Modal-set estimation with an application to clustering

Heinrich Jiang [1]

Samory Kpotufe
Princeton University

Abstract

We present a procedure that can estimate – with statistical consistency guarantees – any local-maxima of a density, under benign distributional conditions. The procedure estimates all such local maxima, or *modal-sets*, of any bounded shape or dimension, including usual point-modes. In practice, modal-sets can arise as dense low-dimensional structures in noisy data, and more generally serve to better model the rich variety of locally dense structures in data.

The procedure is then shown to be competitive on clustering applications, and moreover is quite stable to a wide range of settings of its tuning parameter.

1 INTRODUCTION

Mode estimation is a basic problem in data analysis. Modes, i.e. points of locally high density, serve as a measure of central tendency and are therefore important in unsupervised problems such as outlier detection, image or audio segmentation, and clustering in particular (as cluster cores). In the present work, we are interested in capturing a wider generality of *modes*, i.e. general structures (other than single-points) of locally high density, that can arise in modern data.

For example, application data in $\mathbb{R}^d$ (e.g. speech, vision, medical imaging) often display locally high-density structures of arbitrary shape. Such an example is shown in Figure 1. While there are many quite reasonable ways of modeling such locally high-density structures, we show that the simple model investigated in this work yields a practically successful procedure with strong statistical guarantees. In particular, our model simply allows the unknown density f to be locally maximal (or approximately maximal), not only at point-modes, but also on nontrivial subsets of $\mathbb{R}^d$. In other words, we make no a priori assumption on the nature of local maxima of the ground-truth f.

Figure 1: Medical imaging: high-resolution image of an eye with structures of blood capillaries. The aim in diabetic-retinopathy is to detect and delineate such capillary structures towards medical diagnostics.

We will refer to connected subsets of $\mathbb{R}^d$ where the unknown density f is locally maximal as *modal-sets* of f. A modal-set can be of any bounded shape and dimension, from 0-dimensional (point modes), to full dimensional surfaces, and aim to capture the rich variety of dense structures in data.

Our main contribution is a procedure, M(odal)-cores, that consistently estimates all such modal-sets from data, of general shape and dimension, with minimal assumption on the unknown f. If the ground-truth f is locally maximal at a point-mode, we return an estimate that converges to a point; if instead a modal-set is a surface, we consistently estimate that surface. Furthermore we have no a priori assumption on the number of modal-sets, besides that it is finite. Figure 2 shows a typical simulation on some arbitrary high-density structures.

The procedure builds on recent developments in topological data analysis [1, 2, 3, 4, 5, 6, 7], and works by carefully traversing a hierarchy of k-NN graphs which encode level sets of a k-NN density estimate. We show that, under mild uniform continuity assumptions on f, the Hausdorff distance between any modal-set and its estimate vanishes as $n \to \infty$ (Theorem 1, Section 2);

[1] Much of this work was done when this author was at Princeton University.

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

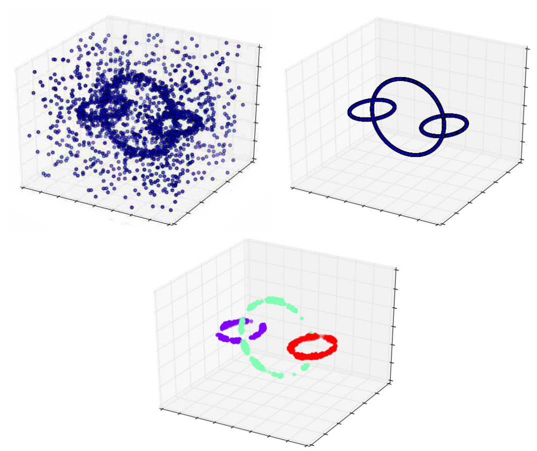

Figure 2: A first simulation on arbitrary shapes. (Top-Left) Points-cloud generated as three 1-dimensional rings + noise. (Top-Right) The 3 rings, and (Bottom) their estimate (as modal-sets) by M-cores.

the estimation rate for point-modes matches (up to $\log n$) the known minimax rates. Furthermore, under mild additional conditions on f (Hölder continuity), *false* structures (due to empirical variability) are correctly identified and pruned. We know of no other such general consistency guarantees in the context of mode estimation, especially for a practical procedure.

Finally, the present procedure is easy to implement and yields competitive scores on many clustering applications (Section 4); here, as in *mode-based clustering*, clusters are simply defined as regions of high-density of the data, and the estimated modal-sets serve as the centers of these regions, i.e., as *cluster-cores*. A welcome aspect of the resulting clustering procedure is its stability to tuning settings of the parameter k (from k-NN): it maintains high clustering scores (computed with knowledge of the ground-truth) over a wide range of settings of k, for various datasets. Such stability to tuning is of practical importance, since typically the ground-truth is unknown, making it difficult to tune the various hyperparameters of clustering procedures in practice. The performance of the present procedure seems rather robust to choices of its hyperparameters.

In the next section we put our result in context with respect to previous work on mode estimation and density-based clustering in general.

Related Work

• Much theoretical work on mode-estimation is concerned with understanding the statistical difficulty of the problem, and as such, often only considers the case of densities with single point-modes [8, 9, 10, 11, 12, 13]. The more practical case of densities with multiple point-modes has received less attention in the theoretical literature. However there exist practical estimators, e.g., the popular *Mean-Shift* procedure (which doubles as a clustering procedure), which are however harder to analyze. Recently, [14] shows the consistency of a variant of Mean-Shift. Other recent work of [15] derives a method for pruning false-modes obtained by mode-seeking procedures. Also recent, the work of [16] shows that point-modes of a k-NN density estimate f_k approximate the true modes of the unknown density f, assuming f only has point-modes and bounded Hessian at the modes; their procedure, therefore operates on level-sets of f_k (similar to ours), but fails in the presence of more general high-density structures such as modal-sets. To handle such general structures, we have to identify more appropriate level-sets to operate on, the main technical difficulty being that local-maxima of f_k can be relatively far (in Hausdorff) from those of f, for instance single-point modes rather than more general modal-sets, due to data-variability. The present procedure handles general structures, and is consistent under the much weaker conditions of continuity (of f) on a compact domain.

A related line of work, which seeks more general structures than point-modes, is that of *ridge* estimation (see e.g. [17, 18]). A ridge is typically defined as a lower-dimensional structure away from which the density curves (in some but not all directions), and can serve to capture various lower-dimensional patterns apparent in point clouds. Also, related to ridge-estimation, is the area of *manifold-denoising* which seeks to understand noise-conditions under which a low-dimensional manifold can be recovered from noisy observations (see e.g. [19] for a nice overview), often with knowledge of the manifold dimension. In contrast to this line of work, the modal-sets defined here can be full-dimensional and are always local maxima of the density. Also, unlike in ridge estimation, we do not require local differentiability of the unknown f, nor knowledge of the dimension of the structure, thus targeting a different set of practical structures.

• A main application of the present work, and of mode-estimation in general, is *density-based clustering*. Such clustering was formalized in early work of [20, 21, 22], and can take various forms, each with their advantage.

In its hierarchical version, one is interested in estimating the connected components (CCs) of *all* level sets $\{f \geq \lambda\}_{\lambda>0}$ of the unknown density f. Many recent works analyze approaches that consistently estimate such a hierarchy under quite general conditions, e.g. [1, 2, 3, 4, 5, 6, 7].

In the *flat* clustering version, one is interested in estimating the CCs of $\{f \geq \lambda\}$ for a single λ, somehow appropriately chosen [23, 24, 25, 26, 27, 28]. The popular DBSCAN procedure [29] can be viewed as estimating such single level set. The main disadvantage here is in the ambiguity in the choice of λ, especially when the levels λ of f have different numbers of clusters (CCs).

Another common flat clustering approach, most related to the present work, is *mode-based* clustering. The approach clusters points to estimated modes of f, a fixed target, and therefore does away with the ambiguity in choosing an appropriate level λ of f [30, 31, 32, 33, 34, 35, 36, 37]. As previously discussed, these approaches are however hard to analyze in that mode-estimation is itself not an easy problem. Popular examples are extensions of k-Means to categorical data [38], and the many variants of Mean-Shift which cluster points by gradient ascent to the closest mode. Notably, the recent work [39] analyzes clustering error of Mean-Shift in a general high-dimensional setting with potentially irrelevant features. The main assumption is that f only has point-modes.

2 OVERVIEW OF RESULTS

2.1 Basic Setup and Definitions

We have samples $X_{[n]} = \{X_1, ..., X_n\}$ drawn i.i.d. from a distribution $\mathcal{F}$ over $\mathbb{R}^d$ with density f. We let $\mathcal{X}$ denote the support of f. Our main aim is to estimate all local maxima of f, or *modal-sets* of f, as we will soon define.

We first require the following notions of distance between sets.

Definition 1. *For $M \subset \mathcal{X}$, $x \in \mathcal{X}$, let $d(x, M) := \inf_{x' \in M} \|x - x'\|$. The* **Hausdorff** *distance between $A, B \subset \mathcal{X}$ is defined as $d(A, B) := \max\{\sup_{x\in A} d(x, B), \sup_{y \in B} d(y, A)\}$.*

A modal set, defined below, extends the notion of a point-mode to general subsets of $\mathcal{X}$ where f is locally maximal. These can arise for instance, as discussed earlier, in applications where high-dimensional data might be modeled as a (disconnected) manifold $\mathcal{M}$ + ambient noise, each connected component of which induces a modal set of f in ambient space $\mathbb{R}^D$ (see e.g. Figure 2).

Definition 2. *For any $M \subset \mathcal{X}$ and $r > 0$, define the* envelope *$B(M, r) := \{x : d(x, M) \leq r\}$. A connected set M is a* **modal-set** *of f if $\forall x \in M$, $f(x) = f_M$ for some fixed f_M, and there exist $r > 0$ such that $f(x) < f_M$ for all $x \in B(M, r) \backslash M$.*

Remark 1. *The above definition can be relaxed to ϵ_0-*modal sets, *i.e., to allow f to vary by a small ϵ_0 on M. Our results extend easily to this more relaxed definition, with minimal changes to some constants. This is because the procedure operates on f_k, and therefore already needs to account for variations in f_k on M. This is described in the Appendix.*

2.2 Estimating Modal-sets

The algorithm relies on nearest-neighbor density estimate f_k, defined as follows.

Definition 3. *Let $r_k(x) := \min\{r : |B(x, r) \cap X_{[n]}| \geq k\}$, i.e., the distance from x to its k-th nearest neighbor. Define the* **k-NN density estimate** *as*

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

where v_d is the volume of a unit ball in $\mathbb{R}^d$.

Furthermore, we need an estimate of the level-sets of f; various recent work on cluster-tree estimation (see e.g. [6]) have shown that such level sets are encoded by subgraphs of certain *modified* k-NN graphs. Here however, we directly use k-NN graphs, simplifying implementation details, but requiring a bit of additional analysis.

Definition 4. *Let $G(\lambda)$ denote the (mutual)* **k-NN graph** *with vertices $\{x \in X_{[n]} : f_k(x) \geq \lambda\}$ and an edge between x and x' iff $||x - x'|| \leq \min\{r_k(x), r_k(x')\}$.*

$G(\lambda)$ can be viewed as approximating the λ-level set of f_k, hence approximates the λ-level set of f (implicit in the connectedness result in the Appendix).

Algorithm 1 (M-cores) estimates the modal-sets of the unknown f. It is based on various insights described below. A basic idea, used for instance in point-mode estimation [16], is to proceed top-down on the level sets of f_k (i.e. on $G(\lambda)$, $\lambda \to 0$), and identify new modal-sets as they appear in separate CCs of a level λ.

Here however, we have to be more careful: the CCs of $G(\lambda)$ (modal-sets of f_k for some λ) might be singleton points (since f_k might take unique values over samples $x \in X_{[n]}$) while the modal-sets to be estimated might be of any dimension and shape. Fortunately, if a data point x, locally maximizes f_k, and belongs to some modal-set M of f, then the rest of $M \cap X_{[n]}$ must be at a nearby level; Algorithm 1 therefore proceeds by checking a nearby level $(\lambda - 9\beta_k\lambda)$ from which it picks a specific set of points as an estimate of M. The main parameter here is β_k which is worked out explicitly in terms of k and requires no a priori knowledge of distributional parameters. The essential algorithmic parameter is therefore just k, which, as we will show, can be chosen over a wide range (w.r.t. n) while ensuring statistical consistency.

Definition 5. *Let* $0 < \delta < 1$. *Define* $C_{\delta,n} := 16\log(2/\delta)\sqrt{d\log n}$, *and define* $\beta_k = 4\frac{C_{\delta,n}}{\sqrt{k}}$.

If $\delta = 1/n$, then $C_{\delta,n} \approx \sqrt{d}\cdot(\log n)^{3/2}$. We note that the above definition of β_k is somewhat conservative (needed towards theoretical guarantees). Since $C_{\delta,n}$ behaves like a constant, it turns out to have little effect in implementation.

