Quickshift++: Provably Good Initializations for Sample-Based Mean Shift

Heinrich Jiang [1] **Jennifer Jang** [2] **Samory Kpotufe** [3]

Abstract

We provide initial seedings to the Quick Shift clustering algorithm, which approximate the locally high-density regions of the data. Such seedings act as more *stable* and *expressive* cluster-cores than the singleton modes found by Quick Shift. We establish statistical consistency guarantees for this modification. We then show strong clustering performance on real datasets as well as promising applications to image segmentation.

1. Introduction

Quick Shift (Vedaldi & Soatto, 2008) is a mode-seeking based clustering algorithm that has a growing popularity in computer vision. It proceeds by repeatedly moving each sample to its closest sample point that has higher empirical density if one exists within a τ-radius ball, otherwise we stop. Thus each path ends at a point which can be viewed as a local mode of the empirical density. Then, points that end up at the same mode are assigned to the same cluster. The most popular choice of empirical density function is the Kernel Density Estimator (KDE) with Gaussian Kernel. The algorithm also appears in Rodriguez & Laio (2014).

Quick Shift was designed as a faster alternative to the well-known Mean Shift algorithm (Cheng, 1995; Comaniciu & Meer, 2002). Mean Shift is equivalent to performing a gradient ascent of the KDE starting at each sample until convergence (Arias-Castro et al., 2016). Samples that correspond to the same points of convergence are in the same cluster and the points of convergence are taken to be the estimates of the modes. Thus, both procedures hill-climb to the local modes of the empirical density function and cluster based on these modes. The key differences are that Quick Shift restricts the steps to sample points (and thus is a sample-based version of Mean Shift) and has the extra τ parameter which allows it to merge close segments together.

[1]Google Research, Mountain View, CA [2]Uber Inc, San Francisco, CA [3]Princeton University, Princeton, NJ. Correspondence to: Heinrich Jiang <heinrich.jiang@gmail.com>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

One of the drawbacks of these two procedures, as well as many mode-seeking based clustering algorithms, is that the point-modes of the density functions are often poor representations of the clusters. This will happen when the high-density regions within a cluster are of arbitrary shape and have some variations causing the underlying density function to have possibly many apparent, but not so salient modes. In this case, such procedures asymptotically recover all of the modes separately, leading to over-segmentation. To combat this effect, practitioners often increase the kernel bandwidth, which makes the density estimate more smooth. However, this can cause the density estimate to deviate too far from the original density we are intending to cluster based on.[1] Thus, practitioners may not wish to identify the clusters based on the point-modes of the density function, but rather identify them based on *locally high density regions* of the dataset (See Figure 1).[2]

We propose modeling these locally high-density regions as *cluster-cores* (to be precisely defined later), which can be of arbitrary shape, size, and density level, and are thus better suited at capturing the possibly complex topological properties of clusters that can arise in practice. In other words, these cluster-cores are better at expressing the clusters and are more stable as they are less sensitive to the small fluctuations that can arise in the empirical density function. We parameterize the cluster-core by β where $0 < \beta < 1$, which determines how much the density is allowed to vary within the cluster-core. We estimate them from a finite sample using a minor modification of the MCores algorithm of Jiang & Kpotufe (2017).

We introduce Quickshift++, which first estimates these cluster-cores, and then runs the Quick Shift based hill-climbing procedure on each remaining sample until it reaches a cluster-core. Samples that end up in the same cluster-core are assigned to the same cluster; thus,

[1]KDE with Gaussian kernel and bandwidth h approximates the underlying density convolved with a Gaussian with mean 0 and covariance $h^2\mathbf{I}$. Thus, the higher h is, the more the KDE deviates from the original density.

[2]Over-segmentation is also dealt with in Quick Shift via the τ parameter, but a threshold for the distance between two modes which should be clustered together is hard to determine in practice. Moreover, there may not even be a good setting of τ which works everywhere in the input space.

the cluster-cores can be seen as representing the high-confidence regions within each cluster. We utilize the k-NN density estimator as our empirical density.

Despite the simplicity of our approach, we show that Quickshift++ considerably outperforms the popular density-based clustering algorithms, while being efficient. Another desirable property of Quickshift++ is that it is simple to tune its two hyperparameters β and k.[3] We show that a few settings of β turn out to work for a wide range of applications and that the procedure is stable in choices of k.

We then give a novel statistical consistency analysis for Quickshift++ which provides guarantees that points within a cluster-core's attraction regions (to be described later) are correctly assigned. We also show promising results on image segmentation, which further validates the desirability of using cluster-cores on real-data applications.

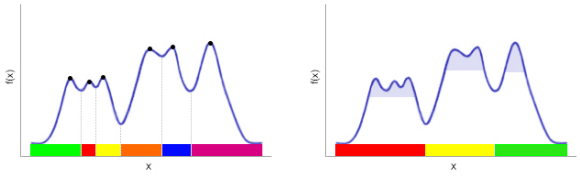

Figure 1. It can often be the case that the locally high-density regions are of arbitrary shape and fluctuations within them lead to many apparent modes. **Left**: Mode-seeking clustering procedures often lead to over-segmentation. **Right**: It may be more desirable to use cluster-cores (shaded), which allows fluctuations within arbitrarily-shaped regions of locally high density.

2. Related Works and Contributions

We show that Quickshift++ is a new and powerful addition to the family of clustering procedures known as *density-based* clustering, which most notably includes DBSCAN (Ester et al., 1996) and Mean Shift (Cheng, 1995). Such procedures operate on the estimated density function based on a finite sample to recover structures in the density function that ultimately correspond to the clusters. There are several advantages of density-based clustering over classical objective-based procedures such as k-means and spectral clustering. Density-based procedures can *automatically* detect the number of clusters, while objective-based procedures typically require this as an input. Density-based clustering algorithms also make little assumptions on the shapes of the clusters as well as their relative positions.

