Label consistency in overfitted generalized k-means

Linfan Zhang
Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90095
linfanz@g.ucla.edu

Arash A. Amini
Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90095
aaamini@ucla.edu

Abstract

We provide theoretical guarantees for label consistency in generalized k-means problems, with an emphasis on the overfitted case where the number of clusters used by the algorithm is more than the ground truth. We provide conditions under which the estimated labels are close to a refinement of the true cluster labels. We consider both exact and approximate recovery of the labels. Our results hold for any constant-factor approximation to the k-means problem. The results are also model-free and only based on bounds on the maximum or average distance of the data points to the true cluster centers. These centers themselves are loosely defined and can be taken to be any set of points for which the aforementioned distances can be controlled. We show the usefulness of the results with applications to some manifold clustering problems.

1 Introduction

Consider the problem of clustering data points sampled according to some probability measure μ from a normed space $\mathcal{X}$ with norm $\|\cdot\|_{\mathcal{X}}$. In the ideal setting, the generalized k-means clustering minimizes the population cost function

$$Q(\xi;\mu) := \Big(\int \min_{1\le \ell\le L} \|x-\xi_\ell\|_{\mathcal{X}}^p d\mu(x)\Big)^{1/p} \tag{1}$$

where $\xi = (\xi_1,\dots,\xi_L) \in \mathcal{X}^L$ is a set of L vectors in $\mathcal{X}$, for some fixed integer L. In practical data analysis, we are given a sample $\{x_1,\dots,x_n\}$ drawn from μ and solve an empirical version of (1), namely,

$$\widehat{Q}(\xi) = Q(\xi;\mathbb{P}_n) := \Big(\frac{1}{n}\sum_{i=1}^n \min_{1\le \ell\le L} \|x_i-\xi_\ell\|_{\mathcal{X}}^p\Big)^{1/p}. \tag{2}$$

Here, $\mathbb{P}_n := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ is the empirical measure associated with the sample and δ_x is the point mass measure at x. The minimizer of $\widehat{Q}(\cdot)$ over $\mathcal{X}^L$ is denoted as $\widehat{\xi} = (\widehat{\xi}_1,\dots,\widehat{\xi}_L)$ and each point x_i is assigned a cluster label $\widehat{z}_i := \operatorname{argmin}_\ell \|x_i - \widehat{\xi}_\ell\|_{\mathcal{X}}$.

Meanwhile, we assume that each data point x_i also has a true cluster label $z_i \in [K] := \{1,\dots,K\}$ which is determined solely by an unknown data-generating process. These true labels are not necessarily related to the optimal solutions of (1) or (2). To distinguish the two, we refer to the clustering induced by (z_i) as the **true clustering**, while a clustering that minimizes the generalized k-means cost function (2), i.e., the clustering induced by $(\widehat{z}_i)$, is referred to as an **optimal k-means clustering**. In this paper, we consider the *label consistency* problem, that is, how close the labels $(\widehat{z}_i)$ estimated by k-means clustering are to the true labels (z_i). Note that we allow the number of k-means clusters L to be different from the true number of clusters K.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

In the above formulation, the case where $p = 2$, $\mathcal{X} = \mathbb{R}^d$ and $\|\cdot\|_{\mathcal{X}}$ is the Euclidean norm leads to the classical and widely used k-means problem. Much of the theoretical analysis of k-means has been performed in this case. Early work has focused on how close the optimization problems based on the empirical and ideal cost functions (2) and (1) are to each other, where the closeness is measured in terms of the recovered centers (i.e., $\widehat{\xi}$ and ξ) or the optimal value of the objective function.

Such consistency results are proved, for the global minimizers of (2), in the early work of [22, 29] and also in [30, 19] from the vector quantization perspective. These classical results do not directly apply to the performance of the k-means in practice, mainly because solving (2) is NP-hard and approximation methods are usually applied to deal with it. Also, considerations of the label consistency problem are absent from this line of work since no true clustering, external to the k-means problem, is assumed to exist.

More recently, there has been more interest in the consistency of practical k-means algorithms [14, 21] as well as the label consistency problem. Lu and Zhou [21] obtain sharp bounds on the label consistency of the Lloyd's algorithm [20] under a sub-Gaussian mixture model. Semidefinite programming (SDP) relaxation is another popular technique for deriving polynomial-time approximations to the k-means problem [28]. Its label consistency has been studied when data is generated from the stochastic ball model [4, 10], sub-Gaussian mixtures [25, 8, 9], the Stochastic Block Model (SBM) [9] and general models [18]. Convex clustering is another relaxation method whose label consistency has been discussed in [34, 27, 11, 31]. The literature on community detection in SBM, a network clustering problem, is also mainly focused on label consistency and inspires our work here; see [1, 33] for a review of those results. For label consistency in kernel spectral clustering, see [2].

In this paper, we study the label consistency of approximate solutions of the generalized k-means problem (2) when $L \geq K$. Our focus will be on the overfitted case where $L > K$. This is often relevant in practice since the data-generating process may have a natural number of clusters K that is unknown a priori. An example is the sub-Gaussian mixture with K components. More interesting examples are given in Section 3. All the aforementioned works on label consistency exclusively consider the correctly-fitted case $L = K$. We show that when the data-generating process admits a set of centers that satisfy certain separation conditions, estimated labels with $L \geq K$ clusters, are close to a *refinement* of the true labels. These bounds reduce to the label consistency criteria for $L = K$, but have no counterpart in the literature for $L > K$.

Overfitting in k-means is considered in [32, 23] where it is shown to improve the approximation factor (see Assumption 1(b)) of certain polynomial-time k-means algorithms. Analysis of the approximation factor is concerned with how close one can get to the optimal value of the k-means objective function. In contrast, we are concerned with the label recovering problem and not directly concerned with how well the objective function is approximated. Our work is also aligned with the recent trend of *beyond worst case* analysis of the NP-hard problems [6], where the performance of the algorithms are considered assuming that there are some meaningful structures in the data (e.g., true clusters). We refer to Remark 1 for a more detailed comparison with this literature.

Our results are algorithm-free in the sense that they apply to any algorithm that achieves a constant-factor approximation to the optimal objective. They are also model-free in the sense that we do not make any explicit assumption on the data-generating process. This is important in practice, since many common data models, such as sub-Gaussian mixtures, are often too simplified to capture real clustering problems. We provide examples of more complicated data models in Section 3 and show how our general results can provide new insights for these models. Since k-means clustering often appears as a building block in many sophisticated clustering algorithms, we believe our results will be of broad interest in understanding the performance of clustering algorithms in overfitted settings.

Notation. $Q(\xi; \mu)$ is only dependent on the set of values among the coordinates of ξ. Although we view ξ as a vector (for which the order of elements matter), with some abuse of notation, we view $Q(\cdot; \mu)$ as a set function (mapping $2^{\mathcal{X}}$ to $\mathbb{R}$) that is only sensitive to the set of values represented by ξ. This justifies using the the same symbol for the function irrespective of the number of coordinates of ξ, i.e., the number of clusters. The reason to keep ξ as an (ordered) vector is to make the cluster labels well-defined. For simplicity, let $\|\cdot\| = \|\cdot\|_{\mathcal{X}}$. For the case where $\mathcal{X} \subset \mathbb{R}^d$, one often takes $\|\cdot\|$ to be the Euclidean norm, but our results are valid for any norm on $\mathbb{R}^d$, and more broadly any normed space $\mathcal{X}$.

2 Main Results

We first state assumptions about the k-means clustering algorithm.

Assumption 1. *Consider an algorithm for the generalized k-means problem* (2)*, referred to as ALG(p) hereafter, and let $\widehat{\xi}^{(L)} \in \mathcal{X}^L$ and $\widehat{\xi}^{(K)} \in \mathcal{X}^K$ be its estimated centers when applied with L and K clusters, respectively. Let $L \geq K$. Assume that ALG(p) has the following properties, for all input sequences (x_i):*

(a) Efficiency: The Voronoi cell of every estimated center $\widehat{\xi}^{(L)}_\ell$ contains at least one of (x_i).

(b) κ-approximation: $\widehat{Q}(\widehat{\xi}^{(K)}) \leq \kappa \cdot \min_{\xi \in \mathcal{X}^K} \widehat{Q}(\xi)$, and similarly with K replaced by L.

Efficiency can be achieved by substituting centers whose Voronoi cells have an empty intersection with $\{x_i\}$, by those having the opposite property. For κ-approximation, the factor κ can depend on the number of clusters K (or L). For example, the k-means++ algorithm has $\kappa = O(\log K)$, with high probability over the initialization [3]. However, there are also constant-factor approximation algorithms for k-means where $\kappa = O(1)$ independent of K (or L) [24, 13, 15]. For example, with local search, k-means++ can achieve a constant-factor approximation [16]. In addition, κ-approximation is not required for all inputs. That is, we are not concerned with the worst-case approximation factor. The κ in Assumption 1(b) is the approximation factor of the algorithm on the specific data under consideration. It is enough for an algorithm to achieve good approximation only on the data of interest.

For some of the results, Assumption 1(b) can be replaced with the following modified version: (b') κ-approximation only for K clusters plus a mononoticity assumption: $\widehat{Q}(\widehat{\xi}^{(L)}) \leq \widehat{Q}(\widehat{\xi}^{(K)})$. Mononoticity is also a reasonable requirement and obviously true for the exact k-means solutions.

Next, we extend the definition of the misclassification rate to the overfitted case.

Definition 1. The (generalized) misclassification rate between two label vectors $z \in [K]^n$ and $\widehat{z} \in [L]^n$, with $K \leq L$, is

$$\mathrm{Miss}(z, \widehat{z}) = \min_{\omega} \frac{1}{n} \sum_{i=1}^{n} 1\{z_i \neq \omega(\widehat{z}_i)\},$$

where the minimization ranges over all surjective maps $\omega : [L] \to [K]$.

When $L = K$, a surjective map ω is necessarily a bijection and the above becomes the usual definition of misclassification rate when the number of clusters is correctly identified. In this case, $\mathrm{Miss}(z, \widehat{z}) = 0$ means that there is a one-to-one correspondence between the estimated and true clusters. The generalized definition above allows us to extend this notion of exact recovery to the case $L > K$. In particular, $\mathrm{Miss}(z, \widehat{z}) = 0$ when $L > K$, if and only if $\widehat{z}$ is a *refinement* of z. To see this, note that $\mathrm{Miss}(z, \widehat{z}) = 0$ implies the existence of a map $\omega : [L] \to [K]$ such that $\omega(\widehat{z}_i) = z_i$ for all i. This in turn is equivalent to: $\widehat{z}_i = \widehat{z}_{i'} \implies z_i = z_{i'}$, which is the refinement relation for the associated clusters. In general, $\mathrm{Miss}(z, \widehat{z})$ is small if $\widehat{z}$ is close to a refinement of z.

We also use the (optimal) matching distances between elements of two vectors viewed as sets.

Definition 2. For $\xi \in \mathcal{X}^L$ and $\xi^* \in \mathcal{X}^K$, define the ℓ_∞ and ℓ_p optimal matching distances as

$$d_\infty(\xi, \xi^*) = \min_{\sigma} \max_{1 \leq k \leq K} \|\xi_{\sigma(k)} - \xi^*_k\|, \quad d_p(\xi, \xi^*) = \min_{\sigma} \Big(\sum_{k=1}^{K} \|\xi_{\sigma(k)} - \xi^*_k\|^p \Big)^{1/p},$$

where $\sigma : [K] \to [L]$ ranges over all injective maps.

For $K = L$, d_∞ is an upper bound on the Hausdorff distance between the two sets. Obviously, we have $d_\infty \leq d_p$ for any $p \geq 1$.

2.1 Distance to true centers

Let $z = (z_i)_{i=1}^n \in [K]^n$ be a given set of true labels for the data points $(x_i)_{i=1}^n$. In addition, our results are stated in terms of a set of vectors $\xi^* = (\xi^*_k)_{k=1}^K$ which we refer to as the "true centers".

