Journal of Machine Learning Research 26 (2025) 1-62 Submitted 10/22; Revised 11/23; Published 05/25

Bagged k-Distance for Mode-Based Clustering Using the Probability of Localized Level Sets

Hanyuan Hang HANYUAN0725@GMAIL.COM
Hong Kong Research Institute
Contemporary Amperex Technology (Hong Kong) Limited
Hong Kong Science Park, New Territories, Hong Kong

Editor: Maya Gupta

Abstract

In this paper, we propose an ensemble learning algorithm named *bagged k-distance for mode-based clustering* (*BDMBC*) by putting forward a new measure called the *probability of localized level sets* (*PLLS*), which enables us to find all clusters for varying densities with a global threshold. On the theoretical side, we show that with a properly chosen number of nearest neighbors k_D in the bagged k-distance, the sub-sample size s, the bagging rounds B, and the number of nearest neighbors k_L for the localized level sets, BDMBC can achieve optimal convergence rates for mode estimation. It turns out that with a relatively small B, the sub-sample size s can be much smaller than the number of training data n at each bagging round, and the number of nearest neighbors k_D can be reduced simultaneously. Moreover, we establish fast convergence rates for the level set estimation of the PLLS in terms of Hausdorff distance, which reveals that BDMBC can find localized level sets for varying densities and thus enjoys local adaptivity. On the practical side, we conduct numerical experiments to empirically verify the effectiveness of BDMBC for mode estimation and level set estimation, which demonstrates the promising accuracy and efficiency of our proposed algorithm.

Keywords: Modal clustering, mode-based clustering, probability of localized level sets, k-distance, bagging, convergence rates, ensemble learning, learning theory

1. Introduction

In the field of *density-based clustering*, the common assumption that all clusters have similar levels of densities is shared by many algorithms. In detail, those algorithms employ a global threshold for densities to define the high-density regions and categorize them as clusters. Due to the algorithmic simplicity, such paradigm, also named as *single-level* density-based clustering, attracts lots of attention in the early stage of clustering researches (Ester et al., 1996; Rehioui et al., 2016; Hinneburg and Gabriel, 2007; Idrissi et al., 2015; Jang and Jiang, 2019). However, with the rapid development of information technology, the assumption is hard to hold as the number of clusters and the size of data keeps growing. It has also been proved in experiments that the well-performed single-level clustering algorithms are less effective in encountering datasets that have varying densities for different clusters (Zhu et al., 2021; McInnes et al., 2017). Consequently, a more general setting for density-based clustering called *multi-level density* clustering comes into vogue (Zhu et al., 2021; Mistry et al., 2021; Chazal et al., 2013) and is applied in various subjects including computer vision

(Pla-Sacristán et al., 2019; Cesario et al., 2021; Hu et al., 2022), medicine and biometrics (Liu et al., 2015; Ranjbarzadeh and Saadi, 2020), etc.

To solve the multi-level density clustering problem, a primary idea is to expand the solutions proposed in single-level clustering problems. Some researchers therefore hold the opinion that increasing the number of thresholds can help seek clusters with different densities, and propose the paradigm called *hierarchical density-based clustering*. The term *hierarchical* means that the algorithm follows either an agglomerative (bottom-up) or a divisive (top-down) order to move the threshold, estimates the clusters with each threshold, and finally grows a cluster tree based on the clustering results. A cluster tree is a diagram that organizes data into hierarchical clusters. And by carefully selecting the nodes in the cluster tree, the hierarchical methods can obtain promising results in multi-level density situations (McInnes et al., 2017; Malzer and Baum, 2020; Amini et al., 2016). For example, McInnes et al. (2017) takes the advantage of DBSCAN and proposes an automatic framework called HDBSCAN to decide which thresholds are better according to some pre-determined informational metric; In addition, Malzer and Baum (2020) also studied how to choose the optimal clustering results from a cluster tree. Nevertheless, the hierarchical methods are criticized for the heavy computational cost of growing the cluster tree.

Therefore, to improve the computation efficiency and directly obtain the clustering result, another part of the research aims at finding a suitable transforming for the current density measure to balance the levels of densities for all clusters into a similar level (Cheng, 1995; Jang and Jiang, 2021; Zhu et al., 2021, 2016; Mitra and Nandy, 2011). To be specific, such algorithms care little about the absolute density value of samples. Instead, they attach more importance to the relative information of samples in a local area. For example, Cheng (1995) proposes the mean shift method to find the density hill of clusters by iteratively searching the center of mass from several randomly chosen initial points. And in Zhu et al. (2021), the estimated probability density function is transformed to a new measure called density ratio to help balance the density measure. Since they result in seeking the *bump* or *hill* in the distribution, they are also called mode-based clustering algorithms.

Although mode-based methods largely increase the computational efficiency of multi-level density clustering problems, they still suffer from two inevitable shortcomings. Firstly, many mode-based algorithms require the estimation of the probability density function, e.g. Zhu et al. (2021). Nevertheless, density estimation value can be too small or too large when the dimension of the input variables is high, which means estimating a satisfactory density function will be much harder. Hence, it is hard to perform the mode-based clustering algorithm on high-dimensional datasets. Secondly, the computational efficiency of the mode-based algorithms may not be as satisfactory as expected. On the one hand, mode-based algorithms require much training time when the sample size is large. On the other hand, the procedure of searching an optimal combination of parameters can be even more tiresome.

Under such background, in this paper, we propose an ensemble learning algorithm called *bagged k-distance for mode-based clustering* (*BDMBC*) to solve the multi-level density clustering problems. To be specific, we first introduce a new measure called *probability of localized level sets* (*PLLS*) to deal with the multi-level density problems. PLLS represents the local rank of the density which makes it possible to employ a global threshold to recognize the multi-level density clusters. Secondly, to deal with high-dimensional data, we introduce the *k-distance* as the measurement of density which is then plugged into the localized level set

estimation. This approach allows us to avoid explicitly calculate the k-NN density value during the localized level set estimation process.

Last but not least, we further employ the bagging technique to enhance the computational efficiency in calculating the k-distance. In particular, when dealing with large-scale datasets, the bagging technique can accelerate the algorithm with a small sampling ratio and thus uses a much smaller training dataset in each bagging iteration. Since the size of the training dataset in each iteration is largely decreased by sub-sampling, the searching grid for sample-size-based hyper-parameters can also be simplified, preventing the practitioners from tedious hyper-parameter tuning.

The theoretical and experimental contributions of this paper are summarized as follows:

(i) From the theoretical perspective, we first conduct a learning theory analysis of the bagged k-distance by introducing the *hypothetical density estimation*. Under the Hölder smoothness of the density function, with properly chosen k, we establish optimal convergence rates of the hypothetical density estimation in terms of the L_∞-norm. It is worth pointing out that our finite sample results demonstrate the explicit relationship among bagging rounds B, the number of nearest neighbors k, and the sub-sample size s.

Then we propose a novel mode estimation built from the probability of a localized level set. Based on the convergence rates of the hypothetical density estimation, we show that under mild assumptions on modes, we obtain optimal convergence rates for mode estimation with properly chosen parameters. We show that the bagging technique helps to reduce the subsample size and the number of neighbors simultaneously for mode estimation and thus increases the computational efficiency.

Moreover, under mild assumptions on the density function, we establish convergence results of level set estimation for the probability of localized level set in terms of Hausdorff distance. Compared to previous works on level set estimation in clustering that focus on a single threshold, our results reveal level sets for varying densities. This reveals the local adaptivity of our BDMBC in multi-level density clustering.

(ii) From the experimental perspective, we conduct numerical experiments to illustrate the properties of our proposed BDMBC. Firstly, we verify our theoretical results about mode estimation by conducting the experiment of mode estimation on synthetic datasets. We demonstrate that our BDMBC can detect all modes successfully and thus can cluster all mode-based clusters. Secondly, we verify our theoretical results about level-set estimation by conducting numerical comparisons with other competing methods. We show the promising accuracy and efficiency of our proposed algorithm compared with other density-based, cluster-tree-based, and mode-based methods. Thirdly, we conduct parameter analysis on our proposed BDMBC, and empirically demonstrate that with a relatively small subsample ratio, bagging can significantly narrow the search-grid of parameters. Moreover, we compare the bagging and non-bagging version of the BDMBC on large-scale synthetic datasets and verify that bagging can largely shorten the computation time without sacrificing accuracy.

The remainder of this paper is organized as follows. Section 2 is a warm-up section for the introduction of some notations and the new measure, the *probability of localized level sets* (*PLLS*). Then we propose *the bagged k-distance for mode-based clustering* (*BDMBC*) in Section 2. In Section 3, we first present our main results on the convergence rates for mode estimation and level set estimation. Then we provide some comments and discussions concerning the main results in this section. In Section 4, we conduct the error analysis

for the bagged k-distance and calculate its computational complexity. Section 5 presents experimental results on both real and synthetic data. We also conduct scalability experiments to show the computational efficiency of our algorithm in this section. In Section 7, we demonstrate the details of proofs. Finally, we summarize our work in Section 6.

2. Methodology

In this section, we briefly recall some necessary notations and algorithms as preliminaries in Section 2.1. Then, to avoid the drawback of classical density-based clustering methods, we propose a new measure, the *probability of localized level sets* (*PLLS*) in Section 2.2, introduce the bagged k-distance in Section 2.3, and construct the corresponding density-based clustering algorithm called *bagged k-distance for mode-based clustering* (*BDMBC*) in Section 2.4.

2.1 Preliminaries

First, we introduce some basic notations that will be frequently used in this paper. We use the notation $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$. For any $x \in \mathbb{R}$, let $\lfloor x \rfloor$ denote the largest integer less than or equal to x and $\lceil x \rceil$ the smallest integer greater than or equal to x. Recall that for $1 \leq p < \infty$, the ℓ_p-norm is defined as $\|x\|_p := (x_1^p + \cdots + x_d^p)^{1/p}$, and the ℓ_∞-norm is defined as $\|x\|_\infty := \max_{i=1,\ldots,d} |x_i|$. Let $(\Omega, \mathcal{A}, \mu)$ be a probability space. We denote $L_p(\mu)$ as the space of (equivalence classes of) measurable functions $g : \Omega \to \mathbb{R}$ with finite L_p-norm $\|g\|_p$. For any $x \in \mathbb{R}^d$ and $r > 0$, denote $B(x, r) := \{x' \in \mathbb{R}^d : \|x' - x\|_2 \leq r\}$ as the closed ball centered at x with radius r. For a set $A \subset \mathbb{R}^d$, the cardinality of A is denoted by $\#(A)$ and the indicator function on A is denoted by $\mathbf{1}_A$ or $\mathbf{1}\{A\}$.

In the sequel, the notations $a_n \lesssim b_n$ and $a_n = \mathcal{O}(b_n)$ denote that there exists some positive constant $c \in (0, 1)$, such that $a_n \leq c b_n$ and $a_n \gtrsim b_n$ denotes that there exists some positive constant $c \in (0, 1)$, such that $a_n \geq c^{-1} b_n$. Moreover, the notation $a_n \asymp b_n$ means that there hold $a_n \lesssim b_n$ and $b_n \lesssim a_n$ simultaneously. Let P be a probability distribution on $\mathbb{R}^d$ with the underlying density f which has a compact support $\mathcal{X} \subset [-R, R]^d$ for some $R > 0$. Suppose that the data $D_n = (X_1, \ldots, X_n) \in \mathcal{X}^n$ is drawn from P in an i.i.d. fashion. With a slight abuse of notation, in this paper, c, c', C will be used interchangeably for positive constants while their values may vary across different lemmas, propositions, theorems, and corollaries.

The k-nearest neighbor (k-NN) graph is constructed in the way that two vertices, denoted as p and q, are connected by an edge if the distance between p and q is one of the k smallest distances from p to other objects within the dataset D_n. It can serve as an approximation of the connectivity among the level sets of the underlying probability distribution. By selecting a suitable value for k, the connected components within the k-NN graph align with the actual clusters defined by the level sets.

2.2 Probability of Localized Level Sets

One of the main drawbacks of density-based clustering based on density estimation is that it can not find all clusters with varying densities using a global threshold, see, e.g., Chacón (2015); Zhu et al. (2016); Chacón (2020). Here we give a simple univariate example of

this phenomenon in Figure 1. For the univariate trimodal density, there are three different clusters that are visually identifiable, yet none of the level sets of the density has three connected components. In fact, if the level is chosen too low, the two clusters of high densities will be merged into a single cluster. If the density level is chosen too high, the other cluster exhibiting a lower density will be lost. Clearly, in such case, the clusters derived from a single density level cannot completely describe the inherent clustering structure of the data set.

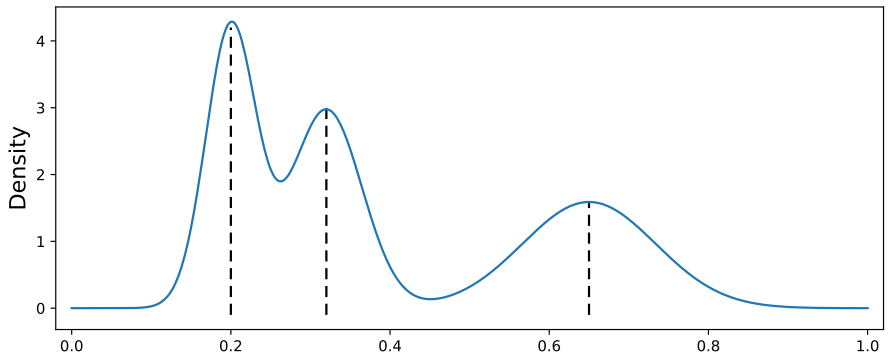


Figure 1: Univariate trimodal density for which it is not possible to capture its whole cluster structure using a global threshold.

To deal with this issue, we propose a local measure named the *probability of localized level sets* (*PLLS*) to implement the density-based clustering.

Definition 1 (Probability of Localized Level Sets) *Let $x \in \mathcal{X}$ and $\eta(x) > 0$ be the local radius parameter. Given the true density function $f : \mathcal{X} \to \mathbb{R}$, the probability of localized level sets (PLLS) is defined by*

$$p_\eta(x) = P\big(f(X) \leq f(x) | X \in B(x, \eta(x))\big) = \frac{P\big(f(X) \leq f(x), X \in B(x, \eta(x))\big)}{P(X \in B(x, \eta(x)))}. \quad (1)$$

Note that the PLLS is the conditional probability of the event where the density of the instance is larger than that of its neighborhood. To explain the advantages of the PLLS over the original probability density function for clustering, we point out two critical observations from (1). On the one hand, if x is a mode of f, then $f(y) \leq f(x)$ for all $y \in B(x, \eta(x))$. This yields that $p_\eta(x) = 1$. On the other hand, if $f(x)$ is a local minimum of the density, i.e., if $f(y) \geq f(x)$ for all $y \in B(x, \eta(x))$, then we have $p_\eta(x) = 0$. Therefore, the PLLS figures out the relative positions to the modes of the density f unlike the probability density function. As a result, we can deal with the variation in density across different clusters and thus allow for a single threshold to identify all the modes and the corresponding clusters simultaneously.

2.3 Bagged k-Distance

In this section, we introduce the bagged k-distance, which represents the density implicitly, for the construction of mode-based clustering. For any $x \in \mathbb{R}^d$, any subset $D \subset D_n$ and

$1 \leq k \leq |D|$, we denote $X_{(k)}(x) := X_{(k)}(x; D)$ as the k-th nearest neighbor of x in D. Then we denote $R_k(x; D)$ as the distance between x and $X_{(k)}(x; D)$, termed as the k-distance of x in D. Specifically, we let $R_k(x) := R_k(x; D_n)$.

We first recall k-nearest neighbor (k-NN) for density estimation. To be specific, denote $\mu(B(x, r))$ as the volume (described in the Lebesgue measure) of the ball $B(x, r)$. Then the k-NN density estimator (Biau and Devroye, 2015, Definition 3.1) is defined by

$$f_k(x) = \frac{k/n}{\mu(B(x, R_k(x)))} = \frac{k/n}{V_d R_k(x)^d}, \tag{2}$$

where $V_d := \pi^{d/2}/\Gamma(d/2 + 1)$ is the volume of the unit ball.

However, in practice, a major problem is the numerical issues when computing k-NN density estimation for high-dimensional data where samples in a finite dataset can distribute quite sparsely. As a consequence, the target density can be extremely small in areas of the input space. In this case, $R_k(x)^d$ in (2) can be extremely large when d is large, leading to an extremely small density estimate value. Therefore, we tend to avoid calculating k-NN density value and use the k-distance to implicitly represent the density. On the other hand, for large-scale datasets, the computational burden of searching for k-nearest neighbors can be heavy. To deal with this problem, in this work, we adopt the bagging technique to reduce the number of nearest neighbors to search, and investigate a bagged variant of k-distance, called *bagged k-distance*. For a fixed subsampling size $s \leq n$, we randomly draw s samples from the dataset D_n without replacement. These s samples form a new subset, denoted as D_1. We repeat this process for B times and then we get B subsets $D_1, D_2, \ldots, D_B$. Given a fixed $k \leq s$, we define the bagged k-distance as

$$R_k^B(x) = \frac{1}{B} \sum_{b=1}^{B} R_k(x; D_b). \tag{3}$$

For the following theoretical analysis, we show that the bagged k-distance can be used to construct a *hypothetical density estimator*,

$$f_B(x) := \frac{\left(\sum_{i=1}^{n} p_i (i/n)^{1/d}\right)^d}{V_d R_k^B(x)^d} \tag{4}$$

with the weights

$$\begin{aligned} p_i &:= P(X_{(i)}(x) \text{ is the } k\text{-th nearest neighbor of } x \text{ in } D_b) \\ &= \begin{cases} \binom{i-1}{k-1}\binom{n-i}{s-k}/\binom{n}{s}, & \text{if } k \leq i \leq n-s+k \\ 0, & \text{if } i \leq k \text{ or } i > n-s+k, \end{cases} \end{aligned} \tag{5}$$

where s denotes the subsample size of bagging. In the following, we briefly explain why the probability p_i in (5) is the ratio of two combinatorial numbers. To be specific, the denominator of p_i is the number of all possible cases of drawing s samples from n samples, $\binom{n}{s}$. The nominator of p_i should be the number of subsampling cases corresponding to the event that the i-th nearest neighbor of x in D is the k-th nearest neighbor of x in the subset D_b, i.e. $X_{(i)}(x) = X_{(k)}(x; D_b)$. The event implies that among the first $(i-1)$-th nearest

neighbor points of x in D, $(k-1)$ points are drawn. In the meantime, among the $(n-i)$-th farthest neighbor points of x in D, $(s-k)$ points are drawn. Therefore, the corresponding number is the combinatorial number $\binom{i-1}{k-1}\binom{n-i}{s-k}$.

The terminology *hypothetical* derives from the observation that it can be quite complicated to calculate the p_i's in practice due to the large amount of calculations of combinations for large sample size n. That is, rather than for practical use, the hypothetical density estimator is only for understanding the bagged k-distance and thus the theoretical analysis.

The above definition acts as a bridge between bagged k-distance and hypothetical density estimator (4), where $f_B(x)$ is proportional to $R_k^B(x)^{-d}$. We show that $f_B(x)$ has the same relative magnitude as $R_k^B(x)^{-1}$, that is, for a given x, larger bagged k-distance $R_k^B(x)$ indicates smaller hypothetical density estimation $f_B(x)$. We delay the discussions that $f_B(x)$ is indeed a desired estimator of the underlying density function f to Proposition 8 in Section 4.1.4.

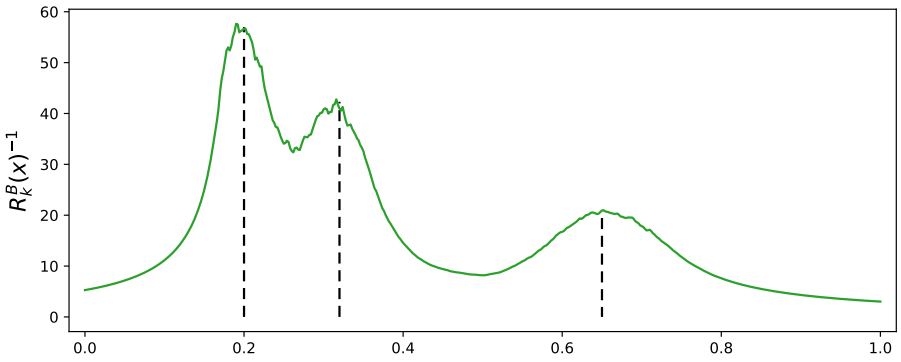


Figure 2: Examples of the inverse of bagged k-distance $R_k^B(x)^{-1}$ with $B = 25$, $s = 0.9n$, $k_D = 300$. Three vertical dash lines present modes of these three Gaussian distributions. The sample size $n = 2000$.

In Figure 2, we empirically illustrate the relationship between the inverse of bagged k-distance $R_k^B(x)^{-1}$ and the true density $f(x)$. Here, we use a one-dimensional synthetic dataset named `3Mix` containing three Gaussian distributions $\mathcal{N}(0.20, 0.001)$, $\mathcal{N}(0.32, 0.002)$, and $\mathcal{N}(0.65, 0.007)$ with equal mixture component weights. Figure 2 illustrates a one-dimensional example of the bagged k-distance. The true density which is a mixture of three Gaussian distributions is provided in Figure 1. With $d = 1$, we show that the inverse of bagged k-distance $R_k^B(x)^{-1}$ in Figure 2 is proportional to the underlying density $f(x)$ in Figure 1.

2.4 Bagged k-Distance for Mode-Based Clustering

The reformulation in (1) inspires us to empirically estimate both the numerator and denominator term of $p_\eta(x)$ respectively. More specifically, let $k_L \in \mathbb{N}$ be the number of nearest neighbors for localized level sets and $\eta(x) := R_{k_L}(x)$, then $p_\eta(x)$ can be estimated by

$$\widehat{p}_{k_L}(x) = \frac{\sum_{i=1}^n \mathbf{1}\{X_i \in B(x, R_{k_L}(x)), f(X_i) \le f(x)\}}{\sum_{i=1}^n \mathbf{1}\{X_i \in B(x, R_{k_L}(x))\}} = \frac{1}{k_L}\sum_{i=1}^{k_L} \mathbf{1}\{f(X_{(i)}(x)) \le f(x)\}. \quad (6)$$

To derive a computationally efficient estimator for PLLS, we use the bagged k-distance in Section 2.3 to define the *empirical PLLS* by

$$\widehat{p}^{B}_{k_L}(x) := \frac{1}{k_L}\sum_{i=1}^{k_L} \mathbf{1}\{R^{B}_{k_D}(X_{(i)}(x)) \geq R^{B}_{k_D}(x)\}. \tag{7}$$

In fact, $\widehat{p}^{B}_{k_L}(x)$ denotes the proportion of instances in the k_L nearest neighbors whose density estimates are smaller than that of x.

With respect to the bagged k-distance plotted in Figure 2, here we plot the empirical PLLS in Figure 3 which shows that PLLS pushes all density peaks towards 1 and forces all density valleys towards 0. This enlarges the difference between peaks and valleys, and therefore it is easier to use a global threshold to separate high-density regions and low-density regions. Moreover, note that three vertical dash lines in Figure 3 present modes of these three Gaussian distributions, the figures show that the PLLS based on bagged k-distance can find out all modes with varied densities at a time. Specifically, in Figure 1, there are three density peaks (modes) which are close to 0.2, 0.32, and 0.65, respectively. Two density valleys are located nearby $x = 0.25$ and $x = 0.50$. By introducing the PLLS, we see from Figure 3 that, the values of density peaks are close to 1, and the values of density valleys are close to 0. The score difference between the density peak in $x = 0.32$ and the density valley in $x = 0.25$ is significantly enlarged. The score difference between the density peak in $x = 0.65$ and the density valley in $x = 0.50$ is also significantly enlarged.

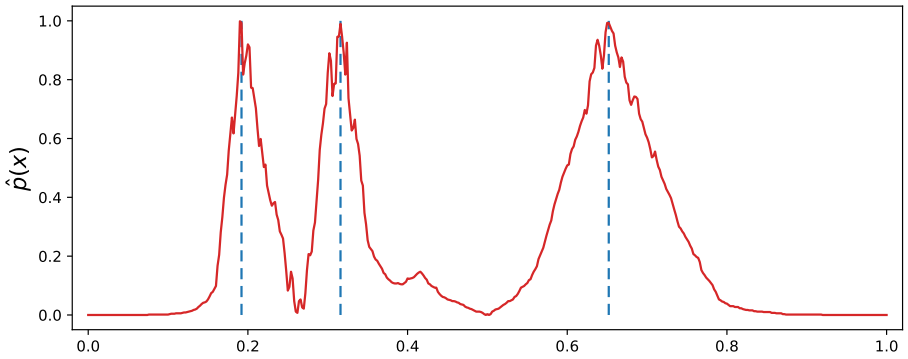


Figure 3: Examples of empirical PLLS with $B = 25$, $s = 0.9n$, $k_D = 300$, and $k_L = 750$. Three vertical dash lines are modes of this distribution, showing that our method can find out all modes with varied densities at a time and make density-based clustering much easier. The sample size $n = 2000$.

Now we can use a global threshold to recover the clusters with the following density-based clustering algorithm named *bagged k-distance for mode-based clustering* (*BDMBC*), which is summarized in Algorithm 1. To obtain clusters, we firstly build a k-NN graph G of all training samples D. Then, we construct the subgraph of the graph G by using all samples whose *probability of localized level sets* (*PLLS*) are in the λ level-set, denoted as $G_B(\lambda)$. Finally, we obtain clusters by finding the connected components of the subgraph $G_B(\lambda)$. By this approach, we can find the regions of *locally* high density based on the upper level set of $\widehat{D}_B(\lambda)$ defined by (8). Then we recover the clusters according to the k-NN graph which utilizes the local density information to connect points. We mention that in Algorithm 1, we only consider the instances with $\widehat{p}^{B}_{k_L}(x) \geq \lambda$ since these instance are regarded as more

important following from the statistically-principled approach in Hartigan (1975). Those instances not in the graph $G_B(\lambda)$ can be assigned to their closest clusters, see e.g. Jang and Jiang (2019).

Algorithm 1: Bagged k-Distance for Mode-Based Clustering (BDMBC)

Input: A dataset $D := D_n := \{X_1, \ldots, X_n\}$;
Bagging size B and subsample size s;
Nearest neighbor k_D for hypothetical density estimation;
Nearest neighbor k_L for localized level set;
Level-set threshold λ;
Nearest neighbor k_G for graph.

