ around its mean, we can easily obtain a result similar to the one we obtain in Lemma 13.

**Lemma 15** *Let $h > 0$, $\varepsilon > 0$, $\eta \in (0,1)$ and $t$ be such that*

$$t(1-\eta) \ge r_{h,\varepsilon,\alpha} + 4K_1 e^{-K_2 nh^d\varepsilon^2} + 16ne^{-n\varepsilon^2/32},$$

*where $r_{h,\varepsilon,\alpha} = P(\{u\colon |p_h(u) - \lambda_{h,\alpha}| \le 2\widetilde{\varepsilon}\})$, with $\widetilde{\varepsilon} = \varepsilon(A\kappa_3+1) + A\kappa_3/n$. Then, for all $n \ge n(\varepsilon,h)$,*

$$\mathbb{P}_{X,Y,Z}(|\Xi_{\alpha,n}(h) - \xi_{\alpha,n}(h)| > t) \le e^{-ntC_\eta} + 4K_1 e^{-K_2 nh^d\varepsilon^2} + 16ne^{-n\varepsilon^2/32},$$

*where*

$$C_\eta = 9(1-\eta)\left(\frac{3-2\eta}{3(1-\eta)} - \sqrt{\frac{3-\eta}{3(1-\eta)}}\right).$$

The proof is basically the same as the proof of Lemma 13, except that we have to restrict our analysis to the event described in Equation (29). We omit the details.

### 4.3 Stability for Density Cluster Trees

The stability properties of the cluster tree can be easily derived from the results we have established so far. To this end, for a fixed $h > 0$, define the level set of $p_h$

$$L_h(\lambda) = \{u : p_h(u) > \lambda\}$$

and recall its estimator based on the kernel density estimator $\widehat{p}_{h,X}$:

$$\widehat{L}_{h,X}(\lambda) = \{u : \widehat{p}_{h,X}(u) > \lambda\}.$$

Let $N_h(\lambda)$, $\widehat{N}_{h,X}(\lambda)$ be the number of connected components of the sets $L_h(\lambda)$ and $\widehat{L}_{h,X}(\lambda)$, respectively. Notice that $\widehat{L}_{h,X}(\lambda)$ is a random set. Also, denote with $C_1, \ldots, C_{N_h(\lambda)}$ and $\widehat{C}_1, \ldots, \widehat{C}_{\widehat{N}_{h,X}(\lambda)}$ the connected components of $L_h(\lambda)$ and $\widehat{L}_{h,X}(\lambda)$, respectively.

When building cluster trees, the value of the bandwidth $h$ is kept fixed and the values of the level $\lambda$ vary instead. It has been observed empirically (see, for instance Stuetzle and Nugent, 2009) that the uncertainty of cluster tree estimators depend on the particular value of $\lambda$ at which the tree is observed. In order to characterize the behavior of the cluster tree, we propose the following definition, which formalize the case in which the clusters $C_1, \ldots, C_{N_h(\lambda')}$ persist for each $\lambda'$ in a neighborhood of $\lambda$.

**Definition 16** *A level set value $\lambda$ is $(h, \varepsilon)$-stable, with $\varepsilon > 0$ and $h > 0$, if*

$$N_h(\lambda) = N_h(\lambda'), \quad \forall \lambda' \in (\lambda - \varepsilon, \lambda + \varepsilon)$$

*and, for any $\lambda - \varepsilon < \lambda_1 < \lambda_2 < \lambda + \varepsilon$,*

$$C_i(\lambda_2) \subseteq C_i(\lambda_1), \quad \forall i = 1, \ldots, N_h(\lambda).$$

If the level $\lambda$ is $(h, \varepsilon)$-stable, then the cluster tree estimate at level $\lambda$ is an accurate estimate of the true cluster tree, in a sense made precise by the following result, whose proof follows easily from the proofs of our previous results and Lemma 2 in Rinaldo and Wasserman (2010).

**Lemma 17** *If $\lambda$ is $(h, \varepsilon)$-stable, then, for all large $n \geq n(\varepsilon, \lambda)$, with probability at least $1 - \frac{1}{n}$,*

1. *$N_h(\lambda) = \widehat{N}_{h,X}(\lambda)$;*
2. *there exists a permutation $\sigma$ on $\{1, \ldots, N_h(\lambda)\}$ such that, for every connected component $C_j$ of $L_h(\lambda - \varepsilon)$ there exists one $\widehat{C}_{\sigma(j)}$ for which*

$$C_j \subseteq \widehat{C}_{\sigma(j)};$$

3. *$P(\widehat{L}_{h,X}(\lambda) \Delta L_h(\lambda)) \leq P(\{u : |p_h(u) - \lambda| < \varepsilon\})$.*

**Remarks.**

1. If $p_h$ is smooth (which is the case if, for instance, the kernel or $p$ are smooth), the values of $\lambda$ which are not $(h,\varepsilon)$-stable are values for which the set $U_{\lambda',h,\varepsilon}$ contains critical points of $p_h$, that is
$$\inf_{u\in U_{\lambda',h,\varepsilon}} \|\nabla p_h(u)\| = 0 \quad \text{for some } \lambda' \in (\lambda-\varepsilon,\lambda+\varepsilon),$$
where $\nabla p_h$ denotes the gradient of $p$. For those values, the probability of $N_h(\lambda) \neq \widehat{N}_{h,X}(\lambda)$ can be quite large, since the set $\widehat{L}_{h,X}\Delta L_h(\lambda)$ may have a relatively large $P$-mass.

2. Conversely, if $p_h$ is smooth (which is the case if, for instance, the kernel or $p$ are smooth) and $\inf_{u\in U_{\lambda,h,\varepsilon}} \|\nabla p_h(u)\| > \delta$, then $\lambda$ is $(h,\varepsilon)$-stable for a small enough $\varepsilon$.

The above result has a somewhat limited practical value, because the notion of a $(h,\varepsilon)$-stable $\lambda$ depends on the unknown density $p_h$. In order to get a better sense of which $\lambda$'s are $(h,\varepsilon)$-stable or not, we once again resort to evaluate the instability of the clustering solution via data splitting. In fact, essentially all of our previous results about instability from section 4.1 carry over to these new settings by treating $h$ fixed and letting $\lambda$ vary. To express this changes explicitly, we will adopt a slightly different notation for quantities we have already considered. In particular, we let

$$\begin{aligned} U_{\lambda,\varepsilon} &= \{u\colon |p_h(u)-\lambda| < \varepsilon\}, \\ r_{\lambda,\varepsilon} &= P(U_{\lambda,\varepsilon}), \\ \pi_\lambda(u) &= \mathbb{P}_X(\widehat{p}_{h,X}(u) > \lambda), \\ \overline{A}_{\lambda,\varepsilon} &= \sup_{u\in U_{\lambda,\varepsilon}} 2\pi_\lambda(u)(1-\pi_\lambda(u)) \end{aligned}$$

and

$$\underline{A}_{\lambda,\varepsilon} = \inf_{u\in U_{\lambda,\varepsilon}} 2\pi_\lambda(u)(1-\pi_\lambda(u)).$$

We divide the sample size into three distinct groups, $X$, $Y$ and $Z$, of equal sizes $n$. For a fixed bandwidth $h$, we define the instability of the density cluster tree as the random function $T_{h,n}\colon \mathbb{R}_{\geq 0} \mapsto [0,1]$ given by

$$\lambda \to \widehat{\mathbb{P}}_Z(\widehat{L}_{h,X}(\lambda)\Delta\widehat{L}_{h,Y}(\lambda))$$

and denote its expectation by

$$\tau_{h,n}(\lambda) = \mathbb{E}_{X,Y,Z}[T_{h,n}(\lambda)].$$

For any fixed $h$, the behavior of $T_{h,n}(\lambda)$ and $\tau_{h,n}(\lambda)$ is essentially governed by $r_{\lambda,\varepsilon}$. The following result describes some of the properties of the density tree instability. We omit its proof, because it relies essentially on the same arguments from the proofs of the results described in section 4.1.

**Corollary 18**

1. *For any $\lambda > 0$, the expected cluster tree instability can be expressed as*
$$\tau_{h,n}(\lambda) = 2\int \pi_\lambda(u)(1-\pi_\lambda(u))dP(u).$$

2. *For any $\varepsilon > 0$ and $\lambda > 0$,*
$$\underline{A}_{\lambda,\varepsilon} r_{\lambda,\varepsilon} \leq \tau_{h,n}(\lambda) \leq \underline{A}_{\lambda,\varepsilon} r_{\lambda,\varepsilon} + 2K_1 e^{-K_2 n h^d \varepsilon^2},$$
*for all $n$ large enough.*

3. *Assume that $K$ is the spherical kernel. For any $\lambda > 0$, let $0 < \varepsilon \leq \frac{\lambda}{2}$ and let*

$$\delta = 1 - \sup_u P(B(u,h)).$$

*Then,*

$$\overline{A}_{\lambda,\varepsilon} \leq 2\left(1 - \Phi\left(-\sqrt{nh^d}\varepsilon\frac{2v_d}{3\lambda}\right) + \frac{C(\delta,\lambda)}{\sqrt{nh^d}}\right)^2,$$

*and*

$$\underline{A}_{\lambda,\varepsilon} \geq 2\left(1 - \Phi\left(\sqrt{nh^d}\varepsilon\frac{2v_d}{\delta\lambda}\right) - \frac{C(\delta,\lambda)}{\sqrt{nh^d}}\right)^2,$$

*where $\Phi$ denote the cumulative distribution function of a standard normal random variable and*

$$C(\delta,\lambda) = \frac{33}{4}\sqrt{\frac{2}{\delta v_d \lambda}}.$$

4. *For any $h > 0$, $\varepsilon > 0$, $\eta \in (0,1)$ let $t$ by such that*

$$t(1-\eta) \geq r_{\lambda,\varepsilon} + 2K_1 e^{-K_2 n\varepsilon^2 h^d},$$

*Then, for all $n \geq n(\varepsilon,h)$,*

$$\mathbb{P}_{X,Y,Z}\left(|T_{h,n}(\lambda) - \tau_{h,n}(\lambda)| > t\right) \leq e^{-ntC_\eta} + 2K_1 e^{-nK_2 h^d \varepsilon^2}.$$

*with*

$$C_\eta = 9(1-\eta)\left(\frac{3-2\eta}{3(1-\eta)} - \sqrt{\frac{3-\eta}{3(1-\eta)}}\right).$$

Collectively, the results above results show that the cluster tree of $p_h$ can be estimated more accurately for values of $\lambda$ for which the quantity $r_{\lambda,\varepsilon}$ remain small, with $\varepsilon$ a term vanishing in $n$. In particular, the level sets $\lambda$ with larger instability are then the ones that are close to a critical level of $p_h$ or for which the gradient of $p_h$ is not defined, vanishes of has infinite norm for some points in $\{x\colon\ p_h(x) = \lambda\}$. This suggests that the sample complexity for accurately reconstructing of the cluster tree may vary significantly depending on the particular level of the tree, with levels closer to a branching point exhibiting a higher degree of uncertainty and, therefore, requiring larger sample sizes.

### 4.4 Total Variation Stability

In the previous section, we established stability of the cluster tree for a fixed $h$ and all levels $\lambda$ that are $(h,\varepsilon)$-stable. A more complete measure of stability would be to establish stability of the entire cluster tree. However, it appears that this is not feasible. Here we investigate an interesting alternative: we compare the entire distribution $\widehat{p}_{h,X}$ to the entire distribution $\widehat{p}_{h,Y}$. The idea is that if these two distributions are stable over all measurable sets, then this implies it is stable over any class of subsets, including all clusters.

More precisely, we consider the stronger notion of instability corresponding to the total variation stability as defined in (7). Recall that we assume that the data have sample size $2n$ and we randomly

split them into two sets of size $n$, $X$ and $Y$, with which we compute the he kernel density estimates $\widehat{p}_{h,X}$ and $\widehat{p}_{h,Y}$, for a given value of the bandwidth $h$. Then, the total variation stability is defined as

$$\Gamma_n(h) \equiv \sup_{B\in\mathcal{B}} \left| \int_B \widehat{p}_{h,X}(u)du - \int_B \widehat{p}_{h,Y}(u)du \right| = \frac{1}{2}\int |\widehat{p}_{h,X}(u) - \widehat{p}_{h,Y}(u)|\, du$$

where where the supremum is over all Borel sets $B$ and the second equality is a standard identity. Requiring $\Gamma_n(h)$ to be small is a more demanding type of stability. In particular, $\mathcal{B}$ includes all level sets for all $\lambda$. Thus, when $\Gamma_n(h)$ is small, the entire cluster tree is stable. Note that $\Gamma_n(h)$ is easy to interpret: it is the maximum difference in probability between the two density estimators. And of course $0 \leq \Gamma_n(h) \leq 1$. The bottom graph in Figure 2 shows the total variation instability for our example distribution in Figure 1. Note that $\Gamma_n(h)$ first drops drastically as $h$ increases and then continues to smoothly decrease.

We now discuss the properties of $\Gamma_n(h)$. Note first that $\Gamma_n(h) \approx 1$ for small $h$ so the behavior as $h$ gets large is most relevant.

**Theorem 19** *Let $\mathcal{H}_n$ be a finite set of bandwidths such that $|\mathcal{H}_n| = Hn^a$, for some positive $H$ and $a \in (0,1)$. Fix a $\delta \in (0,1)$.*

1. *(Upper bound.) There exists a constant $C$ such that, for all $n \geq n_0 \equiv n_0(\delta, H, a)$, and such that $\delta > H/n$,*

$$\mathbb{P}_{X,Y}\left(\Gamma_n(h) \leq t_h \text{ for all } h \in \mathcal{H}_n\right) > 1-\delta,$$

   *where $t_h = \sqrt{\frac{C\log n}{nh^d}}$.*

2. *(Lower bound.) Suppose that $K$ is the spherical kernel and that the probability distribution $P$ satisfies the conditions*

$$a_1 h^d v_d \leq \inf_{u\in S} P(B(u,h)) \leq \sup_{u\in S} P(B(u,h)) \leq h^d v_d a_2, \quad \forall h > 0, \tag{10}$$

   *for some positive constants $a_1 < a_2$, where $S$ denotes the support of $P$. Let $h_*$ be such that $\sup_u P(B(u,h_*)) < 1-\delta$. There exists a $t$, depending on $\delta$ but not on $h$, such that, for all $h < h_*$ and for $n \geq n_0 \equiv n_0(a, a_1, a_2, h, delta)$*

$$\mathbb{P}_{X,Y}\left(\Gamma_n(h) \geq t\sqrt{\frac{1}{nh^d}}\right) > 1-\delta.$$

3. *$\Gamma_n(0) = 1$ and $\Gamma_n(\infty) = 0$.*

**Remarks.**

1. Note that the upper bound is uniform in $h$ while the lower bound is pointwise in $h$. Making the lower bound uniform is an open problem. However, if we place a nonzero lower bound on the bandwidths in $\mathcal{H}_n$ then the bound could be made uniform. The latter approach was used in Chaudhuri and Marron (2000).