Density-based clustering procedures can roughly be classified into two categories: hill-climbing based approaches (discussed earlier, which includes both Mean Shift and Quick Shift) and density-level set based approaches. We now discuss the latter approach, which takes the connected components of the density-level set defined by $\{x : f(x) \geq \lambda\}$ for some density level λ as the clusters. This statistical notion of clustering traces back to Hartigan (1975). Since then, there has been extensive work done, e.g. Tsybakov et al. (1997); Cadre (2006); Rigollet et al. (2009); Singh et al. (2009); Chaudhuri & Dasgupta (2010); Rinaldo & Wasserman (2010); Kpotufe & von Luxburg (2011); Balakrishnan et al. (2013); Chaudhuri et al. (2014); Chen et al. (2017). More recently, Sriperumbudur & Steinwart (2012); Jiang (2017a) show that the popular DBSCAN algorithm turns out to converge to these clusters. However, one of the main drawbacks of this approach is that the density-level λ is fixed and thus such methods perform poorly when the clusters are at different density-levels. Moreover, the question of how to choose λ remains (e.g. Steinwart (2011)).

Jiang & Kpotufe (2017) provide an alternative notion of clusters, called modal-sets, which are regions of flat density which are local maximas of the density. They can be of arbitrary shape, dimension, or density. They provide a procedure, MCores, which estimates these with consistency guarantees. Our notion of cluster-core is similar to modal-sets, but the density within a cluster-core is allowed to vary by a substantial amount in order to capture such variations seen in real data as a the flat density of modal-sets may be too restrictive in practice. It turns out that a small modification of MCores allows us to estimate these cluster-cores. Thus Quickshift++ has the advantage over DBSCAN in that clusters can be at any density level and that furthermore, the density levels are chosen adaptively.

Mcores however consists of an over-simplistic final clustering: it simply assigns each point to its closest modal-set, while in practice, clusters tend not to follow the geometry induced by the Euclidean metric. Quickshift++ on the other hand clusters the remaining points by a hill-climbing method which we show is far better in practice.

Thus, Quickshift++ combines the strengths of both density-based clustering approaches while avoiding many of their weaknesses. In addition to the general advantages of density-based clustering algorithms shared by both approaches, it is also able to both (1) recover clusters at varying density levels and (2) not suffer from the over-segmentation issue described in Figure 1. To our knowledge, no other procedure has been shown to have this property.

For our theoretical analysis, we give guarantees about Quickshift++'s ability to recover the clusters based on attraction regions defined by the gradient flows. Wasserman et al. (2014); Arias-Castro et al. (2016) showed that Mean Shift's iterates approximate the gradient flows. Some progress has been made in understanding Quick Shift (Jiang, 2017b; Verdinelli & Wasserman, 2018). There are also related lines

[3] The τ parameter from Quick Shift is unnecessary here because we climb until we reach a cluster-core as our stopping condition.

of work in mode clustering e.g. (Li et al., 2007; Chacón, 2012; Genovese et al., 2016; Chen et al., 2016). In this paper, we show that Quickshift++ recovers the interior of its attraction region, thus adding to our statistical understanding of hill-climbing based clustering procedures.

3. Algorithm

3.1. Basic Definitions

Let $X_{[n]} = \{x_1, ..., x_n\}$ be n i.i.d. samples drawn from an unknown density f, defined over the Lebesgue measure on $\mathbb{R}^d$. Suppose that f has compact support $\mathcal{X}$.

Our procedure will operate on the k-NN density estimator:

Definition 1. *Let $r_k(x) := \inf\{r > 0 : |B(x, r) \cap X_{[n]}| \geq k\}$, i.e., the distance from x to its k-th nearest neighbor. Define the k-NN density estimator 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 $\mathbb{R}^d$.

3.2. Cluster-Cores

We define the cluster core with respect to fixed fluctuation parameter β as follows.

Definition 2. *Let $0 < \beta < 1$. Closed and connected set $M \subset \mathcal{X}$ is a cluster-core if M is a connected component (CC) of $\{x \in \mathcal{X} : f(x) \geq (1 - \beta) \cdot \max_{x' \in M} f(x')\}$.*

Note that when $\beta \to 0$, then the cluster-cores become the modes or local-maximas of f. When $\beta \to 1$, then the cluster-core becomes the entire support $\mathcal{X}$. We next give a very basic fact about cluster-cores, that they do not overlap.

Lemma 1. *Suppose that M_1, M_2 are distinct cluster-cores of f. Then $M_1 \cap M_2 = \emptyset$.*

Proof. Suppose otherwise. We have that M_1 and M_2 are CCs of $\{x \in \mathcal{X} : f(x) \geq \lambda_1\}$ and $\{x \in \mathcal{X} : f(x) \geq \lambda_2\}$, respectively for some λ_1, λ_2. Clearly, if $\lambda_1 = \lambda_2$, then it follows that $M_1 = M_2$. Then, without loss of generality, let $\lambda_1 < \lambda_2$. Then since the CCs of $\{x \in \mathcal{X} : f(x) \geq \lambda_2\}$ are nested in the CCs of $\{x \in \mathcal{X} : f(x) \geq \lambda_1\}$, then it follows that $M_2 \subseteq M_1$. Then, $\lambda_2 = (1 - \beta) \sup_{x \in M_2} f(x) \leq (1 - \beta) \sup_{x \in M_1} f(x) = \lambda_1$, a contradiction. As desired. □

Algorithm 1 is a simple modification of MCores by Jiang & Kpotufe (2017). The difference is that we use a multiplicative fluctuation parameter β, while Jiang & Kpotufe (2017) uses an additive one. The latter requires knowledge of the scale of the density function, which is difficult to determine in practice. Moreover, the multiplicative fluctuation adapts to clusters at varying density levels more reasonably than a fixed additive fluctuation. It uses the levels of the mutual k-NN graph of the sample points, defined below.

