Recovery Guarantees for Kernel-based Clustering under Non-parametric Mixture Models

Leena C. Vankadara[1]   **Sebastian Bordt**[1,2]   **Ulrike von Luxburg**[1,2]   **Debarghya Ghoshdastidar**[3]

University of Tübingen[1]   Max Planck Institute   Technical University of
for Intelligent Systems, Tübingen[2]   Munich[3]

Abstract

Despite the ubiquity of kernel-based clustering, surprisingly few statistical guarantees exist beyond settings that consider strong structural assumptions on the data generation process. In this work, we take a step towards bridging this gap by studying the statistical performance of kernel-based clustering algorithms under non-parametric mixture models. We provide necessary and sufficient separability conditions under which these algorithms can consistently recover the underlying *true clustering*. Our analysis provides guarantees for kernel clustering approaches without structural assumptions on the form of the component distributions. Additionally, we establish a key equivalence between kernel-based data-clustering and kernel density-based clustering. This enables us to provide consistency guarantees for kernel-based estimators of non-parametric mixture models. Along with theoretical implications, this connection could have practical implications, including in the systematic choice of the bandwidth of the Gaussian kernel in the context of clustering.

1 INTRODUCTION

Clustering refers to the unsupervised task of partitioning a given data sample or the input space into *meaningful* regions. Kernel clustering approaches such as kernel k-means (Dhillon et al., 2004) and kernel spectral clustering (Ng et al., 2002) are widely adopted by practitioners, particularly for partitioning non-spherical

complex cluster structures. Beyond their good practical behavior, kernel methods are appealing due to their amenability to theoretical analysis. However, as an anomaly, kernel clustering has been elusive to theoretical analysis, in particular, under general non-parametric assumptions on the data generation process. One of the principle sources for this gap between theory and practice had been the lack of a universally accepted characterization of the quality of a clustering. One popular notion of the goodness of clustering is defined as the one that consistently partitions the data space. Consistency is, however, only a necessary condition for clustering algorithms. It simply checks if an algorithm asymptotically converges to a limiting partition. The optimality of this limiting partition is not studied under consistency. As an example, spectral clustering has been shown to be consistent (Luxburg et al., 2008) for any similarity function k. However, if one uses a similarity function based on an uninformative kernel such as the identity kernel, then the obtained limiting partition is clearly not guaranteed to be a desirable one. Density based clustering (Hartigan, 1975; Hartigan, 1981; Rinaldo et al., 2010) is another popular line of work with theoretical backing, where clusters are defined as connected components of high-density regions, referred to as density level sets. The imprecise notion of a high-density region is overcome using the so called cluster-tree approach (Chaudhuri et al., 2014; Sriperumbudur and Steinwart, 2012), where a continuum of all level sets is simultaneously considered.

Another systematic approach to overcome the ambiguity concerning the quality of clustering lies in the so called *model-based clustering*, which assumes that the data is generated from a mixture distribution and the goal is to partition the data in congruity with the components that generate the data. However, theoretical analysis of kernel clustering methods have been confined to settings with parametric distributions (Yan et al., 2016; Couillet et al., 2016; Vankadara et al., 2020). Parametric assumptions such as the Gaussian mixture setting, where the components are assumed to

be normally distributed, are extremely restrictive since the data generated under such assumptions are far from a typical dataset for which kernel clustering algorithms are applicable. In contrast, non-parametric assumptions on the data-generation process can be considerably less restrictive, but kernel clustering algorithms have been elusive to theoretical analysis under such assumptions. A primary hurdle in the analysis of clustering approaches under non-parametric assumptions is due to the issue of identifiability of non-parametric mixture models, that is, non-parametric models may be ambiguously defined. There is limited previous work that presents an analysis of kernel-based clustering algorithms under non-parametric mixture models. Schiebinger et al. (2015) provide recovery guarantees for spectral clustering of non-parametric mixtures by analyzing the spectral properties of the Laplacian operator under the assumption that the overlap between the components is small relative to a notion of "indivisibility" of the components. The analysis provided in Schiebinger et al. (2015) is restricted to that of spectral clustering and considerably different from the analysis in this paper.

1.1 Contributions

Non-parameteric kernel clustering. We provide non-parametric conditions for consistency of certain kernel-based clustering algorithms. To the best of our knowledge, these are among the first theoretical guarantees to kernel-based clustering methods without assumptions on the form of the component distributions.

1. We provide an **impossibility result for kernel k-means**: there exists a mixture distribution with arbitrarily large separation between the components such that for finite samples from this distribution kernel k-means fails to recover the underlying clustering.

2. We establish **sufficient separability conditions** under which kernel-based algorithms such as k-center, farthest-first k-means (FFk-means++), or kernel linkage algorithms can consistently recover the true partition, given finite samples from a mixture distribution.

3. We establish **necessary conditions for consistency** of the kernel FFk-means++ and kernel linkage algorithms and show that these separability conditions are optimal, that is, the sufficient conditions match the necessary conditions.

Kernel-based data clustering as distribution clustering. We establish a key equivalence between kernel-based data clustering and kernel-based density clustering. In particular:

4. We show that Gaussian kernel-based data clustering is equivalent to density clustering, where, each data point is first represented by a Gaussian probability density function and the densities are then clustered using the maximum mean discrepancy metric (with respect to a Gaussian kernel).

5. In addition to theoretical implications, this connection could also have practical implications in matters such as choosing the bandwidth of the Gaussian kernel for clustering which has not been systematically studied in literature so far. Our analysis reveals that the bandwidth of the kernel used for clustering needs to decrease with n but, perhaps surprisingly, asymptotically remain non-zero.

Non-parametric estimation of mixture models. Due to this relationship between kernel data clustering and distribution clustering, any standard Gaussian kernel clustering algorithm can be used to define an estimation procedure of the mixture model. Therefore, in addition to our primary contributions to kernel clustering, we also make contributions related to non-parametric *estimation* of mixture models.

6. We provide conditions under which the *estimation procedures* corresponding to the kernel-based clustering algorithms can consistently estimate the true mixture model.