2. Conditions (10) are quite standard in support set estimation. In particular, when the lower bound holds, the support $S$ is said to be *standard*. See, for instance, Cuevas and Rodríguez-Casal (2004).

In low dimensions, we can compute $\Gamma_n(h)$ by numerically evaluating the integral

$$\frac{1}{2}\int |\widehat{p}_{h,X}(u) - \widehat{p}_{h,Y}(u)|\, du.$$

In high dimensions it may be easier to use importance sampling as follows. Let $g(u) = (1/2)(\widehat{p}_{h,X}(u) + \widehat{p}_{h,Y}(u))$. Then,

$$\Gamma_n(h) = \frac{1}{2}\int \frac{|\widehat{p}_{h,X}(u) - \widehat{p}_{h,Y}(u)|}{g(u)} g(u)du \approx \frac{1}{N}\sum_{i=1}^{N} \frac{|\widehat{p}_{h,X}(U_i) - \widehat{p}_{h,Y}(U_i)|}{|\widehat{p}_{h,X}(U_i) + \widehat{p}_{h,Y}(U_i)|},$$

where $U_1, \ldots, U_N$ is a random sample from $g$. We can thus estimate $\Gamma_n(h)$ with the following algorithm:

---

1. Draw Bernoulli(1/2) random variables $Z_1, \ldots, Z_N$.
2. Draw $U_1, \ldots, U_N$ as follows:
   (a) If $Z_i = 1$: draw $X$ randomly from $X_1, \ldots, X_n$. Draw $W \sim K$. Set $U_i = X + hW$.
   (b) If $Z_i = 0$: draw $Y$ randomly from $Y_1, \ldots, Y_n$. Draw $W \sim K$. Set $U_i = Y + hW$.
3. Set
$$\widehat{\Gamma}_n(h) = \frac{1}{N}\sum_{i=1}^{N} \frac{|\widehat{p}_{h,X}(U_i) - \widehat{p}_{h,Y}(U_i)|}{|\widehat{p}_{h,X}(U_i) + \widehat{p}_{h,Y}(U_i)|}.$$

---

It is easy to see that $U_i$ has density $g$ and that $\widehat{\Gamma}_n(h) - \Gamma_n(h) = O_P(1/\sqrt{N})$ which is negligible for large $N$.

## 5. Examples

We present results for two examples where, although the dimensionality is low, estimating the connected components of the true level sets is surprisingly difficult. For the first example, we begin by illustrating how the instability changes for given values of $\lambda, \alpha$ and then split each data set 200 times to find point-wise confidence bands for $\Xi_{\lambda,n}(h)$ for fixed $\lambda, \alpha$ and for $\Gamma_n(h)$. We then present selected results for a bivariate example.

### 5.1 Instability as Function of $h$ for Fixed $\lambda$

Returning to the example distribution in Section 1, 600 observations were sampled from the following mixture of normals: $(4/7)N(0,1) + (2/7)N(3.5,1) + (1/7)N(7,1)$. The original sample is randomly split into three samples of 200. All kernel density estimates use the Epanechnikov kernel

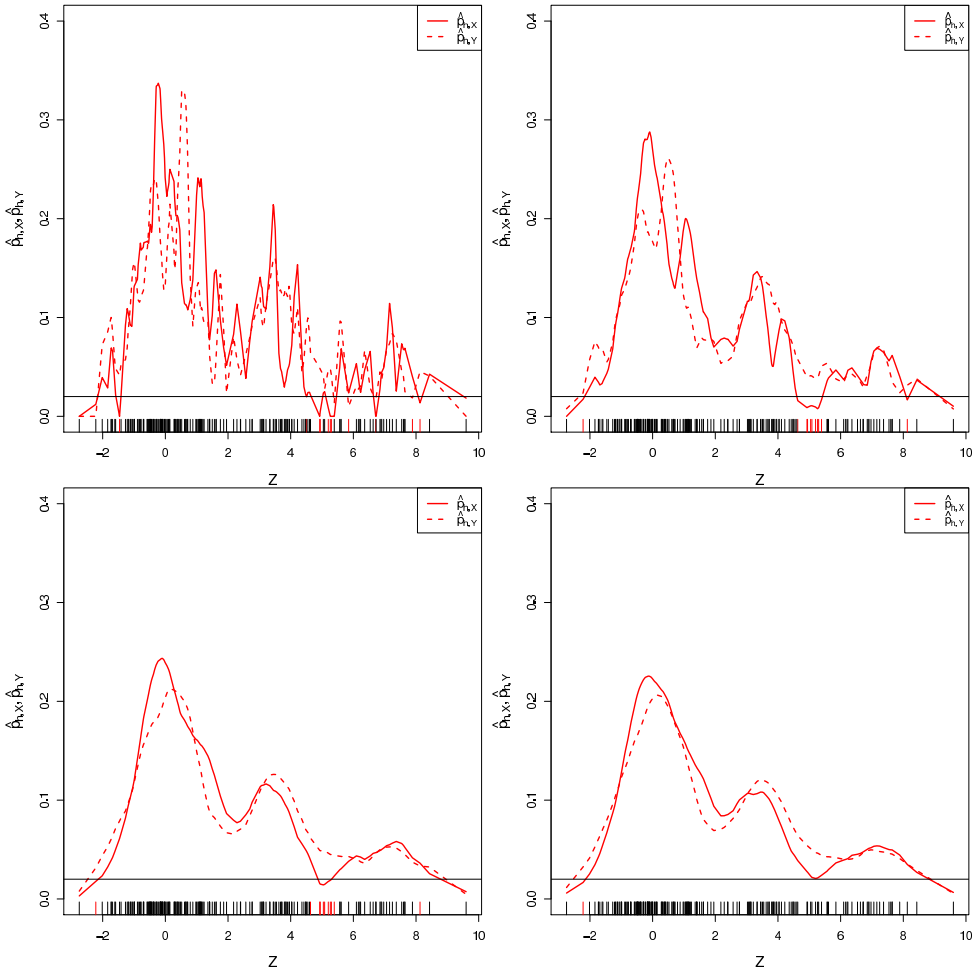


Figure 4: Comparing $\widehat{L}_{h,X}(0.02)$ and $\widehat{L}_{h,Y}(0.02)$ with $h = 0.15$ (top left), $h = 0.35$ (top right), $h = 0.75$ (bottom left) and $h = 0.95$ (bottom right) for data sampled from the mixture distribution of Figure 1. The two kernel density estimates are obtained using the $X$ sample (solid line) and the $Y$ sample (dotted line). Points in the $Z$ sample are showed as short vertical lines on the $x$-axis, and are colored in red when they belong to $\widehat{L}_{h,X}(\lambda)\Delta\widehat{L}_{h,Y}(\lambda)$.

(Scott, 1992). We examine the stability at $\lambda = 0.02$, a height at which the true density's connected components should be unambiguous, and $\lambda = 0.09$, the height used in our earlier motivating graphs.

We start by illustrating the instability for selected values of $h$ in Figures 4, 5. In each subfigure, $\widehat{p}_{h,X}, \widehat{p}_{h,Y}$ are graphed for the $Z$ set of observations. Levels $\lambda = 0.02, 0.09$ are marked respectively with a horizontal line. Those observations in $Z$ that belong to $\widehat{L}_{h,X}(\lambda)$ and not to $\widehat{L}_{h,Y}(\lambda)$ (or vice versa) are marked in red; the overall fraction of these observations is $\Xi_{\lambda,n}(h)$. In general, we can see that as $h$ increases, the number of the red $Z$ observations decreases. For $\lambda = 0.02$, note that the location that most contributes to the instability is the valley around $Z = 5$. Once $h$ is large enough to smooth this valley to have height above $\lambda = 0.02$, the instability is negligible. Turning to $\lambda = 0.09$ (Figure 5), even for larger values of $h$, the differences between the two density estimates can be

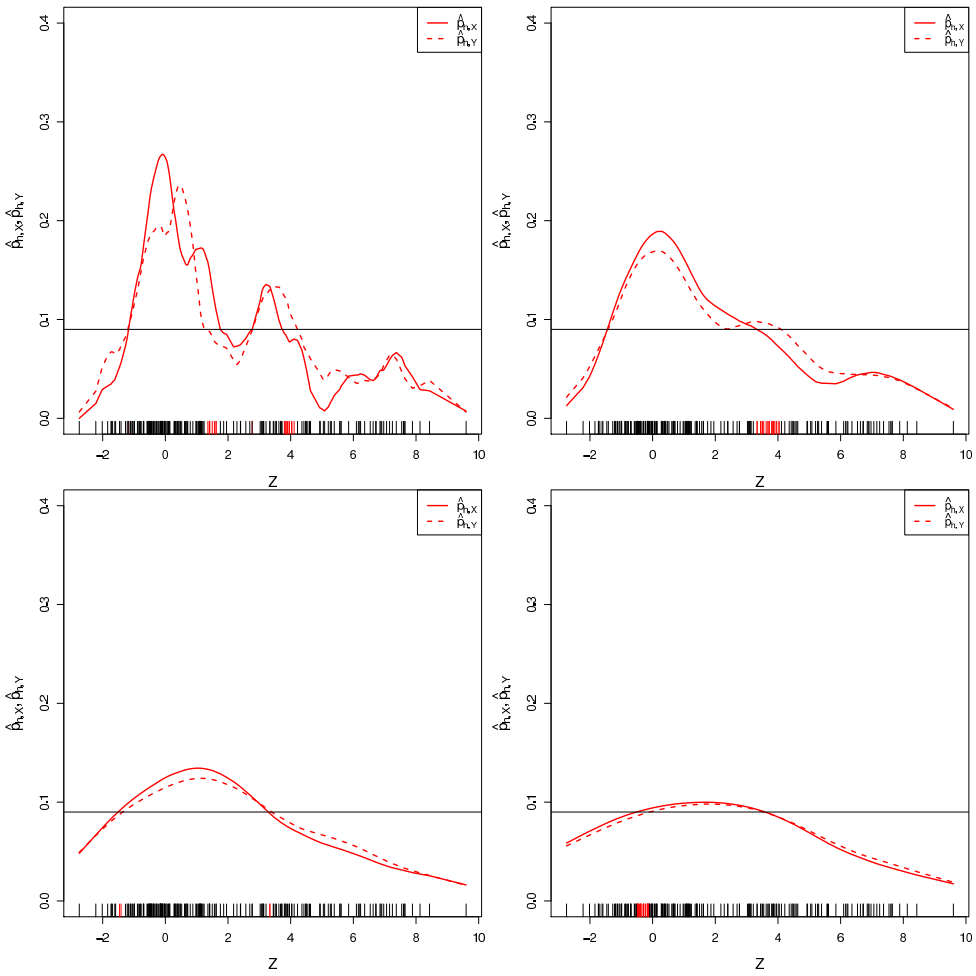


Figure 5: Comparing $\widehat{L}_{h,X}(0.09)$ and $\widehat{L}_{h,Y}(0.09)$ for $h = 0.5$ (top left), $h = 1.75$ (top right), $h = 3.75$ (bottom left) and $h = 6$ (bottom right) for data sampled from the mixture distribution of Figure 1. The two kernel density estimates are obtained using the $X$ sample (solid line) and the $Y$ sample (dotted line). Points in the $Z$ sample are showed as short vertical lines on the $x$-axis, and are colored in red when they belong to $\widehat{L}_{h,X}(\lambda)\Delta\widehat{L}_{h,Y}(\lambda)$.

quite large. When $h$ is large enough such that both density estimates lie entirely below $\lambda = 0.09$, our instability drops to and remains at zero.

Figure 6 shows the overall behavior of $\Xi_{\lambda,n}(h)$ as a function of $h$. As expected, for $\lambda = 0.02$, $\Xi_{\lambda,n}(h)$ jumps for the first non-zero $h$ and then quickly drops to almost zero by $h = 1$ (Figure 6, left). At $\lambda = 0.09$, a height with a wide range of possible level sets (depending on the density estimate and the value of $h$), $\Xi_{\lambda,n}(h)$ first drops and then oscillates as previously described as $h$ increases, indicating multi-modality (Figure 6, right).

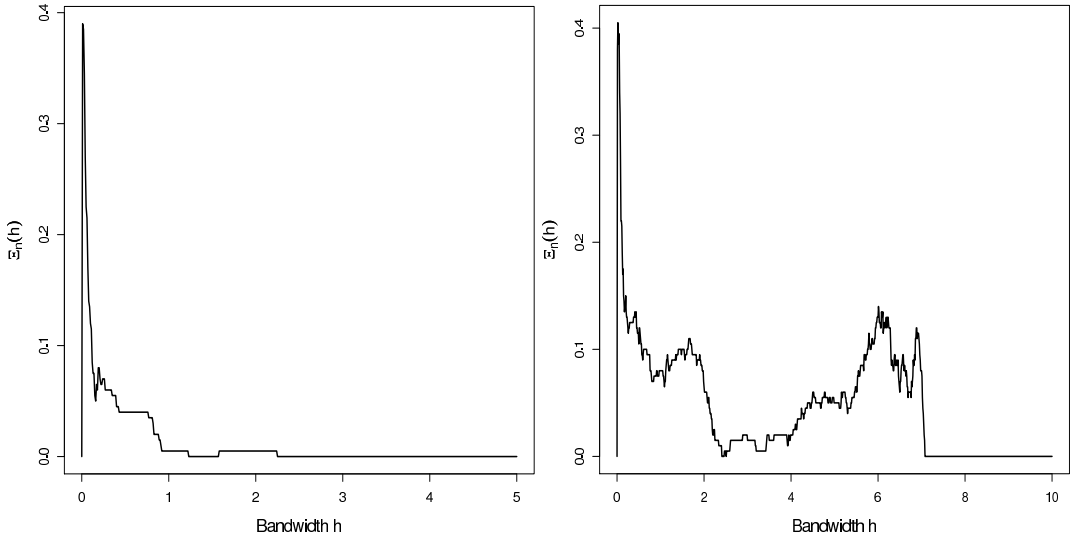


Figure 6: $\Xi_{\lambda,n}(h)$ as a function of the bandwidth $h$ for $\lambda = 0.02$ (left) and 0.09 (right) for data sampled from the mixture distribution of Figure 1.