A further algorithmic difficulty is that a level $G(\lambda)$ might have too many CCs w.r.t. the ground truth. For example, due to variability in the data, f_k might have more modal-sets than f, inducing too many CCs at some level $G(\lambda)$. Fortunately, it can be shown that the nearby level $\lambda - 9\beta_k\lambda$ will likely have the right number of CCs. Such lookups down to lower-level act as a way of *pruning false modal-sets*, and trace back to earlier work [3] on pruning cluster-trees. Here, we need further care: we run the risk of over-estimating a given M if we look too far down (aggressive pruning), since a CC at lower level might contain points *far outside* of a modal-set M. Therefore, the main difficulty here is in figuring out how far down to look and yet not over-estimate *any* M (to ensure consistency). In particular our lookup *distance* of $9\beta_k\lambda$ is adapted to the level λ unlike in aggressive pruning.

Finally, for clustering with M-cores, we can simply assign every data-point to the closest estimated modal-set (acting as cluster-cores).

Algorithm 1 M-cores (estimating modal-sets).

```
Initialize $\widehat{\mathcal{M}} := \emptyset$. Define $\beta_k = 4\frac{C_{\delta,n}}{\sqrt{k}}$.
Sort the $X_i$'s in decreasing order of $f_k$ values (i.e. $f_k(X_i) \geq f_k(X_{i+1})$).
for $i = 1$ to $n$ do
    Define $\lambda := f_k(X_i)$.
    Let $A \equiv$ CC of $G(\lambda - 9\beta_k\lambda)$ containing $X_i$.    (i)
    if $A$ is disjoint from all cluster-cores in $\widehat{\mathcal{M}}$ then
        Add $\widehat{M} := \{x \in A : f_k(x) > \lambda - \beta_k\lambda\}$ to $\widehat{\mathcal{M}}$.
    end if
end for
return $\widehat{\mathcal{M}}$.    // Each $\widehat{M} \in \widehat{\mathcal{M}}$ is a cluster-core estimating a modal-set of the unknown $f$.
```

2.3 Consistency Results

Our consistency results rely on the following mild assumptions.

Assumption 1. *f is continuous with compact support $\mathcal{X}$. Furthermore f has a finite number of modal-sets all in the interior of its support $\mathcal{X}$.*

We will express the convergence of the procedure explicitly in terms of quantities that characterize the behavior of f at the boundary of every modal set. The first quantity has to do with how *salient* a modal-set, i.e, whether it is sufficiently *separated* from other modal sets. We start with the following definition of *separation.*

Definition 6. *Two sets $A, A' \subset \mathcal{X}$ are r-**separated**, if there exists a set S such that every path from A to A' crosses S and* $\sup_{x\in B(S,r)} f(x) < \inf_{x\in A\cup A'} f(x)$.

In simple terms, the definition says that there is a sufficiently wide valley in the density f between A and A'. For modal-set estimation, we also need such valleys to be sufficiently deep, which will be captured by the *curvature* of f at any modal-set.

The next quantities characterize the *change* in f in a neighborhood of a modal set M. The existence of a proper such neighborhood A_M, and appropriate functions u_M and l_M capturing smoothness and curvature, follow from the above assumptions on f. This is stated in the next proposition.

Proposition 1. *Let M be a modal-set of f. Then there exists a CC A_M of some level-set $\mathcal{X}^{\lambda_M} := \{x : f(x) \geq \lambda_M\}$, containing M, such that the following holds.*

- A_M isolates M by a valley*: A_M does not intersect any other modal-set; and A_M and $\mathcal{X}^{\lambda_M}\backslash A_M$ are r_s-separated (by some S_M) for some $r_s > 0$ independent of M.*
- A_M is full-dimensional*: A_M contains an envelope $B(M, r_M)$ of M, for some $r_M > 0$.*
- f is both *smooth* and has *curvature* around M*: there exist functions u_M and l_M, increasing and continuous on $[0, r_M]$, $u_M(0) = l_M(0) = 0$, such that $\forall x \in B(M, r_M)$,*

$$l_M(d(x,M)) \leq f_M - f(x) \leq u_M(d(x,M)).$$

Finally, our consistency guarantees require the following admissibility condition on $k = k(n)$. This condition results, roughly, from needing the density estimate f_k to properly approximate the behavior of f in the neighborhood of a modal-set M. In particular, we intuitively need f_k values to be smaller for points far from M than for points close to M, and this should depend on the smoothness and curvature of f around M (as captured by u_M and l_M).

Definition 7. *k is* **admissible** *for a modal-set M if (we let u_M^{-1} denote the inverse of u_M):*

$$\max\left\{\left(\frac{24\sup_{x\in\mathcal{X}} f(x)}{l_M(\min\{r_M, r_s\}/2)}\right)^2, 2^{7+d}\right\}\cdot C_{\delta,n}^2$$
$$\leq k \leq \frac{v_d\cdot f_M}{2^{2+2d}}\left(u_M^{-1}\left(f_M\frac{C_{\delta,n}}{2\sqrt{k}}\right)\right)^d\cdot n.$$

Remark 2. *The admissibility condition on k, although seemingly opaque, allows for a wide range of settings of k. For example, suppose $u_M(t) = ct^\alpha$ for some $c, \alpha > 0$. These are polynomial tail conditions common in mode estimation, following e.g. from Hölder assumptions on f. Admissibility then (ignoring $\log(1/\delta)$), is immediately seen to correspond to the wide range*

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

where C_1, C_2 are constants depending on M, but independent of k and n. It's clear then that even the simple choice $k = \Theta(log^2 n)$ is always admissible for any M *for n sufficiently large.*

Main theorems. We then have the following two main consistency results for Algorithm 1. Theorem 1 states a rate (in terms of l_M and u_M) at which any modal-set M is approximated by some estimate in $\widehat{\mathcal{M}}$; Theorem 2 establishes *pruning* guarantees.

Theorem 1. *Let $0 < \delta < 1$. The following holds with probability at least $1 - \delta$, simultaneously for all modal-sets M of f. Suppose k is admissible for M. Then there exists $\widehat{M} \in \widehat{\mathcal{M}}$ such that the following holds. Let l_M^{-1} denote the inverse of l_M.*

$$d(M, \widehat{M}) \le l_M^{-1}\left(\frac{8C_{\delta,n}}{\sqrt{k}} f_M\right),$$

which goes to 0 as $C_{\delta,n}/\sqrt{k} \to 0$.

If k is admissible for all modal-sets M of f, then $\widehat{\mathcal{M}}$ estimates all modal-sets of f at the above rates. These rates can be instantiated under the settings in Remark 2: suppose $l_M(t) = c_1 t^{\alpha_1}$, $u_M(t) = ct^\alpha$, $\beta_1 \ge \beta$; then the above bound becomes $d(M, \widehat{M}) \lesssim k^{-1/2\alpha_1}$ for admissible k. As in the remark, $k = \Theta(\log^2 n)$ is admissible, simultaneously for all M (for n sufficiently large), and therefore all modal-sets of f are recovered at the above rate. In particular, taking large $k = O(n^{2\alpha/(2\alpha+d)})$ optimizes the rate to $O(n^{-\alpha/(2\alpha_1\alpha+\alpha_1 d)})$. Note that for $\alpha_1 = \alpha = 2$, the resulting rate ($n^{-1/(4+d)}$) is tight (see e.g. [12] for matching lower-bounds in the case of point-modes $M = \{x\}$.).

Finally, Theorem 2 (pruning guarantees) states that any estimated modal-set in $\widehat{\mathcal{M}}$, at a sufficiently high level (w.r.t. to k), corresponds to a *true* modal-set of f at a similar level. Its proof consists of showing that if two sets of points are wrongly disconnected at level λ, they remain connected at nearby level $\lambda - 9\beta_k\lambda$ (so are reconnected by the procedure). The main technicality is the dependence of the nearby level on the empirical λ; the proof is less involved and given in the Appendix.

Theorem 2. *Let $0 < \delta < 1$. There exists $\lambda_0 = \lambda_0(n, k)$ such that the following holds with probability at least $1 - \delta$. All modal-set estimates in $\widehat{\mathcal{M}}$ chosen at level $\lambda \ge \lambda_0$ can be injectively mapped to modal-sets $\left\{M : \lambda_M \ge \min_{\{x \in \mathcal{X}_{[n]} : f_k(x) \ge \lambda - \beta_k\lambda\}} f(x)\right\}$, provided k is admissible for all such M.*

In particular, if f is Hölder-continuous, (i.e. $||f(x) - f(x')|| \le c||x - x'||^\alpha$ for some $0 < \alpha \le 1$, $c > 0$) then $\lambda_0 \xrightarrow{n\to\infty} 0$, provided $C_1 \log n \le k \le C_2 n^{2\alpha/(2\alpha+d)}$, for some C_1, C_2 independent n.

Remark 3. *Thus with little additional smoothness ($\alpha \approx 0$) over uniform continuity of f, any estimate above level $\lambda_0 \to 0$ corresponds to a true modal-set of f. We note that these pruning guarantees can be strengthened as needed by implementing a more aggressive pruning: simply replace $G(\lambda - 9\beta_k\lambda)$ in the procedure (on line (*i)) with $G(\lambda - 9\beta_k\lambda - \tilde{\epsilon})$ using a *pruning parameter $\tilde{\epsilon} \ge 0$.* This allows $\lambda_0 \to 0$ faster. However the rates of Theorem 1 (while maintained) then require a larger initial sample size n. This is discussed in the Appendix.

3 ANALYSIS OVERVIEW

The bulk of the analysis is in establishing Theorem 1. The key technicalities are in bounding distances from estimated cores to an unknown number of modal-sets of general shape, dimension and location.

The analysis considers each modal-set M of f separately, and only combines results in the end into the uniform consistency statement of Theorem 1. The following notion of *distance* from the sample $X_{[n]}$ to a modal-set M will be crucial.

Definition 8. *For any $x \in \mathcal{X}$, let $r_n(x) := d(\{x\}, X_{[n]})$, and $r_n(M) := \sup_{x\in M} r_n(x)$.*

We require a notion of a region $\mathcal{X}_M$ that *isolates* a modal-set M from other modal-sets. In other words, $\mathcal{X}_M$ containing only points *close* to M but far from other modes. To this end, let S_M denote the separating set from Proposition 1.

Definition 9. $\mathcal{X}_M := \{x : \exists$ *a path $\mathcal{P}$ from x to $x' \in M$ such that $\mathcal{P} \cap S_M = \emptyset$*$\}$.

For each M, define $\hat{x}_M := \text{argmax}_{x\in\mathcal{X}_M\cap X_{[n]}} f_k(x)$, a local maximizer of f_k on the modal-set M. The analysis (concerning each M) proceeds in the following steps:

- *Isolation of M*: when processing $\hat{x}_M$, the procedure picks an estimate $\widehat{M}$ that contains no point from (or close to) modal-sets other than M.

- *Integrality of M*: the estimate $\widehat{M}$ picks all of the envelope $B(M, r_n(M)) \cap X_{[n]}$.

- *Consistency of $\widehat{M}$*: it can then be shown that $\widehat{M} \to M$ in *Hausdorff* distance. This involves two directions: the first direction (that points of M are close to $\widehat{M}$) follows from integrality; the second direction is to show that points in $\widehat{M}$ are close to M.

The following gives an upper-bound on the distance from a modal-set to the closest sample point. It follows from Bernstein-type VC concentration on masses of balls. The proof is given in the Appendix.

Lemma 1 (Upper bound on r_n)**.** *Let M be a modal-set with density f_M and suppose that k is admissible. With probability at least $1 - \delta$,*

$$r_n(M) \le \left(\frac{2C_{\delta,n}\sqrt{d \log n}}{n \cdot v_d \cdot f_M} \right)^{1/d}.$$

From Lemma 1 above, for any modal-set M of f, for n sufficiently large, there is a sample point in $\mathcal{X}_M$, i.e., $\mathcal{X}_M \cap X_{[n]} \neq \emptyset$. The next lemma conditions on the existence of such a sample.

The proofs for Lemma 2 and 3 are given in the Appendix but we give proof sketches here.