Throughout, ξ^* will be only vaguely specified. The only requirement on ξ^* is that together with the observed data points (x_i) and the true labels (z_i), they satisfy the deviation bounds in each theorem, e.g., $\max_{1 \le i \le n} \|x_i - \xi^*_{z_i}\| \le \eta$ in Theorem 1, etc. Let $\pi_k = \sum_{i=1}^n 1\{z_i = k\}/n$ be the proportion of observed data points in true cluster k and let $\pi_{\min} = \min_k \pi_k$.

We let $\widehat{\xi}$ be a solution of the k-means algorithm with $L \ge K$ centers and let $\widehat{z}_i \in \operatorname{argmin}_\ell \|x_i - \widehat{\xi}_\ell\|$ be the corresponding estimated labels. Our first result provides guarantees for exact label recovery, in the extended sense of recovering a refinement of the true partition when $L > K$ and the exact partition when $L = K$.

Theorem 1 (Exact recovery)**.** *Consider a vector of (true) centers $\xi^* \in \mathcal{X}^K$ and labels $(z_i)_{i=1}^n \in [K]^n$. Pick $\eta, \delta > 0$ such that $\max_{1 \le i \le n} \|x_i - \xi^*_{z_i}\| \le \eta$, and*

$$\min_{(k,k') \colon k \ne k'} \|\xi^*_k - \xi^*_{k'}\| \ge \delta. \tag{3}$$

Consider an algorithm ALG(p) for problem (2)*, satisfying Assumption 1, and let $(\widehat{z}_i)_{i=1}^n \in [L]^n$ and $\widehat{\xi} \in \mathcal{X}^L$ be the estimated labels and centers of ALG(p) applied with the $L \ge K$. Then,*

$$\frac{\delta}{\eta} > 2\frac{(1+\kappa)}{\pi_{\min}^{1/p}} + 4 \quad \Longrightarrow \quad \mathrm{Miss}(z, \widehat{z}) = 0, \quad d_p(\widehat{\xi}, \xi^*) \le \frac{(1+\kappa)\eta}{\pi_{\min}^{1/p}}. \tag{4}$$

When $L = K$, the assertion $\mathrm{Miss} = 0$ means that there is a permutation σ on $[K]$ such that $\sigma(\widehat{z}_i) = z_i$ for all i, that is, we have the exact recovery of labels (z_i) in the classical sense. When $L > K$, Theorem 1 guarantees the exact recovery of a refinement of the true labels (z_i).

Example 1 (Stochastic Ball Model)**.** Assume that data are generated from the stochastic ball model considered in [26], where $x_i = \xi^*_{z_i} + r_i$ with r_i sampled independently from a distribution supported on the unit ball in $\mathbb{R}^d$. Here, $\{\xi^*_k\}_{k=1}^K \subset \mathbb{R}^d$ are a fixed set of centers. Clearly, we can take $\eta = 1$ in Theorem 1. Then, any κ-approximate k-means algorithm achieves exact recovery when $\delta > 2 + 2(1+\kappa)/\sqrt{\pi_{\min}}$ for $L = K$. In the overfitted case, when $\delta > 4 + 2(1+\kappa)/\sqrt{\pi_{\min}}$, the estimated label vector is an exact refinement of the true labels (z_i). □

In the above example, although it is intuitively clear that for a sufficiently large δ, the solution of the k-means problem should achieve exact label recovery (in the extended sense), Theorem 1 allows us to provide a provable guarantee for any constant-factor approximation, with an explicit bound on the separation parameter δ.

We now turn to approximate recovery where the misclassification rate is small.

Theorem 2 (Approximate Recovery)**.** *Consider a vector of (true) centers $\xi^* \in \mathcal{X}^K$ and labels $(z_i)_{i=1}^n \in [K]^n$. Pick $\varepsilon, \delta > 0$ such that $(\frac{1}{n}\sum_{i=1}^n \|x_i - \xi^*_{z_i}\|^p)^{1/p} \le \varepsilon$, and* (3) *holds. Consider an algorithm ALG(p) for problem* (2)*, satisfying Assumption 1, and let $(\widehat{z}_i)_{i=1}^n \in [L]^n$ and $\widehat{\xi} \in \mathcal{X}^L$ be the estimated labels and centers of ALG applied with the $L \ge K$. Then, for any $c > 2$,*

$$\frac{\delta}{\varepsilon} > \frac{(1+\kappa)c}{\pi_{\min}^{1/p}} \quad \Longrightarrow \quad \mathrm{Miss}(z, \widehat{z}) < K(1+\kappa)^p c^p \Big(\frac{\varepsilon}{\delta}\Big)^p, \quad d_p(\widehat{\xi}, \xi^*) \le \frac{(1+\kappa)\varepsilon}{\pi_{\min}^{1/p}}. \tag{5}$$

The key difference between Theorems 1 and 2 is the bounds assumed on the deviations $D_i := \|x_i - \xi^*_{z_i}\|, i \in [n]$. Theorem 1 assumes a bound on the maximum distance to true centers, $\max_i D_i$, while Theorem 2 assumes a bound on an average distance, $(\frac{1}{n}\sum_i D_i^p)^{1/p}$, leading to a more relaxed condition.

Theorem 2 provides an upper bound on the misclassification rate when a certain separation condition is satisfied. To simplify, consider the case $K = \kappa = p = 2$ and take $c = 2.1$. Then, Theorem 2 implies the following: For every $\beta > 0$, there exists a constant $c_1(\beta, \pi_{\min}) > 0$ such that if

$$\delta/\varepsilon \ge c_1(\beta, \pi_{\min}), \tag{6}$$

then any 2-factor k-means algorithm will have $\mathrm{Miss} \le \beta$ to the target labels. The next proposition shows that condition (6) is sharp up to constants.

Proposition 1. *There exists a family of datasets $\{(x_i, z_i)\}_{i=1}^n$, with $K = 2$ balanced true clusters (i.e., $\pi_{\min} = 1/2$) and parameterized by true center separation δ and $\varepsilon = (\frac{1}{n}\sum_{i=1}^n \|x_i - \xi^*_{z_i}\|^2)^{1/2}$*

with the following property: Given any constant $\beta \in (0, 1/2)$, there exists a constant $c_2(\beta) > 0$, such that if $\delta/\varepsilon < c_2(\beta)$, then any 2-factor k-means approximation algorithm with $L = 2$ clusters has misclassification rate satisfying $\frac{1}{2} - \beta \leq$ Miss $\leq \frac{1}{2}$. Moreover, any 2-factor k-means approximation algorithm with $L = 4$ clusters will recover a perfect refinement of the original clusters in the above setting.

The proof of Proposition 1 can be found in the Supplementary Material. This proposition shows that if the separation condition (6) is reversed, one can force the performance of any k-means algorithm to be arbitrarily close to that of random guessing. The true centers in Proposition 1 are the natural centers implied by the k-means cost function for the true labels, that is, $\xi_k^* = \frac{1}{n}\sum_i x_i 1\{z_i = k\}$ for $k = 1, 2$. One can take $c_1(\beta, \pi_{\min}) = 6.3 \max(1/\pi_{\min}, 2/\beta)^{1/2}$ and $c_2(\beta) = \sin(\tan^{-1}(\sqrt{\beta/45}))$ for the constants in (6) and Proposition 1.

Remark 1. The separation condition (6) is related to the *distribution stability* introduced in [5]. Roughly speaking distribution stability plus the following property implies our condition:

(D1) For every pair of distinct clusters C_k and C_ℓ with centers ξ_k^* and ξ_ℓ^*, there is a point $x \in C_\ell$ such that $\|x - \xi_k^*\| \leq \|\xi_\ell^* - \xi_k^*\|$.

That is, every cluster C_ℓ has points which are closer than ξ_ℓ^* to the centers of other clusters. This property is quite mild and one expects it to hold with high probability if the distribution of the points have positive density w.r.t. to the (full-dimensional) Lebesgue measure in a ball around the center. The above seems to suggest that distribution stability is slightly weaker than our condition (6). However, in the presence of (D1), we can also significantly relax distribution stability to arrive at our condition, the details of which are provided in the Supplementary Material. In this sense, these two notions are closely related but not directly comparable, i.e., neither is weaker than the other in general.

Example 2 (Sub-Gaussian mixtures)**.** Let us assume that the data is generated from a K-component sub-Gaussian mixture model $x_i = \xi_{z_i}^* + d^{-1/2} w_i$ where $w_i = (w_{i1}, \dots, w_{id}) \in \mathbb{R}^d$ is a zero mean sub-Gaussian noise vector with sub-Gaussian parameter σ_i, and $z_i \in [K]$ is the latent label of the ith observation. This is an extension of the sub-Gaussian mixture model considered in [7]. Determining whether $(\xi_k^*)_{k=1}^K$ is actually the solution of the population problem (1) is, itself, challenging and the answer may depend on the exact distribution of $\{w_i\}$. Nevertheless, our results allow us to treat (ξ_k^*) as the true centers. Below we sketch how Theorem 2 applies in this case. The details of the arguments, including the exact definition of a sub-Gaussian vector are provided in the Supplementary Material. Let $\sigma_{\max} = \max_i \sigma_i$ and set $\alpha_i^2 := \mathbb{E}\|d^{-1/2} w_i\|_2^2$ and $\bar{\alpha}_n^2 := \frac{1}{n}\sum_{i=1}^n \alpha_i^2$. Assume that there is a numerical constant $C > 0$ such that $\bar{\alpha}_n^2 \leq C\sigma_{\max}^2$. Then, one can show that

$$\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n \|x_i - \xi_{z_i}^*\|^2 > 2\bar{\alpha}_n^2\Big) \leq e^{-c_1 n \bar{\alpha}_n^4/\sigma_{\max}^4} =: p_n$$

for some numerical constant $c_1 > 0$. Taking $\varepsilon^2 = 2\bar{\alpha}_n^2$ and $p = 2$ in Theorem 2, we have that with probability at least $1 - p_n$,

$$\frac{\delta^2}{2\bar{\alpha}_n^2} > \frac{(1+\kappa)^2 c^2}{\pi_{\min}} \qquad \Longrightarrow \qquad \mathrm{Miss}(z, \widehat{z}) \leq 2K(1+\kappa)^2 c^2 \Big(\frac{\bar{\alpha}_n}{\delta}\Big)^2,$$

where δ is as in (3) and $c > 2$. In a general problem, $\bar{\alpha}_n$, $\sigma_{\max}$ and δ all can vary with n. In order to have label consistency for an ALG(2) algorithm, it is enough to have $\bar{\alpha}_n/\delta = o(1)$ and $n\bar{\alpha}_n^4/\sigma_{\max}^4 \to \infty$. The consistency here is based on the extended Definition 1 and includes the overfitted case in which a refinement of the true labels is consistently recovered. We note that the model in this example includes a very general Gaussian mixture model as a special case, namely the case $w_i \sim N(0, \Sigma_i)$ where the covariance matrices $\Sigma_i \in \mathbb{R}^{d\times d}$ are allowed to vary with each data point. In this case, one can take $\sigma_{\max} = \max_{1\leq i\leq n} \|\Sigma_i\|_{\mathrm{op}}$ where $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm, and $\bar{\alpha}_n^2 := \frac{1}{n}\sum_{i=1}^n \mathrm{tr}(\Sigma_i)/d$. □

2.2 Distance to fake centers

We now extend Theorem 2, to allow for "fake" centers $\{\widetilde{\xi}_\ell\}_{\ell=1}^L$ and the corresponding labels $\{\widetilde{z}_i\}$. These can be constructed to reduce the required distance to the data points (x_i).

