Density Level Set Estimation on Manifolds with DBSCAN

Heinrich Jiang [1]

Abstract

We show that DBSCAN can estimate the connected components of the λ-density level set $\{x : f(x) \geq \lambda\}$ given n i.i.d. samples from an unknown density f. We characterize the regularity of the level set boundaries using parameter $\beta > 0$ and analyze the estimation error under the Hausdorff metric. When the data lies in $\mathbb{R}^D$ we obtain a rate of $\widetilde{O}(n^{-1/(2\beta+D)})$, which matches known lower bounds up to logarithmic factors. When the data lies on an embedded unknown d-dimensional manifold in $\mathbb{R}^D$, then we obtain a rate of $\widetilde{O}(n^{-1/(2\beta+d\cdot\max\{1,\beta\})})$. Finally, we provide adaptive parameter tuning in order to attain these rates with no a priori knowledge of the intrinsic dimension, density, or β.

1. Introduction

DBSCAN (Ester et al., 1996) is one of the most popular clustering algorithms amongst practitioners and has had profound success in a wide range of data analysis applications. However, despite this, its statistical properties have not been fully understood. The goal of this work is to give a theoretical analysis of the procedure and to the best of our knowledge, provide the first analysis of density level-set estimation on manifolds. We also contribute ideas to related areas that may be of independent interest.

DBSCAN aims at discovering clusters which turn out to be the high-density regions of the dataset. It takes in two hyperparameters: minPts and ε. It defines a point as a *core-point* if there are at least minPts sample points in its ε-radius neighborhood. The points within the ε-radius neighborhood of a core-point are said to be *directly reachable* from that core-point. Then, a point q is *reachable* from a core-point p if there exists a path from q to p where each point is directly reachable from the next point. It is now clear that this definition of reachable gives a partitioning of the dataset (and remaining points not reachable from any core-point are considered noise). This partitioning is the clustering that is returned by DBSCAN.

The problem of analyzing DBSCAN has recently been explored in (Sriperumbudur & Steinwart, 2012). Their analysis is for a modified version of DBSCAN and is not focused on estimating a fixed density level. Their results have many desirable properties, but are not immediately applicable for what this paper tries to address. Using recent developments in topological data analysis along with some tools we develop in this paper, we show that it is now possible to analyze the original procedure.

The clusters DBSCAN aims at discovering can be viewed as approximations of the connected components of the level sets $\{x : f(x) \geq \lambda\}$ where f is the density and λ is some density level. We provide the first comprehensive analysis in tuning minPts and ε to estimate the density level set for a particular level. Here, the density level λ is known to the algorithm while the density remains unknown. Density level set estimation has been studied extensively. e.g., (Carmichael et al., 1968; Hartigan, 1975; Polonik, 1995; Cuevas & Fraiman, 1997; Walther, 1997; Tsybakov et al., 1997; Baıllo et al., 2001; Cadre, 2006; Willett & Nowak, 2007; Biau et al., 2008; Rigollet & Vert, 2009; Maier et al., 2009; Singh et al., 2009; Rinaldo & Wasserman, 2010; Steinwart, 2011; Rinaldo et al., 2012; Steinwart et al., 2015; Chen et al., 2016; Jiang, 2017). However approaches that obtain state-of-art consistency results are largely unpractical (i.e. unimplementable). Our work shows that in actuality, DBSCAN, a procedure known for decades and has since been used widely, can achieve the strongest known results. Also, unlike much of the existing work, we show that DBSCAN can also recover the connected components of the level sets separately and bijectively.

Our work begins with the insight that DBSCAN behaves like an ε-neighborhood graph, which is different from, but related to the k-nearest neighbor graph. The latter has been heavily used for cluster-tree estimation (Chaudhuri & Dasgupta, 2010; Stuetzle & Nugent, 2010; Kpotufe & von Luxburg, 2011; Chaudhuri et al., 2014; Jiang & Kpotufe, 2017) and in this paper we adapt some of these ideas for ε-neighborhood graphs.

[1] Google. Correspondence to: Heinrich Jiang <heinrich.jiang@gmail.com>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

Cluster-tree estimation aims at discovering the hierarchical tree structure of the connected-components as the levels vary. Balakrishnan et al. (2013) extends results by Chaudhuri & Dasgupta (2010) to the setting where the data lies on a lower dimensional manifold and provide consistency results depending on the lower dimension and independent of the ambient dimension. Here we are instead interested in how to set minPts and ε in order to estimate a particular level and provide rates on the Hausdorff distance error. This is different from works on cluster tree estimation which focuses on how to recover the tree structure rather than recovering a particular level. In that regard, we also require density estimation bounds in order to get a handle on the true density-levels and the empirical ones.

Dasgupta & Kpotufe (2014) gives us optimal high-probability finite-sample k-NN density estimation bounds which hold *uniformly*; this is key to obtaining optimal level-set estimation rates under the Hausdorff error. Much of the previous works on density level-set estimation, e.g. (Rigollet & Vert, 2009) provide rates under risk measures such as symmetric set difference. These metrics are considerably weaker than the Hausdorff metric; the latter is a uniform guarantee. There are such bounds for the histogram density estimator. This allowed Singh et al. (2009) to obtain optimal rates under Hausdorff metric, while having a fully adaptive procedure. This was a significant breakthrough for level set estimation, as discussed by Chazal et al. (2015). We believe this to be the strongest consistency results obtained thus far. However, a downside is that the histogram density estimator has little practical value. Here, aided with the desired bounds on the k-NN density estimator, we can actually obtain similar results to Singh et al. (2009) but with the clearly practical DBSCAN.

We extend the k-NN density estimation results of Dasgupta & Kpotufe (2014) to the manifold case, as the bulk our analysis is about the more general case that the data lies on a manifold. Density-based procedures perform poorly in high-dimensions since the number of samples required increases exponentially in the dimension– the so called curse of dimensionality. Thus, the consequences of handling the manifold case are of practical significance. Since the estimation rates we obtain depend only on the intrinsic dimension, it explains why DBSCAN can do well in high dimensions if the data has low intrinsic dimension (i.e. the manifold hypothesis). Given the modern capacity of systems to collect data of increasing complexity, it has become ever more important to understand the feasibility of *practical* algorithms in high dimensions.

To analyze DBSCAN, we write minPts and ε in terms of the d, unknown manfold dimension; k, which controls the density estimator; and λ, which determines which level to estimate. We assume knowledge of λ with the goal of estimating the λ-level set of the density. We give a range of k in terms of n and corresponding consistency guarantees and estimation rates for such choices. We then adaptively tune d and k in order to attain close to optimal performance with no a priori knowledge of the distribution. Adaptivity is highly desirable because it allows for automatic tuning of the hyper-parameters, which is a core tenet of unsupervised learning. To solve for the unknown dimension, we use an estimator from Farahmand et al. (2007), which we show to have considerably better finite-sample behavior than previously thought. More details and discussion of related works is in the main text. We then provide a new method of choosing k such that it will asymptotically approach a value that provides near-optimal level set estimation rates.

