Supplementary to Non-parametric kernel clustering

A Equivalence between Kernel-based data clustering and Kernel-based density clustering.

A.1 Proof of Lemma 1

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.

Proof. Squared MMD $\rho^2(\psi_i, \psi_j)$ with respect to the Gaussian kernel g_ζ can be decomposed as follows:

$$\rho^2(\psi_i, \psi_j) = ||\mu_{\psi_i}||^2_{\mathcal{H}_{g_\zeta}} + ||\mu_{\psi_j}||^2_{\mathcal{H}_{g_\zeta}} - 2\langle \mu_{\psi_i}, \mu_{\psi_j} \rangle_{\mathcal{H}_{g_\zeta}}, \tag{1}$$

where μ_{ψ_j} denotes the kernel mean embedding of ψ_i with respect to the Gaussian kernel function g_ζ which can be computed in closed form as shown in (2).

$$\begin{aligned} \mu_{\psi_j}(\cdot) &= \int_{\mathbb{R}^d} \frac{1}{(2\pi\beta^2)^{d/2}} \exp\left(\frac{-\|x - \cdot\|^2}{\zeta}\right) \exp\left(\frac{-\|x_j - \cdot\|^2}{2\beta^2}\right) dx \\ &= \left(\frac{\zeta}{\zeta + 2\beta^2}\right)^{d/2} \exp\left(-\frac{\|x_j - \cdot\|^2}{2\beta^2 + \zeta}\right) \end{aligned} \tag{2}$$

By means of theorem 1 which provides a spectral characterization of the Gaussian RKHS and the inner-product within, we compute $\langle \mu_{\psi_i}, \mu_{\psi_j} \rangle_{\mathcal{H}_{g_\zeta}}, \forall i, j \in [n]$. The computation uses the closed form expressions of Fourier transforms of the kernel function and the kernel mean embeddings of the component kde distributions given in (3). The closed form expression for the inner product between the kernel mean embeddings of any two component kde distributions is given in Equation (4).

$$\begin{aligned} \mathcal{F}[g_\zeta](\omega) &= \left(\frac{\zeta}{2}\right)^{d/2} \exp\left(\frac{-\|\omega\|^2 \zeta}{4}\right). \\ \mathcal{F}[\mu_{\psi_i}](\omega) &= \left(\frac{\zeta}{2}\right)^{d/2} \exp\left(\frac{-\|\omega\|^2 (2\beta^2 + \zeta)}{4}\right) \exp\left(\mathbf{i} \sum_{l \in [d]} x_i^l \omega^l\right), \end{aligned} \tag{3}$$

where $\mathbf{i}$ denotes the imaginary unit and satisfies $\mathbf{i}^2 = -1$.

$$\begin{aligned}\langle \mu_{\psi_i}, \mu_{\psi_j} \rangle_{\mathcal{H}_{g_\zeta}} &= \frac{1}{(2\pi)^{d/2}} \int \frac{\mathcal{F}[\mu_{\psi_i}](\omega)\overline{\mathcal{F}[\mu_{\psi_i}](\omega)}}{\mathcal{F}[g_\zeta](\omega)} d\omega \\ &= \left(\frac{\zeta}{4\beta^2+\zeta}\right)^{d/2} \exp\left(\frac{-\|x_i - x_j\|^2}{4\beta^2+\zeta}\right)\end{aligned} \tag{4}$$

Substituting the values of $\langle \mu_{\psi_i}, \mu_{\psi_j} \rangle_{\mathcal{H}_{g_\zeta}}$ for any $i, j \in [n]$ we obtain

$$\rho^2(\psi_i, \psi_j) = 2\left(\frac{\zeta}{4\beta^2+\zeta}\right)^{d/2} (1 - g(x_i, x_j)) \tag{5}$$

□

The following result given by Kimeldorf et al. (1970) and Wendland (2004) provides a spectral characterization of the RKHS corresponding to any translation-invariant kernel.

Theorem 1 (**Spectral characterization of RKHS. (Kimeldorf et al., 1970; Wendland, 2004)**). *Let k be a translation-invariant kernel on $\mathbb{R}^d$ such that $k(x, y) := \psi(x - y)$ where $\Phi \in C(\mathbb{R}^d) \cap L_1(R^d)$. Then the corresponding RKHS $\mathcal{H}$ is given by*

$$\mathcal{H} = \left\{ f \in L_2(\mathbb{R}^d) \cap C(\mathbb{R}^d) : \ \|f\|^2_{\mathcal{H}_g} = \frac{1}{(2\pi)^{d/2}} \int \frac{|\mathcal{F}[f](\omega)|^2}{\mathcal{F}[\psi](\omega)} d\omega < \infty \right\}, \tag{6}$$

where $|\cdot|$ denotes the magnitude of the enclosed quantity and $\mathcal{F}[f](\omega)$ denotes the Fourier transform of the function f. The inner product on $\mathcal{H}$ is defined as $\langle f, g \rangle_{\mathcal{H}} = \frac{1}{(2\pi)^{d/2}} \int \frac{\mathcal{F}[f](\omega)\overline{\mathcal{F}[g](\omega)}}{\mathcal{F}[\psi](\omega)} d\omega, \quad f, g \in \mathcal{H}$, where $\overline{\mathcal{F}[g](\omega)}$ denotes the complex conjugate of $\mathcal{F}[g](\omega)$.

A.2 Proof of Theorem 4

Theorem 4 immediately follows from Lemma 1. For any data clustering algorithm with respect to the Gaussian kernel $\eta > 0$, decompose η into any two positive quantities $\beta, \zeta > 0$ satisfying $\eta = 4\beta^2 + \zeta$. Due to Lemma 1, the kernel clustering algorithm equivalently defines a clustering of the component kde distributions $\{\psi_i\}_{i=1}^n$.

B Algorithms

For completeness, we briefly describe the kernel-based clustering algorithms ($\mathcal{A}_{\text{KMN}}$, $\mathcal{A}_{\text{CTR}}$, $\mathcal{A}_{\text{FFK}}$, and $\mathcal{A}_{\text{LNK}}$) here. In each of the algorithms, we describe the standard kernel data clustering procedure as well as the equivalent kernel density clustering procedures (see Theorem 4). The component kde distributions are defined in the usual way with respect to the bandwidth parameter $\beta > 0$ and ρ is defined with respect to the Gaussian kernel with bandwidth parameter $\zeta > 0$.

B.1 Kernel k-means ($\mathcal{A}_{\mathbf{KMN}}$)

Algorithm - Kernel k-means

- Given: A sample $X = \{x_1, x_2, \cdots x_n\} \subset \mathbb{R}^d$ and for some $\beta, \zeta > 0$ the Gaussian kernel function $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with bandwidth parameter $4\beta^2 + \zeta$.
- Find the partition

$$\widehat{\sigma} = \underset{\sigma:[n]\to[K]}{\arg\max} \sum_{k\in[K]} \sum_{i,j\in c_k} g(x_i, x_j) = \underset{\sigma:[n]\to[K]}{\arg\min} \sum_{k\in[K]} \sum_{i\in c_k} \rho(\mu_{\psi_i}, \frac{1}{|c_k|} \sum_{j\in c_k} \mu_{\psi_j})^2 \tag{7}$$

B.2 FFk-means++ ($\mathcal{A}_{\mathbf{FFK}}$)

Algorithm - Farthest first Kernel k-means ++

Phase one: Initializing the centers

- Given: A sample $X = \{x_1, x_2, \cdots x_n\} \subset \mathbb{R}^d$ and for some $\beta, \zeta > 0$ the Gaussian kernel function $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with bandwidth parameter $4\beta^2 + \zeta$.
- Choose an initial center c_1 uniformly at random and set $C = \{c_1\}$.
- While $t < K$:
  - let $C = \{c_1, c_2, \cdots c_{t-1}\}$ be the current set of centers,
  - for each $x \in X$, compute $d(x) = \min_{c\in C} k(x, c) = \max_{c\in C} \rho(\psi_x, \psi_c)$
  - pick the new center $c_t = \underset{x\in X}{\arg\max}\, d(x)$, and set $C = C \cup \{c_t\}$.
- For each $k \in [K]$:
  - set

$$\begin{aligned} C_k &= \{x \in X : k(x, c_k) \geq k(x, c_{k'}) \ \forall k \neq k' \in [K]\} \\ &= \left\{x \in X : \rho(\psi_x, \psi_{c_k}) \leq \rho(\psi_x, \psi_{c'_k}) \ \forall k \neq k' \in [K]\right\} \end{aligned}$$