### 5.2 Instability as Function $h$ for Fixed $\alpha$

In Section 4.2 we consider the sample instability $\Xi_{\alpha,n}(h)$ as a function of $h$ and $\alpha$. As done before, we show $\Xi_{\alpha,n}(h)$ for selected values of $h$ and $\alpha = 0.50$ and 0.95 in Figure 7. In each subfigure, $\widehat{p}_{h,X}, \widehat{p}_{h,Y}$ again are graphed for the $Z$ set of observations. The probability content of the density estimates are respectively indicated on the left and right axes. The values $\alpha = 0.50, 0.95$ are also marked with solid and dashed horizontal lines for the two density estimates. Those observations in $Z$ that belong to $\widehat{M}_{h,X}(\alpha)$ and not to $\widehat{M}_{h,Y}(\alpha)$ (or vice versa) are marked in red; the overall fraction of these observations is $\Xi_{\alpha,n}(h)$. In general, we can see that as $h$ increases (for both values of $\alpha$), the number of red $Z$ observations decreases. This decrease happens more quickly for higher values of $\alpha$ (as expected).

In Figure 8, we display $\Xi_n(h,\alpha)$ as a function of $h$ for $\alpha = 0.50, 0.95$. For level sets that contain at least 50% probability content, such as $\widehat{M}_{h,X}(0.50)$, the instability quickly drops as $h$ increases and then oscillates as $h$ approaches values that correspond to density estimates with uncertainty at those levels. Again, this ambiguity occurs due to the presence of the second mode (we would see similar behavior with respect to the smallest mode if $\alpha \approx 0.80$). As $h$ continues to increase, the density estimates become smooth enough that there is very little difference between $M_{h,X}(0.50)$, $M_{h,Y}(0.50)$. This behavior also occurs when $\alpha = 0.95$ albeit more quickly (Figure 8, top right) since level sets that contain at least 95% probability content occur at lower heights and are more stable.

Figure 8c is the corresponding heat map for $\alpha = 0, 0.01, \ldots, 1.0$ and $h = 0, 0.01, \ldots, 10$. White sections indicate $\Xi_{\alpha,n}(h) \approx 0$; black sections indicate higher instability values. In this particular example, the maximum instability of 0.425 is found at $h = 0.03, \alpha = 0.46$. Note that around $h = 3$, we have very low instability values for almost all values of $\alpha$, and hence this value of kernel bandwidth would be a good choice that yields stable clustering.

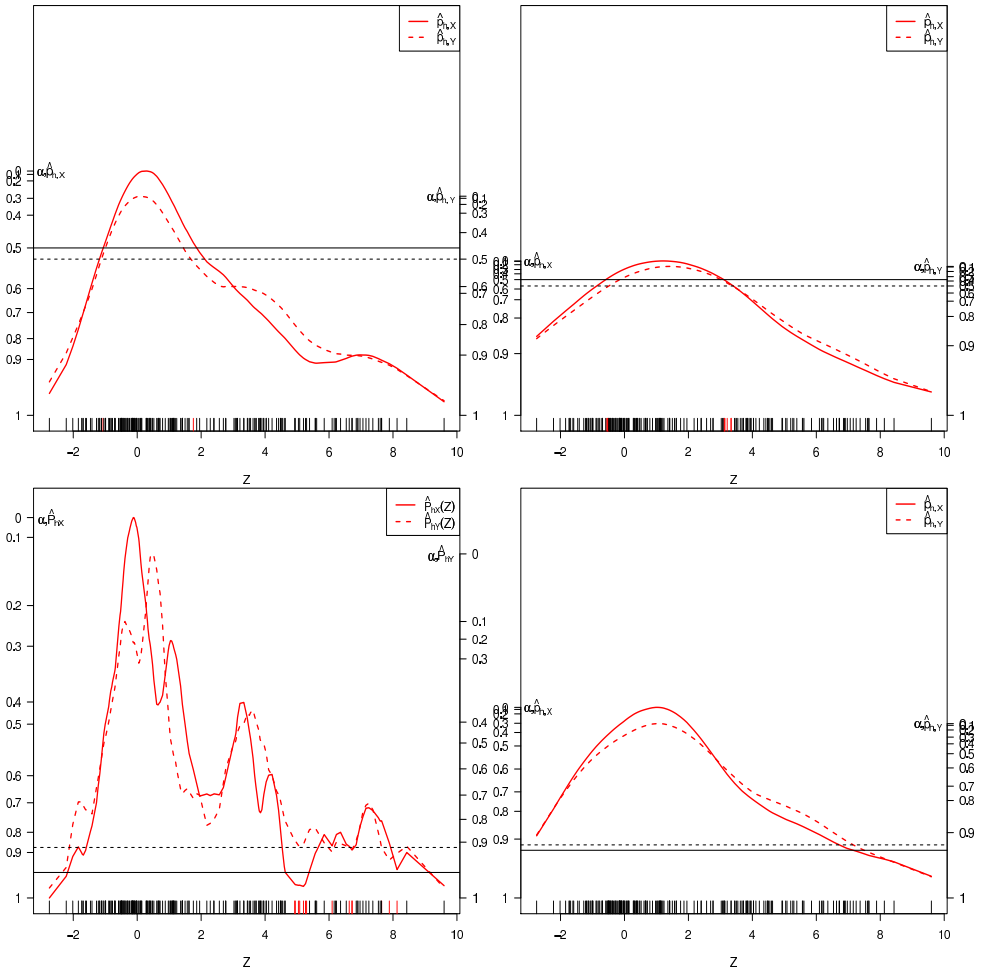


Figure 7: Top: comparing $\widehat{M}_{h,X}(0.50)$ and $\widehat{M}_{h,Y}(0.50)$ for $h = 2$ (left) and $h = 5$ right). Bottom: comparing $\widehat{M}_{h,X}(0.95)$ and $\widehat{M}_{h,Y}(0.95)$ for $h = 0.4$ (left) and $h = 3.5$ (right). The data were sampled from the mixture distribution of Figure 1. The two kernel density estimates are obtained using the $X$ sample (solid line) and the $Y$ sample (dotted line). Points in the $Z$ sample are showed as short vertical lines on the $x$-axis, and are colored in red when they belong to $\widehat{M}_{h,X}(\alpha)\Delta\widehat{M}_{h,Y}(\alpha)$.

### 5.3 Instability Confidence Bands

The results in the previous subsections were for splitting the original sample one time into three groups of 200 observations. Here we briefly include a snapshot of what the distribution of our instability measures look like over repeated splits. For computational reasons, we used the binned kernel density estimate, again with the Epanechnikov kernel, and discretize the feature space over 200 bins; see Wand (1994). Increasing the number of bins improves the approximation to the kernel density estimate; the use of two hundred bins was found to give almost identical results to the original kernel density estimate (results not shown). We split the original sample 200 times and find 95% point-wise confidence intervals for $\Xi_{\lambda,n}(h)$, $\Gamma_n(h)$, and $\Xi_{\alpha,n}(h)$ for $\alpha = 0.50, 0.95$ and as a function of $h$. The results are depicted in Figure 9. The confidence bands are plotted in red, the

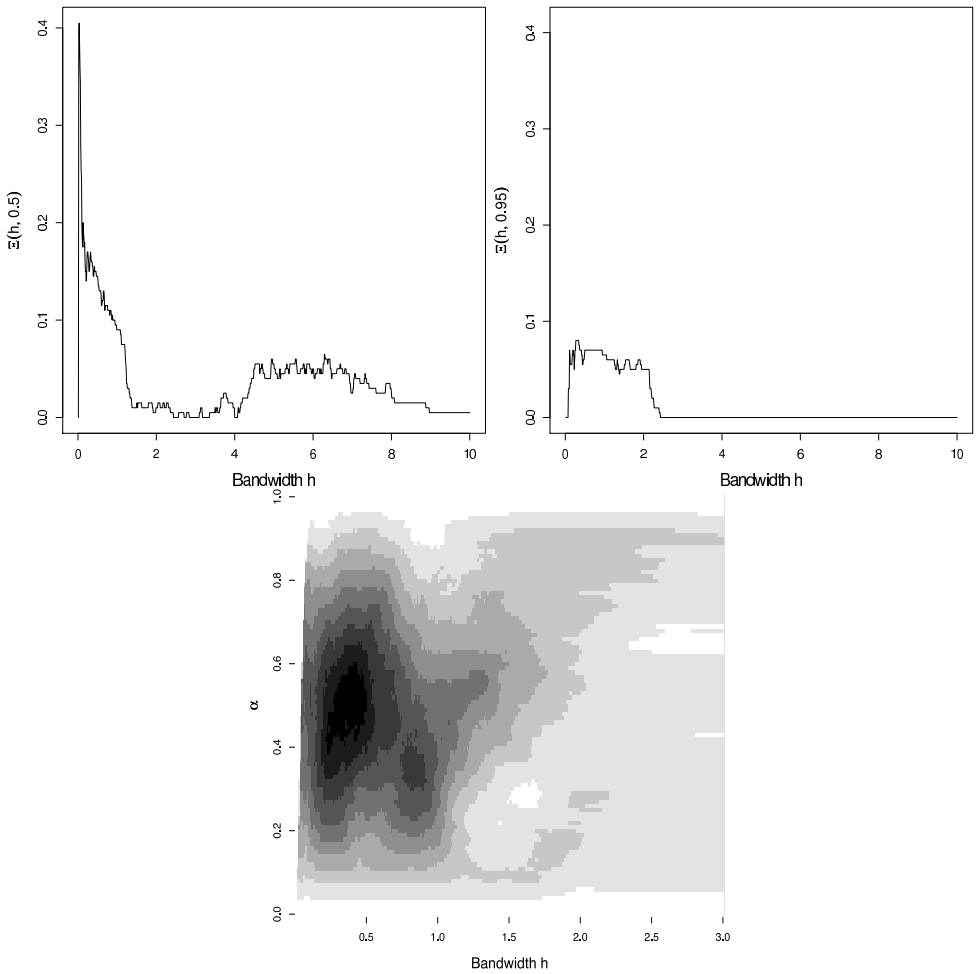


Figure 8: Top: $\Xi_n(h, \alpha = 0.50)$ (left) and $\Xi_n(h, \alpha = 0.95)$ (right) as a function of $h$. Bottom: heat map of $\Xi_{\alpha,n}(h)$ as function of $h, \alpha$ for the example of Figure 1. The data were sampled from the mixture distribution of Figure 1.

medians in black. The distribution of the instability measures for each value of $h$ is also plotted using density strips (see Jackson, 2008); on the grey-scale, darker colors indicate more common instability values. The density strips allow us to see how the distribution changes (not just the 50, 95% percentiles). For example, for the plot on the top left in Figure 9, note that right before $h = 2$, the upper half of the distribution of $\Xi_{\lambda,n}(h)$ is more concentrated. This shift corresponds to the increase in instability in the presence of the additional modes.

### 5.4 Bivariate Moons

We also include a bivariate example with two equal-sized moons; this data set with seemingly simple structure can be quite difficult to analyze. The scatterplot of the data on the left in Figure 10 show two clusters, each shaped like a half moon. Each cluster contains 300 data points. The plot on the right in Figure 10b shows a two-dimensional kernel density estimate using a Epanechnikov kernel with $h = 0.60$ (for illustrative purposes) and 10,000 evaluation points. We can see that while levels

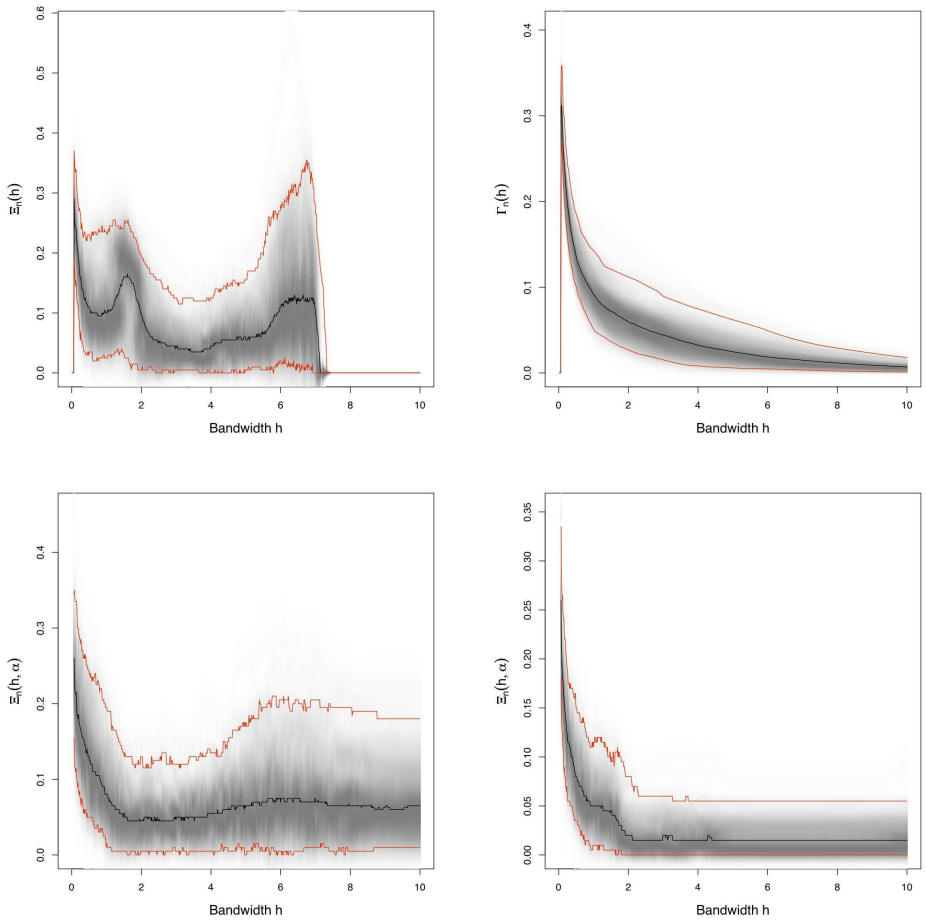


Figure 9: 95% point-wise confidence bands for $\Xi_{\lambda,n}(h)$ (top left), $\Gamma_n(h)$ (top right), $\Xi_n(h,\alpha=0.50)$ (bottom left) and $\Xi_n(h,\alpha=0.95)$ (bottom right) for data sampled from the mixture distribution of Figure 1.

around $\lambda = 0.012$ show clear multi-modality, the connectedness of the level sets around $\lambda = 0.01$ is less clear.