2. Overview

We start by analyzing the procedure under the manifold assumption. The end of the paper will discuss the full-dimensional setting. The bulk of our contribution lies in analyzing the former situation, while the analysis of the latter uses a subset of those techniques.

- Section 3 proves that the clusters returned by DBSCAN are close to the connected components of certain ε-neighborhood graphs (Lemma 2). This is significant because these graphs can be shown to estimate density level sets.
- Section 4 introduces the manifold setting and provides supporting results including k-nearest neighbor density estimation bounds (Lemma 5 and Lemma 6) that are useful later on.
- Section 5 provides a range of parameter settings under which for each true cluster, there exists a corresponding cluster returned by DBSCAN (Lemma 7 and Lemma 8), and a rate for the Hausdorff distance between them (Theorem 1).
- Section 6 shows how one can apply DBSCAN a second time to remove false clusters from the first application, thus completing a bijection between the estimates and the true clusters (Theorem 2).
- Section 7 explains how to adaptively tune the parameters so that they fall within the theoretical ranges. The main contributions of this section are a stronger result about a known k-nearest neighbor based approach to estimating the unknown dimension (Theorem 3) and a new way to tune k to approach an optimal choice of k (Theorem 4).
- Section 8 gives the result when the data lives in $\mathbb{R}^D$ without the manifold assumption.

3. The connection to neighborhood graphs

This section is dedicated towards the understanding of the clusters produced by DBSCAN. The algorithm can be found in (Ester et al., 1996) and is not shown here since Lemma 1 characterizes what DBSCAN returns.

We have n i.i.d. samples $X = \{x_1, ..., x_n\}$ drawn from a distribution $\mathcal{F}$ over $\mathbb{R}^D$.

Definition 1. *Define the k-NN radius of $x \in \mathbb{R}^D$ as*

$$r_k(x) := \inf\{r > 0 : |X \cap B(x,r)| \geq k\},$$

where $B(x, r)$ denotes the Euclidean ball of radius r centered at x. Let $G(k, \varepsilon)$ denote the ε-neighborhood level graph of X with vertices $\{x \in X : r_k(x) \leq \varepsilon\}$ and an edge between x and x' iff $||x - x'|| \leq \varepsilon$.

Remark 1. *This is slightly different from ε-neighborhood graph, which includes all vertices. Here we exclude vertices below certain empirical density level (i.e. $r_k(x) > \varepsilon$).*

The next definition is relevant to DBSCAN and is from (Ester et al., 1996) but in the notation of Definition 1.

Definition 2. *The following is with respect to fixed $\varepsilon > 0$ and minPts $\in \mathbb{N}$.*

- *p is a core-point if $r_{minPts}(p) \leq \varepsilon$.*
- *q is directly density-reachable from p if $|p - q| \leq \varepsilon$ and p is a core-point.*
- *q is density-reachable from p if there exists a sequence $q = p_1, p_2, ..., p_m = p$ such that p_i is directly density-reachable from p_{i+1} for $i = 1, .., m - 1$.*

The following result is paraphrased from Lemmas 1 and 2 from (Ester et al., 1996), which characterizes the clusters learned by DBSCAN.

Lemma 1. *(Ester et al., 1996) Let $\mathcal{C}$ be the clusters returned by DBSCAN(minPts, ε). For any core-point x, there exists $C \in \mathcal{C}$ with $x \in C$. On the other hand, for any $C \in \mathcal{C}$, there exists core-point x such that $C = \{x' : x'$ is density-reachable from $x\}$.*

We now show the following result relating the ε-neighborhood level graphs and the clusters obtained from DBSCAN. Such an interpretation of DBSCAN has been given in previous works such as Campello et al. (2015).

Lemma 2 (DBSCAN and ε-neighborhood level graphs)**.** *Let $\mathcal{C}$ be the clusters obtained from DBSCAN$(minPts, \varepsilon)$ on X. Let $\mathcal{K}$ be the connected components of $G(minPts, \varepsilon)$. Then, there exists a one-to-one correspence between $\mathcal{C}$ and $\mathcal{K}$ such that if $C \in \mathcal{C}$ and $K \in \mathcal{K}$ correspond, then*

$$K \subseteq C \subseteq \cup_{x \in K} B(x, \varepsilon) \cap X.$$

Proof. Take any $K \in \mathcal{K}$. Each point in K is a core-point and by Lemma 1 and the definition of density-reachable, each point in K belongs to the same $C \in \mathcal{C}$. Thus, $K \subseteq \{x \in C : r_k(x) \leq \varepsilon\}$. Next we show that $K = \{x \in C : r_k(x) \leq \varepsilon\}$.

Suppose there exists core-point $x \in C$ but $x \notin K$ and let $y \in K$. By Lemma 1, there exists core-point $c \in C$ such that all points in C are directly reachable from c. Then there exists a path of core-points from x to c with pairwise edges of length at most ε. The same holds for c to y. Thus there exists such a path of core-points from x to y, which means that x, y are in the same CC of $G(minPts, \varepsilon)$, contradicting the assumption that $x \notin K$ and $y \in K$. Thus, in fact $K = \{x \in C : r_k(x) \leq \varepsilon\}$. The result now follows since C consists of points that are at most ε from its core-points. □

We can now see that DBSCAN's clusterings can be viewed as the connected components (CCs) of an appropriate ϵ-neighborhood level graph. Using a neighborhood graph to approximate the level-set has been studied in (Rinaldo & Wasserman, 2010). The difference is that they use a kernel density estimator instead of a k-NN density estimator and study the convergence properties under different settings.

4. Manifold Setting

4.1. Setup

We make the following regularity assumptions which are standard among works on manifold learning e.g. (Baraniuk & Wakin, 2009; Genovese et al., 2012; Balakrishnan et al., 2013).

Assumption 1. *$\mathcal{F}$ is supported on M where:*

- *M is a d-dimensional smooth compact Riemannian manifold without boundary embedded in compact subset $\mathcal{X} \subseteq \mathbb{R}^D$.*
- *The volume of M is bounded above by a constant.*
- *M has condition number $1/\tau$, which controls the curvature and prevents self-intersection.*

Let f be the density of $\mathcal{F}$ with respect to the uniform measure on M.

Assumption 2. *f is continuous and bounded.*

4.2. Basic Supporting Bounds

The following result bounds the empirical mass of Euclidean balls to the true mass under f. It is a direct consequence of Lemma 6 of Balakrishnan et al. (2013).