Phase two: Standard kernel k-means algorithm

1. For each $k \in [K]$, set $C_k = \{x \in X : \text{condition (8) holds}\}$

$$\frac{1}{|C_k|^2} \sum_{y,z\in C_k} k(x, z) - \frac{1}{|C_k|} \sum_{y\in C_k} k(y, x) \leq \frac{1}{|C_l|^2} \sum_{y,z\in l} k(y, z) - \frac{1}{|C_l|} \sum_{y\in C_l} k(y, x) \ \forall l \neq k \in [K]. \tag{8}$$

$$(8) \iff \rho(\psi_x, \frac{1}{|C_k|} \sum_{x'\in C_k} \psi_{x'}) \leq \rho(\psi_x, \frac{1}{|C_l|} \sum_{x'\in C_l} \psi_{x'}) \ \ \forall l \neq k \in [K]. \tag{9}$$

2. Repeat step (1) until convergence, that is, the set of centers C do not change anymore.

B.3 Kernel K-center($\mathcal{A}_{\mathbf{CTR}}$)

Algorithm - Kernel K-center

- Given: A sample $X = \{x_1, x_2, \cdots x_n\} \subset \mathbb{R}^d$ and for some $\beta, \zeta > 0$ the Gaussian kernel function $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with bandwidth parameter $4\beta^2 + \zeta$.
- Find the partition

$$\begin{aligned}\widehat{\sigma} &= \underset{\sigma:[n]\to[K]}{\arg\max} \inf_{l\in[n]} \frac{-1}{|c_k^l|^2} \sum_{i,j\in c_k^l} k(x_i, x_j) + \frac{1}{|c_k^l|} \sum_{i\in c_k^l} k(x_i, x_l) \\ &= \underset{\sigma:[n]\to[K]}{\arg\min} \max_{i\in[n]} \rho(\psi_i, \widehat{\gamma}_{\sigma(i),\sigma})\end{aligned}$$

B.4 Agglomerative hierarchical clustering ($\mathcal{A}_{\mathbf{LNK}}$)

Given a sample $X = \{x_1, x_2, \cdots x_n\} \subset \mathbb{R}^d$ and a similarity function $S : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, hierarchical clustering algorithms seek to generate a cluster tree (dendrogram) establishing a hierarchy of relationships between the elements of the sample. Aggolomerative methods, in contrast to divisive methods, seek a bottom up approach, starting out with each point as its own cluster and progressively combining them into larger clusters until there is a single cluster that contains all the elements of the sample X. The criterion for merging hinges on the underlying similarity function, which in our case is the kernel matrix computed on the sample for a given kernel function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. We discuss two of the popular hierarchical clustering algorithms that exist in literature: **single linkage** and **complete linkage** methods. The distinguishing factor across the two methods is the choice of the criterion C used to merge any two clusters $c, c' \subset X$ $(c \cap c' = \varnothing)$, which are given below in 10.

$$C(c, c') = \underbrace{\max_{x\in c, y\in c'} k(x, y) = \min_{x\in c, y\in c'} \rho(\psi_x, \psi_y)}_{\textbf{Single linkage}}, \quad \text{and} \quad \underbrace{\min_{x\in c, y\in c'} k(x, y) = \max_{x\in c, y\in c'} \rho(\psi_x, \psi_y)}_{\textbf{Complete linkage}}. \tag{10}$$

By substituting the different criterion $C(c, c')$ to merge any two clusters c, c' in Algorithm 1, we obtain variants of the corresponding algorithms.

C Impossibility of recovery by kernel k-means(Proof of Theorem 1)

Proof. Fix the kernel bandwidth parameter $\zeta > 0$. Consider the following example in $\mathbb{R}$, where $\mathcal{U}([a, b])$ denotes the uniform distribution on the real interval $[a, b]$. Let

$$\gamma_1 = m\left(\frac{1}{2}\mathcal{U}([-\epsilon, \epsilon]) + \frac{1}{2}\mathcal{U}([r - \epsilon, r + \epsilon])\right) \tag{11}$$

and

$$\gamma_2 = \mathcal{U}([Dr - \epsilon, Dr + \epsilon]). \tag{12}$$

Algorithm 1: Agglomorative hierarchical kernel-clustering.

Given: A sample $X = \{x_1, x_2, \cdots x_n\} \subset \mathbb{R}^d$ and for some $\beta, \zeta > 0$ the Gaussian kernel function $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with bandwidth parameter $4\beta^2 + \zeta$;
Let $\mathcal{S} = \{s_1, \dots s_n\}$ be a collection of singleton trees with the root node of $s_i = \{i\}$.
while $|\mathcal{S}| > 1$ **do**
  Let $s_q, s_r \in \mathcal{S}$ be the pair of trees such that $C(root(s_q), root(s_r))$ is maximal ;
  Generate s_{qr} s.t, $root(s_{qr}) = root(s_q) \cup root(s_r)$, $left, right(s_{qr}) = s_q, s_r$;
  Add s_{qr} and remove s_q and s_r from $\mathcal{S}$;
end
$\hat{\sigma} \leftarrow$ Partition function obtained by cutting the only element in $\mathcal{S}$, a dendrogram at a level such that the resulting partition contains K clusters ;
return $\hat{\sigma}$;

The mixing measure is given by $\Lambda = \lambda_1\gamma_1 + \lambda_2\gamma_2$. The constants $D \gg 2 \gg r \gg \epsilon$ and $\lambda_1 \gg \lambda_2$ are to be chosen later. The idea is that the interval $[Dr - \epsilon, Dr + \epsilon]$ is separated from the rest of the distribution via a large constant D, but the points in $[Dr - \epsilon, Dr + \epsilon]$ will nevertheless be clustered with the points in $[r - \epsilon, r + \epsilon]$ because λ_2 is so small. We first show that Λ satisfies the condition in the theorem, namely that

$$\frac{\rho^2(\gamma_1, \gamma_2)}{\sup\limits_{x \in X_n} \rho^2(\psi_x, \widehat{\gamma}_{\sigma^*(x),\sigma^*})} > K^2. \tag{13}$$

Therefore, consider the numerator, which is simply the squared MMD between γ_1 and γ_2. We have

$$\begin{aligned}\rho^2(\gamma_1, \gamma_2) &= \mathbb{E}_{X\sim\gamma_1, \tilde{X}\sim\gamma_1} g(X, \tilde{X}) + \mathbb{E}_{Y\sim\gamma_2, \tilde{Y}\sim\gamma_2} g(Y, \tilde{Y}) - 2\mathbb{E}_{X\sim\gamma_1, Y\sim\gamma_2} g(X, Y) \\ &\geq \frac{1}{(2\epsilon)^2} \int_{[-\epsilon,\epsilon]^2} e^{-|x-y|^2/\zeta}\, dx\, dy - \frac{2}{(2\epsilon)^2} \int_{[-\epsilon,\epsilon]^2} e^{-|(D-1)r+x-y|^2/\zeta}\, dx\, dy.\end{aligned}$$

At this point, assume that ϵ is sufficiently small compared to the kernel bandwidth parameter ζ, namely that $4\epsilon^2 < \eta$. This allows us to lower bound the first integral by $\frac{1}{e}$. Similarly, choosing D large enough in comparison to r allows us to make the second term arbitrarily small, whence we conclude that

$$\rho^2(\gamma_1, \gamma_2) \geq \frac{1}{e} - \frac{1}{2e} \geq \frac{1}{2e},$$

i.e. the numerator is at least $\frac{1}{2e}$. Now consider the denominator, which is the maximum squared MMD between an empirical cluster mean and a sampled point belonging to that cluster. This is at most the squared MMD between any two points belonging to the same cluster

$$\sup_{x \in X_n} \rho^2(\psi_x, \frac{1}{|\sigma^*(x)|} \sum_{y \in \sigma^*(x)} \psi_y) \leq \sup_{x,y \in X_n, \sigma^*(x) = \sigma^*(y)} \rho^2(\psi_x, \psi_y) \tag{14}$$

which can be bound, independently of the sample X_n, by

$$\begin{aligned}\rho^2(\psi_0, \psi_{r+2\epsilon}) &= 2\sqrt{\frac{\zeta}{4\beta^2 + \zeta}} \left(1 - e^{\frac{-(r+2\epsilon)^2}{4\beta^2+\zeta}}\right) \\ &\leq 2\frac{(r + 2\epsilon)^2}{\zeta} + o(r^4).\end{aligned} \tag{15}$$