Lemma 2 (Isolation)**.** *Let M be a modal-set and k be admissible for M. The following holds with probability at least $1 - \delta$. Suppose there exists a sample point in $\mathcal{X}_M$, and define $\hat{x}_M := \text{argmax}_{x \in \mathcal{X}_M \cap X_{[n]}} f_k(x)$. When $\hat{x}_M$ is processed in Algorithm 1, an estimate $\widehat{M}$ is added to $\widehat{\mathcal{M}}$ satisfying $\widehat{M} \subset \mathcal{X}_M$.*

Proof sketch. First we choose $\bar{r} > 0$ depending on the smoothness and decay around M. It suffices to show that (i) $B(M, \bar{r})$ and $\mathcal{X} \backslash \mathcal{X}_M$ are disjoint in the k-NN graph when $\hat{x}_M$ is being processed in Algorithm 1 and (ii) $\hat{x}_M \in B(M, \bar{r})$. To show (i), we use the k-NN bounds to justify that the k-NN graph does not contain any points from $B(S_M, \bar{r})$ or $\mathcal{X}_M \backslash B(M, \bar{r})$. Thus, any path from $\hat{x}_M$ to $\mathcal{X}_M \backslash B(M, \bar{r})$ must contain an edge with length greater than $\bar{r}$. We then show there is no such edge. To show (ii), we argue that $\hat{x}_M$ cannot be far away from M and is in fact within distance $\bar{r}$ from M. □

The above Lemma 2 establishes the existence of an estimate $\widehat{M}$ of M containing no point from other modes. The next lemma establishes that such an estimate $\widehat{M}$ contains all of $M \cap X_{[n]}$.

Lemma 3 (Integrality)**.** *Let M be a modal-set with density f_M, and suppose k is admissible for M. The following holds with probability at least $1 - \delta$. First, $\mathcal{X}_M$ does contain a sample point. Define $\hat{x}_M := \text{argmax}_{x \in \mathcal{X}_M \cap X_{[n]}} f_k(x)$. When processing $\hat{x}_M$ in Algorithm 1, suppose we add $\widehat{M}$ to $\widehat{\mathcal{M}}$, then $B(M, r_n(M)) \cap X_{[n]} \subseteq \widehat{M}$.*

Proof sketch. Let $A := B(M, r_n(M))$ and $\lambda_0 := (1 - \frac{C_{\delta,n}}{\sqrt{k}})^2 f_M$. We begin by showing that $A \cap X_{[n]}$ is connected in $G(\lambda_0)$ as follows. Let $r_\lambda := (k/(nv_d\lambda_0))^{1/d}$ which is the k-NN radius corresponding to a point with f_k value λ_0, and $r_o := (k/(2nv_d f_M))^{1/d}$ which will be smaller than the k-NN radius of any sample point. Next, we show that $B(x, r_o)$ contains a sample point for any $x \in B(A, r_\lambda)$. Now, for any two points $x, x' \in A \cap X_{[n]}$, we use this fact to argue for the existance of a sequence of sample points starting with x and ending with x' such that the distance between adjacent points is less than r_o and all the points in the sequence lie in $B(A, r_o)$. We then show that each pair of adjacent points is an edge in $G(\lambda_0)$ and thus $A \cap X_{[n]}$ is connected in $G(\lambda_0)$. Finally, we argue that $\lambda \le \lambda_0$ and thus $A \cap X_{[n]}$ is connected in $G(\lambda)$ as well. □

Combining isolation and integrality, we obtain:

Corollary 1 (Identification)**.** *Suppose the conditions of Lemmas 2 and 3 hold for modal-set M. Define $\widehat{f}_M := \max_{x \in \mathcal{X}_M \cap X_{[n]}} f_k(x)$. With probability at least $1 - \delta$, there exists $\widehat{M} \in \widehat{\mathcal{M}}$ such that $B(M, r_n(M)) \cap X_{[n]} \subseteq \widehat{M} \subseteq \{x \in \mathcal{X}_M \cap X_{[n]} : f_k(x) \ge \widehat{f}_M - \beta_k \widehat{f}_M\}$.*

Proof. By Lemma 2, there exists $\widehat{M} \in \widehat{\mathcal{M}}$ which contains only points in $\mathcal{X}_M$ with maximum f_k value of $\widehat{f}_M$. Thus, we have $\widehat{M} \subseteq \{x \in \mathcal{X}_M \cap X_{[n]} : f_k(x) \ge \widehat{f}_M - \beta_k \widehat{f}_M\}$. By Lemma 3, $B(M, r_n(M)) \cap X_{[n]} \subseteq \widehat{M}$. □

Here, we give a sketch of the proof for Theorem 1 which can be found in the Appendix.

Proof sketch of Theorem 1. Let $\tilde{r} = l_M^{-1}\left(\frac{8C_{\delta,n}}{\sqrt{k}} f_M\right)$. There are two directions to show: $\max_{x \in \widehat{M}} d(x, M) \le \tilde{r}$ and $\sup_{x \in M} d(x, \widehat{M}) \le \tilde{r}$.

For the first direction, by Corollary 1 we have $\widehat{M} \subseteq \{x \in \mathcal{X}_M : f_k(x) \ge \widehat{f}_M - \beta_k \widehat{f}_M\}$ where $\widehat{f}_M := \max_{x \in \mathcal{X}_M \cap X_{[n]}} f_k(x)$. Thus, it suffices to show

$$\inf_{x \in B(M, r_n(M))} f_k(x) \ge \sup_{\mathcal{X}_M \backslash B(M, \tilde{r})} f_k(x) + \beta_k \widehat{f}_M. \quad (1)$$

Using known upper and lower bounds on f_k in terms of f, we can lower bound the LHS by approximately $f_M - u_M(r)$ (for some $r < \tilde{r}$) and upper bound the first term on the RHS by approximately $f_M - l_M(\tilde{r})$.

The remaining difficulty is carefully choosing an appropriate r.

For the other direction, by Corollary 1, $\widehat{M}$ contains all sample points in $B(M, r_n(M))$. Lemma 1 and the admissibility of k implies that $r_n(x) \leq \tilde{r}$ which easily gives us the result. □

4 EXPERIMENTS

4.1 Practical Setup

The analysis prescribes a setting of $\beta_k = O(1/\sqrt{k})$. Throughout the experiments we simply fix $\beta_k = 2/\sqrt{k}$, and let our choice of k be the essential parameter. As we will see, M-cores yields competitive and stable performance for a wide-range of settings of k. The implementation can be done efficiently and is described in the Appendix.

A Python/C++ implementation of the code at [40].

4.2 Qualitative Experiments on General Structures

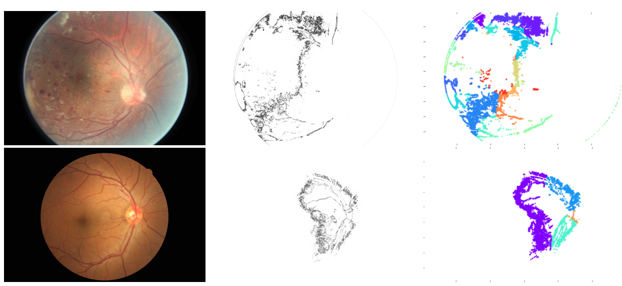

Figure 3: Diabetic Retinopathy: (Top 3 figures) An unhealthy eye, (Bottom 3 figures) A healthy eye. In both cases, shown are (1) original image, (2) a filter applied to the image, (3) modal-sets (structures of capillaries) estimated by M-cores on the corresponding filtered image. The unhealthy eye is characterized by a proliferation of damaged capillaries, while a healthy eye has visually fewer capillaries. The analysis task is to automatically discover the higher number of capillary-structures in the unhealthy eye. M-cores discovers 29 structures for unhealthy eye vs 6 for healthy.

We start with a qualitative experiment highlighting the flexibility of the procedure in fitting a large variety of high-density structures. For these experiments, we use $k = \frac{1}{2} \cdot \log^2 n$, which is within the theoretical range for admissible values of k (see Theorem 1 and discussion of Remark 2).

We consider a medical imaging problem. Figure 3 displays the procedure applied to the Diabetic Retinopathy detection problem [41]. While this is by no means an end-to-end treatment of this detection problem, it gives a sense of M-cores' versatility in fitting real-world patterns. In particular, M-cores automatically estimates a reasonable number of clusters, independent of shape, while pruning away (most importantly in the case of the healthy eye) false clusters due to noisy data. As a result, it correctly picks up a much larger number of clusters in the case of the unhealthy eye.

4.3 Clustering applications

We now evaluate the performance of M-cores on clustering applications, where for **clustering:** we assign every point $x_i \in X_{[n]}$ to $\text{argmin}_{\widehat{M} \in \widehat{\mathcal{M}}} d(x_i, \widehat{M})$, i.e. to the closest estimated modal-set.

We compare M-cores to two common density-based clustering procedures, DBSCAN and Mean-Shift, as implemented in the *sci-kit-learn* package. Mean-Shift clusters data around point-modes, i.e. local-maxima of f, and is therefore most similar to M-cores in its objective.

Clustering scores. We compute two established scores which evaluate a clustering against a labeled ground-truth. The *rand-index*-score is the 0-1 accuracy in grouping pairs of points, (see e.g. [42]); the *mutual information*-score is the (information theoretic) mutual-information between the distributions induced by the clustering and the ground-truth (each cluster is a mass-point of the distribution, see e.g. [43]). For both scores we report the *adjusted* version, which adjusts the score so that a random clustering (with the same number of clusters as the ground-truth) scores near 0 (see e.g. [42], [43]).

Datasets. Phonemes [44], and UCI datasets: Glass, Seeds, Iris, and Wearable Computing. They are described in the table below.

Dataset	n	d	*Labels*	*Description*
Phonemes	4509	256	5	Log-periodograms of spoken phonemes
Glass	214	7	6	Properties of different types of glass
Seeds	210	7	3	Geometric measurements of wheat kernels
Iris	150	4	3	Various measurements over species of flowers
Wearable	10000	12	5	4 sensors on a human body, recording body posture and activity

Results. Figure 4 reports the performance of the procedures for each dataset. Rather than reporting the performance of the procedures under *optimal-tuning*, we report their performance *over a range* of hyperparameter settings, mindful of the fact that optimal-

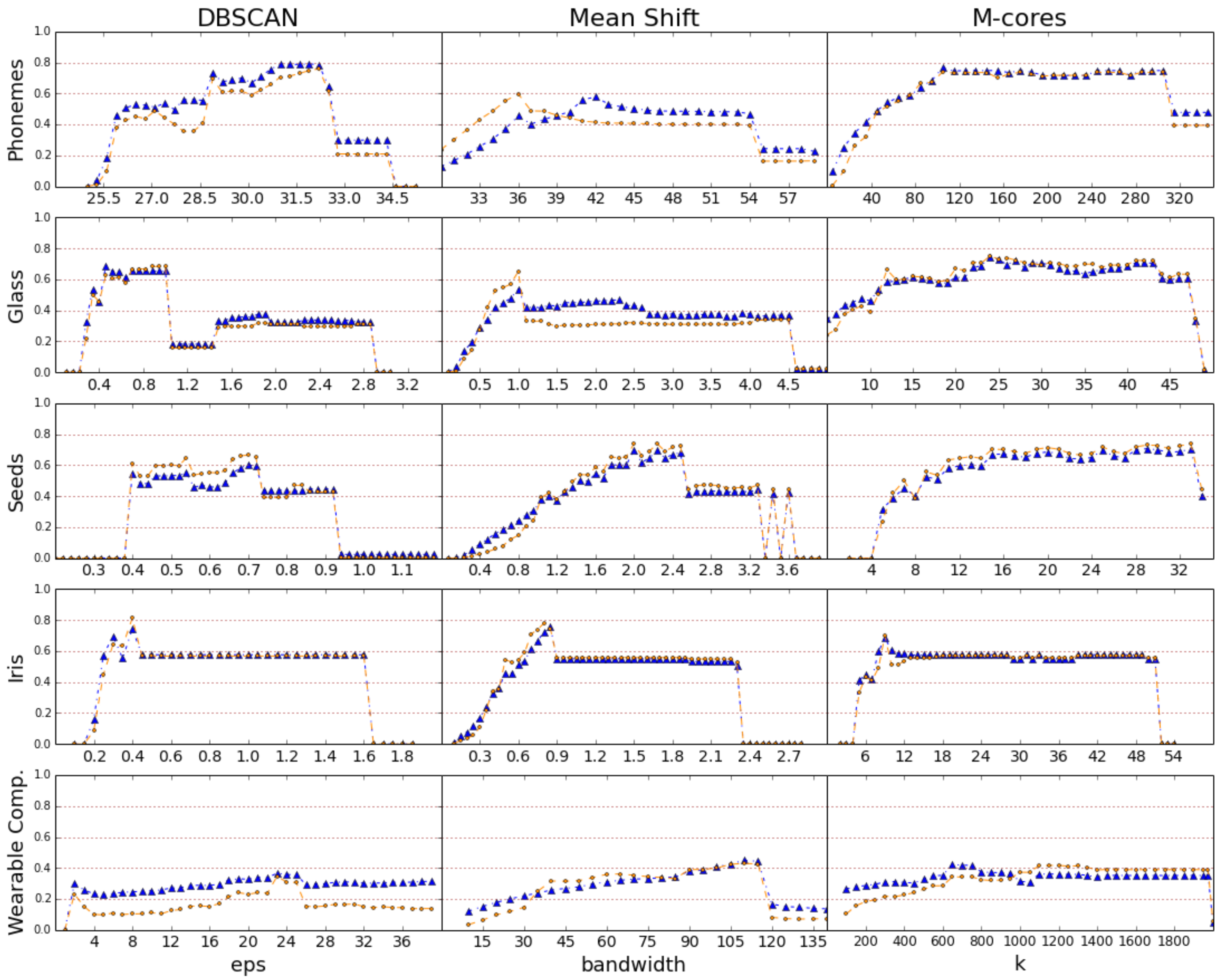

Figure 4: Comparison on real datasets (along the rows) across different hyperparameter settings for each algorithm (along the columns). The hyperparameters being tuned are displayed at the bottom of the figure for each clustering algorithm. Scores: the blue line with triangular markers is Adjusted-Mutual-Information, and the dotted red line is Adjusted-Rand-Index.

tuning is hardly found in practice (this is a general problem in clustering given the lack of ground-truth to guide tuning).