1. Subsample s points as $\{D_b\}_{b=1}^B$ from D_n without replacement.
2. Compute the bagged k-distance $R_{k_D}^B$ by (3) based on $(D_b)_{b=1}^B$.
3. Compute the empirical PLLS $\widehat{p}_{k_L}^B(x)$ by (7) for each X_i, $i = 1, \ldots, n$.
4. Construct k_G-nearest neighbor graph G of all training samples D.
5. Construct the subgraph $G_B(\lambda)$ retaining the core-samples

$$\widehat{D}_B(\lambda) = \{X_i \in D : \widehat{p}_{k_L}^B(X_i) \geq \lambda\} \tag{8}$$

and the mode set

$$\widehat{\mathcal{M}} = \{X_i \in D : \widehat{p}_{k_L}^B(X_i) = 1\}. \tag{9}$$

6. Compute the cluster estimators $\mathcal{C}_B(\lambda)$ that is the connected components of $G_B(\lambda)$.

Output: The proper cluster estimator $\mathcal{C}_B(\lambda)$.

From the definition of the empirical PLLS, we mention that it is critical to choose a proper number of nearest neighbors k_L for localized level sets. On the one hand, if k_L is too large, the neighborhood will contain more than one mode and thus can not reflect the local behavior of the densities. On the other hand, if k_L is too small, there will be too few instances in the neighborhood, which leads to an unreliable estimator for clustering.

When we replace the probability density function to the empirical PLLS $\widehat{p}_{k_L}^B(x)$ in (7), the set of modes can be naturally estimated by (9). From (9), we see that the mode set $\widehat{\mathcal{M}}$ picks the point with minimal bagged k_D-distance out of the k_L nearest neighbors. In this case, the difference in densities between mode estimations at dense and sparse regions can be reduced to zero with an appropriate k_L. We mention that our mode estimation is different from gradient-based mode-seeking algorithms in the literature. Examples of such procedures include the mean shift algorithm (Fukunaga and Hostetler, 1975; Comaniciu and Meer, 2002), the modal EM (Li et al., 2007), and the quick shift algorithm (Jiang, 2017a).

Moreover, we highlight the role of mode estimation in the density-based clustering algorithm. As pointed out in Hartigan (1975), in mode-based clustering, clusters are associated to density modes by the statistically-principled approach.

3. Theoretical Results

In this section, we establish theoretical results related to our algorithm BDMBC. As pointed out in Section 2.4, the ability of mode detection plays a fundamental role in density-based clustering, so we begin with the convergence rates of mode estimators based on the probability of localized level set in Section 3.1. More specifically, we present the convergence rates of BDMBC for mode estimation in Section 3.1. Our results reveal the benefits of bagging to reduce the number of nearest neighbors in bagged k-distance at each round. Then we further show the convergence rates of the level set estimation in Section 3.2. Moreover, we show that BDMBC can find all clusters with varying densities using a single threshold. Finally, we compare our studies with other existing ones in the literature in Section 3.3.

We first introduce the general assumptions needed throughout our theoretical analysis. We first make assumptions about the underlying density function in Assumption 1.

Assumption 1 *Assume that P has a Lebesgue density f with the support $\mathcal{X} = [0,1]^d$.*

(i) ***[Boundedness]*** *There exist constants $\underline{c}, \overline{c} > 0$ such that $\underline{c} \leq f \leq \overline{c}$.*

(ii) ***[Smoothness]*** *f is α-Hölder continuous, where $0 < \alpha \leq 1$, i.e., for all $x, x' \in \mathcal{X}$, there exists a constant $c_L > 0$ such that $|f(x) - f(y)| \leq c_L \|x - y\|_2^\alpha$.*

Then we need to make the following assumption on the modes. Before we proceed, we denote the set of modes of f by

$$\mathcal{M} := \{x \in \mathcal{X} : \exists r > 0, \forall x' \in B(x, r), f(x') \leq f(x)\}.$$

Assumption 2 (Twice Differentiability around Modes) *Assume that there exists some $r_{\mathcal{M}} > 0$ such that f is twice continuously differentiable around the disjoint neighborhood $B(m_i, r_{\mathcal{M}})$ of each $m_i \in \mathcal{M}$, $i = 1, \ldots, \#(\mathcal{M})$. Denote the gradient and Hessian of f by ∇f and H, respectively, and assume that $H(x)$ is negative definite at all $x \in \mathcal{M}$.*

The above mode assumption is widely adopted for mode estimation (Dasgupta and Kpotufe, 2014; Jiang et al., 2018; Jang and Jiang, 2021), which requires that the density f near the modes is concave (Saxena et al., 2017; Xu and Lange, 2019). Compared to Assumption 2, this condition requires second-order smoothness of the density function near the modes. Here we exclude modes at the boundary of support of f, where f can not be continuously differentiable. In fact, this problem can be handled under an additional boundary smoothness assumption. This approach only complicates the analysis, while the main insights remain the same for interior modes. We refer the reader to Dasgupta and Kpotufe (2014) for more discussions.

We mention that Assumption 2 holds for a large non-parametric class of functions including *Morse* density functions, which are widely used in the density-based clustering and mode estimation and topological data analysis; see, e.g., Chacón (2015); Arias-Castro et al. (2016); Chacón (2020) and the references therein. A map f is a Morse function if its critical points are non-degenerate, i.e., the Hessian of f at each critical point is non-singular. As pointed out in Matsumoto (2002, Corollary 1.12), critical points of Morse functions are isolated. It thus follows that Morse functions on compact sets have finitely many critical points, which implies that Morse density functions satisfy Assumption 2 if we choose a sufficiently small $r_{\mathcal{M}}$.

3.1 Convergence Rates of BDMBC for Mode Estimation

To derive the convergence rates of our BDMBC for mode estimation, we need the following assumption under which clusters can be separated with respect to distinct modes.

Assumption 3 (Unflatness) *Assume there exist constants $\gamma > 0$, $c_\gamma > 0$, $\epsilon_0 > 0$ such that for all $\theta \in [0, \bar{c}]$ and $\epsilon \in (0, \epsilon_0]$, we have $P(x : |f(x) - \theta| \leq \epsilon) \leq c_\gamma \epsilon^\gamma$.*

Assumption 3 is a well-known condition introduced by Polonik (1995) for the level set estimation problem. Clearly, the larger the γ, the more steeply f must approach λ from above. In fact, Assumption 3 ensures there are no such flat regions where there is no change in density. It is commonly adopted in cluster analysis (Steinwart, 2015; Jiang et al., 2018).

The following theorem presents the convergence rates of the mode recovery based on the PLLS with respect to the Euclidean distance.

Theorem 2 *Let Assumptions 1, 2 and 3 hold with $2\alpha\gamma \leq 4+d$ and $\widehat{\mathcal{M}}$ be the mode estimator as in (9). Then for every mode $m_i \in \mathcal{M}$ and $\lambda \geq c$ with the constant c which will be specified in the proof, by choosing*

$$\begin{gathered} k_{D,n} \asymp \log n, \qquad s_n \asymp n^{\frac{d}{4+d}} (\log n)^{\frac{4}{4+d}}, \qquad B_n \geq n^{\frac{3}{4+d}} (\log n)^{\frac{d+1}{4+d}}, \\ k_{G,n} \asymp \log n, \qquad n^{1-\frac{\alpha\gamma}{4+d}} (\log n)^{1+\frac{\alpha\gamma}{4+d}} \lesssim k_{L,n} \lesssim n, \end{gathered} \tag{10}$$

there exists a mode estimate $\widehat{m}_i$ such that with probability at least $1 - 3/n^2$, there holds

$$\|\widehat{m}_i - m_i\|_2 \lesssim (\log n / n)^{\frac{1}{4+d}}.$$

Moreover, there exist distinct cluster estimators $\widehat{C}_i \in \mathcal{C}_B(\lambda)$, $1 \leq i \leq \#(\mathcal{M})$, such that $\widehat{m}_i \in \widehat{C}_i$.

Existing works such as Dasgupta and Kpotufe (2014) and Jiang (2017a) showed that under Assumption 2 on modes, the mode estimation has the minimax optimal convergence rates $n^{-1/(4+d)}$, up to a logarithmic factor. Therefore, Theorem 2 together with Theorem 9 implies that up to a logarithm factor, the convergence rate of BDMBC turns out to be minimax optimal for mode estimation, if we choose the sub-sample size s, the number of nearest neighbors k_D and k_L, and the bagging rounds B according to (10), respectively. In other words, when the bagging technique is combined with the k-distances for mode estimation, the optimal convergence rate is obtainable. Moreover, if we choose the number of nearest neighbors k_G properly, then we can obtain k different cluster estimators that correspond to the modes.

Notice that for a given dataset, (10) yields that k_D and B is proportional to s and k_D/s, respectively. Therefore, only a few independent bootstrap samples are required to use for the computation of k-distance at each bagging round. As a result, $k_{D,n}$ is reduced to $\mathcal{O}(\log n)$ in (10), instead of $\mathcal{O}(n^{4/(4+d)}(\log n)^{d/(4+d)})$ in the following Theorem 9 in Section 7.1 for DMBC, the special case of BDMBC without bagging, i.e. $B = 1$ and $s = n$.

Finally, we remark that the proof of Theorem 2 indicates that there is no bias-variance trade-off associated with the parameter k_L in the convergence rates. Moreover, the proof shows that when the value of k_L falls within a suitable range as in (10), we can achieve the convergence rates for mode estimation, and these rates remain in the same order and do not depend on the specific value of k_L in the interval.

3.2 Convergence Rates of BDMBC for Level Set Estimation

In this section, we establish convergence rates of level set estimation for the PLLS of our BDMBC algorithm. Before we proceed, we need to introduce the population version of $R_i(x)$, namely $\overline{R}_i(x)$ defined by

$$\overline{R}_i(x) := \inf\{r \geq 0 : P(B(x,r)) \geq i/n\}. \tag{11}$$

For $k_L \in \mathbb{N}$, we define the population version of the probability of the localized level set,

$$p_{k_L}(x) := P(f(y) \leq f(x) | y \in B(x, \overline{R}_{k_L}(x)). \tag{12}$$

where $\overline{R}_{k_L}(x)$ is defined by (11). Compared with the empirical version defined by (6), the local radius function in (12) relies on the population version of the k_L-distance.

Then for $k_L \in \mathbb{N}$ and $\lambda \in [0, \overline{c}]$, we define the level set of $p_{k_L}(x)$ by $L_{k_L}(\lambda) := \{x : p_{k_L}(x) \geq \lambda\}$. Then the level set estimation of our BDMBC is $\widehat{L}_{k_L}(\lambda) := \{x : \widehat{p}^B_{k_L}(x) \geq \lambda\}$ with $\widehat{p}^B_{k_L}(x)$ defined by (7).

To further conduct our analysis, we need the following assumption introduced in Jiang (2017b); Jiang et al. (2019) on the behavior of level set boundaries.

Assumption 4 (β-regularity) *Assume that there exist constants $c_\beta > 0$ and $\epsilon_0 > 0$ such that for all $\lambda > 1/2$ and $k_L \lesssim n$, we have $c_\beta d(x, L_{k_L}(\lambda))^\beta \leq \lambda - p_{k_L}(x)$ for all x satisfying $\{x : |p(x) - \lambda| \leq \epsilon_0\}$, where $d(x, A) := \inf_{y \in A} d(x, y)$.*

The β-regularity in Assumption 4 ensures that there is a sufficient decay around level set boundaries so that the level sets are salient enough to be detected. The next theorem gives the estimation rate in terms of the Hausdorff distance $d_{\text{Haus}}(A, A') = \max\{\sup_{x \in A} d(x, A'), \sup_{x' \in A'} d(x', A)\}$.

Theorem 3 *Let Assumptions 1, 3 and 4 hold with $\gamma > d/(2\alpha + d)$ and $\alpha\gamma \geq \beta$. By choosing*

$$\begin{aligned} k_{D,n} &\asymp \log n, \qquad & s_n &\asymp n^{\frac{d}{2\alpha+d}} (\log n)^{\frac{2\alpha}{2\alpha+d}}, \\ B_n &\geq n^{\frac{1+\alpha}{2\alpha+d}} (\log n)^{\frac{\alpha+d-1}{2\alpha+d}}, \qquad & n^{1-\frac{\alpha\gamma-\beta}{2\alpha+d}} (\log n)^{\frac{\alpha\gamma-\beta}{2\alpha+d}} &\lesssim k_{L,n} \lesssim n, \end{aligned} \tag{13}$$

then with probability P^n at least $1 - 3/n^2$, there holds

$$d_{\text{Haus}}\Big(\widehat{L}_{k_L}(\lambda), L_{k_L}(\lambda)\Big) \lesssim (\log n/n)^{\frac{1}{2\alpha+d}}.$$

Note that the choice of $k_{D,n}$ in Theorem 3 is the same as that in Theorem 2, whereas the choice of s_n and B_n are different. In fact, compared with Theorem 2, we take Hölder smoothness assumptions in Theorem 3 instead of the twice differentiability in Assumption 2, and thus larger subsample size s is required. Moreover, the convergence rate in Theorem 3 turns out to be $\mathcal{O}(n^{-1/(2\alpha+d)})$ up to a logarithm factor, which matches the lower bound established in Tsybakov (1997); Jiang (2017b). Similar to the comments regarding Theorem 2, when k_L falls within an appropriate range as in (13), we can obtain fast convergence rates for PLLS level set estimation. Moreover, the proof of Theorem 3 indicates that there exists no bias-variance trade-off related to k_L values in the convergence rates.

3.3 Comments and Discussions

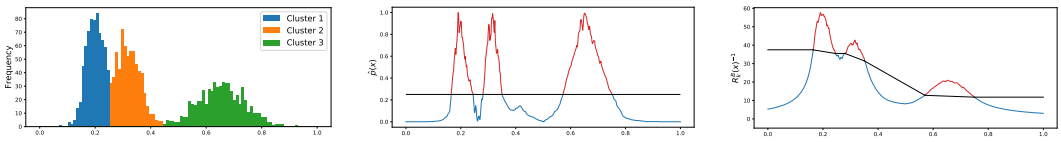
3.3.1 Comments on Convergence Rates for Mode Estimation

Existing modal clustering algorithms use gradient ascent or borrow from work in cluster tree estimation to seek modes. To the best of our knowledge, Dasgupta and Kpotufe (2014) first gives a procedure that recovers multiple modes of a density by using a top-down traversal of the density levels. The best known practical approach for mode estimation is the mean-shift procedure and its variants (Fukunaga and Hostetler, 1975; Li et al., 2007; Chen, 2018; Ghassabeh and Rudzicz, 2018) consisting of gradient ascent of the appropriately smooth density estimator f_D. For the theoretical analysis, Arias-Castro et al. (2016) shows that mean-shift's updates converge to the correct gradient ascent steps. More recently, Jiang (2017a); Jiang et al. (2018) show that Quick Shift and its variants can attain strong statistical guarantees without the second-order density assumption required to analyze mean-shift. However, most of these methods need a proper smooth density estimator as preliminaries. Thus these clustering methods can be very sensitive to user-defined parameters.

3.3.2 Comments on Convergence Rates for Level Set Estimation

We show in Theorem 3 the convergence rate turns out to be $\mathcal{O}(n^{-1/(2\alpha+d)})$, which matches the lower bound established in Tsybakov (1997); Jiang (2017b). The level set estimations by using a single threshold in previous studies are inadequate for multiple modes with varying densities. Considering the density levels of clusters, the explanations are twofold.

1. When the differences in the density levels of clusters are not significant, either the original level set estimation or PLLS level set estimation can distinguish the clusters with a single threshold. Meanwhile, though the core samples returned by the level set algorithms may have slight differences, the convergence rates established for both are nearly equivalent.
2. Secondly, on the condition that the clusters are of varying density levels, as stated in our paper, it is obvious that the original level set estimation with a single threshold is less effective to correctly seek the core samples for the clusters. Hence, one can only obtain the correct core samples corresponding to the clusters by using PLLS level set estimation with a single threshold, whereas the convergence rates w.r.t. the original level set estimation is meaningless.



(a) Ground-True Clusters (b) Empirical PLLS (c) BDMBC Density Estimation

Figure 4: An illustrative example of level set clustering for three density-varying clusters. (a) ground-truth clusters, (b) the curve of the empirical PLLS and its level set estimation, (c) the curve of the hypothetical density estimation and the corresponding original level set estimation.

To better illustrate, we use a one-dimensional synthetic dataset which is visualized in Figure 4(a) to illustrate the necessity of analyzing the PLLS level set estimation rather than the original level set estimation. Figure 4(b) shows the empirical PLLS curve along with the boundary of the PLLS level set estimation in the black horizontal line, whereas Figure 4(c) shows the hypothetical density estimation and the corresponding original level set estimation in black lines. Since PLLS estimation is a non-linear transformation of the hypothetical density estimation, the PLLS level set estimation with a single threshold in Figure 4(b) corresponds to a complex original level set estimation in Figure 4(c). In other words, in the context of multi-level density clusters, we can convert the complex original level set estimation to the PLLS level set estimation with a single threshold. Thus, it is more convenient to use the PLLS level set estimation to obtain density-varying clusters compared to the original level set estimation.

4. Error and Complexity Analysis

In this section, we first conduct error analysis related to the bagged k-distance in Section 4.1. We mention that the theoretical results for mode estimation and level set estimation in Section 3 are all built upon the results for bagged k-distance in this Section. To be specific, in Section 4.1, we first conduct error decomposition for the hypothetical density estimation. Then, in Subsections 4.1.1-4.1.3, we present the upper bounds for the bagging error, estimation error, and approximation error, respectively. With these preparations, we establish in Section 4.1.4 the uniform convergence rates for the hypothetical density estimation under mild smoothness Assumption 1. Moreover, in this Section, we further establish faster convergence rates for the hypothetical density estimation around the modes under mode Assumption 2. Finally, we conduct algorithm complexity analysis in Section 4.2 to demonstrate the efficiency of our algorithm.

4.1 Error Analysis for Bagged k-Distance

The bagged k-distance can not be analyzed directly since it is not of the form of commonly used estimators. According to (4), the problem of analyzing the bagged k-distance can be reduced to the problem of analyzing the hypothetical density estimation. Then we can apply standard techniques to the analysis of $f_B(x)$ and then use it for our bagged k-distance.

Let us turn to the empirical probability of the localized level set defined in (7). By using the hypothetical density estimation (4), $\widehat{p}^B_{k_L}(x)$ can be re-expressed as

$$\widehat{p}^B_{k_L}(x) := \frac{1}{k_L}\sum_{i=1}^{k_L} \mathbf{1}\{f_B(X_{(i)}(x)) \geq f_B(x)\}. \tag{14}$$

Then we conduct the following error decomposition of hypothetical density estimation

$$\begin{aligned}\big|f_B(x) - f(x)\big| &= \left|\frac{\big(\sum_{i=1}^n p_i (i/n)^{1/d}\big)^d}{V_d\big(R^B_k(x)\big)^d} - f(x)\right| \\ &= \left|\frac{\big(\sum_{i=1}^n p_i \big((i/n)/(V_d f(x))\big)^{1/d}\big)^d - \big(R^B_k(x)\big)^d}{\big(R^B_k(x)\big)^d}\right| \cdot f(x)\end{aligned}$$

$$= \left| \frac{\sum_{i=1}^{n} p_i \big((i/n)/(V_d f(x))\big)^{1/d} - R_k^B(x)}{\big(R_k^B(x)\big)^d} \right| \cdot f(x) \cdot$$
$$\cdot \sum_{j=0}^{d-1} \left(\sum_{i=1}^{n} p_i ((i/n)/(V_d f(x)))^{1/d} \right)^j \big(R_k^B(x)\big)^{d-1-j}. \tag{15}$$

Let us consider the first term of the product on the right-hand side of the decomposition above. The numerator term is regarded as the difference between the weighted k-distance $\sum_{i=1}^{n} p_i \big((i/n)/(V_d f(x))\big)^{1/d}$ and the bagged k-distance, while the denominator term is the bagged k-distance $R_k^B(x)$ to the power d.

To conduct theoretical analysis for the bagged k-distance, we need to consider the estimator with infinite bagging rounds, which can be expressed as

$$\widetilde{R}_k^B(x) := \mathbb{E}_{P_Z}^B [R_k^B(x) | \{X_i\}_{i=1}^n], \tag{16}$$

where P_Z denotes the sub-sampling probability distribution.

Note that the bagged k-distance can be re-expressed as a weighted k-distance, which is amenable to statistical analysis. To be specific, let $X_{(i)}(x)$ be the i-th nearest neighbor of x in D_n w.r.t. the Euclidean distance and $R_i(x) = \|x - X_{(i)}(x)\|$. For $1 \le b \le B$, we can re-express the k-distance with respect to the set D_b as

$$R_k(x, D_b) = \sum_{i=1}^{n} p_i^b R_i(x)$$

with $p_i^b := \mathbf{1}\{X_{(i)}(x)$ is the k-th nearest neighbor of x in $D_b\}$. Then the bagged k-distance in (3) can be re-expressed as

$$R_k^B(x) = \frac{1}{B} \sum_{b=1}^{B} \sum_{i=1}^{n} p_i^b R_i(x).$$

Therefore, we have the estimator with infinite bagging rounds

$$\widetilde{R}_k^B(x) = \sum_{i=1}^{n} p_i R_i(x). \tag{17}$$

with p_i defined by (5).

Finally, we are able to make the following error decomposition on the numerator as

$$\left| R_k^B(x) - \sum_{i=1}^{n} p_i ((i/n)/(V_d f(x)))^{1/d} \right|$$
$$\le \big| R_k^B(x) - \widetilde{R}_k^B(x) \big| + \left| \sum_{i=1}^{n} p_i \big(R_i(x) - \overline{R}_i(x)\big) \right| + \left| \sum_{i=1}^{n} p_i \big(\overline{R}_i(x) - ((i/n)/(V_d f(x)))^{1/d}\big) \right|. \tag{18}$$

The three terms on the right-hand side are called *bagging error*, *estimation error*, and *approximation error*, respectively. More specifically, since we are not able to repeat the sampling

strategy an infinite number of times, the bagging procedure brings about the first error term. The second term is called the *estimation error* since it is associated with the empirical measure D_n and the last term is called *approximation error* since it indicates how the error is propagated by the bagged k-distance for hypothetical density estimation. In the next three sections, we will bound these three terms respectively.

4.1.1 Bounding the Bagging Error

The next proposition shows that the bagging error term is determined by the number of bagging rounds B and the ratio k/s.

Proposition 4 *Let Assumption 1 hold. Moreover, let $R_k^B(x)$ and $\widetilde{R}_k^B(x)$ be defined by* (3) *and* (17)*, respectively. Then for all $x \in \mathcal{X}$, with probability $P_Z^B \otimes P^n$ at least $1 - 1/n^2$, there holds*

$$\left|R_k^B(x) - \widetilde{R}_k^B(x)\right| \lesssim \sqrt{(k/s)^{2/d} \log n/B} + \log n/B.$$

4.1.2 Bounding the Estimation Error

We now establish the upper bound of the estimation error of weighted k-distance. This oracle inequality will be crucial in establishing the convergence results of the estimator.

Proposition 5 *Let Assumption 1 hold. Furthermore, let $R_k(x)$ be the k-nearest neighbor distance of x and $\overline{R}_k(x)$ be the quantile diameter function of x defined by* (11)*. Moreover, let p_i be the probability as in* (5)*. Then for all $x \in \mathcal{X}$, with probability P^n at least $1 - 2/n^2$, there holds*

$$\left|\sum_{i=1}^n p_i\big(R_i(x) - \overline{R}_i(x)\big)\right| \lesssim (k/s)^{1/d-1/2}(\log n/n)^{1/2}.$$

4.1.3 Bounding the Approximation Error

The following result on bounding the approximation error term shows that the approximation error can be small by choosing the ratio k/s appropriately.

Proposition 6 *Let Assumption 1 hold. Moreover, let p_i be the probability as in* (5) *and $\overline{R}_i(x)$ be the quantile diameter function of x defined by* (11)*. Then for all $x \in \mathcal{X}$ we have*

$$\left|\sum_{i=1}^n p_i \overline{R}_i(x) - \sum_{i=1}^n p_i((i/n)/(V_d f(x)))^{1/d}\right| \lesssim (k/s)^{(1+\alpha)/d}.$$

4.1.4 Convergence Rates for Hypothetical Density Estimation

The next proposition presents the convergence rates of the hypothetical density estimator induced by the bagged k-distance.

Proposition 7 *Let Assumption 1 hold. Moreover, let $f_B(x)$ be the hypothetical density estimator as in* (4)*. By choosing*

$$k_{D,n} \asymp \log n, \quad s_n \asymp n^{\frac{d}{2\alpha+d}} (\log n)^{\frac{2\alpha}{2\alpha+d}}, \quad B_n \geq n^{\frac{1+\alpha}{2\alpha+d}} (\log n)^{\frac{\alpha+d-1}{2\alpha+d}},$$

then for all $x \in \mathcal{X}$*, with probability* $P^n \otimes P_Z^B$ *at least* $1 - 3/n^2$*, there holds*

$$|f_B(x) - f(x)| \lesssim (\log n/n)^{\frac{\alpha}{2\alpha+d}}.$$

We establish the following finite sample bounds of the hypothetical density estimation near the modes in terms of L_∞-norm.

Proposition 8 *Let Assumptions 1 and 2 hold. Moreover, let* $f_B(x)$ *be the hypothetical density estimator as in* (4)*. By choosing*

$$k_{D,n} \asymp \log n, \qquad s_n \asymp n^{\frac{d}{4+d}} (\log n)^{\frac{4}{4+d}}, \qquad B_n \geq n^{\frac{3}{4+d}} (\log n)^{\frac{d+1}{4+d}},$$

then for all $x \in \mathcal{M}_{r/2}$*, with probability* $P^n \otimes P_Z^B$ *at least* $1 - 3/n^2$*, there holds*

$$|f_B(x) - f(x)| \lesssim (\log n/n)^{\frac{2}{4+d}}.$$

We compare our results with previous theoretical analysis of the k-NN for density estimation. Biau et al. (2011) introduced a weighted version of the k-nearest neighbor density estimate and establish pointwise consistency results. Recently, Zhao and Lai (2020) analyzed the L_α and L_∞ convergence rates of k nearest neighbor density estimation method including two different cases depending on whether the support set is bounded or not. It is worth pointing out that our analysis of the bagged k-distance presents in this study is essentially different from that in the previous works.

