K-Medoids Clustering of Data Sequences With Composite Distributions

Tiexing Wang, Qunwei Li, *Student Member, IEEE*, Donald J. Bucci, Yingbin Liang, *Senior Member, IEEE*, Biao Chen, *Fellow, IEEE*, and Pramod K. Varshney, *Life Fellow, IEEE*

***Abstract*—This paper studies clustering of data sequences using the k-medoids algorithm. All the data sequences are assumed to be generated from *unknown* continuous distributions, which form clusters with each cluster containing a composite set of closely located distributions (based on a certain distance metric between distributions). The maximum intracluster distance is assumed to be smaller than the minimum intercluster distance, and both values are assumed to be known. The goal is to group the data sequences together if their underlying generative distributions (which are unknown) belong to one cluster. Distribution distance metrics based k-medoids algorithms are proposed for known and unknown number of distribution clusters. Upper bounds on the error probability and convergence results in the large sample regime are also provided. It is shown that the error probability decays exponentially fast as the number of samples in each data sequence goes to infinity. The error exponent has a simple form regardless of the distance metric applied when certain conditions are satisfied. In particular, the error exponent is characterized when either the Kolmogrov–Smirnov distance or the maximum mean discrepancy are used as the distance metric. Simulation results are provided to validate the analysis.**

***Index Terms*—Kolmogorov–Smirnov distance, maximum mean discrepancy, unsupervised learning, error probability, k-medoids clustering, composite distributions.**

I. Introduction

THIS paper aims to cluster sequences generated by *unknown* continuous distributions into classes so that each class contains all the sequences generated from the same (composite) distribution cluster. By sequence, we mean a set of feature observations generated by an underlying probability distribution. Here each distribution cluster contains a set of distributions that are close to each other, whereas different clusters are assumed to be far away from each other based on a certain distance metric between distributions. To be more concrete, we assume that the maximum intra-cluster distance (or its upper bound) is smaller than the minimum inter-cluster distance (or its lower bound). This assumption is necessary for the clustering problem to be meaningful. It should be emphasized that the assumption is for distribution clusters containing underlying distributions rather than the empirical distributions corresponding to the data sequences. The problem of clustering empirical distributions is of interest in market segmentation [1], image clustering [2], [3], and meteorological parameters characterization [4]–[6], among others.

Such unsupervised learning problems have been widely studied [7], [8]. The problem is typically solved by applying either centroid-based clustering algorithms, e.g., k-means clustering [9]–[11] and k-medoids clustering [12]–[14], or connectivity-based clustering algorithms, e.g., single-linkage clustering algorithm [15], and complete-linkage clustering algorithm [16], where the data sequences are viewed as multivariate data with Euclidean distance as the distance metric. The partitions of centroid-based clustering algorithms depend on the distances between sequences and the center of each cluster, whereas the connectivity-based ones assign a sequence to a cluster based on the distances between every existing sequence in the cluster and the one to be assigned.

The centroid-based clustering algorithms usually require the knowledge of the number of clusters and they differ in how the initial centers are determined. One reasonable way is to choose a data sequence as a center if it has the largest minimum distances to all the existing centers [17]–[19]. Alternatively, all the initial centers can be randomly chosen [6]. With the number of clusters unknown, there are typically two alternative approaches for clustering. One starts with a small number of clusters, e.g., 1, which is an underestimate of the true number, and proceed to split the existing clusters until convergence [19], [20]. The authors in [20] assumed a maximum number of clusters and the threshold for clustering depended on a pre-determined significance level of the two sample Kolmogorov-Smirnov (KS) test whereas the algorithm proposed in [19] did not assume a maximum number of clusters and the threshold for clustering was a function of the intra-cluster and inter-cluster distances.

Manuscript received August 1, 2018; revised November 6, 2018 and January 8, 2019; accepted February 5, 2019. Date of publication February 25, 2019; date of current version March 11, 2019. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sotirios Chatzis. This work was supported in part by the Defense Advanced Research Projects Agency under Contract HR0011-16-C-0135 and in part by the Dynamic Data Driven Applications Systems program of AFOSR under Grant FA9550-16-1-0077. The work of Y. Liang was supported by the National Science Foundation under Grant CCF-1801855. The work of B. Chen was supported by the National Science Foundation under Grant CNS-1731237. This paper was presented in part at the IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, AB, Canada, April 15–20, 2018 and in part at 52th Annual Conference on Information Sciences and Systems, Princeton, NJ, USA, March 21–23, 2018. *(Corresponding author: Tiexing Wang.)*

T. Wang, Q. Li, B. Chen, and P. K. Varshney are with the Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY 13244 USA (e-mail: twang17@syr.edu; qli33@syr.edu; bichen@syr.edu; varshney@syr.edu).

D. J. Bucci is with the Lockheed Martin - Advanced Technology Labs, Cherry Hill, NJ 08002 USA (e-mail: Donald.J.Bucci.Jr@lmco.com).

Y. Liang is with the Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: liang.889@osu.edu).

Digital Object Identifier 10.1109/TSP.2019.2901370

1053-587X © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

Alternatively, one may start with an overestimated the number of clusters, e.g., every sequence is treated as a cluster, and proceed to merge clusters that are deemed close to each other [19]. The algorithms in [6], [17], [20] were all validated by simulation results without carrying out an analysis of the error probability.

There are some key differences between the k-means algorithm and the k-medoids algorithm. The k-means algorithm minimizes a sum of squared Euclidean distances. Meanwhile, the k-medoids algorithm assigns data sequences as centers and minimizes a sum of arbitrary distances, which makes it more robust to outliers and noise [21], [22]. Moreover, the k-means algorithm requires updating the distances between data sequences and the corresponding centroids in every iteration whereas the k-medoids algorithm only requires the pairwise distances of the data sequences, which can be computed before hand. Thus, the k-medoids algorithm outperforms the k-means algorithm in terms of computational complexity as the number of sequences increases [23].

Most prior research focused on computational complexity analysis, whereas the error probability and the performance comparison of different clustering algorithms were typically studied through simulations, e.g., [13], [14], [23], [24]. This paper attempts to theoretically analyze the error probability for the k-medoids algorithm especially in the asymptotic region. Furthermore, in contrast to previous studies, which frequently used vector norms as the distance metric, e.g., Euclidean distance, our study adopts the distance metrics between distributions for clustering in order to capture the statistical models of data sequences considered in this paper. This formulation based on a distributional distance metric is uniquely suited to the proposed clustering problem, where each data point, i.e., each data sequence, represents a probability distribution and each cluster is a collection of closely related distributions, i.e., composite hypotheses.

Various distance metrics that take the distribution properties into account can be reasonable choices, e.g., KS distance [6], [18]–[20] and maximum mean discrepancy (MMD) [25]. Our previous work [18], [19] has shown that the KS distance based k-medoids algorithms are exponentially consistent for both known and unknown number of clusters. Exponential consistency, whose definition will be formally given in Section II, implies that the error probability of clustering algorithms decays at least exponentially fast as the sample size becomes large enough. In this paper, we consider a much more general framework and instead of focusing on a specific distance metric such as KS distance in our prior work [18], [19], we consider clustering sequences generated from distributions satisfying (3a) and under any distance metric that satisfies (3c)–(3d) in Assumption 1. The rationale is the following: with any distance metric that captures the statistical model of data sequences, one is likely to observe that, for a large sample size, 1) sequences generated from the same distribution cluster are statistically close to each other, and 2) sequences generated from different distribution clusters are statistically far away from each other. Thus, if the minimum distance of different distribution clusters is greater than the maximum diameter of all the distribution clusters defined in Section II, then it becomes unlikely to make a clustering error as the sample size increases. In this paper, we develop k-mediods distribution clustering algorithms where the distances between distributions are selected to capture the underlying statistical model of the data. We analyze the error probability of the proposed algorithms, which takes the form of exponential decay as the sample size increases under Assumption 1. Both the KS distance and the MMD are examined in the present work and we show that they both satisfy Assumption 1 so that the error probability of the proposed algorithms decays exponentially fast if the KS distance or MMD is used as the distance metrics.

We note that recent studies [26]–[30] of anomaly detection problems and classification also took into account the statistical model of the data sequences. Our focus here is on the clustering problem, leading to error performance analysis that is substantially different from that in [26]–[30].

The rest of the paper is organized as follows. In Section II, the system model of the clustering problem, the preliminaries of the KS distance and MMD, and notations are introduced. The clustering algorithm given the number of clusters and the corresponding upper bound on the error probability are provided in Section III, followed by the results of the clustering algorithms with an unknown number of clusters in Section IV. The simulation results for the KS and MMD based algorithms are provided in Section V. Section VI concludes the paper.

II. System Model and Preliminaries

A. Clustering Problem

Suppose there are K distribution clusters denoted by $\mathcal{P}_k$ for $k = 1, \ldots, K$, where K is fixed. Define the intra-cluster distance of $\mathcal{P}_k$ and the inter-cluster distance between $\mathcal{P}_k$ and $\mathcal{P}_{k'}$ for $k \neq k'$ respectively as

$$\begin{aligned} d\big(\mathcal{P}_k\big) &= \sup_{p_i, p_{i'} \in \mathcal{P}_k} d\big(p_i, p_{i'}\big), \\ d\big(\mathcal{P}_k, \mathcal{P}_{k'}\big) &= \inf_{p_i \in \mathcal{P}_k, p_{i'} \in \mathcal{P}_{k'}} d\big(p_i, p_{i'}\big), \end{aligned} \tag{1}$$

where $d(\cdot, \cdot)$ is a distance metric between distributions, e.g., the KS distance or MMD defined later in (5) and (6) respectively. Then $d(\mathcal{P}_k)$ and $d(\mathcal{P}_k, \mathcal{P}_{k'})$ represent the diameter of $\mathcal{P}_k$ and the distance between $\mathcal{P}_k$ and $\mathcal{P}_{k'}$, respectively. In the following discussion, $d(\cdot, \cdot)$ is also used to denote the distance between data sequences if no ambiguity exists, e.g., the MMD statistic defined in (7). Define

$$\begin{aligned} d_L &= \max_{k=1,\ldots,K} d\big(\mathcal{P}_k\big), \\ d_H &= \min_{k \neq k'} d\big(\mathcal{P}_k, \mathcal{P}_{k'}\big), \\ \Sigma &= d_H + d_L, \\ \Delta &= d_H - d_L. \end{aligned} \tag{2}$$

Table I summarizes the notations of the generalized form of distances defined in (1) and (2) which will be used in the following discussion.