Here $r + 2\epsilon$ is the maximum distance of any two points belonging to the same cluster and we used (5). Thus, choosing a small r allows us to make the denominator arbitrarily small, and the fraction in (13) can become larger than any fixed K^2.
Now, we show that k-means does w.h.p. not recover the planted partition. The idea is to choose $\lambda_1 \gg \lambda_2$. In our sample X_n from $m(\Lambda)$, denote the number of points within $[-\epsilon, \epsilon]$ by N_1, the number of points within within $[r-\epsilon, r+\epsilon]$ by N_2, and the number of points within $[Dr-\epsilon, Dr+\epsilon]$ by N_3. Assume that n is large enough s.t. $N_1, N_2, N_3 > 0$. We rely on the equivalence between kernel-based data clustering and kernel-based density clustering and directly consider the MMD between component distributions ψ_{x_i} (compare section B.1). That is we consider k-means w.r.t. the norm $\|\cdot\|^2 = <\cdot,\cdot>_{\mathcal{H}_{g_\zeta}}$. The k-means objective of the planted partition is at least

$$N_1 \left\| \mu_{\psi_\epsilon} - \frac{N_1\,\mu_{\psi_{-\epsilon}} + N_2\,\mu_{\psi_{r-\epsilon}}}{N_1+N_2} \right\|^2 + N_2 \left\| \mu_{\psi_{r-\epsilon}} - \frac{N_1 \mu_{\psi_\epsilon} + N_2 \mu_{\psi_{r+\epsilon}}}{N_1+N_2} \right\|^2 \geq \frac{N_1 N_2}{N_1+N_2} \left\| \mu_{\psi_\epsilon} - \mu_{\psi_{r-\epsilon}} \right\|^2 + O(\epsilon).$$

Similarly, the k-means objective of the alternative partition where the points in $[r-\epsilon, r+\epsilon]$ and $[Dr-\epsilon, Dr+\epsilon]$ form a cluster is at most

$$\begin{aligned}
&N_1 \left\| \mu_{\psi_0} - \mu_{\psi_{2\epsilon}} \right\|^2 + N_2 \left\| \mu_{\psi_{r-\epsilon}} - \frac{N_2 \mu_{\psi_{r+\epsilon}} + N_3 \mu_{\psi_{Dr+\epsilon}}}{N_2+N_3} \right\|^2 + N_3 \left\| \mu_{\psi_{Dr+\epsilon}} - \frac{N_2 \mu_{\psi_{r-\epsilon}} + N_3 \mu_{\psi_{Dr-\epsilon}}}{N_2+N_3} \right\|^2 \\
\leq & N_1 \left\| \mu_{\psi_0} - \mu_{\psi_{2\epsilon}} \right\|^2 + \frac{N_2 N_3}{N_2+N_3} \left\| \mu_{\psi_{r-\epsilon}} - \mu_{\psi_{Dr+\epsilon}} \right\|^2 + O(\epsilon).
\end{aligned}$$

Thus, k-means will choose the alternative partition if

$$\begin{aligned}
N_1 \left\| \mu_{\psi_0} - \mu_{\psi_{2\epsilon}} \right\|^2 + \frac{N_2 N_3}{N_2+N_3} \left\| \mu_{\psi_{r-\epsilon}} - \mu_{\psi_{Dr+\epsilon}} \right\|^2 + O(\epsilon) &\leq \frac{N_1 N_2}{N_1+N_2} \left\| \mu_{\psi_\epsilon} - \mu_{\psi_{r-\epsilon}} \right\|^2 \\
\iff \frac{\left\| \mu_{\psi_{r-\epsilon}} - \mu_{\psi_{Dr+\epsilon}} \right\|^2}{\left\| \mu_{\psi_\epsilon} - \mu_{\psi_{r-\epsilon}} \right\|^2} + O(\epsilon) &\leq \frac{N_1}{N_3} \frac{N_2+N_3}{N_1+N_2} - \frac{N_1(N_2+N_3)}{N_2 N_3} \frac{\left\| \mu_{\psi_0} - \mu_{\psi_{2\epsilon}} \right\|^2}{\left\| \mu_{\psi_\epsilon} - \mu_{\psi_{r-\epsilon}} \right\|^2} \\
\iff \frac{\left\| \mu_{\psi_{r-\epsilon}} - \mu_{\psi_{Dr+\epsilon}} \right\|^2}{\left\| \mu_{\psi_\epsilon} - \mu_{\psi_{r-\epsilon}} \right\|^2} + O(\epsilon) &\leq \frac{N_1}{N_3} \left(\frac{N_2+N_3}{N_1+N_2} - \left(1 + \frac{N_3}{N_2}\right) \frac{\left\| \mu_{\psi_0} - \mu_{\psi_{2\epsilon}} \right\|^2}{\left\| \mu_{\psi_\epsilon} - \mu_{\psi_{r-\epsilon}} \right\|^2} \right).
\end{aligned} \tag{16}$$

□

First note that the norms in equation (16) are deterministic quantities that depend on ϵ, r and D. The N_i are Binomial random variables parametrized by λ_1 and λ_2, i.e. $N_1 \sim \text{Binom}(n, \lambda_1/2)$, $N_2 \sim \text{Binom}(n, \lambda_1/2)$ and $N_3 \sim \text{Binom}(n, \lambda_2)$. All terms involving $N_i's$ w.h.p. concentrate around their expectation. Thus, choosing $\lambda_1 \gg \lambda_2$ allows us to make the fraction $\frac{N_1}{N_3}$ w.h.p. arbitrarily large. Choosing ϵ small enough (in comparison to r) ensures that the $O(\epsilon)$ term on the LHS is small enough, and that the bracketed term on the RHS is at least $\frac{1}{4}$.

D Sufficient conditions for Consistency of $\mathcal{A}_{\mathbf{CTR}}$, $\mathcal{A}_{\mathbf{FFK}}$, and $\mathcal{A}_{\mathbf{LNK}}$. (Proof of Theorem 2)

Proof of Theorem 2: Consistency of $\mathcal{A}_{CTR}$. Let Λ be any mixing measure for which there exists some $\epsilon > 0$ such that,

$$\mathbb{P}_{X_n}\Big(\frac{1}{4}\inf_{k\neq k'}\rho(\gamma_k,\gamma_{k'}) < \sup_{x\in X_n}\rho(\psi_x,\widehat{\gamma}_{\sigma^*(x),\sigma^*}) + \epsilon\Big) \overset{n\to\infty}{\longrightarrow} 0. \tag{17}$$

Then, with high probability (w.h.p) over the samples X_n,

$$\inf_{k\neq k'}\rho(\gamma_k,\gamma_{k'}) > 4\sup_{x\in X_n}\rho(\psi_x,\widehat{\gamma}_{\sigma^*(x),\sigma^*}) + 4\epsilon. \tag{18}$$

If the bandwidth parameter β is chosen according to (19),

$$\beta \to 0, \quad \frac{n\beta^d}{\log n} \to \infty \text{ as } n \to \infty, \tag{19}$$

it is known that the corresponding kernel density estimate $\hat{f}_n$ converges to the true density f in the l_∞ norm (Giné et al., 2002; Einmahl et al., 2005). Observe that the density functions $\widehat{f}_{k,\sigma^*}$ corresponding to the planted partitions $\widehat{\gamma}_{k,\sigma^*}$ are the kernel density estimates of the density functions corresponding to the component distributions γ_k. Furthermore by assumption, we have that the corresponding component weights λ_k are bounded away from 0. Thus, for each $k \in [K]$, we have

$$\sup_{x\in\mathbb{R}^d}|\widehat{f}_{k,\sigma^*} - f_k| \xrightarrow{\mathbb{P}} 0 \text{ as } n \to \infty.$$