First of all, the core challenge in the analysis of bagged k-distance is that it cannot be analyzed using existing techniques for standard k-nearest neighbor methods. To solve this problem, we consider the hypothetical density function in (4). Under the Hölder continuity assumptions, we derive optimal convergence rates of the hypothetical density function with properly selected parameters. Moreover, our results are different from the previous statistical analysis since it is conducted from a learning theory perspective (Cucker and Zhou, 2007; Steinwart and Christmann, 2008) using techniques such as approximation theory and empirical process theory (van der Vaart and Wellner, 1996; Kosorok, 2008). By exploiting arguments such as Bernstein's concentration inequality from the empirical process theory, we can derive the relationships among the number of bagging rounds B, the number of nearest neighbors k_D and the sub-sample size s (Theorem 2). Moreover, (10) implies that $B = \mathcal{O}(n^{3/(4+d)}(\log n)^{(d+1)/(4+d)})$, which is relatively small especially when d is large.

4.1.5 Comments on Convergence Rates For Bagged k-Distances

The novelty of the error analysis of bagged k-distance can be summarized as follows.

1. We propose a novel error decomposition for the error analysis of the bagged k-distance. In detail, while previous works assume the number of bagging as infinity, we consider the finite situation and introduce the additional bagging error term.
2. Based on the previous error analysis of the bagged k-distance, we further introduce a novel error decomposition for the hypothetical density estimation.
3. The theoretical results of the bagged k-distance hold "with high probability".

Before the detailed explanations of the above novelties, we highlight that different from previous statistical analysis, our theoretical analysis is conducted from a learning theory perspective (Cucker and Zhou, 2007; Steinwart and Christmann, 2008), utilizing techniques such as approximation theory and empirical process theory (van der Vaart and Wellner, 1996; Kosorok, 2008).

Firstly, within the learning theory framework, we put forward a novel error decomposition of bagged k-distance (18) in Section 4.1, The additional bagging error term on the right-hand side of (18) distinguishes my error analysis from previous work. Since it is not practically feasible to perform an infinite number of bagging rounds, the study of the bagging error with finite bagging rounds is of great practical significance and importance. By contrast, previous works such as Biau et al. (2010) and Samworth (2012) only consider the bagged nearest neighbor estimate with an infinite number of bagging rounds.

Secondly, as a direct consequence, we derive the convergence rates of the hypothetical density estimation $f_B(x)$ in Propositions 7 and 8. As stated in Section 4.1, we introduce a novel error decomposition for the hypothetical density estimation $f_B(x)$ as in (15). This decomposition indicates that the upper bound of the hypothetical density estimator $f_B(x)$ can be obtained from the error bound of bagged k-distance in Lemma 17.

Finally, we mention that the theoretical results of the bagged k-distance hold "with high probability." Specifically, Propositions 7 and 8 indicate a level of confidence in the rate at which the hypothetical density estimator $f_B(x)$ converges to the true density $f(x)$ (Steinwart and Christmann, 2008). It shows that for the most realizations of datasets sampled from P^n and most resampling processes from P_Z^B, the hypothetical density estimator has nearly optimal performance. Consequently, our findings present a stronger claim compared to previous results, such as those in Biau et al. (2010) that provide expectations over both the resampling distribution and input data, or results in Samworth (2012) that hold "in probability".

4.2 Algorithm Complexity Analysis

In this subsection, we discuss the computational complexity of Algorithm 1. As preliminaries, Friedman et al. (1977) shows that the k-nearest neighbor search by using the k-d tree data structure has a time complexity of $\mathcal{O}(nd \log_2 n)$ and a space complexity of $\mathcal{O}(nd)$ for the tree construction. Moreover, Friedman et al. (1977) shows that, in the search of the k-th nearest neighbor for a test sample, the time complexity is $\mathcal{O}(k \log_2 n)$.

- *Step 1: Calculating the bagged k-distance.* We denote the subsample ratio $\rho = s/n$. Specifically, we build B k-d trees based on B subsampled datasets, each containing ρn samples. A k-d tree, short for k-dimensional tree, is a data structure used for k-NN search in a efficient manner. Then, for every data point in the original dataset, we search for its k_D-th nearest neighbors in each of the B subsampled datasets. Subsequently, we determine its k_D distances as $R_{k_D}(x, D_b)$, where b ranges from 1 to B. Finally, we calculate the bagged k_D-distance $R_{k_D}^B(x)$ for each point in the dataset. The time complexity of the first step is $\mathcal{O}(Bn(\rho d + k_D) \log(\rho n))$:

  - The time complexity of building k-d trees is $\mathcal{O}(B(\rho n) d \log(\rho n))$.

  - The time complexity of k_D-nearest neighbors search of B kd-trees for each data point is $\mathcal{O}(Bnk_D \log(\rho n))$, since the number of the test samples for the nearest neighbor search is the sample size n.
  - The time complexity of calculating its bagged k_D-distance is $\mathcal{O}(Bn)$.

  The space complexity of the first step is $\mathcal{O}(B(\rho n)d)$.

- *Step 2: Calculating the PLLS.* For every sample point in the dataset, we identify its k_L-nearest neighbors and then compute its PLLS. The time complexity of the second step is $\mathcal{O}(n(d + k_L) \log n)$.

  - The time complexity of building a k-d tree on the whole dataset is $\mathcal{O}(nd \log n)$.
  - The time complexity of finding k_L nearest neighbors is $\mathcal{O}(nk_L \log n)$.
  - The time complexity of calculating PLLS is $\mathcal{O}(k_L n)$.

  The space complexity is $\mathcal{O}(nd)$.

- *Step 3: Identifing the core samples.* The core samples are samples with a PLLS value of at least λ. The time complexity of calculating the core points is $\mathcal{O}(n)$. The space complexity is $\mathcal{O}(n)$.

- *Step 4: Constructing a nearest neighborhood graph.* The time complexity of the second step is $\mathcal{O}(n(d + k_G) \log n)$.

  - The time complexity of building a k-d tree on core samples is $\mathcal{O}(nd \log n)$.
  - The time complexity of finding k_G nearest neighbors for each core sample is $\mathcal{O}(nk_G \log n)$.

  The space complexity is $\mathcal{O}(nd)$.

- *Step 5: Extracting the connected components.* The time complexity of finding the connected components by using a disjoint-set data structure is $\mathcal{O}(k_G n\alpha(n))$, where $\alpha(n)$ denotes the inverse-Ackermann function (Ackermann, 1928) of n. It's important to note that $\alpha(n)$ is an extremely slow-growing function. The space complexity is $\mathcal{O}(n)$.

- *Step 6: Labeling points below the level-set by* 1*-NN classifier.* The time complexity is $\mathcal{O}(nd \log n + n \log n)$.

  - The time complexity of building a k-d tree on core-samples is $\mathcal{O}(nd \log n)$.
  - The time complexity of finding the nearest neighbor for each point below the level-set is $\mathcal{O}(n \log n)$.

  The space complexity is $\mathcal{O}(nd)$.

The time and space complexities are summarized in Table 1, where $\alpha(n)$ denotes the inverse-Ackermann function (Ackermann, 1928) of n. According to the orders of B, ρ, k_D, k_L, and k_G in Theorems 5 and 7, we easily find that the time complexity is dominated by $\mathcal{O}(k_L n \log n)$. Moreover, it can be observed that, as d increases, k_L increases as well.

Table 1: Time and Space Complexity for BDMBC

Steps	Time Complexity	Space Complexity
1. Calculating the bagged k-distance	$\mathcal{O}(Bn(\rho d + k_D)\log(\rho n))$	$\mathcal{O}(B\rho nd)$
2. Calculating the PLLS	$\mathcal{O}(n(d + k_L)\log n)$	$\mathcal{O}(nd)$
3. Identifying the core points	$\mathcal{O}(n)$	$\mathcal{O}(n)$
4. Finding k_G nearest neighbors	$\mathcal{O}(n(d + k_G)\log n)$	$\mathcal{O}(nd)$
5. Extracting the connected components	$\mathcal{O}(nk_G\alpha(n))$	$\mathcal{O}(n)$
6. Labeling points below the level-set by 1-NN	$\mathcal{O}(n(d+1)\log n)$	$\mathcal{O}(nd)$
Worst-Case Total Cost	$\mathcal{O}(k_L n \log n)$	$\mathcal{O}(nd)$

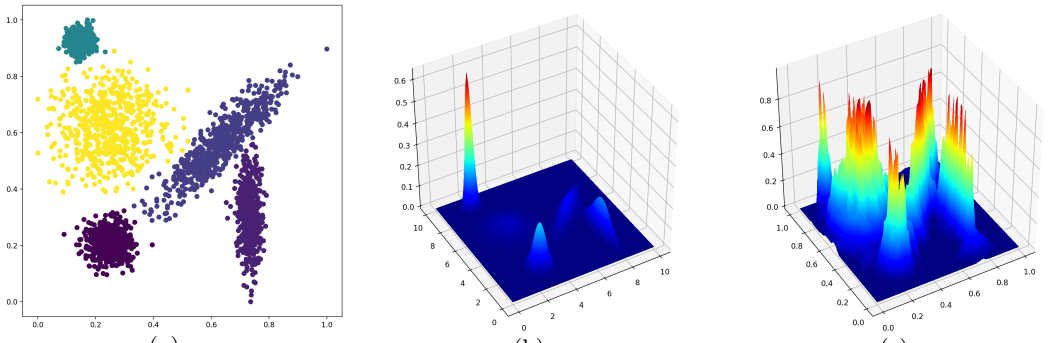
5. Experiments

Firstly, we conduct the experiments of mode estimation on several two-dimensional synthetic datasets in Section 5.1. The ability of the BDMBC algorithm to identify all modes verifies the theoretical results about the mode estimation of the BDMBC. The success of mode estimation is an essential part of our BDMBC for clustering. Then, we evaluate our proposed BDMBC by comparing with other methods on publicly available real-world datasets in Section 5.3. In Section 5.2, we conduct parameter analysis of the BDMBC algorithm, reveal the relationship between the parameter choosing strategies and the performances of BDMBC, and empirically verify the fact that bagging can narrow the searching grid of parameters. We further provide the scalability experiments in Section 5.4 to show that bagging can significantly decrease the computational cost of algorithms without sacrificing accuracy. All the experiments are implemented in Python and run on a machine within a high-performance computing cluster, where one node with 64GB main memory and a 24-core CPU cluster is used.

5.1 Mode Detection

To demonstrate the ability of BDMBC to identify modes so that density-varying mode-based clusters can be detected, we use the following two-dimensional synthetic datasets: The synthetic dataset is generated by a Gaussian mixture model. The Gaussian distribution is consist of five covariance-varying Gaussian distributions with equal mixture component weights. The class of each generated point is the Gaussian mixture component with the highest density.

We generate 3000 points from the distribution, and show the scatter plot of the generated dataset in Figure 5(a). Different clusters are plotted in different colors. We also visualized the probability density function of the Gaussian mixture model in Figure 5(b). Figure 5(b) shows that clusters are density-varied. The densities of the five modes are very different. We apply our BDMBC algorithm to this synthetic dataset, and the estimated PLLS are visualized in Figure 5(c). Compared with Figure 5(b), the local minimums of the estimated PLLS are close to zero, and the local maximums are close to one. As our BDMBC can narrow the density difference of high- and low-density clusters, our BDMBC can successfully distinguish five modes in Figure 5(c). Moreover, Figure 6 on other three additional two-dimensional synthetic datasets (Barton, 2015) also shows that all modes are covered as peaks. Note that we need not provide an accurate estimation of modes. Instead, we use



(a) (b) (c)

Figure 5: Mode detection by the BDMBC algorithm for datasets with density-varying clusters on the Gaussian Mixture Model. (a) Raw Dataset generated from the synthetic distribution. (b) Density of the synthetic distribution. (c) Result of BDMBC on the estimated probability of localized level-set.

non-overlapping clusters to cover modes and each mode is covered by only one cluster. We mention that although our BDMBC may enlarge the difference of densities nearby local maximums in Figures 5 and Figures 6, these fluctuations do not affect the detection of modes and clusters.

5.2 Parameter Analysis

In this subsection, we firstly apply parameter analysis of four hyper-parameters including the number of nearest neighbors for hypothetical density estimation k_D, the number of nearest neighbors for the PLLS k_L, the number of nearest neighbors for graph connection k_G and the level-set threshold λ on a 2-dimensional synthetic dataset `3Clusters`, which is visualized in Figure 9 and has a sample size 2000. Then, we discuss how bagging helps with parameter tuning by comparing the optimal parameters between DMBC and BDMBC on the `3Clusters` dataset and five additional synthetic datasets. Lastly, we summarize the results from synthetic experiments on the selection of hyper-parameters. The synthetic datasets introduced in this subsection are from the `Python` package `sklearn` (Pedregosa et al., 2011) and Iglesias et al. (2019). For each of the six synthetic datasets, we set the sample size to be 2000, the noise rate as 0.05, and visualize the dataset in Figure 9.

5.2.1 Clustering Measures

In our experiments, we use two clustering-based metrics and two classification-based metrics to evaluate the clustering performances of the BDMBC, including Adjusted Rand Index (ARI), Normalized Mutual Infomation (NMI), $F1$ score, and accuracy. The mathematical definition of each measure is defined as follows.

- **ARI**: ARI (Hubert and Arabie, 1985) measures the differences between two clustering results, adjusted for the chance of grouping of elements for Rand Index (RI) (Rand, 1971).

$$\text{ARI} = \frac{RI - E[RI]}{\max(RI) - E[RI]}, \quad RI = \frac{a+b}{\binom{n}{2}},$$

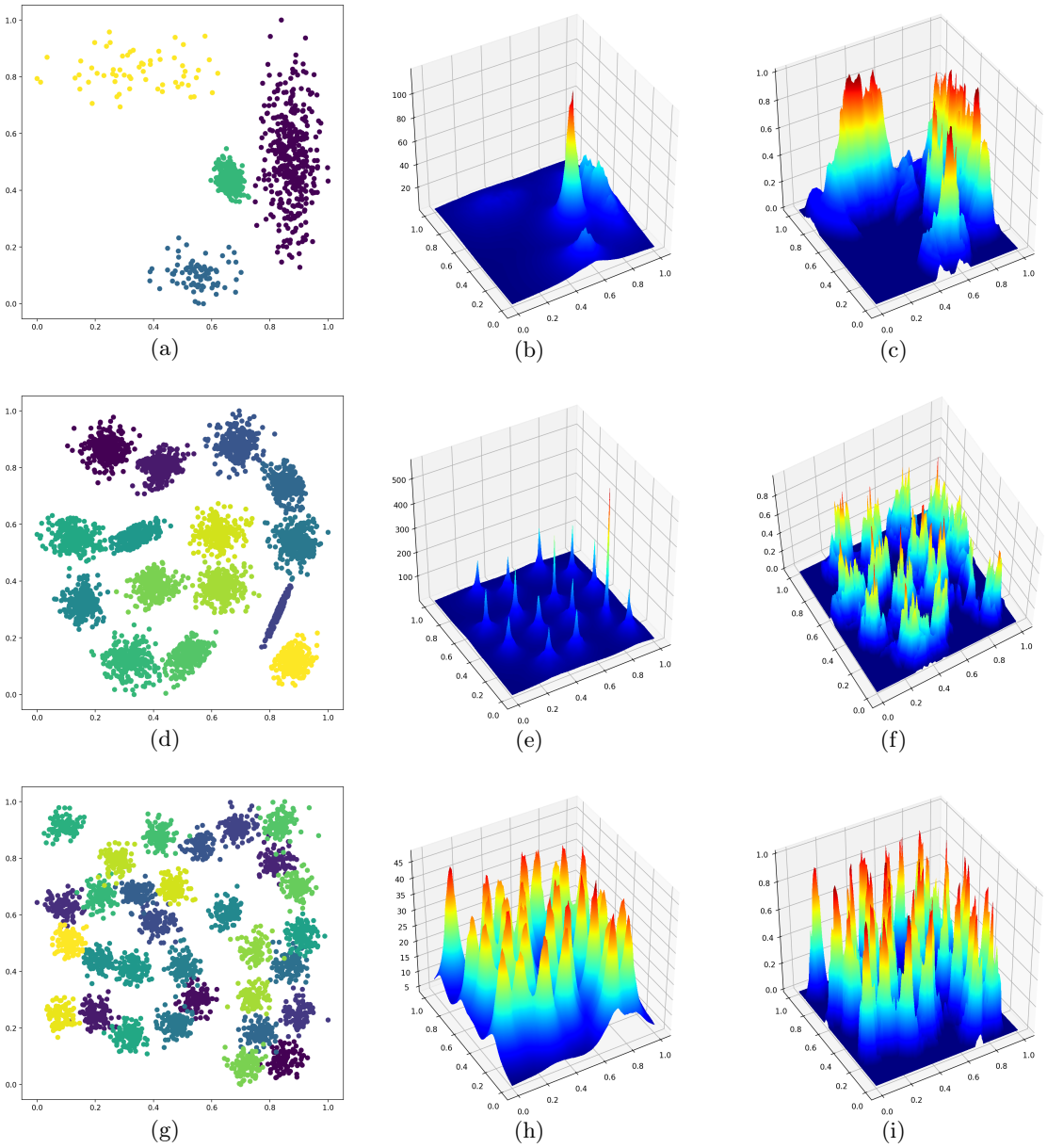


Figure 6: Mode detection by the BDMBC algorithm for datasets with density-varying clusters on other synthetic datasets. (a)(d)(g) Scatter plot of the raw dataset generated from the synthetic distribution. (b)(e)(h) Hypothetical density estimation of the synthetic distribution by bagged k-distance. (c)(f)(i) Result of BDMBC on the estimated probability of localized level-set.

where a is the number of paired objects placed in the same cluster in both partitions, b is the number of paired objects placed in different clusters in both partitions, and the expected value E of the Rand Index for random clusterings are given by $E = [\sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2}]/\binom{n}{2}$, where n_i is the number of samples in cluster i and n_j is the number of samples in cluster j.

- **NMI**: The Mutual Information (MI) (Strehl and Ghosh, 2002) is a symmetric measure that quantifies the mutual dependence between two random variables, or the information that two random variables share.

$$\text{NMI} = \frac{I(Y, P)}{\sqrt{H(Y)H(P)}}$$

where $H(x)$ represents the entropy of x, and $I(Y, P)$ represents the mutual information of Y and P.

On the other hand, as for the classification measure $F1$ measure and accuracy, we have to first use the Kuhn-Munkres (Munkres, 1957; Kuhn, 1955) methods to assign the clustering labels to the underlying labels of instances and then calculate the measure.

- ***F1***: The $F1$ score can be interpreted as a harmonic mean of the precision and recall.

$$F1 = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

where precision describes the ability to only predict really positive samples as samples, denoted as precision $= \frac{TP}{TP+FP}$. And the recall, calculated by recall $= \frac{TP}{TP+FN}$, can be interpreted as the ability of the classifier to find all the positive samples.

- **Accuracy**: The accuracy measures the ratio of correct clustering.

$$ACC = \frac{\#\text{Correct Classification}}{n}$$

5.2.2 Parameter analysis of hyper-parameters k_D and k_L

Firstly, we fix the number of nearest neighbors for graph connection $k_G = 8$ and the level-set threshold $\lambda = 0.5$ which are suitable hyper-parameters for a good clustering performance. We vary the number of nearest neighbors for hypothetical density estimation k_D and for the PLLS k_L. The adjusted rand index (ARI) scores and the number of clusters on `3Clusters` as the function of (k_D, k_L) are visualized in Figure 7. We find that the clustering performance is relatively insensitive to the parameters of k_D and k_L: If k_D and k_L are not too small nor too large, the clustering performance is good, see the dark red filled region on the left side of Figure 7. Moreover, the good clustering performance attributes to the performance of mode estimation. See the right side of Figure 7. A wide range of k_D and k_L can obtain the correct number of clusters (filled in green), which means that all the three modes are detected successfully.

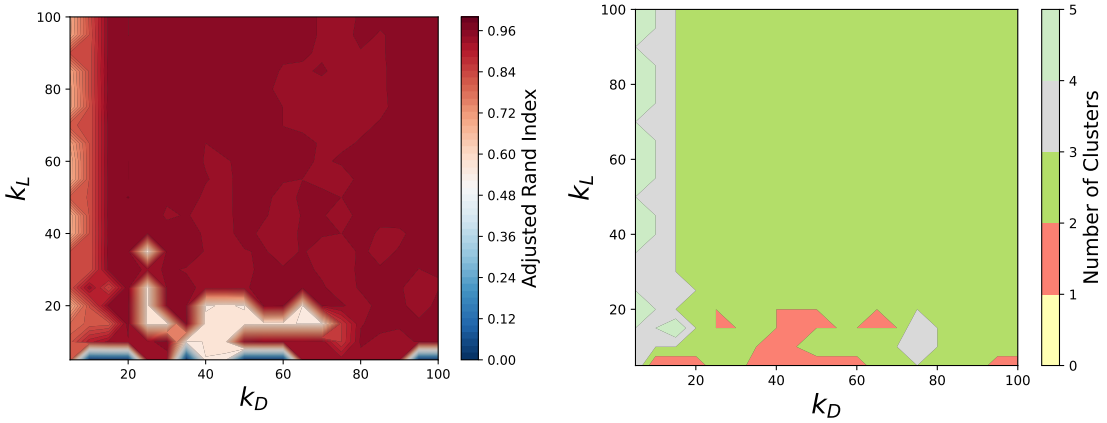


Figure 7: Visualization of `3Clusters` among with ARI scores and the number of clusters as k_D and k_L are changed and $k_G = 8$, $\lambda = 0.5$ are fixed. They shows that a wide range of (k_D, k_L) obtain good clustering performance with correct mode estimation.

5.2.3 Parameter analysis of hyper-parameters k_G and λ

Secondly, we fix the number of nearest neighbors for hypothetical density estimation and localized level-set $k_D = 30$ and $k_L = 60$, and explore the selection of hyper-parameters k_G and λ. In Figure 8, we vary the k_G and λ, and we visualize the ARI scores and the number of clusters on `3Clusters` dataset. See the red filled region on the left figure which means good clustering performance, and we observe that there is a positive linear correlation between optimal k_G-s and optimal λ-s: we can achieve good clustering performance by selecting a pair of relatively small parameters (k_G, λ) or a pair of relatively large parameters (k_G, λ). This can guide the selection of these two hyper-parameters. Similarly, the performance of mode estimation is also good for hyper-parameters with good clustering performance. (See the green-filled region on the right.)

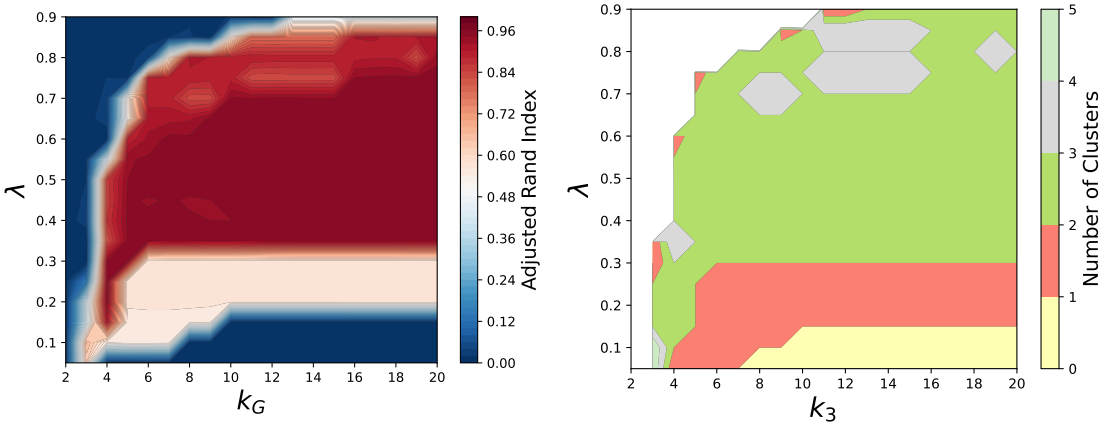
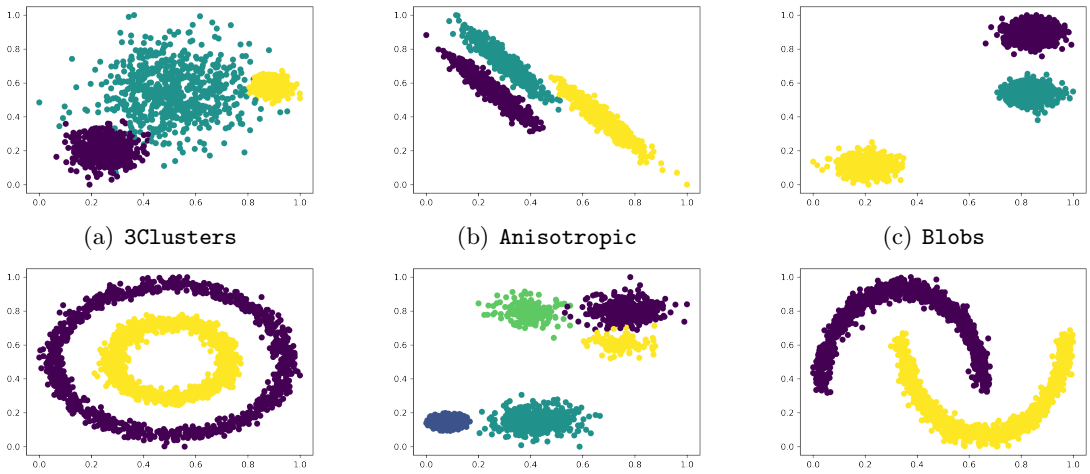


Figure 8: Visualization of `3Clusters` among with ARI scores and the number of clusters as k_G and λ are changed and $k_D = 30$, $k_L = 60$ are fixed. They show that a proper selection of (k_D, k_L) obtains good clustering performance with correct mode estimation.



(a) `3Clusters` (b) `Anisotropic` (c) `Blobs`

(d) `Circles` (e) `MDCGen` (f) `Moons`

Figure 9: Visualization of the synthetic datasets.