Lemma 3 (Uniform convergence of empirical Euclidean balls (Lemma 6 of Balakrishnan et al. (2013)))**.** *Let $\mathcal{N}$ be a minimal fixed set such that each point in M is at most distance $1/n$ from some point in $\mathcal{N}$. There exists a universal* *constant C_0 such that the following holds with probability at least $1-\delta$. For all $x \in X \cup \mathcal{N}$,*

$$\mathcal{F}(B) \geq C_{\delta,n}\frac{\sqrt{d\log n}}{n} \Rightarrow \mathcal{F}_n(B) > 0$$
$$\mathcal{F}(B) \geq \frac{k}{n} + C_{\delta,n}\frac{\sqrt{k}}{n} \Rightarrow \mathcal{F}_n(B) \geq \frac{k}{n}$$
$$\mathcal{F}(B) \leq \frac{k}{n} - C_{\delta,n}\frac{\sqrt{k}}{n} \Rightarrow \mathcal{F}_n(B) < \frac{k}{n}.$$

where $C_{\delta,n} = C_0 \log(2/\delta)\sqrt{d\log n}$, $\mathcal{F}_n$ is the empirical distribution, and $k \geq C_{\delta,n}$.

Remark 2. *For the rest of the paper, many results are qualified to hold with probability at least $1-\delta$. This is precisely the event in which Lemma 3 holds.*

Remark 3. *If $\delta = 1/n$, then $C_{\delta,n} = O((\log n)^{3/2})$.*

Next, we need the following bound on the volume of the intersection Euclidean ball and M; this is required to get a handle on the true mass of the ball under $\mathcal{F}$ in later arguments. The upper and lower bounds follow from Chazal (2013) and Lemma 5.3 of Niyogi et al. (2008). The proof is given in the appendix.

Lemma 4 (Ball Volume)**.** *If $0 < r < \min\{\tau/4d, 1/\tau\}$, and $x \in M$ then*

$$v_d r^d(1-\tau^2r^2) \leq vol_d(B(x,r)\cap M) \leq v_d r^d(1+4dr/\tau).$$

where v_d is the volume of a unit ball in $\mathbb{R}^d$ and vol_d is the volume w.r.t. the uniform measure on M.

4.3. k-NN Density Estimation

Here, we establish density estimation rates for the k-NN density estimator in the manifold setting. This builds on work in density estimation on manifolds e.g. (Hendriks, 1990; Pelletier, 2005; Ozakin & Gray, 2009; Kim & Park, 2013; Berry & Sauer, 2017); thus, it may be of independent interest. The estimator is defined as follows

Definition 3 (k-NN Density Estimator)**.**

$$f_k(x) := \frac{k}{n \cdot v_d \cdot r_k(x)^d}.$$

The following extends previous work of Dasgupta & Kpotufe (2014) to the manifold case. The proofs can be found in the appendix.

Lemma 5 (f_k upper bound)**.** *Suppose that Assumptions 1 and 2 hold. Define the following which charaterizes how much the density increases locally in M:*

$$\hat{r}(\epsilon, x) := \sup\left\{r : \sup_{x' \in B(x,r)\cap M} f(x') - f(x) \leq \epsilon\right\}.$$

Fix $\lambda_0 > 0$ and $\delta > 0$ and suppose that $k \geq C_{\delta,n}^2$. Then there exists constant $C_1 \equiv C_1(\lambda_0, d, \tau)$ such that if

$$k \leq C_1 \cdot C_{\delta,n}^{2d/(2+d)} \cdot n^{2/(2+d)},$$

then the following holds with probability at least $1-\delta$ uniformly in $\epsilon > 0$ and $x \in X$ with $f(x)+\epsilon \geq \lambda_0$:

$$f_k(x) < \left(1 + 3\cdot\frac{C_{\delta,n}}{\sqrt{k}}\right)\cdot(f(x)+\epsilon),$$

provided k satisfies $v_d \cdot \hat{r}(\epsilon,x)^d \cdot (f(x)+\epsilon) \geq \frac{k}{n} - C_{\delta,n}\frac{\sqrt{k}}{n}$.

Lemma 6 (f_k lower bound)**.** *Suppose that Assumptions 1 and 2 hold. Define the following which charaterizes how much the density decreases locally in M:*

$$\check{r}(\epsilon, x) := \sup\left\{r : \sup_{x' \in B(x,r)\cap M} f(x) - f(x') \leq \epsilon\right\}.$$

Fix $\lambda_0 > 0$ and $0 < \delta < 1$ and suppose $k \geq C_{\delta,n}$. Then there exists constant $C_2 \equiv C_2(\lambda_0, d, \tau)$ such that if

$$k \leq C_2 \cdot C_{\delta,n}^{2d/(4+d)} \cdot n^{4/(4+d)},$$

then with probability at least $1-\delta$, the following holds uniformly for all $\epsilon > 0$ and $x \in X$ with $f(x)-\epsilon \geq \lambda_0$:

$$f_k(x) \geq \left(1 - 3\cdot\frac{C_{\delta,n}}{\sqrt{k}}\right)\cdot(f(x)-\epsilon),$$

provided k satisfies $v_d \cdot \check{r}(\epsilon,x)^d \cdot (f(x)-\epsilon) \geq \frac{4}{3}\left(\frac{k}{n} + C_{\delta,n}\frac{\sqrt{k}}{n}\right)$.

Remark 4. *We will often bound the density of points with low density. In low-density regions, there is less data and thus we require more points to get a tight bound. However, in many cases a tight bound is not necessary; thus the purposes of ϵ is to allow some slack. The higher the ϵ, the easier it is for the lemma conditions to be satisified. In particular, if f is α-Hölder continuous (i.e. $|f(x)-f(x')| \leq C_\alpha|x-x'|^\alpha$), we have $\hat{r}(\epsilon,x), \check{r}(\epsilon,x) \geq (\epsilon/C_\alpha)^{1/\alpha}$.*

5. Consistency and Rates

5.1. Level-Set Conditions

Much of the results will depend on the behavior of level set boundaries. Thus, we require sufficient drop-off at the boundaries, as well as separation between the CCs at a particular level set. We give the following notion of separation.

Definition 4. *A, A' are r-separated in M if there exists a set S such that every path from A to A' intersects S and $\sup_{x \in M\cap(S+B(0,r))} f(x) < \inf_{x\in A\cup A'} f(x)$.*

Define the following shorthands for distance from a point to a set, the intersection of M with a neighborhood around a set under the Euclidean distance, and the largest Euclidean distance from a point in a set to its closest sample point.

Definition 5. $d(x, A) := \inf_{x' \in A} |x - x'|$, $C^{\oplus r} := \{x \in M : d(x, C) \leq r\}$, $r_n(C) := \sup_{c \in C} d(c, X)$.

We have the following mild assumptions which ensures that the CCs can be separated from the rest of the density by sufficiently wide valleys and there is sufficient decay around the level set boundaries.