Theorem 3 (Approximate Recovery, II)**.** *For a fixed $L \geq K$, consider a vector of constructed centers $\widetilde{\xi} \in \mathcal{X}^L$, constructed labels $\widetilde{z} = (\widetilde{z}_i)_{i=1}^n \in [L]^n$ and true labels $z = (z_i)_{i=1}^n \in [K]^n$. Assume that $\widetilde{z}$ is a refinement of z, i.e. there is $\widetilde{\omega} : [L] \to [K]$ such that $\widetilde{\omega}(\widetilde{z}_i) = z_i$ for all $i \in [n]$. Pick $\varepsilon, \delta > 0$ such that*

$$\Big(\frac{1}{n}\sum_{i=1}^n \|x_i - \widetilde{\xi}_{\widetilde{z}_i}\|^p\Big)^{1/p} \leq \varepsilon, \qquad \min_{\ell_1 \neq \ell_2,\, \widetilde{\omega}(\ell_1) \neq \widetilde{\omega}(\ell_2)} \|\widetilde{\xi}_{\ell_1} - \widetilde{\xi}_{\ell_2}\| \geq \delta \tag{7}$$

Consider an algorithm ALG(p) for problem (2)*, satisfying Assumption 1, and let $(\widehat{z}_i)_{i=1}^n \in [L]^n$ be the estimated label vector of ALG(p) applied with L clusters. Then, for any $c > 2$,*

$$\frac{\delta}{\varepsilon} > \frac{(1+\kappa)c}{\pi_{\min}^{1/p}} \quad \Longrightarrow \quad \mathrm{Miss}(z, \widehat{z}) < K(1+\kappa)^p c^p \Big(\frac{\varepsilon}{\delta}\Big)^p. \tag{8}$$

The advantage of Theorem 3 is that when the desired number of clusters L increases, the bound on the misclassification rate can go down: In some applications, by carefully constructing the fake centers $\widetilde{\xi}$, we can make ε smaller as L increases, while roughly maintaining the minimum separation among fake centers associated with the true clusters. If successful, this implies that a refinement of the true clustering can be achieved even when it is hard to recover the true clustering itself. In the following section, we show how this strategy can be applied to some manifold clustering problems.

3 The case for overfitting

We now illustrate applications of Theorem 3 in settings where it is hard to recover true clusters, based on the ideal K, but it is possible to obtain accurate refinements by overfitting. The idea is to consider clusters that look like submanifolds of $\mathbb{R}^d$.

3.1 Mixture of curves

We say that a random variable x has a (ρ, σ) *sub-Gaussian curve distribution* if $x = \gamma(t)$ where $t \in \mathbb{R}$ has a sub-Gaussian distribution with parameter σ and $\gamma : \mathbb{R} \to \mathbb{R}^d$ is a locally ρ-Lipschitz map. i.e., $\|\gamma(t) - \gamma(s)\| \leq \rho|t - s|$ for all $t, s \in \mathbb{R}$ such that $|t - s| \leq \frac{1}{\rho}$.

Proposition 2. *Assume that $(x_i)_{i=1}^n$ are independent draws from a K-component mixture of (ρ, σ) sub-Gaussian curve distributions. That is, $x_i = \gamma_{z_i}(t_i)$ where $z_i \in [K]$, $t_i \sim \mathbb{Q}_{z_i}$ independently across i, each $\mathbb{Q}_k$ is a sub-Gaussian distribution on $\mathbb{R}$ with parameter σ, and each γ_k is locally ρ-Lipschitz. Let $\mathcal{C}_k$ be the support of the distribution of $\gamma_k(t)$ where $t \sim \mathbb{Q}_k$. Assume that*

$$\min_{x \in \mathcal{C}_k,\, y \in \mathcal{C}_{k'}} \|x - y\| \geq \delta > 0, \quad \textit{for all } k \neq k'.$$

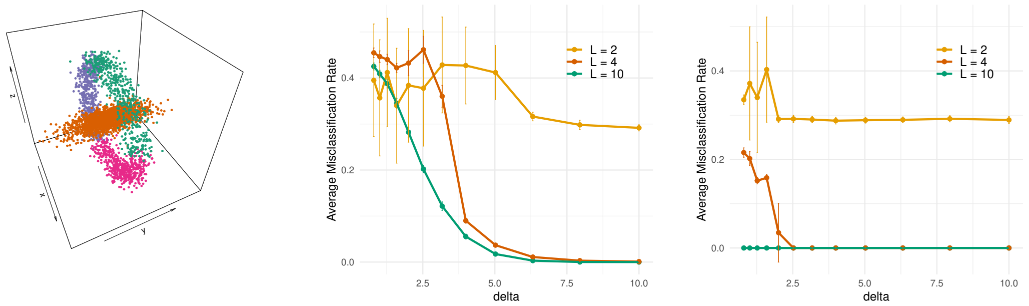
Then, there exist a constant $C = C(K, \delta, \rho, \sigma, \kappa)$ such that any ALG(2) satisfying Assumption 1 applied with $L_n \leq C\sqrt{n \log n}$ clusters recovers a perfect refinement of z with probability $\geq 1 - n^{-1}$.

The significance of this result is that one recovers a perfect refinement with the number of partitions $L_n = o(n)$. It is trivial to obtain a perfect refinement with $L_n = n$, but not so with $L_n/n \to 0$. This is especially the case since one can achieve quite complex cluster configurations with the model in Proposition 2, for some of which applying k-means with K clusters will have a misclassification rate bounded away from zero. Section 3.3 provides some such examples where the true cluster centers coincide, causing any k-means algorithm applied with the true K to incur a substantial error. See also Supplementary Material for a discussion of whether $L_n = O(\sqrt{n \log n})$ can be improved.

Various extensions of Proposition 2 are possible. We have the following extension to the noisy setting.

Corollary 1. *Assume that the data is given by $y_i = x_i + \frac{1}{\sqrt{d}} w_i$ for $i \in [n]$ where (x_i) are as given in Proposition 2 and w_i are sub-Gaussian noise vectors as in Example 2. Then, under the same assumptions as in Proposition 2, ALG(2) applied with $L_n \leq C\sqrt{n \log n}$ achieves a misclassification rate $\lesssim K(\bar{\alpha}_n/\delta)^2 + \frac{1}{n}$ with probability $\geq 1 - p_n - n^{-1}$ where $\bar{\alpha}_n$ and p_n are defined in Example 2.*

Corollary 1 shows that one can achieve consistent clustering (in the generalized sense) with $L_n = o(n)$ clusters assuming that the noise-to-signal ratio $\bar{\alpha}_n/\delta \to 0$ and $n\bar{\alpha}_n^4/\sigma_{\max}^4 \to \infty$; the same conditions needed in the sub-Gaussian mixture example. Again, this result is significant since even in the noiseless case ($\bar{\alpha}_n = 0$), consistent recovery is not possible with $L = K$ for some mixtures of curve models.



(a) Noisy $(L = 4, \delta = 3)$ (b) Noisy (c) Noiseless

Figure 1: Line-circle model: (a) Scatter plot for the noisy version. The colors show the $L = 4$ estimated clusters by k-means. (b) and (c) show the (generalized) misclassification rate versus δ, the radius of the circle, in the noisy and noiseless versions of the model.

3.2 Mixture of higher-order submanifolds

A version of Proposition 2 can be stated for a higher-dimensional version of the mixture-of-curves model, if we consider generalized k-means problems with $p > 2$. We say that a random variable x has a (ρ, σ, r) *sub-Gaussian manifold distribution* if $x = \gamma(t)$ where $t \in \mathbb{R}^r$ has a sub-Gaussian distribution with parameter σ and $\gamma : \mathbb{R}^r \to \mathbb{R}^d$ is a locally ρ-Lipschitz map. i.e., $\|\gamma(t) - \gamma(s)\| \leq \rho\|t - s\|$ for all $t, s \in \mathbb{R}^r$ such that $\|t - s\| \leq \frac{1}{\rho}$.

Proposition 3. *Assume that $(x_i)_{i=1}^n$ are independent draws from a K-component mixture of sub-Gaussian manifold distributions, with parameters (ρ, σ, r_k) for $k \in [K]$, and let $r = \max_{r \in [K]} r_k$. Let $\mathcal{C}_k$ be the support of the distribution of the kth component. Assume that*

$$\min_{x \in \mathcal{C}_k,\, y \in \mathcal{C}_{k'}} \|x - y\| \geq \delta > 0, \quad \textit{for all } k \neq k'.$$

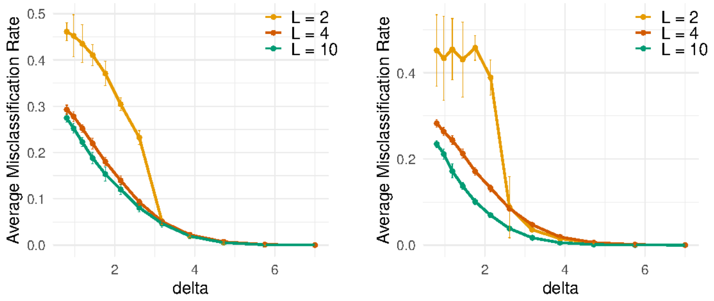
Then, there exist a constant $C = C(K, \delta, \rho, \sigma, r, \kappa)$ such that any ALG(p) satisfying Assumption 1, applied with $L_n \leq C(n^{1/p}\sqrt{\log n})^r$ clusters recovers a perfect refinement of z with probability $\geq 1 - n^{-1}$. In particular, for $p > r$, we have perfect refinement recovery with $L_n = o(n)$ clusters, with high probability.

It is also possible to extend the results to more general distributions on submanifolds via a notion of stochastic covering numbers. For random vector x with distribution $\mu_{\mathcal{C}}$ on a submanifold $\mathcal{C} \subset \mathbb{R}^d$, let $N_{\mu_{\mathcal{C}}}(\varepsilon)$ be the smallest integer for which, there is a high probability ε-cover of x, that is, a finite subset $\mathcal{N} \subset \mathcal{C}$ such that $\mathbb{P}(\min_{y \in \mathcal{N}} \|x - y\| \leq \varepsilon) \geq 1 - n^{-2}$. We state a generalization of Proposition 3 to this setting in the Supplementary Material.

3.3 Numerical experiments

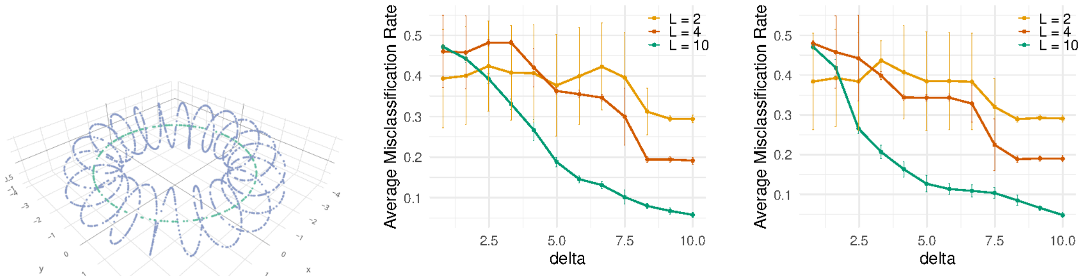
We first consider the (noiseless) line-circle model in $\mathbb{R}^3$, an example of mixture-of-curves. This model has two clusters: (1) The uniform distribution on the circumference of a circle in the xz-plane, centred at the origin, and (2) the standard Gaussian distribution on the y axis. The minimum separation δ between the two clusters is the radius of the circle. We also consider the noisy version of this model where we add $N(0, \sigma^2 I_3)$. We sample data points with equal probability from the two clusters. It is nearly impossible for the k-means to correctly label these two clusters when $L = 2$, since the centers of the two clusters coincide. Figure 1 shows the scatter plot of the data simulated from the noisy line-circle model, with noise level $\sigma = 0.1$, $n = 3000$ and $\delta = 3$. Here, the noise level is set low for better illustration. Different colors are used to label data points based on the output of k-means clustering with $L = 4$, and this demonstrates that each estimated cluster is a subset of a true cluster.

The result aligns with Theorem 3. Although, the true centers coincide (with the origin) when $L = 2$, by increasing L, we can create fake centers on the line and the circle to have separation close to δ and thus get a small missclassification rate. The other two panels in Figure 1 show the average missclassification rate over 32 repetitions versus δ, for both the noiseless and noisy ($\sigma = 1$) line-circle



(a) Noisy (b) Noiseless

Figure 2: Line-Gaussian model: The (generalized) misclassification rate versus δ, the distance of the Gaussian center to the line, in the (a) noisy and (b) noiseless versions of the model.



(a) True clusters (noiseless) (b) Noisy (c) Noiseless

Figure 3: Circle-torus model: (a) Scatter plot for the noiseless version. Colors are used to separate two true clusters. (b) and (c) show the (generalized) misclassification rate versus δ, the radius of the circle, in the noisy and noiseless versions of the model.

model. Both show that the misclassification rate is negatively associated with δ and L when $L > 2$. Similar results are shown for the circle-torus model in Figure 3. Details of this model are discussed in the Supplementary Material.