5.2.4 The Effects of Bagging

In this subsection, we list the optimal parameters for DMBC and BDMBC in various synthetic datasets in Table 2 to demonstrate the effects of bagging on parameter tuning, i.e., with a small sampling ratio, bagging can accelerate the algorithm by narrowing the range of parameter k_D. In the experiments for synthetic datasets, we set the number of bagging iterations as $B = 10$ and the sampling ratio $\rho = 0.1$. As we can see from Table 2, the optimal parameter of k_D for BDMBC is much smaller than that for DMBC. Therefore, bagging enables BDMBC to have a more narrow searching grid of k_D and prevents the algorithm from tedious parameter searching. In addition, bagging with a relatively small ρ can further speed up the algorithm by decreasing the number of training samples in each iteration. To be specific, bagging makes it possible to learn the distributional pattern of training datasets with only a small fraction of samples. Meanwhile, bagging can also increase the randomness and boost the clustering performance. This is empirically verified in Table 4, where the clustering performances of BDMBC are significantly better than DMBC in many cases.

5.2.5 Summary of Parameter Selection

We summarize the experiments concerning synthetic datasets as belows.

- Considering the computational cost, $B = 25$ is acceptable. As for ρ, an empirical rule is $\rho = 0.1$, 0.5 or 0.9 to reduce the parameter grid.
- k_D and k_L are related to the hypothetical density estimation and the corresponding probability of localized level sets. With a small ρ, k_D can be searched from {2, 4, 7}; and the k_L can be searched from {10, 20}.
- k_G and λ are two hyper-parameters for level-set clustering. k_G are quite stable for various dataset, and an empirical rule is $k_G \in \{5, 10, 20\}$. A large λ is more robust to noise samples and noisy density estimates in practice, so an empirical rule is to try

Table 2: The comparison of optimal parameters for DMBC and BDMBC in synthetic datasets

Data	Bagging	ρ	k_D	k_L	λ	k_G
`3Clusters`	No	-	23	9	0.65	17
	Yes	0.1	4	10	0.65	15
`Anisotropic`	No	-	17	17	0.5	15
	Yes	0.1	4	20	0.5	16
`Blobs`	No	-	16	8	0.4	15
	Yes	0.1	4	10	0.5	9
`Circles`	No	-	17	14	0.45	19
	Yes	0.1	7	10	0.55	15
`MDCGen`	No	-	17	26	0.6	19
	Yes	0.1	7	20	0.4	9
`Moons`	No	-	11	9	0.10	16
	Yes	0.1	2	10	0.05	12

various λ from a relatively small $\lambda = 0.10$ to a relatively large $\lambda = 0.50$ or even larger (0.9).

5.3 Comparisons on Real Datasets

We evaluate the clustering performance of our proposed BDMBC by comparing with other competing methods on real-world classification datasets. Before illustrating the experimental results, we demonstrate the basic information of the real datasets, list all comparing methods and provide the parameter settings for each method, and introduce the metrics for clustering performance evaluation.

5.3.1 Datasets

We collect the binary and multi-class classification datasets from UCI Machine Learning Repository, including `wine` (Dua and Graff, 2017), `banknote` (Dua and Graff, 2017), `HTRU2` (Lyon et al., 2016), `iris` (Dua and Graff, 2017), `gisette` (Guyon et al., 2004). In addition, we further collect two image datasets for analyzing the high-dimension situations, including `COIL` (Nene et al., 1996) and `USPS` (Hull, 1994). These datasets are summarized in Table 3, where we list the number of samples n, the number of dimensions d, the number of clusters c, the number of samples falling into the smallest cluster `min_nums`, and the number of samples falling into the largest cluster `max_nums`. Specifically, before experiments, we scale the datasets to the $[0, 1]$ range on each dimension.

5.3.2 Baselines and Parameter Settings

The three baselines include conventional and state-of-the-art clustering algorithms. They are an improved version of DBSCAN called DBSCAN++ (Jang and Jiang, 2019), HDBSCAN (Campello et al., 2013), and an improved version of mean-shift called quickshift++ (Jiang et al., 2018). As a preprocessing, we normalize the datasets within the range $[0, 1]$ using the

Table 3: Descriptions of Datasets

Dataset	n	d	c	min_nums	max_nums
iris	150	4	3	50	50
wine	178	13	3	48	71
seeds	210	7	3	70	70
banknote	1372	4	2	610	762
COIL	1440	1024	20	72	72
gisette	7000	5000	2	3500	3500
USPS	9298	256	10	708	1553
HTRU2	17898	8	2	1639	16259

`MinMaxScaler` from the `Scikit-Learn` package to standardize the parameter tuning process, especially for parameters like the distance-based parameters ε_d and ε_c in DBSCAN++.

Parameter optimization is still an open question for clustering (Gan et al., 2020). For each baseline, we search the parameters according to the author's suggestion or try to search the best parameters within a reasonable range. The specific process of searching the best parameters is that for each evaluation criterion, we find the optimal result within the parameter grid and then record the corresponding optimal parameters. It is worth highlighting that the notation $[a : b : c]$ means the parameter grid starts from a to c with the step size b. We list the details of parameter tuning in the following.

- For DBSCAN++, there are four hyper-parameters including the radius for determining the core points in clusters ε_d, the radius for determining if two clusters are connected ε_c, the sampling fraction p, and the number of neighbors for a point to be labeled as a core point `minPts`. We search ε_d and ε_c in the grid of $[0.05 : 0.05 : 0.95]$. However, since the average distance between points will be considerably larger as the feature dimension increases, we amend the parameter tuning strategies (searching ε_d and ε_c in the grid of $[0.5 : 0.5 : 25]$) for datasets `USPS`, `COIL`, and `Gisette` which own extremely higher dimensions than others. We search the sampling fraction p in the grid of $\{0.1, 0.2, 0.4, 0.6, 0.8\}$, and `minPts` in the grid of $[2 : 2 : 40]$.

- For HDBSCAN, there are two hyper-parameters including the search range of the minimum restriction of clusters `ClusterSize` and the search range of another parameter `MinSamples`. As these two hyper-parameters are dependent on the sample size, we set `ClusterSize` as $p_{\texttt{ClusterSize}} \times$ (sample size) and `MinSamples` as $p_{\texttt{MinSamples}} \times$ (sample size), where the grids of $p_{\texttt{ClusterSize}}$ and $p_{\texttt{MinSamples}}$ are both {0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.10, 0.15}.

- For Quickshift++, there are three hyper-parameters including the number of neighbors to calculate the density k, the threshold for mode detection β, and the minimum restriction ε for connecting clusters. Since k is dependent on the sample size, we set k as $p_k \times$ (sample size), where the grid of p_k is {0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.10, 0.15, 0.20}. We search β in the grid of $[0.1 : 0.1 : 0.9]$ and ε in the grid of $[0.0 : 0.1 : 0.9]$.

- For the proposed BDMBC, the grid of hyper-parameter matches the results in the parameter analysis section. In detail, the number of bagging is set $B = 25$, the sampling rate ρ is searched among a coarse grid $\{0.5, 0.9\}$, and the threshold for the level-set λ is searched among a coarse grid $\{0.1, 0.5, 0.9\}$. There are three hyper-parameters about the number of nearest neighbors, including k_D which is used to construct the k-distance hypothetical density estimator, k_L which is used to constraint the region for calculating the localized density ratio, and k_G which is the number of nearest neighbors for graph connection. Based on empirical observations, it has been determined that low-dimensional datasets (such as `Iris`, `Wine`, `Seeds`, `Banknote`, and `HTRU2`) require larger values of k_D, k_L, and k_G, whereas high-dimensional datasets (including `COIL`, `Gisette`, and `USPS`) benefit from smaller values of these hyper-parameters. Consequently, different parameter grids are employed based on the feature dimensionality of the dataset. Furthermore, it has been observed that the choice of k_L is more dependent on the sample size. Therefore, k_L is determined as the product of a coefficient, denoted as $p_{k_L} \times$ (sample size).
  - k_D is searched in the grid of $\{2, 4, 7, 10, 20\}$ for low-dimensional datasets (for small $n < 2000$, $k_D \in \{2, 4, 7\}$ and for $n > 2000$, $k_D \in \{7, 10, 20\}$) and in the grid of $\{2, 7, 10\}$ for high-dimensional datasets.
  - $p_{k_L} := k_L/n$ is searched in the grid of $\{0.01, 0.03, 0.05, 0.08\}$ for low-dimensional datasets and $p_{k_L} \in \{0.010, 0.015, 0.020\}$ for high-dimensional datasets.
  - k_G is searched in the grid of $\{5, 10, 20\}$ for low-dimensional datasets and $k_G \in \{3, 5, 10\}$ for high-dimensional datasets.

5.3.3 Experimental Results

In this subsection, we compare the performances of our algorithm with the other three baselines with four measures. The results are demonstrated in Table 4. The maximum obtained performances are highlighted in **bold** and the second maximum are highlighted in *italic*. We record the best hyper-parameters of DBSCAN++, HDBSCAN, BDMBC, and Quickshift++ under different criteria in Tables 5, 6, 7, and 10 in Appendix A, respectively.

We find that the majority of the optimal hyper-parameters fall within the parameter grid for the three competing methods, which suggests that the defined parameter grids are wide enough for comparison. Table 4 indicates that BDMBC outperforms other methods in most datasets. The traditional density-based clustering and modal-detecting method show less competitive performances. For some datasets with large sample sizes or high dimensions, such as `Gisette` and `USPS`, we outperform these comparing methods by large margins.

5.4 Scalability Experiments

In this subsection, we use a large-scale synthetic data named `Artset` (Iglesias et al., 2019) to explore the clustering running times of the BDMBC algorithm. In large scale scenarios, besides conducting bagging for the parameter k_D, we can also apply bagging for the parameter k_L to speed up the BDMBC algorithm. To this end, we fix the feature dimension $d = 10$ and the number of clusters $k = 10$, and change the sample size $n \in \{1 \times 10^5, 2 \times 10^5, 5 \times 10^5, 1 \times 10^6\}$. Then we train BDMBC to compare the following two

Table 4: Comparison with baselines on real-world datasets.

Data	Measure	DMBC	BDMBC	DBSCAN++	HDBSCAN	Quickshift++
Iris	ARI	0.8511	**0.9222**	0.8188	0.5681	0.7005
	NMI	0.8448	**0.9011**	0.8411	0.7337	0.7415
	F1	0.9464	**0.9733**	0.9327	0.5556	0.5701
	ACC	0.9466	**0.9733**	0.9333	0.6667	0.8733
Wine	ARI	0.8470	**0.8498**	0.8483	0.4687	0.6287
	NMI	0.8112	**0.8222**	0.8212	0.6125	0.6682
	F1	0.9510	0.9502	**0.9513**	0.5467	0.5359
	ACC	**0.9494**	**0.9494**	**0.9494**	0.6461	0.8483
Seeds	ARI	0.7819	**0.7881**	0.7690	0.6923	0.7209
	NMI	0.7632	**0.7644**	0.7471	0.6922	0.6797
	F1	0.9130	**0.9238**	0.9227	0.5425	0.8941
	ACC	0.9233	**0.9238**	**0.9238**	0.8143	0.8952
Banknote	ARI	0.8757	0.8977	0.8962	**0.9624**	0.9539
	NMI	0.8076	0.8451	0.8199	**0.9317**	0.9194
	F1	0.9676	0.9736	0.9277	**0.9904**	0.9882
	ACC	0.9679	0.9738	0.9555	**0.9905**	0.9883
HTRU2	ARI	0.8086	**0.8123**	0.7359	0.7486	0.5289
	NMI	0.6646	**0.6671**	0.6012	0.5759	0.4125
	F1	0.9166	**0.9184**	0.8823	0.7677	0.4760
	ACC	0.9747	**0.9751**	0.9669	0.9502	0.9461
COIL	ARI	0.7952	0.7630	**0.8265**	0.7874	0.7274
	NMI	0.9311	0.9276	**0.9313**	0.9296	0.8849
	F1	0.7535	0.7461	**0.8099**	0.7988	0.5588
	ACC	0.7993	0.7910	**0.8549**	0.8347	0.7639
Gisette	ARI	0.5361	**0.6164**	0.0555	0.0681	0.0273
	NMI	0.4471	**0.5145**	0.1321	0.1601	0.1715
	F1	0.8656	**0.8924**	0.0664	0.4716	0.0239
	ACC	0.8661	**0.8926**	0.3624	0.5813	0.4996
USPS	ARI	0.7526	**0.7912**	0.1714	0.5923	0.5055
	NMI	0.7955	**0.8104**	0.4953	0.6689	0.6590
	F1	0.7816	**0.7957**	0.1235	0.5594	0.2645
	ACC	0.8437	**0.8541**	0.2724	0.6758	0.6138

For each dataset and each measure, we denote the best performance with **bold**.

Table 5: The Optimal Hyper-parameter of DBSCAN++

Dataset	ARI				NMI				F1				ACC			
	p	ε_d	ε_c	minPts	p	ε_d	ε_c	minPts	p	ε_d	ε_c	minPts	p	ε_d	ε_c	minPts
Iris	0.6	0.05	0.25	2	0.6	0.05	0.25	2	0.6	0.05	0.25	2	0.6	0.05	0.25	2
Wine	0.4	0.40	0.60	3	1.0	0.50	0.40	20	1.0	0.50	0.40	20	0.4	0.40	0.60	3
Seeds	0.8	0.10	0.30	2	0.8	0.10	0.30	2	0.8	0.10	0.30	2	0.8	0.10	0.30	2
Banknote	0.6	0.15	0.10	20	0.6	0.15	0.10	20	0.6	0.05	0.25	12	0.6	0.15	0.10	20
HTRU2	0.2	0.05	0.10	10	0.2	0.05	0.10	10	0.2	0.05	0.10	10	0.2	0.05	0.10	10
COIL	0.6	7.00	4.50	4	0.6	5.00	5.00	2	0.6	4.50	5.00	2	0.6	4.50	4.50	2
Gisette	0.1	20.0	5.0	2	0.1	20.0	5.0	2	0.1	20.0	5.0	2	0.1	20.0	5.0	2
USPS	0.2	6.50	1.00	40	0.2	6.50	1.00	38	0.8	1.00	0.10	2	0.8	1.00	0.10	2

Table 6: The Optimal Hyper-parameter of HDBSCAN

Dataset	ARI		NMI		F1		ACC	
	$p_{\texttt{ClusterSize}}$	$p_{\texttt{MinSamples}}$	$p_{\texttt{ClusterSize}}$	$p_{\texttt{MinSamples}}$	$p_{\texttt{ClusterSize}}$	$p_{\texttt{MinSamples}}$	$p_{\texttt{ClusterSize}}$	$p_{\texttt{MinSamples}}$
Iris	0.0100	0.0200	0.0100	0.0200	0.010	0.0200	0.0100	0.0200
Wine	0.0100	0.0500	0.0100	0.0500	0.010	0.0500	0.0100	0.0500
Seeds	0.0200	0.0200	0.0200	0.0200	0.100	0.0100	0.0200	0.0200
Banknote	0.1000	0.0020	0.1000	0.0020	0.100	0.0020	0.1000	0.0020
HTRU2	0.0005	0.0020	0.0005	0.0020	0.002	0.0005	0.0005	0.0020
COIL	0.0100	0.0020	0.0100	0.0020	0.010	0.0050	0.0200	0.0020
Gisette	0.0005	0.0010	0.0005	0.0010	0.002	0.0005	0.0005	0.0010
USPS	0.0005	0.0005	0.0005	0.0005	0.001	0.0005	0.0010	0.0005

Table 7: The Optimal Hyper-parameter of BDMBC

Dataset	ARI					NMI					F1					ACC				
	ρ	k_D	p_{k_L}	λ	k_G	ρ	k_D	p_{k_L}	λ	k_G	ρ	k_D	p_{k_L}	λ	k_G	ρ	k_D	p_{k_L}	λ	k_G
Iris	0.9	7	0.010	0.9	10	0.9	7	0.010	0.9	10	0.9	7	0.010	0.9	10	0.9	7	0.010	0.9	10
Wine	0.9	4	0.050	0.9	10	0.9	4	0.050	0.9	10	0.9	4	0.050	0.9	10	0.9	4	0.050	0.9	10
Seeds	0.5	2	0.030	0.9	20	0.5	4	0.050	0.9	20	0.9	2	0.050	0.9	20	0.5	2	0.030	0.9	20
Banknote	0.5	10	0.010	0.5	20	0.5	10	0.010	0.5	20	0.5	10	0.010	0.5	20	0.5	10	0.010	0.5	20
HTRU2	0.5	20	0.050	0.5	20	0.9	4	0.080	0.5	20	0.5	20	0.050	0.5	20	0.5	20	0.050	0.5	20
COIL	0.9	7	0.015	0.1	3	0.9	7	0.010	0.1	3	0.9	7	0.020	0.1	3	0.9	7	0.015	0.1	3
Gisette	0.9	2	0.020	0.9	5	0.9	2	0.020	0.9	5	0.9	2	0.020	0.9	5	0.9	2	0.020	0.9	5
USPS	0.5	7	0.015	0.5	3	0.5	7	0.015	0.5	3	0.5	7	0.015	0.5	3	0.9	10	0.010	0.9	10

settings: the first is the bagging version with $B = 10$ and $\rho = 0.001$, and the second is the non-bagging version with $B = 1$ and $\rho = 1.0$. For each setting, we select the optimal parameters including k_D, k_L, λ, K_G, calculate four clustering measures on behalf of the performances, and record the time consumptions of training the k-distance-based PLLS for each setting. The running time is measured in seconds.

Table 8: The comparison between the DMBC and BDMBC algorithms on the large-scale synthetic dataset **Artset** with four different sample sizes. Four clustering measures and the time for training the k-distance-based PLLS are included.

Sample Size	Bagging	ARI	NMI	F1	Accuracy	Time (s)
1×10^5	No	0.9910	0.9800	0.7009	0.9886	27.77
	Yes	0.9910	0.9803	0.7038	0.9897	0.41
2×10^5	No	0.9935	0.9864	0.5922	0.9956	85.30
	Yes	0.9936	0.9865	0.9922	0.9956	0.91
5×10^5	No	0.9875	0.9737	0.8862	0.9805	278.99
	Yes	0.9867	0.9724	0.8975	0.9829	3.94
1×10^6	No	0.9881	0.9733	0.8878	0.9822	1086.36
	Yes	0.9901	0.9787	0.9815	0.9888	13.15

As we can see from Table 8, the clustering performance of the bagging version of the BDMBC algorithm with a very small sampling ratio ρ is comparable with the non-bagging version. However, the time of training the PLLS for the bagging version of BDMBC can be ten times or even a hundred times less than that of the non-bagging version, which

empirically verifies that bagging can improve the computational efficiency of BDMBC by training the k-distance-based PLLS with much fewer samples.

6. Conclusions

In this paper, we propose an ensemble algorithm called *bagged k-distance for mode-based clustering* (*BDMBC*) by putting forward a new measure called the *probability of localized level sets* (*PLLS*), which transforms the multi-level density clustering to the single-level setting. To deal with high-dimensional datasets, we employ the k-distance that can be directly calculated from the data. We further introduce the bagging technique to improve computational efficiency in large-scale situations. To establish solid theoretical guarantees of the proposed algorithm, we first derive optimal convergence rates for mode estimation with properly chosen parameters. It turns out that with a relatively small B, the subsample size s can be much smaller than the number of training data n at each bagging round, and the number of nearest neighbors k_D can be reduced simultaneously. Moreover, by establishing fast convergence rates for the level set estimation of the PLLS in terms of Hausdorff distance, we show that BDMBC can find localized level sets for varying densities and thus enjoys local adaptivity. Finally, we also conducted persuasive experiments on both synthetic and real-world datasets, showing the promising experimental performances of our BDMBC, demonstrating how bagging narrows the searching grid of parameters, and offering advice on how to choose parameters in applications.

It's worth pointing out that compared to other clustering algorithms, our algorithm BDMBC enjoys various advantages. On the one hand, BDMBC is more computationally efficient than hierarchical density-based clustering algorithms. On the other hand, compared with other mode-based clustering algorithms, BDMBC can handle high-dimensional data more effectively and has an easy procedure in the parameter-searching procedure.

7. Proofs

This section presents the proofs concerning the theoretical analysis. We first present the convergence rate of the DMBC algorithm, i.e. the special case of BDMBC with $B = 1$ in Section 7.1. Section 7.2 presents the proofs related to the k-distance and bagged k-distance in Section 4.1. Section 7.3 gives the proofs related to mode estimation in Section 3.1. Section 7.4 provides all proofs related to level set estimation for the proposed probability function PLLS in Section 3.2.

7.1 Convergence Rates of DMBC for Mode Estimation

To demonstrate the benefits of bagging in mode estimation, we consider the DMBC algorithm, which can be viewed as a special case of BDMBC in Algorithm 1 with $B = 1$ and $s = n$. More specifically, we only use k-distance for mode-based clustering without bagging. The procedure of DMBC can be described as follows. Firstly, we compute the empirical PLLS with respect to the k-distance by

$$\widehat{p}_{k_L}^{k_D}(x) := \frac{1}{k_L} \sum_{i=1}^{k_L} \mathbf{1}\big\{R_{k_D}(X_{(i)}(x)) \geq R_{k_D}(x)\big\}. \tag{19}$$

Then we construct the subgraph $G_k(\lambda)$ retaining the core-samples by

$$\widehat{D}_k(\lambda) = \{X_i \in D : \widehat{p}_{k_L}^{k_D}(X_i) \geq \lambda\} \tag{20}$$

and the mode set with respect to the k-distance by

$$\widehat{\mathcal{M}}^k = \{X_i \in D : \widehat{p}_{k_L}^{k_D}(X_i) = 1\}. \tag{21}$$

Finally, we compute the cluster estimators $\mathcal{C}_k(\lambda)$, i.e., the connected components of $G_k(\lambda)$.

The next theorem presents the convergence rates of DMBC, i.e., k-distance for multi-modal distribution under the above mild assumptions.

Theorem 9 *Let Assumptions 1, 2 and 3 hold with $2\alpha\gamma \leq 4 + d$ and $\widehat{\mathcal{M}}^k$ be the mode estimator as in (21). Then for every mode $m_i \in \mathcal{M}$ and $\lambda \geq c$ with the constant c specified in the proof, by choosing*

$$k_{D,n} := n^{\frac{d}{4+d}} (\log n)^{\frac{d}{4+d}}, \qquad k_{G,n} \asymp \log n, \qquad k_{L,n} \gtrsim n^{1-\frac{\alpha\gamma}{4+d}} (\log n)^{1+\frac{\alpha\gamma}{4+d}},$$

there exists a mode estimate $\widehat{m}_i$ such that with probability P^n at least $1 - 2/n^2$, there holds

$$\|\widehat{m}_i - m_i\|_2 \lesssim (\log n/n)^{\frac{1}{4+d}}.$$

Moreover, there exist distinct cluster estimators $\widehat{C}_i \in \mathcal{C}_k(\lambda)$, $1 \leq i \leq k$, such that $\widehat{m}_i \in \widehat{C}_i$.

Theorem 2 shows that if k_D and k_L are chosen properly, then the convergence rate of DMBC matches the lower bound established in Tsybakov (1990) up to a logarithmic factor. Therefore, Theorem 2 coincides with the optimal recovery for multiple modes established in Dasgupta and Kpotufe (2014); Jiang (2017a); Jiang et al. (2018). Finally, we mention that the mode estimation returned by (9) corresponds to the true modes of f in a subjective manner.

7.2 Proofs Related to Section 4.1

In this section, we present the proofs related to the bagged k-distance. To be specific, in Sections 7.2.1-7.2.3, we provide the proofs related to the bagging error, estimation error, and approximation error for the hypothetical density estimation in Sections 4.1.1-4.1.3, respectively. With these preparations, in Section 7.2.4, we provide proofs related to Section 4.1.4, we first establish convergence rates for hypothetical density estimation. Then we propose an important lemma related to Taylor's expansion of the density function around the modes, which supplies the key to proofs of the mode estimation and mode-based clustering. Finally, we derive faster convergence rates of the hypothetical density estimation around the modes using this lemma. These theoretical results play a fundamental role in the proof of mode estimation and level set estimation for BDMBC and DMBC in Sections 7.3 and 7.4.

Before we proceed, we list the well-known Bernstein's inequality that will be used frequently in the proofs. Lemma 10 was introduced in Bernstein (1946) and can be found in many statistical learning textbooks, see e.g., Massart (2007); Cucker and Zhou (2007); Steinwart and Christmann (2008).

Lemma 10 (Bernstein's inequality) *Let $B > 0$ and $\sigma > 0$ be real numbers, and $n \geq 1$ be an integer. Furthermore, let $\xi_1, \ldots, \xi_n$ be independent random variables satisfying $\mathbb{E}_P \xi_i = 0$, $\|\xi_i\|_\infty \leq B$, and $\mathbb{E}_P {\xi_i}^2 \leq \sigma^2$ for all $i = 1, \ldots, n$. Then for all $\tau > 0$, we have*

$$P\left(\frac{1}{n}\sum_{i=1}^{n} \xi_i \geq \sqrt{\frac{2\sigma^2\tau}{n}} + \frac{2B\tau}{3n}\right) \leq e^{-\tau}.$$

7.2.1 Proofs Related to Section 4.1.1

To prove Proposition 4, we need to bound the number of reorderings of the data. To be specific, for fixed $x \in \mathbb{R}^d$, we reorder samples, $X_1, \ldots, X_n$, according to increasing values of $\|X_i - x\|$ with breaking ties by considering indices, i.e., $\|X_{\sigma_1} - x\| \leq \cdots \leq \|X_{\sigma_n} - x\|$, where $(\sigma_1, \ldots, \sigma_n)$ is a permutation of $(1, \ldots, n)$. Then we define the inverse of the permutation, namely the rank Σ_i by $\Sigma_i := \{1 \leq \ell \leq n : X_{\sigma_\ell} = X_i\}$. Since we break ties by considering indices, the rank Σ_i is unique for all $1 \leq i \leq n$. Therefore, the rank vector $(\Sigma_1, \ldots, \Sigma_n)$ for $x \in \mathbb{R}^d$ is well-defined. Let $\mathcal{S} = \{(\Sigma_1, \ldots, \Sigma_n), x \in \mathbb{R}^d\}$ be the set of all rank vectors one can observe by moving x around in space and we use the notation $|\mathcal{S}|$ to represent the cardinality of $\mathcal{S}$.