Assumption 3 (Separation Conditions)**.** *Let $\lambda > 0$ and $\mathcal{C}_\lambda$ be a CCs of $\{x \in M : f(x) \geq \lambda\}$. There exists $\check{C}_\beta, \hat{C}_\beta, \beta, r_s, r_c > 0$ and $0 < \lambda_0 < \lambda$ such that the following holds:*

For each $C \in \mathcal{C}_\lambda$, there exists A_C, a connected component of $M^{\lambda_0} := \{x \in M : f(x) \geq \lambda_0\}$ such that:

- *A_C separates C by a valley: A_C does not intersect with any other CC in $\mathcal{C}_\lambda$; A_C and $M^{\lambda_0} \backslash A_C$ are r_s-separated by some S_C.*
- *$C^{\oplus r_c} \subseteq A_C$.*
- *β-regularity: For $x \in C^{\oplus r_c} \backslash C$, we have*

$$\check{C}_\beta \cdot d(x, C)^\beta \leq \lambda - f(x) \leq \hat{C}_\beta \cdot d(x, C)^\beta.$$

Remark 5. *We can choose any $0 < \beta < \infty$. The β-regularity assumption appears in e.g. (Singh et al., 2009). This is very general and also allows us to make a separate global smoothness assumption.*

Remark 6. *We currently characterize the smoothness w.r.t. the Euclidean distance. One could alternatively use the geodesic distance on M, $d_M(p, q)$. It follows from Proposition 6.3 of Niyogi et al. (2008) that when $|p - q| < \tau/4$, we have $|p - q| \leq d_M(p, q) \leq 2|p - q|$. Since the distances we deal in our analysis with are of such small order, these distances can thus essentially be treated as equivalent. We use the Euclidean distance throughout the paper for simplicity.*

Remark 7. *For the rest of this paper, it will be understood that Assumptions 1, 2, and 3 hold.*

We can define a region which isolates C away from other clusters of $\{x \in M : f(x) \geq \lambda\}$.

Definition 6. $\mathcal{X}_C := \{x :$ *$\exists$ a path $\mathcal{P}$ from x to $x' \in C$ such that $\mathcal{P} \cap S_C = \emptyset$*$\}$.

5.2. Parameter Settings

Fix $\lambda > 0$ and $\delta > 0$. Let k satisfy the following

$$K_l \cdot (\log n)^2 \leq k \leq K_u \cdot (\log n)^{2d/(2+d)} \cdot n^{2\beta'/(2\beta'+d)},$$

where $\beta' := \min\{1, \beta\}$, and K_l and K_u are positive constants depending on $\delta, \check{C}_\beta, \hat{C}_\beta, \beta, \tau, d, ||f||_\infty, \lambda_0, r_s, r_c$ which are implicit in the proofs later in this section.

The remainder of this section will be to show that DBSCAN(minPts, ε) with

$$\text{minPts} = k, \; \varepsilon = \left(\frac{k}{n \cdot v_d \cdot (\lambda - \lambda \cdot C_{\delta,n}^2/\sqrt{k})} \right)^{1/d}$$

will consistently estimate each CC of $\{x \in M : f(x) \geq \lambda\}$. Throughout the text, we denote $\widehat{\mathcal{C}_\lambda}$ as the clusters returned by DBSCAN under this setting.

5.3. Separation and Connectedness

Take $C \in \mathcal{C}_\lambda$. We show that DBSCAN will return an estimated CC $\widehat{C}$, such that $\widehat{C}$ does not contain any points outside of $\mathcal{X}_C$. Then, we show that $\widehat{C}$ contains all the sample points in C. The proof ideas used are similar to that of standard results in cluster trees estimation; they can be found in the appendix.

Lemma 7 (Separation)**.** *There exists K_l sufficiently large and K_u sufficiently small such that the following holds with probability at least $1 - \delta$. Let $C \in \mathcal{C}_\lambda$. There exists $\widehat{C} \in \widehat{\mathcal{C}_\lambda}$ such that $\widehat{C} \subseteq \mathcal{X}_C$.*

Lemma 8 (Connectedness)**.** *There exists K_l sufficiently large and K_u sufficiently small such that the following holds with probability at least $1 - \delta$. Let $C \in \mathcal{C}_\lambda$. If there exists $\widehat{C} \in \widehat{\mathcal{C}_\lambda}$ such that $\widehat{C} \subseteq \mathcal{X}_C$, then $C^{\oplus r_n(C)} \cap X \subseteq \widehat{C}$.*

Remark 8. *These results allow C to have any dimension between 0 to d since we reason with $C^{\oplus r_n(C)}$, which contains samples, instead of simply C.*

5.4. Hausdorff Error

We give the estimation rate under the Hausdorff metric.

Definition 7 (Hausdorff Distance)**.**

$$d_{Haus}(A, A') = \max\{\sup_{x \in A} d(x, A'), \sup_{x' \in A'} d(x', A)\}.$$

Theorem 1. *There exists K_l sufficiently large and K_u sufficiently small such that the following holds with probability at least $1 - \delta$. For each $C \in \mathcal{C}_\lambda$, there exists $\widehat{C} \in \widehat{\mathcal{C}_\lambda}$ such that*

$$d_{Haus}(C, \widehat{C}) \leq 2 \cdot (4\lambda/\check{C}_\beta)^{1/\beta} \cdot C_{\delta,n}^{2/\beta} \cdot k^{-1/2\beta}.$$

Proof. For K_l and K_u appropriately chosen, we have Lemma 7 and Lemma 8 hold. Thus we have for $C \in \mathcal{C}_\lambda$, there exists $\widehat{C} \in \widehat{\mathcal{C}_\lambda}$ such that

$$C^{\oplus r_n(C)} \cap X \subseteq \widehat{C} \subseteq \bigcup_{\substack{x \in \mathcal{X}_C \cap X \\ f_k(x) \geq \lambda - \frac{C_{\delta,n}^2}{\sqrt{k}} \lambda}} B(x, \varepsilon) \cap M.$$

Define $\bar{r} := \left(\frac{4\lambda \cdot C_{\delta,n}^2}{\check{C}_\beta \cdot \sqrt{k}} \right)^{1/\beta}$. We show that $d_{\text{Haus}}(C, \widehat{C}) \leq \bar{r}$, which involves two directions to show from the Hausdroff metric: that $\max_{x\in\widehat{C}} d(x,C) \le \bar{r}$ and $\sup_{x\in C} d(x,\widehat{C}) \le \bar{r}$.