An application of Scheffe's theorem (or Reiz's theorem) (Scheffé, 1947) implies that the corresponding probability measures $\widehat{\gamma}_{k,\sigma^*}$ also converge weakly to γ_k. Simon-Gabriel, Barp, et al. (2020, Theorem 4.2) provide a characterization of the class of kernels that metrize the weak convergence of probability measures on locally compact domains (e.g., $\mathbb{R}^d$). Following Simon-Gabriel and Schölkopf (2016, Corollary 3) and Sriperumbudur et al. (2010, Proposition 5), one can verify that the Gaussian kernel belongs to this class of kernel functions. Therefore, weak convergence of probability measures $\widehat{\gamma}_{k,\sigma^*}$ to γ_k is equivalent to convergence in MMD with respect to (w.r.t) a Gaussian kernel, that is, for every $\epsilon > 0$,

$$\mathbb{P}(\rho(\widehat{\gamma}_{k,\sigma^*},\gamma_k) > \epsilon) \overset{n\to\infty}{\longrightarrow} 0. \tag{20}$$

Let $t = 4\epsilon/2$ and $\delta = 1/n$. Then, for every $k \in [K]$, there exists some $N_t \in \mathbb{N}$ such that $\forall\ n > N_{t,k}$,

$$\mathbb{P}(\rho(\widehat{\gamma}_{k,\sigma^*},\gamma_k) > 4\epsilon/2) < \frac{1}{n}. \tag{21}$$

Let $N_t = \sup_{k\in[K]} N_{t,k}$. For all $n > N_t$, with high probability (w.h.p) over the samples X_n,

$$\inf_{k\neq k'}\rho(\gamma_k,\gamma_{k'}) > 4\sup_{x\in X_n}\rho(\psi_x,\widehat{\gamma}_{\sigma^*(x),\sigma^*}) + 2\rho(\widehat{\gamma}_{k,\sigma^*},\gamma_k). \tag{22}$$

By assumption, we have that λ_k is bounded away from 0 for all $k \in [K]$. Therefore,

$$\mathbb{P}(\min_{k\in[K]}|(\sigma^*)^{-1}(k)| > 0) = \prod_{k=1}^{K}\mathbb{P}(|(\sigma^*)^{-1}(k)| > 0). \tag{23}$$

For any $k \in [K]$, observe that $|(\sigma^*)^{-1}(k)|$ is a binomial random variable, $Bin(n, \lambda_k)$. Using Hoeffding's inequality for binomial random variables,

$$\mathbb{P}(|(\sigma^*)^{-1}(k)| \leq t) < \exp\left(-2n(\lambda_k - \frac{t}{n})^2\right) \tag{24}$$

Setting $t = 0$, for large enough n such that $n/\log n > 1/\lambda_k$, w.p.a.l $1 - 1/n$

$$|(\sigma^*)^{-1}(k)| > 0$$

So w.h.p over the samples,

$$\min_{k\in[K]} |(\sigma^*)^{-1}(k)| > 0$$

From Propositions 1, 2, and 3, we then have that w.h.p over X_n, the algorithms $\mathcal{A}_{\text{CTR}}$, $\mathcal{A}_{\text{FFK}}$, and $\mathcal{A}_{\text{LNK}}$can recover the planted partition σ^* (upto a permutation over the labels). □

D.1 Sufficient conditions for consistency of kernel k-center clustering $\mathcal{A}_{\text{CTR}}$

Proposition 1 (**Conditions for recovery of the true partition by kernel k-center algorithm**). *For any $\Lambda \in \mathcal{P}_K^2$, let $\Gamma = m(\Lambda)$. Let $X = \{x_1, x_2, \cdots x_n\} \sim \Gamma^n$. Define $\widehat{\Gamma} = \sum_{i=1}^{n} \frac{1}{n}\psi_i$ as the probability measure associated with the kde in the usual way. For any partition $\sigma : [n] \to [K]$ such that the following condition holds:*

$$\inf_{k\neq k'} \rho(\gamma_k, \gamma_k') > 4 \sup_{i\in[n]} \rho(\psi_i, \widehat{\gamma}_{\sigma(i),\sigma}) + 2 \sup_{k\in[K]} \rho(\widehat{\gamma}_{k,\sigma}, \gamma_{k,\sigma}), \tag{25}$$

and

$$\inf_{k\in[K]} |\sigma^{-1}(k)| > 0 \tag{26}$$

σ can be recovered by the kernel k-center algorithm on the sample kernel matrix G (defined in section 4 of the main paper).

Proof of Proposition 1. For any sample $X = \{x_1, x_2, \cdots x_n\}$ and a partition σ', let

$$r = \sup_{i\in[n]} \rho(\psi_i, \widehat{\gamma}_{\sigma'(i),\sigma'}) \tag{27}$$

We first show that for any mixing measure satisfying the conditions provided in Equation (25) w.r.t a sample X and a partition σ', then for any $i \neq j \in [n]$,

$$\rho(\psi_i, \psi_j) \leq 2r \iff \sigma'(i) = \sigma'(j)$$
$$\rho(\psi_i, \psi_j) > 2r \iff \sigma'(i) \neq \sigma'(j)$$

1) $\sigma'(i) = \sigma'(j) \implies \rho(\psi_i, \psi_j) \leq 2r$. For any $i \in [n]$, by definition,

$$\rho(\psi_i, \widehat{\gamma}_{\sigma'(i),\sigma'}) \leq r \tag{28}$$

Therefore, for any $i, j \in [n]$,

$$\sigma'(i) = \sigma'(j) \implies \rho(\psi_i, \psi_j) \leq \rho(\psi_i, \widehat{\gamma}_{\sigma'(i),\sigma'}) + \rho(\widehat{\gamma}_{\sigma'(i),\sigma'}, \psi_j) \leq 2r \tag{29}$$

□

2) $\sigma'(i) \neq \sigma'(j) \implies \rho(\psi_i, \psi_j) > 2r$. Let $\sigma'(i) = k \neq k' = \sigma'(j)$. Then, by triangle inequality,

$$\rho(\psi_i, \psi_j) \geq \rho(\gamma_k, \gamma_{k'}) - \rho(\gamma_k, \widehat{\gamma}_{k,\sigma'}) - \rho(\widehat{\gamma}_{k,\sigma'}, \psi_i) - \rho(\psi_j, \widehat{\gamma}_{k',\sigma'}) - \rho(\widehat{\gamma}_{k',\sigma'}, \gamma_{k'}) > 2r \tag{30}$$

Combining Equations (29) and (30), its easy to verify that

$$\begin{aligned} \rho(\psi_i, \psi_j) \leq 2r &\iff \sigma'(i) = \sigma'(j) \\ \rho(\psi_i, \psi_j) > 2r &\iff \sigma'(i) \neq \sigma'(j) \end{aligned}$$

For any partition σ, let

$$L(\sigma) = \sup_{i \in [n]} \rho(\psi_i, \widehat{\gamma}_{\sigma(i),\sigma}). \tag{31}$$

Then the partition $\widehat{\sigma}$ generated by the kernel k-center clustering algorithm is given by

$$\widehat{\sigma} = \operatorname*{arg\,min}_{\sigma:[n]\to[K]} L(\sigma). \tag{32}$$

Then, by definition,

$$L(\widehat{\sigma}) \leq L(\sigma') = r \tag{33}$$

Therefore, from (33),

$$\rho(\widehat{\gamma}_{\sigma'(i),\sigma'}, \widehat{\gamma}_{\widehat{\sigma}(i),\widehat{\sigma}}) \leq \rho(\widehat{\gamma}_{\sigma'(i),\sigma'}, \psi_i) + \rho(\widehat{\gamma}_{\widehat{\sigma}(i),\widehat{\sigma}}) \leq 2r \tag{34}$$

To show that the partitions σ' and $\widehat{\sigma}$ coincide up to a permutation, we show that, for any $i, j \in [n]$, $\sigma'(i) = \sigma'(j) \implies \widehat{\sigma}(i) = \widehat{\sigma}(j)$ and $\sigma'(i) \neq \sigma'(j) \implies \widehat{\sigma}(i) \neq \widehat{\sigma}(j)$.