2 FORMAL SETTING AND BACKGROUND

Consider the Euclidean space $\mathbb{R}^d$ of dimension d as the input domain. Let $\mathcal{P}$ denote the space of all Borel probability measures on $\mathbb{R}^d$ that are absolutely continuous with respect to the Lebesgue measure. In our analysis, we use the framework of mixing measures to define mixture distributions. This is fairly standard in the analysis of non-parametric mixture models (Aragam et al., 2020; Holzmann et al., 2006; Kimeldorf et al., 1970; Nguyen et al., 2013; Teicher, 1963) primarily due to the following reasons:

- Arbitrary mixture distributions are not identifiable. Mixing measures allow for the specification of *true components*. Section 3.1 provides a thorough discussion on identifiability of mixture models.

- In non-parametric clustering, one typically does not make any assumptions on the form of the component distributions. An elegant way to accomplish this is to allow arbitrary component distributions from $\mathcal{P}$ and impose restrictions on the set of admissible mixing measures.

Following the notation of Aragam et al. (2020), we denote the space of all probability distributions (mixing measures) over $\mathcal{P}$ supported on a finite (K) number of elements in $\mathcal{P}$ by $\mathcal{P}_K^2$. Formally,

$$\mathcal{P}_K^2 = \left\{ \sum_{k=1}^K \lambda_k \delta_{\gamma_k} \colon \lambda_k \in \mathbb{R}^+, \ \gamma_k \in \mathcal{P}, \ \sum_{k=1}^K \lambda_k = 1 \right\},$$

where δ_γ denotes the point mass concentrated at $\gamma \in \mathcal{P}$ and $[K]$ denotes the set $\{1, 2, \cdots K\}$ for any $K \in \mathbb{N}$. Furthermore, assume that the coefficients (λ_k) of the component measures (γ_k) are bounded away from 0. Define $m : \mathcal{P}_K^2 \to \mathcal{P}$ to be the mapping that uniquely associates a **mixing measure** to a **mixture distribution**, that is,

$$\forall \, \Lambda \in \mathcal{P}_K^2 : \Lambda = \sum_{k=1}^K \lambda_k \delta_{\gamma_k} \longrightarrow m(\Lambda) = \sum_{k=1}^K \lambda_k \gamma_k.$$

The support of a mixing measure Λ specifies the true components of the corresponding mixture distribution, $\Gamma = m(\Lambda)$.

We now describe the **problem setup**. Let $\Lambda = \sum_{k \in [K]} \lambda_k \delta_{\gamma_k}$ be a mixing measure in $\mathcal{P}_K^2$. Consider a finite sample $X = \{x_1, x_2, \cdots x_n\}$ drawn independently and identically (i.i.d) according to some $\Gamma = m(\Lambda) = \sum_{k=1}^K \lambda_k \gamma_k$. We denote this by $X \sim \Gamma^n$. The component measures γ_k are absolutely continuous with respect to the Lebesgue measure, and therefore admit density functions. We use f_k to denote the density function corresponding to the component measure γ_k and $f = \sum_{k=1}^K \lambda_k f_k$ to denote the density function corresponding to Γ. Given any density function h, we use the term "probability distribution corresponding to h" to denote the measure ψ which is defined as $\psi_i(A) = \int_A h(x) dx$, for any Borel set $A \subseteq \mathbb{R}^d$.

For any sample $X = \{x_1, x_2, \cdots x_n\}$, we use a map $\sigma : [n] \to [K]$ to represent a $K-$partition of X and $c_k(\sigma) = \{x_i \in X : \sigma(i) = k\}$ to denote the kth cluster according to σ for all $k \in [K]$. When it is clear from context, we drop the dependence on σ and simply use c_k to denote $c_k(\sigma)$. Given any $X \sim \Gamma^n$, the "planted partition" and the "Bayes partition" are of particular interest.

Planted partition. Observe that, drawing a sample $X = \{x_1, x_2, \cdots x_n\}$ according to a mixing measure $\Lambda = \sum_{k \in [K]} \lambda_k \delta_{\gamma_k}$ is equivalent to the following procedure. For each $i \in [n]$,

1. sample index $k \in [K]$ using the weights $\lambda_1, \ldots, \lambda_k$,

2. generate a sample x_i from γ_k.

We refer to the partition induced by this process as the planted partition and use σ_X^* or σ_n^* to denote it.

Bayes partition. We refer to the mapping $b^* : X \to [K]$ as the Bayes partition function, given by

$$\sigma_{Bayes}(x) = \arg\max_k \lambda_k f_k(x).$$

We use σ_{Bayes}^X to denote the Bayes partition with respect to a sample $X \sim m(\Lambda)^n$ which is defined as the Bayes partition function restricted to X.

Remark. In this work, any reference to a sample should be understood as drawn i.i.d according to a mixture distribution Γ.

We now describe the main objective of this work: **clustering of non-parametric mixture models.**

Non-parametric clustering. *Given a finite sample $X = \{x_1, x_2, \cdots x_n\}$ drawn i.i.d according to Γ^n, the central objective of non-parametric, model-based clustering is to recover the planted partition up to a permutation over the labels, $[K]$.*

Alternatively, one could also be interested in the consistent estimation of the Bayes partition (Aragam et al., 2020). We present our results with respect to the former notion and they can easily be extended to the latter by means of a simple modification of the algorithms. We discuss this in more detail in Section 5. The primary objective of this paper is to understand the performance of kernel clustering algorithms under the framework of non-parametric clustering. A brief background on kernels is thus warranted for further discourse on our analysis.