We start by proving $\max_{x\in\widehat{C}} d(x,C) \le \bar{r}$. Define $r_0 = \bar{r}/2$. We have

$$r_0 = \frac{1}{2}\left(\frac{4\cdot C_{\delta,n}^2}{\check{C}_\beta\cdot\sqrt{k}}\right)^{1/\beta} \ge \left(\frac{k}{v_d n\lambda_0}\right)^{1/d} \ge \varepsilon,$$

where the first inequality holds when K_u is chosen sufficiently small, and the last inequality holds because $\lambda_0 < \lambda - \frac{C_{\delta,n}^2}{\sqrt{k}}\lambda$. Hence $r_0 + \varepsilon \le \bar{r}$. Therefore, it suffices to show

$$\sup_{x\in(\mathcal{X}_C\backslash C^{\oplus r_0})\cap X} f_k(x) < \lambda - \frac{C_{\delta,n}^2}{\sqrt{k}}\lambda.$$

We have that for $x \in (\mathcal{X}_C\backslash C^{\oplus r_0/2})\cap X$, $f(x) \le \lambda - \check{C}_\beta(r_0/2)^\beta := \lambda'$. Thus, for any $x \in (\mathcal{X}_C\backslash C^{\oplus r_0})\cap X$ and letting $\epsilon = \lambda' - f(x)$, we have

$$\hat{r}(\epsilon, x) \ge r_0/2 \ge (4\lambda_0 C_{\delta,n}/(\sqrt{k}\cdot\check{C}_\beta))^{1/\beta}/2.$$

For K_u chosen sufficiently small, the last equation will be large enough (i.e. of order $(k/v_d n\lambda)^{1/d}$) so that the conditions of Lemma 5 hold. Thus, applying this for each $x \in (\mathcal{X}_C\backslash C^{\oplus r_0})\cap X$, we obtain

$$\sup_{x\in(\mathcal{X}_C\backslash C^{\oplus r_0})\cap X} f_k(x) < \left(1 + 3\frac{C_{\delta,n}}{\sqrt{k}}\right)(\lambda - \check{C}_\beta(r_0/2)^\beta).$$

We have the r.h.s. is at most $\lambda - \frac{C_{\delta,n}^2}{\sqrt{k}}\lambda$ for K_u chosen appropriately and the first direction follows.

We now turn to the other direction, that $\sup_{x\in C} d(x,\widehat{C}) \le \bar{r}$. Let $x \in C$. Then there exists sample point $x' \in B(x, r_n(C))$ by definition of r_n and we have that $x' \in \widehat{C}$. Finally, $r_n(C) \le \bar{r}$ for K_l sufficiently large, and thus $|x' - x| \le \bar{r}$. The result follows. □

Remark 9. *When taking $k \approx n^{2\beta'/(2\beta'+d)}$, we obtain the error rate of $d_{Haus}(C,\widehat{C}) \approx n^{-1/(2\beta+d\cdot\max\{1,\beta\})}$, ignoring logarithmic factors. When $0 < \beta \le 1$, this matches the known lower bound established in Theorem 4 of Tsybakov et al. (1997). However, we do not obtain this rate when $\beta > 1$. In this case, the density estimation error will be of order at least $n^{-1/(2+d)}$ due in part to the error from resolving the geodesic balls with Euclidean balls. This does not arise in the full dimensional setting, which will be described later.*

6. Removal of False Clusters

The result of Theorem 1 guarantees us that for each $C \in \mathcal{C}_\lambda$, there exists $\widehat{C} \in \widehat{\mathcal{C}_\lambda}$ that estimates it. In this section, we show how a second application of DBSCAN (Algorithm 1) can remove the false clusters discovered by the first application of DBSCAN with no additional parameters. This gives us the other direction, that each estimate in $\widehat{\mathcal{C}_\lambda}$ corresponds to a true CC in $\mathcal{C}_\lambda$, and thus DBSCAN can identify with a one-to-one correspondence each CC of the level-set.

Algorithm 1 DBSCAN False CC Removal

As in Section 5.2, let minPts $= k$ and $\varepsilon = \left(\frac{k}{n\cdot v_d\cdot(\lambda-\lambda\cdot C_{\delta,n}^2/\sqrt{k})}\right)^{1/d}$.

Define $\tilde{\varepsilon} := \left(\frac{k}{n\cdot v_d\cdot(\lambda-\lambda\cdot C_{\delta,n}^2/\sqrt[3]{k})}\right)^{1/d}$.

Let $\widehat{\mathcal{C}}_\lambda$ be the clusters returned by DBSCAN(minPts, ε).
Let $\widehat{\mathcal{D}}_\lambda$ be the clusters returned by DBSCAN(minPts, $\tilde{\varepsilon}$).
Let $\widetilde{\mathcal{C}}_\lambda$ be the clusters obtained by merging clusters from $\widehat{C}_\lambda$ which are subsets of the same cluster in $\widehat{\mathcal{D}}_\lambda$.
Return $\widetilde{\mathcal{C}}_\lambda$.

We state our result below. The proof is less involved and is in the appendix.

Theorem 2 (Removal of False CC Estimates)**.** *Define $\gamma = \lambda - \sup_{x\in M\backslash(\cup_{C\in\mathcal{C}_\lambda}\mathcal{X}_C)} f(x)$, which is positive. There exists K_l sufficiently large and K_u sufficiently small depending on γ in addition to the constants mentioned in Section 5.2 so that the following holds with probability at least $1-\delta$. For all $\widehat{C} \in \widetilde{\mathcal{C}}_\lambda$, there exists $C \in \mathcal{C}_\lambda$ such that*

$$d_{Haus}(C,\widehat{C}) \le 2\cdot(4\lambda/\check{C}_\beta)^{1/\beta}\cdot C_{\delta,n}^{2/\beta}\cdot k^{-1/2\beta}.$$

7. Adaptive Parameter Tuning

In this section, we show how to obtain the near optimal rates by estimating d and adaptively choosing k such that $k \approx n^{2\beta'/(2\beta'+d)}$ without knowledge of β.

7.1. Determining d

Knowing the manifold dimension d is necessary to tune the parameters as described in Section 5.2. There has been much work done on estimating the intrinsic dimension as many learning procedures (including this one) require d as an input. Such work in intrinsic dimension estimation include (Kegl, 2002; Levina & Bickel, 2004; Hein & Audibert, 2005). Pettis et al. (1979) and more recently Farahmand et al. (2007) take a k-nearest neighbor approach. We work with the estimate of a dimension at a point proposed in the latter work:

$$\hat{d}(x) = \frac{\log 2}{\log(r_{2k}(x)/r_k(x))}.$$

The main result of Farahmand et al. (2007) gives a high-probability bound for a single sample $X_1 \in X$. Here we give a high-probability bound under more mild smoothness assumptions which hold uniformly for all samples above some density-level given our new knowledge of k-NN density estimation rates. This may be of independent interest.