Definition 3. *Let $G(\lambda)$ denote the λ-level of the mutual k-NN graph with vertices $\{x \in X_{[n]} : f_k(x) \geq \lambda\}$ and an edge between x and x' iff $||x - x'|| \leq \min\{r_k(x), r_k(x')\}$.*

It is known that $G(\lambda)$ approximates the CCs of the λ-level sets of the true density, defined as $\{x \in \mathcal{X} : f(x) \geq \lambda\}$ see e.g. (Chaudhuri & Dasgupta, 2010). Moreover, it can be seen that the CCs of $G(\lambda)$ forms a hierarchical nesting structure as λ decreases.

Algorithm 1 proceeds by performing a top-down sweep of the levels of the mutual k-NN graph, $G(\lambda)$. As λ decreases, it is clear that more nodes appear and that connectivity increases. In other words, as we scan top-down, the CCs of $G(\lambda)$ become larger, some CCs can merge, or new CCs can appear. When a new CC appears at level λ, then intuitively, it should correspond to a local maxima of f, which appears at a density level approximately λ. This follows from the fact that the CCs of $G(\lambda)$ approximates the CCs of $\{x \in \mathcal{X} : f(x) \geq \lambda\}$. Thus, the idea is that when a new CC appears in $G(\lambda)$, then we can take the corresponding CC in $G(\lambda - \beta\lambda)$ (which is the density level $(1 - \beta)$ times that of the highest point in the CC) to estimate the cluster-core.

Algorithm 1 MCores (estimating cluster-cores)

Parameters k, β
Initialize $\widehat{\mathcal{M}} := \emptyset$.
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 be the CC of $G(\lambda - \beta\lambda)$ containing x_i.
  if A is disjoint from all cluster-cores in $\widehat{\mathcal{M}}$ **then**
    Add A to $\widehat{\mathcal{M}}$.
  end if
end for
return $\widehat{\mathcal{M}}$.

Algorithm 2 Quickshift++

Let $\widehat{\mathcal{M}}$ be the cluster-cores obtained by running Algorithm 1.
Initialize directed graph G with vertices $\{x_1, ..., x_n\}$ and no edges.
for $i = 1$ **to** n **do**
  If x_i is not in any cluster-core, then add to G an edge from x_i to its closest sample $x \in X_{[n]}$ such that $f_k(x) > f_k(x_i)$.
end for
For each cluster-core $M \in \widehat{\mathcal{M}}$, let $\widehat{\mathcal{C}}_M$ be the points $x \in X_{[n]}$ such that the directed path in G starting at x ends in M.
return $\{\widehat{\mathcal{C}}_M : M \in \widehat{\mathcal{M}}\}$.

3.3. Quickshift++

Quickshift++ (Algorithm 2) proceeds by first running Algorithm 1 to obtain the cluster-cores, and then moving each sample point to its nearest neighbor that has higher k-NN density until it reaches some cluster-core. All samples that end up in the same cluster-core after the hill-climbing are assigned to the same cluster. Note that since the highest empirical density sample point is contained in a cluster-core, it follows that each sample point not in a cluster-core will eventually be assigned to a unique cluster-core. Thus, Quickshift++ provides a clustering assignment of *every* sample point.

Remark 1. *Although it seems a similar procedure could have been constructed by using Mean Shift in place of Quick Shift, Mean Shift could have convergence outside of the estimated cluster-cores, while Quick Shift guarantees that each sample outside of a cluster-core get assigned to some cluster-core.*

3.4. Implementation

The implementation details for the MCores modification can be inferred from Jiang & Kpotufe (2017). This step runs in $O(nk \cdot \alpha(n))$ where α is the Inverse Ackermann function (Cormen, 2009), in addition to the time it takes to compute the k-NN sets for the n sample points. To cluster the remaining points, for each sample not in a cluster-core, we must find its nearest sample of higher k-NN density. Although this is worst-case $O(n)$ time for each sample point, fortunately we see that in practice (as long as k is not too small) for the vast majority of cases, the nearest sample with higher density is within the k-nearest neighbor set so it only takes $O(k)$ in most cases. It is an open problem whether there the nearest sample with higher empirical density is often in its k-NN set. Code release is at `https://github.com/google/quickshift`.

4. Theoretical Analysis

For the theoretical analysis, we make first the following regularity assumption, that the density is continuously differentiable and lower bounded on $\mathcal{X}$.

Assumption 1. *f is continuously differentiable on $\mathcal{X}$ and there exists $\lambda_0 > 0$ such that $\inf_{x \in \mathcal{X}} f(x) \geq \lambda_0$.*

Let $M_1, ..., M_C$ be the cluster-cores of f. Then we can define the following notion of attraction region for each cluster-core based on the gradient ascent curve or flow. This is similar to notions of attraction regions for some previous analyses of mode-based clustering, such as Wasserman et al. (2014); Arias-Castro et al. (2016), where the intuition is that attraction regions are defined based by following the direction of the gradient of the underlying density. In our situation, instead of an attraction region defined as all points which flow towards a particular point-mode, the attraction region is defined around a cluster-core.