Background on kernels. Every symmetric positive definite (p.d) kernel function $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is associated with a feature map $\phi : \mathbb{R}^d \to \mathcal{H}_g$, where $\mathcal{H}_g$ is a Hilbert space with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_g}$ such that $\langle \phi(x), \phi(y) \rangle_{\mathcal{H}_g} = g(x, y), \ \forall x, y \in \mathbb{R}^d$. $\mathcal{H}_g$ is a **reproducing kernel Hilbert space** (RKHS) if the mapping $f \mapsto f(x)$ is continuous for every $x \in \mathbb{R}^d$, where $f \in \mathcal{H}_g$. The Hilbert space $\mathcal{H}_g$ corresponding to a kernel g is of independent interest while dealing with probability measures since it admits feature representations referred to as the **kernel mean embeddings**. For any probability measure $P \in \mathcal{P}$, the kernel mean embedding with respect to kernel g is defined as $\mu_P(\cdot) = \int_{x \in \mathbb{R}^d} g(x, \cdot) dP$, which is an element of $\mathcal{H}_g$. The RKHS norm $\|\cdot\|_{\mathcal{H}_g}$ associated with $\mathcal{H}_g$ can be used to define a (semi-)metric between the probability measures. Formally, the maximum mean discrepancy (MMD) between two probability measures $P, Q \in \mathcal{P}$ with respect to the kernel g is given by $\rho(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}_g}$. If g is a characteristic kernel, such as the Gaussian kernel,

then ρ is a metric on the space of probability measures $\mathcal{P}$ (Fukumizu et al., 2008; Sriperumbudur, Gretton, et al., 2010). In our analysis, we consider the space $\mathcal{P}$ metrized by the MMD corresponding to a Gaussian kernel function, $g_\zeta : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, where $g_\zeta(x, y) = \exp\left(-\frac{\|x-y\|^2}{\zeta}\right)$ $\forall x, y \in \mathbb{R}^d$ with bandwidth $\zeta > 0$. The MMD metric enjoys several valuable properties, from both a theoretical and practical point of view (Gretton et al., 2012; Muandet et al., 2016). **Kernel density estimation** is a popular non-parametric approach for density estimation. Given any $X = \{x_1, x_2, \cdots x_n\} \sim \Gamma^n$, the kernel density estimate (KDE) of the density function f, with respect to Gaussian kernel g_β with bandwidth $\beta > 0$, is given by

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{f}_i(x); \quad \widetilde{f}_i(x) = \frac{\exp\left(-\frac{\|x-x_i\|^2}{2\beta^2}\right)}{(2\pi\beta^2)^{d/2}}. \quad (1)$$

Let $\widehat{\Gamma}, \psi_i \in \mathcal{P}$ be the probability distributions corresponding to $f, \widetilde{f}_i$ respectively. Under the following conditions on the **bandwidth parameter** β,

$$\beta \to 0, \quad \frac{n\beta^d}{\log n} \to \infty \text{ as } n \to \infty, \quad (2)$$

the kernel density estimate $\hat{f}_n$ converges to the true density f in the l_∞ norm (Giné et al., 2002; Einmahl et al., 2005).

3 RECOVERY GUARANTEES FOR KERNEL-BASED DATA CLUSTERING

Identifiability. A key theoretical question concerning both estimation and clustering under non-parametric mixture models is that of identifiability, that is, any mixture distribution can be decomposed in infinitely many ways into component distributions (Teicher, 1963; Holzmann et al., 2006; Vandermeulen et al., 2015; Miao et al., 2016; Aragam et al., 2020). Therefore, non-parametric clustering and estimation of mixture models are ill-defined, even if the number of components K is assumed to be known. The framework of mixing measures as discussed earlier allow for the specification of the "true components" and the "true planted/Bayes partitions". For any set of mixing measures $\mathfrak{L} \subseteq \mathcal{P}_K^2$, let $m(\mathfrak{L})$ denote the set of mixture distributions corresponding to $\mathfrak{L}$. Clearly, the mapping $\mathfrak{L} \mapsto m(\mathfrak{L})$ is not injective on the whole space $\mathfrak{L} = \mathcal{P}^2$ due to general non-identifiability. This motivates the following definition.

Definition 3.1 (**Identifiablility**). A subset $\mathfrak{L} \subseteq \mathcal{P}_K^2$ is called identifiable if the map $\mathfrak{L} \mapsto m(\mathfrak{L})$ is injective.

The most common approach to deal with identifiability is to make restrictive parametric assumptions on the form of the component distributions, for example, Gaussianity, which renders the mixture model identifiable (Bruni et al., 1985; Teicher, 1963). Recent work by Aragam et al. (2020) uses regularity and separability criteria to achieve identifiability. Our analysis, inspired by Aragam et al. (2020), also uses separability criterion to deal with identifiability. However, our analysis differs from theirs on several fronts since we do not impose any regularity conditions on the mixing measures and also consider a statistical approach to identifiability. Moreover, the focus of their paper (identifiability of non-parametric mixture models) is very different from ours, which is providing recovery guarantees for kernel-based clustering approaches.

Any non-parametric analysis of model-based clustering (or estimation) is typically preceded by an identifiability analysis for the mixture models. We do not explicitly study identifiability, that is, identifying a set $\mathfrak{L} \in \mathcal{P}_K^2$ for which only one mixing measure can generate a mixture distribution. Instead, given finite samples from the mixture distribution, we provide conditions under which a particular algorithm (is biased toward and hence) recovers the true mixing measure/partition. In our analysis of kernel-based clustering algorithms, we show that under appropriate separability conditions, certain algorithms can consistently recover the planted partition. Specifically, we present and analyze the asymptotic behavior of four different kernel-based clustering algorithms.

Algorithms. We present a brief description of the algorithms here for completeness and include detailed descriptions in the supplementary. Consider a finite sample $X = \{x_1, x_2, \cdots x_n\} \sim \Gamma^n$.

- **k-means** ($\mathcal{A}_{\mathbf{KMN}}$). The objective is to find a partition $\widehat{\sigma} : [n] \to [K]$ such that the sum of squared within cluster distances on X is minimized. We consider the optimal solution to the NP-Hard, k-means problem in our analysis.

- **FFk-means++** ($\mathcal{A}_{\mathbf{FFK}}$). This algorithm is a variant of k-means++ where the initial centers are chosen in a deterministic, farthest-first order.

- **k-center** ($\mathcal{A}_{\mathbf{CTR}}$). The objective seeks to obtain a k-partition of X such that the maximal radius of the clusters is minimized. The optimal solution to the NP-Hard k-center problem is analyzed.