Consider $i, j \in [n]$ such that $\sigma'(i) \neq \sigma'(j)$. If $\widehat{\sigma}(i) = \widehat{\sigma}(j)$, then from triangle inequality and (34),

$$\rho(\widehat{\gamma}_{\sigma'(i),\sigma'}, \widehat{\gamma}_{\sigma'(j),\sigma'}) \leq \rho(\widehat{\gamma}_{\sigma'(i),\sigma'}, \widehat{\gamma}_{\widehat{\sigma}(i),\widehat{\sigma}}) + \rho(\widehat{\gamma}_{\sigma'(j),\sigma'}, \widehat{\gamma}_{\widehat{\sigma}(i),\widehat{\sigma}}) \leq 4r. \tag{35}$$

However, from (25) we have that

$$\rho(\widehat{\gamma}_{\sigma'(i),\sigma'}, \widehat{\gamma}_{\sigma'(j),\sigma'}) \geq \rho(\gamma_{\sigma'(i)}, \gamma_{\sigma'(j)}) - \rho(\widehat{\gamma}_{\sigma'(i),\sigma'}, \gamma_{\sigma'(i)}) - \rho(\widehat{\gamma}_{\sigma'(j),\sigma'}, \gamma_{\sigma'(j)}) > 4r, \tag{36}$$

which is a contradiction. Therefore, for any $i, j \in [n]$ such that

$$\sigma'(i) \neq \sigma'(j) \implies \widehat{\sigma}(i) \neq \widehat{\sigma}(j). \tag{37}$$

Consider any $i, j \in [n]$ such that $\sigma'(i) = \sigma'(j)$ but $\widehat{\sigma}(i) \neq \widehat{\sigma}(j)$. From (34) we know that

$$\widehat{\gamma}_{\widehat{\sigma}(i),\widehat{\sigma}} \in B(\widehat{\gamma}_{\sigma'(i),\sigma'}, 2r) \text{ and } \widehat{\gamma}_{\widehat{\sigma}(j),\widehat{\sigma}} \in B(\widehat{\gamma}_{\sigma'(i),\sigma'}, 2r) \tag{38}$$

where $B(x, r) = \{y : \rho(x, y) \leq r\}$ denotes the ball of radius r centered at x.

From the condition (44) that the clusters are non-empty, for each $k \in [K]$, there exists a_k such that $\sigma'(a_k) = k$. Then, for each $k \in [K]$, we know that

$$\widehat{\gamma}_{\widehat{\sigma}(a_k),\widehat{\sigma}} \in B(\widehat{\gamma}_{\sigma'(a_k),\sigma'}, 2r) = B(\widehat{\gamma}_{k,\sigma'}, 2r) \tag{39}$$

Furthermore, observe that for all $k \neq k' \in [K]$,

$$B(\widehat{\gamma}_{k,\sigma'}, 2r) \cap B(\widehat{\gamma}_{k',\sigma'}, 2r) = \varnothing, \tag{40}$$

since otherwise there exists some $x \in B(\widehat{\gamma}_{k,\sigma'}, 2r) \cap B(\widehat{\gamma}_{k',\sigma'}, 2r)$, i.e.,

$$\begin{aligned} &\rho(x, \widehat{\gamma}_{k,\sigma'}) \leq 2r \text{ and } \rho(x, \widehat{\gamma}_{k',\sigma'}) \leq 2r, \\ \implies &\rho(\widehat{\gamma}_{k,\sigma'}, \widehat{\gamma}_{k',\sigma'}) \leq \rho(x, \widehat{\gamma}_{k,\sigma'}) + \rho(x, \widehat{\gamma}_{k',\sigma'}) \leq 4r, \end{aligned}$$

which is a contradiction.

Moreover, by definition, $\sigma'(a_k) \neq \sigma'(a_{k'})$ for all $k, k' \in [K]$, from (37), we have

$$\widehat{\sigma}(a_1) \neq \widehat{\sigma}(a_2) \cdots \neq \widehat{\sigma}(a_K) \tag{41}$$

Since there are only K centers, (39), (40) and (41) imply that

- For any $i \in [n]$, there exists some $k \in [K]$ such that $\widehat{\sigma}(i) = \widehat{\sigma}(a_k)$, and
- $\widehat{\gamma}_{\widehat{\sigma}(a_k),\widehat{\sigma}} \in B(\widehat{\gamma}_{\sigma'(i),\sigma'}, 2r) \implies \widehat{\gamma}_{\widehat{\sigma}(a_{k'}),\widehat{\sigma}} \notin B(\widehat{\gamma}_{\sigma'(i),\sigma'}, 2r)$ for all $k' \neq k \in [K]$.

So, from (38),

$$\sigma'(i) = \sigma'(j) \implies \widehat{\gamma}_{\widehat{\sigma}(i),\widehat{\sigma}} = \widehat{\gamma}_{\widehat{\sigma}(j),\widehat{\sigma}} \implies \widehat{\sigma}(i) = \widehat{\sigma}(j), \tag{42}$$

since, if $\widehat{\sigma}(i) \neq \widehat{\sigma}(j)$, then $\rho(\widehat{\gamma}_{\widehat{\sigma}(i),\widehat{\sigma}}, \widehat{\gamma}_{\widehat{\sigma}(j),\widehat{\sigma}}) > 4r$.

Therefore, the partitions σ' and $\widehat{\sigma}$ coincide up to a permutation over the labels.

D.2 Sufficient conditions for kernel kmeans++ algorithm - proofs

Proposition 2 (**Sufficient conditions for recovery by kernel k-means ++**). *For any $\Lambda \in \mathcal{P}_K^2$, let $\Gamma = m(\Lambda)$. Let $X = \{x_1, x_2, \cdots x_n\} \sim \Gamma^n$. Define $\widehat{\Gamma} = \sum_{i=1}^{n} \frac{1}{n}\psi_i$ as the probability measure associated with the kde in the usual way. For any partition $\sigma' : [n] \to [K]$ such that the following condition holds:*

$$\inf_{k \neq k'} \rho(\gamma_k, \gamma_k') > 4 \sup_{i \in [n]} \rho(\psi_i, \widehat{\gamma}_{\sigma'(i),\sigma'}) + 2 \sup_{k \in [K]} \rho(\widehat{\gamma}_{k,\sigma'}, \gamma_{k,\sigma'}), \tag{43}$$

and

$$\inf_{k \in [K]} |(\sigma')^{-1}(k)| > 0 \tag{44}$$

σ can be recovered by a (deterministic) kernel k-means++ algorithm on the sample kernel matrix G.

Proof of Proposition 2. Let,

$$r = \sup_{i \in [n]} \rho(\psi_i, \widehat{\gamma}_{\sigma'(i),\sigma'}), \text{ and } B_k = B(\widehat{\gamma}_{k,\sigma'}, r) \ \forall k \in [K]. \tag{45}$$

Claim: Let C be the set of centers initialized in phase one of the k-means ++ algorithm as described. Then, for each $k \in [K]$,

$$c_k \in B_k \tag{46}$$

Proof: For every $i \in [n]$, by definition,

$$\rho(\psi_i, \widehat{\gamma}_{\sigma'(i),\sigma'}) \leq r \implies \psi_i \in B_{\sigma'(i)}. \tag{47}$$

Therefore, without loss of generality (W.L.O.G), let $c_1 \in B_1$. For any $t < K$, assume that $C_t = \{c_1, c_2, \cdots c_t\}$ and $c_k \in B_k \ \ \forall k \in [t]$ (upto a permutation over the labels). Note that B_k is non-empty for every $k \in [K]$.
From the proof of Proposition 1, for any mixing measure satisfying the conditions provided in (43),

$$\rho(\psi_i, \psi_j) \leq 2r \iff \sigma'(i) = \sigma'(j) \tag{48}$$
$$\rho(\psi_i, \psi_j) > 2r \iff \sigma'(i) \neq \sigma'(j) \tag{49}$$

Therefore, since $c_k \in B_k$ for all $k \in [K]$, $d(\psi_i) = \rho^2(\psi_i, c_k) \leq 2r$ for all $\sigma'(i) = k$. Therefore,

$$d(\psi_i) \text{ is } \begin{cases} \leq 2r & \forall \psi_i \in B_k, \text{ and } k \leq t, \\ > 2r & \text{otherwise.} \end{cases} \tag{50}$$

Since $c_{t+1} = \arg\max_{\psi_i} d(\psi_i)$, $c_{t+1} \in B_s$ for some $s \notin C_t$.