Figure 2 shows the results for a line-Gaussian mixture model: $x_i = \xi^*_{z_i} + \Sigma_{z_i}^{1/2} w_i \in \mathbb{R}^2$ where $\xi_1^* = (0, \delta)$ and $\xi_2^* = (0, 0)$, $w_i \sim N(0, I_2)$, $\Sigma_1 = I_2$ and $\Sigma_2 = \operatorname{diag}(\sigma^2, 0)$. Here, we have set $\sigma = 5$ and sampled $n = 3000$ data points with equal probability from the two clusters. We also consider its noisy version by setting all the zero elements in Σ_2 to 0.7, which makes the model a general Gaussian mixture. Figure 2 shows the average missclassification rate over 32 repetitions for different L. The results are consistent with Theorem 3 showing that as δ increases, the misclassification rate decreases.

4 Proofs

Let us first recall a fact from functional analysis. Consider the space of functions $f : [n] \to \mathcal{X}$ and let us equip $[n]$ with the uniform probability measure ν_n. Then, from the theory of Lebesgue-Bochner spaces, $\|f\|_p := (\int \|f(\omega)\|_{\mathcal{X}}^p \, d\nu_n(\omega))^{1/p}$ defines a proper norm on this function space, turning it into a Banach space $L^p(\nu_n; \mathcal{X})$. In particular, the triangle inequality holds for this norm. Note that $\|f\|_p = (\frac{1}{n}\sum_{i=1}^n \|f(i)\|_{\mathcal{X}}^p)^{1/p}$. We will frequently invoke the triangle inequality in $L^p(\nu_n, \mathcal{X})$.

Let $\mu^* := \sum_k \pi_k \delta_{\xi_k^*} = \frac{1}{n}\sum_{i=1}^n \delta_{\xi^*_{z_i}}$ be the empirical measure associated with $\{\xi^*_{z_i}\}$. Recalling definition (1) of the population cost function, we have, for any $\xi \in \mathcal{X}^L$,

$$Q(\xi; \mu^*)^p = \sum_{k=1}^K \pi_k \min_{1 \le \ell \le L} \|\xi_k^* - \xi_\ell\|^p = \frac{1}{n}\sum_{i=1}^n \min_{1 \le \ell \le L} \|\xi^*_{z_i} - \xi_\ell\|^p. \tag{9}$$

We start with three lemmas that are proved in the Supplementary Material:

Lemma 1. *Let ALG(p) be a k-means algorithm satisfying Assumption 1(b') and let $\widehat{\xi}$ be its output for L clusters. Furthermore, assume $(\frac{1}{n}\sum_{i=1}^n \|x_i - \xi^*_{z_i}\|^p)^{1/p} \le \varepsilon$. Then $Q(\widehat{\xi};\mu^*) \le (1+\kappa)\varepsilon$.*

Lemma 2 (Curvature)**.** *For every $\xi \subset \mathcal{X}^L$ and $\xi^* \in \mathcal{X}^K$, with $L \ge K$,*

$$Q(\xi;\mu^*) \ge \pi_{\min}^{1/p}\Big(d_p(\xi,\xi^*) \wedge \frac{\delta}{2}\Big).$$

Lemma 3. *Assume that $\max_{1\le i\le n}\|x_i - \xi^*_{z_i}\| \le \eta$ and $d_\infty(\widehat{\xi},\xi^*) \le \gamma$. When $L = K$, if $\delta > 2\gamma + 2\eta$, there exists a bijective function $\omega : [K] \to [K]$ satisfying $\omega(\widehat{z}_i) = z_i,\ \forall i \in [n]$. When $L > K$, if $\delta > 2\gamma + 4\eta$, there exists a surjective function $\omega : [L] \to [K]$ satisfying $\omega(\widehat{z}_i) = z_i,\ \forall i \in [n]$.*

Proof of Theorem 1. As $(\frac{1}{n}\sum_{i=1}^n \|x_i - \xi^*_{z_i}\|^p)^{1/p} \le \max_{1\le i\le n}\|x_i - \xi^*_{z_i}\| \le \eta$, combining Lemma 1 and 2, we have

$$\Big(d_p(\widehat{\xi},\xi^*) \wedge \frac{\delta}{2}\Big) \le \frac{Q(\widehat{\xi},\mu^*)}{\pi_{\min}^{1/p}} \le \frac{(1+\kappa)\eta}{\pi_{\min}^{1/p}}.$$

By the condition on δ in (4), we have $\delta/2 > (1+\kappa)\eta/\pi_{\min}^{1/p}$. Then, $d_\infty(\widehat{\xi},\xi^*) \le d_p(\widehat{\xi},\xi^*) \le \gamma := (1+\kappa)\eta/\pi_{\min}^{1/p}$, which also makes the assumption in Lemma 3 that $\delta > 2\gamma + 4\eta$ valid. Finally, the result follows from Lemma 3. □

Proof of Theorem 2. The argument is similar to one that has appeared in recent literature [17, 12, 33]. From the proof of Lemma 1 (in the Supplementary Material), we have

$$Q(\widehat{\xi};\mu^*) \le \Big(\frac{1}{n}\sum_{i=1}^n \|\xi^*_{z_i} - \widehat{\xi}_{\widehat{z}_i}\|^p\Big)^{1/p} \le (1+\kappa)\varepsilon.$$

By Lemma 2

$$\Big(d_p(\widehat{\xi},\xi^*) \wedge \frac{\delta}{2}\Big) \le \frac{Q(\widehat{\xi},\mu^*)}{\pi_{\min}^{1/p}} \le \frac{(1+\kappa)\varepsilon}{\pi_{\min}^{1/p}}.$$

By the separation assumption in (5), $\delta/2 > (1+\kappa)\varepsilon/\pi_{\min}^{1/p}$. Hence $d_p(\widehat{\xi},\xi^*) \le (1+\kappa)\varepsilon/\pi_{\min}^{1/p}$. Let $\mathcal{C}_k = \{i : z_i = k\}$, $|\mathcal{C}_k| = n_k$, and set $T_k := \{i \in \mathcal{C}_k : \|\xi^*_{z_i} - \widehat{\xi}_{\widehat{z}_i}\| \le \delta/c\}$. Letting $S_k = \mathcal{C}_k \setminus T_k$, we obtain

$$|S_k|\delta^p/c^p < \sum_{i\in S_k}\|\xi^*_{z_i} - \widehat{\xi}_{\widehat{z}_i}\|^p \le \sum_{i=1}^n \|\xi^*_{z_i} - \widehat{\xi}_{\widehat{z}_i}\|^p \le n(1+\kappa)^p\varepsilon^p.$$

Therefore,

$$\frac{|S_k|}{n_k} < \frac{n(1+\kappa)^p c^p \varepsilon^p}{n_k \delta^p} \le 1.$$

The last inequality is by assumption $\delta > (1+\kappa)c\varepsilon/\pi_{\min}^{1/p}$. Hence, T_k is not empty. Furthermore, we argue that if $i \in T_k$ and $j \in T_\ell$ for $k \ne \ell$, i.e. $z_i \ne z_j$, then $\widehat{z}_i \ne \widehat{z}_j$. Assume otherwise, that is, $\widehat{z}_i = \widehat{z}_j$. Then

$$\|\xi^*_k - \xi^*_\ell\| \ \le\ \|\xi^*_k - \widehat{\xi}_{\widehat{z}_i}\| + \|\xi^*_\ell - \widehat{\xi}_{\widehat{z}_j}\| \ \le\ 2\delta/c \ <\ \delta$$

causing a contradiction.

Let $\mathcal{L}_k := \{\widehat{z}_i :\ i \in T_k\}$ and $\mathcal{L} = \bigcup_{k=1}^K \mathcal{L}_k$. Define a function $\omega : \mathcal{L} \to [K]$ by setting $\omega(\ell) = k$ for all $\ell \in \mathcal{L}_k$ and $k \in [K]$. By the property of $\{T_k\}$ shown above, $\mathcal{L}_k, k \in [K]$ are disjoint and nonempty sets. This implies that ω is well-defined and surjective. Extend ω to a surjective $\omega : [L] \to [K]$ by arbitrarily defining it for $[L] \setminus \mathcal{L}$. Note that $\widehat{z}_i \in \mathcal{L}_k$ implies $z_i = k$. It follows that $\omega(\widehat{z}_i) = z_i$ for all $\widehat{z}_i \in \mathcal{L}$, and

$$\frac{1}{n}\sum_{i=1}^n 1\{z_i \ne \omega(\widehat{z}_i)\} \le \frac{n - |\mathcal{L}|}{n} = \sum_{k=1}^K \frac{|S_k|}{n} < \frac{K(1+\kappa)^p c^p \varepsilon^p}{\delta^p}.$$

The result follows. □

Proof of Theorem 3. By assumption, κ-approximation holds for both K and L clusters. Then,

$$\widehat{Q}(\widehat{\xi}) \le \kappa\, \widehat{Q}^{(L)}_{\min}, \quad \text{where} \quad \widehat{Q}^{(L)}_{\min} := \min_{\xi \in \mathcal{X}^L} \widehat{Q}(\xi).$$

Since $\widehat{Q}^{(L)}_{\min} \le (\frac{1}{n}\sum_{i=1}^n \|x_i - \widetilde{\xi}_{\widetilde{z}_i}\|^p)^{1/p} \le \varepsilon$, by the triangle inequality in $L^p(\nu_n, \mathcal{X})$,

$$\Big(\frac{1}{n}\sum_{i=1}^n \|\widetilde{\xi}_{\widetilde{z}_i} - \widehat{\xi}_{\widehat{z}_i}\|^p\Big)^{1/p} \le \Big(\frac{1}{n}\sum_{i=1}^n \|x_i - \widetilde{\xi}_{\widetilde{z}_i}\|^p\Big)^{1/p} + \Big(\frac{1}{n}\sum_{i=1}^n \|x_i - \widehat{\xi}_{\widehat{z}_i}\|^p\Big)^{1/p} \le (1+\kappa)\varepsilon.$$

Let $T_k := \{i \in \mathcal{C}_k : \|\widetilde{\xi}_{\widetilde{z}_i} - \widehat{\xi}_{\widehat{z}_i}\| \le \delta/c\}$ and $S_k = \mathcal{C}_k \setminus T_k$. Then,

$$|S_k|\delta^p/c^p < \sum_{i\in S_k} \|\widetilde{\xi}_{\widetilde{z}_i} - \widehat{\xi}_{\widehat{z}_i}\|^p \le \sum_{i=1}^n \|\widetilde{\xi}_{\widetilde{z}_i} - \widehat{\xi}_{\widehat{z}_i}\|^p \le n(1+\kappa)^p\varepsilon^p.$$

Therefore,

$$\frac{|S_k|}{n_k} < \frac{n(1+\kappa)^p c^p \varepsilon^p}{n_k \delta^p} \le 1$$

The last inequality is by assumption $\delta \ge (1+\kappa)c\varepsilon/\pi_{\min}^{1/p}$. Hence T_k is not empty. Next we argue that if $i \in T_k$, $j \in T_\ell$ for $k \ne \ell$, i.e. $z_i \ne z_j$, then $\widehat{z}_i \ne \widehat{z}_j$. Assume otherwise, that is $\widehat{z}_i = \widehat{z}_j$. Since $\widetilde{z}$ is a refinement of z, $z_i \ne z_j$ implies $\widetilde{z}_i \ne \widetilde{z}_j$ and $\widetilde{\omega}(\widetilde{z}_i) \ne \widetilde{\omega}(\widetilde{z}_j)$. By the triangle inequality,

$$\|\widetilde{\xi}_{\widetilde{z}_i} - \widetilde{\xi}_{\widetilde{z}_j}\| \le \|\widetilde{\xi}_{\widetilde{z}_i} - \widehat{\xi}_{\widehat{z}_i}\| + \|\widetilde{\xi}_{\widetilde{z}_j} - \widehat{\xi}_{\widehat{z}_j}\| \le 2\delta/c < \delta$$

causing a contradiction. The rest of the proof follows that of Theorem 2. □

Proof of Proposition 2. Let m_k be the mean of $\mathbb{Q}_k$. Then, $\mathbb{P}(|t_i - m_{z_i}| > t) \le 2e^{-t^2/2\sigma^2}$. Let $M = \sqrt{6\sigma^2 \log n}$. By union bound, with probability $\ge 1 - 2n^{-2}$ we have $|t_i - m_{z_i}| \le M$ for all $i \in [n]$. We can cover the set $[-M, M] \subset \mathbb{R}$, with $L' = M/\varepsilon$ 1-D balls of radius ε. (Without loss of generality, we assume that L' is an integer for simplicity.) Let $\mathcal{T} = \{\tau_1, \dots, \tau_{L'}\}$ one such cover and note that $m_k + \mathcal{T}$ is an ε-cover of $m_k + [-M, M]$. Let $\pi_k : \mathbb{R} \to (m_k + \mathcal{T})$ be the projection from $\mathbb{R}$ onto $m_k + \mathcal{T}$. Then, $\|\gamma_{z_i}(t_i) - \gamma_{z_i}(\pi_{z_i}(t_i))\| \le \rho|t_i - \pi_{z_i}(t_i)| \le \rho\varepsilon$, assuming that $\varepsilon \le 1/\rho$.