Definition 4 (Attraction Regions)**.** *Let path $\pi_x : \mathbb{R} \to \mathbb{R}^d$ satisfy $\pi_x(0) = x$, $\pi_x'(t) = \nabla f(\pi_x(t))$. For cluster-core M_i, its attraction region $\mathcal{A}_i$ is the set of points $x \in \mathcal{X}$ that satisfy $\lim_{t \to \infty} \pi_x(t) \in M_i$.*

It is clear that these attraction regions are well-defined. The flow path is well-defined since the density is differentiable and since each cluster-core is defined as a CC of a level set, the density must decay around its boundaries. In other words, once an ascent path reaches a cluster-core, it cannot leave the cluster-core.

However, it is in general not the case that the space can be partitioned into attraction regions. For example, if a flow reaches a saddle point, it will get stuck there and thus any point whose flow ends up at a saddle point will not belong to any attraction region. In this paper, we only give guarantees about the clustering of points which are in an attraction region.

The next regularity assumption we make is that the cluster-cores are on the interior of the attraction region (to avoid situations such as when the cluster-cores intersect with the boundary of the input space).

Assumption 2. *There exists $R_0 > 0$ such that $M_i + B(0, R_0) \subseteq \mathcal{A}_i$ for $i = 1, ..., C$, where $M + B(0, r)$ denotes $\{x : \inf_{y \in M} ||x - y|| \leq r\}$.*

Definition 5 (Level Set)**.** *The λ level set of f is defined as $L_f(\lambda) := \{x \in \mathcal{X} : f(x) \geq \lambda\}$.*

The next assumption says that the level sets are continuous w.r.t. the level in the following sense where we denote the ϵ-interior of A as $A^{\ominus \epsilon} := \{x \in A, \inf_{y \in \partial A} ||x - y|| \geq \epsilon\}$ (∂A is the boundary of A):

Assumption 3 (Uniform Continuity of Level Sets)**.** *For each $\epsilon > 0$, there exists $\delta > 0$ such that for $0 < \lambda \leq \lambda' \leq ||f||_\infty$ with $|\lambda - \lambda'| < \delta$, then $L_f(\lambda)^{\ominus \epsilon} \subseteq L_f(\lambda')$.*

This ensures that there are no approximately flat areas in which the procedure may get stuck at. The assumption is borrowed from (Jiang, 2017b). Finally, we need the following regularity condition which ensures that level sets away from cluster-cores do not get arbitrarily thin. This is adapted from standard analyses of level-set estimation (e.g. Assumption B of Singh et al. (2009)).

Assumption 4. *Let μ denote the Lebesgue measure on $\mathbb{R}^d$. For any $r > 0$, there exists $\sigma > 0$ such that the following holds for any connected component A of any level-set of f which is not contained in M_i for any i: $\mu(B(x, r) \cap A) \geq \sigma$ for all $x \in A$.*

For our consistency results, we prove that Quickshift++ can cluster the sample points in the (R, ρ)-interior of an attrac-

tion region (defined below) for each cluster-core properly where $R, \rho > 0$ are fixed and can be chosen arbitrarily small.

Definition 6 ((R,ρ)-interior of Attraction Regions)**.** *Define the (R,ρ)-interior of $\mathcal{A}_i$, denoted as $\mathcal{A}_i^{(R,\rho)}$, as the set of points $x_0 \in \mathcal{A}_i$ such that each path $\mathcal{P}$ from x_0 to any point $y \in \partial\mathcal{A}_i$ satisfies the following.*

$$\sup_{x\in\mathcal{P}} \inf_{x'\in B(x,R)} f(x') \geq \sup_{x'\in B(y,R)} f(x') + \rho.$$

In other words, points in the interior satisfy the property that any path leaving its attraction region must sufficiently decrease in density at some point. This decrease threshold is parameterized by R and ρ.

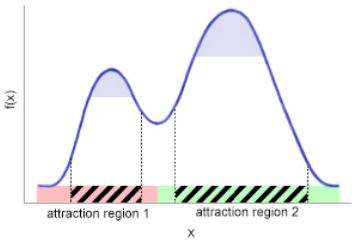

Figure 2. Illustration of interior of attraction region in 1-dimension. The pink and green shaded regions indicated the two attraction regions. The striped parts show the corresponding interiors of the attraction regions.

We first give a guarantee on the first step of MCores recovers, that the cluster-cores are reasonably recovered. The proof follows from the analysis of Jiang & Kpotufe (2017) by replacing modal-sets with cluster-cores, and the results match up to constant factors. The proof is omitted here.

Theorem 1. *[Adapted from Theorem 3, 4 of Jiang & Kpotufe (2017)] Suppose that Assumptions 1, 3, and 4 hold. Let $0 < \beta < 1$, $\epsilon, \delta > 0$ and suppose that $k \equiv k(n)$ is chosen such that $\log^2 n/k \to 0$ and $n^{4/(4+d)}/k \to 0$. Let $M_1, ..., M_C$ be the cluster-cores of f. Then for n sufficiently large depending on f, δ, ϵ, and β, with probability at least $1-\delta$, MCores returns C cluster-core estimates $\widehat{M}_1, ..., \widehat{M}_C$ such that $M_i \cap X_{[n]} \subseteq \widehat{M}_i \subseteq M_i + B(0,\epsilon)$ for $i \in 1, ..., C$.*

Remark 2. *The original result from Jiang & Kpotufe (2017) is about ϵ-approximate modal-set which are defined as level-sets whose density has range ϵ. Our notion of cluster-core is similar, but the range is a β-proportion of the highest density level within the level-set. Using a proportion is more interpretable and thus more useful, as the scale of the density function is difficult to determine in practice.*

In other words, with high probability, MCores estimates each cluster-core bijectively and that for each cluster-core, MCores' estimate contains all of the sample points and that the estimate does not over-estimate by much.