- **Agglomerative linkage** ($\mathcal{A}_{\mathbf{LNK}}$). Given a similarity function (single, average or complete linkage), these algorithms generate a dendrogram establishing a hierarchy of clusters of the data in a bottom up approach, starting out with each point
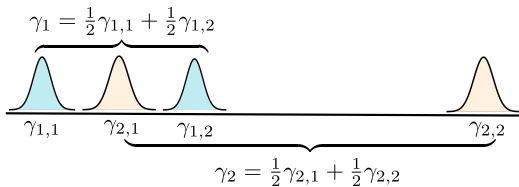
Figure 1: Example to show that simple separation conditions do not suffice to overcome identifiability. As the distribution $\gamma_{2,2}$ moves arbitrarily far from the remaining distributions, the distance between γ_1 and γ_2 also increases arbitrarily. However, without additional assumptions, no clustering algorithm can recover the **desirable clusters** as defined by the **true components** γ_1 and γ_2.

as its own cluster and progressively combining them into larger clusters until there is a single cluster that contains the entire data.

Given a positive definite kernel $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, the kernelized versions of these algorithms are defined by replacing the Euclidean inner product by the inner product $\langle \cdot, \cdot \rangle_g$ induced by g on the input space $\mathbb{R}^d$, which is given by

$$\langle x_i, x_j \rangle_g = g(x_i, x_j).$$

In this paper, we provide necessary and sufficient separability conditions for the kernel-based clustering algorithms $\mathcal{A}_{\text{KMN}}$, $\mathcal{A}_{\text{FFK}}$, $\mathcal{A}_{\text{CTR}}$, and $\mathcal{A}_{\text{LNK}}$.

Main results. For a finite sample $X = \{x_1, x_2, \cdots x_n\} \sim \Gamma^n$, recall that ψ_{x_i} refers to the probability distribution corresponding to $\widetilde{f}_i$ as defined in (1) with bandwidth parameter $\beta > 0$. Given a partition $\sigma : [n] \to [K]$ of probability distributions $\{\psi_i\}_{i=1}^n$, we use $\widehat{\gamma}_{k,\sigma}$ to denote the mean of the kth cluster according to σ, that is,

$$\widehat{\gamma}_{k,\sigma} = \frac{1}{|c_k(\sigma)|} \sum_{x_i \in c_k(\sigma)} \psi_{x_i}.$$

Let ρ denote the MMD corresponding to the Gaussian kernel g_ζ with respect to a bandwidth parameter $\zeta > 0$ and let g denote the Gaussian kernel function with the bandwidth parameter $(4\beta^2 + \zeta)$. For readability, when it is clear from context, we ignore the dependence on the partition function, σ in the notation. We now present one of our key results which establishes the impossibility of cluster recovery for kernel k-means. The result states that there is always a mixing measure with arbitrarily large MMD separation between the component distributions for which, given finite samples from this mixture, kernel k-means fails to recover the planted clustering.

Theorem 1 (Impossibility of clustering recovery by $\mathcal{A}_{\text{KMN}}$). *Fix $\zeta > 0$. Let β be any sequence of*

bandwidth parameters and let g be the Gaussian kernel with bandwidth parameter $4\beta^2 + \zeta$. For all $C > 0$, there exists a mixing measure $\Lambda \in \mathcal{P}_2^2$ such that

$$\rho(\gamma_1, \gamma_2) > C \sup_{x \in X_n} \rho(\psi_x, \widehat{\gamma}_{\sigma^*(x),\sigma^*}) \qquad (3)$$

holds within all finite samples and yet $\mathcal{A}_{KMN}$ with kernel g w.h.p. fails to recover the planted partition σ^*.

Even though kernel k-means fails to provably recover the planted partition for arbitrarily large separation between the components, there is a sufficient separation between the components beyond which kernel-based k-center, FFk-means++, and hierarchical linkage algorithms can provably and consistently recover the planted partition.

Theorem 2 (Sufficient conditions for consistency of $\mathcal{A}_{\text{CTR}}$, $\mathcal{A}_{\text{FFK}}$, and $\mathcal{A}_{\text{LNK}}$). *Fix $\zeta > 0$. Let β be any sequence of bandwidth parameters satisfying (2) and let g be the Gaussian kernel with bandwidth parameter $4\beta^2 + \zeta$. For any $\Lambda \in \mathcal{P}_K^2$, if there exists $\epsilon > 0$ such that*

$$\mathbb{P}_{X_n} \left(\inf_{k \neq k'} \rho(\gamma_k, \gamma_{k'}) > 4 \sup_{x \in X_n} \rho(\psi_x, \widehat{\gamma}_{\sigma^*(x),\sigma^*}) + \epsilon \right)$$
$$\xrightarrow{n \to \infty} 1, \quad (4)$$

then the algorithms $\mathcal{A}_{CTR}$, $\mathcal{A}_{FFK}$, and $\mathcal{A}_{LNK}$ with kernel g can w.h.p. recover the planted partition $\sigma^\star$.

The result states that, for recovery, the distance between any two component distributions in MMD (ρ) needs to be larger than about twice the maximal within cluster distance in the feature space: the RKHS ($\mathcal{H}_g$) corresponding to the kernel g, for clustering defined by the planted partition. The conditions provided here might appear to be weak, but perhaps more consequentially, in Theorem 3 we show that under no additional assumptions the constant $1/4$ is in fact necessary and hence cannot be improved for both $\mathcal{A}_{\text{FFK}}$ and $\mathcal{A}_{\text{LNK}}$.

Theorem 3 (Necessary conditions for $\mathcal{A}_{\text{FFK}}$ and $\mathcal{A}_{\text{LNK}}$ to consistently recovery the planted partition). *Fix $\zeta > 0$. Let β be any sequence of bandwidth parameters and let g be the Gaussian kernel with bandwidth parameter $4\beta^2 + \zeta$. For any $\epsilon > 0$, there exists $\Lambda \in \mathcal{P}_2^2$ such that*

$$\mathbb{P}_{X_n} \left(\rho(\gamma_1, \gamma_2) > 4 \sup_{x \in X_n} \rho(\psi_x, \widehat{\gamma}_{\sigma^*(x),\sigma^*}) - \epsilon \right) \xrightarrow{n \to \infty} 1$$
$$(5)$$

and the algorithms $\mathcal{A}_{FFK}$ and $\mathcal{A}_{LNK}$ with kernel g fail to recover the planted partition σ^* with probability approaching $\frac{1}{2}$ and 1, respectively, as $n \to \infty$.