The next lemma, which plays a crucial role to derive the uniform bound for the proof of Propositions 4, provides the upper bound for the number of reorderings, see also Lemma 20 in Hang et al. (2022).

Lemma 11 *For any $d \geq 1$ and all $n \geq 2d$, there holds $|\mathcal{S}| \leq (25/d)^d n^{2d}$.*

To further our analysis, we first need to recall the definitions of *VC dimension* and *covering number*, which are frequently used in capacity-involved arguments and measure the complexity of the underlying function class (van der Vaart and Wellner, 1996; Kosorok, 2008; Giné and Nickl, 2021).

Definition 12 (VC dimension) *Let $\mathcal{B}$ be a class of subsets of $\mathcal{X}$ and $A \subset \mathcal{X}$ be a finite set. The trace of $\mathcal{B}$ on A is defined by $\{B \cap A : B \subset \mathcal{B}\}$. Its cardinality is denoted by $\Delta^{\mathcal{B}}(A)$. We say that $\mathcal{B}$ shatters A if $\Delta^{\mathcal{B}}(A) = 2^{\#(A)}$, that is, if for every $A' \subset A$, there exists a $B \subset \mathcal{B}$ such that $A' = B \cap A$. For $n \in \mathrm{N}$, let*

$$m^{\mathcal{B}}(n) := \sup_{A \subset \mathcal{X},\, \#(A) = n} \Delta^{\mathcal{B}}(A). \tag{22}$$

Then, the set $\mathcal{B}$ is a Vapnik-Chervonenkis class if there exists $n < \infty$ such that $m^{\mathcal{B}}(n) < 2^n$ and the minimal of such n is called the VC dimension of $\mathcal{B}$, and abbreviate as $\mathrm{VC}(\mathcal{B})$.

Since an arbitrary set of n points $\{x_1, \ldots, x_n\}$ possess 2^n subsets, we say that $\mathcal{B}$ *picks out* a certain subset from $\{x_1, \ldots, x_n\}$ if this can be formed as a set of the form $B \cap \{x_1, \ldots, x_n\}$ for a $B \in \mathcal{B}$. The collection $\mathcal{B}$ *shatters* $\{x_1, \ldots, x_n\}$ if each of its 2^n subsets can be picked out in this manner. From Definition 12 we see that the VC dimension of the class $\mathcal{B}$ is the smallest n for which no set of size n is shattered by $\mathcal{B}$, that is,

$$\mathrm{VC}(\mathcal{B}) = \inf\Big\{n : \max_{x_1, \ldots, x_n} \Delta^{\mathcal{B}}(\{x_1, \ldots, x_n\}) \leq 2^n\Big\},$$

where $\Delta^{\mathcal{B}}(\{x_1,\ldots,x_n\}) = \#\{B \cap \{x_1,\ldots,x_n\} : B \in \mathcal{B}\}$. Clearly, the more refined $\mathcal{B}$ is, the larger its index. Let us recall the definition of the covering number in van der Vaart and Wellner (1996).

Definition 13 (Covering Number) *Let $(\mathcal{X}, d)$ be a metric space and $A \subset \mathcal{X}$. For $\varepsilon > 0$, the ε-covering number of A is denoted as*

$$\mathcal{N}(A, d, \varepsilon) := \min\Big\{ n \geq 1 : \exists x_1,\ldots,x_n \in \mathcal{X} \text{ such that } A \subset \bigcup_{i=1}^{n} B(x_i, \varepsilon) \Big\},$$

where $B(x,\varepsilon) := \{x' \in \mathcal{X} : d(x,x') \leq \varepsilon\}$.

The following Lemma, which is needed in the proof of Lemma 15, provides the covering number of the indicator functions on the collection of balls in $\mathbb{R}^d$, see also Lemma 25 in Hang et al. (2022).

Lemma 14 *Let $\mathcal{B} := \{B(x,r) : x \in \mathbb{R}^d, r > 0\}$ and $\mathbf{1}_{\mathcal{B}} := \{\mathbf{1}_B : B \in \mathcal{B}\}$. Then for any $\varepsilon \in (0,1)$, there exists a universal constant C such that*

$$\mathcal{N}(\mathbf{1}_{\mathcal{B}}, \|\cdot\|_{L_1(Q)}, \varepsilon) \leq C(d+2)(4e)^{d+2}\varepsilon^{-(d+1)}$$

holds for any probability measure Q.

To prove Proposition 4, we need the following lemma, which provides the uniform bound on the distance between any point and its k-th nearest neighbor with high probability.

Lemma 15 *Let Assumption 1 hold. Let $R_k(x) := \|X_{(k)}(x) - x\|$ be the distance from x to its k-th nearest neighbor and $\overline{R}_k(x)$ be the population version defined by (11) for $1 \leq k \leq n$. Then for all $x \in \mathcal{X}$, if $k \geq 32(d+4)\log n$, there holds*

$$R_k(x) \asymp (k/n)^{1/d} \tag{23}$$

with probability P^n at least $1 - 2/n^2$. Moreover, we have

$$|\overline{R}_k^d(x) - R_k^d(x)| \lesssim \sqrt{k\log n}/n. \tag{24}$$

Proof [of Lemma 15] For $x \in \mathcal{X}$ and $q \in [0,1]$, we define the q-quantile diameter

$$\rho_x(q) := \inf\big\{ r : P(B(x,r)) \geq q \big\}.$$

Let us consider the set $\mathcal{B}_k^- := \big\{ B\big(x, \rho_x\big((k - 2\sqrt{\tau k})/n\big)\big) : x \in \mathcal{X} \big\} \subset \mathcal{B}$. Lemma 14 implies that for any probability Q, there holds

$$\mathcal{N}(\mathbf{1}_{\mathcal{B}_k^-}, \|\cdot\|_{L_1(\mathrm{Q})}, \varepsilon) \leq \mathcal{N}(\mathbf{1}_{\mathcal{B}}, \|\cdot\|_{L_1(\mathrm{Q})}, \varepsilon) \leq C(d+2)(4e)^{d+2}\varepsilon^{-(d+1)}. \tag{25}$$

By the definition of the covering number, there exists an ε-net $\{A_j^-\}_{j=1}^J \subset \mathcal{B}_k^-$ with $J := \lfloor C(d+2)(4e)^{d+2}\varepsilon^{-(d+1)} \rfloor$ and for any $x \in \mathcal{X}$, there exists some $j \in \{1,\ldots,J\}$ such that

$$\big\| \mathbf{1}\big\{ B\big(x, \rho_x\big((k - 2\sqrt{\tau k})/n\big)\big)\big\} - \mathbf{1}_{A_j^-} \big\|_{L_1(D)} \leq \varepsilon. \tag{26}$$

For any $i = 1, \ldots, n$, let the random variables ξ_i be defined by $\xi_i = \mathbf{1}_{A_j^-}(X_i) - (k - 2\sqrt{\tau k})/n$. Then we have $\mathbb{E}_P \xi_i = 0$, $\|\xi_i\|_\infty \le 1$, and $\mathbb{E}_P \xi_i^2 \le \mathbb{E}_P \xi_i = (k - 2\sqrt{\tau k})/n$. Applying Bernstein's inequality in Lemma 10, we obtain

$$\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{A_j^-}(X_i) - (k - 2\sqrt{\tau k})/n \ge -\sqrt{2\tau(k - 2\sqrt{\tau k})}/n - 2\tau/(3n)$$

with probability P^n at least $1 - e^{-\tau}$. Then the union bound together with the covering number estimate (25) implies that for any A_j^-, $j = 1, \cdots, J$, there holds

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{A_j^-}(X_i) - (k - 2\sqrt{(\tau + \log J)k})/n \\ &\ge -\sqrt{2(\tau + \log J)\big(k - 2\sqrt{(\tau + \log J)k}\big)}/n - 2(\tau + \log J)/(3n).\end{aligned}$$

This together with (25) yields that for all $x \in \mathcal{X}$, there holds

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i \in \rho_x\big((k - 2\sqrt{\tau k}/n)\big)\} - (k - 2\sqrt{(\tau + \log J)k})/n \\ &\ge -\sqrt{2(\tau + \log J)\big(k - 2\sqrt{(\tau + \log J)k}\big)}/n - 2(\tau + \log J)/(3n) - \varepsilon.\end{aligned}$$

Now, if we take $\varepsilon = 1/n$, then for any $n > (4e) \vee (d+2) \vee C$, there holds $\log J = \log C + \log(d+2) + (d+2)\log(4e) + (d+1)\log n \le (2d+5)\log n$. Let $\tau := 3\log n$. A simple calculation yields that if $k \ge 32(d+4)\log n$, then we have

$$\sqrt{2(\tau + \log J)\big(k - 2\sqrt{(\tau + \log J)k}\big)}/n \le \sqrt{3(\tau + \log J)k}/n.$$

Consequently, for all $n > (4e) \vee (d+2) \vee C$, there holds

$$\sqrt{2(\tau + \log J)\big(k - 2\sqrt{(\tau + \log J)k}\big)}/n + 2(\tau + \log J)/(3n) + 1/n \le 2\sqrt{(\tau + \log J)k}/n.$$

Therefore, for all $x \in \mathcal{X}$, there holds $\frac{1}{n}\sum_{i=1}^n \mathbf{1}\big\{B\big(x, \rho_x\big((k - 2\sqrt{\tau k})/n\big)\big)\big\}(X_i) \ge k/n$ with probability P^n at least $1 - 1/n^3$. By the definition of $R_k(x)$, there holds

$$R_k(x) \ge \rho_x\big((k - 2\sqrt{\tau k})/n\big) \tag{27}$$

with probability P^n at least $1 - 1/n^3$. For any $x \in \mathcal{X}$, we have $P\big(B\big(x, \rho_x\big((k - 2\sqrt{\tau k})/n\big)\big)\big) = (k - 2\sqrt{\tau k})/n$. By Assumption 1, we have

$$P\big(B\big(x, \rho_x\big((k - 2\sqrt{\tau k})/n\big)\big)\big) = (k - 2\sqrt{\tau k})/n \le V_d \bar{c} \rho_x^d\big((k - 2\sqrt{\tau k})/n\big),$$

which yields

$$\rho_x\big((k - 2\sqrt{\tau k})/n\big) \ge \big((k - 2\sqrt{\tau k})/(V_d \bar{c} n)\big)^{1/d} \ge \big((k/(4V_d \bar{c} n)\big)^{1/d}. \tag{28}$$

Combining (27) with (28), we obtain that $R_k(x) \geq \big(k/(4V_d\bar{c}n)\big)^{1/d}$ holds for all $x \in \mathcal{X}$ with probability P^n at least $1 - 1/n^3$. Therefore, a union bound argument yields that for all $x \in \mathcal{X}$, all $k \geq 32(d+4)\log n$, and all sufficiently large n, there holds

$$R_k(x) \geq \rho_x((k - 2\sqrt{\tau k})/n) \geq \big(k/(4V_d\bar{c}n)\big)^{1/d} \tag{29}$$

with probability P^n at least $1 - 1/n^2$. This proves the first inequality of (23).

On the other hand, let us consider the set $\mathcal{B}_k^+ := \big\{B\big(x, \rho_x\big((k+2\sqrt{\tau k})/n\big)\big) : x \in \mathcal{X}\big\} \subset \mathcal{B}$. Similar to the proof of (23), we can show that for all sufficiently large n, there holds

$$R_k(x) \leq \rho_x\big((k + 2\sqrt{\tau k})/n\big) \leq \big((k + 4\sqrt{k\log n})/(\underline{c}n)\big)^{1/d} \leq \big(2k/(\underline{c}n)\big)^{1/d} \tag{30}$$

with probability P^n at least $1 - 1/n^2$.

Finally, combining (29) and (30), we get

$$\rho_x((k - 2\sqrt{\tau k})/n) \leq \rho_x(k/n) = \overline{R}_k(x) \leq \rho_x((k + 2\sqrt{\tau k})/n)$$

and consequently for all $k \geq 32(d+4)\log n$, there holds

$$|P(B(x, \overline{R}_k(x))) - P(B(x, R_k(x)))| \leq 2\sqrt{3k\log n}/n.$$

Therefore, by Assumption 1 with the condition $\mathcal{X} := [0,1]^d$, we have that for all $x \in \mathcal{X}$,

$$|\overline{R}_k^d(x) - R_k^d(x)| \leq 2^d |P(B(x, \overline{R}_k(x))) - P(B(x, R_k(x)))|/\underline{c} \leq 2^{d+1}\sqrt{3k\log n}/(n\underline{c}),$$

which proves (24). This completes the proof of Lemma 15. ■

The following Lemma is needed in the proof of Proposition 4.

Lemma 16 *Let p_i be the probability as in (5). Then we have*

$$\sum_{i=1}^{n} p_i (i/n)^\beta \leq 2(4k/s)^\beta, \qquad \beta \in (0,2] \cup \{3\}. \tag{31}$$

Moreover, if $d \geq 1$, then we have

$$\sum_{i=1}^{n} p_i (i/n)^{1/d} \geq (k/s)^{1/d}/64. \tag{32}$$

Proof [of Lemma 16] Using the substitution $z = i - k$, we get

$$\sum_{i=1}^{n} p_i i^\beta = \sum_{z=0}^{n-s} p_{z+k}(k+z)^\beta. \tag{33}$$

Define the random variable Z by $P(Z = z) = p_{z+k}$, $z = 0, \ldots, n-s$. It is easy to verify that Z follows the beta-binomial distribution with parameters $n-s$, k, and $s-k+1$ by (5). The moments of Z are $\mathbb{E}Z = (n-s)k/(s+1)$, $\mathbb{E}Z^2 = k(n-s)(n-k+kn-ks+1)/((s+1)(s+2))$,

and $\mathbb{E}Z^3 = k(n-s)[(n-s)^2(k^2+3k+2)+3k(n-s)(s-k+1)+(s-k+1)(s-2k+1)]/((s+1)(s+2)(s+3))$. We refer the reader to Johnson et al. (2005) for more discussions on this distribution.

Let us first consider the case $\beta \in (0,1]$. Since $(k/(k+z))^\beta + (z/(z+d))^\beta \geq k/(k+z) + z/(k+z) = 1$, we have $(k+z)^\beta \leq z^\beta + k^\beta$. This together with (33) yields

$$\sum_{i=1}^{n} p_i i^\beta \leq \sum_{z=0}^{n-s} p_{z+k} z^\beta + \sum_{z=0}^{n-s} p_{z+k} k^\beta = k^\beta + \sum_{z=0}^{n-s} p_{z+k} z^\beta.$$

Since the function $g(x) := x^\beta$, $\beta \in (0,1]$, is concave on $[0,\infty)$, by using Jensen's inequality, we get $\sum_{z=0}^{n-s} p_{z+k} z^\beta = \mathbb{E}Z^\beta \leq (\mathbb{E}Z)^\beta \leq (kn/s)^\beta$ and consequently

$$\sum_{i=1}^{n} p_i (i/n)^\beta \leq (k/n)^\beta + (k/s)^\beta \leq 2(4k/s)^\beta, \qquad \beta \in (0,1).$$

Next, let us consider the case $\beta \in (1,2)$ or equivalently $2-\beta \in (0,1)$. Using Hölder's inequality, we get

$$\sum_{i=1}^{n} p_i i^\beta = \sum_{i=1}^{n} (p_i i)^{2-\beta} (p_i i^2)^{\beta-1} \leq \left(\sum_{i=1}^{n} p_i i\right)^{2-\beta} \cdot \left(\sum_{i=1}^{n} p_i i^2\right)^{\beta-1}.$$

With the substitution $z = i-k$ we get $\sum_{i=1}^{n} p_i i = \sum_{z=0}^{n-s} p_{z+k}(z+k) = k + \mathbb{E}Z = k + (n-s)k/(s+1) \leq kn/s$ and $\sum_{i=1}^{n} p_i i^2 = \sum_{z=0}^{n-s} p_{z+k}(z+k)^2 \leq 2\sum_{z=0}^{n-s} p_{z+k}(z^2+k^2) \leq 2k^2 + 2\sum_{z=0}^{n-s} z^2 p_{z+k} = 2k^2 + 2\mathbb{E}Z^2 = 2k^2 + k(n-s)(n-k+kn-ks+1)/((s+1)(s+2)) \leq 4k^2n^2/s^2$. Consequently we obtain

$$\sum_{i=1}^{n} p_i i^\beta \leq (kn/s)^{2-\beta}(4k^2n^2/s^2)^{\beta-1} = 4^{\beta-1}(kn/s)^\beta \leq (4kn/s)^\beta.$$

It is easy to see that this inequality also holds when $\beta = 2$. Therefore, we have

$$\sum_{i=1}^{n} p_i (i/n)^\beta \leq (4k/s)^\beta < 2(4k/s)^\beta, \qquad \beta \in (1,2].$$

Finally, for the case $\beta = 3$, we have

$$\sum_{i=1}^{n} p_i i^3 = \sum_{z=0}^{n-s} p_{z+k}(z+k)^3 \leq \sum_{z=0}^{n-s} p_{z+k} \leq 4\sum_{z=0}^{n-s} p_{z+k}(z^3+k^3) = 4k^3 + 4\mathbb{E}Z^3 \leq 64(kn/s)^3$$

and consequently $\sum_{i=1}^{n} p_i (i/n)^3 \leq 2(4k/s)^3$, which proves (31).

Now we turn to the lower bound (32). Using Hölder's inequality, we get

$$\sum_{i=1}^{n} p_i i \leq \left(\sum_{i=1}^{n} p_i i^{1/d}\right)^{1/2} \left(\sum_{i=1}^{n} p_i i^{2-\frac{1}{d}}\right)^{1/2} \leq \left(\sum_{i=1}^{n} p_i i^{1/d}\right)^{1/2} \left(\sum_{i=1}^{n} p_i i\right)^{\frac{1}{2d}} \left(\sum_{i=1}^{n} p_i i^2\right)^{\frac{d-1}{2d}},$$

which leads to

$$\sum_{i=1}^{n} p_i i^{1/d} \geq \left(\sum_{i=1}^{n} p_i i\right)^{\frac{2d-1}{d}} \left(\sum_{i=1}^{n} p_i i^2\right)^{-\frac{d-1}{d}}.$$

With the substitution $z = i - k$ we get $\sum_{i=1}^{n} p_i i = k + \mathbb{E}Z = k + (n-s)k/(s+1) \geq (k + k(n-s)/(s+1))/2 = nk/(4s)$ and consequently

$$\sum_{i=1}^{n} p_i i^{1/d} \geq (nk/(4s))^{2-1/d}(2kn/s)^{-2+2/d} = 2^{-6+4/d}(nk/s)^{1/d} \geq (k/s)^{1/d}/64,$$

which completes the proof. ■

With the above results, we are in the position of deriving the bound for the bagging error.

Proof [of Proposition 4] By the definition of $R_k^B(x)$ and $\widetilde{R}_k^B(x)$, we have

$$\left|R_k^B(x) - \widetilde{R}_k^B(x)\right| = \left|\frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n} p_i^b R_i(x) - \sum_{i=1}^{n} p_i R_i(x)\right|.$$

For any $b = 1, \ldots, B$, define the random variables $\zeta_b(x)$ by $\zeta_b(x) := \sum_{i=1}^{n}(p_i^b - p_i)R_i(x)$. Then we have $\|\zeta_b\|_\infty \leq (\sum_{i=1}^{n} p_i^b \bigvee \sum_{i=1}^{n} p_i)R_i(x) \leq R_n(x) \leq \mathrm{diam}(\mathcal{X})$. By the definition of p_i in (17), we have $\mathbb{E}_{P_Z}(\zeta_b(x)|D_n) = 0$ and $\mathrm{Var}(\zeta_b(x)|D_n) = \mathrm{Var}(\sum_{i=1}^{n}(p_i^b - p_i)R_i(x)|D_n)$. For $1 \leq i < j \leq n$, we have $\mathrm{Cov}((p_i^b - p_i)R_i(x), (p_j^b - p_j)R_j(x)) = R_i(x)R_j(x)\mathrm{Cov}(p_i^b - p_i, p_j^b - p_j) = R_i(x)R_j(x)\mathrm{Cov}(p_i^b, p_j^b)$. By the definition of p_i^b and p_j^b, we have $p_i^b p_j^b = 0$ and thus $\mathrm{Cov}(p_i^b, p_j^b) = \mathbb{E}(p_i^b p_j^b) - \mathbb{E}p_i^b \cdot \mathbb{E}p_j^b = -p_i p_j \leq 0$, which implies $\mathrm{Cov}\big((p_i^b - p_i)R_i(x), (p_j^b - p_j)R_j(x)\big) \leq 0$ for $1 \leq i < j \leq n$. Consequently we have

$$\mathrm{Var}(\zeta_b(x)|D_n) \leq \sum_{i=1}^{n} R_i^2(x)\mathrm{Var}(p_i^b) = \sum_{i=1}^{n} R_i^2(x)p_i(1-p_i) \leq \sum_{i=1}^{n} p_i R_i^2(x). \tag{34}$$

Let $c_{d,n} := \lceil 32(d+4)\log n \rceil$. Lemma 15 implies that with probability P^n at least $1 - 2/n^2$, there holds

$$\sup_{x\in\mathcal{X}} R_i(x) \lesssim \begin{cases} (\log n/n)^{1/d}, & \text{if } 1 \leq i \leq c_{d,n}, \\ (i/n)^{1/d}, & \text{if } c_{d,n} \leq i \leq n. \end{cases}$$

Consequently we have

$$\sum_{i=1}^{n} p_i R_i^2(x) = \sum_{i=1}^{c_{d,n}} p_i R_i^2(x) + \sum_{i=c_{d,n}}^{n} p_i R_i^2(x) \lesssim \log n(\log n/n)^{2/d} + \sum_{i=1}^{n} p_i (i/n)^{2/d}.$$

Using Lemma 16, we get $\sum_{i=1}^{n} p_i R_i^2(x) \lesssim (\log n)\cdot(\log n/n)^{2/d} + (k/s)^{2/d}$. This together with (34) yields $\mathrm{Var}(\zeta_b|D_n) \lesssim (k/s)^{2/d}$. Applying Bernstein's inequality in Lemma 10, we obtain that for any $\tau > 0$, there holds

$$P_Z^B\left(\left|\frac{1}{B}\sum_{b=1}^{B}\zeta_b(x)\right| \gtrsim \sqrt{\frac{2\tau(k/s)^{2/d}}{B}} + \frac{2\tau\mathrm{diam}(\mathcal{X})}{3B}\middle| D_n\right) \leq e^{-\tau}.$$

Let $\tau := (2d+4)\log n$. Then we have

$$P_Z^B\Big(\big|R_k^B(x) - \widetilde{R}_k^B(x)\big| \lesssim \sqrt{(k/s)^{2/d}\log n/B} + \log n/B \Big| D_n\Big) \geq 1 - 1/n^{2d+4}. \tag{35}$$

In order to derive the uniform upper bound over $\mathcal{X}$, let

$$\mathcal{S} := \big\{(\sigma_1,\ldots,\sigma_n) : \text{all permutations of } (1,\ldots,n) \text{ obtainable by moving } x \in \mathbb{R}^d\big\}$$

and $\varepsilon \asymp \sqrt{(k/s)^{2/d}\log n/B} + \log n/B$. Then we have

$$\begin{aligned}
&P_Z^B\bigg(\sup_{x\in\mathbb{R}^d}\Big(\big|R_k^B(x) - \widetilde{R}_k^B(x)\big| - \varepsilon\Big) > 0 \Big| D_n\bigg) \\
&\leq P_Z^B\bigg(\bigcup_{(\sigma_1,\ldots,\sigma_n)\in\mathcal{S}}\Big|\frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n} p_{i,\sigma}^b R_{i,\sigma}(x) - \sum_{i=1}^{n} p_{i,\sigma} R_{i,\sigma}(x)\Big| > \varepsilon \Big| D_n\bigg) \\
&\leq \sum_{(\sigma_1,\ldots,\sigma_n)\in\mathcal{S}} P_Z^B\bigg(\Big|\frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n} p_{i,\sigma}^b R_{i,\sigma}(x) - \sum_{i=1}^{n} p_{i,\sigma} R_{i,\sigma}(x)\Big| > \varepsilon \Big| D_n\bigg),
\end{aligned}$$

where $p_{i,\sigma}^b := \mathbf{1}\{\|x - X_{\sigma_i}\| = R_k(x; D_b)\}$ and $R_{i,\sigma}(x) := \|x - X_{\sigma_i}\|$. For any $(\sigma_1,\ldots,\sigma_n) \in \mathcal{S}$, (35) implies

$$P_Z^B\bigg(\sup_{x\in\mathcal{X}}\Big|\frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{n} p_{i,\sigma}^b R_{i,\sigma}(x) - \sum_{i=1}^{n} p_{i,\sigma} R_{i,\sigma}(x)\Big| > \varepsilon \Big| D_n\bigg) \leq 2/n^{2d+3}.$$

This together with Lemma 11 yields that for all $n \geq 2d$, there holds

$$P_Z^B\bigg(\sup_{x\in\mathbb{R}^d}\Big(\big|R_k^B(x) - \widetilde{R}_k^B(x)\big| - \varepsilon\Big) > 0 \Big| D_n\bigg) \leq 2(25/d)^d/n^3.$$

Consequently, we obtain

$$P_Z^B \otimes P^n\Big(\big\|R_k^B - \widetilde{R}_k^B\big\|_\infty \lesssim \sqrt{(k/s)^{2/d}\log n/B} + \log n/B\Big) \geq 1 - 1/n^2,$$

which completes the proof. ■

7.2.2 Proofs Related to Section 4.1.2

In this section, we present the proof of the upper bound for the estimation error.