To examine instability, we use a product Epanechnikov kernel density estimate with the same bandwidth $h$ for both dimensions. Figure 11 shows the sample instability $\Xi_{\lambda,n}(h)$ as a function of $h$ for $\lambda = 0.10, 0.20, 0.30$ as well as the total variation instability $\Gamma_n(h)$ as a function of $h$. As expected, the higher the $\lambda$, the more quickly the sample instability drops. We also see the possible presence of multi-modality for all three values of $\lambda$ in $\Xi_{\lambda,n}(h)$. On the other hand, the total variation instability drops smoothly as $h$ increases.

Figure 12 contains the instability as a function of $h$ and probability content $\alpha$ for all values of $h$, $\alpha$ (Figure 12d) and specifically for $\alpha = 0.50, 0.075, 0.95$. Again, as expected, $\Xi_n(h,\alpha)$ drops as $h$ increases for smaller values of $\alpha$. Note that for $\alpha = 0.95$, the instability remains relatively low regardless of the value of $h$. When examining the heat map, we see that for small values of $h$, level sets corresponding to probability content around 0.4-0.6 are very unstable. This behavior

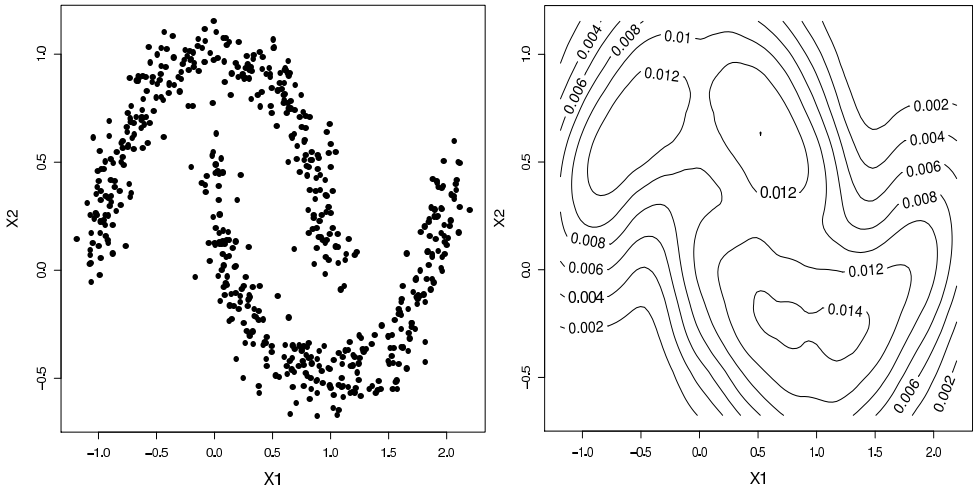


Figure 10: Bivariate moons (left) and contours of a Epanechnikov kernel density estimate (right) for the example discussed in Section 5.4.

is not unexpected given that the moons are of equal sizes and difficult to separate due to sampling variability. We would expect to have difficulty finding stable level sets "in the middle".

## 6. Discussion

We have investigated the properties of the density level set and cluster tree estimator based on kernel density estimates, and we have proposed and analyzed various measures of instability for these quantities. We believe these measures of instability can be of guidance in choosing the bandwidth parameter and also as exploratory tools to gain insights into the properties and shape of the data-generating distribution.

Our analysis leaves some some open questions that we think deserve further attention. First, we have focused on kernel density estimators but the same ideas can be used with other density estimators or more, generally, with other clustering methods for which underlying tuning parameters have to be chosen in a data-driven fashion. See, for instance, Meinshausen and Bühlmann (2010) for a related stability-based approach to clustering.

We have assumed the existence of the Lebesgue density $p$ but this assumption can be relaxed using methods in Rinaldo and Wasserman (2010) to allow for distributions supported on lower-dimensional, well-behaved subsets. This extension is potentially important because it would allows us to include cases where the distribution has positive mass on lower dimensional structures such as points and manifolds.

We have formulated our assumptions and results about stability of the level sets and of the cluster tree in a point-wise manner, for given values of $\lambda$ and $\alpha$. As suggested by a reviewer, it would be desirable to extend them to hold uniformly across level sets. This can be achieved by requiring (A2), (B2) and (B3) to hold uniformly over values of $\lambda$ and $\alpha$. In fact, we believe that it is likely that, for most densities, such uniform assumptions hold for a wide range of $\lambda$'s but certainly they cannot hold for *all* $\lambda$'s. Indeed, our results indicate that these uniformity assumptions are reasonable only for level sets $\lambda$ for which the function $r_{h,\varepsilon}(\lambda)$ remains small and does not fluctuate too wildly.

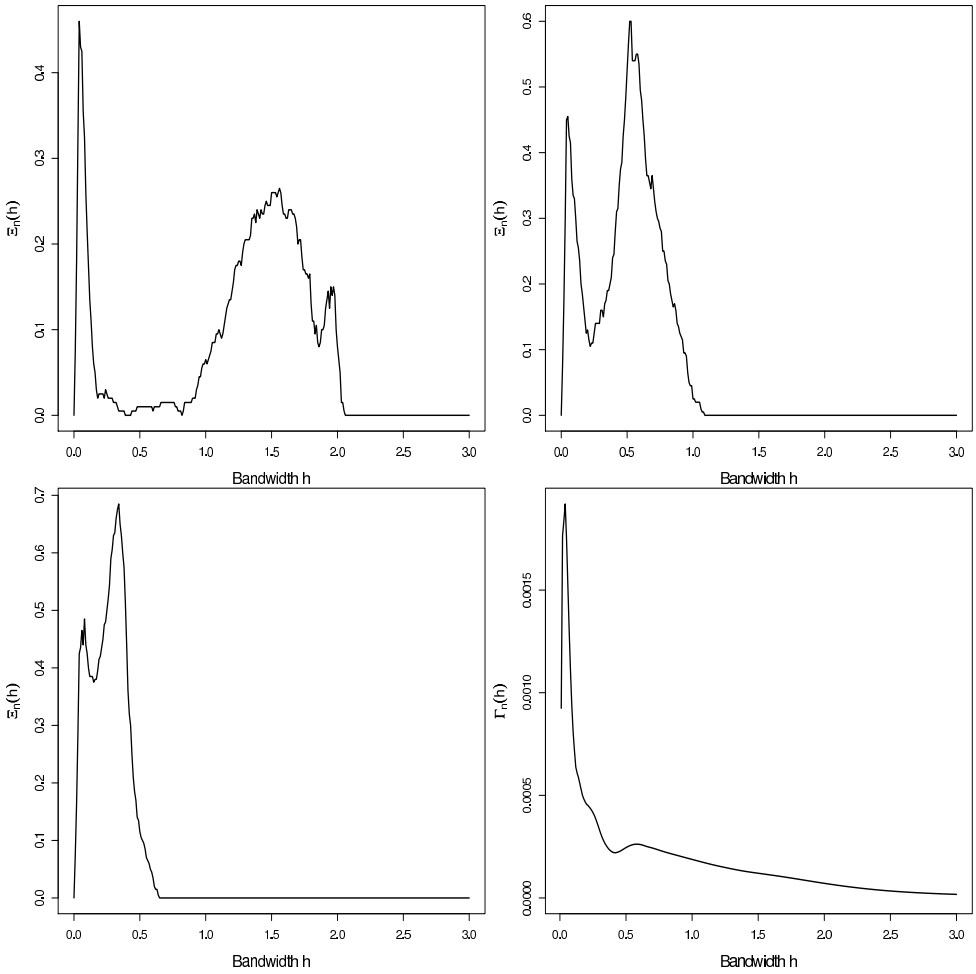


Figure 11: $\Xi_{\lambda,n}(h)$ as a function of $h$ for $\lambda = 0.10$ (top left) 0.20 (top right) and 0.30 (bottom left). $\Gamma_n(h)$ as a function of $h$ (bottom right) for the data depicted in Figure 10.

Finally, in computing the various measures of instability, we have considered just a single split of the data into non-overlapping sub-samples. In fact, one can randomly repeat the splitting process and combine over many splits, which is how we obtained the confidence bands of Figure 9. Though the increase in the computational costs may be significant, repeated sub-sampling would yield a reliable estimate of the uncertainty of the chosen instability measures and would therefore be highly informative about the sample. We believe that the properties of $\Xi_n$ can be established using the theory of U-statistics.

## Acknowledgments

Research supported by NSF grant CCF-0625879 and AFOSR contract FA9550-09-1-0373. The authors thank the referees for helpful comments and suggestions.

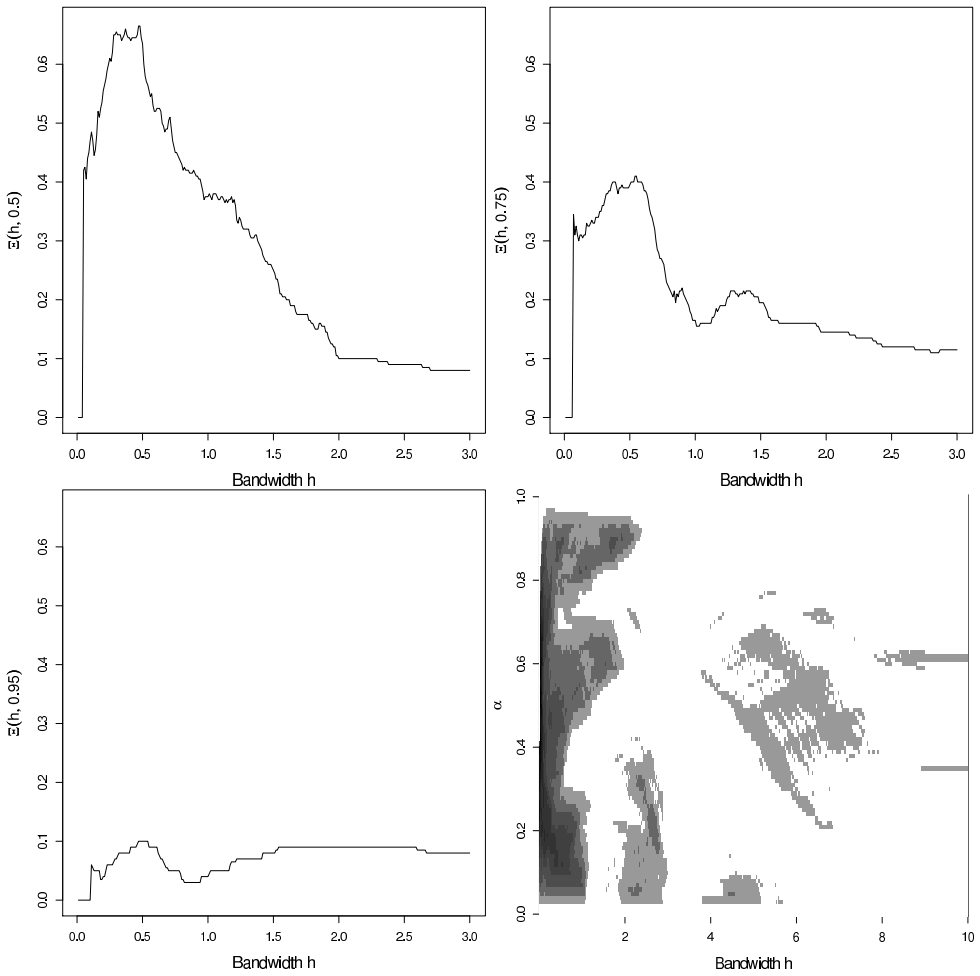


Figure 12: $\Xi_{\alpha,n}(h)$ as a function of $h$ for $\alpha = 0.50$ (top left) 0.75 (top right) and 0.95 (bottom left). Heat Map of $\Xi_{\alpha,n}(h)$ as function of $h$ and $\alpha$; for readability, values of $\Xi_{\alpha,n}(h)$ smaller than 0.045 are displayed in white (bottom right).