Let $z_i' := \operatorname{argmin}_{\ell' \in [L']} |t_i - (m_{z_i} + \tau_{\ell'})|$ so that $\pi_{z_i}(t_i) = m_{z_i} + \tau_{z_i'}$. Then let $L_n = KL'$ and fix a bijection $\phi : [L_n] \to [K] \times [L']$ and define the labels $\widetilde{z}_i = \phi^{-1}(z_i, z_i')$. Also consider the map $\omega_0 : [K] \times [L'] \to [K]$ given by $\omega_0(k, \ell') = k$ and set $\widetilde{\omega} := \omega_0 \circ \phi$ which is a surjective map from $[L_n]$ to $[K]$ satisfying $\widetilde{\omega}(\widetilde{z}_i) = z_i$. For $\ell \in [L_n]$ with $\phi(\ell) = (k, \ell')$, define $\widetilde{\xi}_\ell = \gamma_k(m_k + \tau_{\ell'})$. Then, we have $\widetilde{\xi}_{\widetilde{z}_i} = \gamma_{z_i}(m_{z_i} + \tau_{z_i'}) = \gamma_{z_i}(\pi_{z_i}(t_i))$, hence the above argument gives $\|\gamma(t_i) - \widetilde{\xi}_{\widetilde{z}_i}\| \le \rho\varepsilon$. It is also clear that the the separation condition (7) is satisfied since by construction if $\widetilde{\omega}(\ell_1) \ne \widetilde{\omega}(\ell_2)$ with $\phi(\ell_1) = (k_1, \ell_1')$ and $\phi(\ell_2) = (k_2, \ell_2')$, then $k_1 \ne k_2$ hence $\widetilde{\xi}_{\ell_1}$ and $\widetilde{\xi}_{\ell_2}$ lie on different manifolds (on $\mathcal{C}_{k_1}$ and $\mathcal{C}_{k_2}$). It follows that conclusion (8) of Theorem 3 holds for $p = 2$ and, say, $c = 3$ but with ε replaced with $\rho\varepsilon$. Take $\varepsilon = (c_1\sqrt{n})^{-1}$ for constant c_1 to be determined. Let $c_2 = 3\rho(1+\kappa)/\delta$. As long as $n\pi_{\min} > (c_2/c_1)^2$, the separation condition in (8) is satisfied and we have $\operatorname{Miss}(z, \widehat{z}) \le K(c_2/c_1)^2/n$. Hence, as long as $c_1 > \sqrt{K}c_2$, we will have $\operatorname{Miss}(z, \widehat{z}) < 1/n$ which implies $\operatorname{Miss}(z, \widehat{z}) = 0$. We also need to satisfy $\varepsilon < 1/\rho$ that is $c_1 \ge \rho/\sqrt{n}$. Taking $c_1 = \sqrt{K}c_2 + \rho$ satisfies all the required constraints on c_1. The required number of clusters is

$$L_n = KL' = KM/\varepsilon \le 3K\sigma c_1\sqrt{n\log n},$$

which proves the result with $C = 3K\sigma c_1$. Note that since $c_2/c_1 < 1$ and $n\pi_{\min} \ge 1$, the condition $n\pi_{\min} > (c_2/c_1)^2$ is automatically satisfied. The proof is complete. □

Acknowledgments and Disclosure of Funding

This work was supported by the NSF CAREER grant DMS-1945667. We thank the referees for their helpful comments that lead to the improvement of the paper.

References

[1] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

[2] Arash A Amini and Zahra S Razaee. Concentration of kernel matrices with application to kernel spectral clustering. *The Annals of Statistics*, 49(1):531–556, 2021.

[3] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.

[4] Pranjal Awasthi, Afonso S Bandeira, Moses Charikar, Ravishankar Krishnaswamy, Soledad Villar, and Rachel Ward. Relax, no need to round: Integrality of clustering formulations. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pages 191–200, 2015.

[5] Pranjal Awasthi, Avrim Blum, and Or Sheffet. Stability yields a ptas for k-median and k-means clustering. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 309–318. IEEE, 2010.

[6] Vincent Cohen-Addad and Chris Schwiegelshohn. On the local structure of stable clustering instances. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 49–60. IEEE, 2017.

[7] Noureddine El Karoui. On information plus noise kernel random matrices. *The Annals of Statistics*, 38(5):3191–3216, 2010.

[8] Yingjie Fei and Yudong Chen. Hidden integrality of sdp relaxations for sub-gaussian mixture models. In *Conference On Learning Theory*, pages 1931–1965. PMLR, 2018.

[9] Christophe Giraud and Nicolas Verzelen. Partial recovery bounds for clustering with the relaxed k-means. *Mathematical Statistics and Learning*, 1(3):317–374, 2019.

[10] Takayuki Iguchi, Dustin G Mixon, Jesse Peterson, and Soledad Villar. Probably certifiably correct k-means clustering. *Mathematical Programming*, 165(2):605–642, 2017.

[11] Tao Jiang, Stephen Vavasis, and Chen Wen Zhai. Recovery of a mixture of gaussians by sum-of-norms clustering. *Journal of Machine Learning Research*, 21(225):1–16, 2020.

[12] Jiashun Jin. Fast community detection by score. *Annals of Statistics*, 43(1):57–89, 2015.

[13] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. A local search approximation algorithm for k-means clustering. *Computational Geometry*, 28(2-3):89–112, 2004.

[14] Mieczysław A Kłopotek. On the consistency of k-means++ algorithm. *Fundamenta Informaticae*, 172(4):361–377, 2020.

[15] Amit Kumar, Yogish Sabharwal, and Sandeep Sen. A simple linear time (1+/spl epsiv/)-approximation algorithm for k-means clustering in any dimensions. In *45th Annual IEEE Symposium on Foundations of Computer Science*, pages 454–462. IEEE, 2004.

[16] Silvio Lattanzi and Christian Sohler. A better k-means++ algorithm via local search. In *International Conference on Machine Learning*, pages 3662–3671. PMLR, 2019.

[17] Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *Annals of Statistics*, 43(1):215–237, 2015.

[18] Xiaodong Li, Yang Li, Shuyang Ling, Thomas Strohmer, and Ke Wei. When do birds of a feather flock together? k-means, proximity, and conic programming. *Mathematical Programming*, 179(1):295–341, 2020.

[19] Tamás Linder. Learning-theoretic methods in vector quantization. In *Principles of nonparametric learning*, pages 163–210. Springer, 2002.

[20] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[21] Yu Lu and Harrison H Zhou. Statistical and computational guarantees of lloyd's algorithm and its variants. *arXiv preprint arXiv:1612.02099*, 2016.

[22] James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[23] Konstantin Makarychev, Aravind Reddy, and Liren Shan. Improved guarantees for k-means++ and k-means++ parallel. *Advances in Neural Information Processing Systems*, 33, 2020.

[24] Jirı Matoušek. On approximate geometric k-clustering. *Discrete & Computational Geometry*, 24(1):61–84, 2000.

[25] Dustin G Mixon, Soledad Villar, and Rachel Ward. Clustering subgaussian mixtures with k-means. In *2016 IEEE Information Theory Workshop (ITW)*, pages 211–215. IEEE, 2016.

[26] Abhinav Nellore and Rachel Ward. Recovery guarantees for exemplar-based clustering. *Information and Computation*, 245:165–180, 2015.

[27] Ashkan Panahi, Devdatt Dubhashi, Fredrik D Johansson, and Chiranjib Bhattacharyya. Clustering by sum of norms: Stochastic incremental algorithm, convergence and cluster recovery. In *International conference on machine learning*, pages 2769–2777. PMLR, 2017.

[28] Jiming Peng and Yu Wei. Approximating k-means-type clustering via semidefinite programming. *SIAM journal on optimization*, 18(1):186–205, 2007.

[29] David Pollard. Strong consistency of k-means clustering. *The Annals of Statistics*, pages 135–140, 1981.

[30] David Pollard. Quantization and the method of k-means. *IEEE Transactions on Information theory*, 28(2):199–205, 1982.

[31] Defeng Sun, Kim-Chuan Toh, and Yancheng Yuan. Convex clustering: model, theoretical guarantee and efficient algorithm. *Journal of Machine Learning Research*, 22(9):1–32, 2021.

[32] Dennis Wei. A constant-factor bi-criteria approximation guarantee for k-means++. *Advances in Neural Information Processing Systems*, 29:604–612, 2016.

[33] Zhixin Zhou and Arash A Amini. Analysis of spectral clustering algorithms for community detection: the general bipartite setting. *The Journal of Machine Learning Research*, 20(1):1774–1820, 2019.

[34] Changbo Zhu, Huan Xu, Chenlei Leng, and Shuicheng Yan. Convex optimization procedure for clustering: Theoretical revisit. *Advances in Neural Information Processing Systems*, 27:1619–1627, 2014.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 1.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Our work is theoretical and does not have direct societal impacts. Considering the page limit, we focused on presenting theoretical results.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [Yes]

(b) Did you include complete proofs of all theoretical results? [Yes]

3. If you ran experiments...

   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See the supplemental material.

   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A] We did not do data training.

   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Error bars are included in the plots.

   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Computing details are in the supplemental material.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [N/A]

   (b) Did you mention the license of the assets? [N/A]

   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See the supplemental material.

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix for “Label consistency in overfitted generalized k-means”

Linfan Zhang
Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90095
linfanz@g.ucla.edu

Arash A. Amini
Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90095
aaamini@ucla.edu

This appendix contains further discussion of the results, the remaining proofs, details of some examples and additional numerical experiments.

A Discussion

Proposition 2 and 3 show that perfect refinement for sub-gaussian mixture-of-curves model can be achieved when the number of clusters grows as $L_n = O(\sqrt{n \log n})$. To the best of our knowledge, this is the first such result in the literature, that is, an upper bound on the minimum number of clusters needed to achieve a perfect refinement of the true clusters. What remains for future investigations to determine is how tight this bound is. Empirically, we have found examples of the mixture-of-curves model for which $L_n \asymp 1$ seems to enough, but also an example where $L_n \asymp \sqrt{n \log n}$ seems to be the required scaling. Figure S1(a) shows a noisy circle-torus model (cf. Section D.1) with $R = 10, r = 2$ and $\sigma = 1$ that demonstrates the scaling $L_n \asymp \sqrt{n \log n}$. Here, we plot the average misclassification rate over 32 repetitions vs $L_n/\sqrt{n \log n}$ for various n. The fact that these plots coincide with each other for different n suggests that there is sharp threshold $\tau_n = C_1 \sqrt{n \log n}$ such that with $L_n > \tau_n$, perfect refinement recovery is possible and with $L_n < \tau_n$, impossible. Figure S1(b) shows an example that exhibits $L_n \asymp 1$ threshold: A line-circle model with parameters $\delta = 4$, $\sigma = 1$ and line standard deviation = 7.

The fact that, empirically, there are examples for which L_n has to grow as fast as $\sqrt{n \log n}$ for a perfect refinement recovery, suggests that the result of Proposition 2 may be sharp up to constants, over the class of mixture-of-curves distributions considered.