We now state the main result, which says that as long as the cluster-cores are sufficiently well estimated (up to a certain Hausdorff error) by MCores (via previous theorem), then Quickshift++ will correctly cluster the (R,ρ)-interiors of the attraction regions with high probability.

Theorem 2. *Suppose that Assumptions 1, 2, 3, and 4 hold. Let $0 < R < R_0$ and $\rho, \delta > 0$. Suppose that $k \equiv k(n)$ is chosen such that $\log^2 n/k \to 0$ and $n^{4/(4+d)}/k \to 0$. Suppose that $\widehat{M}_1, ..., \widehat{M}_C$ are the cluster-cores returned by Algorithm 1 and satisfy $M_i \cap X_{[n]} \subseteq \widehat{M}_i \subseteq M_i + B(0, R/4)$ for $i = 1, ..., C$. Then for n sufficiently large depending on f, ρ, δ and R, the following holds with probably at least $1 - 2\delta$ uniformly in $x \in \mathcal{A}_i^{(R,\rho)} \cap X_{[n]}$ and $i \in [C]$: Quickshift++ clusters x to the cluster corresponding to M_i.*

4.1. Proof of Theorem 2

We require the following uniform bound on k-NN density estimator, which follows from Dasgupta & Kpotufe (2014).

Lemma 2. *Let $\delta > 0$. Suppose that f is Lipschitz continuous with compact support $\mathcal{X}$ (e.g. there exists L such that $|f(x) - f(x')| \leq L|x - x'|$ for all $x, x' \in \mathcal{X}$) and f satsifies Assumption 1. Then exists constant C depending on f such that the following holds if $n \geq C_{\delta,n}^2$ with probability at least $1 - \delta$.*

$$\sup_{x\in\mathcal{X}} |f_k(x) - f(x)| \leq C\left(\frac{C_{\delta,n}}{\sqrt{k}} + \left(\frac{k}{n}\right)^{1/d}\right).$$

where $C_{\delta,n} := 16\log(2/\delta)\sqrt{d\log n}$.

We next need the following uniform concentration bound on balls intersected with level-sets, which says that if such a set has large enough probability mass, then it will contain a sample point with high probability.

Lemma 3. *Let $\mathcal{E} := \{B(x,r) \cap L_f(\lambda) : x \in \mathbb{R}^d, r > 0, \lambda > 0\}$. Then the following holds with probability at least $1 - \delta$ uniformly for all $E \in \mathcal{E}$*

$$\mathcal{F}(E) \geq C_{\delta,n}\frac{\sqrt{d\log n}}{n} \Rightarrow E \cap X_n \neq \emptyset.$$

Proof. The indicator functions $1[B(x,f) \cap L_f(\lambda)]$ for $x \in \mathbb{R}^d$, $\lambda > 0$ have VC-dimension $d+1$. This is because the balls over $\mathbb{R}^d$ have VC-dimension $d+1$ and the level-sets $L_f(\lambda)$ has VC-dimension 1 and thus their intersection has VC-dimension $d+1$ (Van Der Vaart & Wellner, 2009). The result follows by applying Theorem 15 of Chaudhuri & Dasgupta (2010). □

Proof of Theorem 2. Suppose that $x_0 \in \mathcal{A}_i^{(R,\rho)} \cap X_{[n]}$ and Quickshift++ gives directed path $x_0 \to x_1 \to \cdots \to x_L$ where $x_1, ..., x_{L-1}$ are outside of cluster-cores and x_L is in a cluster-core but $x_L \notin \mathcal{A}_i$.

We first show that $||x_i - x_{i+1}|| \leq R/2$ for $i = 0, ..., L-1$. By Assumption 3 and 4, we have that there exists $\tau > 0$ and $\sigma > 0$ such that the following holds uniformly for $i = 0, ..., L-1$:

$$\mu\Big(B(x_i, R/2) \cap L_f(f(x_i) + \tau)\Big) \geq \sigma.$$

Hence, since the density is uniformly lower bounded by λ_0, we have

$$\mathcal{F}\Big(B(x_i, R/2) \cap L_f(f(x_i) + \tau)\Big) \geq \sigma\lambda_0.$$

Then by Lemma 3, for n suffiicently large such that $\sigma\lambda_0 > C_{\delta,n}\frac{\sqrt{d\log n}}{n}$, then with probability at least $1-\delta$ there exists sample point x_i' in $B(x_i, R/2) \cap L_f(f(x_i) + \tau)$ for $i = 0, ..., L-1$.

Next, choose n sufficiently large such that by Lemma 2, we have with probability at least $1-\delta$ that

$$\sup_{x \in \mathcal{X}} |f_k(x) - f(x)| \leq \min\{\tau, \rho\}/3.$$

Thus, we have

$$\begin{aligned} f_k(x_i') &\geq f(x_i') - \tau/3 \geq f(x_i) + 2\tau/3 \\ &\geq f_k(x_i) + \tau/3 > f_k(x_i). \end{aligned}$$

Moreover $||x_i - x_i'|| \leq R/2$ and $x_i' \in X_{[n]}$, it follows that $||x_i - x_{i+1}|| \leq R/2$ for $i = 0, ..., L-1$.