## Appendix A. Proofs

**Proof of Theorem 2:** Let $\mathcal{A}_{h_n,\varepsilon_n}$ denote the event that $\|\widehat{p}_{h_n,X} - p_{h_n}\|_\infty \leq \varepsilon_n$. Then, for all $n \geq n_0$, by Equation (3), $\mathbb{P}_X(\mathcal{A}_{h_n,\varepsilon_n}) \geq 1 - \frac{1}{n}$. Also observe that Assumption (A1) implies that, for any $h > 0$, the sup-norm density approximation error can be bounded as

$$\begin{aligned} \|p_h - p\|_\infty &= \sup_x \left| \int \frac{1}{h^d} K\left(\frac{x-y}{h}\right) p(y)dy - p(x) \right| \\ &\leq \sup_x \int \frac{1}{h^d} K\left(\frac{x-y}{h}\right) A\|x-y\|dy \\ &= ADh. \end{aligned} \tag{11}$$

The second step in the previous display follows since $\int K(z)dz = 1$ and using the Lipschitz assumption (A1) on the density, and the last step since $\int \|z\|K(z)dz = D$. Putting the estimation and

approximation error together, and using the triangle inequality, we obtain that, on the event $\mathcal{A}_{h_n,\varepsilon_n}$,

$$\|\widehat{p}_{h_n,X} - p\|_\infty \leq ADh_n + \varepsilon_n, \tag{12}$$

for all $n \geq n_0$. Using Equation (12), we have that, on $\mathcal{A}_{h_n,\varepsilon_n}$ and for all $n \geq n_1(n_0,\lambda)$ so that $ADh_n + \varepsilon_n < \lambda$, the set

$$L(\lambda)\Delta\widehat{L}_{h_n,X}(\lambda) = \{u\colon\, p(u) > \lambda, \widehat{p}_{h_n,X}(u) \leq \lambda\} \cup \{u\colon\, p(u) \leq \lambda, \widehat{p}_{h_n,X}(u) > \lambda\}$$

is contained in

$$\{u\colon\, p(u) > \lambda, p(u) \leq \lambda + ADh_n + \varepsilon_n\} \cup \{u\colon\, p(u) \leq \lambda, p(u) > \lambda - ADh_n - \varepsilon_n\},$$

which is equal to

$$\{u\colon\, |p(u) - \lambda| < ADh_n + \varepsilon_n\}.$$

Then, on $\mathcal{A}_{h_n,\varepsilon_n}$ and for all $n \geq n_1(n_0,\lambda)$ large enough

$$\mathcal{L}(h_n, X, \lambda) = P(L(\lambda)\Delta\widehat{L}_{h_n,X}(\lambda)) \leq r_{h_n,\varepsilon_n,\lambda},$$

so that, $\mathbb{P}_X\left(\mathcal{L}(h_n,X,\lambda) \leq r_n\right) \geq \mathbb{P}_X\left(\mathcal{A}_{h_n,\varepsilon_n}\right) \geq 1 - \frac{1}{n}$, as claimed.

If (A2) is in force for the density level $\lambda$, then for all $n \geq n_2(n_0,\lambda,A,D,\varepsilon_0)$ so that $ADh_n + \varepsilon_n \leq \varepsilon_0$, we have $r_{h_n,\varepsilon_n,\lambda} \leq \kappa_2(ADh_n + \varepsilon_n)$, which proves the second claim.

**Proof of Lemma 4:** Using (A1) and the fact that $\int_{\mathbb{R}^d} K(z)dz = 1$, Equation (11) states that for any $h > 0$

$$\|p_h - p\|_\infty \leq ADh.$$

Then, for any $\alpha \in (0,1)$ and $h > 0$,

$$\{u\colon\, p(u) > \lambda_{h,\alpha} + ADh\} \subseteq \{u\colon\, p_h(u) > \lambda_{h,\alpha}\} \subseteq \{u\colon\, p(u) > \lambda_{h,\alpha} - ADh\}.$$

And as a result,

$$P(\{u\colon\, p(u) > \lambda_{h,\alpha} + ADh\}) \leq P(\{u\colon\, p_h(u) > \lambda_{h,\alpha}\}) \leq P(\{u\colon\, p(u) > \lambda_{h,\alpha} - ADh\}).$$

Since $P(\{u\colon\, p(u) > \lambda_\alpha\}) = \alpha = P(\{u\colon\, p_h(u) > \lambda_{h,\alpha}\})$, we have

$$P(\{u\colon\, p(u) > \lambda_{h,\alpha} + ADh\}) \leq P(\{u\colon\, p(u) > \lambda_\alpha\}) \leq P(\{u\colon\, p(u) > \lambda_{h,\alpha} - ADh\}).$$

Consequently,

$$\lambda_{h,\alpha} + ADh \geq \lambda_\alpha \geq \lambda_{h,\alpha} - ADh.$$

It follows that for any $\alpha \in (0,1)$ and $h > 0$

$$|\lambda_{h,\alpha} - \lambda_\alpha| \leq ADh.$$

**Proof of Lemma 5:** Let $\mathcal{C}_h = \left\{\{u\colon\, p_h(u) > \lambda\}, \lambda > 0\right\}$ denote the class of level sets of $p_h$ and define the events

$$\mathcal{P}_{h,\varepsilon} = \left\{\sup_{C \in \mathcal{C}_h} |\widehat{P}_X(C) - P(C)| \leq \varepsilon\right\} \quad \text{and} \quad \mathcal{A}_{h,\varepsilon} = \left\{||\widehat{p}_{h,X} - p_h||_\infty \leq \varepsilon\right\}.$$

Then, since the $n$-th shatter coefficients (see, for instance, Devroye et al., 1996) of $\mathcal{C}_h$ is $n$,

$$\mathbb{P}_X(\mathcal{P}_{h,\varepsilon}^c) \leq 8ne^{-n\varepsilon^2/32} \quad \text{and} \quad \mathbb{P}_X(\mathcal{A}_{h,\varepsilon}^c) \leq K_1 e^{-K_2 n\varepsilon^2 h^d}, \tag{13}$$

where the first inequality follows from the VC inequality (see, for instance, Devroye et al., 1996) and the second inequality is just (1). Then, on $\mathcal{A}_{h,\varepsilon}$, we obtain

$$\{u\colon p_h(u) > \lambda+\varepsilon\} \subseteq \{u\colon \widehat{p}_{h,X}(u) > \lambda\} \subseteq \{u\colon p_h(u) > \lambda-\varepsilon\}, \quad \forall \lambda > 0.$$

Thus, on $\mathcal{A}_{h,\varepsilon}$,

$$\widehat{P}_X(\{u\colon p_h(u) > \lambda+\varepsilon\}) \leq \widehat{P}_X(\{u\colon \widehat{p}_{h,X}(u) > \lambda\}) \leq \widehat{P}_X(\{u\colon p_h(u) > \lambda-\varepsilon\}),$$

uniformly over all $\lambda > 0$. In particular, the previous inequality hold also for $\widehat{\lambda}_{\alpha,h,X}$ (which is positive with probability one) for any $\alpha \in (0,1)$ and $h > 0$.

Recalling that, by definition,

$$|\widehat{P}_X(\{u\colon \widehat{p}_{h,X}(u) > \widehat{\lambda}_{h,\alpha,X}\}) - \alpha| \leq 1/n,$$

we obtain, on the events $\mathcal{P}_{h,\varepsilon}$ and $\mathcal{A}_{h,\varepsilon}$,

$$P(\{u\colon p_h(u) > \widehat{\lambda}_{h,\alpha,X}+\varepsilon\}) - \frac{1}{n} - \varepsilon \leq \alpha \leq P\{u\colon p_h(u) > \widehat{\lambda}_{h,\alpha,X}-\varepsilon\}) + \frac{1}{n} + \varepsilon. \tag{14}$$

Since $\alpha = P(\{u\colon p_h(u) > \lambda_{h,\alpha}\})$, the first inequality in (14) can be written as

$$\alpha + \frac{1}{n} + \varepsilon = P(\{u\colon p_h(u) > \lambda_{h,\alpha+\frac{1}{n}+\varepsilon}\}) \geq P(\{u\colon p_h(u) > \widehat{\lambda}_{h,\alpha,X}+\varepsilon\})$$

and the second one as

$$\alpha - \frac{1}{n} - \varepsilon = P(\{u\colon p_h(u) > \lambda_{h,\alpha-\frac{1}{n}-\varepsilon}\}) \leq P\{u\colon p_h(u) > \widehat{\lambda}_{h,\alpha,X}-\varepsilon\}),$$

both holding on the events $\mathcal{P}_{h,\varepsilon}$ and $\mathcal{A}_{h,\varepsilon}$. Combining the last two expressions, we obtain, on the same events, for any $\alpha \in (0,1)$ and $h > 0$,

$$\lambda_{h,\alpha+\frac{1}{n}+\varepsilon} - \varepsilon \leq \widehat{\lambda}_{h,\alpha,X} \leq \lambda_{h,\alpha-\frac{1}{n}-\varepsilon} + \varepsilon. \tag{15}$$

We will now show that, for level sets of $p_h$ indexed by $\alpha$ satisfying (B3), and for any $\eta \in (-\eta_0, \eta_0)$ and $0 < h \leq H$,

$$|\lambda_{h,\alpha+\eta} - \lambda_{h,\alpha}| \leq A\kappa_3|\eta|. \tag{16}$$

Recalling that $\varepsilon + 1/n < \eta_0$, Equations (15) and (16) will then imply

$$\lambda_{h,\alpha} - A\kappa_3\left(\varepsilon+\frac{1}{n}\right) - \varepsilon \leq \widehat{\lambda}_{h,\alpha,X} \leq \lambda_{h,\alpha} + A\kappa_3\left(\varepsilon+\frac{1}{n}\right) + \varepsilon,$$

on the events $\mathcal{P}_{h,\varepsilon}$ and $\mathcal{A}_{h,\varepsilon}$, for level sets of $p_h$ indexed by $\alpha$ satisfying (B3) and with $0 < h \leq H$. Finally, using (13), the claim will follow.

In order to show (16), for a set $A \subset \mathbb{R}^d$, let $\partial A$ denote its boundary. Then, notice that, because $p_h$ is Lipschitz and hence continuous, for every $x \in \partial M_h(\alpha)$, $p_h(x) = \lambda_{h,\alpha}$ and, for every $y \in \partial M_h(\alpha+\eta)$, $p_h(y) = \lambda_{h,\alpha+\eta}$. Furthermore, for any point $x \in \partial M_h(\alpha)$, there exists a point $y = y(x) = \inf_{z \in \partial M_h(\alpha+\eta)} \|x-z\|$. Thus, for $|\eta| < \eta_0$,

$$\|x-y\| \leq d_\infty(M_h(\alpha), M_h(\alpha+\eta)) \leq \kappa_3|\eta|,$$

where the last inequality follows for level sets of $p_h$ indexed by $\alpha$ that satisfy (B3) and $0 < h \leq H$. Therefore,

$$|\lambda_{h,\alpha+\eta} - \lambda_{h,\alpha}| = |p_h(y) - p_h(x)| \leq A\|x-y\| \leq A\kappa_3|\eta|,$$

where in the first inequality we used the fact that, by (A1), $p_h$ is Lipschitz with constant $A$. Indeed, for any $x \neq y$, using the Lipschitz assumption (A1) on $p$,

$$|p_h(x) - p_h(y)| \leq \int_{\mathbb{R}^d} |p(x+zh) - p(y+zh)| K(z)dz \leq A\|x-y\| \int_{\mathbb{R}^d} K(z)dz = A\|x-y\|.$$

**Proof of Theorem 7:** Let $\mathcal{A}_{h_n,\varepsilon_n}$ be event defined in the proof of Theorem 2, and recall that for all $n \geq n_0$, by Equation (3), $\mathbb{P}_X(\mathcal{A}^c_{h_n,\varepsilon_n}) \leq 1/n$ and that, Equation (12) states that

$$\|\widehat{p}_{h,X} - p\|_\infty \leq C_{1,n} \tag{17}$$

on that event, for all $n \geq n_0$. Also, let $\mathcal{P}_{h_n,\varepsilon_n}$ be the event defined in Lemma 5 such that $\mathbb{P}_X(\mathcal{P}^c_{h_n,\varepsilon_n}) \leq 8ne^{-n\varepsilon_n^2/32}$. Then from the proof of Lemma 5, we have that on the event $\mathcal{A}_{h_n,\varepsilon_n} \cap \mathcal{P}_{h_n,\varepsilon_n}$, for $h_n = \omega((\log n/n)^{1/d})$ and $h_n \leq H$,

$$|\widehat{\lambda}_{h_n,\alpha,X} - \lambda_\alpha| \leq C_{2,n} \tag{18}$$

for all $n \geq n_3(n_0, \eta_0, K_3)$. Also, since $n$ is large enough, we have

$$8ne^{-n\varepsilon_n^2/32} \leq \frac{1}{n}.$$

Therefore, for all such large $n$, both (17) and (18) hold with probability at least

$$\mathbb{P}_X\left(\mathcal{A}_{h_n,\varepsilon_n} \cap \mathcal{P}_{h_n,\varepsilon_n}\right) \geq 1 - \frac{2}{n}.$$

Thus, on $\mathcal{A}_{h_n,\varepsilon_n} \cap \mathcal{P}_{h_n,\varepsilon_n}$, for $h_n = \omega((\log n/n)^{1/d})$ and $h_n \leq H$, we have that, for all $n \geq n_3(n_0, \eta_0, K_3)$, the set

$$M(\alpha)\Delta\widehat{M}_{h,X}(\alpha) = \{u\colon\, p(u) > \lambda_\alpha, \widehat{p}_{h,X}(u) \leq \widehat{\lambda}_{h,\alpha,X}\} \cup \{u\colon\, p(u) \leq \lambda_\alpha, \widehat{p}_{h,X}(u) > \widehat{\lambda}_{h,\alpha,X}\}.$$

is contained in

$$\{u\colon\, p(u) > \lambda_\alpha, p(u) \leq \widehat{\lambda}_{h,\alpha,X} + C_{1,n}\} \cup \{u\colon\, p(u) \leq \lambda_\alpha, p(u) > \widehat{\lambda}_{h,\alpha,X} - C_{1,n}\}.$$

which, in turn, is a subset of

$$\{u\colon\, p(u) > \lambda_\alpha, p(u) \leq \lambda_\alpha + C_{1,n} + C_{2,n}\} \cup \{u\colon\, p(u) \leq \lambda_\alpha, p(u) > \lambda_\alpha - C_{1,n} - C_{2,n}\}.$$

The final set is just $\{u\colon |p(u)-\lambda_\alpha| \le C_{1,n}+C_{2,n}\}$. Therefore, for for $h_n = \omega((\log n/n)^{1/d})$ and $h_n \le H$, we have, for all $n \ge n_3(n_0,\eta_0,K_3)$,

$$\mathbb{P}_X\left(\mathcal{L}^*(h_n,X,\alpha) \le r_{h_n,\varepsilon_n,\alpha}\right) \ge \mathbb{P}_X\left(\mathcal{A}_{h_n,\varepsilon_n} \cap \mathcal{P}_{h_n,\varepsilon_n}\right) \ge 1-\frac{2}{n}.$$

**Proof of Lemma 9:** We only prove the second claim, since the proof of the limits is straightforward. For simplicity, we will provide the proof for the case of a spherical kernel: $K(x) = 1_{\|x\|\le 1}$, $x \in \mathbb{R}^d$. The extension to other compactly supported kernels is analogous.