B Connection to distribution stability

The distribution stability for the K-means assumes the following [1]:

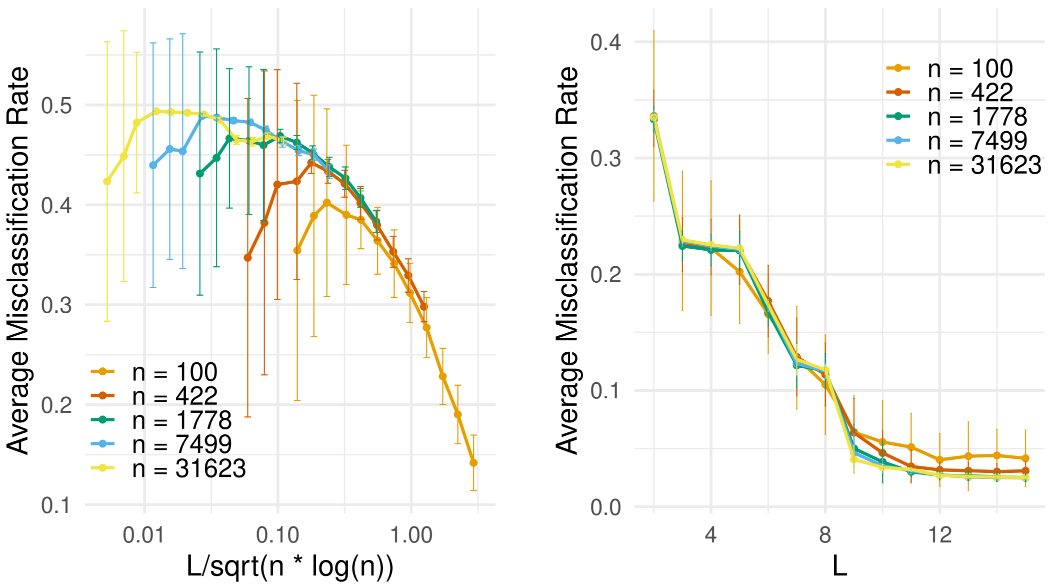
$$\|x - \xi_k^*\|^2 \geq \beta \cdot \frac{\text{OPT}_K}{n_k}, \quad \text{for all } x \notin C_k,$$

where $\text{OPT}_K = \sum_{i=1}^n \|x_i - \xi_{z_i}^*\|^2$ for the K-means optimal cluster labels $\{z_i\} \subset [K]^n$ and optimal centers $\{\xi_k^*\}$. Here, $C_k = \{i : z_i = k\}$ and $n_k = |C_k|$.

In our setting, we do not necessarily need to work with the optimal K-means clustering. So let us generalize the notion as follows: The data $\{x_i\}$ is β-distributed with respect to cluster labels $\{z_i\}$ and centers $\{\xi_k^*\}$ if

$$\|x - \xi_k^*\|^2 \geq \beta \cdot \sum_{i=1}^n \|x_i - \xi_{z_i}^*\|^2 / n_k, \quad \text{for all } x \notin C_k,$$

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.



(a) circle-torus model

(b) line-circle model

Figure S1: Examples of mixture-of-curve models that exhibit (a) $L_n \asymp \sqrt{n \log n}$ and (b) $L_n = O(1)$ refinement recovery threshold.

where $C_k = \{i : z_i = k\}$ and $n_k = |C_k|$. Setting $\frac{1}{n}\sum_{i=1}^n \|x_i - \xi^*_{z_i}\|^2 = \varepsilon^2$ and recalling $\pi_k = n_k/n$, the condition is equivalent to

$$\|x - \xi^*_k\| \geq \frac{\sqrt{\beta}\cdot\varepsilon}{\sqrt{\pi_k}}, \quad \text{for all } x \notin C_k. \tag{S1}$$

Let us strengthen the condition slightly and consider the following notion instead

$$\|x - \xi^*_k\| \geq \frac{\sqrt{\beta}\cdot\varepsilon}{\sqrt{\pi_{\min}}}, \quad \text{for all } x \notin C_k, \tag{S2}$$

where $\pi_{\min} = \min_k \pi_k$. (This is without loss of generality: We could have stated our results with separate center separation parameters for each cluster, i.e., $\delta_k = \min_{\ell \neq k} \|\xi^*_k - \xi^*_\ell\|$, in which case we could directly compare with the original version (S1). We opted for the simpler global center separation in the paper for simplicity.)

Now assume that the data is β-distributed and in addition:

(D1) For all distinct pairs (k, ℓ), there is $x \in C_\ell$ such that $\|x - \xi^*_k\| \leq \|\xi^*_\ell - \xi^*_k\|$.

That is, every cluster C_ℓ has points which are closer than ξ^*_ℓ to the centers of other clusters. Then, it follows that

$$\frac{\delta}{\varepsilon} \geq \frac{\sqrt{\beta}}{\sqrt{\pi_{\min}}} \tag{S3}$$

which is our separation condition. (Recall that $\delta = \min_{k \neq \ell} \|\xi^*_k - \xi^*_\ell\|$).

In fact, in the presence of (D1), we can relax β-distribution stability as follows: Assume (D1) and for the x in (D1) assume that the inequality in (S2) holds. Then, our separation condition (S3) follows. Note that (D1) is quite mild and one expects it to hold almost always if there is some full-dimensional randomness in the distribution of the points in a cluster.

Alternatively, our separation condition can be written equivalently as

$$\|x - \xi^*_k\| \geq \frac{\sqrt{\beta}\cdot\varepsilon}{\sqrt{\pi_{\min}}}, \quad \text{for all } x \in \{\xi^*_\ell\}_{\ell \neq k} \tag{S4}$$

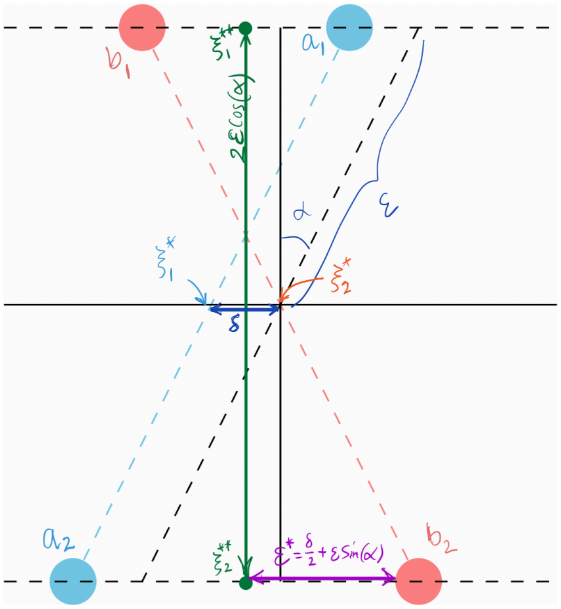


Figure S2: The geometry of the dataset in Proposition 1

Comparing (S4) and (S2), the conditions are somewhat close, but different. Neither condition directly follow from the other one in general.

Note also that although in the discussion above, we refer to ξ_k^* as the center of C_k, in our general setting ξ_k^* need not be the optimal center $\frac{1}{n_k}\sum_{i\in C_k} x_i$.

C Proof of Proposition 1

For $\alpha \in [0, \pi/2)$, consider a constellation of points in $\mathbb{R}^2$ at locations $a_1 = (\varepsilon \sin\alpha - \delta, \varepsilon\cos\alpha)$, $a_2 = (-\varepsilon\sin\alpha - \delta, -\varepsilon\cos\alpha)$, $b_1 = (-\varepsilon\sin\alpha, \varepsilon\cos\alpha)$ and $b_2 = (\varepsilon\sin\alpha, -\varepsilon\cos\alpha)$. Assume that $n/4$ of the data points are on each of the points a_1, a_2, b_1 and b_2. Assume that data points in $\{a_1, a_2\}$ form cluster 1 and points in $\{b_1, b_2\}$ form cluster 2. That is, this is the true cluster labels as specified by an external source. The true cluster centers are then at locations $\xi_1^* = (-\delta, 0)$ and $\xi_2^* = (0, 0)$. We also have $(\frac{1}{n}\sum_i \|x_i - \xi_{z_i}^*\|^2)^{1/2} = \varepsilon$ for true cluster labels $\{z_i\}$. Now take $\delta = \varepsilon\sin\alpha$. Figure S2 shows the geometry of this construction.

To show the result, it is enough to use Theorem 2 with properly chosen (fake) centers on the above dataset. In particular, we are going to show that a 2-factor k-means algorithm has a small misclassification rate with respect to a new clustering that puts points $\{a_1, b_1\}$ in one cluster and $\{a_2, b_2\}$ in another cluster. Consider "fake" centers $\xi_1^{**} = (a_1 + b_1)/2$ and $\xi_2^{**} = (a_2 + b_2)/2$. Then, the new separation is $\delta^* = 2\varepsilon\cos\alpha$ and the new deviation can be taken to be $\varepsilon^* = \delta/2 + \varepsilon\sin\alpha = (3/2)\varepsilon\sin\alpha$ guaranteeing that $(\frac{1}{n}\sum_i \|x_i - \xi_{y_i}^{**}\|^2)^{1/2} \le \varepsilon^*$ where $\{y_i\}$ are labels relative to the new clustering.

Applying Theorem 2 with $\kappa = p = 2$, $c = 2.1$ and $\pi_{\min} = 1/2$, as long as $\delta^*/\varepsilon^* \ge 9 > 3\sqrt{2}c$, the misclassification rate to the new clustering is bounded above as $\text{Miss}^* \le 80(\varepsilon^*/\delta^*)^2$. We have $\varepsilon^*/\delta^* = (3/4)\tan\alpha$. Thus, for $\alpha \le \tan^{-1}(4/27)$ we have $\text{Miss}^* \le 45(\tan\alpha)^2$ w.r.t. to clustering $\{\{a_1, b_1\}, \{a_2, b_2\}\}$. Hence, w.r.t. the original clustering, $\frac{1}{2} \ge \text{Miss} \ge \frac{1}{2} - 45(\tan\alpha)^2$, which can be made arbitrarily close to $\frac{1}{2}$ by choosing α small enough.

To see the last step above, let q_1, q_2, q_3, q_4 be the fractions of misclassified nodes from each of the four categories a_1, a_2, b_1, b_2, w.r.t. to the new clustering (i.e., $\{y_i\}$). The above argument shows that $\frac{1}{4}(q_1 + q_2 + q_3 + q_4) \le 45(\tan\alpha)^2$. The misclassification rate to the original clustering (i.e., $\{z_i\}$) is then

$$\text{Miss} = \frac{1}{n}\Big(\frac{n}{4}(1 - q_{i_1}) + \frac{n}{4}(1 - q_{i_2})\Big) = \frac{1}{2} - \frac{1}{4}(q_{i_1} + q_{i_2}) \ge \frac{1}{2} - 45(\tan\alpha)^2$$

where $\{i_1, i_2\}$ is a pair of distinct elements from $\{1, 2, 3, 4\}$. This proves the lower bound. The upper bound Miss $\leq 1/2$ always holds due to the minimization over permutations in the definition of the misclassification rate.

Since for $\beta \in (0, 1/2)$, $\sqrt{\beta/45} \leq 4/27$, we only need $\alpha \leq \tan^{-1}(\sqrt{\beta/45})$ to have $\frac{1}{2} \geq$ Miss $\geq \frac{1}{2} - \beta$. Recalling that $\delta/\varepsilon = \sin \alpha$, this shows that one can take $c_2(\beta) = \sin(\tan^{-1}(\sqrt{\beta/45}))$ in the statement of the lower bound.

For the claim regarding perfect recovery with $L = 4$ clusters, take $\xi_1^{**} = a_1$, $\xi_2^{**} = b_1$, $\xi_3^{**} = a_2$ and $\xi_4^{**} = b_2$ and apply Theorem 1, noting that $\delta^* = \min_{i \neq j} \|\xi_i^{**} - \xi_j^{**}\| > 0$ while we can take $\varepsilon^* = 0$.