For M-cores we vary the parameter k. For DBSCAN and Mean-Shift, we vary the main parameters, respectively *eps* (choice of level-set), and *bandwidth* (used in density estimation). M-cores yields competitive performance across the board, with stable scores over a large range of values of k (relative to sample size). Such stable performance to large changes in k is quite desirable, considering that proper tuning of hyperparameters remains a largely open problem in clustering.

Conclusion

We presented a theoretically-motivated procedure which can consistently estimate modal-sets, i.e. non-trivial high-density structures in data, under benign distributional conditions. This procedure is easily implemented and yields competitive and stable scores in clustering applications.

Acknowledgements

We are grateful to Sanjoy Dasgupta for useful discussions in the beginning of this project. Finally, part of this work was presented at the Dagstuhl 2016 seminar on Unsupervised Learning, and we are thankful for the useful feedback from participants.

References

[1] 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(2):397–418, 2010.

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

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

[4] A. Rinaldo, A. Singh, R. Nugent, and L. Wasserman. Stability of density-based clustering. *Journal of Machine Learning Research*, 13:905–948, 2012.

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

[6] K. Chaudhuri, S. Dasgupta, S. Kpotufe, and U. von Luxburg. Consistent procedures for cluster tree estimation and pruning. *Arxiv*, 2014.

[7] Justin Eldridge, Yusu Wang, and Mikhail Belkin. Beyond hartigan consistency: Merge distortion metric for hierarchical clustering. *Conference on Learning Theory*, 2015.

[8] Emanuel Parzen et al. On estimation of a probability density function and mode. *Annals of mathematical statistics*, 33(3):1065–1076, 1962.

[9] Herman Chernoff. Estimation of the mode. *Annals of the Institute of Statistical Mathematics*, 16(1):31–41, 1964.

[10] William F Eddy et al. Optimum kernel estimators of the mode. *The Annals of Statistics*, 8(4):870–882, 1980.

[11] Luc Devroye. Recursive estimation of the mode of a multivariate density. *Canadian Journal of Statistics*, 7(2):159–167, 1979.

[12] Aleksandr Borisovich Tsybakov. Recursive estimation of the mode of a multivariate distribution. *Problemy Peredachi Informatsii*, 26(1):38–45, 1990.

[13] Christophe Abraham, Gérard Biau, and Benoît Cadre. On the asymptotic properties of a simple estimate of the mode. *ESAIM: Probability and Statistics*, 8:1–11, 2004.

[14] Ery Arias-Castro, David Mason, and Bruno Pelletier. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *Unpublished Manuscript*, 2013.

[15] Christopher Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Nonparametric inference for density modes. *arXiv preprint arXiv:1312.7567*, 2013.

[16] Sanjoy Dasgupta and Samory Kpotufe. Optimal rates for k-nn density and mode estimation. In *Advances in Neural Information Processing Systems*, pages 2555–2563, 2014.

[17] Umut Ozertem and Deniz Erdogmus. Locally defined principal curves and surfaces. *The Journal of Machine Learning Research*, 12:1249–1286, 2011.

[18] Christopher R Genovese, Marco Perone-Pacifico, Isabella Verdinelli, Larry Wasserman, et al. Nonparametric ridge estimation. *The Annals of Statistics*, 42(4):1511–1545, 2014.

[19] Sivaraman Balakrishnan, Alessandro Rinaldo, Don Sheehy, Aarti Singh, and Larry A Wasserman. Minimax rates for homology inference. In *AISTATS*, volume 9, pages 206–207, 2012.

[20] JW Carmichael, J Alan George, and RS Julius. Finding natural clusters. *Systematic Zoology*, pages 144–150, 1968.

[21] John A Hartigan. *Clustering algorithms*. John Wiley & Sons, Inc., 1975.

[22] J.A. Hartigan. Consistency of single linkage for high-density clusters. *Journal of the American Statistical Association*, 76(374):388–394, 1981.

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

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

[25] M. Maier, M. Hein, and U. von Luxburg. Optimal construction of k-nearest neighbor graphs for identifying noisy clusters. *Theoretical Computer Science*, 410:1749–1764, 2009.

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

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

[28] Bharath K Sriperumbudur and Ingo Steinwart. Consistency and rates for clustering with dbscan. In *International Conference on Artificial Intelligence and Statistics*, pages 1090–1098, 2012.

[29] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases

with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[30] Keinosuke Fukunaga and Larry Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *Information Theory, IEEE Transactions on*, 21(1):32–40, 1975.

[31] Yizong Cheng. Mean shift, mode seeking, and clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(8):790–799, 1995.

[32] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):603–619, 2002.

[33] Jia Li, Surajit Ray, and Bruce G Lindsay. A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research*, 8(8), 2007.

[34] Frédéric Chazal, Leonidas J Guibas, Steve Y Oudot, and Primoz Skraba. Persistence-based clustering in riemannian manifolds. *Journal of the ACM (JACM)*, 60(6):41, 2013.

[35] F. Chazal, B.T. Fasy, F. Lecci, B. Michel, A. Rinaldo, and L. Wasserman. Robust topological inference: Distance to a measure and kernel distance. *arXiv preprint arXiv:1412.7197.*, 2014.

[36] Y.C. Chen, C.R. Genovese, and L. Wasserman. Statistical inference using the morse-smale complex. *arXiv preprint arXiv:1506.08826*, 2015.

[37] Y.C. Chen, C.R. Genovese, and L. Wasserman. A comprehensive approach to mode clustering. *Electronic Journal of Statistics*, 10(1):210–241, 2016.

[38] Anil Chaturvedi, Paul E Green, and J Douglas Caroll. K-modes clustering. *Journal of Classification*, 18(1):35–55, 2001.

[39] Larry Wasserman, Martin Azizyan, and Aarti Singh. Feature selection for high-dimensional clustering. *arXiv preprint arXiv:1406.2240*, 2014.

[40] M-cores code release. `http://github.com/hhjiang/mcores`.

[41] Diabetic retinopathy detection. `https://www.kaggle.com/c/diabetic-retinopathy-detection`.

[42] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[43] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11:2837–2854, 2010.

[44] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

Appendix

As discussed in the main text, the results are easily extended to handle more general modal-sets where the density can vary by ϵ_0. We therefore will be showing such more general results which directly imply the results in the main text.

We give a generalization of modal-sets where the density is allowed to vary by $\epsilon_0 \geq 0$, called ϵ_0-modal sets, which will be defined shortly. In order to estimate the ϵ_0-modal sets, we derive Algorithm 2, which is a simple generalization of Algorithm 1. Algorithm 1 is Algorithm 2 with the setting $\epsilon_0 = 0$ and $\tilde{\epsilon} = 0$. Changing $\tilde{\epsilon}$ to larger values will allow us prune false modal-sets away more aggressively, which will be discussed in Appendix F.

Throughout the Appendix, we restate analogues of the results in the main text for the more general ϵ_0-modal sets and Algorithm 2. It will be understood that these results will imply the results in the main text with the setting $\epsilon_0 = 0$ and $\tilde{\epsilon} = 0$.

In Appendix G we formalize common situations well-modeled by modal-sets. In Appendix H we give implementation details.

A ϵ_0-modal sets

Definition 10. *For $\epsilon_0 \geq 0$, connected set M is an* ϵ_0**-modal set** *of f if there exists $f_M > \epsilon_0$ such that $\sup_{x\in M} f(x) = f_M$ and M is a CC of the level set $\mathcal{X}^{f_M - \epsilon_0} := \{x : f(x) \geq f_M - \epsilon_0\}$.*

We require the following Assumption 2 on ϵ_0-modal sets. Note that under Assumption 1 on modal-sets, Assumption 2 on ϵ_0-modal sets will hold for ϵ_0 sufficiently small.

Assumption 2. *The ϵ_0-modal sets are on the interior of $\mathcal{X}$ and $f_M \geq 2\epsilon_0$ for all ϵ_0-modal sets M.*

Remark 4. *Since each ϵ_0-modal set contains a modal-set, it follows that the number of ϵ_0-modal sets is finite.*

The following extends Proposition 1 to show the additional properties of the regions around the ϵ_0-modal sets necessary in our analysis. The proof is in Appendix B.

Proposition 2 (Extends Proposition 1)**.** *For any ϵ_0-modal set M, there exists $\lambda_M, A_M, r_M, l_M, u_M, r_s, S_M$ such that the following holds. A_M is a CC of $\mathcal{X}^{\lambda_M} := \{x : f(x) \geq \lambda_M\}$ containing M which satisifies the following.*

- A_M isolates M by a valley*: A_M does not intersect any other ϵ_0-modal sets and A_M and $\mathcal{X}^{\lambda_M}\backslash A_M$ are r_s-separated by S_M with $r_s > 0$ where r_s does not depend on M.*
- A_M is full-dimensional*: A_M contains an envelope $B(M, r_M)$ of M, with $r_M > 0$.*
- f is smooth around some maximum modal-set in M*: There exists modal-set $M_0 \subseteq M$ such that f has density f_M on M_0 and $f_M - f(x) \leq u_M(d(x, M_0))$ for $x \in B(M_0, r_M)$*
- f is both *smooth* and has *curvature* around M*: u_M and l_M are increasing continuous functions on $[0, r_M]$, $u_M(0) = l_M(0) = 0$ and $u_M(r), l_M(r) > 0$ for $r > 0$, and*

$$l_M(d(x, M)) \leq f_M - \epsilon_0 - f(x) \leq u_M(d(x, M)) \forall x \in B(M, r_M).$$

Next we give admissibility conditions for ϵ_0-modal sets. The only changes (compared to admissibility conditions for modal-sets) are the constant factors. In particular, when $\epsilon_0 = 0$ and $\tilde{\epsilon} = 0$ it is the admissibility conditions for modal-sets. As discussed in the main text, a larger $\tilde{\epsilon}$ value will prune more aggressively at the cost of requiring a larger number of samples. Furthermore, it is implicit below that $\tilde{\epsilon} < l_M(\min\{r_M, r_s\}/2)$. This ensures that we don't prune too aggressively that the estimated ϵ_0-modal sets merge together.

Definition 11. *k is* **admissible** *for an ϵ_0-modal set M if (letting u_M^{-1}, l_M^{-1} be the inverses of u_M, l_M)*

$$\max\left\{\left(\frac{24C_{\delta,n}(\sup_{x\in\mathcal{X}} f(x) + \epsilon_0)}{l_M(\min\{r_M, r_s\}/2) - \tilde{\epsilon}}\right)^2, 2^{7+d}C_{\delta,n}^2\right\}$$
$$\leq k \leq \frac{v_d \cdot (f_M - \epsilon_0)}{2^{2+2d}}\left(u_M^{-1}\left(\frac{C_{\delta,n}(f_M - \epsilon_0)}{2\sqrt{k}}\right)\right)^d \cdot n.$$

Algorithm 2 M-cores (estimating ϵ_0-modal-sets)

Initialize $\widehat{\mathcal{M}} := \emptyset$. Define $\beta_k = 4\frac{C_{\delta,n}}{\sqrt{k}}$.
Sort the X_i's in descending order of f_k values.
for $i = 1$ **to** n **do**
  Define $\lambda := f_k(X_i)$.
  Let A be the CC of $G(\lambda - 9\beta_k\lambda - \epsilon_0 - \tilde{\epsilon})$ that contains X_i.
  if A is disjoint from all cluster-cores in $\widehat{\mathcal{M}}$ **then**
    Add $\widehat{M} := \{x \in A : f_k(x) > \lambda - \beta_k\lambda - \epsilon_0\}$ to $\widehat{\mathcal{M}}$.
  end if
end for
return $\widehat{\mathcal{M}}$.