Let $h$ be strictly smaller than

$$\min\left\{\min_{i\neq j}||X_i-X_j||, \min_{i\neq j}||Y_i-Y_j||, \min_{i,j}||X_i-Y_j||\right\}.$$

For many distributions, this occurs almost surely for $h = O\left(1/n^d\right)$ (see, e.g., Penrose, 2003; Deheuvels et al., 1988). By the compactness of the support of $K$, for any such $h$, the sets

$$B(X_1,h),\ldots,B(X_n,h),B(Y_1,h),\ldots,B(Y_n,h)$$

are disjoint. Therefore, $\widehat{p}_{h,X}(u) = 1/(nh^d)$ if and only if $u \in B(X_i,h)$ for one $i$ and, similarly, $\widehat{p}_{h,Y}(u) = 1/(nh^d)$ if and only if $u \in B(Y_j,h)$ for one $j$. Furthermore,

$$\widehat{L}_{h,X}\Delta\widehat{L}_{h,Y} = \left(\bigcup_i B(X_i,h)\right)\bigcup\left(\bigcup_j B(Y_j,h)\right).$$

As a result, $\Xi_{\lambda,n}(h)$ is the fraction of $Z_i$'s contained in $(\cup_i B(X_i,h))\bigcup(\cup_i B(Y_i,h))$. Thus,

$$\Xi_{\lambda,n}(h) = \widehat{P}_Z(\widehat{L}_{h,X}\Delta\widehat{L}_{h,Y}|X,Y) \stackrel{d}{=} B/n,$$

where $\stackrel{d}{=}$ denotes equality in distribution and $B \sim \text{Binomial}(n,p_0)$, with $0 \le p_0 \le 2n\ p_{\max} v_d h^d$ and $p_{\max} = \|p\|_\infty$. Therefore, $\mathbb{E}_Z[\Xi_{\lambda,n}(h)|X,Y] \le 2p_{\max}v_d nh^d$ and hence it follows that

$$\xi_{\lambda,n}(h) = \mathbb{E}_{X,Y,Z}[\Xi_{\lambda,n}(h)] \le 2p_{\max}v_d nh^d = O(h^d),$$

as $h \to 0$.

**Proof of Theorem 10:**

1. Since $X$, $Y$ and $Z$ are independent samples from the same distribution, $\widehat{p}_{h,X}(u)$ and $\widehat{p}_{h,Y}(u)$ are independent and identically distributed, for any $u \in \mathbb{R}^d$ and $h > 0$. Also, notice that for every measurable set $A$, $\mathbb{E}_Z(\widehat{P}_Z(A)) = P(A)$. Thus,

$$\begin{aligned}
\xi_{\lambda,n}(h) &= \mathbb{E}_{X,Y,Z}[\widehat{P}_Z(\{u\colon \widehat{p}_{h,X}(u) > \lambda\}\Delta\{u\colon \widehat{p}_{h,Y}(u) > \lambda\})] \\
&= \mathbb{E}_{X,Y}[P(\{u\colon \widehat{p}_{h,X}(u) > \lambda, \widehat{p}_{h,Y}(u) \le \lambda\}) + P(\{u\colon \widehat{p}_{h,X}(u) \le \lambda, \widehat{p}_{h,Y}(u) > \lambda\})] \\
&= 2\mathbb{E}_{X,Y}\left[P(\{u\colon \widehat{p}_{h,X}(u) > \lambda, \widehat{p}_{h,Y}(u) \le \lambda\})\right] \\
&= 2\int_{\mathbb{R}^d} \mathbb{P}_{X,Y}\left(\widehat{p}_{h,X}(u) > \lambda, \widehat{p}_{h,Y}(u) \le \lambda\right) dP(u), \qquad (19)
\end{aligned}$$

where the last identity follows from Fubini theorem. The integrand in the last equation can be written as

$$\begin{aligned}\mathbb{P}_{X,Y}\left(\widehat{p}_{h,X}(u)>\lambda,\widehat{p}_{h,Y}(u)\leq\lambda\right) &= \mathbb{P}_X\left(\widehat{p}_{h,X}(u)>\lambda\right)\mathbb{P}_Y\left(\widehat{p}_{h,Y}(u)\leq\lambda\right)\\ &= \mathbb{P}_X\left(\widehat{p}_{h,X}(u)>\lambda\right)\mathbb{P}_X\left(\widehat{p}_{h,X}(u)\leq\lambda\right)\\ &= \pi_h(u)(1-\pi_h(u)),\end{aligned}$$

from which (8) follows.

2. Let $\mathcal{A}_{h,\varepsilon}$ denote the event

$$\|p_h-\widehat{p}_{h,X}\|_\infty\vee\|p_h-\widehat{p}_{h,Y}\|_\infty\leq\varepsilon. \tag{20}$$

By (1), $\mathbb{P}_{X,Y}(\mathcal{A}_{h,\varepsilon}^c)\leq 2K_1e^{-K_2nh^d\varepsilon^2}$. Letting $1_{\mathcal{A}_{h,\varepsilon}}$ denote the indicator function of the event $\mathcal{A}_{h,\varepsilon}$,

$$\xi_{\lambda,n}(h)\leq\mathbb{E}_{X,Y,Z}[\widehat{P}_Z(\{u\colon\widehat{p}_{h,X}(u)>\lambda\}\Delta\{u\colon\widehat{p}_{h,Y}(u)>\lambda\})1_{\mathcal{A}_{h,\varepsilon}}(X,Y)]+\mathbb{P}_{X,Y}(\mathcal{A}_{h,\varepsilon}^c),$$

and, using the same reasoning that led to (19),

$$\xi_{\lambda,n}(h)\leq 2\int_{\mathbb{R}^d}\mathbb{P}_{X,Y}\left(\{\widehat{p}_{h,X}(u)>\lambda,\widehat{p}_{h,Y}(u)\leq\lambda\}\cap\mathcal{A}_{h,\varepsilon}\right)dP(u)+\mathbb{P}_{X,Y}(\mathcal{A}_{h,\varepsilon}^c)$$

Notice that, on $\mathcal{A}_{h,\varepsilon}$,

$$\{u\colon\widehat{p}_{h,X}(u)>\lambda,\widehat{p}_{h,Y}(u)\leq\lambda\}\subseteq\{u\colon\lambda-\varepsilon\leq p_h(u)\leq\lambda+\varepsilon\}=U_{h,\varepsilon},$$

and therefore, $\operatorname{sign}(\widehat{p}_{h,X}(u)-\lambda)=\operatorname{sign}(p_h(u)-\lambda)$ for all $u\notin U_{h,\varepsilon}$. Thus, the previous expression for $\xi_{\lambda,n}(h)$ is upper bounded by

$$2\int_{U_{h,\varepsilon}}\mathbb{P}_{X,Y}\left(\{\widehat{p}_{h,X}(u)>\lambda,\widehat{p}_{h,Y}(u)\leq\lambda\}\cap\mathcal{A}_{h,\varepsilon}\right)dP(u)+2K_1e^{-K_2nh^d\varepsilon^2}$$

which, using independence, is no larger than

$$2\int_{U_{h,\varepsilon}}\pi_h(u)(1-\pi_h(u))dP(u)+2K_1e^{-K_2nh^d\varepsilon^2}\leq P(U_{h,\varepsilon})\overline{A}_{h,\varepsilon}+2K_1e^{-K_2nh^d\varepsilon^2}.$$

As for the lower bound, from (19) we obtain, trivially,

$$\xi_{\lambda,n}(h)\geq 2\int_{U_{h,\varepsilon}}\pi_h(u)(1-\pi_h(u))dP(u)\geq P(U_{h,\varepsilon})\overline{A}_{h,\varepsilon}.$$

**Proof of Lemma 11.** If $K$ is the spherical kernel, note that $\widehat{p}_{h,X}(u)=n^{-1}\sum_{i=1}^n B_i(u)$, where

$$B_i=h^{-d}K\left(\frac{u-X_i}{h}\right)=\frac{I_{B(u,h)}(X_i)}{(h^d v_d)},$$

with $I_{B(u,h)}(\cdot)$ denoting the indicator function of the ball $B(u,h)$. Let $\sigma^2(u,h)=\operatorname{Var}(B_i(u))$ and $\mu_3(u,h)=\mathbb{E}|B_i(u)-\mu(u,h)|^3$ where $\mu(u,h)=\mathbb{E}(B_i(u))=p_h(u)$. Finally, let $p_{u,h}=P(B(u,h))$. Then,

$$\sigma^2(u,h)=\frac{p_{u,h}(1-p_{u,h})}{(h^d v_d)^2} \tag{21}$$

and

$$\mu_3(u,h) = \frac{p_{u,h}(1-p_{u,h})\left[(1-p_{u,h})^2 + p_{u,h}^2\right]}{(h^d v_d)^3} \leq \frac{p_{u,h}(1-p_{u,h})}{(h^d v_d)^3},$$

where the last inequality holds since $(1-p_{u,h})^2 + p_{u,h}^2 \leq 1$, for all $u$ and $h$. As a result,

$$\frac{\mu_3(u,h)}{\sigma^3(u,h)} \leq (p_{u,h}(1-p_{u,h}))^{-1/2}.$$

By assumption, $h < h(\delta,\varepsilon)$ and $\varepsilon \leq \lambda/2$. In order to avoid trivialities, we further assume that $P(U_{h,\varepsilon}) > 0$. Then, uniformly over all $u$ in $U_{h,\varepsilon}$,

$$(\lambda-\varepsilon)v_d h^d \leq p_{u,h} \leq (\lambda+\varepsilon)v_d h^d$$

and

$$(1-p_{u,h}) \geq \delta.$$

Thus,

$$\frac{\mu_3(u,h)}{\sigma^3(u,h)} \leq \sqrt{\frac{1}{\delta v_d h^d(\lambda-\varepsilon)}} \leq \sqrt{\frac{2}{h^d \delta v_d \lambda}},$$

with the last inequality holding because of our assumption $\varepsilon \leq \lambda/2$. From (21), we then obtain

$$\frac{\delta(\lambda-\varepsilon)}{v_d h^d} \leq \sigma^2(u,h) \leq \frac{(\lambda+\varepsilon)}{v_d h^d}.$$

Thus,

$$\frac{a_1}{h^d} \leq \sigma^2(u,h) \leq \frac{a_2}{h^d},$$

where

$$a_1 = \frac{\delta\lambda}{2v_d} \quad \text{and} \quad a_2 = \frac{3\lambda}{2v_d}, \tag{22}$$

uniformly over $u \in U_{h,\varepsilon}$.

Writing $\sigma^2(u,h) = a(u,h)/h^d$ and using the Berry-Esséen bound (Wasserman, 2004, p. 78), we obtain

$$\sup_t \left| P\left(\frac{\sqrt{nh^d}(\widehat{p}_{h,X}(u) - p_h(u))}{a(u,h)} \leq t\right) - \Phi(t)\right| \leq \frac{33}{4}\frac{\mu_3(u,h)}{\sigma^3(u,h)\sqrt{n}} = \sqrt{\frac{C(\delta,\lambda)}{nh^d}},$$

where $\Phi$ is the cumulative distribution function of the standard Normal distribution.

Now,

$$\pi_h(u) = \mathbb{P}_X(\widehat{p}_{h,X}(u) > \lambda) = \mathbb{P}_X\left(\frac{\sqrt{nh^d}(\widehat{p}_{h,X}(u) - p_h(u))}{a(u,h)} > \frac{\sqrt{nh^d}(\lambda - p_h(u))}{a(u,h)}\right).$$

Hence,

$$1-\Phi\left(\frac{\sqrt{nh^d}(\lambda-p_h(u))}{a(u,h)}\right) - \frac{C(\delta,\lambda)}{\sqrt{nh^d}} \leq \pi_h(u) \leq 1-\Phi\left(\frac{\sqrt{nh^d}(\lambda-p_h(u))}{a(u,h)}\right) + \frac{C(\delta,\lambda)}{\sqrt{nh^d}}.$$

Using the fact that $u \in U_{h,\varepsilon}$, and taking advantage of the uniform bounds $a_1 \leq a(u,h) \leq a_2$, the previous inequalities imply

$$1-\Phi\left(\frac{\sqrt{nh^d}\varepsilon}{a_1}\right)-\frac{C(\delta,\lambda)}{\sqrt{nh^d}} \leq \pi_h(u) \leq 1-\Phi\left(-\frac{\sqrt{nh^d}\varepsilon}{a_2}\right)+\frac{C(\delta,\lambda)}{\sqrt{nh^d}}.$$

Using the inequalities

$$1-\Phi\left(\frac{\sqrt{nh^d}\varepsilon}{a_1}\right)=\Phi\left(-\frac{\sqrt{nh^d}\varepsilon}{a_1}\right) \geq \Phi\left(-\frac{\sqrt{nh^d}\varepsilon}{a_2}\right)$$

and

$$1-\Phi\left(-\frac{\sqrt{nh^d}\varepsilon}{a_2}\right)=\Phi\left(\frac{\sqrt{nh^d}\varepsilon}{a_2}\right) \leq \Phi\left(\frac{\sqrt{nh^d}\varepsilon}{a_1}\right),$$

we obtain the bounds

$$\Phi\left(-\frac{\sqrt{nh^d}\varepsilon}{a_2}\right)-\frac{C(\delta,\lambda)}{\sqrt{nh^d}} \leq \pi_h(u) \leq 1-\Phi\left(-\frac{\sqrt{nh^d}\varepsilon}{a_2}\right)+\frac{C}{\sqrt{nh^d}} \tag{23}$$

and

$$1-\Phi\left(\frac{\sqrt{nh^d}\varepsilon}{a_1}\right)-\frac{C(\delta,\lambda)}{\sqrt{nh^d}} \leq \pi_h(u) \leq \Phi\left(\frac{\sqrt{nh^d}\varepsilon}{a_1}\right)+\frac{C}{\sqrt{nh^d}}, \tag{24}$$

respectively. Thus, uniformly over all $\varepsilon \leq \lambda/2$ and all $h < h(\delta,\varepsilon)$, Equations (23) and (24) yield

$$\overline{A}_{h,\varepsilon} = 2\sup_{u\in U_{h,\varepsilon}} \pi_h(u)(1-\pi_h(u)) \quad \leq \quad 2\left(1-\Phi\left(-\frac{\sqrt{nh^d}\varepsilon}{a_2}\right)+\frac{C(\delta,\lambda)}{\sqrt{nh^d}}\right)^2,$$

and

$$\underline{A}_{h,\varepsilon} = 2\inf_{u\in U_{h,\varepsilon}} \pi_h(u)(1-\pi_h(u)) \quad \geq \quad 2\left(1-\Phi\left(\frac{\sqrt{nh^d}\varepsilon}{a_1}\right)-\frac{C(\delta,\lambda)}{\sqrt{nh^d}}\right)^2,$$

respectively, where $a_1$ and $a_2$ are given in (22).