Let $\pi : [0,1] \to \mathbb{R}^d$ be the piecewise linear path defined by $\pi(j/L) = x_j$ for $j = 0, ..., L$. Let $t_2 = \min\{t \in [0,1] : \pi(t) \in \partial\mathcal{A}_i\}$. Then, by definition of $\mathcal{A}_i^{(R,\rho)}$, there exists $0 \leq t_1 < t_2$ such that $x := \pi(t_1)$ and $y := \pi(t_2)$ satisfies $y \in \partial\mathcal{A}_i$ and

$$\inf_{x' \in B(x,R)} f(x') \geq \sup_{x' \in B(y,R)} f(x') + \rho.$$

Thus, there exists indices $p, q \in \{0, ..., L-1\}$ such that $p \leq q$, $|x_p - x| \leq R$, and $|x_q - y| \leq R$. Thus, we have $f(x_p) \geq f(x_q) + \rho$, but $f_k(x_p) \leq f_k(x_q)$. However, we have

$$\begin{aligned} f_k(x_p) &\geq f(x_p) - \rho/3 \geq f(x_q) + 2\rho/3 \\ &\geq f_k(x_q) + \rho/3 > f_k(x_q), \end{aligned}$$

a contradiction, as desired. □

5. Simulations

Figure 3 provides simple verification that Quickshift++ provides reasonable clusterings in a wide variety of situations where other density-based procedures are known to fail. For instance, in the two rings dataset (first row), we see that Mean Shift and Quick Shift suffer from the over-segmentation issue coupled with the oversized bandwidth which causes them to recover clusters that have points from both the rings even though the rings are separated. In the three Gaussians dataset (third row), we see that DBSCAN fails because the three clusters are of different density levels and thus no matter which density-level we set, DBSCAN will not be able to recover the three clusters.

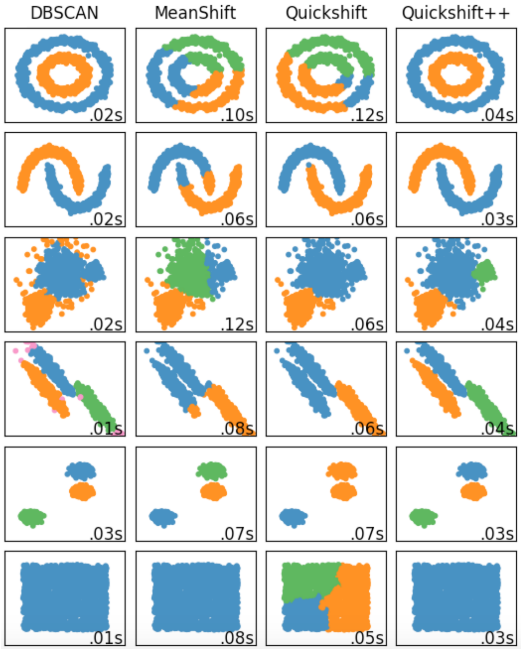

Figure 3. Comparison against other clustering algorithms on toy datasets, adapted from scikit-learn cluster demo. Quickshift++ settings were fixed at $k = 20, \beta = 0.7$ for all the datasets, while the other algorithms were tuned to obtain a reasonable number of clusters.

6. Image Segmentation

In order to apply clustering to image segmentation, we use the following standard approach (see e.g. Felzenszwalb & Huttenlocher (2004)): we transform each pixel into a 5-dimensional vector where two coordinates correspond to the location of the pixel and three correspond to each of the RGB color channels. Then segmentation is done by clustering this 5-dimensional dataset.

We observed that for Quickshift++, setting $\beta = 0.9$ is reasonable across a wide range of images, β was fixed to this value for segmentation here. We compare Quickshift++ to Quick Shift, as the latter is often used for segmentation. Quick Shift often over-segments in some areas and under-segments in other areas under any hyperparameter setting and we showed the settings which provided a reasonable trade-off. On the other hand Quickshift++ gives us reason-

able segmentations in many cases and can capture segments that may be problematic for other procedures.

As shown in the figures, it moreover has the interesting property of being able to recover segments of widely varying shapes and sizes in the same image, which suggests that modelling the dense regions of the segments as cluster-cores instead of point-modes may be useful as we compare to Quick Shift. Although this is only qualitative, it further suggests that Quickshift++ is a versatile algorithm and begins to show its potential application in many more areas.

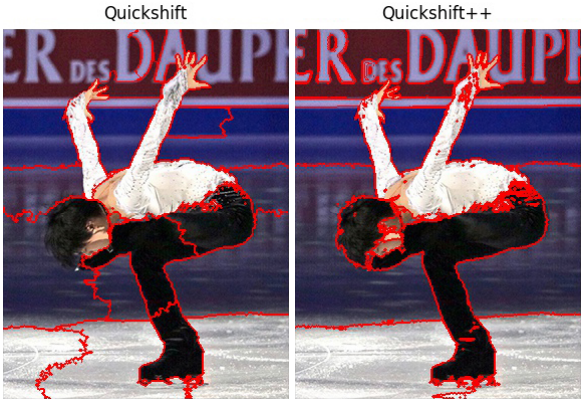

Figure 4. Figure skater Yuzuru Hanyu performs at the 2018 Winter Olympics. Quick Shift was set with bandwidth 10 and Quickshift++ was set with $k = 300$ and $\beta = 0.9$. We see that when compared to Quick Shift, Quickshift++ is able to recover the variations in the background more accurately, including correctly segmenting most of the letters on the wall, while still recovering the structure of Hanyu's costume accurately.