Suppose M_k data sequences are generated from the distributions in $\mathcal{P}_k$, and hence a total of $\sum_{k=1}^{K} M_k = M$ sequences are to be clustered, where $M < +\infty$. Without loss of

TABLE I
NOTATIONS

general	KS	MMD
$d(\mathcal{P}_k)$	$d_{KS}(\mathcal{P}_k)$	$\mathbb{MMD}(\mathcal{P}_k)$
$d(\mathcal{P}_k, \mathcal{P}_{k'})$	$d_{KS}(\mathcal{P}_k, \mathcal{P}_{k'})$	$\mathbb{MMD}(\mathcal{P}_k, \mathcal{P}_{k'})$
d_L	$d_{L,ks}$	$d_{L,mmd}$
d_H	$d_{H,ks}$	$d_{H,mmd}$
Δ	Δ_{ks}	Δ_{mmd}
Σ	Σ_{ks}	Σ_{mmd}

generality, assume that each sequence $\mathbf{x}_{k,j_k} = [\mathbf{x}_{k,j_k}[1], \ldots, \mathbf{x}_{k,j_k}[n]]$ consists of n independently identically distributed (i.i.d.) samples generated from $p_{k,j_k} \in \mathcal{P}_k$ for $k = 1, \ldots, K$ and $j_k \in \{1, \ldots, M_k\}$. The i-th observation in $\mathbf{x}_{k,j_k}$ is denoted by $\mathbf{x}_{k,j_k}[i] \in \mathbb{R}^m$, where $m < \infty$ and $i \in \{1, \ldots, n\}$. Note that for any fixed k, p_{k,j_k}'s are not necessarily distinct. Namely, for a fixed k, some $\mathbf{x}_{k,j_k}$'s can be generated from the same distribution. Although the following discussion assumes that all the data sequences have the same length, our analysis can be easily extended to the case with different sequence lengths by replacing n with the minimum sequence length. In order to analyze the error probability of the clustering algorithm, we introduce an assumption relating to the concentration property of the distance metrics:

Assumption 1: For any distribution clusters $\{\mathcal{P}_1, \ldots, \mathcal{P}_K\}$, any sequences $\mathbf{x}_{k,j_k} \sim p_{k,j_k}$, $\mathbf{x}_{k,j'_k} \sim p_{k,j'_k}$ and $\mathbf{x}_{k',j_{k'}} \sim p_{k',j_{k'}}$, where $k \neq k'$, and sufficiently large n, the following inequalities hold:

$$d_L < d_H, \tag{3a}$$

$$P\big(d(\mathbf{x}_{k,j_k}, \mathbf{x}_{k',j_{k'}}) \leq d_0\big) \leq a_1 e^{-bn} \ \forall d_0 \in (d_L, d_H), \tag{3b}$$

$$P\big(d(\mathbf{x}_{k,j_k}, \mathbf{x}_{k,j'_k}) > d_0\big) \leq a_2 e^{-bn} \ \forall d_0 \in (d_L, d_H), \tag{3c}$$

$$P\big(d(\mathbf{x}_{k,j_k}, \mathbf{x}_{k,j'_k}) \geq d(\mathbf{x}_{k,j_k}, \mathbf{x}_{k',j_{k'}})\big) \leq a_3 e^{-bn}, \tag{3d}$$

where a_i's are some constants independent of distributions, b (>0) is a function of d_0 and n is the sample size. □

Here (3a) guarantees that the lower bound of inter-cluster distances is greater than the upper bound of intra-cluster distances. (3b) guarantees that the probability that the distance between two sequences generated from different distribution clusters is smaller than d_H decays exponentially fast. (3c) guarantees that the probability that the distance between two sequences generated from the same distribution cluster is greater than d_L decays exponentially fast. (3d) guarantees that given two sequences generated from the same cluster and a third sequence generated from another distribution cluster, the probability that the first sequence is actually closer to the third sequence decays exponentially fast.

A clustering output is said to be incorrect if the sequences generated by different distribution clusters are assigned to the same cluster, or sequences generated by the same distribution cluster are assigned to more than one cluster. Let a trial be the event of applying a clustering algorithm to M data sequences generated by $p_{1,1}, \ldots, p_{1,M_1}, \ldots, p_{K,M_K}$. The error probability of the clustering algorithm is defined as

$$P_{e,p} = \lim_{N \to +\infty} \frac{N_e}{N},$$

where N is the number of trials of which N_e trials result in incorrect clustering outputs. The error probability of a clustering algorithm given $\mathcal{P}_1, \ldots, \mathcal{P}_K$ is defined as

$$P_{e,\mathcal{P}} = \sup_{p_{k,j_k} \in \mathcal{P}_k : k=1,\ldots,K,\ j_k=1,\ldots,M_k} P_{e,p}.$$

The error probability of a clustering algorithm given d_L and d_H is defined as

$$P_e(d_L, d_H) = \sup_{\{\mathcal{P}_1,\ldots,\mathcal{P}_K\} \subset \mathcal{P}^K : d_L,\, d_H} P_{e,\mathcal{P}},$$

where $\mathcal{P}$ is the set of all the continuous distributions on $\mathbb{R}^m$. In the following discussion, the notation $P_e(d_L, d_H)$ will be replaced by P_e for simplicity if it does not cause any ambiguity. We also note that P_e is a function of the sample size n.

A clustering algorithm is said to be *consistent* if for any $0 \leq d_L < d_H$,

$$\lim_{n \to \infty} P_e = 0.$$

The algorithm is said to be *exponentially consistent* if for any $0 \leq d_L < d_H$,

$$B = \lim_{n \to \infty} -\frac{1}{n} \log P_e > 0.$$

For the case where a clustering algorithm is exponentially consistent, we are also interested in characterizing the error exponent B.

B. Preliminaries of KS Distance

Suppose $\mathbf{x}$ is generated by the distribution p, where $\mathbf{x}[i] \in \mathbb{R}$. Then the empirical cumulative distribution function (c.d.f.) induced by $\mathbf{x}$ is given by

$$F_{\mathbf{x}}(a) = \frac{1}{n} \sum_{i=1}^{n} 1_{[-\infty, a]}(\mathbf{x}[i]), \tag{4}$$

where $1_{[-\infty,x]}(\cdot)$ is the indicator function. Let the c.d.f. of p evaluated at a be $F_p(a)$. The KS distance is defined as

$$d_{KS}(p, q) = \sup_{a \in \mathbb{R}} |F_p(a) - F_q(a)|, \tag{5}$$

where $F_p(a)$ and $F_q(a)$ can be either c.d.f's of distributions or empirical c.d.f.'s induced by sequences.

Proposition 1: The KS distance satisfies (3b)–(3d) if $d_{L,ks} < d_{H,ks}$.

Proof: See Lemmas A.5, A.3, and A.7 in Appendix A. ■

C. Preliminaries of MMD

Suppose $\mathcal{P}$ includes a class of probability distributions, and suppose $\mathcal{H}$ is the reproducing kernel Hilbert space (RKHS) associated with a kernel $g(\cdot, \cdot)$. Define a mapping from $\mathcal{P}$ to $\mathcal{H}$ such that each distribution $p \in \mathcal{P}$ is mapped into an element in $\mathcal{H}$ as follows

$$\mu_p(\cdot) = \mathbb{E}_p[g(\cdot, x)] = \int g(\cdot, x)\, dp(x),$$

where $\mu_p(\cdot)$ is referred to as the *mean embedding* of the distribution p into the Hilbert space $\mathcal{H}$. Due to the reproducing property of $\mathcal{H}$, it is clear that $\mathbb{E}_p[f] = \langle \mu_p, f \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$.

It has been shown in [31]–[34] that for many RKHSs such as those associated with Gaussian and Laplace kernels, the mean embedding is injective, i.e., each $p \in \mathcal{P}$ is mapped to a unique element $\mu_p \in \mathcal{H}$. In this way, many machine learning problems with unknown distributions can be solved by studying mean embeddings of probability distributions without actually estimating the distributions, e.g., [28], [29], [35], [36]. In order to distinguish between two distributions p and q, the authors in [37] introduced the following notion of MMD based on the mean embeddings μ_p and μ_q of p and q, respectively:

$$\mathbb{MMD}(p, q) := \|\mu_p - \mu_q\|_{\mathcal{H}}. \tag{6}$$

A biased estimator of $\mathbb{MMD}(p, q)$ based on $\mathbf{x}$ and $\mathbf{y}$ of sample lengths n and m, respectively, is given by

$$\mathbb{MMD}(\mathbf{x}, \mathbf{y}) = \left[\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} g(\mathbf{x}[i], \mathbf{x}[j]) + \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} g(\mathbf{y}[i], \mathbf{y}[j]) - \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} g(\mathbf{x}[i], \mathbf{y}[j]) \right]^{\frac{1}{2}}, \tag{7}$$

where $g(x, y)$ is the kernel function which is assumed to be bounded, i.e., $0 \le g(x, y) \le \mathbb{G} < +\infty$.

Proposition 2: The MMD statistic satisfies (3b)–(3d) if $d_{L,mmd} < d_{H,mmd}$.

Proof: See Lemmas A.6, A.4, and A.8 in Appendix A. □

D. Additional Notations

The following notations are used in the algorithms and the corresponding proofs. Let $\mathcal{C}_l^t$ be the l-th cluster obtained at the t-th cluster update step and let $\mathbf{c}_l^{t,a}$, $\mathbf{c}_l^{t,e}$ and $\mathbf{c}_l^{t,s}$ be the centers of the l-th cluster obtained by the center update step, merge step and split step of the t-th iteration respectively for $t \ge 1$. Let $\mathcal{C}_l^0$ be the l-th cluster obtained at the initialization step and $\mathbf{c}_l^{0,a}$ be the corresponding center. Let $\hat{K}^t$ ($t \ge 1$) be the number of centers before the t-th cluster update step and the t-th split step. Moreover, use $\hat{K}^0$ to denote the number of centers obtained at the center initialization step. For simplicity, all the superscripts are omitted in the following discussion when there is no ambiguity.

To further simplify the notation in the algorithms and the proofs, let $\{\mathbf{y}_i\}_{i=1}^M$ denote the data sequence set $\{\mathbf{x}_{k,j_k}\}_{k=1,j_k=1}^{K,M_k}$. However, the one-to-one mapping from $\{\mathbf{y}_i\}_{i=1}^M$ onto $\{\mathbf{x}_{k,j_k}\}_{k=1,j_k=1}^{K,M_k}$ is not fixed, i.e., given a fixed i, $\mathbf{y}_i$ can be any sequence in $\{\mathbf{x}_{k,j_k}\}_{k=1,j_k=1}^{K,M_k}$ unless other constraints are imposed. Denote by $\mathbf{y}_i \sim \mathcal{P}_k$ if $\mathbf{y}_i$ is generated from a distribution $p \in \mathcal{P}_k$. Furthermore, we define the following index set

$$I_{k_1}^{k_2} = \{k_1, \ldots, k_2\},$$

where $k_1,\ k_2 \in \mathbb{Z}^+$ and $k_1 < k_2$.

Algorithm 1: Initialization With Known K.

Input: Data sequences $\{\mathbf{y}_i\}_{i=1}^M$, number of clusters K.
Output: Partitions $\{\mathcal{C}_k\}_{k=1}^K$.
{Center initialization}
Arbitrarily choose one $\mathbf{y}_i$ as $\mathbf{c}_1$.
for $k = 2$ to K **do**
    $\mathbf{c}_k \leftarrow \arg\max_{\mathbf{y}_i} \left(\min_{l \in I_1^{k-1}} d(\mathbf{y}_i, \mathbf{c}_l) \right)$
end for
{Cluster initialization}
Set $\mathcal{C}_k \leftarrow \emptyset$ for $1 \le k \le K$.
for $i = 1$ to M **do**
    $\mathcal{C}_l \leftarrow \mathcal{C}_l \cup \{\mathbf{y}_i\}$, where $l = \arg\min_{l \in I_1^K} d(\mathbf{y}_i, \mathbf{c}_l)$
end for
Return $\{\mathcal{C}_k\}_{k=1}^K$

Algorithm 2: Clustering With Known K.

Input: Data sequences $\{\mathbf{y}_i\}_{i=1}^M$, number of clusters K.
Output: Partition set $\{\mathcal{C}_k\}_{k=1}^K$.
Initialize $\{\mathcal{C}_k\}_{k=1}^K$ by Algorithm 1.
while not converge **do**
    {Center update}
    for $k = 1$ to K **do**
        $\mathbf{c}_k \leftarrow \arg\min_{\mathbf{y}_i \in \mathcal{C}_k} \sum_{\mathbf{y}_{j'} \in \mathcal{C}_k} d(\mathbf{y}_i, \mathbf{y}_{j'})$
    end for
    {Cluster update}
    for $i = 1$ to M **do**
        if $\mathbf{y}_i \in \mathcal{C}_{k'}$ and $d(\mathbf{y}_i, \mathbf{c}_k) < d(\mathbf{y}_i, \mathbf{c}_{k'})$ **then**
            $\mathcal{C}_k \leftarrow \mathcal{C}_k \cup \{\mathbf{y}_i\}$ and $\mathcal{C}_{k'} \leftarrow \mathcal{C}_{k'} \setminus \{\mathbf{y}_i\}$.
        end if
    end for
end while
Return $\{\mathcal{C}_k\}_{k=1}^K$

III. Known Number of Clusters

In this section, we study the clustering algorithm for known K, the number of clusters. The method proposed in [17] is used for center initialization, as described in Algorithm 1. The initial K centers are chosen sequentially such that the center of the k-th cluster is the sequence that has the largest minimum distance to the previous $k-1$ centers. The clustering algorithm itself is presented in Algorithm 2. Given the centers, each sequence is assigned to the cluster for which the sequence has the minimum distance to the center. For a given cluster, a sequence is assigned as the center subsequently if the sum of its distances to all the sequences in the cluster is the smallest. The algorithm continues until the clustering result converges.