**Proof of Lemma 12.** Letting $1_i = 1_{\{Z_i \in \widehat{L}_{X,h}\Delta\widehat{L}_{Y,h}\}}$, we have

$$\Xi_{\lambda,n}(h) = \frac{1}{n}\sum_{i=1}^{n} 1_i.$$

where, conditionally on $X$ and $Y$, the $1_i$'s are independent and identically distributed Bernoulli random variables with $\mathbb{E}_Z[1_i|X,Y] = P(\widehat{L}_{h,X}\Delta\widehat{L}_{h,Y})$. Thus

$$\begin{aligned}
\mathbb{V}\left[\Xi_{\lambda,n}(h)\right] &= \mathbb{E}_{X,Y,Z}\left[\Xi_n^2(h)\right]-\xi^2(h) \\
&= \tfrac{1}{n^2}\mathbb{E}_{XY}\left[\mathbb{E}_Z\left[\left(\textstyle\sum_{i=1}^n 1_i+\sum_{j\neq k}1_j1_k\right)|X,Y\right]\right]-\xi^2(h) \\
&= \tfrac{\xi_{\lambda,n}(h)}{n}+\tfrac{n-1}{2n}\mathbb{E}_{X,Y}\left[P^2(\widehat{L}_{h,X}\Delta\widehat{L}_{h,Y})\right]-\xi^2(h) \\
&\leq \tfrac{\xi_{\lambda,n}(h)}{n}+\tfrac{n-1}{2n}\mathbb{E}_{X,Y}\left[P(\widehat{L}_{h,X}\Delta\widehat{L}_{h,Y})\right]-\xi^2(h) \\
&= \tfrac{\xi_{\lambda,n}(h)}{n}+\tfrac{n-1}{2n}\xi_{\lambda,n}(h)-\xi^2(h) \\
&= \xi_{\lambda,n}(h)\left(\tfrac{n+1}{2n}-\xi_{\lambda,n}(h)\right).
\end{aligned}$$

**Proof of Lemma 13.**

Let $\xi(h,X,Y) = \mathbb{E}_Z[\Xi_{\lambda,n}(h)|X,Y]$ and let $A_{h,\varepsilon}$ be the event given in (20), where $\varepsilon, h > 0$, so that $\mathbb{P}_{X,Y}(\mathcal{A}_{h,\varepsilon}^c) \leq 2K_1 \exp\left\{-nK_2 h^d \varepsilon^2\right\}$ by (1). Then, we can write

$$\mathbb{P}_{X,Y,Z}\left(\left|\Xi_{\lambda,n}(h) - \xi_{\lambda,n}(h)\right| > t\right) = \mathbb{P}_{X,Y,Z}\left(\left|\Xi_{\lambda,n}(h) - \xi(h,X,Y) + \xi(h,X,Y) - \xi_{\lambda,n}(h)\right| > t\right),$$

which is therefore upper bounded by

$$\mathbb{P}_{X,Y,Z}\left(\left|\Xi_{\lambda,n}(h) - \xi(h,X,Y) + \xi(h,X,Y) - \xi_{\lambda,n}(h)\right| > t; \mathcal{A}_{h,\varepsilon}\right) + 2K_1 \exp\left\{-nK_2 h^d \varepsilon^2\right\}.$$

The first term in the previous expression is no larger than the sum of

$$\mathbb{E}_{X,Y}\left[\mathbb{P}_Z\left(\left|\Xi_{\lambda,n}(h) - \xi(h,X,Y)\right| > t\eta \middle| X,Y\right); \mathcal{A}_{h,\varepsilon}\right], \tag{25}$$

and

$$\mathbb{P}_{X,Y}\left(\left|\xi(h,X,Y) - \xi_{\lambda,n}(h)\right| > t(1-\eta); \mathcal{A}_{h,\varepsilon}\right), \tag{26}$$

for any $\eta \in (0,1)$. We will first show that, if (9) is satisfied, the probability (26) is zero. Indeed, first observe that

$$\mathbb{E}_Z[\Xi_{\lambda,n}(h)|X,Y] = P(\widehat{L}_{h,X} \Delta \widehat{L}_{h,Y})$$

and that, on $\mathcal{A}_{h,\varepsilon}$,

$$\begin{aligned}
\widehat{L}_{h,X} \Delta \widehat{L}_{h,Y} &= \{u\colon \widehat{p}_{h,X}(u) > \lambda, \widehat{p}_{h,Y}(u) \leq \lambda\} \cup \{u\colon \widehat{p}_{h,X}(u) \leq \lambda, \widehat{p}_{h,Y}(u) > \lambda\} \\
&\subseteq \{u\colon p_h(u) > \lambda - \varepsilon, p_h(u) \leq \lambda + \varepsilon\} \\
&= \{u\colon |p_h(u) - \lambda| \leq \varepsilon\} \\
&= U_{h,\varepsilon},
\end{aligned}$$

Therefore, on $\mathcal{A}_{h,\varepsilon}$,

$$\xi(h,X,Y) = \mathbb{E}_Z[\Xi_{\lambda,n}(h)|X,Y] \leq r_{h,\varepsilon} \leq t(1-\eta). \tag{27}$$

By part 2 of Theorem 10, (9) further implies that $t(1-\eta) \geq \xi_{\lambda,n}(h)$. As a result, on $\mathcal{A}_{h,\varepsilon}$, $\left|\xi(h,X,Y) - \xi_{\lambda,n}(h)\right| \leq t(1-\eta)$, which yields

$$\mathbb{P}_{X,Y}\left(\left|\xi(h,X,Y) - \xi_{\lambda,n}(h)\right| > t(1-\eta); \mathcal{A}_{h,\varepsilon}\right) = 0,$$

as claimed.

We now proceed to bound from above (25). Since

$$\Xi_{\lambda,n}(h) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Z_i \in \widehat{L}_{h,X} \Delta \widehat{L}_{h,Y}\}},$$

Bernstein's inequality (see, for instance, Massart, 2006, Proposition 2.9) yields that, for any $t > 0$ and conditionally on $X$ and $Y$,

$$\mathbb{P}_Z\left(\left|\Xi_{\lambda,n}(h) - \xi(h,X,Y)\right| > t\eta \middle| X,Y\right) \leq \exp\left\{-9\sigma^2(X,Y,h) g\left(\frac{nt\eta}{3\sigma^2(X,Y,h)}\right)\right\} \tag{28}$$

where $g(u) = 1+u-\sqrt{1+2u}$ for all $u > 0$, and

$$\sigma^2(X,Y,h) = \mathrm{Var}_Z[\Xi_{\lambda,n}(h)|X,Y].$$

It is easy to see that

$$\sigma^2(X,Y,h) \leq \mathbb{E}_Z\left[\Xi_{\lambda,n}(h)|X,Y\right] = n\xi(h,X,Y)$$

and, therefore, restricting to the event $\mathcal{A}_{h,\varepsilon}$, $\sigma^2(X,Y,h) \leq nt(1-\eta)$, just like in (27).

Using the fact that $e^{-9xg\left(\frac{nt}{3x}\right)}$ is increasing in $x$ for $x > 0$, we conclude that, on the event $\mathcal{A}_{h,\varepsilon}$, the right hand side of (28) is bounded from above by

$$\exp\left\{-9nt(1-\eta)g\left(\frac{\eta}{3(1-\eta)}\right)\right\},$$

which is independent of $X$ and $Y$. Thus, the previous expression is an upper bound for (25) and, therefore, for $\mathbb{P}_{X,Y,Z}\left(\left|\Xi_{\lambda,n}(h)-\xi_{\lambda,n}(h)\right| > t\right)$. The claim now follows from simple algebra.

**Proof of Theorem 14.**

1. The proof is almost the same as the proof of part 1 of Theorem 10 and is therefore omitted.

2. Let $\mathcal{A}_{h,\widetilde{\varepsilon}}$ denote the event

$$\max\left\{||\widehat{p}_{h,X}-p_h||_\infty, |\lambda_{h,\alpha}-\widehat{\lambda}_{h,\alpha,X}|, ||\widehat{p}_{h,Y}-p_h||_\infty, |\lambda_{h,\alpha}-\widehat{\lambda}_{h,\alpha,Y}|\right\} \leq \widetilde{\varepsilon}, \tag{29}$$

where $\widetilde{\varepsilon} = \varepsilon(A\kappa_3+1)+A\kappa_3/n$. Then, using (1), (5) and the fact that $\varepsilon < \widetilde{\varepsilon}$, the union bound yields

$$\mathbb{P}_{X,Y}(\mathcal{A}^c_{h,\widetilde{\varepsilon}}) \leq 4K_1 e^{-K_2 nh^d\varepsilon^2} + 16ne^{-n\varepsilon^2/32} \equiv C(h,\varepsilon,n) \tag{30}$$

Now, on $\mathcal{A}_{h,\widetilde{\varepsilon}}$, $\{u : \widehat{p}_{h,X}(u) > \widehat{\lambda}_{h,\alpha,X}, \widehat{p}_{h,Y}(u) \leq \widehat{\lambda}_{h,\alpha,Y}\}$ is a subset of

$$\{u : p_h(u) > \widehat{\lambda}_{h,\alpha,X}-\widetilde{\varepsilon}, p_h(u) \leq \widehat{\lambda}_{h,\alpha,Y}+\widetilde{\varepsilon}\},$$

which is equal to

$$\{u : |p_h(u)-\lambda_{h,\alpha}| \leq 2\widetilde{\varepsilon}\} = U_{h,\widetilde{\varepsilon},\alpha}.$$

Therefore, $\mathrm{sign}(\widehat{p}_{h,X}(u)-\widehat{\lambda}_{h,\alpha,X}) = \mathrm{sign}(p_h(u)-\lambda_{h,\alpha})$ for all $u \notin U_{h,2\widetilde{\varepsilon},\alpha}$. Next, just like in the proof of part 2 of theorem 10, using this fact and the result of the first part we have that $\xi_{\alpha,n}(h)$ is no larger than

$$\mathbb{E}_{X,Y,Z}[\widehat{P}_Z(\{u\colon \widehat{p}_{h,X}(u) > \widehat{\lambda}_{h,\alpha,X}\}\Delta\{u\colon \widehat{p}_{h,Y}(u) > \widehat{\lambda}_{h,\alpha,Y}\})1_{\mathcal{A}_{h,\widetilde{\varepsilon}}}(X,Y)] + \mathbb{P}_{X,Y}(\mathcal{A}^c_{h,\widetilde{\varepsilon}}).$$

The previous expression can be written as

$$2\int_{\mathbb{R}^d} \mathbb{P}_{X,Y}(\{\widehat{p}_{h,X}(u) > \widehat{\lambda}_{h,\alpha,X}, \widehat{p}_{h,Y}(u) \leq \widehat{\lambda}_{h,\alpha,Y}\} \cap \mathcal{A}_{h,\widetilde{\varepsilon}})dP(u) + \mathbb{P}_{X,Y}(\mathcal{A}^c_{h,\widetilde{\varepsilon}}),$$

which is less than

$$2\int_{U_{h,2\widetilde{\varepsilon},\alpha}} \mathbb{P}_{X,Y}(\{\widehat{p}_{h,X}(u) > \widehat{\lambda}_{h,\alpha,X}, \widehat{p}_{h,Y}(u) \leq \widehat{\lambda}_{h,\alpha,Y}\} \cap \mathcal{A}_{h,\widetilde{\varepsilon}})dP(u) + C(h,\varepsilon,n).$$

This quantity is bounded from above by

$$2\int_{U_{h,2\widetilde{\varepsilon},\alpha}} \mathbb{P}_{X,Y}(\widehat{p}_{h,X}(u) > \widehat{\lambda}_{h,\alpha,X}, \widehat{p}_{h,Y}(u) \leq \widehat{\lambda}_{h,\alpha,Y}) dP(u) + C(h,\varepsilon,n),$$

which is finally smaller than

$$2\int_{U_{h,2\widetilde{\varepsilon},\alpha}} \pi_{h,\alpha}(u)(1-\pi_{h,\alpha}(u)) dP(u) + C(h,\varepsilon,n) \leq P(U_{h,2\widetilde{\varepsilon},\alpha})\overline{A}_{h,\varepsilon,\alpha} + C(h,\varepsilon,n).$$

As for the lower bound, from the result of first part we obtain, trivially,

$$\begin{aligned} \xi_{\alpha,n}(h) &\geq 2\int_{U_{h,2\widetilde{\varepsilon},\alpha}} \pi_{h,\alpha}(u)(1-\pi_{h,\alpha}(u)) dP(u) \\ &\geq P(U_{h,2\widetilde{\varepsilon},\alpha})\overline{A}_{h,\varepsilon,\alpha}. \end{aligned}$$

3. To compute an upper bound for $\overline{A}_{h,\varepsilon,\alpha}$ and a lower bound for $\underline{A}_{h,\varepsilon,\alpha}$, we use the Berry-Esséen bound and the stated assumptions. The proof is very similar to the proof of lemma 11, except that the result holds only on the event $\mathcal{A}_{h,\widetilde{\varepsilon}}$. Therefore, we only provide a sketch of the arguments.