The proofs for the results appear in the supplementary. For the kernel k-center problem, we can indeed show
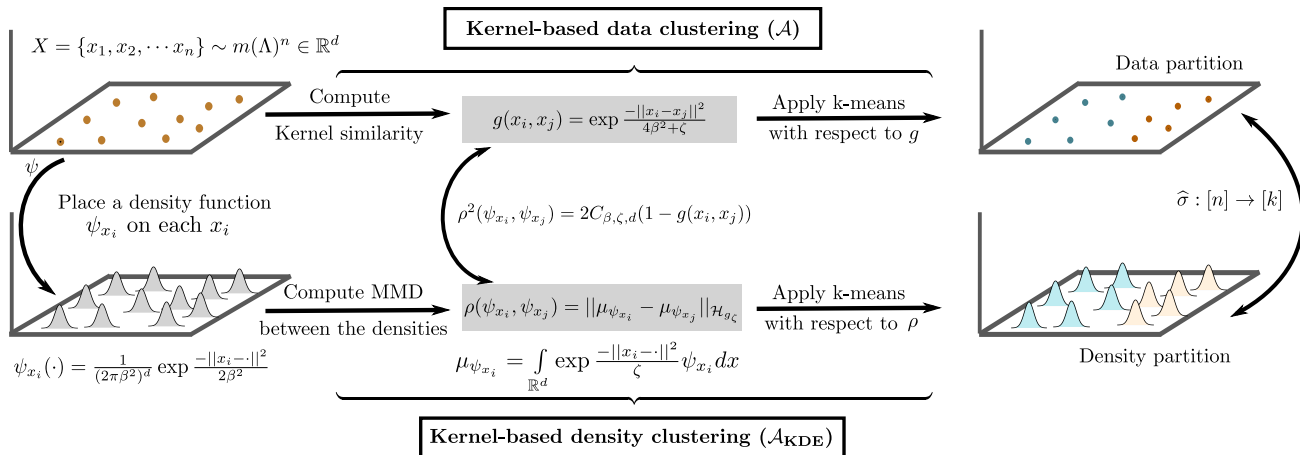
Figure 2: Illustration of the equivalence between kernel-based data clustering and distribution clustering. For Gaussian kernel clustering algorithm $\mathcal{A}$ using a bandwidth parameter $\eta > 0$, decompose η to obtain *any* $\beta > 0$ and $\zeta > 0$ satisfying $4\beta^2 + \zeta = \eta$. Then $\mathcal{A}$ can equivalently be reformulated as a kernel-based density clustering procedure as shown in the figure.

that the constant in the sufficient conditions (4) can further be improved to $1/3$ when $K = 2$. However, we believe that for any arbitrary K, the conditions provided in (4) cannot be further improved. This can be shown for a linear kernel and we leave the more general case of the Gaussian kernel as a conjecture. Our results not only show that certain kernel-based clustering algorithms can exploit separability to recover the planted clustering but also clearly show that under no additional assumptions very strong separability conditions are necessary to obtain recovery guarantees for kernel-based clustering. Furthermore, due to reasons of identifiability, simple separation conditions between the component distributions do not suffice to derive consistent recovery guarantees. For instance consider a simple example of a mixture distribution shown in Figure 1. As $\gamma_{2,2}$ moves arbitrarily far from the remaining distributions, the distance between the two component distributions, γ_1, γ_2 becomes arbitrarily far. However, without additional assumptions, it is not possible for a clustering algorithm to recover the desirable clustering even if we see infinite amount of data. Therefore, the separability conditions on the component distributions are necessarily dependent on the geometric properties of the distribution and not merely on the sample size or the dimension of the input space as it often is in the parametric setting. Our results, providing necessary and sufficient recovery conditions for kernel-based data clustering algorithms (Theorems 1, 2 , and 3), are obtained by analyzing an equivalent density/distribution clustering procedure which is considerably easier to analyze. Specifically, this equivalence allows us to exploit the metric geometry of the space of probability measures on the Euclidean space. Our proof techniques are motivated by the work of Aragam et al. (2020). We

now describe this relationship between kernel-based data clustering and kernel-based density clustering.

4 EQUIVALENCE BETWEEN KERNEL-BASED DATA CLUSTERING AND DISTRIBUTION CLUSTERING

In this section, we present a density clustering procedure and describe its close relationship to kernel-based data clustering. Given a finite sample X, the density clustering procedure clusters the component probability distributions (ψ_i) of the kernel density estimate with respect to X using MMD as the metric between the distributions. This procedure is illustrated in Figure 2. As shown in Figure 2, the partition obtained by this density clustering procedure can be used to define a partition on the sample X. This partition can alternatively be obtained by using a simple kernel-based data clustering procedure. We now describe this density clustering procedure, which we denote by $\mathcal{A}_{\mathrm{KDE}}$.

Kernel-based density clustering $\mathcal{A}_{\mathrm{KDE}}$. Consider Gaussian kernel g_ζ for some $\zeta > 0$. Given sample $X \sim \Gamma^n$:

- Estimate the density of Γ by $\widehat{f} = \frac{1}{n} \sum_{i=1}^{n} \widetilde{f}_i$ as in (1) with a bandwidth parameter $\beta > 0$.