■

Claim: Kernel k-means algorithm does not affect the centers obtained in Phase one of the algorithm.
Proof: From claim 1, in phase one of the algorithm, the centers $C = \{c_1, c_2, \cdots c_K\}$ are obtained such that $c_k \in B_k$ for all $k \in [K]$. For each $k \in [K]$, clusters $\{C_1, C_2, \cdots C_K\}$ are then defined as follows.

$$C_k = \left\{ i \in [n] : \rho^2(c_k, \psi_i) \geq \rho^2(c_{k'}, \psi_i) \quad \forall k \neq k' \in [K] \right\} \tag{51}$$

From (48), we have that

$$\rho^2(\psi_i, c_k) \leq 4r^2 \quad \text{if } \sigma'(i) = k$$
$$\rho^2(\psi_i, c_k) > 4r^2 \quad \text{otherwise .}$$

Therefore, the partition obtained in the Phase 1 of the algorithm coincides with σ' up to a permutation over the labels, that is,

$$C_k = \{\psi_i \in X : \sigma'(i) = k\}, \tag{52}$$

and

$$\sum_{i:\sigma'(i)=k} \psi_i = \widehat{\gamma}_{k,\sigma'} \in B_k. \tag{53}$$

Clearly,

$$\rho(\psi_i, \widehat{\gamma}_{\sigma'(i),\sigma'}) \leq 2r \leq \rho(\psi_i, \widehat{\gamma}_{k,\sigma'}) > 2r \ \forall k \neq \sigma'(i).$$

Therefore, the clusters obtained in the phase 1 of the algorithm do not change in the Phase 2 of the algorithm and the partition obtained by $\mathcal{A}_{\text{FFK}}$coincides with that of σ' up to a permutation over the labels.

■ □

D.3 Sufficient conditions for kernel linkage clustering algorithms (Proof of Theorem 2 - Part III)

Proposition 3 (Recovery by single linkage clustering). *For any $\Lambda \in \mathcal{P}_K^2$, let $\Gamma = m(\Lambda)$. Let $X_n = \{x_1, x_2, \cdots x_n\} \sim \Gamma^n$ be a sample. Define $\widehat{\Gamma} = \sum_{i=1}^{n} \frac{1}{n} \psi_i$ as the probability measure associated with the kde in the usual way. For any partition σ_n such that the following condition holds:*

$$\inf_{k \neq k'} \rho(\gamma_k, \gamma_k') > 3 \sup_k \sup_{l \neq l' \in \sigma_n^{-1}(k)} \rho(\psi_l, \psi_{l'}) + 2 \sup_{k \in [K]} \rho(\widehat{\gamma}_{k,\sigma_n}, \gamma_{k,\sigma_n}), \tag{54}$$

σ_n can be recovered by the kernel single (and complete) linkage clustering algorithms with respect to the Gaussian kernel with bandwidth para using the sample kernel matrix G (defined in section 4 of the main paper).

Proof of proposition 3. For any partition σ, let

$$\delta = \sup_{k\in[K]} \sup_{i,j'\in\sigma^{-1}(k)} \rho(\psi_i, \psi_j).$$

We first show that for any partition σ satisfying the conditions stated in Proposition 3,

$$\begin{aligned} \forall l, l' \in [n] \qquad \sigma(l) = \sigma(l') &\iff \rho(\psi_l, \psi_{l'}) \leq \delta, \\ \sigma(l) \neq \sigma(l') &\iff \rho(\psi_l, \psi_{l'}) > \delta. \end{aligned}$$

Observe that, by definition,

$$\forall l \neq l' \in [n], \quad \sigma(l) = \sigma(l') \implies \rho(\psi_l, \psi_{l'}) \leq \delta. \tag{55}$$

By subadditivity of ρ, for any $l, l' \in [n]$ such that $\sigma(l) = k$, $\sigma(l') = k'$, and $k \neq k'$,

$$\rho(\gamma_k, \gamma_{k'}) < \rho(\gamma_k, \widehat{\gamma}_k) + \rho(\widehat{\gamma}_k, \psi_l) + \rho(\psi_l, \psi_{l'}) + \rho(\psi_{l'}, \widehat{\gamma}_{k'}) + \rho(\widehat{\gamma}_{k'}, \gamma_{k'}). \tag{56}$$

Substituting (54) in (56), we obtain

$$\sigma(l) \neq \sigma(l') \implies \rho(\psi_l, \psi_{l'}) > \delta. \tag{57}$$

Using the fact that $\rho(\cdot, \cdot) \geq 0$, from (55) and (57), we have

$$\begin{aligned} \forall l, l' \in [n] \qquad \sigma(l) = \sigma(l') &\iff \rho^2(\psi_l, \psi_{l'}) \leq \delta^2, \\ \sigma(l) \neq \sigma(l') &\iff \rho^2(\psi_l, \psi_{l'}) > \delta^2. \end{aligned}$$

All three linkage algorithms based on the matrix of squared MMD evaluations between the component distributions $\{\psi_l\}_{l=1}^n$ or alternatively using the sample kernel matrix G (see Lemma 1) would first group the components within the same cluster according to σ before grouping components belonging to different clusters according to σ. Therefore, thresholding the dendrogram to obtain exactly K clusters would recover the underlying partition σ upto a permutation over the labels. With a minor modification of the proof, it is easy to see that the Proposition also holds under separbility conditions provided in (43).

□

Proof of Theorem 5: Consistent recovery of the planted partition by $\mathcal{A}_{LNK}$. Let Λ be any mixing measure for which there exists some $\epsilon > 0$ such that,

$$\mathbb{P}_{X_n}\left(\sup_{\substack{x,x'\in X_n: \\ \sigma^*(x)=\sigma^*(x')}} \rho(\psi_x, \psi_{x'}) > \frac{1}{3} \inf_{k\neq k'} \rho(\gamma_k, \gamma_{k'}) - \epsilon \right) \overset{n\to\infty}{\longrightarrow} 0, \tag{58}$$

Then, with high probability (w.h.p) over the samples X_n,

$$\inf_{k\neq k'} \rho(\gamma_k, \gamma_{k'}) > 3 \sup_{\substack{x,x'\in X_n: \\ \sigma^*(x)=\sigma^*(x')}} \rho(\psi_x, \psi_{x'}) + 3\epsilon. \tag{59}$$

Furthermore, we know that for every $\epsilon > 0$,

$$\mathbb{P}(\rho(\widehat{\gamma}_{k,\sigma^*}, \gamma_k) > \epsilon) \overset{n\to\infty}{\longrightarrow} 0. \tag{60}$$

Let $t = 3\epsilon/2$ and $\delta = 1/n$. Then, for every $k \in [K]$, there exists some $N_t \in \mathbb{N}$ such that $\forall\ n > N_{t,k}$,

$$\mathbb{P}(\rho(\widehat{\gamma}_{k,\sigma^*}, \gamma_k) > 3\epsilon/2) < \frac{1}{n}. \tag{61}$$

Let $N_t = \sup_{k\in[K]} N_{t,k}$. For all $n > N_t$, with high probability (w.h.p) over the samples X_n,

$$\inf_{k\neq k'} \rho(\gamma_k, \gamma_{k'}) > 3 \sup_{\substack{x,x'\in X_n:\\ \sigma^*(x)=\sigma^*(x')}} \rho(\psi_x, \psi_{x'}) + 2\rho(\widehat{\gamma}_{k,\sigma^*}, \gamma_k). \tag{62}$$

From Proposition 3, we have that w.h.p over X_n, kernel single linkage clustering algorithm recovers the true partition σ^* (upto a permutation over the labels).