The assumptions that $\widetilde{\varepsilon} \leq \inf_h \frac{\lambda_{\alpha,h}}{4}$, implies that, for any $u \in U_{h,2\widetilde{\varepsilon},\alpha}$,

$$\frac{1}{h^d}\frac{\delta\lambda_{\alpha,h}}{2v_d} \leq \frac{\delta(\lambda_{\alpha,h}-2\widetilde{\varepsilon})}{h^d v_d} \leq \sigma^2(u,h) \leq \frac{(\lambda_{\alpha,h}+2\widetilde{\varepsilon})}{h^d v_d} \leq \frac{1}{h^d}\frac{3\lambda_{\alpha,h}}{2v_d}.$$

Because of this and the fact that, on $\mathcal{A}_{h,\widetilde{\varepsilon}}$, $|p_h(u) - \widehat{\lambda}_{h,\alpha,X}| \leq 3\widetilde{\varepsilon}$ for all $u \in U_{h,2\widetilde{\varepsilon},\alpha}$, the same Berry-Esseen arguments used in the proof of lemma 11 yield

$$1-\Phi\left(\frac{3\widetilde{\varepsilon}\sqrt{nh^d}}{a_1}\right) - \frac{C(\delta,\lambda_{h,\alpha})}{\sqrt{nh^d}} \leq \pi_{h,\alpha,\widetilde{\varepsilon}}(u) \leq 1-\Phi\left(-\frac{3\widetilde{\varepsilon}\sqrt{nh^d}}{a_2}\right) + \frac{C(\delta,\lambda_{h,\alpha})}{\sqrt{nh^d}}.$$

where $\pi_{h,\alpha,\widetilde{\varepsilon}}(u) = \mathbb{P}_X\left(\{\widehat{p}_{h,X}(u) > \widehat{\lambda}_{h,\alpha,X}\} \cap \mathcal{A}_{h,\widetilde{\varepsilon}}\right)$, $a_1 = \delta\lambda_{h,\alpha}/(2v_d)$, $a_2 = 3\lambda_{h,\alpha}/(2v_d)$, and $C(\delta,\lambda_{h,\alpha}) = \frac{33}{4}\sqrt{\frac{2}{\delta v_d \lambda_{h,\alpha}}}$. Now notice that

$$\pi_{h,\alpha}(u) \geq \pi_{h,\alpha,\widetilde{\varepsilon}}(u) \geq 1-\Phi\left(\frac{3\widetilde{\varepsilon}\sqrt{nh^d}}{a_1}\right) - \frac{C(\delta,\lambda_{h,\alpha})}{\sqrt{nh^d}}$$

and

$$\pi_{h,\alpha}(u) \leq \pi_{h,\alpha,\widetilde{\varepsilon}}(u) + P(\mathcal{A}^c_{h,\widetilde{\varepsilon}}) \leq 1-\Phi\left(-\frac{3\widetilde{\varepsilon}\sqrt{nh^d}}{a_2}\right) + \frac{C(\delta,\lambda_{h,\alpha})}{\sqrt{nh^d}} + C(h,\varepsilon,n).$$

where $C(h,\varepsilon,n)$ is defined in (30). Therefore,

$$\overline{A}_{h,\varepsilon,\alpha} \leq 2\left(1-\Phi\left(-\frac{3\widetilde{\varepsilon}\sqrt{nh^d}}{a_2}\right) + \frac{C(\delta,\lambda_{h,\alpha})}{\sqrt{nh^d}} + C(h,\varepsilon,n)\right)^2,$$

and

$$\underline{A}_{h,\varepsilon,\alpha} \geq 2\left(1-\Phi\left(\frac{3\widetilde{\varepsilon}\sqrt{nh^d}}{a_1}\right) - \frac{C(\delta,\lambda_{h,\alpha})}{\sqrt{nh^d}} - C(h,\varepsilon,n)\right)^2.$$

**Proof of Theorem 19.** (1) Since the sample space is compact, $\mu(S) < \infty$, where $S$ denotes the support of $P$ and $\mu$ denotes the Lebesgue measure. Therefore, we obtain the inequality

$$\begin{aligned}\Gamma_n(h) &\leq \frac{\mu(S)}{2}||\widehat{p}_{h,X} - \widehat{p}_{h,Y}||_\infty \leq \frac{\mu(S)}{2}||\widehat{p}_{h,X} - p_h||_\infty + \frac{\mu(S)}{2}||\widehat{p}_{h,Y} - p_h||_\infty \\ &\stackrel{d}{=} \mu(S)||\widehat{p}_{h,X} - p_h||_\infty.\end{aligned}$$

Next, let $C = \frac{(\mu(S))^2(a+2)}{K_2}$, so that for $n > K_1$

$$t_h > \sqrt{\frac{\mu(S)^2 \log(n^{a+1}K_1)}{K_2 n h^d}}.$$

Then,

$$\begin{aligned}\mathbb{P}_{X,Y}\left(\Gamma_n(h) > t_h \text{ for some } h \in \mathcal{H}_n\right) &\leq \mathbb{P}_X\left(||\widehat{p}_{h,X} - p_h||_\infty > \frac{t_h}{\mu(S)} \text{ for some } h \in \mathcal{H}_n\right) \\ &\leq \sum_{h\in\mathcal{H}_n} \mathbb{P}_X\left(||\widehat{p}_{h,X} - p_h||_\infty > \frac{t_h}{\mu(S)}\right) \\ &\leq \sum_{h\in\mathcal{H}_n} K_1 \exp\{-K_2 n t_h^2 h^d/(\mu(S)^2)\} \\ &\leq H n^a \frac{1}{n^{a+1}} = \frac{H}{n} \\ &\leq \delta,\end{aligned}$$

where the third inequality stems from (1) and the assumption that $n \geq n_0$ is large enough, and the last inequality follows from the assumed condition on $\delta$.
(2) Consider any $h \leq h_*$. Note that

$$\Gamma_n(h) \geq \Gamma_{n,S}(h) \equiv \frac{1}{2}\int_S |\widehat{p}_{h,X}(u) - \widehat{p}_{h,Y}(u)|du.$$

Let

$$D(u) = \sqrt{nh^d}(\widehat{p}_{h,X}(u) - \widehat{p}_{h,Y}(u)).$$

The variance of $D(u)$ is

$$\begin{aligned}\mathrm{Var}\left(\sqrt{nh^d}(\widehat{p}_{h,X}(u) - \widehat{p}_{h,Y}(u))\right) &= nh^d\left(\mathrm{Var}(\widehat{p}_{h,X}(u)) + \mathrm{Var}(\widehat{p}_{h,Y}(u))\right) \\ &= 2nh^d \mathrm{Var}(\widehat{p}_{h,X}(u)) \\ &= 2nh^d \mathrm{Var}\left(\frac{1}{nh^d v_d}\sum_{i=1}^n I(||X_i - u|| \leq h)\right) \\ &= \frac{2n^2h^d}{n^2h^{2d}v_d^2}\mathrm{Var}(I(||X_i - u|| \leq h)) \\ &= \frac{2}{v_d^2 h^d}P(B(u,h))(1 - P(B(u,h))).\end{aligned}$$

Now, for $u \in S$, by (10),

$$P(B(u,h))(1-P(B(u,h))) \leq P(B(u,h)) \leq a_2 h^d v_d$$

and

$$P(B(u,h))(1-P(B(u,h))) \geq P(B(u,h))\delta \geq a_1 h^d v_d \delta.$$

Hence,

$$2a_1 v_d \delta \leq \mathrm{Var}(D(u)) \leq 2a_2 v_d, \quad \forall u \in S,$$

which shows that the variance of $D(u)$ is bounded above and below by positive functions that do not depend on $h$. By a similar calculation, $\mathrm{Cov}(D(u),D(v))$ is bounded above and below by functions that do not depend on $h$, for all $u,v \in S$.

Now, for any $u$,

$$D(u) = D_1(u) - D_2(u) \equiv \sqrt{nh^d}(P_n - P)(f_u) - \sqrt{nh^d}(Q_n - P)(f_u)$$

where $P_n$ is the empirical measure based on $X_1,\ldots,X_n$, $Q_n$ is the empirical measure based on $Y_1,\ldots,Y_n$, and $f_u(\cdot) = h^{-d}K(||u-\cdot||/h)$. Note that $D_1$ and $D_2$ are independent, mean 0 stochastic processes. We can regard $\{\sqrt{nh^d}(P_n-P)(f) : f \in \mathcal{F}\}$ as an empirical process, where $\mathcal{F} = \{f_u : u \in S\}$ and similarly for $\{\sqrt{nh^d}(Q_n-P)(f) : f \in \mathcal{F}\}$. For fixed $h$, the collection $\mathcal{F}$ is a Donsker class. Hence, for every $u \in S$, $D_1(u)$ and $D_2(u)$ converge to two independent mean 0 Gaussian processes. By the continuous mapping theorem, for every $u \in S$, $D(u)$ converges to a mean 0 Gaussian process $\mathbb{G}$ with some covariance kernel $\kappa$. By the calculations above, there exist positive bounded functions $r(u,v) \leq s(u,v)$ such that $r(u,v) \leq \kappa(u,v) \leq s(u,v)$ and such that neither $r$ nor $s$ depend on $h$. Hence

$$\begin{aligned}
\mathbb{P}_{X,Y}\left(\Gamma_n(h) \geq t\sqrt{\frac{1}{nh^d}}\right) &\geq \mathbb{P}_{X,Y}\left(\Gamma_{n,S}(h) \geq t\sqrt{\frac{1}{nh^d}}\right) = \mathbb{P}_{X,Y}\left(\sqrt{nh^d}\Gamma_{n,S}(h) \geq t\right) \\
&= \mathbb{P}_{X,Y}\left(\frac{1}{2}\int_S |D(u)|du \geq t\right) \\
&= \mathbb{P}\left(\frac{1}{2}\int |\mathbb{G}(u)|du \geq t\right) + o(1),
\end{aligned}$$

where the last probability is the law of the Gaussian process $\mathbb{G}$. The $o(1)$ term is less than $\delta/2$ when $n \geq n_0$. Since $\mathbb{G}$ has strictly positive variance, $\mathbb{P}(\int|\mathbb{G}| \geq 0) = 1$. Clearly, $\mathbb{P}(\int|\mathbb{G}| \geq 2t)$ is decreasing in $t$. Hence, for each $\delta$, there is a positive $t$ such that $\mathbb{P}\left(\frac{1}{2}\int|\mathbb{G}| \geq t\right) \geq 1-\delta/2$.
(3) The proof of this part is straightforward and is omitted.

## References

S. Ben-David, U. von Luxburg, and D. Pál. A sober look at clustering stability. In *COLT*, pages 5–19, 2006.

A. Ben-Hur, A. Elisseef, and I. Guyon. A stability based method for discovering structure in clustered data. *Pacific Symposium on Biocomputing*, pages 6–17, 2002.

B. Cadre, B. Pelletier, and P. Pudlo. Clustering by estimation of density level sets at a fixed probability, 2009. Manuscript available at http://w3.bretagne.ens-cachan.fr/math/people/benoit.cadre/fichiers/tlevel.pdf.

G. Carlsson and F. Memoli. Characterization, stability and convergence of hierarchical clustering methods. *Journal of Machine Learning Research*, 11:1425–1470, 2010.

K. Chaudhuri and S. Dasgupta. Rates of convergence for the cluster tree. In *Neural Information Processing Systems (NIPS)*, December 2010.

P. Chaudhuri and S.J. Marron. Scale space view of curve estimation. *Annals of Statistics*, 28(2): 408–428, 2000.

A. Cuevas and A. Rodríguez-Casal. On boundary estimation. *Advances in Applied Probability*, 36: 340–354, 2004.

A. Cuevas, W. González-Manteiga, and A. Rodríguez-Casal. Plug-in estimation of general level sets. *Australian and New Zealand Journal of Statistics*, 48(1):7–19, 2006.

P. Deheuvels, J. Einmahl, D. Mason, and F. F. Ruymgaart. The almost sure behavior of maximal and minimal multivariate kn-spacings. *Journal of Multivariate Analysis*, 24:155–176, 1988.

L. Devroye, , L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

B. Fischer and J. M. Buhmann. Bagging for path-based clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:1411–1415, 2003.

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

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

C. Jackson. Displaying uncertainty with shading. *The American Statistician*, 62(4):340–347, 2008.

S. Kpotufe and U. von Luxburg. Pruning nearest neighbor cluster trees. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. Omnipress, 2011.

T. Lange, V. Roth, M. Braun, and J.Buhmann. Stability-based validation of clustering solutions. *Neural Computation*, 16:1299–1323, 2004.

E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *Annals of Statistics*, 27(6): 1808–1829, 1999.

P. Massart. *Concentration Inequalities and Model Selection*. Number 1896 in Springer Lecture Notes in Mathematics. Springer, 2006.

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72:417–473, 2010.

M. Penrose. *Random Geometric Graphs*. Oxford University Press, 2003.

W. Polonik. Measuring mass concentration and estimating density contour clusters–an excess mass approach. *Annals of Statistics*, 32(3):855–881, 1995.

P. Rigollet and R. Vert. Optimal rates for plug-in estimators of density level sets. *Bernoulli*, 15: 1154–1178, 2009.

A. Rinaldo and L. Wasserman. Generalized density clustering. *The Annals of Statistics*, 38(5): 2678–2722, 2010.

D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Series in Probability and Mathematical Statistics: Applied Probability and Statistics. Wiley, 1992.

A. Singh, C. Scott, and R. Nowak. Adaptive hausdorff estimation of level sets. *The Annals of Statistics*, 37(5B):2760–2782, 2009.

I. Steinwart. Adaptive density level set clustering. In *Proceedings of the Twenty-Fourth Annual Conference on Learning Theory (COLT'11)*. Omnipress, 2011.

W. Stuetzle and R. Nugent. A generalized single linkage method for estimating the cluster tree of a density. *Journal of Computational and Graphical Statistics*, 19:1–22, 2009.

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

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32 (1):135–166, 2004.

U. von Luxburg. Clustering stability: An overview. *Foundations and Trends in Machine Learning*, 2:235–274, 2009.

M. P. Wand. Fast computation of multivariate kernel estimators. *Journal of Computational and Graphical Statistics*, 3(4):433–445, 1994.

L. Wasserman. *All of Statistics*. Springer, New York, N.Y., 2004.