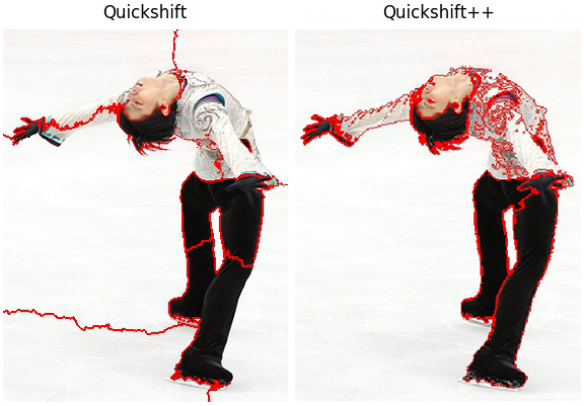

Figure 5. Yuzuru Hanyu at the 2017 Rostelecom Cup. Quick Shift was set with bandwidth 15 and Quickshift++ was set with $k = 50$ and $\beta = 0.9$. Quickshift++ can recover the homogeneous background as a whole, and reasonably separates Hanyu's light-colored costume from the background.

Figure 6. Assorted fruit in a metal bowl. For Quick Shift, bandwidth was set to 8 and for Quickshift++, $k = 100$ and $\beta = 0.9$. Quickshift++ is able to segment most of the fruits in the bowl, while recovering the details of the bowl as well as the structures in the background.

7. Clustering Experiments

We ran Quickshift++ against other clustering algorithms on the various real datasets and scored against the ground-truth using the adjusted rand index and the adjusted mutual information scores.

Dataset	n	d	Clusters
(A) seeds	210	7	4
(B) phonemes	4509	258	5
(C) iris	150	4	3
(D) banknote	1372	4	2
(E) images	210	19	7
(F) letters	20000	16	26
(G) MNIST	1000	784	10
(H) page blocks	5473	10	5
(I) glass	214	19	7

Figure 8. Summary of datasets used, including dataset size (n), number of features (d) and number of clusters.

Datasets Used: Summary of the datasets can be found in Figure 8. Seeds, glass, and iris are standard UCI datasets (Lichman, 2013) used for clustering. Banknote is another UCI dataset which involves identifying whether a banknote is forged or not, based on various statistics of an image of the banknote. Page Blocks is a UCI dataset which involves determining the type of a portion of a page (e.g. text, image, etc) based on various statistics of an image of the portion. Phonemes (Friedman et al., 2001) is a dataset which involves the log periodograms of spoken phonemes. Images is a UCI dataset called Statlog, based on features extracted from various images, and letters is the UCI letter recognition dataset. We also used a small subset of MNIST (LeCun et al., 2010) for our experiments.

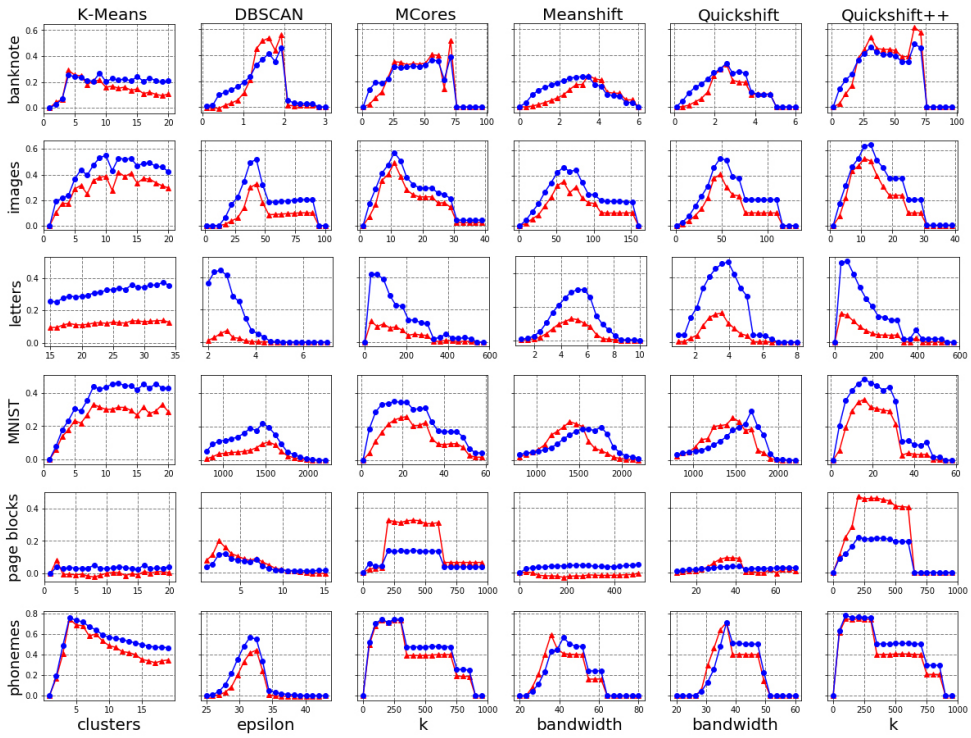

Figure 7. For each algorithm, we show clustering performance as a function of its respective hyperparameter setting. The blue line is adj. mutual information and the red line is adj. rand index. Notice that for Quickshift++, we show a wide range of k (relative to n), while for the popular procedures, their respective parameters had to be carefully tuned to find the region where the scores are non-trivial.