The following theorem provides the convergence guarantee for Algorithm 2 via an upper bound on the error probability.

Theorem III.1: Algorithm 2 converges after at most $\binom{M}{K} K^{(M-K)}$ iterations. Moreover, under Assumption 1, the error probability of Algorithm 2 after T iterations is upper bounded as follows

$$P_e \le M^2 (a_1 + a_2 + (T+1) a_3) e^{-bn},$$

where a_1, a_2, a_3 and b are as defined in Assumption 1 and

$$T \leq \binom{M}{K} K^{(M-K)}.$$

Proof: The idea of proving the upper bound on the error probability is as follows. We first prove that the error probability at the initialization step decays exponentially. Note that the event that an error occurs during the first T iterations is the union of the event that an error occurs at the t-th step and the previous $t-1$ iterations are correct for $t = 1, \ldots, T$. Thus, if we prove that the error probability at the t-th step *given* correct updates from the previous iterations decays exponentially, then so does the error probability of the algorithm by the union bound argument. See Appendix B1 for details. □

Theorem III.1 shows that for any given K, any distance metric satisfying Assumption 1 yields an exponentially consistent k-medoids clustering algorithm with the error exponent b.

Corollary III.1.1: Suppose the KS distance and the MMD statistic are used for Algorithms 1 and 2, then for n sufficiently large,

$$P_e^{KS} \leq M^2 (6T + 14) \exp\left(-\frac{n\Delta_{ks}^2}{8}\right),$$

$$P_e^{MMD} \leq M^2 (4T + 8) \exp\left(-\frac{n\Delta_{mmd}^2}{64\mathbb{G}}\right).$$

Proof: By Propositions 1 and 2, the upper bound on the error probability of Algorithm 2 in Theorem III.1 applies to the KS distance and the MMD statistic. Thus, the corollary is obtained by substituting the values specified in Lemmas A.3–A.8 in the upper bound. □

Corollary III.1.1, combined with the fact that T is finitely bounded for finite M and K, implies that Algorithm 2 is exponentially consistent under both the KS and MMD distance metrics with an error exponent no smaller than $\frac{\Delta_{ks}^2}{8}$ and $\frac{\Delta_{mmd}^2}{64\mathbb{G}}$, respectively.

IV. Unknown Number of Clusters

In this section, we propose the merge- and split-based algorithms for estimating the number of clusters as well as grouping the sequences.

A. Merge Step

If a distance metric satisfies (3c) and two sequences generated by distributions within the same cluster are assigned as centers, then, with high probability, the distance between the two centers is small. This is the premise of the clustering algorithm based on merging centers that are close to each other.

The proposed approach is summarized in Algorithms 3 and 4. There are two major differences between Algorithms 3 and 4 and Algorithms 1 and 2. First, the center initialization step of Algorithm 3 keeps generating an increasing number of centers until all the sequences are close to one of the existing centers. Second, an additional Merge Step in Algorithm 4 helps to combine clusters if the corresponding centers have small distances between each other.

Algorithm 3: Merge-Based Initialization With Unknown K.

Input: Data sequences $\{\mathbf{y}_i\}_{i=1}^M$ and threshold d_{th}.
Output: Partitions $\{\mathcal{C}_k\}_{k=1}^{\hat{K}}$.
{Center initialization}
Arbitrarily choose one $\mathbf{y}_i$ as $\mathbf{c}_1$ and set $\hat{K} = 1$.
while $\max_{i \in I_1^M} \left(\min_{k \in I_1^{\hat{K}}} d(\mathbf{y}_i, \mathbf{c}_k)\right) > d_{th}$ **do**
    $\mathbf{c}_{\hat{K}+1} \leftarrow \arg\max_{\mathbf{y}_i} \left(\min_{k \in I_1^{\hat{K}}} d(\mathbf{y}_i, \mathbf{c}_k)\right)$
    $\hat{K} \leftarrow \hat{K} + 1$
end while
Clustering initialization specified in Algorithm 1.
Return $\{\mathcal{C}_k\}_{k=1}^{\hat{K}}$

Algorithm 4: Merge-Based Clustering With Unknown K.

Input: Data sequences $\{\mathbf{y}_i\}_{i=1}^M$ and threshold d_{th}.
Output: Partition set $\{\mathcal{C}_k\}_{k=1}^{\hat{K}}$.
Initialize $\{\mathcal{C}_k\}_{k=1}^{\hat{K}}$ by Algorithm 3.
while not converge **do**
    Center update specified in Algorithm 2.
    {Merge Step}
    for $k_1, k_2 \in \{1, \ldots, \hat{K}\}$ and $k_1 \neq k_2$ **do**
        if $d(\mathbf{c}_{k_1}, \mathbf{c}_{k_2}) \leq d_{th}$ **then**
            if $\sum_{\mathbf{y}_i \in \mathcal{C}_{k_1}} d(\mathbf{c}_{k_2}, \mathbf{y}_i) < \sum_{\mathbf{y}_i \in \mathcal{C}_{k_2}} d(\mathbf{c}_{k_1}, \mathbf{y}_i)$ **then**
                $\mathcal{C}_{k_2} \leftarrow \mathcal{C}_{k_1} \cup \mathcal{C}_{k_2}$ and delete $\mathbf{c}_{k_1}$ and $\mathcal{C}_{k_1}$.
            else
                $\mathcal{C}_{k_1} \leftarrow \mathcal{C}_{k_1} \cup \mathcal{C}_{k_2}$ and delete $\mathbf{c}_{k_2}$ and $\mathcal{C}_{k_2}$.
            end if
            $\hat{K} \leftarrow \hat{K} - 1$.
        end if
    end for
    Cluster update specified in Algorithm 2.
end while
Return $\{\mathcal{C}_k\}_{k=1}^{\hat{K}}$

Theorem IV.1: Algorithm 4 converges after at most $T_{\max}$ iterations, where

$$T_{\max} = \sum_{\hat{K}=1}^{M} \binom{M}{\hat{K}} \hat{K}^{(M-\hat{K})}.$$

Moreover, under Assumption 1, if $d_{th} \in (d_L, d_H)$, then the error probability of Algorithm 4 after T iterations is upper bounded as follows

$$P_e \leq M^2 \left((T+1) a_1 + a_2 + (T+1) a_3\right) e^{-bn},$$

where a_1, a_2, a_3 and b are as defined in Assumption 1 and $T \leq T_{\max}$.

Proof: The proof shares the same idea as that of Theorem III.1. See Appendix VI-B2 for details. □

Algorithm 5: Split-Based Clustering With Unknown K.

Input: Data sequences $\{\mathbf{y}_i\}_{i=1}^M$ and threshold d_{th}.
Output: Partition set $\{\mathcal{C}_k\}_{k=1}^{\hat{K}}$.
$\mathcal{C}_1 = \{\mathbf{y}_i\}_{i=1}^M$, $\hat{K} = 1$ and find $\mathbf{c}_1$ by center update specified in Algorithm 2.
while not converge **do**
  {Split Step}
  if $\max_{k \in I_1^{\hat{K}},\, \mathbf{y}_i \in \mathcal{C}_k} d(\mathbf{c}_k, \mathbf{y}_i) > d_{th}$ **then**
    $\hat{K} \leftarrow \hat{K} + 1$.
    $k = \arg\max_{k \in I_1^{\hat{K}}} \left(\max_{\mathbf{y}_i \in \mathcal{C}_k} d(\mathbf{c}_k, \mathbf{y}_i)\right)$
    $\mathbf{c}_{\hat{K}} \leftarrow \arg\max_{\mathbf{y}_i \in \mathcal{C}_k} d(\mathbf{c}_k, \mathbf{y}_i)$
  end if
  Cluster update specified in Algorithm 2.
end while
Return $\{\mathcal{C}_k\}_{k=1}^{\hat{K}}$

Theorem IV.1 shows that the merge-based algorithm is exponentially consistent under any distance metric satisfying Assumption 1 with the error exponent b.

Corollary IV.1.1: Suppose the KS distance and the MMD statistic are used with $d_{th} = \frac{\Sigma_{ks}}{2}$ and $d_{th} = \frac{\Sigma_{mmd}}{2}$. Then for n sufficiently large, the error probability of Algorithm 4 after T iterations is upper bounded as follows

$$P_e^{KS} \leq M^2 (10T + 14) \exp\left(-\frac{n\Delta_{ks}^2}{8}\right),$$

$$P_e^{MMD} \leq M^2 (6T + 8) \exp\left(-\frac{n\Delta_{mmd}^2}{64\mathbb{G}}\right).$$

Proof: By Propositions 1 and 2, the upper bound on the error probability of Algorithm 4 in Theorem IV.1 applies to the KS distance and the MMD statistic. Thus, the corollary is obtained by substituting the values specified in Lemmas A.3–A.8 in the upper bound. □

Corollary IV.1.1, combined with the fact that T is finitely bounded for finite M and K, implies that Algorithm 4 is exponentially consistent under both the KS and MMD distance metrics with an error exponent no smaller than $\frac{\Delta_{ks}^2}{8}$ and $\frac{\Delta_{mmd}^2}{64\mathbb{G}}$, respectively.

B. Split Step

Suppose a cluster contains sequences generated by different distributions and the center is generated from $p \in \mathcal{P}_k$. Then if the distance metric satisfies (3b), the probability that the distances between sequences generated from distribution clusters other than $\mathcal{P}_k$ and the center is small decays as the sample size increases. Therefore, it is reasonable to begin with one cluster and then split a cluster if there exists a sequence in the cluster that has a large distance to the center. The corresponding algorithm is summarized in Algorithm 5.

Definition IV.1.1: Suppose Algorithm 5 obtains $\hat{K}$ clusters at the t-th iteration, where $\hat{K} < K$ and $\hat{K} = t$ or $t + 1$. Then the correct clustering update result is that each cluster contains all the sequences generated from the distribution cluster that generates the center.

Theorem IV.2: Algorithm 5 converges after at most M iterations. Moreover, under Assumption 1, if $d_{th} \in (d_L, d_H)$, then the error probability of Algorithm 5 after T iterations is upper bounded as follows

$$P_e \leq M^2 T (a_1 + a_2 + a_3) e^{-bn},$$

where a_1, a_2, a_3 and b are as defined in Assumption 1 and $T \leq M$.

Proof: An error occurs at the t-th iteration if and only if the $\hat{K}$-th center is generated from distribution clusters that generated the previous centers or the clustering result is incorrect. Note that the error event of the first T iterations is the union of the events that an error occurs at the t-th iteration while the clustering results in the previous $t - 1$ iterations are correct for $t = 1, \ldots, T$. Similar to the proof of Theorem III.1, the error probability is bounded by the union bound. See Appendix VI-B3 for more details. □

Theorem IV.2 shows that the split-based algorithm is exponentially consistent under any distance metric satisfying Assumption 1 with the error exponent b.