Theorem 3. *Suppose that f is α-Hölder continuous for some $0 < \alpha \leq 1$. Choose $\bar{\lambda}_0 > 0$ and $\delta > 0$. Then there exists constants C_1, C_2 depending on $\delta, C_\alpha, \alpha, \tau, d, \bar{\lambda}_0$ such that if k satisfies*

$$C_1 \cdot (\log n)^2 \leq k \leq C_2 \cdot n^{2\alpha/(2\alpha+d)},$$

then with probability at least $1-\delta$,

$$|\hat{d}(x) - d| \leq 20d \cdot ||f||_\infty \cdot \frac{C_{\delta,n}}{\sqrt{k}},$$

uniformly for all $x \in X$ with $f_k(x) \geq \bar{\lambda}_0$.

Proof. We have for $x \in X$ such that if $f_k(x) \geq \bar{\lambda}_0$, then $f(x) \geq \lambda_0 := \bar{\lambda}_0/2$ by Lemma 5 for C_1 chosen appropriately large and C_2 chosen appropriately small.

$$\hat{d}(x) = \frac{\log 2}{\log(r_{2k}(x)/r_k(x))} = \frac{d \log 2}{\log 2 + \log(f_k(x)/f_{2k}(x))}.$$

We now try to get a handle on $f_k(x)/f_{2k}(x)$ and show it is sufficiently close to 1. Applying Lemma 5 and 6 with $\epsilon = \frac{C_{\delta,n}}{\sqrt{k}} f(x)$ and C_1, C_2 appropriately chosen so that the conditions for the two Lemmas hold (remember that here we have $\hat{r}(\epsilon, x), \check{r}(\epsilon, x) \geq (\epsilon/C_\alpha)^{1/\alpha}$), we obtain

$$\begin{aligned}\frac{f_k(x)}{f_{2k}(x)} &\geq \frac{(1 - 3C_{\delta,n}/\sqrt{k})(1 - C_{\delta,n}/\sqrt{k}) \cdot f(x)}{(1 + 3C_{\delta,n}/\sqrt{k})(1 + C_{\delta,n}/\sqrt{k}) \cdot f(x)} \\ &\geq 1 - 9 \cdot \frac{C_{\delta,n}}{\sqrt{k}},\end{aligned}$$

where the last inequality holds when C_1 is chosen sufficiently large so that $C_{\delta,n}/\sqrt{k}$ is sufficiently small. On the other hand, we similarly obtain (for C_1 and C_2 appropriately chosen):

$$\begin{aligned}\frac{f_k(x)}{f_{2k}(x)} &\leq \frac{(1 + 3C_{\delta,n}/\sqrt{k})(1 + C_{\delta,n}/\sqrt{k}) \cdot f(x)}{(1 - 3C_{\delta,n}/\sqrt{k})(1 - C_{\delta,n}/\sqrt{k}) \cdot f(x)} \\ &\leq 1 + 9 \cdot \frac{C_{\delta,n}}{\sqrt{k}}.\end{aligned}$$

It is now clear that by the expansion $\log(1-r) = -r - r^2/2 - r^3/3 - \cdots$, and for K_l chosen sufficently large so that $C_{\delta,n}/\sqrt{k}$ is sufficiently small, we have

$$\left|\log\left(\frac{f_k(x)}{f_{2k}(x)}\right)\right| \leq 10 \cdot \frac{C_{\delta,n}}{\sqrt{k}}.$$

The result now follows by combining this with the earlier established expression for $\hat{d}(x)$, as desired. □

Remark 10. *In Farahmand et al. (2007), it is the case that $\alpha = 1$; under this setting, we match their bound with an error rate of $n^{1/(2+d)}$ with $k \approx n^{2/(2+d)}$ being the optimal choice for k (ignoring log factors).*

7.2. Determining k

After determining d, the next parameter we look at is k. In particular, to obtain the optimal rate, we must choose $k \approx n^{2\beta'/(2\beta'+d)}$ without knowledge of β. We present a consistent estimator for β.

We need the following definition. The first characterizes how much f varies in balls of a certain radius along the boundaries of the λ-level set (where $\partial\mathcal{C}_\lambda$ denotes the boundary of $\mathcal{C}_\lambda$). The second is meant to be an estimate of the first, which can be computed from the data alone. The final is our estimate of β.

$$\begin{aligned} D_r &= \inf_{x_0 \in \partial\mathcal{C}_\lambda} \sup_{x \in B(x_0, r)} |\lambda - f(x)| \\ \hat{D}_{r,k} &= \min_{\substack{x_0 \in X \\ B(x_0,r)\cap X \neq \emptyset}} \max_{x \in B(x_0,r)\cap X} |\lambda - f_k(x)| \\ \hat{\beta} &= \log_r(\hat{D}_{r,k}) \end{aligned}$$

The next is a result of how $\hat{D}_{r,k}$ estimates D_r.

Lemma 9. *Suppose that f is α-Hölder continuous for some $0 < \alpha \leq 1$. Let $k = \lfloor (\log n)^5 \rfloor$ and $r = 1/\sqrt{\log n}$. Then there exists positive constants $\tilde{C}$ and N depending on $d, \tau, \alpha, C_\alpha, \lambda_0, ||f||_\infty, r_c$ such that when $n \geq N$, then the following holds with probability at least $1 - 1/n$.*

$$|D_r - \hat{D}_{r,k}| \leq \tilde{C}/(\log n)^2.$$

Proof sketch. Suppose that the value of D_r is attained at $x_0 = p$ and the value of $\hat{D}_{r,k}$ is attained at $x_0 = q$. Let y, z be the points that maximize $|\lambda - f(x)|$ on $B(p,r)$ and $B(q,r)$, respectively. Let $\hat{y}, \hat{z}$ be the sample points that maximize $|\lambda - f_k(x)|$ on $B(p,r)$ and $B(q,r)$, respectively. Now, we have

$$\begin{aligned} D_r - \hat{D}_{r,k} &= |\lambda - f(y)| - |\lambda - f_k(\hat{z})| \\ &\leq |\lambda - f(z)| - |\lambda - f_k(\hat{z})| \leq |f(z) - f_k(\hat{z})| \\ &\leq \max\{f(z) - f_k(z), f_k(\hat{z}) - f(\hat{z})\}. \end{aligned}$$

Now let z' be the closest sample point to z in $B(q,r)$. Then,

$$\begin{aligned} &\leq \max\{f(z') - f_k(z'), f_k(\hat{z}) - f(\hat{z})\} + |f(z) - f(z')| \\ &+ |f_k(z) - f_k(z')| \leq \max_{x \in X, f(x) \geq \lambda_0} |f(x) - f_k(x)| \\ &+ C_\alpha |z - z'|^\alpha + |f_k(z) - f_k(z')|. \end{aligned}$$

On the other hand, we have

$$\begin{aligned} \hat{D}_{r,k} - D_r &= |\lambda - f_k(\hat{z})| - |\lambda - f(y)| \\ &\leq |\lambda - f_k(\hat{y})| - |\lambda - f(y)| \leq |f(y) - f_k(\hat{y})| \\ &\leq \max\{f(y) - f_k(y), f_k(\hat{y}) - f(\hat{y})\}. \end{aligned}$$