- Consider MMD corresponding to the Gaussian kernel g_ζ as the metric between the distributions. Cluster the probability distributions $\{\psi_i\}_{i=1}^{n}$ corresponding to $\left\{\widetilde{f}_i\right\}_{i=1}^{n}$ by means of a distance based clustering algorithm (for example, k-means) to
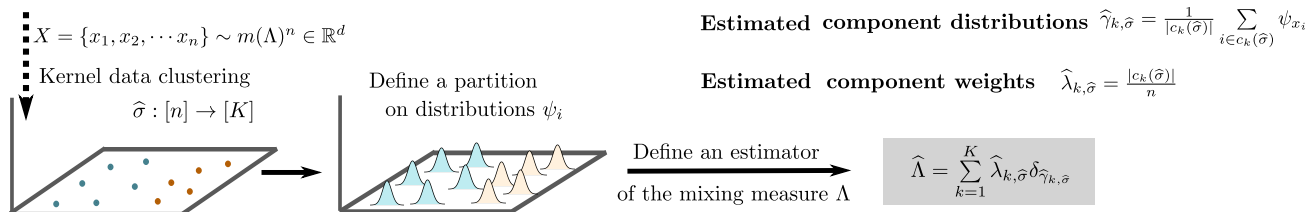
Figure 3: Illustration of the estimation procedure defined with respect to a kernel clustering algorithm. Any Gaussian kernel clustering algorithm can be used to define a partition on component density functions $\{\psi_i\}_{i=1}^n$ which can in turn be used to define an estimator of the mixing measure Λ.

obtain a partition function $\widehat{\sigma}$.

This procedure is also illustrated in Figure 2. We show that for appropriately chosen bandwidth parameters, any kernel-based data clustering algorithm can be equivalently formulated as a density clustering procedure ($\mathcal{A}_{\mathrm{KDE}}$). Recall that β and ζ are the bandwidth parameters of the Gaussian kernels used in $\mathcal{A}_{\mathrm{KDE}}$ for kernel density estimation and for defining the MMD respectively. Then, let $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be the Gaussian kernel with bandwidth parameter $4\beta^2 + \zeta$. The following lemma shows that the maximum mean discrepancy between the component distributions (ψ_i) is closely related to kernel evaluations on the input data.

Lemma 1 (**MMD between components is closely related to kernel evaluations between input data.**). *Given any sample $X \in \mathbb{R}^d$, let the component KDE distributions (ψ_i) be defined in the usual way. For all $x_i, x_j \in X$,*

$$\rho^2(\psi_i, \psi_j) = C_{\beta,\zeta,d}(1 - g(x_i, x_j))$$

where $C_{\beta,\zeta,d}$ is a constant dependent on the bandwidths β, ζ and the input dimension d.

We obtain this result by explicitly computing the MMD between the component distributions. Theorem 4 is then an immediate consequence of Lemma 1, which states that every kernel based data-clustering algorithm can equivalently be formulated as a kernel-based density clustering procedure (see Figure 2).

Theorem 4 (**Equivalence between kernel data–clustering and $\mathcal{A}_{\mathrm{KDE}}$**). *Any Gaussian kernel-based (data) clustering algorithm can equivalently be formulated as a clustering of the component KDE distributions with respect to the MMD metric corresponding to a Gaussian kernel for appropriately chosen bandwidth parameters.*

This simple result is consequential for practical considerations such as in the choice of bandwidth parameter for kernel data clustering (see Section 6) as well as for theoretical considerations. As it turns out, the density clustering procedure ($\mathcal{A}_{\mathrm{KDE}}$) of the component KDE

distributions can be used to define an estimator of the true mixing measure, that is, true component distributions and the corresponding weights. The equivalence between the two procedures, therefore, allows us to derive consistency guarantees for the estimators by analyzing the corresponding kernel-based clustering algorithms.

5 CONSISTENCY OF ESTIMATING MIXTURE MODELS

Estimation procedure. By an estimation procedure, we refer to any algorithm that takes a sample X drawn according to some mixing measure Λ, that is, $X \sim m(\Lambda)^n$ and provides an estimate $\widehat{\Lambda}$ of Λ.

Identifiability. Identifiability is also a key issue for estimation. Similar to our analysis of non-parametric clustering, we circumvent an *explicit* analysis of identifiability. Moreover, in the preceding discussion, identifiability is defined as a deterministic property of a set of mixing measures. We introduce a ***statistical*** notion of identifiability which can be defined as a property of either a mixing measure or a set of mixing measures. Additionally, in contrast to identifiability, statistical identifiability is defined with respect to an algorithm and therefore, it is a more intuitive and natural definition in the analysis of estimation procedures. Intuitively, the set of all mixing measures which are identifiable with respect to an estimation procedure $\mathcal{E}$ encodes the ***inductive bias*** of $\mathcal{E}$.

Definition 5.1 (**Statistical identifiability**). Let ϱ be some metric defined on the space of all mixing measures $\mathcal{P}_K^2$. A mixing measure Λ is statistically identifiable with respect to an estimation procedure $\mathcal{E}$ if the sequence of mixing measures $\left\{\widehat{\Lambda}_n = \mathcal{E}(X_n)\right\}$ converges in probability to Λ, where given $X_n \sim m(\Lambda)^n$.

Furthermore, a set of mixing measures $\mathfrak{L} \subset \mathcal{P}_K^2$ is said to be statistically identifiable with respect to estimation procedure $\mathcal{E}$ if every mixing measure $\Lambda \in \mathfrak{L}$ is statistically identifiable with respect to $\mathcal{E}$.

Remark. The convergence of the mixing measures can be defined with respect to any metric on $\mathcal{P}_K^2$. In our results, we show convergence with respect to the Wasserstien distance between mixing measures (see the supplementary for a definition).

Estimation procedure based on kernel-based data clustering. We describe the procedure to define an estimator of the true mixing measure Λ. This procedure is illustrated in Figure 3. As usual, for some $\beta, \zeta > 0$, denote the Gaussian kernel with bandwidth parameter $4\beta^2 + \zeta > 0$ by g. The component probability distributions of the KDE ψ_i are also defined in the usual way with respect to the bandwidth parameter $\beta > 0$. Given a sample $X_n \sim m(\Lambda)^n$,

(a) By means of a kernel-based data clustering procedure, with respect to g, obtain a partition $\widehat{\sigma} : [n] \to [K]$ of X_n.

(b) Use $\widehat{\sigma}$ to define a partition of component KDE distributions $\{\psi_i\}_{i=1}^n$.

(c) The estimator is defined as $\widehat{\Lambda}_n = \sum_{i=1}^{K} \widehat{\lambda}_{k,\widehat{\sigma}} \delta_{\widehat{\gamma}_{k,\widehat{\sigma}}}$, where $\widehat{\gamma}_{k,\widehat{\sigma}} = \frac{1}{|c_k|} \sum_{x_i \in c_k} \psi_i$ and $\widehat{\lambda}_{k,\widehat{\sigma}} = \frac{|c_k|}{n}$.