Proof [of Proposition 5] Let $c_{d,n} := \lceil 32(d+4)\log n\rceil$. Using the triangular inequality, we get

$$\Big|\sum_{i=1}^{n} p_i\big(R_i(x) - \overline{R}_i(x)\big)\Big| \leq \sum_{i=1}^{n} p_i\big|R_i(x) - \overline{R}_i(x)\big|$$

$$= \sum_{i=c_{d,n}}^{n} p_i \big| R_i(x) - \overline{R}_i(x) \big| + \sum_{i=1}^{c_{d,n}-1} p_i \big| R_i(x) - \overline{R}_i(x) \big|. \tag{36}$$

Let us consider the first term of (36). Lemma 15 implies that for all $x \in \mathbb{R}^d$ and $i \geq 32(d+4)\log n$, with probability P^n at least $1 - 2/n^2$, there holds

$$\big| \overline{R}_i^d(x) - R_i^d(x) \big| \lesssim \sqrt{i \log n}/(n V_d \underline{c}) \tag{37}$$

and

$$R_i(x) \gtrsim (i/n)^{1/d}. \tag{38}$$

By Assumption 1, we have $i/n = P(B(x, \overline{R}_i(x))) \leq \bar{c} V_d \overline{R}_i(x)^d$ and consequently $\overline{R}_i(x) \geq (i/(\bar{c} V_d n))^{1/d}$. This together with (38) yields

$$\sum_{j=0}^{d-1} R_i^j(x) \overline{R}_i^{d-1-j}(x) \geq \sum_{j=0}^{d-1} (i/n)^{j/d} \cdot (i/(\bar{c} V_d n))^{(d-1-j)/d} \gtrsim (i/n))^{1-1/d}. \tag{39}$$

Combining (37) and (39), we obtain

$$\big| R_i(x) - \overline{R}_i(x) \big| = \frac{\big| \overline{R}_i^d(x) - R_i^d(x) \big|}{\sum_{j=0}^{d-1} R_i^j(x) \overline{R}_i^{d-1-j}(x)} \lesssim i^{1/d-1/2} n^{-1/d} (\log n)^{1/2}.$$

Consequently, we have

$$\begin{aligned} \sum_{i=c_{d,n}}^{n} p_i \big| R_i(x) - \overline{R}_i(x) \big| &\lesssim \sum_{i=c_{d,n}}^{n} p_i i^{1/d-1/2} n^{-1/d} (\log n)^{1/2} \\ &\lesssim n^{-1/d} (\log n)^{1/2} \sum_{i=1}^{n} p_i i^{1/d-1/2} \lesssim (k/s)^{1/d-1/2} (\log n/n)^{1/2}. \end{aligned} \tag{40}$$

Next, let us consider the second term on the right-hand side of (36). Notice that (5) tells us that $p_i = 0$ for $i \leq k$. Therefore, we have

$$\sum_{i=1}^{c_{d,n}-1} p_i \big| R_i(x) - \overline{R}_i(x) \big| = 0. \tag{41}$$

Combining (36), (40), (41), we obtain

$$\left| \sum_{i=1}^{n} p_i \big(R_i(x) - \overline{R}_i(x) \big) \right| \lesssim (k/s)^{1/d-1/2} (\log n/n)^{1/2}$$

for all $x \in \mathcal{X}$, which completes the proof. ■

7.2.3 Proofs Related to Section 4.1.3

In this section, we present the proof of the upper bound for the approximation error.

Proof [of Proposition 6] By Assumption 1, we have that for all $x \in \mathcal{X}$,

$$\left|\overline{R}_i^d(x) - \frac{i/n}{V_d f(x)}\right| = \left|\frac{i/n - V_d f(x)\overline{R}_i^d(x)}{V_d f(x)}\right| \leq \left|\frac{i/n - V_d f(x)\overline{R}_i^d(x)}{V_d \underline{c}}\right|. \tag{42}$$

By the definition of $\overline{R}_i(x)$ and the Hölder continuity in Assumption 1, we have

$$\begin{aligned}\left|i/n - V_d f(x)\overline{R}_i^d(x)\right| &= \left|\int_{B(x,\overline{R}_i(x))} f(x')\,dx' - \int_{B(x,\overline{R}_i(x))} f(x)\,dx'\right| \\ &\leq \int_{B(x,\overline{R}_i(x))} |f(x') - f(x)|\,dx' \leq c_L \int_{B(x,\overline{R}_i(x))} \|x' - x\|^\alpha\,dx' \leq c_d c_L \overline{R}_i^{d+\alpha}(x), \end{aligned} \tag{43}$$

where c_d is a constant depending only on d. Moreover, by Assumption 1 and the definition of $\overline{R}_i(x)$, we have $\underline{c}V_d\overline{R}_i^d(x)/2^d \leq P(B(x,\overline{R}_i(x))) = i/n \leq V_d\overline{c}\overline{R}_i^d(x)$ and consequently

$$((i/n)/(V_d\overline{c}))^{1/d} \leq \overline{R}_i(x) \leq 2((i/n)/(\underline{c}V_d))^{1/d}. \tag{44}$$

Combining (44) and (43), we get $|i/n - V_d f(x)\overline{R}_i^d(x)| \leq 2^{d+\alpha} c_d c_L ((i/n)/(\underline{c}V_d))^{(d+\alpha)/d}$. This together with (42) yields $|\overline{R}_i^d(x) - (i/n)/(V_d f(x))| \leq (2^{d+\alpha} c_d c_L/(V_d\underline{c}))\cdot((i/n)/(\underline{c}V_d))^{(d+\alpha)/d}$. The first inequality of (44) implies

$$\begin{aligned}&\sum_{j=0}^{d} \overline{R}_i(x)^j ((i/n)/(V_d f(x)))^{(d-i-j)/d} \\ &\geq \sum_{j=0}^{d} ((i/n)/(V_d\overline{c}))^{j/d} ((i/n)/(V_d\underline{c}))^{(d-1-j)/d} \geq ((i/n)/(V_d\overline{c}))^{(d-1)/d}.\end{aligned}$$

Using the equality $x^d - y^d = (x-y)(\sum_{i=0}^{d-1} x^i \cdot y^{d-1-i})$, we get

$$\left|\overline{R}_i(x) - ((i/n)/(V_d f(x)))^{1/d}\right| = \frac{\left|\overline{R}_i^d(x) - (i/n)/(V_d f(x))\right|}{\sum_{j=0}^{d} \overline{R}_i(x)^j ((i/n)/(V_d f(x)))^{(d-1-j)/d}} \lesssim (i/n)^{(1+\alpha)/d}$$

and consequently

$$\begin{aligned}&\left|\sum_{i=1}^{n} p_i \overline{R}_i(x) - \sum_{i=1}^{n} p_i ((i/n)/(V_d f(x)))^{1/d}\right| \\ &\lesssim \sum_{i=1}^{n} p_i \left|\overline{R}_i(x) - \sum_{i=1}^{n} p_i ((i/n)/(V_d f(x)))^{1/d}\right| \lesssim \sum_{i=1}^{n} p_i (i/n)^{(1+\alpha)/d}.\end{aligned}$$

Lemma 16 implies $\sum_{i=1}^{n} p_i (i/n)^{(1+\alpha)/d} \lesssim (k/s)^{(1+\alpha)/d}$ and thus we have

$$\left|\sum_{i=1}^{d} p_i \overline{R}_i(x) - \sum_{i=1}^{n} p_i ((i/n)/(V_d f(x)))^{1/d}\right| \lesssim (k/s)^{(1+\alpha)/d},$$

which completes the proof. ■

7.2.4 Proofs Related to Section 4.1.4

The following Lemma, which is needed in the proof of Proposition 7, bounds the difference between the bagged k-distance and its infinite version.

Lemma 17 *Let Assumption 1 hold. Furthermore, let $R_k^B(x)$ and p_i be as in (3) and (5), respectively. Then for all $x \in \mathbb{R}^d$, with probability $P_Z^B \otimes P^n$ at least $1 - 3/n^2$, there holds*

$$\begin{aligned}&\left|R_k^B(x) - \sum_{i=1}^n p_i((i/n)/(V_d f(x)))^{1/d}\right| \\ &\lesssim \sqrt{(k/s)^{2/d}\log n/B} + \log n/B + (k/s)^{1/d-1/2}(\log n/n)^{1/2} + (k/s)^{(1+\alpha)/d}.\end{aligned}$$

Proof [of Lemma 17] Using the triangle inequality, we get

$$\begin{aligned}&\left|R_k^B(x) - \sum_{i=1}^n p_i((i/n)/(V_d f(x)))^{1/d}\right| \\ &\le \left|R_k^B(x) - \widetilde{R}_k^B(x)\right| + \left|\sum_{i=1}^n p_i\big(R_i(x) - \overline{R}_i(x)\big)\right| + \left|\sum_{i=1}^n p_i\big(\overline{R}_i(x) - ((i/n)/(V_d f(x)))^{1/d}\big)\right|.\end{aligned}$$

Then Propositions 4, 5, and 6 yield that for all $x \in \mathcal{X}$, with probability at least $1 - 3/n^2$, there holds

$$\begin{aligned}&\left|R_k^B(x) - \sum_{i=1}^n p_i((i/n)/(V_d f(x)))^{1/d}\right| \\ &\lesssim \sqrt{(k/s)^{2/d}\log n/B} + \log n/B + (k/s)^{1/d-1/2}(\log n/n)^{1/2} + (k/s)^{(1+\alpha)/d},\end{aligned}$$

which finishes the proof. ■

Now, we are in the position of presenting the proof of the convergence rates of the hypothetical density estimation.

Proof [of Proposition 7] If we choose $k_{D,n} \asymp \log n$, $s_n \asymp n^{d/(2\alpha+d)}(\log n)^{2\alpha/(2\alpha+d)}$ and $B_n \ge n^{(1+\alpha)/(2\alpha+d)}(\log n)^{(\alpha+d-1)/(2\alpha+d)}$, then by applying Lemma 17, we obtain that for all $x \in \mathcal{X}$, there holds

$$\begin{aligned}&\left|R_k^B(x) - \sum_{i=1}^n p_i((i/n)/(V_d f(x)))^{1/d}\right| \\ &\lesssim \sqrt{(k_{D,n}/s_n)^{2/d}\log n/B_n} + \log n/B_n + (k_{D,n}/s_n)^{1/d-1/2}(\log n/n)^{1/2} + (k_{D,n}/s_n)^{(1+\alpha)/d} \\ &\lesssim (\log n/n)^{\frac{1+\alpha}{2\alpha+d}} \qquad (45)\end{aligned}$$

with probability $P_Z^B \otimes P^n$ at least $1 - 3/n^2$. Therefore, for all sufficiently large n and $x \in \mathcal{X}$, we have

$$\left|R_k^B(x) - \sum_{i=1}^n p_i((i/n)/(V_d f(x)))^{1/d}\right| \le (\log n/n)^{1/(2\alpha+d)}(V_d \underline{c})^{-1/d}/128. \qquad (46)$$

Lemma 16 together with Assumption 1 yields that for all $x \in \mathcal{X}$, there hold

$$\sum_{i=1}^{n} p_i((i/n)/(V_d f(x)))^{1/d} \geq (V_d \underline{c})^{-1/d} \sum_{i=1}^{n} p_i (i/n)^{1/d} \geq (\log n/n)^{1/(2\alpha+d)} (V_d \underline{c})^{-1/d}/64 \quad (47)$$

and

$$\sum_{i=1}^{n} p_i((i/n)/(V_d f(x)))^{1/d} \leq (V_d \overline{c})^{-1/d} \sum_{i=1}^{n} p_i (i/n)^{1/d} \lesssim (\log n/n)^{1/(2\alpha+d)}. \quad (48)$$

Combining (46), (47) and (48), we find

$$R_k^B(x) \geq \sum_{i=1}^{n} p_i((i/n)/(V_d f(x)))^{1/d} - \left| R_k^B(x) - \sum_{i=1}^{n} p_i((i/n)/(V_d f(x)))^{1/d} \right| \gtrsim (\log n/n)^{\frac{1}{2\alpha+d}} \quad (49)$$

and

$$R_k^B(x) \leq \sum_{i=1}^{n} p_i((i/n)/(V_d f(x)))^{1/d} + \left| R_k^B(x) - \sum_{i=1}^{n} p_i((i/n)/(V_d f(x)))^{1/d} \right| \lesssim (\log n/n)^{\frac{1}{2\alpha+d}}. \quad (50)$$

Combining (48) and (50), we get

$$\sum_{j=0}^{d} \big(R_k^B(x)\big)^j \left(\sum_{i=1}^{n} p_i((i/n)/(V_d f(x)))^{1/d} \right)^{d-1-j} \lesssim (\log n/n)^{\frac{d-1}{2\alpha+d}}.$$

This together with (46) yields

$$\begin{aligned}
&\left| \big(R_k^B(x)\big)^d - \left(\sum_{i=1}^{n} p_i((i/n)/(V_d f(x)))^{1/d} \right)^d \right| \\
&\lesssim \left| R_k^B(x) - \sum_{i=1}^{n} p_i((i/n)/(V_d f(x)))^{1/d} \right| \cdot \sum_{j=0}^{d} \big(R_k^B(x)\big)^j \left(\sum_{i=1}^{n} p_i((i/n)/(V_d f(x)))^{1/d} \right)^{d-1-j} \\
&\lesssim (\log n/n)^{\frac{\alpha+d}{2\alpha+d}}. \quad (51)
\end{aligned}$$

Combining (49) and (51), we obtain that for all $x \in \mathcal{X}$ and all sufficiently large n, there holds

$$\begin{aligned}
\left| \frac{\big(\sum_{i=1}^{n} p_i (i/n)^{1/d}\big)^d}{V_d \big(R_k^B(x)\big)^d} - f(x) \right| &= \left| \frac{\big(\sum_{i=1}^{n} p_i (i/(V_d f(x) n))^{1/d}\big)^d - \big(R_k^B(x)\big)^d}{\big(R_k^B(x)\big)^d} \right| \cdot f(x) \\
&\lesssim (\log n/n)^{\frac{\alpha}{2\alpha+d}} \quad (52)
\end{aligned}$$

with probability $P_Z^B \otimes P^n$ at least $1 - 3/n^2$. This finishes the proof. ■

The following lemma, which will be used several times in the sequel, supplies the key to proofs of mode estimation and mode-based clustering.

Lemma 18 *Let Assumption 2 hold. Moreover, let $x \in \mathcal{M}_{r_\mathcal{M}}$ and $H(x)$ be the corresponding Hessian matrix. Then there exist two constants $c_1 \geq c_2 > 0$ such that for any $y \in \mathbb{R}^d$, there holds $-c_1\|y\|^2 \leq y^\top H(x)y \leq -c_2\|y\|^2$. Moreover, for all $1 \leq i \leq \#(\mathcal{M})$ and all $x, y \in B(X_i, r_\mathcal{M})$, we have*

$$f(y) \leq f(x) + \nabla f(x)^\top (y-x) - c_2\|y-x\|^2/2,$$
$$f(y) \geq f(x) + \nabla f(x)^\top (y-x) - c_1\|y-x\|^2/2.$$

Proof [of Lemma 18] For $x \in \mathcal{M}_{r_\mathcal{M}}$, let $\lambda_i(x)$, $1 \leq i \leq n$ be the eigenvalues of $H(x)$. By Assumption 2, f is twice continuously differentiable in $\mathcal{M}_{r_\mathcal{M}}$. Consequently, $\lambda_i(x)$ is continuous in $\mathcal{M}_{r_\mathcal{M}}$. Applying the extreme value theorem to $\lambda_i(x)$, there exist two constant c_2' and c_1' such that

$$c_1' \leq \lambda_i(x) \leq c_2', \qquad x \in \mathcal{M}_{r_\mathcal{M}}. \tag{53}$$

By Assumption 2, $H(x)$ is negative definite. Thus, we have $\lambda_i(x) < 0$ for all $x \in B(X_i, r_\mathcal{M})$, $1 \leq i \leq \#(\mathcal{M})$. This together with (53) yields

$$c_1' \leq \lambda_i(x) \leq c_2' < 0, \qquad x \in \mathcal{M}_{r_\mathcal{M}}. \tag{54}$$

Since $H(x)$ is negative definite for all $x \in \mathcal{M}_{r_\mathcal{M}}$, there exists an orthogonal matrix T such that $T^\top H(x) T = \mathrm{diag}\{\lambda_1(x), \ldots, \lambda_n(x)\}$. With $\widetilde{y} := Ty$ we then have

$$y^\top H(x) y = \widetilde{y}^\top \mathrm{diag}\{\lambda_1(x), \ldots, \lambda_n(x)\}\widetilde{y} = \sum_{i=1}^n \lambda_i(x)\widetilde{y}_i^2. \tag{55}$$

Combining (55) and (54), we obtain $c_1'\|\widetilde{y}\|^2 \leq y^\top H(x) y \leq c_2'\|\widetilde{y}\|^2$. Since $\|\widetilde{y}\|^2 = y^\top T^\top T y = \|y\|^2$, by choosing $c_1 = -c_1'$ and $c_2 = -c_2'$, we obtain

$$-c_1\|y\|^2 \leq y^\top H(x) y \leq -c_2\|y\|^2. \tag{56}$$

By Taylor's expansion, we have $f(y) = f(x) + \nabla f(x)^\top (y-x) + (y-x)^\top H(\xi)(y-x)/2$ for all $x, y \in B(m_i, r_\mathcal{M})$. This together with (56) yields $-c_1\|y-x\|^2 \leq f(y) - f(x) - \nabla f(x)^\top (y-x) \leq -c_2\|y-x\|^2$, which completes the proof. ■

The next proposition, which is need in the proof of Proposition 20, provides a tighter bound for the approximation error due to the higher order of smoothness around the modes.

Proposition 19 *Let Assumptions 1 and 2 hold. Moreover, let p_i be the probability as in (5) and $\overline{R}_i(x)$ be the quantile diameter function of x as in (11). Then for any $x \in \mathcal{M}_{r/2}$, we have*

$$\left| \sum_{i=1}^n p_i \overline{R}_i(x) - \sum_{i=1}^n p_i ((i/n)/(V_d f(x)))^{1/d} \right| \lesssim (k/s)^{3/d}.$$

Proof [of Proposition 19] Let $c_{n,r} := \lfloor (r/2)^d n \underline{c} V_d \rfloor$. Using the triangular inequality, we get

$$\left|\sum_{i=1}^{n} p_i \overline{R}_i(x) - \sum_{i=1}^{n} p_i ((i/n)/(V_d f(x)))^{1/d}\right| \le \left|\sum_{i=1}^{c_{n,r}} p_i \big(\overline{R}_i(x) - ((i/n)/(V_d f(x)))^{1/d}\big)\right| + \left|\sum_{i=c_{n,r}}^{n} p_i \big(\overline{R}_i(x) - ((i/n)/(V_d f(x)))^{1/d}\big)\right|.$$

The boundedness of f in Assumption 1 *(i)* implies that for all $x \in \mathcal{X}$, there holds

$$\big|\overline{R}_i^d(x) - (i/n)/(V_d f(x))\big| = \left|\frac{i/n - V_d f(x) \overline{R}_i^d(x)}{V_d f(x)}\right| \le \left|\frac{i/n - V_d f(x) \overline{R}_i^d(x)}{V_d \underline{c}}\right|.$$

By the definition of $\overline{R}_i(x)$, we have

$$\begin{aligned} \big|i/n - V_d f(x) \overline{R}_i^d(x)\big| &= \left|\int_{B(x,\overline{R}_i(x))} f(x')\,dx' - \int_{B(x,\overline{R}_i(x))} f(x)\,dx'\right| \\ &= \left|\int_{B(x,\overline{R}_i(x))} (f(x') - f(x))\,dx'\right|. \end{aligned} \tag{57}$$

By Assumption 1, we have $\underline{c} V_d \overline{R}_i^d(x)/2^d \le P(B(x, \overline{R}_i(x))) = i/n \le V_d \overline{c} \overline{R}_i^d(x)$ for all $x \in \mathcal{X}$, which yields

$$((i/n)/(V_d \overline{c}))^{1/d} \le \overline{R}_i(x) \le 2((i/n)/(\underline{c} V_d))^{1/d}, \qquad \forall x \in \mathcal{X}. \tag{58}$$

If $i \le c_{n,r}$, then we have $\overline{R}_i(x) \le r/2$. Consequently, for all $x \in \mathcal{M}_{r/2}$ and $x' \in B(x, \overline{R}_i(x))$, there exists an $m_i \in \mathcal{M}$ such that $\|x' - m_i\| \le \|x' - x\| + \|x - m_i\| \le r$. Therefore, we have $x' \in \mathcal{M}_r$. Using Taylor's expansion, we get

$$f(x') = f(x) + \nabla f(x)^\top (x' - x) + (x' - x)^\top H(x_\xi)(x' - x)/2.$$

Then Lemma 18 implies

$$\begin{aligned} \left|\int_{B(x,\overline{R}_i(x))} \big(f(x') - f(x)\big)\,dx'\right| &= \left|\int_{B(x,\overline{R}_i(x))} \big(\nabla f(x)^\top (x' - x) + (x' - x)^\top H(x_\xi)(x' - x)/2\big)\,dx'\right| \\ &= \left|\int_{B(x,\overline{R}_i(x))} (x' - x)^\top H(x_\xi)(x' - x)/2\,dx'\right| \le c_1 \int_{B(x,\overline{R}_i(x))} \|x' - x\|^2\,dx' \\ &\lesssim c_d \overline{R}_i^{d+2}(x) \lesssim (i/n)^{1+2/d}. \end{aligned} \tag{59}$$

This together with (57) yields that $\big|i/n - V_d f(x) \overline{R}_i^d(x)\big| \lesssim (i/n)^{1+2/d}$ holds for all $i \le c_{n,r}$ and consequently

$$\big|i/n - V_d f(x) \overline{R}_i^d(x)\big|/(V_d \underline{c}) \lesssim (i/n)^{1+2/d}, \qquad i \le c_{n,r}. \tag{60}$$

The first inequality of (58) implies

$$\sum_{j=0}^{d} \overline{R}_i(x)^j ((i/n)/(V_d f(x)))^{(d-i-j)/d}$$

$$\geq \sum_{j=0}^{d}((i/n)/(V_d\bar{c}))^{j/d}((i/n)/(V_d\underline{c}))^{(d-1-j)/d} \geq ((i/n)/(V_d\bar{c}))^{(d-1)/d}. \tag{61}$$

This together with (60) yields

$$\left|\overline{R}_i(x) - ((i/n)/(V_d f(x)))^{1/d}\right| \lesssim (i/n)^{3/d}, \qquad i \leq c_{n,r}, \tag{62}$$

where we used the equality $x^d - y^d = (x-y)(\sum_{i=0}^{d-1} x^i \cdot y^{d-1-i})$.

On the other hand, the Hölder continuity in Assumption 1 implies

$$\begin{aligned}\left|i/n - V_d f(x)\overline{R}_i^d(x)\right| &\leq \int_{B(x,\overline{R}_i(x))} |f(x') - f(x)|\, dx' \\ &\leq c_L \int_{B(x,\overline{R}_i(x))} \|x' - x\|^\alpha\, dx' \lesssim \overline{R}_i^{d+\alpha}(x) \lesssim (i/n)^{(\alpha+d)/d},\end{aligned}$$

where the last inequality follows from (58). This together with (61) yields

$$\left|\overline{R}_i(x) - ((i/n)/(V_d f(x)))^{1/d}\right| \lesssim (i/n)^{(\alpha+1)/d}, \qquad i > c_{n,r}, \tag{63}$$

where we use the equality $x^d - y^d = (x-y)(\sum_{i=0}^{d-1} x^i y^{d-1-i})$. Combining (62) and (63), we obtain that for all $x \in \mathcal{M}_{r/2}$, there holds

$$\begin{aligned}&\left|\sum_{i=1}^{n} p_i\overline{R}_i(x) - \sum_{i=1}^{n} p_i((i/n)/(V_d f(x)))^{1/d}\right| \leq \sum_{i=1}^{n} p_i|\overline{R}_i(x) - (i/(nV_d f(x)))^{1/d}| \\ &\lesssim \sum_{i=1}^{c_{n,r}} p_i(i/n)^{3/d} + \sum_{i=c_{n,r}+1}^{n} p_i(i/n)^{(1+\alpha)/d} \lesssim n^{-3/d}\sum_{i=1}^{c_{n,r}} p_i i^{3/d} + n^{-(1+\alpha)/d}\sum_{i=c_{n,r}+1}^{n} p_i i^{(1+\alpha)/d} \\ &\lesssim n^{-3/d}\sum_{i=1}^{n} p_i i^{3/d} + n^{-(1+\alpha)/d} c_{n,r}^{(\alpha-2)/d}\sum_{i=1}^{n} p_i i^{3/d} \leq n^{-3/d}(kn/s)^{3/d} = (k/s)^{3/d},\end{aligned}$$

which completes the proof. ■

The next proposition, which is needed in the proof of Proposition 20, presents the error between the bagged k-distance and its infinite version around the modes. This result is in fact an improvement of Proposition 17.

Proposition 20 *Let Assumptions 1 and 2 hold. Furthermore, let $R_k^B(x)$ and p_i be defined in (3) and (5), respectively. Then for all $x \in \mathcal{M}_{r/2}$, there holds*

$$\begin{aligned}&\left|R_k^B(x) - \sum_{i=1}^{n} p_i((i/n)/(V_d f(x)))^{1/d}\right| \\ &\lesssim \sqrt{(k/s)^{2/d}\log n/B} + \log n/B + (k/s)^{1/d-1/2}(\log n/n)^{1/2} + (k/s)^{3/d}\end{aligned}$$

with probability $P_Z^B \otimes P^n$ at least $1 - 3/n^2$.

Proof [of Proposition 20] The proof is similar to that of Proposition 17 by replacing the approximation error bound with the bound in Proposition 19. Thus we omit the proof. ■

With the above results, we are able to present the proof of the convergence rates for the hypothetical density estimation around modes.

Proof [of Proposition 8] Similar to the proof of Proposition 7, we can show the desired assertion by applying Proposition 20. Therefore, we omit the proof. ■

7.3 Proofs Related to Section 3.1

In this section, we provide proofs related to mode estimation. We give details of proofs for BDMBC, whereas DMBC can be dealt with similarly. To derive the convergence rates of mode estimation for BDMBC, we first show that the hypothetical density estimation around the modes is no less than the supremum of that far away from the modes in Proposition 21, which implies that the local maximum of hypothetical density estimation is close to the modes. Then by using Bernstein's inequality in Lemma 10, we establish concentration inequality for the localized level sets in Lemma 24 and derive the distance between the empirical PLLS and the population version of PLLS. Furthermore, we show that those points which are far away from the modes have a small population PLLSs. Hence we can show that the points with lower PLLS are not included in the level sets and thus we can find cluster estimators corresponding to the modes in a subjective manner.

The next proposition, which plays a key role in the proofs related to mode estimation, is needed in the proof of Theorem 9.