D Experiment details

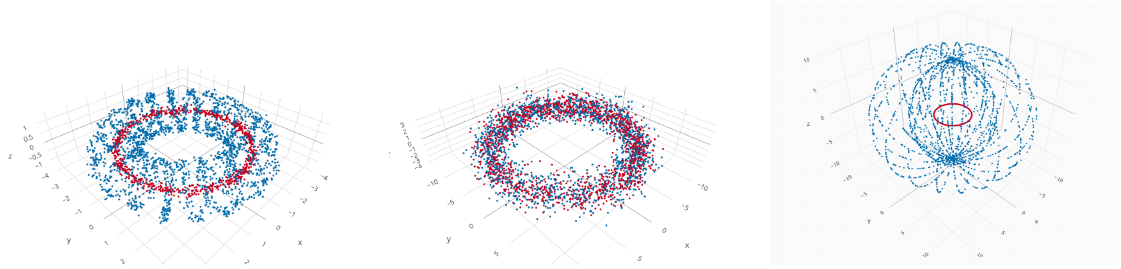
The code for numerical experiments are executed in R [3] version 4.0.3 on a Linux system with 48 CPU cores. The code is provided as a ZIP file as part of the supplementary material. We use the `kmeans` function in base R and go with the default algorithm of Hartigan and Wong [2]. We set the number of random starts to `nstart = 20` and the maximum number of iterations allowed to `iter.max = 200`.

D.1 Circle-torus model

The circle-torus model is a mixture of two parts: (1) The uniform distribution on the circumference of a circle in the xy-plane, at the origin, and (2) a torus built around this circle. Parametrically, these two clusters can be defined via the following equations,

$$\begin{aligned} x_1 &= R\cos(t) & \qquad & x_2 = (R + r\cos(mt))\cos(t) \\ y_1 &= R\sin(t) & \text{and} \qquad & y_2 = (R + r\cos(mt))\sin(t) \\ z_1 &= 0 & \qquad & z_2 = r\sin(mt). \end{aligned} \tag{S5}$$

Here R is the radius of the circle on the plane and also the distance from the center of the tube to the center of the torus. r is the radius of the tube and it is also the minimal distance between two clusters. We also created a noisy version by adding $N(0, \sigma^2 I_3)$ to the model. Figure 3 shows the geometry of the two clusters in the case $R = 3, r = 1$ and $\sigma = 0$. The other two panels in Figure 3 show the average missclassification rate over 32 repetitions versus $\delta := r$, for both the noiseless and noisy ($\sigma = 1$) circle-torus model. In both cases, we let $R = 3$ and very r (i.e., δ), from 0.1 to 10. In Figure S3, we include additional scatter plots of the circle-torus model for various settings of the parameters (R, r, σ). Figure S3(a) is the noisy version of Figure 3(a) with noise level $\sigma = 0.1$. Figure S3(b) shows that for sufficiently small r and high noise, the two clusters are nearly indistinguishable. Figure S3(c) shows the scatter plot for $R = 3$ and $r = 10$; it is an example of how the model looks like when $R < r$.



(a) $R = 3$, $r = 1$ and $\sigma = 0.1$. (b) $R = 10$, $r = 1$ and $\sigma = 1$. (c) $R = 3$, $r = 10$ and $\sigma = 0$.

Figure S3: Scatter plots for the circle-torus model. True clusters are distinguished by their color.

E Norm of a sub-gaussian vector

We first recall the definition of a sub-gaussian vector: A random vector $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$ is sub-gaussian if the one-dimensional marginals $u^T X$ are sub-gaussian random variables for all $u \in$

$\mathbb{R}^d$ [4, Definition 3.4.1]. The sub-gaussian norm of X is defined as $\|X\|_{\psi_2} = \sup_{u \in S^{d-1}} \|u^T X\|_{\psi_2}$, where $\|\cdot\|_{\psi_2}$ denotes the sub-gaussian norm of a random variable and S^{d-1} the unit sphere in $\mathbb{R}^d$. Alternatively, we can define a sub-gaussian vector with parameter σ, as a random vector satisfying $\mathbb{P}(|u^T X| \geq t) \leq 2\exp(-\frac{t^2}{2\sigma^2})$ for all $u \in S^{d-1}$ and $t \geq 0$. We will have $\sigma \asymp \|X\|_{\psi_2}$. We also use $\|\cdot\|_{\psi_1}$ for the sub-exponential norm of a random variable. For any random variable, we have $\|Y^2\|_{\psi_1} = \|Y\|_{\psi_2}^2$ [4, Lemma 2.7.6]. Below we apply this fact with $Y = \|X\| = (\sum_{i=1}^d X_i^2)^{1/2}$, leading to the following useful lemma.

Lemma S1. *Assume that $X \in \mathbb{R}^d$ is a sub-gaussian random vector with parameter σ. Then, $\|X\|$ is sub-gaussian with parameter $\lesssim \sigma\sqrt{d}$. In fact, for some universal constant $C > 0$,*

$$\|\|X\|\|_{\psi_2} \leq C\sigma\sqrt{d}, \quad \|\|X\|^2\|_{\psi_1} \leq C^2\sigma^2 d.$$

Proof. We have $\|\|X\|^2\|_{\psi_1} \leq \sum_{i=1}^d \|X_i^2\|_{\psi_1} = \sum_{i=1}^d \|X_i\|_{\psi_2}^2 \leq dC^2\sigma^2$, for some universal constant $C^2 > 0$. The first inequality is the triangle inequality for $\|\cdot\|_{\psi_1}$ and the second by the equivalence of the sub-gaussian norm and sub-gaussian parameter. Next, we note that $\|\|X\|\|_{\psi_2} = \sqrt{\|\|X\|^2\|_{\psi_1}}$ and the result follows. □

F Details of the sub-gaussian mixture example

By assumption, w_i is a sub-gaussian vector with parameter σ_i. Then, by Lemma S1, $\|w_i\|^2/d$ is sub-exponential with sub-exponential norm $\lesssim \sigma_i^2$. By the Bernstein inequality for sub-exponential variables [4, Corollary 2.8.3],

$$\mathbb{P}\Big(\frac{1}{n}\Big(\sum_{i=1}^n \frac{\|w_i\|^2}{d} - \alpha_i^2\Big) > t\Big) \leq \exp\Big(-cn\min\Big(\frac{t^2}{\sigma_{\max}^4}, \frac{t}{\sigma_{\max}^2}\Big)\Big).$$

Let $t = \bar{\alpha}_n$, and recall that $\bar{\alpha}_n^2/\sigma_{\max}^2 \leq C$. Then, for a constant $c_1 > 0$,

$$\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n \frac{\|w_i\|^2}{d} > 2\bar{\alpha}_n^2\Big) \leq \exp\Big(-c_1 n\frac{\bar{\alpha}_n^2}{\sigma_{\max}^4}\Big).$$

In the Gaussian case $w_i \sim N(0, \Sigma_i)$, it is not hard to see that w_i is a sub-guassian vector with parameter $\|\Sigma_i\|_{\text{op}}$. Therefore, in Gaussian mixtures, we have $\sigma_{\max} = \max_i \|\Sigma_i\|_{\text{op}}$ and $\bar{\alpha}_n^2 = \sum_{i=1}^n \frac{1}{n}\operatorname{tr}(\Sigma_i)/d$.

G Extension of Proposition 3

For random vector x with distribution $\mu_{\mathcal{C}}$ on some subset $\mathcal{C} \subset \mathbb{R}^d$, let $N_{\mu_{\mathcal{C}}}(\varepsilon)$ be the smallest integer for which there is a high probability ε-cover of x, that is, a finite subset $\mathcal{N} \subset \mathcal{C}$ such that $\mathbb{P}(\min_{y \in \mathcal{N}} \|x - y\| \leq \varepsilon) \geq 1 - n^{-2}$. We refer to $N_{\mu_{\mathcal{C}}}(\varepsilon)$ as the stochastic covering number of $\mu_{\mathcal{C}}$. We have the following extension of Proposition 3.

Proposition S1. *Assume that $\{x_i\}_{i=1}^n$ are independent draws from a K-mixture where the kth component is a distribution $\mu_{\mathcal{C}_k}$ on a subset $\mathcal{C}_k \subset \mathbb{R}^{r_k}$. Let z_i be the label of x_i so that $x_i \mid z_i = k \sim \mu_{\mathcal{C}_k}$. Assume that*

$$\min_{x \in \mathcal{C}_k,\, y \in \mathcal{C}_{k'}} \|x - y\| \geq \delta > 0, \quad \textit{for all } k \neq k'.$$

Let $N_{\mu_{\mathcal{C}_k}}(\varepsilon)$ be the stochastic covering number of $\mu_{\mathcal{C}_k}$. Then, there exist a constant $C = C(K, \delta, \kappa)$ such that any ALG(p), satisfying Assumption 1, applied with $L_n = \sum_{k=1}^K N_{\mu_{\mathcal{C}_k}}(Cn^{-1/p})$ clusters, recovers a perfect refinement of $z = (z_i)$ with probability $\geq 1 - n^{-1}$.

Proof. Let $\mathcal{N}_k \subset \mathcal{C}_k$ be the ε-net that realizes the stochastic ε-covering number of $\mu_{\mathcal{C}_k}$ and let $\pi_k : \mathcal{C}_k \to \mathcal{N}_k$ be the corresponding projection operator. Then, for any $i \in [n]$ for which $z_i = k$, we have $\mathbb{P}(\|x_i - \pi_k(x_i)\| > \varepsilon) \leq n^{-2}$. By union bound, we have $\|x_i - \pi_{z_i}(x_i)\| \leq \varepsilon$ for all $i \in [n]$ with probability at least $1 - n^{-1}$. The collection of the fake centers $\{\widetilde{\xi}_\ell\}_{\ell=1}^{L_n}$ can be taken to be the union of the nets $\bigcup_{k=1}^K \mathcal{N}_k$ with cardinality $L_n = \sum_k N_{\mu_{\mathcal{C}_k}}(\varepsilon)$. The rest of the proof follows those of Propositions 2 and 3 with $\varepsilon = (c_1 n^{1/p})^{-1}$, $c_2 = 3(1 + \kappa)/\delta$ and $c_1 = K^{1/p}c_2$. (Note that there is no condition $\varepsilon < 1/\rho$ that needs to be satisfied in this case.) □

H Remaining proofs

Proof of Corollary 1. We first construct fake centers $(\widetilde{\xi}_\ell)$ for (x_i) as in the proof of Proposition 2 and treat them as the fake centers for y_i. By the triangle inequality,

$$\Big(\frac{1}{n}\sum_{i=1}^n \|y_i - \widetilde{\xi}_{\widetilde{z}_i}\|^2\Big)^{1/2} \le \Big(\frac{1}{n}\sum_{i=1}^n \|x_i - \widetilde{\xi}_{\widetilde{z}_i}\|^2\Big)^{1/2} + \Big(\frac{1}{n}\sum_{i=1}^n \|w_i/\sqrt{d}\|^2\Big)^{1/2} \le \varepsilon + \sqrt{2}\bar{\alpha}_n$$

holds with probability at least $1 - p_n - n^{-1}$. The result follows by applying Theorem 3. □

Proof of Proposition 3. The proof follows that of Proposition 2. We only highlight the differences. When $z_i = k$, by Lemma S1, $\|t_i - m_k\|$ is sub-gaussian with parameter $\le c_0\sigma\sqrt{r_k}$ for some universal constant $c_0 > 0$. Thus, we have $\mathbb{P}(\|t_i - m_k\| \ge t) \le 2e^{-c_0 t^2/(r_k\sigma^2)}$. Let $M = \sqrt{3c_0 r\sigma^2 \log n}$. By union bound, with probability at least $1 - 2n^{-2}$, we have $\|t_i - m_{z_i}\| \le M$ for all $i \in [n]$. The ε-cover has to be constructed for $\{u : \|u\| \le M\}$ in the ℓ_2 norm, which can be done with a net of size at most $L' = (1 + 2M/\varepsilon)^r$. Take $\varepsilon = (c_1 n^{1/p})^{-1}$ and let $c_2 = 3\rho(1+\kappa)/\delta$. As long as $n\pi_{\min} > (c_2/c_1)^p$, the separation condition in (8) is satisfied and we have $\mathrm{Miss}(z, \widehat{z}) \le K(c_2/c_1)^p/n$. Hence, as long as $c_1 > K^{1/p}c_2$, we will have $\mathrm{Miss}(z, \widehat{z}) < 1/n$ which implies $\mathrm{Miss}(z, \widehat{z}) = 0$. We also need to satisfy $\varepsilon < 1/\rho$ that is $c_1 \ge \rho/\sqrt{n}$. Taking $c_1 = K^{1/p}c_2 + \rho$ satisfies all the required constraints on c_1. The required number of clusters is

$$\begin{aligned} L_n = KL' = K(1 + 2M/\varepsilon)^r &= K(1 + 2c_1\sqrt{3c_0 r\sigma^2} n^{1/p}\sqrt{\log n})^r \\ &\le C(n^{1/p}\sqrt{\log n})^r \end{aligned}$$

for $C = K(2 + 2c_1\sqrt{3c_0 r\sigma^2})^r$. Here, we have used $1 \le 2n^{1/p}\sqrt{\log n}$ for $n \ge 2$. Note that since $c_2/c_1 < 1$ and $n\pi_{\min} \ge 1$, the condition $n\pi_{\min} > (c_2/c_1)^p$ is automatically satisfied. The proof is complete. □