Let $\mathcal{E}_{\text{CTR}}$, $\mathcal{E}_{\text{FFK}}$, and $\mathcal{E}_{\text{LNK}}$ denote the estimation procedures corresponding to the kernel data clustering algorithms, $\mathcal{A}_{\text{CTR}}$, $\mathcal{A}_{\text{FFK}}$, and $\mathcal{A}_{\text{LNK}}$ respectively: the estimation procedure that uses the respective kernel clustering algorithm to obtain a partition $\widehat{\sigma}$ in (a). Theorem 5 then immediately follows from the recovery guarantees for the corresponding kernel-based clustering algorithms (Theorem 2) and the equivalence between kernel data clustering and density clustering established in Theorem 4. We show that any mixing measure satisfying the conditions provided in (4) is statistically identifiable with respect to the estimation procedures corresponding to $\mathcal{A}_{\text{CTR}}$, $\mathcal{A}_{\text{FFK}}$, and $\mathcal{A}_{\text{LNK}}$.

Theorem 5 (Statistical identifiability with respect to $\mathcal{E}_{\text{CTR}}$, $\mathcal{E}_{\text{FFK}}$, and $\mathcal{E}_{\text{LNK}}$). *Let ζ and β be bandwidth parameters satisfying the conditions provided in Theorem 2. Then any $\Lambda \in \mathcal{P}_K^2$ satisfying the conditions provided in (4) is statistically identifiable with respect to $\mathcal{E}_{CTR}$, $\mathcal{E}_{FFK}$, and $\mathcal{E}_{LNK}$.*

Estimating the Bayes partition. For theoretical considerations, it might be of interest to analyze conditions under which kernel-clustering algorithms can consistently estimate the Bayes partition. Given a finite sample $X = \{x_1, x_2, \cdots x_n\}$, let $\widehat{\sigma}$ denote the partition generated by a kernel clustering algorithm $\mathcal{A}$. We can define an estimator of the Bayes partition function $\widehat{\sigma}_b : \mathbb{R}^d \to [K]$ in the natural way:

$$\widehat{\sigma}_b(x) = \arg\sup_{k \in [K]} \sum_{j:\widehat{\sigma}(j)=k} G_\beta(x, x_j) \stackrel{(*)}{=} \arg\sup_{k \in [K]} \widehat{\lambda}_{k,\widehat{\sigma}} \widehat{f}_{k,\widehat{\sigma}}(x) \tag{6}$$

where $(*)$ follows from Lemma 1. Due to the equivalence between kernel clustering and density-based clustering, we can show that if a kernel-based algorithm $\mathcal{A}$ can consistently recover the planted partition, then by means of a single reassignment step given by (6), the algorithm consistently recovers the Bayes partition.

Exceptional set. Given $\Lambda = \sum_{k \in [K]} \lambda_k \delta_{\gamma_k}$, for any $t > 0$, we define the exceptional set

$$E(t) = \bigcup_{k \neq k'} \left\{ x \in \mathbb{R}^d : |\lambda_k f_k(x) - \lambda_{k'} f_{k'}(x)| \leq t \right\}.$$

Theorem 6 (Estimating the Bayes partition). *Let ζ, and β be bandwidth parameters satisfying the conditions provided in Theorem 2. Let $\Lambda \in \mathcal{P}_K^2$ satisfying the conditions provided in (4). For $X = \{x_1, x_2, \cdots x_n\} \sim m(\Lambda)^n$ and let $\widehat{\sigma}_{b,n}$ be the partition function obtained by $\mathcal{A}_{CTR}$, $\mathcal{A}_{FFK}$ or $\mathcal{A}_{LNK}$ followed by the reassignment step in (6). Then, w.h.p over the samples, there exists a sequence $\{t_n\} \stackrel{n \to \infty}{\longrightarrow} 0$ such that $\widehat{\sigma}_n(x) = \sigma_{Bayes}(x)$ for all $x \in \mathbb{R}^d - E_0(t_n)$.*

6 DISCUSSION AND FUTURE WORK

We show in this work that certain kernel-based clustering algorithms can exploit separability conditions to overcome identifiability. Our results also show that strong separability conditions are indeed necessary for provable recovery guarantees for clustering methods under non-parametric conditions. To further elaborate, we highlight a conceptually interesting insight from our results, which is surprising on the first glance. Even though kernel-based FFk-means++, which is a relaxation of the NP-Hard kernel k-means can provably recover the true clusters under the sufficient separability conditions (Theorem 2), our impossibility result (Theorem 3) shows that the NP-Hard kernel k-means algorithm fails to (provably) do so. This clearly shows that for better recovery guarantees for a clustering algorithm $\mathcal{A}$, in the non-parametric setting, it is essential to thoroughly characterize the inductive bias of the $\mathcal{A}$, that is, the set of mixing measures for which $\mathcal{A}$ can recover the true clustering.

We also established a key connection between kernel data clustering and distribution clustering when using Gaussian kernels and MMD as a metric between the distributions. As a consequence, we can interpret any standard Gaussian kernel clustering algorithm as a distribution clustering procedure. This is particularly useful in theoretical analysis since, for instance, we can

analyze kernel clustering algorithms by analyzing the corresponding distribution clustering procedure and vice versa. This connection could also have practical implications on matters such as bandwidth selection for kernel clustering.

Extending our results beyond the Gaussian kernel. We believe that the relationship between kernel data clustering and density clustering can indeed be established for a larger class of kernel functions. For instance, choosing kernel functions from conjugate families is one way in which the analysis could possibly be extended to other kernels, that is, choosing the MMD kernel function as the conjugate prior of the kernel function used for density estimation. It would also be of significant interest to characterize the class of kernels for which the equivalence can be established. However, a detailed study in this direction is reserved for future work.