Corollary IV.2.1: Suppose the KS distance and the MMD statistic are used with $d_{th} = \frac{\Sigma_{ks}}{2}$ and $d_{th} = \frac{\Sigma_{mmd}}{2}$. Then for n sufficiently large, the error probability of Algorithm 5 after T iterations is upper bounded as follows

$$P_e^{KS} \leq 14M^2 T \exp\left(-\frac{n\Delta_{ks}^2}{8}\right),$$

$$P_e^{MMD} \leq 8M^2 T \exp\left(-\frac{n\Delta_{mmd}^2}{64\mathbb{G}}\right).$$

Proof: By Propositions 1 and 2, the upper bound on the error probability of Algorithm 5 in Theorem IV.2 applies to the KS distance and the MMD statistic. Thus, the corollary is obtained by substituting the values specified in Lemmas A.3–A.8 in the upper bound. □

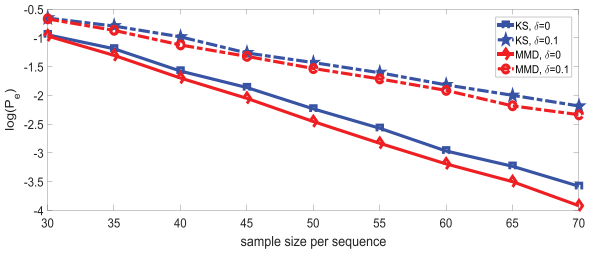
Corollary IV.2.1, combined with the fact that T is finitely bounded for finite M, implies that Algorithm 5 is exponentially consistent under both the KS and MMD with an error exponent no smaller than $\frac{\Delta_{ks}^2}{8}$ and $\frac{\Delta_{mmd}^2}{64\mathbb{G}}$, respectively.

V. Numerical Results

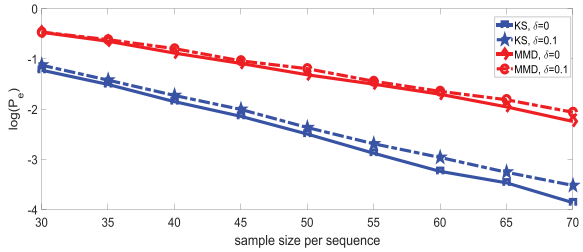
In this section, we provide some simulation results given $K = 5$ and $M_k = 3$ for $k = 1, \ldots, 5$, and $\mathbf{x}_{k,j_k}[i] \in \mathbb{R}$. Gaussian distributions $\mathcal{N}(\mu_{k,j_k}, \sigma^2)$ and Gamma distributions $\Gamma(a_{k,j_k}, b)$ are used in the simulations. The probability density function (p.d.f.) of $\Gamma(a, b)$ is defined as

$$f(x; \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right) \quad (x > 0),$$

where $\alpha > 0$, $\beta > 0$ and $\Gamma(\cdot)$ is the Gamma function, respectively. Specifically, the parameters of the distributions are $\mu_{k,j_k} \in \{k - \delta, k, k + \delta\}$, $\sigma^2 = 1$, $\alpha_{k,j_k} \in \{2.5k + 1 - \delta, 2.5k + 1, 2.5k + 1 + \delta\}$ and $\beta = 1$, where $\delta = 0$ and 0.1. Note that when $\delta = 0$, all the sequences generated from the

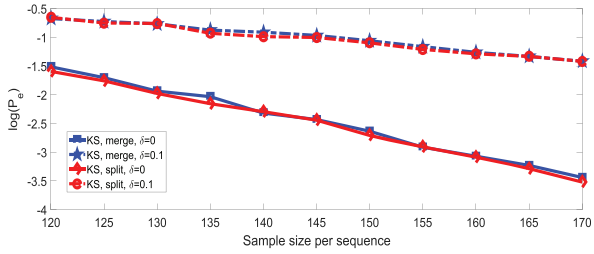


(a) Gaussian distributions

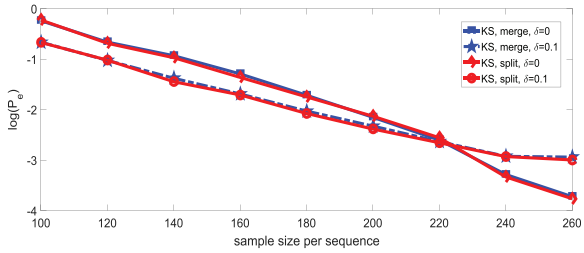


(b) Gamma distributions

Fig. 1. Performance of Algorithm 2.



(a) Gaussian distributions



(b) Gamma distributions

Fig. 2. Performance of Algorithms 4 and 5 for the KS distance.

same distribution cluster are generated from a single distribution. The squared exponential kernel function is used in the simulations for the MMD distance, i.e.,

$$g(x,y) = e^{-\frac{(x-y)^2}{2}}. \tag{8}$$

The simulation for a given sample size keeps running until the number of trials that provide incorrect clustering output reaches 1000.

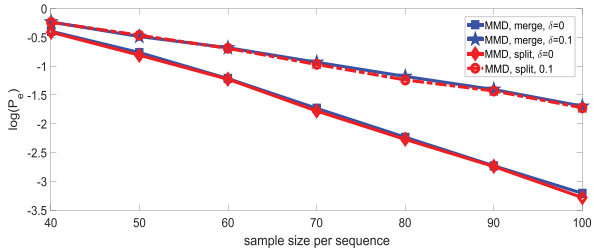
A. Known Number of Clusters

Simulation results for a known number of clusters are shown in Fig. 1. One can observe from the figures that by using both the KS distance and MMD, $\log P_e$ is a linear function of the sample size, i.e., P_e is exponentially consistent. Moreover, the logarithmic slope of P_e with respect to n, i.e., the quantity $-\frac{\log P_e}{n}$, increases as δ becomes smaller, which, in the current simulation setting, implies a larger Δ.

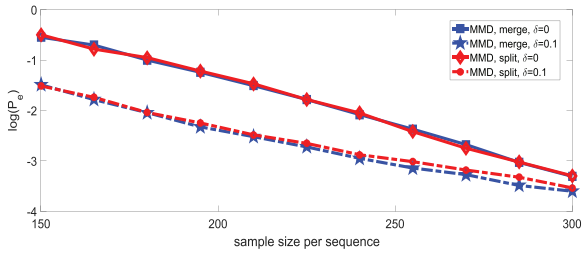
Furthermore, a good distance metric for Algorithm 2 depends on the underlying distributions. The kernel function in (8) is a good choice given symmetric p.d.f.s whereas the KS distance which relates to the order statistics becomes a better choice when the p.d.f.s are skewed.

B. Unknown Number of Clusters

With an unknown number of distribution clusters, the threshold d_{th} specified in Corollaries IV.1.1 and IV.2.1 are used in the simulation. The performance of Algorithms 4 and 5 for the KS distance and MMD are shown in Figs. 2 and 3, respectively. Given the KS distance and MMD, $\log P_e$'s are linear functions of the sample size when the sample size is large and larger Δ implies a larger slope of $\log P_e$. Furthermore, given the same value of δ, Algorithms 4 and 5 have similar performance under both the KS distance and MMD.



(a) Gaussian distributions



(b) Gamma distributions

Fig. 3. Performance of Algorithms 4 and 5 for MMD.

Intuitively, smaller δ implies larger Δ in the current simulation setting, thereby should result in better clustering performance for a given sample size. Figs. 2(b) and 3(b) indicates that Algorithms 4 and 5 with the KS distance and MMD performs better with $\delta = 0.1$ than that with $\delta = 0$ when the sample size is small. This is likely due to the fact that 1) the KS distance between the two sequences is always lower bounded by $\frac{1}{n}$, 2) the MMD estimator in (7) always has a positive bias, 3) the Gaussan kernel in (8) may not be a good choice for skewed p.d.f.s. Thus, with small sample sizes, Algorithms 4 and 5 are likely to overestimate the number of clusters. As the sample size increases, P_e with $\delta = 0$ indeed becomes smaller than that with $\delta = 0.1$ in Fig. 2(b). Meanwhile, since $\log P_e$ with $\delta = 0$

TABLE II
$\min \hat{K} / \max \hat{K}$ IN FIG. 2(A)

	Merge		Split	
n	$\delta = 0$	$\delta = 0.1$	$\delta = 0$	$\delta = 0.1$
120	5/8	5/9	5/9	5/9
125	5/8	5/8	5/8	5/9
130	5/8	5/9	5/8	5/9
135	5/8	5/8	5/8	5/9
140	5/8	5/9	5/8	5/9
145	5/7	5/9	5/7	5/9
150	5/7	5/9	5/7	5/8
155	5/7	5/9	5/7	5/8
160	5/7	5/8	5/7	5/9
165	5/7	5/8	5/7	5/9
170	5/7	5/8	5/7	5/9

TABLE III
$\min \hat{K} / \max \hat{K}$ IN FIG. 2(B)

	Merge		Split	
n	$\delta = 0$	$\delta = 0.1$	$\delta = 0$	$\delta = 0.1$
100	5/10	5/9	5/10	5/9
120	5/9	5/9	5/9	5/8
140	5/9	5/8	5/9	5/8
160	5/8	5/8	5/8	5/8
180	5/8	5/7	5/7	5/7
200	5/8	5/7	5/9	5/7
220	5/7	5/8	5/8	5/7
240	5/7	5/7	5/7	5/7
260	5/7	5/7	5/7	5/7

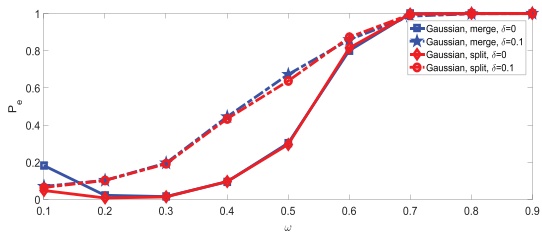
TABLE IV
$\min \hat{K} / \max \hat{K}$ IN FIG. 3(A)

	Merge		Split	
n	$\delta = 0$	$\delta = 0.1$	$\delta = 0$	$\delta = 0.1$
40	4/9	4/10	5/10	5/10
50	5/10	4/9	5/10	5/10
60	5/8	4/9	5/8	5/9
70	4/8	4/8	5/8	5/8
80	5/7	5/8	4/8	5/8
90	5/7	5/8	5/8	5/9
100	5/7	5/8	5/7	5/8

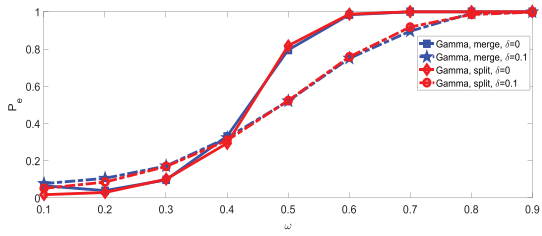
TABLE V
$\min \hat{K} / \max \hat{K}$ IN FIG. 3(B)

	Merge		Split	
n	$\delta = 0$	$\delta = 0.1$	$\delta = 0$	$\delta = 0.1$
150	5/10	4/8	5/9	5/8
165	5/8	4/9	5/9	5/8
180	5/8	4/7	5/9	4/7
195	5/9	4/8	5/8	5/8
210	5/8	4/7	5/8	5/8
225	5/8	4/7	5/8	5/7
240	5/8	4/7	5/7	5/8
255	5/8	4/7	5/7	5/7
270	5/7	4/7	5/8	5/7
285	5/7	5/7	5/7	5/7
300	5/7	4/7	5/7	5/7

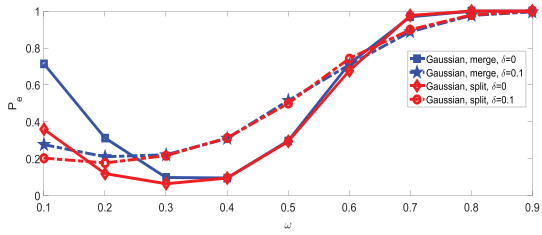
has a larger slope in Fig. 3(b), it should eventually become less than $\log P_e$ with $\delta = 0.1$. In Tables II–V, the maximum and minimum estimated number of clusters by Algorithms 4 and 5 corresponding to Figs. 2 and 3 are provided. Tables II and III show that under the KS distance the algorithms tend



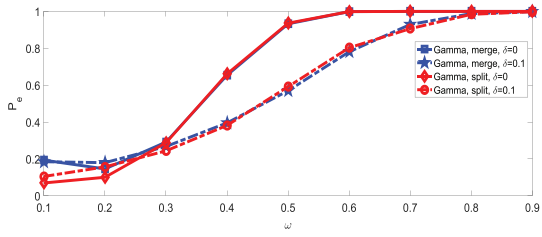
(a) KS distance ($n = 100$, Gaussian)



(b) KS distance ($n = 100$, Gamma)



(c) MMD ($n = 80$, Gaussian)



(d) MMD ($n = 80$, Gamma)

Fig. 4. Performance of Algorithms 4 and 5 given different ω.