□

E Necessary conditions for consistency of $\mathcal{A}_{\mathbf{FFK}}$ and $\mathcal{A}_{\mathbf{LNK}}$. (Proof of Theorem 3)

E.1 Proof for $\mathcal{A}_{\mathbf{FFK}}$

Fix the kernel bandwidth parameter $\zeta > 0$. Let r, ϵ and K be small constants that satisfy $1 > r > 2K > 16\epsilon$. Consider the following example in $\mathbb{R}$, where $\mathcal{U}([a, b])$ denotes the uniform distribution on the real interval $[a, b]$. Let

$$\gamma_1 = m\left(\frac{1}{2}\mathcal{U}([-\epsilon, \epsilon]) + \frac{1}{2}\mathcal{U}([r-\epsilon, r+\epsilon])\right) \tag{63}$$

and

$$\gamma_2 = m\left(\frac{1}{2}\mathcal{U}([2r-K-\epsilon, 2r-K+\epsilon]) + \frac{1}{2}\mathcal{U}([3r-K-\epsilon, 3r-K+\epsilon])\right). \tag{64}$$

The mixing measure is given by $\Lambda = \frac{1}{2}\gamma_1 + \frac{1}{2}\gamma_2$. The idea is that because $K > 0$, the two clusters are just not separated enough.
To see that $\mathcal{A}_{\text{FFK}}$ fails to recover the planted partition with probability approaching $\frac{1}{2}$, consider the case where the first cluster center is initialized with a point $c_1 \in [r-\epsilon, r+\epsilon]$. The farthest first heuristic then chooses a second cluster center $c_2 \in [3r - K\epsilon, 3r - K + \epsilon]$. Since $K > 4\epsilon$, the initial clusters will be given by

$$C_1 = \{x : x \leq 2r - K + \epsilon\} \quad \text{and} \quad C_2 = \{x : x \geq 3r - K - \epsilon\}.$$

Consequently, in the first iteration of phase two of the algorithm (compare section B.2), the new cluster centers satisfy

$$\tilde{c}_1 \geq \frac{rN_2 + (2r-K)N_3}{N_1 + N_2 + N_3} - \epsilon \quad \text{and} \quad \tilde{c}_2 \geq 3r - K - \epsilon,$$

where N_i denotes the number of points within the respective intervals. Now the clusters themselves do not change if

$$(2r-K)+\epsilon-\tilde{c}_1 \leq \tilde{c}_2-(2r-K)-\epsilon$$

$$\iff \frac{2N_1+N_2}{N_1+N_2+N_3}r-\frac{N_1+N_2}{N_1+N_2+N_3}K \leq r-4\epsilon,$$

an event that occurs asymptotically almost surely as the N_i concentrate around their expectation. Conditional on this event, the algorithm terminates with clusters C_1 and C_2, i.e. it does not recover the planted partition. Due to symmetry, the same holds if the first cluster center is initialized with a point in $[2r-K-\epsilon, 2r-K+\epsilon]$. As $n\to\infty$, the probability to initialize the first cluster center with a point in either $[r-\epsilon, r+\epsilon]$ or $[2r-K-\epsilon, 2r-K+\epsilon]$ approaches $\frac{1}{2}$.
We now show that the condition in the theorem is satisfied, namely that as $n\to\infty$, it holds that

$$\frac{\rho(\gamma_1,\gamma_2)}{\sup\limits_{x\in X_n}\rho(\psi_x,\widehat{\gamma}_{\sigma^*(x),\sigma^*})} > 4-\hat{\epsilon}. \tag{65}$$

A simple way to evaluate the LHS is to express both numerator and denominator as sums of inner products between Gaussians. We have

$$\rho(\gamma_1,\gamma_2) \geq \rho(\widehat{\gamma}_{1,\sigma^*},\widehat{\gamma}_{2,\sigma^*})-\rho(\gamma_1,\widehat{\gamma}_{1,\sigma^*})-\rho(\gamma_2,\widehat{\gamma}_{2,\sigma^*}),$$

and as $n\to\infty$ and $\beta\to 0$, the latter two terms converge in probability to 0. Hence, for all $\epsilon_1>0$, it holds that

$$\rho^2(\gamma_1,\gamma_2) \geq \rho^2(\widehat{\gamma}_{1,\sigma^*},\widehat{\gamma}_{2,\sigma^*})-\epsilon_1.$$

Furthermore, since ρ^2 is bounded, for all n large enough

$$\rho^2(\gamma_1,\gamma_2) \geq \mathbb{E}\left[\rho^2(\widehat{\gamma}_{1,\sigma^*},\widehat{\gamma}_{2,\sigma^*})\right]-2\epsilon_1.$$

A straightforward if somewhat lengthy calculation shows that

$$\mathbb{E}\left[\rho^2(\hat{\gamma}_{1,\sigma^*},\hat{\gamma}_{2,\sigma^*})\right] \geq \frac{2}{\zeta}(2r-K)^2+O(\epsilon)+o(r^4). \tag{66}$$

Similarly, for the denominator,

$$\sup_{x\in X_n}\rho^2(\psi_x,\widehat{\gamma}_{\sigma^*(x),\sigma^*}) \leq \frac{2}{\zeta}\frac{1}{4}r^2+O(\epsilon). \tag{67}$$

Hence,

$$\begin{aligned}\frac{\rho^2(\hat{\gamma}_{1,\sigma^*},\hat{\gamma}_{2,\sigma^*})}{\sup\limits_{x\in X_n}\rho^2(\psi_x,\widehat{\gamma}_{\sigma^*(x),\sigma^*})} &\geq \frac{(2r-K)^2+O(\epsilon)+o(r^4)-2\epsilon_1}{\frac{1}{4}r^2+O(\epsilon)}\\ &\geq \frac{16-2\frac{K}{r}+O\left(\frac{\epsilon}{r^2}\right)+o(r^2)+\frac{2\epsilon_1}{r^2}}{1+O\left(\frac{\epsilon}{r^2}\right)}.\end{aligned}$$

Thus, in order to satisfy (65), we have to choose r small enough, and K, ϵ and ϵ_1 small enough in comparison to r. We now derive the expression for the numerator. First define the sets $I_1=\{x\in$

$X_n : x \in [-\epsilon, \epsilon]\}$, $I_2 = \{x \in X_n : x \in [r-\epsilon, r+\epsilon]\}$, $I_3 = \{x \in X_n : x \in [2r-K-\epsilon, 2r-K+\epsilon]\}$ and $I_4 = \{x \in X_n : x \in [3r-K-\epsilon, 3r-K+\epsilon]\}$. Denote $N_i = |I_i|$. We have

$$
\begin{aligned}
\rho^2(\hat{\gamma}_{1,\sigma_n^*}, \hat{\gamma}_{2,\sigma_n^*}) = &< \hat{\gamma}_{1,\sigma_n^*}, \hat{\gamma}_{1,\sigma_n^*} > + < \hat{\gamma}_{2,\sigma_n^*}, \hat{\gamma}_{2,\sigma_n^*} > -2 < \hat{\gamma}_{1,\sigma_n^*}, \hat{\gamma}_{2,\sigma_n^*} > \\
=&\frac{\sum_{x,y\in I_1} < \psi_x, \psi_y > +2\sum_{x\in I_1, y\in I_2} < \psi_x, \psi_y > + \sum_{x,y\in I_2} < \psi_x, \psi_y >}{(N_1+N_2)^2} \\
&+ \frac{\sum_{x,y\in I_3} < \psi_x, \psi_y > +2\sum_{x\in I_3, y\in I_4} < \psi_x, \psi_y > + \sum_{x,y\in I_4} < \psi_x, \psi_y >}{(N_3+N_4)^2} \\
&- 2\frac{\sum_{x\in I_1, y\in I_3} < \psi_x, \psi_y > + \sum_{x\in I_1, y\in I_4} + \sum_{x\in I_2, y\in I_3} < \psi_x, \psi_y > + \sum_{x\in I_2, y\in I_4} < \psi_x, \psi_y >}{(N_1+N_2)(N_3+N_4)} \\
\geq&\sqrt{\frac{\zeta}{\eta}}\Bigg[\frac{N_1^2(1-\frac{4\epsilon^2}{\eta}) + 2N_1N_2(1-\frac{(r+2\epsilon)^2}{\eta}) + N_2^2(1-\frac{4\epsilon^2}{\eta})}{(N_1+N_2)^2} \\
&+ \frac{N_3^2(1-\frac{4\epsilon^2}{\eta}) + 2N_3N_4(1-\frac{(r+2\epsilon)^2}{\eta}) + N_4^2(1-\frac{4\epsilon^2}{\eta})}{(N_3+N_4)^2} \\
&- 2\frac{N_1N_3(1-\frac{(2r-K-2\epsilon)^2}{\eta}) + N_1N_4(1-\frac{(3r-K-2\epsilon)^2}{\eta})}{(N_1+N_2)(N_3+N_4)} \\
&- 2\frac{N_2N_3(1-\frac{(r-K-2\epsilon)^2}{\eta}) + N_2N_4(1-\frac{(2r-K-2\epsilon)^2}{\eta})}{(N_1+N_2)(N_3+N_4)}\Bigg] + o(r^4)
\end{aligned}
$$