B Supporting lemmas and propositions

Proof of Proposition 2. Let M be an ϵ_0-modal set with maximum density f_M and minimum density $f_M - \epsilon_0$ (i.e. $f_M - \epsilon_0 \le f(x) \le f_M$ for $x \in M$). Define $\mathcal{X}^\lambda := \{x : f(x) \ge \lambda\}$. Let $A_1, ..., A_m$ be the CCs of $\mathcal{X}^{f_M - \epsilon_0}$ (there are a finite number of CCs since each CC contains at least one modal-set and the number of modal-sets is finite). Define $r_{\min} := \min_{A_i \neq A_j} \inf_{x \in A_i, x' \in A_j} |x - x'|$, which is the minimum distance between pairs of points in different CCs. Next, define the one-sided Hausdorff distance for closed sets A, B: $d_{H'}(A, B) := \max_{x \in A} \min_{x \in B} |x - y|$. Then consider $g(t) := d_{H'}(\mathcal{X}^{f_M - \epsilon_0 - t}, \mathcal{X}^{f_M - \epsilon_0})$.

Since f is continuous and has a finite number of modal-sets, g has a finite number of points of discontinuity (i.e. when $f_M - \epsilon_0 - t$ is the density of some modal-set) and we have $g(t) \to 0$ as $t \to 0$. Thus, there exists $0 < \lambda_M < f_M - \epsilon_0$ such that $g(f_M - \epsilon_0 - \lambda_M) < \frac{1}{4} r_{\min}$ and there are no modal-sets or ϵ_0-modal sets with minimum density in $[\lambda_M, f_M - \epsilon_0)$. For each A_i, there exists exactly one CC of $\mathcal{X}^{\lambda_M}$, A_i', such that $A_i \subset A_i'$. Since $g(f_M - \epsilon_0 - \lambda_M) < \frac{1}{4} r_{\min}$, it follows that $A_i' \subseteq B(A_i, \frac{1}{4} r_{\min})$. Thus, the A_i''s are pairwise separated by distance at least $\frac{1}{2} r_{\min}$. Moreover, there are no other CCs in $\mathcal{X}^{f_M - \epsilon_0}$ because there are no modal-sets with density in $[\lambda_M, f_M - \epsilon_0)$.

Then, let A_M be the CC of $\mathcal{X}^{\lambda_M}$ containing M. Then A_M contains no other ϵ_0-modal sets and it is $\frac{1}{5} r_{\min}$-separated by $\mathcal{X}^{\lambda_M} \backslash M$ by some set S_M (i.e. take $S_M := \{x : d(x, A_M) = \frac{1}{5} r_{\min}\}$). Since there is a finite number of modal-sets, it suffices to take r_s to be the minimum of the corresponding $\frac{1}{5} r_{\min}$ for each ϵ_0-modal set. This resolves the first part of the proposition.

Let $h(r) := \inf_{x \in B(M,r)} f(x)$. Since f is continuous, h is continuous and decreasing with $h(0) = f_M - \epsilon_0 > \lambda_M$. Take $r_M > 0$ sufficiently small so that $h(r_M) > \lambda_M$. This resolves the second part of the proposition.

Take M_0 to be some modal-set with density f_M in M. One must exist since M has local-maxima at level f_M. For each r, let $u_M(r) := \max\{f_M - \epsilon_0 - \inf_{x \in B(M,r)} f(x), f_M - \inf_{x \in B(M_0,r)} f(x)\}$. Then, we have $f_M - f(x) \le u_M(d(x, M_0))$ and $f_M - \epsilon_0 - f(x) \le u_M(d(x, M))$. Clearly u_M is increasing on $[0, r_M]$ with $u_M(0) = 0$ and continuous since f is continuous. If u_M is not strictly increasing then we can replace it with a strictly increasing continuous function while still having $u_M(r) \to 0$ as $r \to 0$ (i.e. by adding an appropriate strictly increasing continuous function). This resolves the third part of the proposition and the upper bound in the fourth part of the proposition.

Now, define $g_M(t) := d(\mathcal{X}^{f_M - \epsilon_0 - t} \cap A_M, M)$ for $t \in [0, \frac{1}{2}(f_M - \epsilon_0 - \lambda_M)]$. Then, g_M is continuous, $g_M(0) = 0$ and is strictly increasing. Define l_M to be the inverse of g_M. Clearly l_M is continuous, strictly increasing, and $l_M(r) \to 0$ as $r \to 0$. From the definition of g_M, it follows that for $x \in B(M, r_M)$, $f_M - \epsilon_0 - f(x) \ge l_M(d(x, M))$ as desired. ☐

We need the following result giving guarantees on the empirical balls.

Lemma 4 ([2]). *Pick $0 < \delta < 1$. Assume that $k \ge d \log n$. Then with probability at least $1 - \delta$, for every ball*

$B \subset \mathbb{R}^d$ we have

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

Remark 5. *Throughout the paper, there are results which are qualified to hold with probability at least $1-\delta$. This is understood to be precisely the event that Lemma 4 holds.*

Lemma 5 of [16] establishes convergence rates for f_k.

Definition 12. *For $x \in \mathbb{R}^d$ and $\epsilon > 0$, define $\hat{r}(\epsilon, x) := \sup\left\{r : \sup_{x' \in B(x,r)} f(x') - f(x) \leq \epsilon\right\}$ and $\check{r}(\epsilon, x) := \sup\left\{r : \sup_{x' \in B(x,r)} f(x) - f(x') \leq \epsilon\right\}$.*

Lemma 5 (Bounds on f_k)**.** *Suppose that $\frac{C_{\delta,n}}{\sqrt{k}} < \frac{1}{2}$. Then the follow two statements each hold with probability at least $1-\delta$:*

$$f_k(x) < \left(1 + 2\frac{C_{\delta,n}}{\sqrt{k}}\right)(f(x) + \epsilon),$$

for all $x \in \mathbb{R}^d$ and all $\epsilon > 0$ provided k satisfies $v_d \cdot \hat{r}(\epsilon, x)^d \cdot (f(x) + \epsilon) \geq \frac{k}{n} - C_{\delta,n}\frac{\sqrt{k}}{n}$.

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

for all $x \in \mathbb{R}^d$ and all $\epsilon > 0$ provided k satisfies $v_d \cdot \check{r}(\epsilon, x)^d \cdot (f(x) - \epsilon) \geq \frac{k}{n} + C_{\delta,n}\frac{\sqrt{k}}{n}$.

Lemma 6 (Extends Lemma 1)**.** *(Upper bound on r_n) Let M be an ϵ_0-modal set with maximum density f_M and suppose that k is admissible. With probability at least $1-\delta$,*

$$r_n(M) \leq \left(\frac{2C_{\delta,n}\sqrt{d\log n}}{n \cdot v_d \cdot (f_M - \epsilon_0)}\right)^{1/d}.$$

Proof of Lemma 6. Define $r_0 := \left(\frac{2C_{\delta,n}\sqrt{d\log n}}{nv_d\cdot(f_M-\epsilon_0)}\right)^{1/d}$ and $r := (4k/(nv_d f_M))^{1/d}$. Since k is admissible, we have that $u_M(r_0) \leq u_M(r) \leq (f_M - \epsilon_0)/2$. We have

$$\mathcal{F}(B(x, r_0)) \geq v_d {r_0}^d (f_M - \epsilon_0 - u_M(r_0)) \geq v_d {r_0}^d (f_M - \epsilon_0)/2 = \frac{C_{\delta,n}\sqrt{d\log n}}{n}.$$

By Lemma 4, this implies that $\mathcal{F}_n(B(x, r_0)) > 0$ with probability at least $1-\delta$ and therefore we have $r_n(x) \leq r_0$. □

C Isolation Results

The following extends Lemma 2 to handle more general ϵ_0-modal sets and pruning parameter $\tilde{\epsilon}$.

Lemma 7 (Extends Lemma 2)**.** *(Isolation) Let M be an ϵ_0-modal set and k be admissible for M. Suppose $0 \leq \tilde{\epsilon} < l_M(\min\{r_M, r_s\}/2)$ and there exists a sample point in $\mathcal{X}_M$. Define $\hat{x}_M := \mathrm{argmax}_{x \in \mathcal{X}_M \cap X_{[n]}} f_k(x)$. Then the following holds with probability at least $1-\delta$: when processing sample point $\hat{x}_M$ in Algorithm 2 we will add $\widehat{M}$ to $\widehat{\mathcal{M}}$ where $\widehat{M}$ does not contain points outside of $\mathcal{X}_M$.*

Proof. Define $\widehat{f}_M := f_k(\hat{x}_M)$, $\lambda = \widehat{f}_M$ and $\bar{r} := \min\{r_M, r_s\}/2$. It suffices to show that (i) $\mathcal{X}\backslash\mathcal{X}_M$ and $B(M, \bar{r})$ are disconnected in $G(\lambda - 9\beta_k\lambda - \epsilon_0 - \tilde{\epsilon})$ and (ii) $\hat{x}_M \in B(M, \bar{r})$.

In order to show (i), we first show that $G(\lambda - 9\beta_k\lambda - \epsilon_0 - \tilde{\epsilon})$ contains no points from $B(S_M, r_s/2)$ and no points from $\mathcal{X}_M \backslash B(M, \bar{r})$. Then, all that will be left is showing that there are no edges between $B(M, \bar{r})$ and $\mathcal{X} \backslash \mathcal{X}_M$.

We first prove bounds on f_k that will help us show (i) and (ii). Let $\bar{F} := f_M - \epsilon_0 - l_M(\bar{r}/2)$. Then for all $x \in \mathcal{X}_M \backslash B(M, \bar{r})$, we have $\hat{r}(\bar{F} - f(x), x) \geq \bar{r}/2$. Thus the conditions for Lemma 5 are satisfied by the admissibility of k and hence $f_k(x) < \left(1 + 2\frac{C_{\delta,n}}{\sqrt{k}}\right)\bar{F}$. Now,

$$\sup_{x \in \mathcal{X}_M \backslash B(M, \bar{r})} f_k(x) < (1 + 2\frac{C_{\delta,n}}{\sqrt{k}})\bar{F} = (1 + 2\frac{C_{\delta,n}}{\sqrt{k}})(f_M - \epsilon_0 - l_M(\bar{r}/2))$$
$$\leq (1 + 2\frac{C_{\delta,n}}{\sqrt{k}})^3 \widehat{f}_M - (1 + 2\frac{C_{\delta,n}}{\sqrt{k}}) \cdot (\epsilon_0 + l_M(\bar{r}/2)) \leq \lambda - 9\beta_k\lambda - \epsilon_0 - \tilde{\epsilon},$$

where the second inequality holds by using Lemma 5 as follows. Choose $x \in M_0$ and $\epsilon = \frac{C_{\delta,n}}{2\sqrt{k}} f_M$. Then $\check{r}(\epsilon, x) \geq u^{-1}(\epsilon)$. The conditions for Lemma 5 hold by the admissibility of k and thus $\widehat{f}_M \geq f_k(x) \geq (1 - C_{\delta,n}/\sqrt{k})^2 f_M$. Furthermore it follows from Lemma 5 that $\widehat{f}_M < (1 + 2C_{\delta,n}/\sqrt{k}) f_M$; combine this admissibility of k to obtain the last inequality. Finally, from the above, we also have $\sup_{x \in \mathcal{X}_M \backslash B(M, \bar{r})} f_k(x) < \widehat{f}_M$, implying (ii).

Next, if $x \in B(S_M, r_s/2)$, then $\hat{r}(\bar{F} - f(x), x) \geq \bar{r}/2$ and the same holds for $B(S_M, r_s/2)$:

$$\sup_{x \in B(S_M, r_s/2)} f_k(x) < \lambda - 9\beta_k\lambda - \epsilon_0 - \tilde{\epsilon}.$$

Thus, $G(\lambda - 9\beta_k\lambda - \epsilon_0 - \tilde{\epsilon})$ contains no point from $B(S_M, r_s/2)$ and no point from $\mathcal{X}_M \backslash B(M, \bar{r})$.

All that remains is showing that there is no edge between $B(M, \bar{r})$ and $\mathcal{X} \backslash \mathcal{X}_M$. It suffices to show that any such edge will have length less than r_s since $B(S_M, r_s/2)$ separates them by a width of r_s. We have for all $x \in B(M, \bar{r})$,

$$\mathcal{F}(B(x, \bar{r})) \geq v_d \bar{r}^d \inf_{x' \in B(x, 2\bar{r})} f(x') \geq \frac{k}{n} + C_{\delta,n}\frac{\sqrt{k}}{n}.$$

Thus by Lemma 4, we have $r_k(x) \leq \bar{r} < r_s$, establishing (i).

□

D Integrality Results

The goal is to show that the $\widehat{M} \in \widehat{\mathcal{M}}$ refered to above contains $B(M, r_n(M))$. We give a condition under which $B(M, r_n(M)) \cap X_{[n]}$ would be connected in $G(\lambda)$ for some λ. It is adapted from arguments in Theorem V.2 in [6].