TABLE VI
$\hat{K} < K / \hat{K} = K / \hat{K} > K$ IN FIG. 3(A)

	Merge		Split	
n	$\delta = 0$	$\delta = 0.1$	$\delta = 0$	$\delta = 0.1$
40	0/0.36/0.64	0/0.25/0.77	0/0.35/0.65	0/0.25/0.75
50	0/0.56/0.44	0/0.42/0.58	0/0.57/0.43	0/0.41/0.59
60	0/0.72/0.28	0/0.53/0.47	0/0.72/0.28	0/0.52/0.48
70	0/0.83/0.17	0/0.64/0.36	0/0.83/0.16	0/0.65/0.35
80	0/0.90/0.10	0/0.71/0.29	0/0.90/0.10	0/0.73/0.27
90	0/0.94/0.06	0/0.77/0.23	0/0.94/0.06	0/0.78/0.22
100	0/0.96/0.04	0/0.83/0.17	0/0.96/0.04	0/0.83/0.16

to overestimate the number of clusters given $K = 5$. This can be mitigated by the increased threshold d_{th} to control merging/splitting of cluster centers, which is verified by Fig. 4. The frequencies of the cases where $\hat{K} < K$, $\hat{K} = K$ and $\hat{K} > K$ corresponding to Fig. 3 are provided in Tables VI and VII. Those frequencies that are smaller than 0.005 are set to zero. Combining Tables IV and V , we can conclude that

TABLE VII
$\hat{K} < K / \hat{K} = K / \hat{K} > K$ IN FIG. 3(B)

	Merge		Split	
n	$\delta = 0$	$\delta = 0.1$	$\delta = 0$	$\delta = 0.1$
150	0/0.42/0.58	0/0.78/0.22	0/0.39/0.61	0/0.78/0.22
165	0/0.51/0.49	0/0.83/0.17	0/0.54/0.46	0/0.83/0.17
180	0/0.63/0.37	0/0.87/0.18	0/0.61/0.39	0/0.87/0.13
195	0/0.71/0.29	0/0.90/0.10	0/0.70/0.30	0/0.89/0.11
210	0/0.78/0.22	0/0.92/0.08	0/0.77/0.23	0/0.92/0.08
225	0/0.83/0.17	0/0.93/0.07	0/0.83/0.17	0/0.93/0.07
240	0/0.88/0.12	0/0.95/0.05	0/0.87/0.13	0/0.94/0.06
255	0/0.91/0.09	0/0.96/0.04	0/0.91/0.09	0/0.95/0.05
270	0/0.93/0.07	0/0.96/0.04	0/0.94/0.06	0/0.96/0.04
285	0/0.95/0.05	0/0.97/0.03	0/0.95/0.05	0/0.96/0.04
300	0/0.96/0.04	0/0.97/0.03	0/0.96/0.04	0/0.97/0.03

under MMD, the proposed algorithms also tend to overestimate the number of clusters. However, unlike the KS distance, the overestimation may not be mitigated by simply increasing d_{th}, which is verified by Fig. 4(c).

C. Choice of d_{th}

Note that in general $d_{th} = \omega d_L + (1-\omega) d_H$, where $\omega \in (0,1)$. Theorems IV.1 and IV.2 only establish the exponential consistency of Algorithms 4 and 5, respectively. We now investigate the impact on performance given different ω's. One can observe from Fig. 4 that the choice of d_{th} has a significant impact on the performance of Algorithms 4 and 5. The optimal d_{th} depends on both the value of δ and the underlying distributions. Moreover, from Figs. 4(a)-(b), we can see that a smaller ω which implies larger d_{th} results in better performance for KS distance and the two algorithms always have similar performance. Similarly, from Figs. 4(c)-(d), smaller ω also results in better performance for MMD except the case of Gaussian distributions with $\delta = 0$.

D. Computational Complexity

Assume that the complexity of the sum and point-wise max/min operations is linear in the argument cardinality. The complexity of other operations is assumed to be $O(1)$. The computational complexities of the center initialization step and the cluster initialization step in Algorithm 1 are $O(K^2 M)$ and $O(KM)$, respectively. The computational complexity of the center update step and cluster update step in Algorithm 2 are $O(KM^2)$ and $O(KM)$. Thus, the computational complexity of Algorithm 2 is $O(\binom{M}{K} K^{(M-K+1)} M^2)$

Similarly, one can verify that the computational complexities of the center initialization step and the cluster initialization step in Algorithm 3 are $O(M^3)$ and $O(M^2)$, respectively. The computational complexity of the center update step, the merge step and the cluster update step in Algorithm 4 are $O(M^3)$, $O(M^3)$ and $O(M^2)$. Thus, the computational complexity of Algorithm 4 is $O(M^3 T_{max})$.

The computational complexities of the finding $\mathbf{c}_1^1$, the split step and the cluster update step in Algorithm 5 are $O(M^2)$, $O(M)$ and $O(M^2)$, respectively. Thus, the computational complexity of Algorithm 5 is $O(M^3)$.

VI. Conclusion

This paper studied the k-medoids algorithm for clustering data sequences generated from composite distributions. The convergence of the proposed algorithms and the upper bound on the error probability were analyzed for both known and unknown number of clusters. The exponential decay of error probabilities of the proposed algorithms was established for distance metrics satisfying certain properties. In particular, the KS distance and MMD were shown to satisfy the required condition, and hence the corresponding algorithms were exponentially consistent. Note that the assumption of knowing d_L and d_H (or their bounds) can be justified because the empirical KS distance and MMD can be constructed, which converge to the true KS distance and MMD. Thus these thresholds or their bounds can be obtained from historical data.

One possible generalization of the current work is to investigate the exponential consistency of other clustering algorithms given distributional distance metrics that satisfy the properties similar to that in Assumption 1.

Appendix

A. Technical Lemmas

The following technical lemmas are used to prove Corollaries III.1.1, IV.1.1 and IV.2.1. All the data sequences in Lemmas A.3–A.8 are assumed to consist of n i.i.d. samples.

Lemma A.1 [Dvoretzky-Kiefer-Wolfowitz Inequality [38]]: Suppose $\mathbf{x}$ consists of n i.i.d. samples generated from p. Then

$$P\big(d_{KS}(\mathbf{x}, p) > \epsilon\big) \le 2\exp\left(-2n\epsilon^2\right).$$

Theorem A.2 [Theorem 7 in [37]]: Suppose $\mathbf{x} \sim p$, $\mathbf{y} \sim q$, where $\mathbf{x}$ and $\mathbf{y}$ have m and n samples, respectively. Given $0 \le g(x,y) \le \mathbb{G}$, the following inequality holds:

$$P\left(|\mathbb{MMD}(\mathbf{x},\mathbf{y}) - \mathbb{MMD}(p,q)| > f(\mathbb{G},m,n) + \epsilon\right) \le 2\exp\left(-\frac{\epsilon^2 mn}{2\mathbb{G}\,(m+n)}\right).$$

where $f(\mathbb{G},m,n) = 2(\sqrt{\frac{\mathbb{G}}{m}} + \sqrt{\frac{\mathbb{G}}{n}})$.

Lemmas A.3–A.8 establish that the KS distance and the MMD statistic obtained by (7) satisfy Assumption 1. Moreover, the lemmas provided in [18] are special cases of Lemmas A.3, A.5 and A.7 with $d_L = 0$.

Lemma A.3: Suppose $\mathbf{x}_j \sim p_j$ for $j = 1, 2$, where $p_j \in \mathcal{P}$ and $d_{KS}(\mathcal{P}) \le d_{L,ks}$. Then for any $d_0 > d_{L,ks}$,

$$P\big(d_{KS}(\mathbf{x}_1, \mathbf{x}_2) > d_0\big) \le 4\exp\left(-\frac{n(d_0 - d_{L,ks})^2}{2}\right).$$

Proof: Consider

$$\begin{aligned}
&P\big(d_{KS}(\mathbf{x}_1,\mathbf{x}_2) > d_0\big)\\
&\quad\le P\big(d_{KS}(\mathbf{x}_1,p_1) + d_{KS}(p_1,p_2) + d_{KS}(\mathbf{x}_2,p_2) > d_0\big)\\
&\quad\le P\big(d_{KS}(\mathbf{x}_1,p_1) + d_{L,ks} + d_{KS}(\mathbf{x}_2,p_2) > d_0\big)\\
&\quad\le P\bigg(d_{KS}(\mathbf{x}_1,p_1) > \frac{\hat{d}}{2}\bigg) + P\bigg(d_{KS}(\mathbf{x}_2,p_2) > \frac{\hat{d}}{2}\bigg)\\
&\quad\le 4\exp\bigg(-\frac{n\hat{d}^2}{2}\bigg),
\end{aligned}$$

where $\hat{d} = d_0 - d_{L,ks}$. The first inequality is due to the triangle inequality of the L_1-norm and the property of the supremum, and the last inequality is due to Lemma A.1. Therefore, we have

$$P\big(d_{KS}(\mathbf{x}_1,\mathbf{x}_2) > d_0\big) \le 4\exp\bigg(-\frac{n(d_0 - d_{L,ks})^2}{2}\bigg).$$

□

Lemma A.3 implies that the KS distance satisfies (3c) for $d > d_{L,ks}$.

Lemma A.4: Suppose $\mathbf{x}_j \sim p_j$ for $j = 1, 2$, where $p_j \in \mathcal{P}$ and $\mathbb{MMD}(\mathcal{P}) \le d_{L,mmd}$. Then for any $d_0 > d_{L,mmd}$ and sufficiently large n,

$$P\big(\mathbb{MMD}(\mathbf{x}_1,\mathbf{x}_2) > d_0\big) \le 2\exp\left(-\frac{n\left(d_0 - d_{L,mmd}\right)^2}{16\mathbb{G}}\right).$$

Proof: Since $\mathbb{MMD}(p_1,p_2) \le d_{L,mmd}$, we have

$$\begin{aligned}
&P\left(\mathbb{MMD}(\mathbf{x}_1,\mathbf{x}_2) > d_0\right)\\
&\quad\le P\left(\mathbb{MMD}(\mathbf{x}_1,\mathbf{x}_2) - \mathbb{MMD}(p_1,p_2) > d_0 - d_{L,mmd}\right)\\
&\quad\le P\left(\left|\mathbb{MMD}(\mathbf{x}_1,\mathbf{x}_2) - \mathbb{MMD}(p_1,p_2)\right| > d_0 - d_{L,mmd}\right).
\end{aligned}$$

Choose $\epsilon = \frac{d_0 - d_{L,mmd}}{2}$ and n sufficiently large such that $f\left(\mathbb{G},n,n\right) + \epsilon < d_0 - d_{L,mmd}$. By Theorem A.2, we have,

$$P\left(\mathbb{MMD}(\mathbf{x}_1,\mathbf{x}_2) > d_0\right) \le 2\exp\left(-\frac{n\left(d_0 - d_{L,mmd}\right)^2}{16\mathbb{G}}\right).$$

□

Lemma A.4 implies that the MMD statistc satisfies (3c) for $d > d_{L,mmd}$.