Let y' be the closest sample point to y in $B(p,r)$. Then,

$$\begin{aligned} &\leq \max\{f(y') - f_k(y'), f_k(\hat{y}) - f(\hat{y})\} + |f(y) - f(y')| \\ &+ |f_k(y) - f_k(y')| \leq \max_{x \in X, f(x) \geq \lambda_0} |f(x) - f_k(x)| \\ &+ C_\alpha |y - y'|^\alpha + |f_k(y) - f_k(y')|. \end{aligned}$$

Thus it suffices to bound $\max_{x \in X, f(x) \geq \lambda_0} |f(x) - f_k(x)|, |y - y'|, |z - z'|, |f_k(y) - f_k(y')|, |f_k(z) - f_k(z')|$. First take $\delta = 1/n$ and use Lemma 5 and 6 for $\max_{x \in X, f(x) \geq \lambda_0} |f(x) - f_k(x)|$. Using Lemma 3, we can show that $r_n := |y - y'| \lesssim (\log n/n)^{1/d}$. Next we bound $|f_k(y) - f_k(y')|$. $y' \in X$ so we have guarantees on its f_k value. Note that $r_k(y') - r_n \leq r_k(y) \leq r_k(y') + r_n$. Let $r_k = r_k(y')$. This implies that $f_k(y')(r_k/(r_k + r_n))^d \leq f_k(y) \leq f_k(y')(r_k/(r_k - r_n))^d$. Now since $r_k \approx (k/n)^{1/d}$, we have $|f_k(y) - f_k(y')| \lesssim \log n/k$. The same holds for the bounds related to z, z'. □

Theorem 4 ($\hat{\beta} \to \beta$ in probability)**.** *Suppose f is α-Hölder continuous for some α with $0 < \alpha \leq \beta'$. Let $k = \lfloor (\log n)^5 \rfloor$ and $r = 1/\sqrt{\log n}$. Then for all $\epsilon > 0$,*

$$\lim_{n \to \infty} \mathbb{P}\left(|\hat{\beta} - \beta| \geq \epsilon\right) = 0.$$

Proof. Based on the β-regularity assumption, we have for $r < r_c$:

$$\check{C}_\beta r^\beta \leq D_r \leq \hat{C}_\beta r^\beta.$$

Combining this with Lemma 9, we have with probability at least $1 - 1/\sqrt{n}$ that

$$\check{C}_\beta r^\beta - \tilde{C}/(\log n)^2 \leq \hat{D}_{r,k} \leq \hat{C}_\beta r^\beta + \tilde{C}/(\log n)^2.$$

Thus with probability at least $1 - 1/n$,

$$\beta - \hat{\beta} \geq \frac{\log(1 - \tilde{C}/(\hat{D}_{r,k} \cdot (\log n^2)))}{\log r} - \frac{\log \hat{C}_\beta}{\log r}$$
$$\beta - \hat{\beta} \leq \frac{\log(1 + \tilde{C}/(\hat{D}_{r,k} \cdot (\log n^2)))}{\log r} + \frac{\log \check{C}_\beta}{\log r}.$$

It is clear that these expressions go to 0 as $n \to \infty$ and the result follows. □

Remark 11. *We can then take $k = n^{\hat{\beta}'/(2\hat{\beta}'+d)}$ with $\hat{\beta}' = \min\{1, \hat{\beta} - \epsilon_0\}$ for some $\epsilon_0 > 0$ so that $\hat{\beta}' < \beta'$ for n sufficiently large and thus k lies in the allowed ranges described in Section 5.2 asymptotically. The settings of ε and MinPts are implied by this choice of k and our estimate of d.*

7.3. Rates with Data-driven Tuning

Putting this all together, along with Theorems 1 and 2, gives us the following consequence about level set recovery with adaptive tuning. It shows that we can obtain rates arbitrarily close to those obtained as if the smoothness parameter β and intrinsic dimension were known.

Corollary 1. *Suppose that $0 < \delta < 1$ and f is α-Hölder continuous for some $0 < \alpha \leq 1$ and suppose the data-driven choices of parameters described in Remark 11 are used for DBSCAN. For any $\epsilon > 0$, there exists $N_{\epsilon,\delta,f} \equiv N(\epsilon, \delta, f)$ and $C_\delta \equiv C_\delta(\delta, f)$ such that the following holds. If $n \geq N_{\epsilon,\delta,f}$, then with probability at least $1 - \delta$ simulatenously for each $C \in \mathcal{C}_\lambda$, there exists $\widehat{C} \in \widehat{\mathcal{C}_\lambda}$ such that*

$$d_{Haus}(C, \widehat{C}) \leq C_\delta \cdot n^{-\frac{1}{2\beta + d\max\{1,\beta\}} + \epsilon}.$$

Moreover, using Algorithm 2, there is a one-to-one correspondence between $\mathcal{C}_\lambda$ and $\widehat{\mathcal{C}_\lambda}$.

8. Full Dimensional Setting

Here we instead take f to be the density of $\mathcal{F}$ over the uniform measure on $\mathbb{R}^D$. Let

$$\text{minPts} = k,\ \varepsilon = \left(\frac{k}{n \cdot v_D \cdot (\lambda - \lambda \cdot C_{\delta,n}^2/\sqrt{k})}\right)^{1/D},$$

where k satisfies

$$K_l \cdot (\log n)^2 \leq k \leq K_u \cdot (\log n)^{2D/(2+D)} \cdot n^{2\beta/(2\beta+D)},$$

and K_l and K_u are positive constants depending $\delta, \check{C}_\beta, \hat{C}_\beta, \beta, \tau, D, ||f||_\infty, \lambda_0, r_s, r_c$.

Then Theorem 1 and 2 hold (replacing d with D in Algorithm 1) for this setting of DBSCAN and thus taking $k \approx n^{2\beta/(2\beta+D)}$ gives us the optimal estimation rate of $O(n^{-1/(2\beta+D)})$. A straightforward modification of Corollary 1 also holds. This is discussed further in the Appendix.

9. Conclusion

We proved that DBSCAN can obtain Hausdorff level-set recovery rates of $\widetilde{O}(n^{-1/(2\beta+D)})$ when the data is in $\mathbb{R}^D$, and $\widetilde{O}(n^{-1/(2\beta+d\cdot\max\{1,\beta\})})$ when the data lies on an embedded d-dimensional manifold. The former rate is optimal up to log factors and the latter matches known d-dimensional lower bounds for $0 < \beta \leq 1$ up to log factors. Moreover, we provided a fully data-driven procedure to tune the parameters to attain these rates.

This shows that the procedure's ability to recover density level sets matches the strongest known consistency results attained for this problem. Furthermore, we developed the necessary tools and give the first analysis of density level-set estimation on manifolds, let alone with a practical procedure such as DBSCAN.