Lemma 8. *(Connectedness) Let M be an ϵ_0-modal set and k be admissible for M. Then with probability at least $1 - \delta$, $\mathcal{X}_M$ contains a sample point and $B(M, r_n(M)) \cap X_{[n]}$ is connected in $G(\lambda)$ if*

$$\lambda \leq \left(1 - \frac{C_{\delta,n}}{\sqrt{k}}\right)^2 (f_M - \epsilon_0).$$

Proof. We first show that $B(M, r_n(M)) \cap X_{[n]}$ is connected in $G(\lambda)$ when $\lambda \leq \left(1 - \frac{C_{\delta,n}}{\sqrt{k}}\right)^2 (f_M - \epsilon_0)$. For simplicity of notation, let $A := B(M, r_n(M))$. It suffices to prove the result for $\lambda = (1 - \frac{C_{\delta,n}}{\sqrt{k}})^2 (f_M - \epsilon_0)$. Define $r_\lambda = (k/(nv_d\lambda))^{1/d}$ and $r_o = (k/(2nv_d f_M))^{1/d}$. First, we show that each $x \in B(A, r_\lambda)$, there is a sample point in $B(x, r_o)$. We have for $x \in B(A, r_\lambda)$,

$$\mathcal{F}(B(x, r_o)) \geq v_d r_o^d \inf_{x' \in B(x, r_o + r_\lambda)} f(x') \geq v_d r_o^d (f_M - \epsilon_0 - u_M(r_o + r_\lambda + r_n(M)))$$
$$\geq v_d r_o^d (f_M - \epsilon_0)\left(1 - \frac{C_{\delta,n}}{\sqrt{k}}\right) \geq C_{\delta,n}\frac{\sqrt{d \log n}}{n}.$$

Thus by Lemma 4 we have that with probability at least $1-\delta$, $B(x, r_o)$ contains a sample uniformly over $x \in B(A, r_\lambda)$.

Now, let x and x' be two points in $A \cap X_{[n]}$. We now show that there exists $x = x_0, x_1, ..., x_p = x'$ such that $||x_i - x_{i+1}|| < r_o$ and $x_i \in B(A, r_o)$. For arbitrary $\gamma \in (0,1)$, we can choose $x = z_0, z_1, ..., z_p = x'$ where $||z_{i+1} - z_i|| \leq \gamma r_o$. Next, choose γ sufficiently small such that

$$v_d \left(\frac{(1-\gamma)r_o}{2}\right)^d \inf_{z \in B(A, r_0)} f(z) \geq \frac{C_{\delta,n}\sqrt{d \log n}}{n},$$

then there exists a sample point x_i in $B(z_i, (1-\gamma)r_o/2)$. Moreover we obtain that

$$||x_{i+1} - x_i|| \leq ||x_{i+1} - z_{i+1}|| + ||z_{i+1} - z_i|| + ||z_i - x_i|| \leq r_o.$$

All that remains is to show $(x_i, x_{i+1}) \in G(\lambda)$. We see that $x_i \in B(A, r_o)$. However, for each $x \in B(A, r_o)$, we have

$$\mathcal{F}(B(x, r_\lambda)) \geq v_d r_\lambda^d \inf_{x' \in B(x, r_o + r_\lambda)} f(x') \geq v_d r_\lambda^d (f_M - \epsilon_0)\left(1 - \frac{C_{\delta,n}}{\sqrt{k}}\right) \geq \frac{k}{n} + \frac{C_{\delta,n}\sqrt{k}}{n}.$$

Thus $r_k(x_i) \leq r_\lambda$ for all i. Therefore, $x_i \in G(\lambda)$ for all x_i. Finally, $||x_{i+1} - x_i|| \leq r_o \leq \min\{r_k(x_i), r_k(x_{i+1})\}$ and thus $(x_i, x_{i+1}) \in G(\lambda)$. Therefore, $A \cap X_{[n]}$ is connected in $G(\lambda)$, as desired.

Finally, to show that $\mathcal{X}_M$ contains a sample point, it suffices to have $r_n(x) \leq r_M$ for some $x \in M$. Indeed, we have $r_n(M) \leq r_k(x) \leq (k/(2nv_d f_M))^{1/d} < r_M$ where the second inquality follows from the earlier fact that $B(x, r_0)$ contains a sample point for all points in A and the last inequality follows from the admissibility of k. ☐

The following extends Lemma 3 handle more general ϵ_0-modal sets.

Lemma 9 (Extends Lemma 3)**.** *(Integrality) Let M be an ϵ_0-modal set with density f_M, and suppose k is admissible for M. Let $\hat{x}_M := \text{argmax}_{x \in \mathcal{X}_M \cap X_{[n]}} f_k(x)$. Then the following holds with probability at least $1-\delta$. First, $\mathcal{X}_M$ does contain a sample point. When processing sample point $\hat{x}_M$ in Algorithm 1, if we add $\widehat{M}$ to $\widehat{\mathcal{M}}$, then $B(M, r_n(M)) \cap X_{[n]} \subseteq \widehat{M}$.*

Proof. By Lemma 8, $\mathcal{X}_M$ contains a sample point and $B(M, r_n(M)) \cap X_{[n]}$ is connected in $G(\lambda_0)$ when $\lambda_0 \leq (1 - \frac{C_{\delta,n}}{\sqrt{k}})^2 (f_M - \epsilon_0)$. Define $\widehat{f}_M := f_k(\hat{x}_M)$ and $\lambda := \widehat{f}_M$. It suffices to show that $B(M, r_n(M)) \cap X_{[n]}$ is connected in $G(\lambda - 9\beta_k\lambda - \tilde{\epsilon})$. Indeed, we have that

$$\begin{aligned}\left(1 - \frac{C_{\delta,n}}{\sqrt{k}}\right)^2 (f_M - \epsilon_0) &\geq \widehat{f}_M \left(1 - \frac{C_{\delta,n}}{\sqrt{k}}\right)^2 \Big/ \left(1 + 2\frac{C_{\delta,n}}{\sqrt{k}}\right) - \left(1 - \frac{C_{\delta,n}}{\sqrt{k}}\right)^2 \epsilon_0 \\ &\geq \lambda - \beta_k \lambda - \epsilon_0 \geq \lambda - 9\beta_k\lambda - \epsilon_0 - \tilde{\epsilon},\end{aligned}$$

where the first inequality follows from Lemma 5, as desired. ☐

E Theorem 1

Combining the isolation and integrality, we obtain the following extention of Corollary 1.

Corollary 2 (Extends Corollary 1)**.** *(Identification) Suppose we have the assumptions of Lemmas 7 and 9 for ϵ_0-modal set M. Define $\widehat{f}_M := \max_{x \in \mathcal{X}_M \cap X_{[n]}} f_k(x)$ and $\lambda := \widehat{f}_M$. With probability at least $1-\delta$, there exists $\widehat{M} \in \widehat{\mathcal{M}}$ such that $B(M, r_n(M)) \cap X_{[n]} \subseteq \widehat{M} \subseteq \{x \in \mathcal{X}_M \cap X_{[n]} : f_k(x) \geq \lambda - \beta_k\lambda - \epsilon_0\}$*

Proof. By Lemma 7, there exists $\widehat{M} \in \widehat{\mathcal{M}}$ which contains only points in $\mathcal{X}_M$ with maximum f_k value of $\widehat{f}_M$. Thus, we have $\widehat{M} \subseteq \{x \in \mathcal{X}_M \cap X_{[n]} : f_k(x) \geq \widehat{f}_M - \beta_k \widehat{f}_M - \epsilon_0\}$. By Lemma 9, $B(M, r_n(M)) \cap X_{[n]} \subseteq \widehat{M}$. ☐

The following extends Theorem 1 to handle more general ϵ_0-modal sets and pruning parameter $\tilde{\epsilon}$.

Theorem 3 (Extends Theorem 1)**.** *Let $\delta > 0$ and M be an ϵ_0-modal set. Suppose k is admissible for M and $0 \le \tilde{\epsilon} < l_M(\min\{r_M, r_s\}/2)$. Then with probability at least $1-\delta$, there exists $\widehat{M} \in \widehat{\mathcal{M}}$ such that*

$$d(M, \widehat{M}) \le l_M^{-1}\left(\frac{8C_{\delta,n}}{\sqrt{k}} f_M\right),$$

which goes to 0 as $C_{\delta,n}/\sqrt{k} \to 0$.

Proof. Define $\tilde{r} = l_M^{-1}\left(\frac{8C_{\delta,n}}{\sqrt{k}} f_M\right)$. There are two directions to show: $\max_{x\in\widehat{M}} d(x, M) \le \tilde{r}$ and $\sup_{x\in M} d(x, \widehat{M}) \le \tilde{r}$ with probability at least $1-\delta$.

We first show $\max_{x\in\widehat{M}} d(x, M) \le \tilde{r}$. By Corollary 2 we have $\widehat{M} \in \widehat{\mathcal{M}}$ such that $\widehat{M} \subseteq \{x \in \mathcal{X}_M : f_k(x) \ge \widehat{f}_M - \beta_k \widehat{f}_M - \epsilon_0\}$ where $\widehat{f}_M := \max_{x\in\mathcal{X}_M \cap X_{[n]}} f_k(x)$. Hence, it suffices to show

$$\inf_{x\in B(M_0, r_n(M))} f_k(x) \ge \sup_{\mathcal{X}_M \setminus B(M,\tilde{r})} f_k(x) + \beta_k \widehat{f}_M + \epsilon_0. \tag{2}$$

Define $r := (4/f_M v_d)^{1/d}(k/n)^{1/d}$. For any $x \in B(M_0, r + r_n(M))$, $f(x) \ge f_M - u_M(r + r_n(M)) := \check{F}$. Thus, for any $x \in B(M_0, r_n(M))$ we can let $\epsilon = f(x) - \check{F}$ and thus $\check{r}(\epsilon, x) \ge r$ and hence the conditions for Lemma 5 are satisfied. Therefore, with probability at least $1-\delta$,

$$\inf_{x\in B(M_0, r_n(M))} f_k(x) \ge \left(1 - \frac{C_{\delta,n}}{\sqrt{k}}\right)(f_M - u_M(r + r_n(M))). \tag{3}$$

For any $x \in \mathcal{X}_M \setminus B(M, \tilde{r}/2)$, $f(x) \le f_M - \epsilon_0 - l_M(\tilde{r}/2) := \hat{F}$. Now, for any $x \in \mathcal{X} \setminus B(M, \tilde{r})$, let $\epsilon := \hat{F} - f(x)$. We have $\hat{r}(\epsilon, x) \ge \tilde{r}/2 = l_M^{-1}(8C_{\delta,n}/\sqrt{k})/2 \ge l_M^{-1}(u_M(2r))/2 \ge r$ (since l_M is increasing and $l_M \le u_M$) and thus the conditions for Lemma 5 hold. Hence, with probability at least $1-\delta$,

$$\sup_{x\in\mathcal{X}_M \setminus B(M,\tilde{r})} f_k(x) \le \left(1 + 2\frac{C_{\delta,n}}{\sqrt{k}}\right)(f_M - \epsilon_0 - l_M(\tilde{r})). \tag{4}$$

Thus, by (3) and (4) applied to (2) it suffices to show that

$$\left(1 - \frac{C_{\delta,n}}{\sqrt{k}}\right)(f_M - u_M(r + r_n(M))) \ge \left(1 + 2\frac{C_{\delta,n}}{\sqrt{k}}\right)(f_M - \epsilon_0 - l_M(\tilde{r})) + \beta_k \widehat{f}_M + \epsilon_0, \tag{5}$$

which holds when

$$l_M(\tilde{r}) \ge u_M(r + r_n(M)) + \frac{3C_{\delta,n}}{\sqrt{k}} f_M + \beta_k \widehat{f}_M. \tag{6}$$

The admissibility of k ensures that $r_n(M) \le r \le r_M/2$ so that the regions of $\mathcal{X}$ we are dealing with in this proof are confined within $B(M_0, r_M)$ and $B(M, r_M)\setminus M$.

By the admissibility of k, $u_M(2r) \le \frac{C_{\delta,n}}{2\sqrt{k}} f_M$. This gives

$$l_M(\tilde{r}) = \frac{8C_{\delta,n}}{\sqrt{k}} f_M \ge u_M(2r) + \frac{15C_{\delta,n}}{2\sqrt{k}} f_M \ge u_M(r + r_n(M)) + \frac{3C_{\delta,n}}{\sqrt{k}} f_M + \beta_k \widehat{f}_M,$$

where the second inequality holds since $C_{\delta,n}/\sqrt{k} < 1/16$, u is increasing, $r \ge r_n(M)$, and $\widehat{f}_M \le \left(1 + 2\frac{C_{\delta,n}}{\sqrt{k}}\right) f_M$ by Lemma 5. Thus, showing (6), as desired.

This shows one direction of the Hausdorff bound. We now show the other direction, that $\sup_{x\in M} d(x, M) \le \tilde{r}$.

It suffices to show for each point $x \in M$ that the distance to the closest sample point $r_n(x) \le \tilde{r}$ since $\widehat{M}$ contains these sample points by Corollary 2. However, by Lemma 6 and the admissibility of k, $r_n(x) \le \tilde{r}$ as desired. $\square$

F Theorem 2

We need the following Lemma 10 which gives guarantees us that given points in separate CCs of the pruned graph, these points will also be in separate CCs of f at a nearby level. [6] gives a result for a different graph and the proof can be adapted to give the same result for our graph (but slightly different assumptions on k).