Lemma A.5: Suppose two distribution clusters $\mathcal{P}_1$ and $\mathcal{P}_2$ satisfy (3a) under the KS distance. Assume that for $j = 1, 2$, $\mathbf{x}_j \sim p_j$ where $p_j \in \mathcal{P}_j$. Then for any $d_0 < d_{H,ks}$,

$$P\big(d_{KS}(\mathbf{x}_1,\mathbf{x}_2) \le d_0\big) \le 4\exp\bigg(-\frac{n(d_{H,ks} - d_0)^2}{2}\bigg).$$

Proof: Similar to the proof of theorem A.3, we have

$$\begin{aligned}
&P\big(d_{KS}(\mathbf{x}_1,\mathbf{x}_2) \le d_0\big)\\
&\quad\le P\big(-d_{KS}(\mathbf{x}_1,p_1) + d_{KS}(p_1,p_2) - d_{KS}(\mathbf{x}_2,p_2) \le d_0\big)\\
&\quad\le P\big(-d_{KS}(\mathbf{x}_1,p_1) + d_2 - d_{KS}(\mathbf{x}_2,p_2) < d_0\big)\\
&\quad\le P\bigg(d_{KS}(\mathbf{x}_1,p_1) > \frac{\hat{d}}{2}\bigg) + P\bigg(d_{KS}(\mathbf{x}_2,p_2) > \frac{\hat{d}}{2}\bigg)\\
&\quad\le 4\exp\bigg(-\frac{n\hat{d}^2}{2}\bigg),
\end{aligned}$$

where $d_0 < d_2 < d_{H,ks}$, $\hat{d} = d_2 - d_0$ and $\lim_{d_2 \uparrow d_{H,ks}} = d_{H,ks} - d_0$. The last inequality is due to Lemma A.1. Therefore, by the continuity of the exponential function, we have

$$P\big(d_{KS}(\mathbf{x}_1,\mathbf{x}_2) \le d_0\big) \le 4\exp\bigg(-\frac{n(d_{H,ks} - d_0)^2}{2}\bigg).$$

□

Lemma A.5 implies that the KS distance satisfies (3b) for $d > d_{H,ks}$.

Lemma A.6: Suppose two distribution clusters $\mathcal{P}_1$ and $\mathcal{P}_2$ satisfy (3a) under MMD. Assume that for $j = 1, 2$, $\mathbf{x}_j \sim p_j$, where $p_j \in \mathcal{P}_j$. Then for any $d_0 < d_{H,mmd}$ and sufficiently large n,

$$P\big(\mathbb{MMD}(\mathbf{x}_1,\mathbf{x}_2) \le d_0\big) \le 2\exp\left(-\frac{n\left(d_{H,mmd} - d_0\right)^2}{16\mathbb{G}}\right).$$

Proof: Similar to the proof of Lemma A.4, we have

$$\begin{aligned}
&P\left(\mathbb{MMD}(\mathbf{x}_1,\mathbf{x}_2) \le d_0\right)\\
&\quad\le P\left(\mathbb{MMD}(p_1,p_2) - \mathbb{MMD}(\mathbf{x}_1,\mathbf{x}_2) \ge d_{H,mmd} - d_0\right)\\
&\quad\le P\left(\left|\mathbb{MMD}(\mathbf{x}_1,\mathbf{x}_2) - \mathbb{MMD}(p_1,p_2)\right| > \hat{d}\right)
\end{aligned}$$

where $\hat{d} = d_3 - d_0$ and $d_0 < d_3 < d_{H,mmd}$. Choose $\epsilon = \frac{\hat{d}}{2}$ and n sufficiently large such that $f(\mathbb{G},n,n) + \epsilon < \hat{d}$. By Theorem A.2, we have

$$P\left(\mathbb{MMD}(\mathbf{x}_1,\mathbf{x}_2) > d_0\right) \le 2\exp\left(-\frac{n\hat{d}^2}{16\mathbb{G}}\right).$$

Let $\lim_{d_3 \uparrow d_{H,ks}} = d_{H,ks} - d_0$. Then by the continuity of the exponential function, we have for n sufficiently large,

$$P\left(\mathbb{MMD}(\mathbf{x}_1,\mathbf{x}_2) \le d_0\right) \le 2\exp\left(-\frac{n\left(d_{H,mmd} - d_0\right)^2}{16\mathbb{G}}\right).$$

□

Lemma A.6 implies that MMD satisfies (3b) for $d > d_{H,mmd}$.

Lemma A.7: [39] Suppose two distribution clusters $\mathcal{P}_1$ and $\mathcal{P}_2$ satisfy (3a) under the KS distance. Assume that for $j = 1, 2$, $\mathbf{x}_j \sim p_j$ with length n where $p_j \in \mathcal{P}_j$. Then for any $\mathbf{x}_3 \sim p_3$ with length n where $p_3 \in \mathcal{P}_1$,

$$P\big(d_{KS}(\mathbf{x}_1,\mathbf{x}_3) \ge d_{KS}(\mathbf{x}_2,\mathbf{x}_3)\big) \le 6\exp\bigg(-\frac{n\Delta_{ks}^2}{8}\bigg).$$

Lemma A.7 implies that the KS distance satisfies (3d) for $d \in (d_{L,ks}, d_{H,ks})$.

Lemma A.8: Suppose two distribution clusters $\mathcal{P}_1$ and $\mathcal{P}_2$ satisfy (3a) under MMD. Assume that for $j = 1, 2$, $\mathbf{x}_j \sim p_j$ where $p_j \in \mathcal{P}_j$. Then for any $\mathbf{x}_3 \sim p_3$ where $p_3 \in \mathcal{P}_1$, where n is sufficiently large,

$$P\big(\mathbb{MMD}(\mathbf{x}_1, \mathbf{x}_3) \geq \mathbb{MMD}(\mathbf{x}_2, \mathbf{x}_3)\big) \leq 4 \exp\left(-\frac{n\Delta_{mmd}^2}{64\mathbb{G}}\right)$$

Proof: Let $\hat{\Delta} \in (0, \Delta_{mmd})$. Similar to the proof of Lemmas A.4 and A.6, we have

$$\begin{aligned}
&P\left(\mathbb{MMD}(\mathbf{x}_1, \mathbf{x}_3) \geq \mathbb{MMD}(\mathbf{x}_2, \mathbf{x}_3)\right) \\
&\quad \leq P\big(\mathbb{MMD}(\mathbf{x}_1, \mathbf{x}_3) - \mathbb{MMD}(p_1, p_3) + \mathbb{MMD}(p_2, p_3) \\
&\qquad - \mathbb{MMD}(\mathbf{x}_2, \mathbf{x}_3) \geq \Delta_{mmd}\big) \\
&\quad \leq P\big(\big|\mathbb{MMD}(\mathbf{x}_1, \mathbf{x}_3) - \mathbb{MMD}(p_1, p_3)| + |\mathbb{MMD}(p_2, p_3) \\
&\qquad - \mathbb{MMD}(\mathbf{x}_2, \mathbf{x}_3)\big| > \hat{\Delta}\big) \\
&\quad \leq P\left(\big|\mathbb{MMD}(\mathbf{x}_1, \mathbf{x}_3) - \mathbb{MMD}(p_1, p_3)\big| > \frac{\hat{\Delta}}{2}\right) \\
&\qquad + P\left(\big|\mathbb{MMD}(\mathbf{x}_2, \mathbf{x}_3) - \mathbb{MMD}(p_2, p_3)\big| > \frac{\hat{\Delta}}{2}\right),
\end{aligned}$$

where the last inequality is due to the union bound. Choose $\epsilon = \frac{\hat{\Delta}}{4}$ and n sufficiently large such that $f(\mathbb{G}, n, n) + \epsilon < \frac{\hat{\Delta}}{2}$. By Theorem A.2, we have

$$P\left(\mathbb{MMD}(\mathbf{x}_1, \mathbf{x}_3) \geq \mathbb{MMD}(\mathbf{x}_2, \mathbf{x}_3)\right) \leq 4 \exp\left(-\frac{n\hat{\Delta}^2}{64\mathbb{G}}\right).$$

Let $\hat{\Delta} \uparrow \Delta_{mmd}$. By the continuity of the exponential function, we have for n sufficiently large,

$$P\left(\mathbb{MMD}(\mathbf{x}_1, \mathbf{x}_3) \geq \mathbb{MMD}(\mathbf{x}_2, \mathbf{x}_3)\right) \leq 4 \exp\left(-\frac{n\Delta_{mmd}^2}{64\mathbb{G}}\right).$$

□

Lemma A.8 implies that MMD satisfies (3d) for $d \in (d_{L,mmd}, d_{H,mmd})$.

B. Proof of Main Results

Define the following three events:

$$\begin{aligned}
S_1(d_{th}) &= \big\{\exists k, k' \in I_1^K, k \neq k', j \in I_1^{M_k}, j' \in I_1^{M_{k'}}, \\
&\qquad \text{s.t. } d(\mathbf{x}_{k,j}, \mathbf{x}_{k',j'}) \leq d_{th}\big\}, \\
S_2(d_{th}) &= \big\{\exists k \in I_1^K,\ j, j' \in I_1^{M_k} \text{ s.t. } d(\mathbf{x}_{k,j}, \mathbf{x}_{k,j'}) > d_{th}\big\}, \\
S_3 &= \big\{\exists k, k' \in I_1^K,\ k \neq k',\ j_1, j_2 \in I_1^{M_k},\ j' \in I_1^{M_{k'}}, \\
&\qquad \text{s.t. } d(\mathbf{x}_{k,j_1}, \mathbf{x}_{k,j_2}) \geq d(\mathbf{x}_{k,j_1}, \mathbf{x}_{k',j'})\big\},
\end{aligned}$$

where $d_{th} \in (d_L, d_H)$. Assume that the sequences $\mathbf{x}_{k,j}$'s and the corresponding distribution clusters $\mathcal{P}_k$'s satisfy Assumption 1. By (3b)–(3d) and the union bound, we have

$$P\big(S_1(d_{th})\big) \leq \sum_{k=1}^{K} \sum_{\substack{k'=1 \\ k' \neq k}}^{K} \sum_{j_k=1}^{M_k} \sum_{j_{k'}=1}^{M_{k'}} a_1 e^{-bn} \leq M^2 a_1 e^{-bn}, \tag{9a}$$

$$P\big(S_2(d_{th})\big) \leq \sum_{k=1}^{K} \sum_{j_k=1}^{M_k} \sum_{j_{k'}=1}^{M_{k'}} a_2 e^{-bn} \leq M^2 a_2 e^{-bn}, \tag{9b}$$

$$P\big(S_3\big) \leq \sum_{k=1}^{K} \sum_{j_k=1}^{M_k} \sum_{j_{k'}=1}^{M_{k'}} a_3 e^{-bn} \leq M^2 a_3 e^{-bn}. \tag{9c}$$

The main idea of the proofs of Theorems III.1, IV.1 and IV.2 is to show that the error event at each iteration is a subset of $S_1(d_{th}) \cup S_2(d_{th}) \cup S_3$.

1) Proof of Theorem III.1: The convergence of Algorithm 2 results from the design of the algorithm. Consider the $(t-1)$-th clustering step and the t-th center update step. We have for $t \geq 1$,

$$\sum_{k=1}^{K} \sum_{\mathbf{y}_i \in \mathcal{C}_k^{t-1,a}} d(\mathbf{y}_i, \mathbf{c}_k^{t-1,a}) \geq \sum_{k=1}^{K} \sum_{\mathbf{y}_i \in \mathcal{C}_k^{t-1,a}} d(\mathbf{y}_i, \mathbf{c}_k^{t,a}). \tag{10}$$

Moreover, for the t-th center update and the t-th cluster update, we have for $t \geq 1$,

$$\sum_{k=1}^{K} \sum_{\mathbf{y}_i \in \mathcal{C}_k^{t-1}} d(\mathbf{y}_i, \mathbf{c}_k^{t,a}) \geq \sum_{l=1}^{K} \sum_{\mathbf{y}_i \in \mathcal{C}_k^{t}} d(\mathbf{y}_i, \mathbf{c}_k^{t,a}). \tag{11}$$

The equalities in (10) and (11) hold if and only if $\mathcal{C}_k^{t-1} = \mathcal{C}_k^t$ and $\mathbf{c}_k^{t-1,a} = \mathbf{c}_k^{t,a}$ for $k = 1, \ldots, K$ respectively which implies the convergence of the algorithm.