Proposition 21 *Let Assumptions 1 and 2 hold. Moreover, let $f_B(x)$ be the hypothetical density estimator as in* (4)*. By choosing*

$$k_{D,n} \asymp \log n, \qquad s_n \asymp n^{\frac{d}{4+d}} (\log n)^{\frac{4}{4+d}}, \qquad B_n \geq n^{\frac{3}{4+d}} (\log n)^{\frac{d+1}{4+d}},$$

then with probability P^n at least $1 - 2/n^2$, there holds

$$\inf\{f_B(x) : x \in B(m_i, c' r_n)\} > \sup\{f_B(x) : x \in B(m_i, r_{\mathcal{M}}/2) \setminus B(m_i, r_n)\}$$

where $c' := (c_2/(2c_1))^{1/2}$ with the constants c_1 and c_2 specified as in Lemma 18.

Proof [of Proposition 21] Proposition 8 yields that there exists a constant $c > 0$ such that for all sufficiently large n, with probability P^n at least $1 - 2/n^2$, for all $x \in \mathcal{M}_{r/2}$, there holds

$$|f_B(x) - f(x)| \leq c(\log n/n)^{\frac{2}{4+d}}. \tag{64}$$

The following arguments will be made on the good event E in which (64) holds.

Let $r_n := (8c/c_2)^{1/2}(\log n/n)^{1/(4+d)}$. Then we have $r_n \leq r_{\mathcal{M}}/2$ for sufficiently large n. By Lemma 18, we have $f(m_i) - c_1\|x - m_i\|^2/2 \leq f(x) \leq f(m_i) - c_2\|x - m_i\|^2/2$ for all

$x \in B(m_i, r_{\mathcal{M}})$. Consequently, we have $\sup\{f(x) : x \in B(m_i, r_{\mathcal{M}}/2) \setminus B(m_i, r_n)\} \leq f(m_i) - c_2/2r_n^2$. This together with (64) yields that $\sup\{f_B(x) : x \in B(m_i, r_{\mathcal{M}}/2) \setminus B(m_i, r_n)\} \leq f(m_i) - c_2 r_n^2 + c(\log n/n)^{2/(4+d)}$. On the other hand, by Lemma 18, we have $\inf\{f(x) : x \in B(m_i, c'r_n)\} \geq f(m_i) - c_1(c'r_n)^2/2$. This together with (64) yields that $\inf\{f_B(x) : x \in B(m_i, c'r_n)\} \geq f(m_i) - c_1(c'r_n)^2/2 - c(\log n/n)^{2/(4+d)}$. Consequently we obtain

$$\begin{aligned}\inf\{f_B(x) : x \in B(m_i, c'r_n)\} &\geq f(m_i) - c_1(c'r_n)^2/2 - c(\log n/n)^{2/(4+d)} \\ &= f(m_i) - c_2 r_n^2/2 + c(\log n/n)^{2/(4+d)} \\ &\geq \sup\{f_B(x) : x \in B(m_i, r_{\mathcal{M}}/2) \setminus B(m_i, r_n)\},\end{aligned}$$

which completes the proof. ■

The following Lemma, which is need in the proof of Theorem 2, presents the uniform concentration bounds on the empirical mass of balls in $\mathbb{R}^d$.

Lemma 22 *Let P be a probability measure on $\mathbb{R}^d$ with a bounded Lebesgue density f and $\eta : \mathbb{R}^d \to (0, \infty)$ be the local radius parameter function. Then for all $x \in \mathbb{R}^d$ and $n \geq 1$, with probability P^n at least $1 - 2/n^2$, there holds*

$$\left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \in B(x, \eta(x))\} - P(B(x, \eta(x))) \right| \lesssim \sqrt{\|\eta\|_\infty^d \log n/n} + \log n/n.$$

Proof [of Lemma 22] Let us consider the set $\mathcal{B}_\eta := \{B(x, \eta(x)) : x \in \mathbb{R}^d\} \subset \mathcal{B}$. Lemma 14 implies that for any probability Q, there holds

$$\mathcal{N}(\mathbf{1}_{\mathcal{B}_\eta}, \|\cdot\|_{L_1(\mathrm{Q})}, \varepsilon) \leq \mathcal{N}(\mathbf{1}_{\mathcal{B}}, \|\cdot\|_{L_1(\mathrm{Q})}, \varepsilon) \leq C(d+2)(4e)^{d+2}\varepsilon^{-(d+1)}. \tag{65}$$

By the definition of the covering number, there exists an ε-net $\{A_j\}_{j=1}^J \subset \mathcal{B}_\eta$ with $J := \lfloor C(d+2)(4e)^{d+2}\varepsilon^{-(d+1)} \rfloor$ and for any $x \in \mathcal{X}$, there exists some $j \in \{1, \ldots, J\}$ such that

$$\left\| \mathbf{1}\{B(x, \eta(x))\} - \mathbf{1}_{A_j} \right\|_{L_1(D)} \leq \varepsilon. \tag{66}$$

For any $i = 1, \ldots, n$, let the random variables ξ_i be defined by $\xi_i = \mathbf{1}_{A_j}(X_i) - P(A_j)$. Then we have $\mathbb{E}_P \xi_i = 0$, $\|\xi_i\|_\infty \leq 1$, and $\mathbb{E}_P \xi_i^2 \leq P(A_j) \leq \bar{c} V_d \eta(x)^d \leq \bar{c} V_d \|\eta\|_\infty^d$. Applying Bernstein's inequality in Lemma 10, we obtain

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{A_j}(X_i) - P(A_j) \leq \sqrt{2\bar{c} V_d \|\eta\|_\infty^d \tau/n} + 2\tau/(3n)$$

with probability P^n at least $1 - e^{-\tau}$. Then the union bound together with the covering number estimate (65) implies that for any A_j, $j = 1, \cdots, J$, there holds

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{A_j}(X_i) - P(A_j) \leq \sqrt{2\bar{c} V_d \|\eta\|_\infty^d (\tau + \log J)/n} + 2(\tau + \log J)/(3n).$$

This together with (66) yields that for all $x \in \mathcal{X}$, there holds

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \in B(x, \eta(x))\} - P(B(x, \eta(x))) \\ \leq \sqrt{2\overline{c}V_d\|\eta\|_\infty^d(\tau + \log J)/n} + 2(\tau + \log J)/(3n) + \varepsilon.$$

Now, if we take $\varepsilon = 1/n$, then for any $n > (4e) \vee (d+2) \vee C$, there holds $\log J = \log C + \log(d+2) + (d+2)\log(4e) + (d+1)\log n \leq (2d+5)\log n$. Let $\tau := 2\log n$. Then we have

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \in B(x, \eta(x))\} - P(B(x, \eta(x))) \\ \leq \sqrt{2(2d+7)\overline{c}V_d\|\eta\|_\infty^d \log n/n} + 2(2d+7)\log n/(3n) + 1/n. \tag{67}$$

On the other hand, let $\xi_i' = -\xi_i$. Then we have $\mathbb{E}_P\xi_i' = 0$ and $\mathbb{E}_P\xi_i'^2 = \mathbb{E}_P\xi_i^2$. Similarly, we can show that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \in B(x, \eta(x))\} - P(B(x, \eta(x))) \\ \geq -\sqrt{2(2d+7)\overline{c}V_d\|\eta\|_\infty^d \log n/n} - 2(2d+7)\log n/(3n) - 1/n$$

holds with probability P^n at least $1 - 1/n^2$. This together with (67) yields the assertion. ■

The following Lemma, which is needed in the proof of Lemma 24, presents the covering number of the indicator functions of localized level sets.

Lemma 23 *Let P be a probability measure on $\mathbb{R}^d$ with a bounded Lebesgue density f and $\eta : \mathbb{R}^d \to (0, \infty)$ be the local radius parameter function. For $\lambda > 0$, let $\widetilde{L}_f(\lambda) := \{x \in \mathbb{R}^d : f(x) \leq \lambda\}$ be the lower level set. Moreover, let $\mathcal{B}_{\eta,L} := \{\mathbf{1}\{B(x, \eta(x)) \cap \widetilde{L}_f(\lambda)\}, x \in \mathbb{R}^d\}$ be the collection of sets. Then $\mathcal{B}_{\eta,L}$ is a uniformly bounded VC class satisfying*

$$\mathcal{N}(\mathcal{B}_{\eta,L}, L_1(D), \varepsilon) \leq W(d+3)(4e)^{d+3}(1/\varepsilon)^{d+1},$$

where $W > 0$ is a universal constant.

Proof [of Lemma 23] We first show that the collection of sets $\widetilde{\mathcal{L}}_f := \{\widetilde{L}_f(\lambda), \lambda > 0\}$ are nested with VC dimension 2 by contradiction. Suppose that $\mathrm{VC}(\widetilde{\mathcal{L}}_f) > 2$. Then there exists two distinct points $x_1, x_2 \in \mathbb{R}^d$ that can be shattered by $\widetilde{\mathcal{L}}_f$, i.e, $\widetilde{L}_f(\lambda_1) \cap \{x_1, x_2\} = x_1$ and $\widetilde{L}_f(\lambda_2) \cap \{x_1, x_2\} = x_2$ for some $\lambda_1, \lambda_2 > 0$. Consequently we have $f(x_1) \leq \lambda_1 < f(x_2)$ and $f(x_2) \leq \lambda_2 < f(x_1)$, which leads to a contradiction. Therefore, we have $\mathrm{VC}(\widetilde{\mathcal{L}}_f) = 2$.

On the other hand, for the collection of balls $\mathcal{B}_\eta := \{B(x, \eta(x)) : x \in \mathbb{R}^d\}$, Dudley (1979) shows that for any set $A \in \mathbb{R}^d$ of $d+2$ points, not all subsets of A can be formed as a set of the form $B \cap A$ for a $B \in \mathcal{B}_\eta$. In other words, $\mathcal{B}_\eta$ can not pick out all subsets from $A \in \mathbb{R}^d$ of $d+2$ points. Therefore, the collection $\mathcal{B}_\eta$ fails to shatter A. Consequently, according

to Definition 12, we have $\mathrm{VC}(\mathcal{B}_\eta) = d + 2$. By Lemma 9.7 in Kosorok (2008), we have $\mathrm{VC}(\widetilde{\mathcal{L}}_f \cap \mathcal{B}_\eta) \leq \mathrm{VC}(\widetilde{\mathcal{L}}_f) + \mathrm{VC}(\mathcal{B}_\eta) - 1 \leq d + 3$. Then our assertion follows directly from Theorem 2.6.4 in van der Vaart and Wellner (1996). ■

To prove Proposition 26, we need the following Lemma which presents the uniform concentration bounds on the empirical mass of localized levels sets.

Lemma 24 *Let P be a probability measure on $\mathbb{R}^d$ with a bounded Lebesgue density f and $\eta : \mathbb{R}^d \to (0, \infty)$ be the local radius parameter function. Moreover, for $\lambda > 0$, let $\widetilde{L}_f(\lambda) := \{x \in \mathbb{R}^d : f(x) \leq \lambda\}$ be the lower level sets. Then for all $x \in \mathbb{R}^d$, $n \geq 1$, and $\lambda > 0$, with probability P^n at least $1 - 2/n^2$, there holds*

$$\left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \in B(x, \eta(x)) \cap \widetilde{L}_f(\lambda)\} - P(B(x, \eta(x)) \cap \widetilde{L}_f(\lambda)) \right| \lesssim \sqrt{\|\eta\|_\infty^d \log n / n} + \log n / n.$$

Proof [of Lemma 24] The proof is similar to that of Lemma 22 and hence is omitted. ■

The following technical Lemma is needed in the proof of Proposition 30.

Lemma 25 *Let P be a probability measure on $\mathbb{R}^d$ and $A_i \subset \mathbb{R}^d$, $1 \leq i \leq 4$, be four sets. Then we have $|P(A \cap B) - P(C \cap D)| \leq P(A \triangle C) + P(B \triangle D)$.*

Proof [of Lemma 25] We first show that for any $x \in \mathbb{R}^d$, there holds

$$\mathbf{1}\{A \cap B\} - \mathbf{1}\{C \cap D\} \leq \mathbf{1}\{A \triangle C\} + \mathbf{1}\{B \triangle D\}. \tag{68}$$

It is clear to see that (68) holds if $\mathbf{1}\{A \cap B\} - \mathbf{1}\{C \cap D\} \leq 0$. Therefore, it remains to consider the case $\mathbf{1}\{A \cap B\} - \mathbf{1}\{C \cap D\} = 1$. In this case, we have $\mathbf{1}\{A \cap B\} = 1$ and $\mathbf{1}\{C \cap D\} = 0$, which implies that $x \in A$, $x \in B$ and $x \notin C \cap D$. Consequently, if $x \notin C$, we have $\mathbf{1}\{A \triangle C\} = 1$. On the other hand, if $x \notin D$, we have $\mathbf{1}\{B \triangle D\} = 1$. Therefore, we always have $\mathbf{1}\{A \triangle C\} + \mathbf{1}\{B \triangle D\} \geq 1$. This shows (68). Now taking expectation with respect to P on both sides of (68), we obtain

$$P(A \cap B) - P(C \cap D) \leq P(A \triangle B) + P(C \triangle D). \tag{69}$$

Using the same arguments, we can show that

$$P(C \cap D) - P(A \cap B) \leq P(A \triangle B) + P(C \triangle D). \tag{70}$$

Combing (69) and (70), we obtain the assertion. ■

The next proposition, which is needed in the proof of Theorem 9, provides the difference between the empirical PLLS w.r.t. to the k-distance and the population version.

Proposition 26 *Let Assumptions 1, 2 and 3 hold and suppose that $2\alpha\gamma \leq 4+d$. Moreover, let $p^B_{k_L}(x)$ be defined as in (7). By choosing*

$$k_{D,n} \asymp \log n, \quad s_n \asymp n^{\frac{d}{4+d}} (\log n)^{\frac{4}{4+d}}, \quad B_n \geq n^{\frac{3}{4+d}} (\log n)^{\frac{d+1}{4+d}}, \quad k_{L,n} \gtrsim n^{1-\frac{\alpha\gamma}{4+d}} (\log n)^{1+\frac{\alpha\gamma}{4+d}},$$

then for all $x \in \mathcal{X}$, with probability P^n at least $1 - 3/n^2$, there holds

$$|\widehat{p}^B_{k_L}(x) - p_{k_L}(x)| \lesssim (\log n)^{-1}.$$

Proof [of Proposition 26] Following similar analysis to (52), by choosing

$$k_{D,n} \asymp \log n, \quad s_n \asymp n^{\frac{d}{4+d}} (\log n)^{\frac{4}{4+d}}, \quad B_n \geq n^{\frac{3}{4+d}} (\log n)^{\frac{d+1}{4+d}},$$

we can show that $|f_B(x) - f(x)| \lesssim (\log n/n)^{\alpha/(4+d)}$ holds for all $x \in \mathcal{X}$ with probability $P^n \otimes P_Z^B$ at least $1 - 3/n^2$. The following arguments will be made on this event.

Let $u_n := (\log n/n)^{\alpha/(4+d)}$. Then from (7) we get $\widehat{p}^B_{k_L}(x) = \frac{1}{k_L} \sum_{i=1}^{k_L} \mathbf{1}\{f_B(X_i) \leq f_B(x)\} \leq \frac{1}{k_L} \sum_{i=1}^{k_L} \mathbf{1}\{f(X_i) \leq f(x) + 2u_n\}$ and $\widehat{p}^B_{k_L}(x) \geq \frac{1}{k_L} \sum_{i=1}^{k_L} \mathbf{1}\{f(X_i) \leq f(x) - 2u_n\}$. Write $f^+(x) := f(x) + 2u_n$, $f^-(x) := f(x) - 2u_n$, and denote

$$\widehat{p}^+(x) := \frac{1}{k_L} \sum_{i=1}^{k_L} \mathbf{1}\{f(X_i) \leq f^+(x)\} \quad \text{and} \quad \widehat{p}^-(x) := \frac{1}{k_L} \sum_{i=1}^{k_L} \mathbf{1}\{f(X_i) \leq f^-(x)\}.$$

Then we have $\widehat{p}^-(x) \leq \widehat{p}^B_{k_L}(x) \leq \widehat{p}^+(x)$ and consequently

$$|\widehat{p}^B_{k_L}(x) - p_{k_L}(x)| \leq |\widehat{p}^+(x) - p_{k_L}(x)| \vee |\widehat{p}^- - p_{k_L}(x)|. \tag{71}$$

Let us consider the first term $|\widehat{p}^+(x) - p_{k_L}(x)|$. By the definition of $p_{k_L}(x)$, for all $x \in \mathcal{X}$, we have

$$|\widehat{p}^+(x) - p_{k_L}(x)| = \left| \sum_{i=1}^{k_L} \frac{\mathbf{1}\{f(X_i) \leq f^+(x)\}}{k_L} - \frac{P(X \in \widetilde{L}_f(x) \cap B(x, \overline{R}_{k_L}(x)))}{P(X \in B(x, \overline{R}_{k_L}(x)))} \right|.$$

Since $P(y \in B(x, \overline{R}_{k_L}(x))) = k_L/n$, we have

$$\begin{aligned}
&|\widehat{p}^+(x) - p_{k_L}(x)| \\
&\leq \frac{n}{k_L} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\big\{X_i \in \widetilde{L}_f(f^+(x)) \cap B(x, R_{k_L}(x))\big\} - P\big(X \in \widetilde{L}_f(f(x)) \cap B(x, \overline{R}_{k_L}(x))\big) \right| \\
&\leq \frac{n}{k_L} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\big\{X_i \in \widetilde{L}_f(f^+(x)) \cap B(x, R_{k_L}(x))\big\} - P\big(X \in \widetilde{L}_f(f^+(x)) \cap B(x, R_{k_L}(x))\big) \right| \\
&\quad + \frac{n}{k_L} \big| P\big(y \in \widetilde{L}_f(f^+(x)) \cap B(x, R_{k_L}(x))\big) - P\big(X \in \widetilde{L}_f(f(x)) \cap B(x, \overline{R}_{k_L}(x))\big) \big| \\
&=: (I) + (II).
\end{aligned}$$

Lemma 15 yields that for all sufficiently large n, there holds $(k/(4V_d \bar{c} n))^{1/d} \leq R_k(x) \leq (2k/(\underline{c} n))^{1/d}$. For the first term (I), by applying Lemma 24, we obtain that

$$(I) \lesssim (n/k_L) \Big(\sqrt{R_{k_L}(x)^d \log n/n} + \log n/n \Big) \lesssim \sqrt{\log n/k_L} \tag{72}$$

holds with probability P^n at least $1-1/n^2$. For the second term (II), by applying Lemma 25 and Assumption 3, we get

$$\begin{aligned}(II) &\le (n/k_L)\big|P\big(\widetilde{L}_f(f^+(x))\big) - P\big(\widetilde{L}_f(f(x))\big)\big| \\ &\quad + (n/k_L)\big|P\big(B(x, R_{k_L}(x))\big) - P\big(B(x, \overline{R}_{k_L}(x))\big)\big| \\ &\le (n/k_L)\big|f^+(x) - f(x)\big|^\gamma + (\overline{c}n/k_L)|\overline{R}^d_{k_L}(x) - R^d_{k_L}(x)|.\end{aligned}$$

By applying Lemma 15 and Assumption 1, we have

$$(II) \lesssim nu_n^\gamma/k_L + \sqrt{\log n/k_L}.$$

This together with (72) yields that $|\widehat{p}^+(x) - p_{k_L}(x)| \le (I) + (II) \lesssim nu_n^\gamma/k_L + \sqrt{\log n/k_L}$. On the other hand, we can show that $|\widehat{p}^-(x) - p_{k_L}| \lesssim nu_n^\gamma/k_L + \sqrt{\log n/k_L}$ in a similar way. Thus from (71), we get

$$|\widehat{p}^B_{k_L}(x) - p_{k_L}(x)| \lesssim n(\log n/n)^{\frac{\alpha\gamma}{4+d}}/k_L + \sqrt{\log n/k_L}.$$

Since $k_L \le n$, the assumption $2\alpha\gamma \le 4+d$ yields $\sqrt{\log n/k_L} \lesssim n(\log n/n)^{\alpha\gamma/(4+d)}/k_L$ and thus the desired assertion. ■

The following Lemma, which is needed in the proof of Lemma 28, shows that the instance with PLLS equal to 1 is a mode of the density function.

Lemma 27 *Let Assumption 1 hold and $p_{k_L}(x)$ be defined by (12). If $p_{k_L}(x) = 1$ for some $k_L \in \mathbb{N}$, then we have $x \in \mathcal{M}$.*

Proof [of Lemma 27] Since $p_{k_L}(x) = P(f(y) \le f(x) | y \in B(x, \overline{R}_{k_L}(x))) = 1$, we have

$$f(y) \le f(x), \qquad y \in B(x, \overline{R}_{k_L}(x)) \setminus \mathcal{C} \tag{73}$$

with $\mathcal{C}$ of measure zero. For any $x \in \mathcal{C}$, there exists a sequence $\{x_i\}_{i=1}^n \in B(x, \overline{R}_{k_L}(x)) \setminus \mathcal{C}$ such that $\{x_i\}_{i=1}^n \to y$ and

$$f(x_i) \le f(x), \qquad i \ge 1. \tag{74}$$

By Condition (i) in Assumption 1, f is a continuous function on $B(x, \overline{R}_{k_L}(x))$. Consequently, (74) yields that $f(y) \le f(x)$ for $y \in \mathcal{C}$. This together with (73) yields that

$$f(y) \le f(x), \qquad y \in B(x, \overline{R}_{k_L}(x)).$$

Therefore, x is a mode of f. Hence we complete the proof. ■

The following lemma, which is needed in the proof of Theorem 9, shows that the PLLS can not be too large for the instance far away from the modes.

Lemma 28 *Let assumption 1 and 2 hold. Let $p_{k_L}(x)$ be defined by (12). Then there exists a constant $0 < c < 1$ such that for all $x \in \mathcal{X} \setminus \mathcal{M}_{r_\mathcal{M}}$, we have $p_{k_L}(x) \le c$.*

Proof [of Lemma 28] Let $\mathcal{M}_{r_\mathcal{M}}^\circ$ denotes the interior of $\mathcal{M}_{r_\mathcal{M}}$ and $A := \mathcal{X} \setminus \mathcal{M}_{r_\mathcal{M}}^\circ$. Then A is a compact set following from the compactness of $\mathcal{X}$. By the condition (i) in Assumption 1, f is a continuous function on $\mathcal{X}$. Thus, $p_{k_L}(x)$ is a continuous function on $\mathcal{X}$. Therefore, applying extreme value theorem to $p_{k_L}(x)$ on A, there exists an $x' \in A$, such that

$$c = p_{k_L}(x') = \max_{x \in A} p_{k_L}(x). \tag{75}$$

Suppose that $c = 1$, then by Lemma 27, we have $x' \in \mathcal{M}$, which contradicts with $x' \in A$. Therefore, we have $c < 1$ by a contradiction. This completes the proof. ■

Now, we are in the position of presenting the proof of BDMBC for mode estimation.

Proof [of Theorem 2] By Lemma 28, there exists a constant $c > 0$ such that $p_{k_L}(x) \leq c$ for all $x \in \mathcal{X} \setminus \mathcal{M}_{r_\mathcal{M}}$. The following proof will be made in the case $\lambda > (c+1)/2$. Let $c'' := \sqrt{c_2/(2c_1)} \wedge 1$ with the constants c_1 and c_2 specified as in Lemma 18 and r_n specified as in Proposition 21. Lemma 22 with $\eta(x) := c'' r_n$ and $\tau := 2 \log n$ yields that for all $1 \leq i \leq \#(\mathcal{M})$, there holds

$$\left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \in B(m_i, c'' r_n)\} - P(B(m_i, r_n)) \right| \leq \sqrt{r_n^d \log n / n} + \log n / n + 1/n \tag{76}$$

with probability at least $1 - 2/n^2$. Since $r_n \asymp (\log n / n)^{1/(d+4)}$, we have

$$P(B(m_i, c'' r_n)) \geq \underline{c} V_d (c'' r_n)^d / 2^d \gtrsim \sqrt{r_n^d \log n / n} + \log n / n + 1/n,$$

where the first inequality follows from Assumption 1 *(i)*. This together with (76) yields that $\sum_{i=1}^n \mathbf{1}\{X_i \in B(m_i, c'' r_n)\} > 0$. Consequently, $D \cap B(m_i, c'' r_n)$ is a non-empty set. In other words, there exists an $\widetilde{m}_i \in B(m_i, c'' r_n)$. Since $c'' \leq 1$, we have $\widetilde{m}_i \in B(m_i, r_n)$, which implies that $D \cap B(m_i, r_n) \neq \emptyset$. Therefore, we can pick $\widehat{m}_i$ with maximal f_B out of the finite sample $D \cap B(m_i, r_n)$, i.e.,

$$\widehat{m}_i := \operatorname*{arg\,max}_{X_i \in B(m_i, r_n)} f_B(X_i). \tag{77}$$

Next, we show that $\widehat{p}_{k_L}^{k_D}(\widehat{m}_i) = 1$. Proposition 21 implies that with probability P^n at least $1 - 2/n^2$, there holds $\inf\{f_B(x) : x \in B(m_i, c' r_n)\} > \sup\{f_B(x) : x \in B(m_i, r_\mathcal{M}) \setminus B(m_i, r_n)\}$ with c' specified in Proposition 21, which implies that $f_B(\widetilde{m}_i) \geq \inf\{f_B(x) : x \in B(m_i, c' r_n)\} \geq \sup\{f_B(x) : x \in B(m_i, r_\mathcal{M}) \setminus B(m_i, r_n)\}$. Consequently, by the definition of $\widehat{m}_i$ in (77), we have

$$f_B(\widehat{m}_i) \geq f_B(\widetilde{m}_i) > \sup\{f_B(x) : x \in B(m_i, r_\mathcal{M}) \setminus B(m_i, r_n)\}. \tag{78}$$

This together with (77) yields

$$\widehat{m}_i = \operatorname*{arg\,max}_{X_i \in B(m_i, r_\mathcal{M})} f_B(X_i). \tag{79}$$

For any $X_j \in B(\widehat{m}_i, R_{k_L}(x))$, we have $\|X_j - m_i\|_2 \le \|X_j - \widehat{m}_i\|_2 + \|\widehat{m}_i - m_i\|_2 \le R_{k_L}(x) + r_n$, where $R_{k_L}(x)$ denotes the k_L-distance. By Lemma 15, for all sufficiently large n, we have $R_{k_L}(x) \lesssim (k_L/n)^{1/d} \le r_{\mathcal{M}/2}$. Consequently, we get $\|X_j - m_i\|_2 \le r_{\mathcal{M}}$, This together with (79) implies that $f_B(X_j) \le f_B(\widehat{m}_i)$, $X_j \in B(\widehat{m}_i, R_{k_L}(x))$. Therefore, we get $\widehat{p}^B_{k_L}(\widehat{m}_i) = 1$. This implies that $\widehat{m}_i \in \widehat{\mathcal{M}}$. Moreover, we have $\|\widehat{m}_i - m_i\|_2 \le r_n \lesssim (\log n/n)^{1/(4+d)}$.