Lemma 10 (Separation of level sets under pruning, [6])**.** *Fix $\epsilon > 0$ and let $r(\epsilon) := \inf_{x\in\mathbb{R}^d} \min\{\hat{r}(\epsilon, x), \check{r}(\epsilon, x)\}$. Define $\Lambda := \max_{x\in\mathbb{R}^d} f(x)$ and assume $\tilde{\epsilon}_0 \geq 2\epsilon + \beta_k(\lambda_f + \epsilon)$ and let $\tilde{G}(\lambda)$ be the graph with vertices in $G(\lambda)$ and edges between pairs of vertices if they are connected in $G(\lambda - \tilde{\epsilon}_0)$. Then the following holds with probability at least $1 - \delta$.*

Let $\tilde{A}_1$ and $\tilde{A}_2$ denote two disconnected sets of points $\tilde{G}(\lambda)$. Define $\lambda_f := \inf_{x\in\tilde{A}_1\cup\tilde{A}_2} f(x)$. Then $\tilde{A}_1$ and $\tilde{A}_2$ are disconnected in the level set $\{x \in \mathcal{X} : f(x) \geq \lambda_f\}$ if k satisfies

$$v_d(r(\epsilon)/2)^d(\lambda_f - \epsilon) \geq \frac{k}{n} + C_{\delta,n}\frac{\sqrt{k}}{n}$$

and

$$k \geq \max\{8\Lambda C_{\delta,n}^2/(\lambda_f - \epsilon), 2^{d+7}C_{\delta,n}^2\}.$$

Proof. We prove the contrapositive. Let A be a CC of $\{x \in \mathcal{X} : f(x) \geq \lambda_f\}$ with $\lambda_f = \min_{x\in A\cap X_{[n]}} f(x)$. Then it suffices to show $A \cap X_{[n]}$ is connected in $G(\lambda')$ for $\lambda' := \min_{x\in A\cap X_{[n]}} f_k(x) - \tilde{\epsilon}_0$.

We first show $A \cap X_{[n]}$ is connected in $G(\lambda)$ for $\lambda = (\lambda_f - \epsilon)/(1 + C_{\delta,n}/\sqrt{k})$ and all that will remain is showing $\lambda' \leq \lambda$.

Define $r_o := (k/(2nv_d f_M))^{1/d}$ and $r_\lambda := (k/(nv_d\lambda))^{1/d}$. Then from the first assumption on k, it follows that $r_\lambda \leq r(\epsilon)/2$. Now for each $x \in B(A, r_\lambda)$, we have

$$\mathcal{F}(B(x, r_o)) \geq v_d r_o^d \inf_{x'\in B(x,r_o+r_\lambda)} f(x') \geq v_d r_o^d(\lambda_f - \epsilon) \geq C_{\delta,n}\frac{\sqrt{d\log n}}{n}.$$

Thus, by Lemma 4 we have with probability at least $1 - \delta$ that $B(x, r_0)$ contains a sample point.

Now, in the same way shown as in Lemma 8, we have the following. If x and x' be two points in $A \cap X_{[n]}$ then there exists $x = x_0, x_1, ..., x_p = x'$ such that $||x_i - x_{i+1}|| < r_o$ and $x_i \in B(A, r_o)$.

Next is showing $(x_i, x_{i+1}) \in G(\lambda)$. We see that $x_i \in B(A, r_o)$. However, for each $x \in B(A, r_o)$, we have

$$\mathcal{F}(B(x, r_\lambda)) \geq v_d r_\lambda^d \inf_{x'\in B(x,r_o+r_\lambda)} f(x') \geq v_d r_\lambda^d(\lambda_f - \epsilon) \geq \frac{k}{n} + \frac{C_{\delta,n}\sqrt{k}}{n}.$$

Thus $r_k(x_i) \leq r_\lambda$ for all i. Therefore, $x_i \in G(\lambda)$ for all x_i. Finally, $||x_{i+1} - x_i|| \leq r_o \leq \min\{r_k(x_i), r_k(x_{i+1})\}$ and thus $(x_i, x_{i+1}) \in G(\lambda)$. Therefore, $A \cap X_{[n]}$ is connected in $G(\lambda)$.

All that remains is showing $\lambda' \leq \lambda$. We have

$$\lambda' = \min_{x\in A\cap X_{[n]}} f_k(x) - \tilde{\epsilon}_0 \leq \left(1 + 2\frac{C_{\delta,n}}{\sqrt{k}}\right)(\lambda_f + \epsilon) - \tilde{\epsilon}_0 \leq \lambda,$$

where the first inequality holds by Lemma 5, and the second inequality holds from the assumption on $\tilde{\epsilon}_0$, as desired.

□

We state the pruning result for more general choices of $\tilde{\epsilon}$. Its proof is standard and given here for completion. (See e.g. [16]).

Theorem 4 (Extends Theorem 2)**.** *Let $0 < \delta < 1$ and $\tilde{\epsilon} \geq 0$. There exists $\lambda_0 = \lambda_0(n, k)$ such that the following holds with probability at least $1 - \delta$. All ϵ_0-modal set estimates in $\widehat{\mathcal{M}}$ chosen at level $\lambda \geq \lambda_0$ can be injectively mapped to ϵ_0-modal sets $\left\{M : \lambda_M \geq \min_{\{x \in \mathcal{X}_{[n]} : f_k(x) \geq \lambda - \beta_k \lambda\}} f(x)\right\}$, provided k is admissible for all such M.*

In particular, if f is Hölder-continuous, (i.e. $||f(x) - f(x')|| \leq c||x - x'||^\alpha$ for some $0 < \alpha \leq 1$, $c > 0$) and $\tilde{\epsilon} = 0$, then $\lambda_0 \to 0$ as $n \to \infty$, provided $C_1 \log n \leq k \leq C_2 n^{2\alpha/(2\alpha+d)}$, for some C_1, C_2 independent of n.

Proof. Define $r(\epsilon) := \inf_{x \in \mathbb{R}^d} \min\{\hat{r}(\epsilon, x), \check{r}(\epsilon, x)\}$. Since f is uniformly continuous, it follows that $r(0) = 0$, r is increasing, and $r(\epsilon) > 0$ for $\epsilon > 0$.

Thus, there exists $\tilde{\lambda}_{n,k,\tilde{\epsilon}} > 0$ such that

$$\tilde{\lambda}_{n,k,\tilde{\epsilon}} = \frac{k}{n \cdot v_d \cdot r((8\beta_k \tilde{\lambda}_{n,k,\tilde{\epsilon}} + \tilde{\epsilon})/3)}.$$

Define

$$\lambda_0 := \max\{\tilde{\lambda}_{n,k,\tilde{\epsilon}}, 32\beta_k \sup_{x \in \mathcal{X}} f(x) + 4\tilde{\epsilon}\}.$$

Let us identify each estimated ϵ_0-modal set $\widehat{M}$ with the point $\hat{x}_M := \max_{x \in \widehat{M}} f_k(x)$. Let us call these points modal-points. Then it suffices to show that there is an injection from modal-points to the ϵ_0-modal sets.

Define $G'(\lambda)$ to be the graph with vertices in $G(\lambda - \beta_k \lambda)$ and edges between vertices if they are in the same CC of $G(\lambda - 9\beta_k \lambda - \tilde{\epsilon})$ and $X_{[n]}^\lambda := \{x : f_k(x) \geq \lambda\}$. Let $\tilde{A}_{i,\lambda} := \tilde{A}_i \cap X_{[n]}^\lambda$ for $i = 1, ..., m$ to be the vertices of the CCs of $G'(\lambda)$ which do not contain any modal-points chosen thus far as part of estimated modal-sets.

Fix level $\lambda > 0$ such that $\lambda_f := \inf_{x \in X_{[n]}^\lambda} f(x) \geq \lambda_0/2$. Then the conditions are satisified for Lemma 10 with $\epsilon = (8\beta_k \lambda + \tilde{\epsilon})/3$. Suppose that $\tilde{A}_{1,\lambda}, ..., \tilde{A}_{m,\lambda}$ are in ascending order according to $\lambda_{i,f} := \min_{x \in \tilde{A}_{i,\lambda}} f(x)$. Starting with $i = 1$, by Lemma 10, $\mathcal{X}^{\lambda_1, f}$ can be partitioned into disconneced subsets A_1 and $\mathcal{X}^{\lambda_1, f} \backslash A_1$ containing respectively $\tilde{A}_{1,\lambda}$ and $\cup_{i=2}^m \tilde{A}_{i,\lambda}$. Assign the modal-point $\text{argmax}_{x \in \tilde{A}_{1,\lambda}} f_k(x)$ to any ϵ_0-modal set in A_1. Repeat the same argument successively for any $\tilde{A}_{i,\lambda}$ and $\cup_{j=i+1}^m \tilde{A}_{j,\lambda}$ until all modal-points are assigned to distinct ϵ_0-modal sets in disjoint sets A_i.

Now by Lemma 10, $\mathcal{X}^{\lambda_f}$ can be partitioned into disconnected subsets A and $\mathcal{X}^{\lambda_f} \backslash A$ containing respectively $\tilde{A}_\lambda := \cup_{i=1}^m \tilde{A}_{i,\lambda}$ and $\mathcal{X}_{[n]}^{\lambda_f} \backslash \tilde{A}_\lambda$. Thus, the modal-points in $\tilde{A}_\lambda$ were assigned to ϵ_0-modal sets in A.

Now we repeat the argument for all $\lambda' > \lambda$ to show that the modal-points in $X_{[n]}^\lambda \backslash \tilde{A}_\lambda$ can be assigned to distinct ϵ_0-modal sets in $\mathcal{X}^\lambda \backslash A^\lambda$. (We have $\lambda'_f := \min_{x \in X_{[n]}^{\lambda' - \beta_k \lambda'}} f(x) \geq \lambda_f$).

Finally, it remains to show that $\lambda \geq \lambda_0$ implies $\lambda_f \geq \lambda_0/2$. We have $\lambda_0/4 \geq 8\beta_k \lambda + \tilde{\epsilon}$, thus $r(\lambda_0/4) \geq r(8\beta_k \lambda + \tilde{\epsilon})$. It follows that

$$v_d (r(\lambda_0/4))^d \cdot (\lambda_0/4) \geq \frac{k}{n} + C_{\delta,n} \frac{\sqrt{k}}{n}.$$

Hence, for all x such that $f(x) \leq \lambda_0/2$, we have

$$f_k(x) \leq (1 + 2\frac{C_{\delta,n}}{\sqrt{k}})(f(x) + \lambda_0/4) \leq \lambda_0.$$

To see the second part, suppose we have $C_1, C_2 > 0$ such that $C_1 \log n \leq k \leq C_2 n^{2\alpha/(2\alpha+d)}$. This combined with the fact that $r(\epsilon) \geq (\epsilon/C)^{1/\alpha}$ implies $\lambda_0 \to 0$, as desired.

□

G Point-Cloud Density

Here we formalize the fact that modal-sets can serve as good models for high-density structures in data, for instance a low-dimensional structure M + noise.

Lemma 11. *(Point Cloud with Gaussian Noise) Let $M \subseteq \mathbb{R}^d$ be compact (with possibly multiple connected-components of differing dimension). Then there exists a density f over $\mathbb{R}^d$ such that the density is uniform in M and has Gaussian decays around M i.e.*

$$f(x) = \frac{1}{Z}\exp(-d(x,M)^2/(2\sigma^2)),$$

where $\sigma > 0$ and $Z > 0$ depends on M, σ. Thus, the modal-sets of f are the connected-components of M.

Proof. Since M is compact in $\mathbb{R}^d$, it is bounded. Thus there exists $R > 0$ such that $M \subseteq B(0,R)$. It suffices to show that for any $\sigma > 0$,

$$\int_{\mathbb{R}^d} \exp(-d(x,M)^2/(2\sigma^2))dx < \infty.$$

By a scaling of x by σ, it suffices to show that

$$\int_{\mathbb{R}^d} g(x)dx < \infty,$$

where $g(x) := \exp(-\frac{1}{2}d(x,M)^2)$. Consider level sets $\mathcal{X}^\lambda := \{x \in \mathbb{R}^d : g(x) \geq \lambda\}$. Note that $\mathcal{X}^\lambda \subseteq B(M, \sqrt{2\log(1/\lambda)})$ based on the decay in g around M. Clearly the image of g is $(0,1]$ so consider partitioning this range into intervals $[1, 1/2], [1/2, 1/3], ...$. Then it follows that

$$\begin{aligned}\int_{\mathbb{R}^d} g(x)dx &\leq \sum_{n=2}^\infty \text{Vol}(\mathcal{X}^{1/n})\left(\frac{1}{n-1} - \frac{1}{n}\right) \leq \sum_{n=2}^\infty \frac{\text{Vol}(B(M, \sqrt{2\log(n)}))}{(n-1)n} \\ &\leq \sum_{n=2}^\infty \frac{\text{Vol}(B(0, R + \sqrt{2\log(n)}))}{(n-1)n} = \sum_{n=2}^\infty \frac{v_d(R + \sqrt{2\log(n)})^d}{(n-1)n} \\ &\leq v_d \cdot 2^{d-1}\sum_{n=2}^\infty \frac{R^d + (2\log(n))^{d/2}}{(n-1)n} < \infty,\end{aligned}$$

where the last inequality holds by AM-GM. As desired. □

H Implementation

In this section, we explain how to implement Algorithm 2 (which supersedes Algorithm 1) efficiently. Here we assume that for our sample $X_{[n]}$, we have the k-nearest neighbors for each sample point. In our implementation, we simply use kd-tree, although one could replace it with any method that can produce the k-nearest neighbors for all the sample points. In particular, one could use approximate k-NN methods if scale is an issue.