Bandwidth. There is little to no literature that provides a systematic approach to bandwidth selection for kernel-based clustering. In contrast to kernel clustering, bandwidth selection is a well studied problem in the context of kernel density estimation (Giné et al., 2002; Einmahl et al., 2005; Goldenshluger et al., 2011; Chacón et al., 2013). By appropriating bandwidth selection strategies from this work, we provide the following guidance in **bandwidth selection for kernel-based data clustering**. As it would be expected, our analysis suggests that the bandwidth parameter used for kernel-clustering $(4\beta^2 + \zeta)$ needs to decrease with n since our sufficient conditions for recovery require that $\beta \xrightarrow{n \to \infty} 0$. Interestingly, however, it suggests that the bandwidth parameter can asymptotically remain non-zero since ζ is chosen to be a fixed parameter greater than 0. We note that these conditions are asymptotic and a more thorough analysis of the convergence rates of the estimators is necessary to provide the rate at which the bandwidth needs to reduce with sample size. Moreover, the range of the bandwidth parameter, which depends on the constant terms, could be be data-dependent. We conducted few small-sample experiments, and observed that the dependence of clustering performance on bandwidth is complex and requires more thorough investigation. We leave this analysis for future work.

References

Aragam, Bryon, Chen Dan, Eric P Xing, Pradeep Ravikumar, et al. (2020). "Identifiability of nonparametric mixture models and bayes optimal clustering". In: *The Annals of Statistics* 48.4, pp. 2277–2302.

Bruni, Carlo and Giorgio Koch (1985). "Identifiability of continuous mixtures of unknown Gaussian distributions". In: *The Annals of Probability*, pp. 1341–1357.

Chacón, José E and Tarn Duong (2013). "Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting". In: *Electronic Journal of Statistics* 7, pp. 499–532.

Chaudhuri, Kamalika, Sanjoy Dasgupta, Samory Kpotufe, and Ulrike Von Luxburg (2014). "Consistent procedures for cluster tree estimation and pruning". In: *IEEE Transactions on Information Theory* 60.12, pp. 7900–7912.

Couillet, Romain and Florent Benaych-Georges (2016). "Kernel spectral clustering of large dimensional data". In: *Electronic Journal of Statistics* 10.1, pp. 1393–1454.

Dhillon, Inderjit S, Yuqiang Guan, and Brian Kulis (2004). "Kernel k-means: spectral clustering and normalized cuts". In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 551–556.

Einmahl, Uwe and David Mason (2005). "Uniform in bandwidth consistency of kernel-type function estimators". In: *The Annals of Statistics* 33.3, pp. 1380–1403.

Fukumizu, Kenji, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf (2008). "Kernel measures of conditional dependence". In: *Advances in Neural Information Processing Systems*, pp. 489–496.

Giné, Evarist and Armelle Guillou (2002). "Rates of strong uniform consistency for multivariate kernel density estimators". In: *Annales de l'Institut Henri Poincare (B) Probability and Statistics*. Vol. 38. 6. Elsevier, pp. 907–921.

Goldenshluger, Alexander and Oleg Lepski (2011). "Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality". In: *The Annals of Statistics* 39.3, pp. 1608–1632.

Gretton, Arthur, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola (2012). "A kernel two-sample test". In: *Journal of Machine Learning Research* 13.Mar, pp. 723–773.

Hartigan, John A (1975). *Clustering algorithms*. John Wiley & Sons, Inc.

Hartigan, John A (1981). "Consistency of single linkage for high-density clusters". In: *Journal of the American Statistical Association* 76.374, pp. 388–394.

Holzmann, Hajo, Axel Munk, and Tilmann Gneiting (2006). "Identifiability of finite mixtures of elliptical distributions". In: *Scandinavian Journal of Statistics* 33.4, pp. 753–763.

Kimeldorf, George S and Grace Wahba (1970). "A correspondence between Bayesian estimation on stochastic processes and smoothing by splines". In: *The Annals of Mathematical Statistics* 41.2, pp. 495–502.

Luxburg, Ulrike von, Mikhail Belkin, and Olivier Bousquet (2008). "Consistency of spectral clustering". In: *The Annals of Statistics*, pp. 555–586.

Miao, Wang, Peng Ding, and Zhi Geng (2016). "Identifiability of normal and normal mixture models with non-ignorable missing data". In: *Journal of the American Statistical Association* 111.516, pp. 1673–1683.

Muandet, Krikamol, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf (2016). "Kernel mean embedding of distributions: A review and beyond". In: *arXiv preprint arXiv:1605.09522*.

Ng, Andrew Y, Michael I Jordan, and Yair Weiss (2002). "On spectral clustering: Analysis and an algorithm". In: *Advances in Neural Information Processing Systems*, pp. 849–856.

Nguyen and XuanLong (2013). "Convergence of latent mixing measures in finite and infinite mixture models". In: *The Annals of Statistics* 41.1, pp. 370–400.

Rinaldo, Alessandro, Larry Wasserman, et al. (2010). "Generalized density clustering". In: *The Annals of Statistics* 38.5, pp. 2678–2722.

Schiebinger, Geoffrey, Martin J Wainwright, Bin Yu, et al. (2015). "The geometry of kernelized spectral clustering". In: *Annals of Statistics* 43.2, pp. 819–846.

Sriperumbudur, Bharath, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert Lanckriet (2010). "Hilbert space embeddings and metrics on probability measures". In: *Journal of Machine Learning Research* 11.Apr, pp. 1517–1561.

Sriperumbudur, Bharath and Ingo Steinwart (2012). "Consistency and rates for clustering with dbscan". In: *Artificial Intelligence and Statistics*, pp. 1090–1098.

Teicher, Henry (1963). "Identifiability of finite mixtures". In: *The Annals of Mathematical statistics*, pp. 1265–1269.

Vandermeulen, Robert A and Clayton D Scott (2015). "On the identifiability of mixture models from grouped samples". In: *arXiv preprint arXiv:1502.06644*.

Vankadara, Leena C and Debarghya Ghoshdastidar (2020). "On the optimality of kernels for high-dimensional clustering". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 2185–2195.

Yan, Bowei and Purnamrita Sarkar (2016). "On robustness of kernel clustering". In: *Advances in Neural Information Processing Systems*, pp. 3098–3106.