Note that $\widehat{p}^B_{k_L}(\widehat{m}_i) = 1$ implies that $\widehat{m}_i \in \widehat{D}_B(\lambda)$, where $\widehat{D}_B(\lambda)$ is defined by (8). Therefore, we can pick a cluster estimator $\widehat{C}_i$ out of $\mathcal{C}_B(\lambda)$ such that $\widehat{m}_i \in \widehat{C}_i$ for $1 \le i \le \#(\mathcal{M})$. Next, we will show that for any $1 \le i < j \le \#(\mathcal{M})$, there holds $\widehat{C}_i \ne \widehat{C}_j$ by contradiction. Suppose that there exists $1 \le i < j \le \#(\mathcal{M})$ such that $\widehat{C}_i = \widehat{C}_j$. Then the distinct mode estimations $\widehat{m}_i$ and $\widehat{m}_j$ with $\widehat{m}_i \in B(m_i, c'r_n)$ and $\widehat{m}_j \in B(m_j, c'r_n)$ are contained in the same connected components of the subgraph $G_B(\lambda)$. Consequently, there exists a sequence $X'_1 \ldots, X'_\ell \in \widehat{D}_B(\lambda)$ such that X'_i and X'_{i+1} are connected in the subgraph $G_B(\lambda)$, $1 \le i \le \ell$, where we set $x'_0 := \widehat{m}_i$ and $x'_{\ell+1} := \widehat{m}_j$. This together with Lemma 26 yields that

$$\widehat{p}^B_{k_L}(x) \le (c+1)/2 < \lambda, \quad x \in \mathcal{X} \setminus \mathcal{M}_{r_{\mathcal{M}}}.$$

for all sufficiently large n, where the last inequality follows from the choice of λ. Since $X'_i \in \widehat{D}_B(\lambda)$ for $1 \le i \le \ell$, we have

$$X'_i \in \mathcal{M}_{r_{\mathcal{M}}}, \quad 1 \le i \le \ell. \tag{80}$$

Let $v := \sup_{0 \le i \le \ell+1}\{i : X'_i \in B(m_i, r_{\mathcal{M}})\}$. Since $X'_0 = \widehat{m}_i \in B(m_i, r_{\mathcal{M}})$ and $X'_{\ell+1} = \widehat{m}_j \notin B(m_i, r_{\mathcal{M}})$, we have $0 \le v \le \ell$. From the definition of the supremum and (80), there exists $i' \ne i$ such that $X'_{v+1} \in B(m_{i'}, r_{\mathcal{M}})$. This together with $X'_v \in B(m_i, r_{\mathcal{M}})$ yields that

$$\|X'_v - X'_{v+1}\| \ge \|m_i - m_{i'}\| - 2r_{\mathcal{M}} \ge \min_{1 \le i < j \le \#(\mathcal{M})} \|m_i - m_j\| - 2r_{\mathcal{M}}. \tag{81}$$

On the other hand, since X'_v and X'_{v+1} are in the connected components of the subgraph $G_B(\lambda)$, we have $\|X'_v - X'_{v+1}\| \le R_{k_G}(X'_v) \wedge R_{k_G}(X'_{v+1})$, where $R_k(x)$ represents the k-distance of x. By Lemma 15, for all sufficiently large n, we have $R_{k_G}(X'_v) \wedge R_{k_G}(X'_{v+1}) \lesssim (k_G/n)^{1/d} \lesssim (\log n/n)^{1/d}$. Therefore, we get

$$\|X'_v - X'_{v+1}\| < \min_{1 \le i < j \le \#(\mathcal{M})} \|m_i - m_j\| - 2r_{\mathcal{M}}$$

for all sufficiently large n, which leads contradiction to (81). Consequently we have $\widehat{C}_i \ne \widehat{C}_j$ for $1 \le i < j \le \#(\mathcal{M})$. This completes the proof. ■

Next, we present the proof of DMBC for mode estimation.

Proof [of Theorem 9] By the triangle inequality, we have

$$\big|k/n - f(x)V_d R_k^d(x)\big| \le \big|k/n - P(B(x, R_k(x)))\big| + \big|P(B(x, R_k(x))) - f(x)V_d R_k^d(x)\big|. \tag{82}$$

By (24) in Lemma 15, we get $\big|k/n - P(B(x, R_k(x)))\big| \lesssim \sqrt{k \log n}/n$. Similar to the analysis of (59), we can show that $\big|P(B(x, R_k(x))) - V_d f(x) R_k^d(x)\big| \lesssim R_k^{d+2}(x)$ from Lemma 18.

This together with (23) in Lemma 15 and (82) yields $\left|k/n - f(x)V_d R_k^d(x)\right| \leq \sqrt{k \log n/n} + R_k^{d+2}(x)$. Then using (23) in Lemma 15 and choosing $k_{D,n} \asymp n^{\frac{4}{4+d}} (\log n)^{\frac{d}{4+d}}$, we get

$$\left| \frac{k}{nV_d R_k^d(x)} - f(x) \right| = \left| \frac{k/n - f(x)V_d R_k^d(x)}{V_d R_k^d(x)} \right| \leq \sqrt{\log n/k} + (k/n)^{2/d}.$$

Similar analysis to that in the proof of Theorem 9 yields the desired assertion. Thus we omit the proof of Theorem 9 here. ■

7.4 Proofs Related to Section 3.2

The following Lemma is needed in the proof of Theorem 3.

Lemma 29 *Let Assumption 1 and 3 hold. Moreover, let $p_{k_L}(x)$ be as in (12). Then for any $x, y \in \mathbb{R}^d$, we have $|p_{k_L}(x) - p_{k_L}(y)| \leq c_\gamma n \|x - y\|^{\alpha\gamma}/k_L$.*

Proof [of Lemma 29] For any $x, y \in \mathbb{R}^d$, there holds

$$\begin{aligned}
&\left|p_{k_L}(x) - p_{k_L}(y)\right| \\
&= \left| \frac{P(f(X) \leq f(x), X \in B(x, \overline{R}_{k_L}(x)))}{P(B(x, \overline{R}_{k_L}(x)))} - \frac{P(f(X) \leq f(y), X \in B(y, \overline{R}_{k_L}(y)))}{P(B(y, \overline{R}_{k_L}(y)))} \right| \\
&= (k_L/n) \left| P(f(X) \leq f(x), X \in B(x, \overline{R}_{k_L}(x))) - P(f(X) \leq f(y), X \in B(y, \overline{R}_{k_L}(y))) \right|,
\end{aligned} \tag{83}$$

where we use $P(x, \overline{R}_i(x)) = i/n$ when the density function is continuous by Assumption 1. By Lemma 25, we have

$$\begin{aligned}
&\left|P(f(X) \leq f(x), X \in B(x, \overline{R}_{k_L}(x))) - P(f(X) \leq f(y), X \in B(y, \overline{R}_{k_L}(y)))\right| \\
&\leq |P(\{X : f(X) \leq f(x)\} \triangle \{X : f(X) \leq f(y)\})| + |P(B(x, \overline{R}_{k_L}(x))) - P(B(y, \overline{R}_{k_L}(y)))| \\
&= |P(\{X : f(X) \leq f(x)\} \triangle \{X : f(X) \leq f(y)\})|,
\end{aligned}$$

where we use $P(x, \overline{R}_i(x)) = i/n$. By Assumption 1 *(ii)* and 3, we have

$$|P(\{X : f(X) \leq f(x)\} \triangle \{X : f(X) \leq f(y)\})| \leq c_\gamma |f(y) - f(x)|^\gamma \leq c_\gamma \|x - y\|^{\alpha\gamma}.$$

This together with (83) yields $|p_{k_L}(x) - p_{k_L}(y)| \leq c_\gamma n \|x - y\|^{\alpha\gamma}/k_L$, which completes the proof. ■

The next proposition, which provides the difference between the empirical PLLS and the population version w.r.t. the bagged k-distance, supplies the key to the proof of Theorem 3.

Proposition 30 *Let Assumptions 1 and 3 hold and suppose that $2\alpha\gamma \leq 2\alpha + d$. Choosing*

$$k_{D,n} \asymp \log n, \quad s_n \asymp n^{\frac{d}{2\alpha+d}} (\log n)^{\frac{2\alpha}{2\alpha+d}}, \quad B_n \geq n^{\frac{1+\alpha}{2\alpha+d}} (\log n)^{\frac{\alpha+d-1}{2\alpha+d}},$$

then with probability $P^n \otimes P_Z^B$ at least $1 - 3/n^2$, for all $x \in \mathcal{X}$, there holds

$$|\widehat{p}_{k_L}^B(x) - p_{k_L}(x)| \lesssim n (\log n/n)^{\frac{\alpha\gamma}{2\alpha+d}}/k_L.$$

Proof [of Proposition 30] The proof is similar to that of Proposition 26 and hence we omit it here. ■

Next, we present the proof of the level set estimation of BDMBC.

Proof [of Theorem 3] The desired assertion involves two directions to show from the Hausdorff metric:

$$(I) := \max\big\{d(x, L_{k_L}(\lambda)) : x \in \widehat{L}_{k_L}(\lambda)\big\}, \qquad (II) := \sup\big\{d(x, \widehat{L}_{k_L}(\lambda)) : x \in L_{k_L}(\lambda)\big\}.$$

Proposition 30 yields that with probability P^n at least $1 - 3/n^2$, for all $x \in \mathcal{X}$, there holds

$$|\widehat{p}^{B}_{k_L}(x) - p_{k_L}(x)| \lesssim n(\log n/n)^{\frac{\alpha\gamma}{2\alpha+d}}/k_L := \delta_n. \tag{84}$$

The following arguments will be made on the event that (84) holds.

For any $x \in \widehat{L}_{k_L}(\lambda)$, we have $\widehat{p}^{B}_{k_L}(x) \geq \lambda$. This together with (84) yields

$$p_{k_L}(x) \geq \lambda - \delta_n. \tag{85}$$

If $p_{k_L}(x) \geq \lambda$, i.e., $x \in L_{k_L}(\lambda)$, then we have $d(x, L_{k_L}(\lambda)) = 0$. Otherwise if $p_{k_L}(x) < \lambda$, then (85) yields $\lambda - \delta_n \leq p_{k_L}(x) < \lambda$. By Assumption 4, we then have $(I) \leq d(x, L_{k_L}(\lambda)) \leq (\delta_n/c_\beta)^{1/\beta}$.

Next, let us consider (II). We first show that for any $x \in L_{k_L}(\lambda)$, there exists some $y \in L_{k_L}(\lambda + 2\delta_n)$ such that $\|x - y\| \leq (2\delta_n)^{1/\beta}$. Indeed, if $x \in L_{k_L}(\lambda + 2\delta_n)$, then we can choose $y = x$ and we have $\|y - x\| = 0$. Otherwise if $x \notin L_{k_L}(\lambda + 2\delta_n)$, then we have $\lambda \leq p_{k_L}(x) \leq \lambda + 2\delta_n$. By Assumption 4, we have $d(x, L_{k_L}(\lambda + 2\delta_n)) \leq (2\delta_n)^{1/\beta}$. Therefore, we can choose $y \in L_{k_L}(\lambda + 2\delta_n)$ such that $\|x - y\| \leq (2\delta_n)^{1/\beta}$.

Lemma 22 with $r_n = \big((\delta_n k_L)/(c_\gamma n)\big)^{1/\alpha\gamma}$ and $\tau := 2\log n$ implies that for all $y \in \mathbb{R}^d$, there holds

$$\left|\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \in B(y, r_n)\} - P(B(y, r_n))\right| \lesssim \sqrt{r_n^d \log n/n} + \log n/n \tag{86}$$

with probability at least $1 - 1/n^2$. Assumption 1 *(ii)* together with the definition of r_n and the condition $\gamma > d/(2\alpha + d)$ yields that for all sufficiently large n, we have $r_n \leq 1$ and

$$P(B(y, h)) \geq \underline{c} V_d r_n^d \geq \sqrt{r_n^d \log n/n} + \log n/n.$$

This together with (86) yields $\sum_{i=1}^{n} \mathbf{1}\{X_i \in B(y, r_n)\} > 0$. Therefore, we can pick an $X_i \in D$ such that $\|y - X_i\| \leq r_n$. By Lemma 29, we have $p_{k_L}(X_i) \geq p_{k_L}(y) - c_\gamma n\|X_i - y\|^{\alpha\gamma}/k_L \geq \lambda + \delta_n$. This together with (84) yields that $\widehat{p}^{B}_{k_L}(X_i) \geq \lambda$, which implies that $X_i \in \widehat{L}_{k_L}(\lambda)$. By the triangular inequality, we have $\|x - X_i\| \leq \|x - y\| + \|y - X_i\| \leq (2\delta_n)^{1/\beta} + r_n$, which yields $(II) := \sup\big\{d(x, \widehat{L}_{k_L}(\lambda)) : x \in L_{k_L}(\lambda)\big\} \leq (2\delta_n)^{1/\beta} + r_n$. Therefore, we have

$$d_{\mathrm{Haus}}(\widehat{L}_{k_L}(\lambda), L_{k_L}(\lambda)) \leq (I) \vee (II) \lesssim (\log n/n)^{\frac{1}{2\alpha+d}} + (n/k_L)^{1/\beta}(\log n/n)^{\frac{\alpha\gamma}{(2\alpha+d)\beta}}.$$

By the condition $\alpha\gamma \geq \beta$ and the selection of $n^{1+(\beta-\alpha\gamma)/(2\alpha+d)}(\log n)^{(\alpha\gamma-\beta)/(2\alpha+d)} \lesssim k_{L,n} \lesssim n$, there holds $d_{\mathrm{Haus}}(\widehat{L}_{k_L}(\lambda), L_{k_L}(\lambda)) \lesssim (\log n/n)^{1/(2\alpha+d)}$. Thus, we obtain the desired assertion. ■

Appendix A. Additional Tables

Table 9: The Optimal Hyper-parameter of DMBC

Dataset	ARI				NMI				F1				ACC			
	k_D	p_{k_L}	λ	k_G	k_D	p_{k_L}	λ	k_G	k_D	p_{k_L}	λ	k_G	k_D	p_{k_L}	λ	k_G
Iris	2	0.08	0.9	20	2	0.08	0.9	20	2	0.08	0.9	20	2	0.08	0.9	20
Wine	7	0.08	0.9	10	7	0.08	0.9	10	7	0.08	0.9	10	7	0.08	0.9	10
Seeds	10	0.03	0.9	10	10	0.03	0.9	10	10	0.03	0.9	10	10	0.03	0.9	10
Banknote	2	0.05	0.05	10	2	0.05	0.05	10	2	0.05	0.05	10	2	0.05	0.05	10
HTRU2	20	0.05	0.5	20	20	0.05	0.5	20	20	0.05	0.5	20	20	0.05	0.5	20
COIL	7	0.01	0.1	3	7	0.01	0.1	3	7	0.02	0.1	3	7	0.01	0.1	3
Gisette	2	0.015	0.9	5	2	0.015	0.9	5	2	0.015	0.9	5	2	0.015	0.9	5
USPS	10	0.01	0.9	10	10	0.015	0.5	3	10	0.015	0.9	10	10	0.015	0.9	10

Table 10: The Optimal Hyper-parameter of Quickshift++

Dataset	ARI			NMI			F1			ACC		
	p_k	β	ε	p_k	β	ε	p_k	β	ε	p_k	β	ε
Iris	0.100	0.1	0.1	0.100	0.1	0.1	0.05	0.9	0.0	0.100	0.1	0.1
Wine	0.100	0.1	0.0	0.100	0.1	0.0	0.10	0.9	0.0	0.100	0.1	0.0
Seeds	0.050	0.5	0.8	0.150	0.1	0.0	0.05	0.6	0.6	0.050	0.5	0.8
Banknote	0.050	0.1	0.9	0.050	0.1	0.9	0.05	0.1	0.9	0.050	0.1	0.9
HTRU2	0.002	0.2	0.8	0.002	0.2	0.8	0.05	0.1	0.0	0.002	0.2	0.8
COIL	0.020	0.1	0.0	0.020	0.1	0.0	0.02	0.1	0.1	0.020	0.1	0.1
Gisette	0.002	0.1	0.5	0.001	0.8	0.4	0.02	0.1	0.1	0.020	0.1	0.0
USPS	0.010	0.2	0.7	0.002	0.9	0.2	0.02	0.6	0.8	0.010	0.2	0.7

References

Wilhelm Ackermann. Zum Hilbertschen aufbau der reellen zahlen. *Mathematische Annalen*, 99(1):118–133, 1928.

Amineh Amini, Hadi Saboohi, Tutut Herawan, and Teh Ying Wah. MuDi-Stream: A multi density clustering algorithm for evolving data stream. *Journal of Network and Computer Applications*, 59:370–385, 2016.

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. *The Journal of Machine Learning Research*, 17(1):1487–1514, 2016.

Tomas Barton. The clustering benchmark repository, 2015. URL `https://github.com/deric/clustering-benchmark`.

Sergei N. Bernstein. *The Theory of Probabilities*. Gastehizdat Publishing House, Moscow, 1946.

Gérard Biau and Luc Devroye. *Lectures on the Nearest Neighbor Method*. Springer, 2015.

Gérard Biau, Frédéric Cérou, and Arnaud Guyader. On the rate of convergence of the bagged nearest neighbor estimate. *The Journal of Machine Learning Research*, 11(2): 687–712, 2010.

Gérard Biau, Frédéric Chazal, David Cohen-Steiner, Luc Devroye, and Carlos Rodriguez. A weighted k-nearest neighbor density estimate for geometric inference. *Electronic Journal of Statistics*, 5:204–237, 2011.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 160–172. Springer, 2013.

Eugenio Cesario, Andrea Vinci, and Shabnam Zarin. Towards parallel multi-density clustering for urban hotspots detection. In *The 29th Euromicro International Conference on Parallel, Distributed and Network-Based Processing*, pages 245–248. IEEE, 2021.

José E Chacón. A population background for nonparametric density-based clustering. *Statistical Science*, 30(4):518–532, 2015.

José E Chacón. The modal age of statistics. *International Statistical Review*, 88(1):122–141, 2020.

Frédéric Chazal, Leonidas J Guibas, Steve Y Oudot, and Primoz Skraba. Persistence-based clustering in Riemannian manifolds. *Journal of the ACM*, 60(6):1–38, 2013.

Yen-Chi Chen. Modal regression using kernel density estimation: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(e1431):1–14, 2018.

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

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

Felipe Cucker and Ding-Xuan Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, 2007.

Sanjoy Dasgupta and Samory Kpotufe. Optimal rates for k-NN density and mode estimation. *Advances in Neural Information Processing Systems*, 27:2555–2563, 2014.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Richard M. Dudley. Balls in $\mathbb{R}^k$ do not cut all subsets of $k+2$ points. *Advances in Mathematics*, 31(3):306–308, 1979.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, volume 96, pages 226–231, 1996.

Jerome H. Friedman, Jon Louis Bentley, and Raphael Ari Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software*, 3(3):209–226, 1977.

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

Guojun Gan, Chaoqun Ma, and Jianhong Wu. *Data Clustering: Theory, Algorithms, and Applications*. Society for Industrial and Applied Mathematics, 3600 University City Science Center Philadelphia, PA, United States, 2020.

Youness Aliyari Ghassabeh and Frank Rudzicz. Modified mean shift algorithm. *IET Image Processing*, 12(12):2172–2177, 2018.

Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, 2021.

Isabelle Guyon, Steve Gunn, Asa Ben-Hur, and Gideon Dror. Result analysis of the NIPS 2003 feature selection challenge. *Advances in Neural Information Processing Systems*, 17, 2004.

Hanyuan Hang, Yuchao Cai, Hanfang Yang, and Zhouchen Lin. Under-bagging nearest neighbors for imbalanced classification. *The Journal of Machine Learning Research*, 23 (118):1–63, 2022.

John A Hartigan. *Clustering Algorithms*. John Wiley & Sons, Inc., 1975.

Alexander Hinneburg and Hans-Henning Gabriel. Denclue 2.0: Fast clustering based on kernel density estimation. In *International Symposium on Intelligent Data Analysis*, pages 70–80. Springer, 2007.

Lihua Hu, Yaoyao Nie, Jifu Zhang, and Sulan Zhang. GMC_FM: A grid and multi-density-based method for matching ancient Chinese architectural images. *Machine Vision and Applications*, 33(2):1–13, 2022.

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

Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994.

Abdellah Idrissi, Hajar Rehioui, Abdelquoddouss Laghrissi, and Sara Retal. An improvement of DENCLUE algorithm for the data clustering. In *The 5th International Conference on Information & Communication Technology and Accessibility*, pages 1–6. IEEE, 2015.

Félix Iglesias, Tanja Zseby, Daniel Ferreira, and Arthur Zimek. MDCGen: Multidimensional dataset generator for clustering. *Journal of Classification*, 36(3):599–618, 2019.

Jennifer Jang and Heinrich Jiang. DBSCAN++: Towards fast and scalable density clustering. In *International Conference on Machine Learning*, pages 3019–3029. PMLR, 2019.

Jennifer Jang and Heinrich Jiang. MeanShift++: Extremely fast mode-seeking with applications to segmentation and object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4102–4113, 2021.

Heinrich Jiang. On the consistency of quick shift. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 45–54, 2017a.

Heinrich Jiang. Density level set estimation on manifolds with DBSCAN. In *International Conference on Machine Learning*, pages 1684–1693. PMLR, 2017b.

Heinrich Jiang, Jennifer Jang, and Samory Kpotufe. Quickshift++: Provably good initializations for sample-based mean shift. In *International Conference on Machine Learning*, pages 2294–2303. PMLR, 2018.

Heinrich Jiang, Jennifer Jang, and Ofir Nachum. Robustness guarantees for density clustering. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3342–3351. PMLR, 2019.

Norman L Johnson, Samuel Kotz, and Adrienne W Kemp. *Univariate Discrete Distributions*. John Wiley & Sons, 2005.

Michael R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics. Springer, New York, 2008.

Harold W Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

Jia Li, Surajit Ray, and Bruce G. Lindsay. A nonparametric statistical approach to clustering via mode identification. *The Journal of Machine Learning Research*, 8(59):1687–1723, 2007.

Zhi Liu, Jing Liu, Xiaoyan Xiao, Hui Yuan, Xiaomei Li, Jun Chang, and Chengyun Zheng. Segmentation of white blood cells through nucleus mark watershed operations and mean shift clustering. *Sensors*, 15(9):22561–22586, 2015.

Robert J Lyon, BW Stappers, Sally Cooper, John Martin Brooke, and Joshua D Knowles. Fifty years of pulsar candidate selection: From simple filters to a new principled real-time classification approach. *Monthly Notices of the Royal Astronomical Society*, 459(1): 1104–1123, 2016.

Claudia Malzer and Marcus Baum. A hybrid approach to hierarchical density-based cluster selection. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, pages 223–228. IEEE, 2020.

Pascal Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007.

Yukio Matsumoto. *An Introduction to Morse theory*, volume 208. American Mathematical Soc., 2002.

Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205, 2017.

Vidhi Mistry, Urja Pandya, Anjana Rathwa, Himani Kachroo, and Anjali Jivani. AEDBSCAN—Adaptive Epsilon Density-Based Spatial Clustering of Applications with Noise. In *Progress in Advanced Computing and Intelligent Engineering*, pages 213–226. Springer, 2021.

Sushmita Mitra and Jay Nandy. KDDclus: A simple method for multi-density clustering. In *Proceedings of International Workshop on Soft Computing Applications and Knowledge Discovery*, pages 72–76, 2011.

James Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957.

Sameer A Nene, Shree K Nayar, and Hiroshi Murase. COIL-20: Columbia object image library. 1996.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

Eduardo Pla-Sacristán, Iván González-Díaz, Tomás Martínez-Cortés, and Fernando Díaz-de María. Finding landmarks within settled areas using hierarchical density-based clustering and meta-data from publicly available images. *Expert Systems with Applications*, 123: 315–327, 2019.

Wolfgang Polonik. Measuring mass concentrations and estimating density contour clusters-an excess mass approach. *The Annals of Statistics*, 23(3):855–881, 1995.

William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

Ramin Ranjbarzadeh and Soroush Baseri Saadi. Automated liver and tumor segmentation based on concave and convex points using fuzzy c-means and mean shift clustering. *Measurement*, 150:107086, 2020.

Hajar Rehioui, Abdellah Idrissi, Manar Abourezq, and Faouzia Zegrari. DENCLUE-IM: A new approach for big data clustering. *Procedia Computer Science*, 83:560–567, 2016.

Richard J. Samworth. Optimal weighted nearest neighbour classifiers. *The Annals of Statistics*, 40(5):2733–2763, 2012.

Amit Saxena, Mukesh Prasad, Akshansh Gupta, Neha Bharill, Om Prakash Patel, Aruna Tiwari, Meng Joo Er, Weiping Ding, and Chin-Teng Lin. A review of clustering techniques and developments. *Neurocomputing*, 267:664–681, 2017.

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

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3(Dec): 583–617, 2002.

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

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

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

Jason Xu and Kenneth Lange. Power k-means clustering. In *International Conference on Machine Learning*, pages 6921–6931. PMLR, 2019.

Puning Zhao and Lifeng Lai. Analysis of KNN density estimation. *arXiv preprint arXiv:2010.00438*, 2020.

Ye Zhu, Kai Ming Ting, and Mark J Carman. Density-ratio based clustering for discovering clusters with varying densities. *Pattern Recognition*, 60:983–997, 2016.

Ye Zhu, Kai Ming Ting, Mark J Carman, and Maia Angelova. CDF Transform-and-Shift: An effective way to deal with datasets of inhomogeneous cluster densities. *Pattern Recognition*, 117:107977, 2021.