I Proofs of the lemmas

Proof of Lemma 1. Recall that $\widehat{\xi}$ is the output of ALG for L clusters. Let $\widehat{\xi}^{(K)}$ be the output of the ALG for K clusters. Then, since $L \ge K$,

$$\widehat{Q}(\widehat{\xi}) \le \widehat{Q}(\widehat{\xi}^{(K)}) \le \kappa\, \widehat{Q}^{(K)}_{\min}, \quad \text{where} \quad \widehat{Q}^{(K)}_{\min} := \min_{\xi \in \mathcal{X}^K} \widehat{Q}(\xi).$$

The first inequality is by the monotonicity of ALG and the second by its constant-factor approximation property. Since by assumption $\xi^* \in \mathcal{X}^K$, we have

$$\widehat{Q}^{(K)}_{\min} \le \widehat{Q}(\xi^*) \le \Big(\frac{1}{n}\sum_{i=1}^n \|x_i - \xi^*_{z_i}\|^p\Big)^{1/p} \le \varepsilon.$$

It follows that $\widehat{Q}(\widehat{\xi}) \le \kappa\varepsilon$. Recalling (9) and noting that $\widehat{Q}(\widehat{\xi}) = \big(\frac{1}{n}\sum_{i=1}^n \|x_i - \widehat{\xi}_{\widehat{z}_i}\|^p\big)^{1/p}$, we have

$$\begin{aligned} Q(\widehat{\xi}; \mu^*) = \Big(\frac{1}{n}\sum_{i=1}^n \min_{\ell \in [L]} \|\xi^*_{z_i} - \widehat{\xi}_\ell\|^p\Big)^{1/p} &\le \Big(\frac{1}{n}\sum_{i=1}^n \|\xi^*_{z_i} - \widehat{\xi}_{\widehat{z}_i}\|^p\Big)^{1/p} \\ &\le \Big(\frac{1}{n}\sum_{i=1}^n \|x_i - \xi^*_{z_i}\|^p\Big)^{1/p} + \Big(\frac{1}{n}\sum_{i=1}^n \|x_i - \widehat{\xi}_{\widehat{z}_i}\|^p\Big)^{1/p} \\ &\le \varepsilon + \kappa\varepsilon \qquad \text{(S6)} \end{aligned}$$

where the second line is the triangle inequality in the aforementioned $L^p(\nu_n, \mathcal{X})$ space. The proof is complete. □

Proof of Lemma 2. Consider the partition of the space by the Voronoi cells of $\xi = (\xi_\ell)$. Assume first that there is a Voronoi cell that contains at least two distinct elements of ξ^*, e.g., $\xi^*_{k_1}$ and $\xi^*_{k_2}$, with $k_1 \neq k_2$, both belonging to the Voronoi cell of ξ_{ℓ_0}. That is, $\min_\ell \|\xi^*_k - \xi_\ell\| = \|\xi^*_k - \xi^*_{\ell_0}\|$ for

$k = k_1, k_2$. As $\|\xi^*_{k_1} - \xi^*_{k_2}\| \le \|\xi^*_{k_2} - \xi^*_{\ell_0}\| + \|\xi^*_{k_1} - \xi^*_{\ell_0}\|$, at least one of the $k = k_1, k_2$ satisfy $\|\xi^*_k - \xi^*_{\ell_0}\| \ge \|\xi^*_{k_1} - \xi^*_{k_2}\|/2$, and assume this is true for $k = k_1$, we have

$$Q(\xi; \mu^*) \ge \pi^{1/p}_{\min} \|\xi^*_{k_1} - \xi_{\ell_0}\| \ge \frac{\pi^{1/p}_{\min}}{2} \|\xi^*_{k_1} - \xi^*_{k_2}\| \ge \frac{\pi^{1/p}_{\min}}{2} \delta.$$

Otherwise, each Voronoi cell of ξ contains at most one element of ξ^*. On the other hand, each element of ξ^* belongs to at least one Voronoi cell of ξ, since the union of Voronoi cells is the whole space. It follows that there are K distinct Voronoi cells of ξ, each of which contains exactly one element of ξ^*. Thus, there is an injective map $\sigma : [K] \to [L]$ such that ξ^*_k belongs to Voronoi cell of $\xi_{\sigma(k)}$, that is, $\min_\ell \|\xi^*_k - \xi_\ell\| = \|\xi^*_k - \xi_{\sigma(k)}\|$. Then,

$$Q(\xi; \mu^*) \ge \pi^{1/p}_{\min} \Big(\sum_{k=1}^{K} \|\xi^*_k - \xi_{\sigma(k)}\|^p \Big)^{1/p} \ge \pi^{1/p}_{\min}\, d_p(\xi, \xi^*).$$

The proof is complete. □

Proof of Lemma 3. By assumption, there exists an injective map $\sigma : [K] \to [L]$ such that

$$\max_{k \in K} \|\xi^*_k - \widehat{\xi}_{\sigma(k)}\| \le \gamma.$$

Then, σ is invertible on $\mathrm{Im}(\sigma) := \{\sigma(k) : k \in [K]\}$, with an inverse denoted as σ^{-1}. We obtain

$$\|\xi^*_{\sigma^{-1}(\ell)} - \widehat{\xi}_\ell\| \le \gamma, \quad \forall \ell \in \mathrm{Im}(\sigma). \tag{S7}$$

First assume that $\widehat{z}_i \in \mathrm{Im}(\sigma)$. We prove that $\sigma(z_i) = \widehat{z}_i$ by contradiction. Suppose that $\sigma(z_i) \ne \widehat{z}_i$. Then, we show that $\|x_i - \widehat{\xi}_{\sigma(z_i)}\| < \|x_i - \widehat{\xi}_{\widehat{z}_i}\|$ contradicting $\widehat{z}_i = \operatorname*{argmin}_\ell \|x_i - \widehat{\xi}_\ell\|^2$. By the triangle inequality

$$\|x_i - \widehat{\xi}_{\sigma(z_i)}\| \le \|x_i - \xi^*_{z_i}\| + \|\widehat{\xi}_{\sigma(z_i)} - \xi^*_{z_i}\| \le \eta + \gamma. \tag{S8}$$

Since $\widehat{z}_i \in \mathrm{Im}(\sigma)$ and $\sigma(z_i) \ne \widehat{z}_i$, we have $\sigma^{-1}(\widehat{z}_i) \ne z_i$. By (S7), $\|\widehat{\xi}_{\widehat{z}_i} - \xi^*_{\sigma^{-1}(\widehat{z}_i)}\| \le \gamma$. Therefore,

$$\begin{aligned}
\|x_i - \widehat{\xi}_{\widehat{z}_i}\| &\ge \|\widehat{\xi}_{\widehat{z}_i} - \xi^*_{z_i}\| - \|x_i - \xi^*_{z_i}\| \\
&\ge \|\xi^*_{z_i} - \xi^*_{\sigma^{-1}(\widehat{z}_i)}\| - \|\widehat{\xi}_{\widehat{z}_i} - \xi^*_{\sigma^{-1}(\widehat{z}_i)}\| - \eta \\
&\ge \delta - \gamma - \eta.
\end{aligned} \tag{S9}$$

Since by assumption $\delta > 2\gamma + 2\eta$, the claimed contradiction follows by combining (S8) and (S9). Hence, we have $\sigma(z_i) = \widehat{z}_i$ when $\widehat{z}_i \in \mathrm{Im}(\sigma)$. Define $\omega(\cdot) = \sigma^{-1}(\cdot)$ on $\mathrm{Im}(\sigma) \subset [L]$. Then, ω satisfies $\omega(\widehat{z}_i) = z_i$ whenever $\widehat{z}_i \in \mathrm{Im}(\sigma)$. This finishes proof for the case $L = K$.

Next, we define ω for $\ell_0 \notin \mathrm{Im}(\sigma)$. Since $\widehat{\xi}$ is an efficient solution, there exists at least one $i \in [n]$ such that $\widehat{z}_i = \ell_0$. When there is only one such i, we can just let $\omega(\ell_0) = \omega(\widehat{z}_i) = z_i$. When there are at least two data points x_i and x_j such that $\widehat{z}_i = \widehat{z}_j = \ell_0$, we are going to show, by contradiction, that their true cluster labels must be the same, i.e., $z_i = z_j$. Suppose that $z_i \ne z_j$, then we will show that $\|x_i - \widehat{\xi}_{\ell_0}\| > \|x_i - \widehat{\xi}_{\sigma(z_i)}\|$ which contradicts x_i being in the Voronoi cell of $\widehat{\xi}_{\ell_0}$. Inequality (S8) still holds in this case. Furthermore

$$\begin{aligned}
\|x_i - \widehat{\xi}_{\ell_0}\| &\ge \|\widehat{\xi}_{\ell_0} - \xi^*_{z_i}\| - \|x_i - \xi^*_{z_i}\| \\
&\ge \|x_j - \xi^*_{z_i}\| - \|x_j - \widehat{\xi}_{\ell_0}\| - \eta \\
&\ge \|\xi^*_{z_i} - \xi^*_{z_j}\| - \|x_j - \xi^*_{z_j}\| - \|x_j - \widehat{\xi}_{\ell_0}\| - \eta \\
&\ge \delta - 2\eta - \|x_j - \widehat{\xi}_{\ell_0}\|.
\end{aligned} \tag{S10}$$

Since x_j is in the Voronoi cell of $\widehat{\xi}_{\ell_0}$ and $\ell_0 \notin \mathrm{Im}(\sigma)$, we have $\ell_0 \ne \sigma(z_j)$. Therefore,

$$\begin{aligned}
\|x_j - \widehat{\xi}_{\ell_0}\| &\le \|x_j - \widehat{\xi}_{\sigma(z_j)}\| \\
&\le \|x_j - \xi^*_{z_j}\| + \|\widehat{\xi}_{\sigma(z_j)} - \xi^*_{z_j}\| \\
&\le \eta + \gamma.
\end{aligned} \tag{S11}$$

Combining inequalities (S8), (S10) and (S11) and using the assumption $\delta > 2\gamma + 4\eta$, we get

$$\|x_i - \widehat{\xi}_{\ell_0}\| \;\geq\; \delta - 3\eta - \gamma \;>\; \eta + \gamma \;\geq\; \|x_i - \widehat{\xi}_{\sigma(z_i)}\|$$

which is the claimed contradiction. Therefore, we can define ω on $[L] \setminus \mathrm{Im}(\sigma)$ so that $\omega(\widehat{z}_i) = z_i$ when $\widehat{z}_i \notin \mathrm{Im}(\sigma)$. Combining with the definition of ω on $\mathrm{Im}(\sigma)$, we have successfully constructed a surjective map $\omega : [L] \to [K]$ satisfying $\omega(\widehat{z}_i) = z_i$ for all $i \in [n]$. The proof is complete. $\square$

References

[1] Pranjal Awasthi, Avrim Blum, and Or Sheffet. Stability yields a ptas for k-median and k-means clustering. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 309–318. IEEE, 2010.

[2] John A Hartigan and Manchek A Wong. A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):100–108, 1979.

[3] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[4] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.