Where we used (4) and the Taylor expansion $e^x = 1 + x + o(x^2)$. The inequality sign stems from the fact that we have replaced the exact locations of sampled points with interval boundaries. Taking expectations,

$$
\begin{aligned}
\mathbb{E}\left[\rho^2(\hat{\gamma}_{1,\sigma_n^*}, \hat{\gamma}_{2,\sigma_n^*})\right] \geq&\sqrt{\frac{\zeta}{\eta}}\frac{1}{\eta}\Bigg[\frac{-4\epsilon^2 - 2(r+2\epsilon)^2 - 4\epsilon^2}{4} + \frac{-4\epsilon^2 - 2(r+2\epsilon)^2) - \epsilon^2}{4} \\
&+ 2\frac{(2r-K-2\epsilon)^2 + (3r-K-2\epsilon)^2}{4} + 2\frac{(r-K-2\epsilon)^2 + (2r-K-2\epsilon)^2}{4}\Bigg] + o(r^4) \\
=&\frac{2}{\eta}\sqrt{\frac{\zeta}{\eta}}(2r-K)^2 + O(\epsilon) + o(r^4).
\end{aligned}
$$

We now derive the expression for the denominator. By symmetry, it suffices to consider the case

$x \in [-\epsilon, \epsilon]$.

$$
\begin{aligned}
&\rho\left(\psi_x, \frac{1}{N_1+N_2}\left(\sum_{x'\in[-\epsilon,\epsilon]}\psi_{x'} + \sum_{x'\in[r-\epsilon,r+\epsilon]}\psi_{x'}\right)\right)\\
=&\frac{1}{N_1+N_2}||\sum_{x'\in[-\epsilon,\epsilon]}(\psi_{x'}-\psi_x) + \sum_{x'\in[r-\epsilon,r+\epsilon]}(\psi_{x'}-\psi_x)||\\
\leq&\frac{N_1}{N_1+N_2}\rho(\psi_{-\epsilon},\psi_{\epsilon}) + \frac{N_2}{N_1+N_2}\rho(\psi_{-\epsilon},\psi_{r+\epsilon})\\
\leq&\rho(\psi_{-\epsilon},\psi_{+\epsilon}) + \frac{N_2}{N_1+N_2}\rho(\psi_0,\psi_r)\\
=&\sqrt{2\sqrt{\frac{\zeta}{\eta}}\left(1-e^{-\frac{4\epsilon^2}{\eta}}\right)} + \frac{N_2}{N_1+N_2}\sqrt{2\sqrt{\frac{\zeta}{\eta}}\left(1-e^{-\frac{r^2}{\eta}}\right)}\\
\leq&\frac{N_2}{N_1+N_2}r\sqrt{\frac{2}{\eta}}\sqrt[4]{\frac{\zeta}{\eta}} + O(\epsilon)
\end{aligned}
$$

where we used (5) and the inequality $1-e^{-x} \leq x$. It follows that asymptotically almost surely

$$\sup_{x\in X_n}\rho^2(\psi_x, \widehat{\gamma}_{\sigma^*(x),\sigma^*}) \leq \frac{2}{\eta}\sqrt{\frac{\zeta}{\eta}}\frac{1}{4}r^2 + O(\epsilon).$$

E.2 Proof for $\mathcal{A}_{\mathbf{LNK}}$

Consider the same example as in the above proof for $\mathcal{A}_{\mathrm{FFK}}$. At first, a hierarchical linkage algorithm (compare section B.4) will merge all points within 2ϵ-intervals. This leaves us with 4 trees. Then, the linkage algorithm does *not* return the planted partition if the trees belonging to the intervals $[r-\epsilon, r+\epsilon]$ and $[2r-K\epsilon, 2r-K+\epsilon]$ are merged in the next step. For $r \gg K \gg \epsilon$, it can be easily seen that this is the case.

F Statistical identifiability with respect to $\mathcal{E}_{\mathbf{CTR}}$, $\mathcal{E}_{\mathbf{FFK}}$, and $\mathcal{E}_{\mathbf{LNK}}$

Proof of Theorem 5: Consistency implies statistical identifiability. Let Λ be
For appropriate choice of bandwidths, we know that

$$\lim_{n\to\infty}\rho(\widehat{\gamma}_{k,\sigma_n^*}, \gamma_k) \overset{\mathbb{P}}{=} 0 \qquad \text{and} \qquad \lim_{n\to\infty}|\widehat{\lambda}_{k,\sigma_n^*} - \lambda_k| \overset{\mathbb{P}}{=} 0. \tag{68}$$

From Aragam et al. (2020, Lemma A.3), convergence of component measures and the corresponding component weights implies that the sequence of estimators defined by $\widehat{\Lambda} = \sum_{i=1}^{K}\widehat{\lambda}_{k,\sigma_n^*}\delta_{\widehat{\gamma}_{k,\sigma_n^*}}$ converges in probability to the true mixing measure Λ w.r.t the Wasserstein metric. □

G Estimating 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:

$$\widehat{\sigma}_b(x) = \underset{k\in[K]}{\arg\sup} \sum_{j:\widehat{\sigma}(j)=k} G_\beta(x, x_j) \overset{(*)}{=} \underset{k\in[K]}{\arg\sup}\, \widehat{\lambda}_{k,\widehat{\sigma}} \widehat{f}_{k,\widehat{\sigma}}(x) \tag{69}$$

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 (69), 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 2 (**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 (17). 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 (69). Then, w.h.p over the samples, there exists a sequence $\{t_n\} \overset{n\to\infty}{\longrightarrow} 0$ such that $\widehat{\sigma}_n(x) = \sigma_{Bayes}(x)$ for all $x \in \mathbb{R}^d - E_0(t_n)$.*

Proof of Theorem 2. The proof of this Proposition is adapted with minor changes from the proof of Aragam et al. (2020, Theorem 5.2). For this reason, we borrow some of the notation from Aragam et al. (2020). Since Λ satisfies the separability conditions given in equation (58), from Theorem 2, we know that w.h.p over the samples the algorithms $\mathcal{A}_{\text{CTR}}$, $\mathcal{A}_{\text{FFK}}$, and $\mathcal{A}_{\text{LNK}}$recover the planted partition up to a permutation over the labels, that is, $\widehat{\sigma} = \sigma^*$. For appropriate choice of bandwidths, we know that w.h.p over the samples,

$$\lim_{n\to\infty} f_{k,\sigma^*} \overset{\mathbb{P}}{=} f_k, \tag{70}$$

where the convergence is defined pointwise and uniformly over $\mathbb{R}^d$.
Let,

$$t_n = 2 \sup_{k\in[K]} \sup_{x\in\mathbb{R}^d} |\widehat{\lambda}_{k,\sigma_n^*} \widehat{f}_{k,\sigma_n^*}(x) - \lambda_k f_k(x)| \geq 0. \tag{71}$$

From (70), we know that $t_n \xrightarrow{\mathbb{P}} 0$. Moreover, by definition, we have that

$$|\lambda_k f_k(x) - \lambda_{k'} f_{k'}(x)| > t_n \implies \lambda_{\sigma_{Bayes}(x)} f_{\sigma_{Bayes}(x)}(x) > \lambda_k f_k(x) + t_n \;\forall x \notin E_0(t_n),\; k \neq \sigma_{Bayes}(x). \tag{72}$$

Therefore, it follows that for any $x \in \mathbb{R}^D - E_0(t_n)$ and any $k \neq \sigma_{Bayes}(x)$,

$$\widehat{\lambda}_{\sigma_{Bayes}(x),\sigma_n^*} \widehat{f}_{\sigma_{Bayes}(x),\sigma_n^*}(x) \overset{(1)}{>} \lambda_{\sigma_{Bayes}(x)} f_{\sigma_{Bayes}(x)} - \frac{t_n}{2} \overset{(2)}{>} \lambda_k f_k(x) + \frac{t_n}{2} \overset{(3)}{>} \widehat{\lambda}_{k,\sigma_n^*} \widehat{f}_{k,\sigma_n^*}(x), \tag{73}$$

where, (1) and (3) follow from (71) and (2) follows from (72). This implies that $\widehat{\sigma}_b(x) = \underset{k\in[K]}{\arg\sup}\, \widehat{\lambda}_{k,\sigma^*} \hat{f}_{k,\sigma^*}(x) = \sigma_{Bayes}(x)$ for all $x \notin E_0(t_n)$. □