Suppose there are K sequences assigned as cluster centers, and as a result $M - K$ remaining sequences are to be assigned to cluster centers. The order in which cluster centers are chosen does not matter, so there are a total of $\binom{M}{K}$ permutations of them. Since each of the remaining $M - K$ sequences can be assigned to one and only one cluster center, there are a total of $K^{(M-K)}$ possible assignments. Therefore the total number of valid partitions is $\binom{M}{K} K^{(M-K)}$. By (10) and (11), Algorithm 2 is guaranteed to visit each possible partition at most once except the one coinciding with the clustering output. Hence the maximum number of algorithm iterations is always upper bounded as

$$T \leq \binom{M}{K} K^{(M-K)}.$$

Define for $t \geq 1$,

$$\begin{aligned}
E^t = \{&\text{After } t\text{-th iteration, there are } K_1 \text{ centers generated} \\
&\text{from } K_2 \text{ distribution clusters}\}.
\end{aligned}$$

where

$$K_1 \begin{cases} > K_2 & \text{if } K_2 = K, \\ \geq K_2 & \text{if } K_2 < K. \end{cases}$$

Similarly, define

$$E^0 = \{\text{The center initialization obtains } K_1 \text{ centers generated from } K_2 \text{ distribution clusters}\}.$$

Then E^t for $t \geq 0$ denotes the error event that centers are incorrectly chosen at the center initialization or the t-th center update. We first consider the error occurs at the initialization step. For Algorithm 2,

$$\begin{aligned} E^0 &= \{\text{The center initialization results in } K \text{ centers generated from } K_2 \ (< K) \text{ distribution clusters centers.}\} \\ &= \{\exists k, l, l' \in I_1^K,\ l \neq l' \text{ s.t. } \mathbf{c}_l^{0,a}, \mathbf{c}_{l'}^{0,a} \sim \mathcal{P}_k\}. \end{aligned}$$

Moreover, define

$$E_1^0 = E^0 \cap \big\{\exists l, l' \in \{1, \ldots, K\} \text{ s.t. } d(\mathbf{c}_l^{0,a}, \mathbf{c}_{l'}^{0,a}) \leq d_{th}\big\},$$

$$E_2^0 = E^0 \cap \big\{\exists l, l' \in \{1, \ldots, K\} \text{ s.t. } d(\mathbf{c}_l^{0,a}, \mathbf{c}_{l'}^{0,a}) > d_{th}\big\}.$$

Then $E^0 = E_1^0 \cup E_2^0$. Without loss of generality, assume that $\mathbf{c}_1^{0,a}, \ldots, \mathbf{c}_K^{0,a}$ are chosen sequentially at the center initialization step and $l < l'$. Then E_1^0 implies that for all the sequences $\mathbf{z} \in \{\mathbf{y}_i\}_{i=1}^M \setminus \{\mathbf{c}_m^{0,a}\}_{m=1}^{l'}$,

$$\min_{m \in \{1, \ldots, l'-1\}} d(\mathbf{c}_m^{0,a}, \mathbf{z}) \leq d_{th}.$$

Thus, $E_1^0 \subset S_1(d_{th})$. Then by (9a), we have

$$P\big(E_1^0\big) \leq P\big(S_1(d_{th})\big) \leq M^2 a_1 e^{-bn}.$$

Moreover, since $E_2^0 \subset S_2(d_{th})$, by (9b), we have

$$P\big(E_2^0\big) \leq P\big(S_2(d_{th})\big) \leq M^2 a_2 e^{-bn}.$$

Thus, the error probability at the center initialization step is bounded as follows

$$P\big(E^0\big) \leq M^2(a_1 + a_2)e^{-bn}. \tag{12}$$

We now consider the assignment step. Define for $t \geq 1$,

$$H^t = \{\text{The clustering result after the } t\text{-th cluster update is incorrect}\},$$

Moreover, define

$$H^0 = \{\text{The clustering initialization is incorrect}\}.$$

Since $E^t \subset H^{t-1}$ for $t \geq 1$, it is sufficient to obtain an upper bound on $P\big(H^t\big)$ which serves as the upper bound of $P(H^t \cup E^t)$. Define

$$\hat{H}_1^t = \begin{cases} H^0 \setminus E^0 & \text{for } t = 0, \\ H^t \setminus \big(E^0 \cup \big(\cup_{l=0}^{t-1} \big(H^l\big)\big)\big) & \text{for } t \geq 1. \end{cases}$$

Then $E^0 \cup \big(\cup_{t=1}^T H^t\big) = E^0 \cup \big(\cup_{t=0}^T \hat{H}_1^t\big)$, which is the event that Algorithm 2 makes an error before the first T iterations complete. Moreover, $\hat{H}_1^t$ implies the event that an error occurs at the t-th cluster update step *given* correct center update in the same iteration which is denoted by

$$\bar{H}_1^t = \big\{\exists k, k', l, l' \in I_1^K,\ k \neq k',\ j_k \in I_1^{M_k} \text{ s.t. } d(\mathbf{x}_{k,j_k}, \mathbf{c}_l^{t,a}) \geq d(\mathbf{x}_{k,j_k}, \mathbf{c}_{l'}^{t,a}) :\ \mathbf{c}_l^{t,a} \sim \mathcal{P}_k,\ \mathbf{c}_{l'}^{t,a} \sim \mathcal{P}_{k'}\big\}.$$

Then $P(\hat{H}_1^t) \leq P\big(\bar{H}_1^t\big)$. Moreover, since $\bar{H}_1^t \subset S_3$, we have

$$P(\hat{H}_1^t) \leq P(\bar{H}_1^t) \leq P\big(S_3\big) \leq M^2 a_3 e^{-bn}. \tag{13}$$

Therefore, by (12), (13) and the union bound, the error probability of Algorithm 2 after T iterations is bounded by

$$\begin{aligned} P_e &= P\big(E^0 \cup \big(\cup_{t=0}^T \hat{H}_1^t\big)\big) \\ &\leq M^2\big(a_1 + a_2 + (T+1)a_3\big)e^{-bn}. \end{aligned} \tag{14}$$

2) Proof of Theorem IV.1: If no merge step is executed and $\hat{K}$ clusters are found by Algorithm 3, then similar to the proof of Theorem III.1 Algorithm 4 converges after at most T_0 iterations, where

$$T_0 = \binom{M}{\hat{K}} \hat{K}^{(M-\hat{K})}.$$

If the merge step is executed, the valid partitions before and after the merge step are mutually exclusive since the number of clusters is strictly decreasing. Therefore, Algorithm 4 converges after at most $T_{\max}$ iterations, where

$$T_{\max} = \sum_{\hat{K}=1}^{M} \binom{M}{\hat{K}} \hat{K}^{(M-\hat{K})}.$$

In conclusion, Algorithm 4 converges after at most $T_{\max}$ iterations since $T_0 < T_{\max}$.

We then analyze the error probability of Algorithm 4. We first consider the initialization step. Define

$$E_3^0 = E^0 \cap \big\{K_2 < K\big\},$$

$$E_4^0 = E^0 \cap \big\{K_2 = K\big\}.$$

Then $E^0 = E_3^0 \cup E_4^0$. Moreover, since

$$E_3^0 \subset \big\{\exists k, k' \in I_1^K,\ j_k \in I_1^{M_k},\ j_{k'} \in I_1^{M_{k'}} \text{ s.t. } d(\mathbf{x}_{k,j_k}, \mathbf{x}_{k',j_{k'}}) \leq d_{th}\big\},$$

$$E_4^0 \subset \big\{\exists k \in I_1^K,\ j_k, j_k' \in I_1^{M_k},\ \text{s.t. } d(\mathbf{x}_{k,j_k}, \mathbf{x}_{k,j_k'}) > d_{th}\big\},$$

then $E_3^0 \subset S_1(d_{th})$ and $E_4^0 \subset S_2(d_{th})$. Thus, by (9a), (9b), we have

$$P\big(E_3^0\big) \leq P\big(S_1(d_{th})\big) \leq M^2 a_1 e^{-bn},$$

$$P\big(E_4^0\big) \leq P\big(S_2(d_{th})\big) \leq M^2 a_2 e^{-bn}.$$

Therefore, by the union bound, the probability that an error occurs at the center initialization step is bounded by

$$P\big(E^0\big) \leq P\big(E_3^0\big) + P\big(E_4^0\big) \leq M^2 a_1 e^{-bn} + M^2 a_2 e^{-bn}. \tag{15}$$

We now consider the error that occurs during iterations. $E^t \subset H^{t-1}$ for $t \geq 1$ still holds. Furthermore, define an incorrect merge as the event that the distance between two centers generated from different distribution clusters is smaller than d_{th}. Let D^t be the event that incorrect merges occur at the t-th ($t \geq 1$) merge step. Thus we only need to bound $P\big(H^t\big)$ and $P\big(D^t\big)$. Let $B_{t_1,t_2} = \big(\cup_{l=1}^{t_1} D^l\big) \cup \big(\cup_{l=0}^{t_2} H^l\big)$ for $t_1 \geq 1$ and

$t_2 \geq 1$. Define

$$\hat{D}^t = \begin{cases} D^1 & \text{for } t = 1 \\ D^t \setminus \big(E^0 \cup B_{t-1,t-1}\big) & \text{for } t > 1 \end{cases},$$

$$\hat{H}_2^t = \begin{cases} H^0 \setminus E^0 & \text{for } t = 0 \\ H^t \setminus \big(E^0 \cup B_{t,t-1}\big) & \text{for } t \geq 1 \end{cases}.$$

Then

$$\begin{aligned} &E^0 \cup \big(\cup_{t=1}^T D^t\big) \cup \big(\cup_{t=0}^T H^t\big) \\ &\quad = E^0 \cup \big(\cup_{t=1}^T \hat{D}^t\big) \cup \big(\cup_{t=0}^T \hat{H}_1^t\big), \end{aligned}$$

which denotes the event that an error occurs before T iterations complete. Note that $\hat{D}^t$ implies the event that an error occurs at the t-th merge step *given* correct center update in the same iteration, which is denoted by

$$\begin{aligned} \bar{D}^t = \big\{&\exists k, k' \in I_1^K,\ k \neq k',\ l \in I_1^{\hat{K}^{t-1}},\ \text{s.t. } d(\mathbf{c}_l^{t,a}, \mathbf{c}_{l'}^{t,a}) \\ &\leq d_{th} :\ \mathbf{c}_l^{t,e} \sim \mathcal{P}_k, \mathbf{c}_{l'}^{t,e} \sim \mathcal{P}_{k'}\big\}. \end{aligned}$$

Then $P\big(\hat{D}^t\big) \leq P\big(\bar{D}^t\big)$ and $\bar{D}^t \subset S_1(d_{th})$. Thus, by (9a), we have

$$P\big(\hat{D}^t\big) \leq P\big(\bar{D}^t\big) \leq P\big(S_1(d_{th})\big) \leq M^2 a_1 e^{-bn}. \tag{16}$$

Moreover, we have $P\big(\hat{H}_2^t\big) \leq P\big(\bar{H}_2^t\big)$, where

$$\begin{aligned} \bar{H}_2^t = \big\{&\exists k, k' \in I_1^K,\ k \neq k',\ j_k \in I_1^{M_k},\ l, l' \in I_1^{\hat{K}^t},\ \text{s.t.} \\ &d(\mathbf{x}_{k,j_k}, \mathbf{c}_l^{t,e}) \geq d(\mathbf{x}_{k,j_k}, \mathbf{c}_{l'}^{t,e}) :\ \mathbf{c}_l^{t,e} \sim \mathcal{P}_k,\ \mathbf{c}_{l'}^{t,e} \sim \mathcal{P}_{k'}\big\}. \end{aligned}$$

Note that $P(\bar{H}_2^t)$ has the same upper bound as $P(\bar{H}_1^t)$ in (13). Therefore, by (15), (13) and (16), the error probability after T iterations is bounded by

$$\begin{aligned} P_e &= P\big(Y^0 \cup \big(\cup_{t=0}^T \hat{H}_2^t\big) \cup \big(\cup_{t=1}^T \hat{D}^t\big)\big) \\ &\leq M^2\big((T+1)a_1 + a_2 + (T+1)a_3\big)e^{-bn}. \end{aligned} \tag{17}$$

3) Proof of Theorem IV.2: Note that in the extreme case, splitting results in each cluster containing only one sequence, i.e., splitting can happen at most $M-1$ times. Therefore, Algorithm 5 converges after at most M iterations. Furthermore, if $\hat{K}$ does not change from the $(t-1)$-th to the t-th iteration, then $\mathcal{C}_k^{t-1} = \mathcal{C}_k^t$ and $\mathbf{c}_k^{t-1} = \mathbf{c}_k^t$ for $k = 1, \ldots, \hat{K}$, which implies the convergence of the algorithm.