This section now concerns with what remains: constructing a data structure that maintains the CCs of the mutual k-NN graph as we traverse down the levels. At level λ in Algorithm 2, we must keep track of the mutual k-NN graph for points x such that $f_k(x) \geq \lambda - 9\beta_k\lambda - \epsilon_0 - \tilde{\epsilon}$. Thus as λ decreases, we add more vertices (and corresponding edges to the mutual k-nearest neighbors). Algorithm 3 shows what functions this data structure must support. Namely, adding nodes and edges, getting CCs of nodes, and checking if a CC intersects with the current estimates of the ϵ_0-modal sets.

We implement this data structure as a disjoint-set forest data structure. The CCs can be represented as disjoint-sets of forests. Adding a node corresponds to making a set while adding an edge corresponds to a union operation. We can identify the verticies with the roots of the corresponding set's trees and thus getConnectedComponent and componentSeen can be implemented in a straightforward way.

In sum, the bulk of the time complexity is in preprocessing the data. This consists of obtaining the initial k-NN graph, i.e. distances to nearest neighbors; this one time operation is of worst-case order $O(n^2)$, similar to usual clustering procedures (e.g. Mean-Shift, K-Means, Spectral Clustering), but average case $O(nk\log n)$. After this preprocessing step, the estimation procedure itself requires just $O(nk)$ operations, each with amortized $O(\alpha(n))$ where α is the inverse Ackermann function. Thus, the implementation provided in Algorithm 4 is near-linear in k and n.

Algorithm 3 Interface for Mutual k-NN graph construction

```
InitializeGraph()                          // Creates an empty graph
addNode(G, node)                           // Adds a node
addEdge(G, node1, node2)                   // Adds an edge
getConnectedComponent(G, node)             // Get the vertices in node's CC
componentSeen(G, node)  // checks whether node's CC intersects with the estimates. If not, then marks the component as seen.
```

Algorithm 4 Implementation of M-cores (Algorithm 2)

Let kNNSet(x) be the k-nearest neighbors of $x \in X_{[n]}$.
$\widehat{\mathcal{M}} \leftarrow \{\}$
$G \leftarrow$ InitializeGraph()
Sort points in descending order of f_k values
Let $p \leftarrow 1$
for $i = 1, ..., n$ **do**
  $\lambda \leftarrow f_k(X_i)$
  while $p < n$ and $f_k(X_p) \geq \lambda - 9\beta_k\lambda - \epsilon_0 - \tilde{\epsilon}$ **do**
    addNode(G, X_p)
    for $x \in$ kNNSet(X_p) $\cap\, G$ **do**
      addEdge(G, x, X_p)
    end for
    $p \leftarrow p + 1$
  end while
  if not componentSeen(G, X_i) **then**
    toAdd $\leftarrow$ getConnectedComponent(G, X_i)
    Delete all x from toAdd where $f_k(x) < \lambda - \beta_k\lambda$
    $\widehat{\mathcal{M}} \leftarrow \widehat{\mathcal{M}} + \{\text{toAdd}\}$
  end if
end for
return $\widehat{\mathcal{M}}$

I Additional Simulations

I.1 Modal-sets

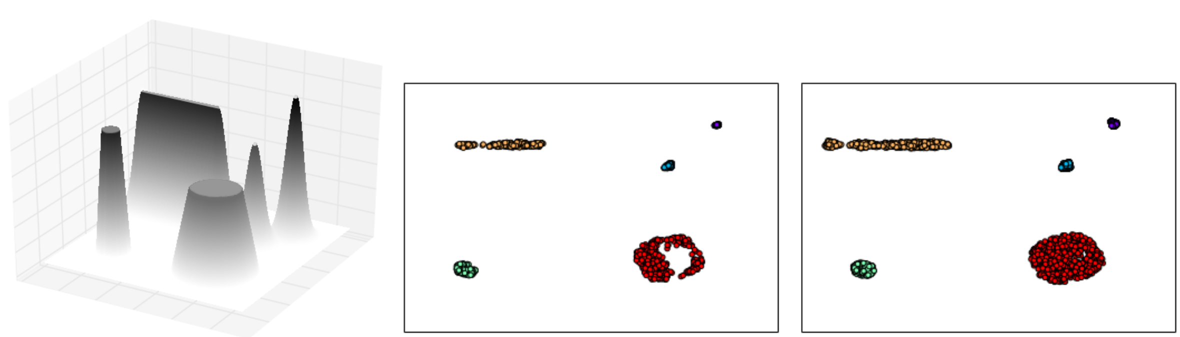

Figure 5: Density (left image) with modal-sets of varying dimension, shape, and level. Middle image displays the estimated modal-sets, and the right image displays the estimated modal-sets when $\beta_k = 3/\sqrt{k}$ (instead of the default value $\beta_k = 2/\sqrt{k}$). Parameter settings: $n = 20000, k = 200$.

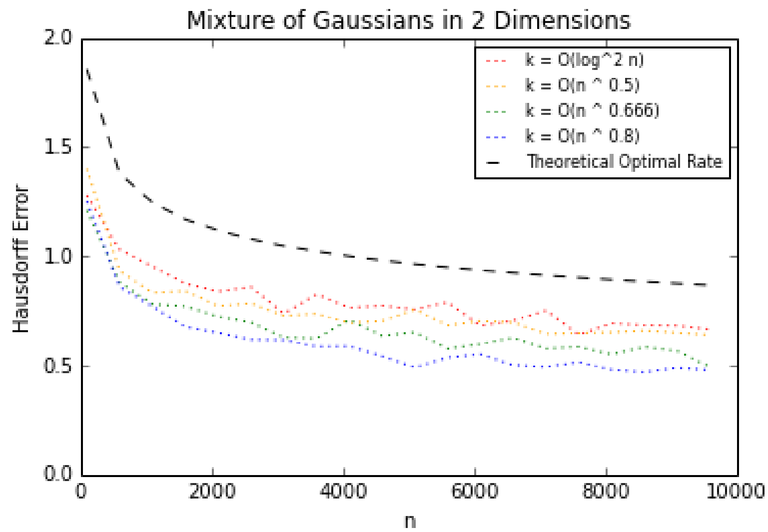

Figure 6: Simulation on a simple mixture of two gaussians. Shown are the performances of MCores estimating two point modes under different settings of k as a function of n. Plotted are the theoretical minimum and maximum settings for k ($k = \log^2 n$ and $k = n^{2/3}$, respectively), as well as in a setting in between ($k = 0.5 \cdot \sqrt{n}$), and a setting faster ($k = 0.54 \cdot n^{0.8}$). The constants were chosen so that the values for k were identitical at the starting setting of n. Also shown is the theoretically optimal estimation rate $n^{-1/6}$; the actual value is unknown since the constant factors are unknown, but it serves to illustrate the expected rate of decay. Each datapoint is the average of 10 different random samplings from the distribution.

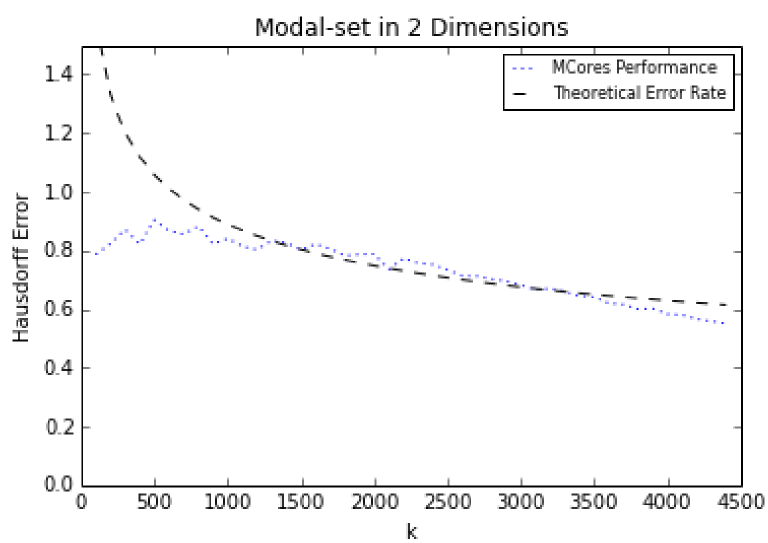

Figure 7: Simulation on a density which is generated by sampling a point on a one-dimensional curve (an arc that is 1/6 of a unit circle) and adding orthogonal Gaussian noise. Shown are the performances of MCores estimating the modal-set (one-dimensional curve) under different settings of k with fixed $n = 5000$. Also shown is the theoretical error bound $k^{-1/4}$; again it is only for purposes of displaying expected rates of decay. Each datapoint is the average of 10 different random samplings from the distribution.

I.2 Clustering

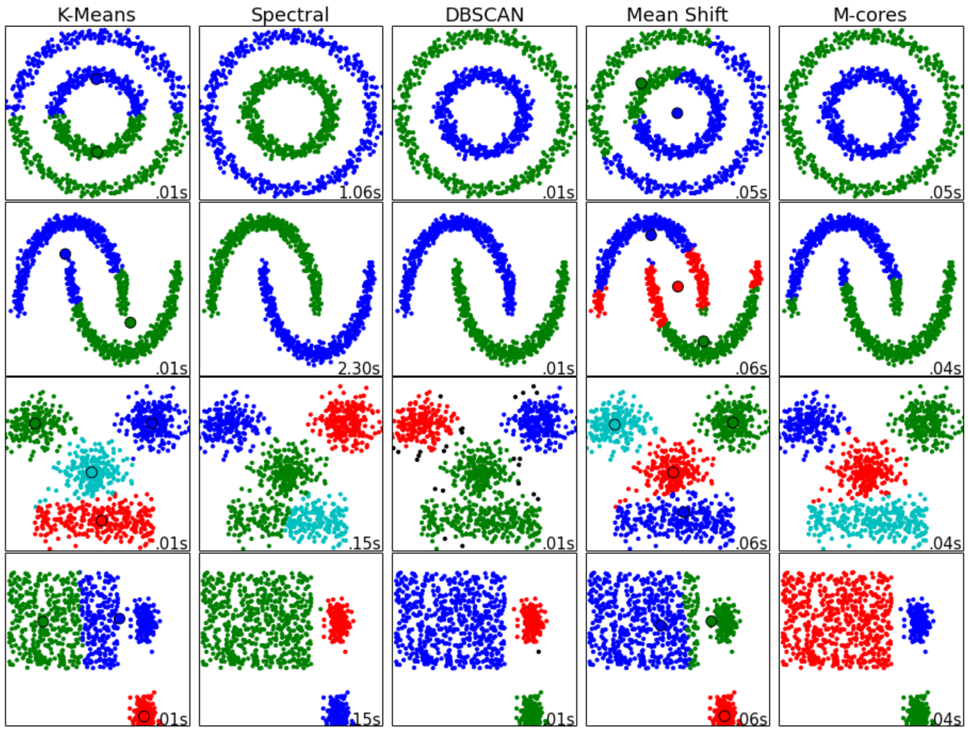

Figure 8: Toy cluster-shapes handled by various procedures. The means or modes (used as cluster centers) by K-Means or Mean Shift are shown as larger dotted points. Parameter settings: $n = 1000, k = 50$. We compare M-cores to state-of-the-art procedures, under *default settings*, on standard toy cluster shapes. This is motivated by common practice where clustering procedures are used as *black-box* routines, i.e. under default settings (implementing various rule of thumbs) provided by software packages; this is because it is difficult to properly tune clustering procedures, given the absence of labels. The data and procedures are from the popular *sci-kit-learn* software package. M-cores uses fixed settings as described earlier. For the other procedures, we employed the automatic tuning used by the *sci-kit-learn* package; K-Means and Spectral Clustering are additionally given the *correct* number of clusters. Results are shown in Figure 8 along with running times. M-cores compares well across the board, with relatively good time efficiency (all the procedures are in Python and C++).