	KMns	DScn	MCrs	MSft	QSft	QS++
A	.7092	.4473	**.7327**	**.7319**	.6715	.7261
	.6738	.4429	**.6872**	.6769	.6360	**.7085**
B	**.7432**	.4458	.7361	.5974	.7165	**.7530**
	.7574	.5731	.7479	.5700	.7149	**.7870**
C	.7294	.5898	**.7261**	.7028	.6203	**.7399**
	.7418	.5865	.7265	.6106	.5836	**.7424**
D	.2893	**.5584**	.5145	.2434	.3318	**.6152**
	.2690	**.4594**	.3857	.2351	.3397	**.4866**
E	.4177	.3313	**.5008**	.3497	.4077	**.5359**
	.5497	.5264	**.5814**	.4656	.5364	**.6456**
F	.1384	.0705	.1284	.1287	**.1793**	**.1766**
	.3741	.4422	.4217	.3027	**.4940**	**.5001**
G	**.3320**	.1070	.2584	.2281	.2503	**.3606**
	.4629	.2164	.3483	.1958	.2911	**.4806**
H	.0830	.1962	**.3251**	.0028	.0925	**.4727**
	.0524	.1179	**.1363**	.0526	.0397	**.2192**
I	.2770	.2844	.2647	.2790	**.2929**	**.2849**
	.3865	.3542	.3523	.3858	**.4195**	**.4250**

Figure 9. For each dataset, the first row is the adjusted rand index scores and the second row is the adjusted mutual information scores. Bolded are **highest** and **second highest** scores. The procedures were tuned in their respective essential hyperparameter: k-means number of clusters, DBSCAN epsilon, MCores k, mean shift bandwidth, quick shift bandwidth, Quickshift++ k.

We evaluate performance under the Adjusted Mutual Information and Rand Index scores (Vinh et al., 2010) which are metrics to compare clusterings. Not only do we show that Quickshift++ considerably outperforms the popular density-based clustering procedures under optimal tuning (Figure 9), but that it is also robust in its hyperparameter k (Figure 7), all while fixing $\beta = 0.3$ for all but one of the datasets. Such robustness to its tuning parameters is highly desirable since optimal tuning is usually not available in practice.

8. Conclusion

We presented Quickshift++, a new density-based clustering procedure that first estimates the cluster-cores of the density, which are locally high-density regions. Then remaining points are assigned to its appropriate cluster-core using a hill-climbing procedure based on Quick Shift. Such cluster-cores turn out to be more stable and expressive representations of the possibly complex clusters than point-modes. As a result, Quickshift++ enjoys the advantages of the popular density-based clustering algorithms while avoiding many of their respective weaknesses. We then gave guarantees for cluster recovery. Finally, we showed that the algorithm has *strong and robust* performance on real datasets and has promising applications to image segmentation.

Acknowledgements

We thank the anonymous reviewers for their helpful feedback.

References

Arias-Castro, E., Mason, D., and Pelletier, B. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *Journal of Machine Learning Research*, 17(43):1–28, 2016.

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.

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

Chacón, J. E. Clusters and water flows: a novel approach to modal clustering through morse theory. *arXiv preprint arXiv:1212.1384*, 2012.

Chaudhuri, K. and Dasgupta, S. Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems*, pp. 343–351, 2010.

Chaudhuri, K., Dasgupta, S., Kpotufe, S., and von Luxburg, U. Consistent procedures for cluster tree estimation and pruning. *IEEE Transactions on Information Theory*, 60 (12):7900–7912, 2014.

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

Chen, Y.-C., Genovese, C. R., and Wasserman, L. Density level sets: Asymptotics, inference, and visualization. *Journal of the American Statistical Association*, pp. 1–13, 2017.

Cheng, Y. Mean shift, mode seeking, and clustering. *IEEE transactions on pattern analysis and machine intelligence*, 17(8):790–799, 1995.

Comaniciu, D. and Meer, P. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619, 2002.

Cormen, T. H. *Introduction to algorithms*. MIT press, 2009.

Dasgupta, S. and Kpotufe, S. Optimal rates for k-nn density and mode estimation. In *Advances in Neural Information Processing Systems*, pp. 2555–2563, 2014.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pp. 226–231, 1996.

Felzenszwalb, P. F. and Huttenlocher, D. P. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2):167–181, 2004.

Friedman, J., Hastie, T., and Tibshirani, R. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. Non-parametric inference for density modes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):99–126, 2016.

Hartigan, J. A. *Clustering algorithms*, volume 209. Wiley New York, 1975.

Jiang, H. Density level set estimation on manifolds with dbscan. *arXiv preprint arXiv:1703.03503*, 2017a.

Jiang, H. On the consistency of quick shift. In *Neural Information Processing Systems (NIPS)*, 2017b.

Jiang, H. and Kpotufe, S. Modal-set estimation with an application to clustering. In *Artificial Intelligence and Statistics*, pp. 1197–1206, 2017.

Kpotufe, S. and von Luxburg, U. Pruning nearest neighbor cluster trees. *arXiv preprint arXiv:1105.0540*, 2011.

LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, 2, 2010.

Li, J., Ray, S., and Lindsay, B. G. A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research*, 8(Aug):1687–1723, 2007.

Lichman, M. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

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

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

Rodriguez, A. and Laio, A. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 2014.

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

Sriperumbudur, B. and Steinwart, I. Consistency and rates for clustering with dbscan. In *Artificial Intelligence and Statistics*, pp. 1090–1098, 2012.

Steinwart, I. Adaptive density level set clustering. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 703–738, 2011.

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

Van Der Vaart, A. and Wellner, J. A. A note on bounds for vc dimensions. *Institute of Mathematical Statistics collections*, 5:103, 2009.

Vedaldi, A. and Soatto, S. Quick shift and kernel methods for mode seeking. *Computer vision–ECCV 2008*, pp. 705–718, 2008.

Verdinelli, I. and Wasserman, L. Analysis of a mode clustering diagram. *arXiv preprint arXiv:1805.04187*, 2018.

Vinh, N. X., Epps, J., and Bailey, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854, 2010.

Wasserman, L., Azizyan, M., and Singh, A. Feature selection for high-dimensional clustering. *arXiv preprint arXiv:1406.2240*, 2014.