Let A^t be the event that the error occurs at the t-th split step. Then $A^t = A_1^t \cup A_2^t$, where

$A_1^t = \big\{$The algorithm fails to split any cluster containing sequences generated by diffierent distribution clusters at the t-th iteration$\big\}$,

$A_2^t = \big\{$The algorithm splits a cluster containing sequences generated by one distribution clusters at the t-th iteration$\big\}$.

Let V^t denote the event that the clustering result at the t-th cluster update is incorrect. Then $A^t \cup V^t$ denotes the event that an error occurs at the t-th iteration. Define $\hat{A}^t = \hat{A}_1^t \cup \hat{A}_2^t$, where

$$\hat{A}_i^t = \begin{cases} A^1 & \text{for } t = 1, \\ A_i^t \setminus \big((\cup_{l=1}^{t-1} A^l) \cup (\cup_{l=1}^{t-1} V^l)\big) & \text{for } t > 1, \end{cases}$$

for $i = 1, 2$. Moreover, define

$$\hat{V}^t = \begin{cases} V^1 \setminus A^1 & \text{for } t = 1 \\ V^t \setminus \big((\cup_{l=1}^{t-1} V^l) \cup (\cup_{l=1}^{t} A^l)\big) & \text{for } t > 1 \end{cases}.$$

Then $\big(\cup_{t=1}^T A^t\big) \cup \big(\cup_{t=1}^T V^t\big) = \big(\cup_{t=1}^T \hat{A}^t\big) \cup \big(\cup_{t=1}^T \hat{V}^t\big)$. Since $\hat{A}_1^t \subset S_1(d_{th})$ and $\hat{A}_2^t \subset S_2(d_{th})$, then we have for $t = 1, \ldots, T$,

$$\begin{aligned} P\big(\hat{A}_1^t\big) &\leq P\big(S_1(d_{th})\big) \leq M^2 a_1 e^{-bn}, \\ P\big(\hat{A}_2^t\big) &\leq P\big(S_2(d_{th})\big) \leq M^2 a_2 e^{-bn}. \end{aligned}$$

Moreover, since $P\big(\hat{A}^t\big) = P\big(\hat{A}_1^t \cup \hat{A}_2^t\big)$, by the union bound

$$P\big(\hat{A}^t\big) \leq M^2 a_1 e^{-bn} + M^2 a_2 e^{-bn}. \tag{18}$$

Furthermore, by Definition IV.1.1, $\hat{V}^t$ implies the following event

$$\begin{aligned} \bar{V}^t = \big\{&\exists l, l' \in I_1^{\hat{K}^t},\ k, k' \in I_1^K\ k' \neq k,\ j_k \in I_1^{M_k}\ \text{s.t.} \\ &d(\mathbf{x}_{k,j_k}, \mathbf{c}_l^{t,s}) \geq d(\mathbf{x}_{k,j_k}, \mathbf{c}_{l'}^{t,s}) :\ \mathbf{c}_l^{t,s} \sim \mathcal{P}_k, \\ &\mathbf{c}_{l'}^{t,s} \sim \mathcal{P}_{k'}\big\}. \end{aligned}$$

Then, $P\big(\hat{V}^t\big) \leq P\big(\bar{V}^t\big)$ and $\bar{V}^t \subset S_3$. Thus, we have

$$P\big(\hat{V}^t\big) \leq P\big(\bar{V}^t\big) \leq M^2 a_3 e^{-bn}. \tag{19}$$

Therefore, by (18), (19) and the union bound, the error probability of Algorithm 5 after T iterations is bounded by

$$\begin{aligned} P_e &= P\big(\big(\cup_{t=1}^T \hat{A}^t\big) \cup \big(\cup_{t=1}^T \hat{V}^t\big)\big) \\ &\leq M^2 T\big(a_1 + a_2 + a_3\big)e^{-bn}. \end{aligned} \tag{20}$$

REFERENCES

[1] Y. Sakurai, L. Li, R. Chong, and C. Faloutsos, "Efficient distribution mining and classification," in *Proc. SIAM Int. Conf. Data Mining*, Atlanta, Georgia, USA, Apr. 2008, pp. 632–643.

[2] E. Spellman, B. C. Vemuri, and M. Rao, "Using the KL-center for efficient and accurate retrieval of distributions arising from texture images," in *Proc. IEEE Conf. Comput. Vision, Pattern Recognit.*, San Diego, CA, USA, Jun. 2005, vol. 1, pp. 111–116.

[3] C. Lin, C. Chen, H. Lee, and J. Liao, "Fast k-means algorithm based on a level histogram for image retrieval," *Expert Syst. Appl.*, vol. 41, no. 7, pp. 3276–3283, 2014.

[4] M. Vrac, L. Billard, E. Diday, and A. Chédin, "Copula analysis of mixture models," *Comput. Statist.*, vol. 27, no. 3, pp. 427–457, 2012.

[5] R. Moreno-Sáez, M. S. de Cardona, and L. Mora-López, "Data mining and statistical techniques for characterizing the performance of thin-film photovoltaic modules," *Expert Syst. Appl.*, vol. 40, no. 17, pp. 7141–7150, 2013.

[6] R. Moreno-Sáez and L. Mora-López, "Modelling the distribution of solar spectral irradiance using data mining techniques," *Environmental Model., Softw.*, vol. 53, pp. 163–172, 2014.

[7] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[8] D. Barber, *Bayesian Reasoning and Machine Learning*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[9] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 3, pp. 129–137, Mar. 1982.

[10] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in *Proc. KDD Workshop Text Mining*, 2000.
[11] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, Jul. 2002.
[12] L. Kaufman and P. Rousseeuw, "Clustering by means of medoids," in *Statistical Data Analysis Based on the L1-norm and Related Methods. Cham*, Switzerland: Springer, 1987, pp. 405–416.
[13] M. Laan, K. Pollard, and J. Bryan, "A new partitioning around medoids algorithm," *J. Statistical Comput. Simul.*, vol. 73, no. 8, pp. 575–584, 2003.
[14] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3336–3341, 2009.
[15] J. C. Gower and G. J. S. Ross, "Minimum spanning trees and single linkage cluster analysis," *J. Royal Statist. Soc. Series C (Appl. Statist.)*, vol. 18, no. 1, pp. 54–64, 1969.
[16] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.
[17] I. Katsavounidis, C. C. J. Kuo, and Z. Zhang, "A new initialization technique for generalized Lloyd iteration," *IEEE Signal Process. Lett.*, vol. 1, no. 10, pp. 144–146, Oct. 1994.
[18] T. Wang, D. J. Bucci, Y. Liang, B. Chen, and P. K. Varshney, "Exponentially consistent k-means clustering algorithm based on Kolmogrov-Smirnov test," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Calgary, Canada, Apr. 2018, pp. 2296–2300.
[19] T. Wang, D. J. Bucci, Y. Liang, B. Chen, and P. K. Varshney, "Clustering under composite generative models," in *Proc. Annu. Conf. Inform. Sci. Syst.*, Princeton, NJ, USA, Mar. 2018, pp. 338–343.
[20] L. Mora-López and J. Mora, "An adaptive algorithm for clustering cumulative probability distribution functions using the Kolmogorov-Smirnov two-sample test," *Expert Syst. Appl.*, vol. 42, pp. 4016–4021, 2015.
[21] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Advances Neural Inform. Process. Syst.*, Vancouver, Canada, 2002, pp. 849–856.
[22] L. Kaufman and P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ, USA: Wiley, 2009, vol. 344.
[23] T. Velmurugan and T. Santhanam, "Computational complexity between k-means and k-medoids clustering algorithms for normal and uniform distributions of data points," *J. Comput. Sci.*, vol. 6, no. 3, pp. 363–368, 2010.
[24] W. Sheng and X. Liu, "A genetic k-medoids clustering algorithm," *J. Heuristics*, vol. 12, no. 6, pp. 447–466, Dec. 2006.
[25] M. Gangeh, A. Sadeghi-Naini, M. Diu, H. Tadayyon, M. Kamel, and G. J. Czarnota, "Categorizing extent of tumor cell death response to cancer therapy using quantitative ultrasound spectroscopy and maximum mean discrepancy," *IEEE Trans. Med. Imag.*, vol. 33, no. 6, pp. 1390–1400, Jun. 2014.
[26] Y. Li, S. Nitinawarat, and V. V. Veeravalli, "Universal outlier hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4066–4082, Jul. 2014.
[27] Y. Bu, S. Zou, and V. Veeravalli, "Linear-complexity exponentially-consistent tests for universal outlying sequence detection," in *Proc. IEEE Int. Symp. Inform. Theory*, Aachen, Germany, Jun. 2017, pp. 988–992.
[28] S. Zou, Y. Liang, H. V. Poor, and X. Shi, "Nonparametric detection of anomalous data streams," *IEEE Trans. Signal Process.*, vol. 65, no. 21, pp. 5785–5797, Nov 2017.
[29] S. Zou, Y. Liang, and H. V. Poor, "Nonparametric detection of geometric structures over networks," *IEEE Trans. Signal Process.*, vol. 65, no. 19, pp. 5034–5046, Jun. 2017.
[30] Q. Mai and H. Zou, "The Kolmogorov filter for variable screening in high-dimensional binary classification," *Biometrika*, vol. 100, pp. 229–234, 2012.
[31] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf, "Kernel measures of conditional dependence," in *Proc. Advances Neural Inform. Process. Syst.*, Vancouver, Canada, Dec. 2008, pp. 489–496.
[32] B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf, "Injective Hilbert space embeddings of probability measures," in *Proc. Annu. Conf. Learn. Theory*, Helsinki, Finland, 2008, pp. 111–122.
[33] K. Fukumizu, B. Sriperumbudur, A. Gretton, and B. Schölkopf, "Characteristic kernels on groups and semigroups," in *Proc. Advances Neural Inform. Process. Syst.*, Vancouver, Canada, Dec. 2009, pp. 473–480.
[34] B. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf, "Hilbert space embeddings and metrics on probability measures," *J. Mach. Learn. Res.*, vol. 11, pp. 1517–1561, 2010.
[35] L. Song, A. Gretton, and K. Fukumizu, "Kernel embeddings of conditional distributions," *IEEE Signal Process. Mag.*, vol. 30, no. 4, pp. 98–111, Jul. 2013.
[36] A. Smola, A. Gretton, L. Song, and B. Schölkopf, "A Hilbert space embedding for distributions," in *Proc. Int. Conf. Algorithmic Learn. Theory*, Sendai, Japan, Oct. 2007, pp. 13–31.
[37] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, 2012.
[38] P. Massart, "The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality," *Ann. Probability*, vol. 18, pp. 1269–1283, 1990.
[39] Q. Li, T. Wang, D. J. Bucci, Y. Liang, B. Chen, and P. K. Varshney, "Nonparametric composite hypothesis testing in an asymptotic regime," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 5, pp. 1005–1014, Oct. 2018.

Tiexing Wang, photograph and biography not available at the time of publication.

Qunwei Li (S'16), photograph and biography not available at the time of publication.

Donald J. Bucci, photograph and biography not available at the time of publication.

Yingbin Liang (SM'19), photograph and biography not available at the time of publication.

Biao Chen (S'96–M'99–SM'07–F'15), photograph and biography not available at the time of publication.

Pramod K. Varshney (S'72–M'77–SM'82–F'97–LF'18), photograph and biography not available at the time of publication.