Our density estimation errors however cannot converge faster than $\widetilde{O}(n^{-1/(2+d)})$, which is due in part to the error from resolving geodesic balls with Euclidean balls. Thus it remains an open problem whether the manifold level-set rates are minimax optimal when $\beta > 1$.

Acknowledgements

The author is grateful to Samory Kpotufe for insightful discussions and to the anonymous reviewers for their useful feedback.

References

Baıllo, Amparo, Cuesta-Albertos, Juan A, and Cuevas, Antonio. Convergence rates in nonparametric estimation of level sets. *Statistics & probability letters*, 53(1):27–35, 2001.

Balakrishnan, S., Narayanan, S., Rinaldo, A., Singh, A., and Wasserman, L. Cluster trees on manifolds. In *Advances in Neural Information Processing Systems*, pp. 2679–2687, 2013.

Baraniuk, Richard G and Wakin, Michael B. Random projections of smooth manifolds. *Foundations of computational mathematics*, 9(1):51–77, 2009.

Berry, Tyrus and Sauer, Timothy. Density estimation on manifolds with boundary. *Computational Statistics & Data Analysis*, 107:1–17, 2017.

Biau, Gérard, Cadre, Benoît, and Pelletier, Bruno. Exact rates in density support estimation. *Journal of Multivariate Analysis*, 99(10):2185–2207, 2008.

Cadre, Benoıt. Kernel estimation of density level sets. *Journal of multivariate analysis*, 97(4):999–1023, 2006.

Campello, Ricardo JGB, Moulavi, Davoud, Zimek, Arthur, and Sander, Jörg. Hierarchical density estimates for data clustering, visualization, and outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(1):5, 2015.

Carmichael, J., George, G., and Julius, R. Finding natural clusters. *Systematic Zoology*, 1968.

Chaudhuri, K. and Dasgupta, S. Rates for convergence for the cluster tree. In *Advances in Neural Information Processing Systems*, 2010.

Chaudhuri, K., Dasgupta, S., Kpotufe, S., and von Luxburg, U. Consistent procedures for cluster tree estimation and pruning. *IEEE Transactions on Information Theory*, 2014.

Chazal, F. An upper bound for the volume of geodesic balls in submanifolds of euclidean spaces. 2013.

Chazal, Frédéric, Glisse, Marc, Labruère, Catherine, and Michel, Bertrand. Convergence rates for persistence diagram estimation in topological data analysis. *Journal of Machine Learning Research*, 16:3603–3635, 2015.

Chen, Yen-Chi, Genovese, Christopher R, and Wasserman, Larry. Density level sets: Asymptotics, inference, and visualization. *Journal of the American Statistical Association*, (just-accepted), 2016.

Cuevas, A. and Fraiman, R. A plug-in approach to support estimation. *Annals of Statistics*, pp. 2300–2312, 1997.

Dasgupta, S. and Kpotufe, S. Optimal rates for k-nn density and mode estimation. In *Advances in Neural Information Processing Systems*, 2014.

Ester, M., Kriegel, H., Sander, J., and Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD*, 96(34):226–231, 1996.

Farahmand, A., Szepesvari, C., and Audibert, J. Manifold-adaptive dimension estimation. *ICML*, 2007.

Genovese, Christopher, Perone-Pacifico, Marco, Verdinelli, Isabella, and Wasserman, Larry. Minimax manifold estimation. *Journal of machine learning research*, 13(May):1263–1291, 2012.

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

Hein, Matthias and Audibert, Jean-Yves. Intrinsic dimensionality estimation of submanifolds in rd. *ICML*, 2005.

Hendriks, Harrie. Nonparametric estimation of a probability density on a riemannian manifold using fourier expansions. *The Annals of Statistics*, pp. 832–849, 1990.

Jiang, Heinrich. Uniform convergence rates for kernel density estimation. *International Conference on Machine Learning (ICML)*, 2017.

Jiang, Heinrich and Kpotufe, Samory. Modal-set estimation with an application to clustering. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Kegl, Balazs. Intrinsic dimension estimation using packing numbers. *NIPS*, 2002.

Kim, Yoon Tae and Park, Hyun Suk. Geometric structures arising from kernel density estimation on riemannian manifolds. *Journal of Multivariate Analysis*, 114: 112–126, 2013.

Kpotufe, S. and von Luxburg, U. Pruning nearest neighbor cluster trees. In *International Conference on Machine Learning*, 2011.

Levina, Elizaveta and Bickel, Peter J. Maximum likelihood estimation of intrinsic dimension. *NIPS*, 2004.

Maier, Markus, Hein, Matthias, and von Luxburg, Ulrike. Optimal construction of k-nearest-neighbor graphs for identifying noisy clusters. *Theoretical Computer Science*, 410(19):1749–1764, 2009.

Niyogi, Partha, Smale, Stephen, and Weinberger, Shmuel. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1-3):419–441, 2008.

Ozakin, Arkadas and Gray, Alexander G. Submanifold density estimation. pp. 1375–1382, 2009.

Pelletier, Bruno. Kernel density estimation on riemannian manifolds. *Statistics & probability letters*, 73(3):297–304, 2005.

Pettis, K., Bailey, T., and Jain, A. An intrinsic dimensionality estimator from near-neighbor information. *IEEE Transactions on PAMI*, 1979.

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

Rigollet, P. and Vert, R. Fast rates for plug-in estimators of density level sets. *Bernoulli*, 15(4):1154–1178, 2009.

Rinaldo, Alessandro and Wasserman, Larry. Generalized density clustering. *The Annals of Statistics*, pp. 2678–2722, 2010.

Rinaldo, Alessandro, Singh, Aarti, Nugent, Rebecca, and Wasserman, Larry. Stability of density-based clustering. *Journal of Machine Learning Research*, 13(Apr): 905–948, 2012.

Singh, Aarti, Scott, Clayton, Nowak, Robert, et al. Adaptive hausdorff estimation of density level sets. *The Annals of Statistics*, 37(5B):2760–2782, 2009.

Sriperumbudur, Bharath K and Steinwart, Ingo. Consistency and rates for clustering with dbscan. pp. 1090–1098, 2012.

Steinwart, I. Adaptive density level set clustering. In *24th Annual Conference on Learning Theory*, 2011.

Steinwart, Ingo et al. Fully adaptive density-based clustering. *The Annals of Statistics*, 43(5):2132–2167, 2015.

Stuetzle, Werner and Nugent, Rebecca. A generalized single linkage method for estimating the cluster tree of a density. *Journal of Computational and Graphical Statistics*, 19(2):397–418, 2010.

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

Walther, Guenther. Granulometric smoothing. *The Annals of Statistics*, pp. 2273–2299, 1997.

Willett, RM and Nowak, Robert D. Minimax optimal level-set estimation. *IEEE Transactions on Image Processing*, 16(12):2965–2979, 2007.