Label consistency in overfitted generalized k-means

Anonymous Author(s)
Affiliation
Address
email

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 points to true cluster centers. These centers themselves are loosely defined and can be taken to be any sets of points for which the aforementioned distances can be controlled. We show the usefulness of the results with some examples.

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 generalized k-means clustering, one attempts to minimize the ideal cost function

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

where $\xi = (\xi_1,\dots,\xi_K) \in \mathcal{X}^K$ is a set of K vectors in $\mathcal{X}$, for some fixed integer K. Minimizing the loss function $Q(\,\cdot\,;\mu)$ over $\mathcal{X}^K$ gives $\xi^* \in \mathcal{X}^K$, the best approximation of μ by K points. Elements of ξ^* serve as the centers of the *true clusters*. The clusters themselves are given by the Voronoi cells of $\{\xi_k^*\}$. Each point $x \in \mathcal{X}$ can be assigned a true cluster label $\mathcal{C}(x;\xi^*) := \operatorname{argmin}_k \|x-\xi_k^*\|_{\mathcal{X}}$.

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. In solving the empirical version, we allow the number of clusters L to be potentially different from the ideal K. A minimizer of $\widehat{Q}(\cdot)$ over $\mathcal{X}^L$, denoted as $\widehat{\xi} = (\widehat{\xi}_1,\dots,\widehat{\xi}_L)$ provides estimates of the true cluster centers ξ^*. Each of the data points x_i will have a true cluster label $z_i := \mathcal{C}(x_i;\xi^*) \in [K]$ and an estimated cluster label $\widehat{z}_i := \mathcal{C}(x_i;\widehat{\xi}) \in [L]$. One can ask how close the estimated centers $\widehat{\xi}$ are to the true centers ξ^*. Another question of interest is how close are the estimated labels $(\widehat{z}_i)$ to the true labels (z_i). This is the *label consistency* problem.

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 the problem has been performed in this case. Early work has focused on recovering the centers and/or the optimal

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

value the cost function. Such consistency results are proved, for the global minimizers of (2), in the early work of [17, 23, 24] and also in [24, 14] 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.

More recently, there has been more interest in the consistency of practical k-means algorithms [10, 16] as well as the label consistency problem. Lu and Zhou [16] obtain sharp bounds on the label consistency of the Lloyd's algorithm [15] under a sub-Gaussian mixture model. Semidefinite programming (SDP) relaxation is another popular technique for deriving polynomial-time approximations to the k-means problem [22]. Its label consistency has been studied when data is generated from the stochastic ball model [3, 7], sub-Gaussian mixtures [19, 5, 6], the Stochastic Block Model (SBM) [6] and general models [13]. Convex clustering is another relaxation method whose label consistency has been discussed in [26, 21, 8, 25]. The literature on community detection in SBM, a network clustering problem, is also mainly focused on label consistency; see [1] for a review of those results.

In this paper, we study 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 could have a natural number of clusters K that is unknown a priori. An example is the sub-Gaussian mixture with K components. All the aforementioned work on label consistency exclusively considers the correctly-fitted case $L = K$. We show that when the data-generating centers satisfy certain separation conditions, estimated labels, with $L \geq K$ clusters, are close to a *refinement* of the true labels. These bounds reduce to label consistency criteria for $L = K$, but have no counterpart in the literature for $L > K$.

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 assumption on the data generation process. In fact, we relax the notion of true centers ξ^* and true labels (z_i) to any collection of centers and labels for which deviation bounds to the sample $\{x_i\}$ can be established. This is important in practice, since getting a handle on the true centers minimizing (1) is itself a challenging task for a given stochastic data model. Since k-means clustering often appears as a building block in many sophisticated clustering algorithms, we believe these 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, in the sequel, we 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 $\widehat{\xi}_\ell^{(L)}$ contains at least one of $\{x_i\}$.

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

Many practical k-means algorithms have properties in Assumption 1. For example, the algorithms in [18, 11, 2] can achieve the κ-approximation of the optimal objective. Efficiency can be achieved by substituting centers whose Voronoi cells have an empty intersection with $\{x_i\}$, by those having the opposite property. 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 misclassiciation 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 labels. 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. 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.

Note that, in the case $K = L$, d_∞ is an upper bound on the Hausdorff distance between the two sets (it is a stronger distance). Obviously, we have $d_\infty \leq d_p$ for any $p \geq 1$.

2.1 Distance to true centers

Let $z = (z_i)_{i=1}^n$ be a given set of true labels for the data points $(x_i)_{i=1}^n$ and let $\xi^* = (\xi^*_k)_{k=1}^K$ be a set of true centers. Throughout, we only vaguely specify z and ξ^*. The only requirement on them is that together with the observed data points (x_i), they satisfy the deviation bounds in each theorem, e.g., $\max_{1 \leq i \leq n} \|x_i - \xi^*_{z_i}\| \leq \eta$ in Theorem 1, etc. In particular, ξ^* is not assumed to be a minimizer of (1). Let $\pi_k = \sum_{i=1}^n 1\{z_i = k\}/n$ be the proportion of observed data points in cluster k and let $\pi_{\min} = \min_k \pi_k$.

We let $\widehat{\xi}$ be a solution of the k-means algorithm with $L \geq K$ centers and $\widehat{z}_i \in \mathrm{argmin}_\ell \|x_i - \widehat{\xi}_\ell\|$ 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 $\max_{1 \leq i \leq n} \|x_i - \xi^*_{z_i}\| \leq \eta$, and*

$$\min_{(k,k'):\, k \neq k'} \|\xi^*_k - \xi^*_{k'}\| \geq \delta. \tag{3}$$

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 $L \geq K$. Then,*

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

When $K = L$, the theorem shows that the clusters of $\{\widehat{z}_i\}$ and $\{z_i\}$ are in one-to-one correspondence, i.e., we have the exact recovery of labels $\{z_i\}$ in the classical sense. In this case, we can also remove the assumption of efficiency for ALG(p) and relax the separation condition in (4) to $\delta/\eta > 2(1+\kappa)/\pi_{\min}^{1/p} + 2$.

Example 1 (Stochastic Ball Model)**.** Assume that data are generated from the stochastic ball model considered in [20], 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), providing provable guarantees for any constant-factor approximation, with an explicit bound on the separation δ, is a nontrivial task.

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 relaxed bound assumed on the deviations $\|x_i - \xi^*_{z_i}\|$.

Example 2 (Sub-Gaussian mixtures)**.** Let us assume that the data is generated from a K component sub-Gaussian mixture model $x_i = \xi^*_{z_i} + \frac{1}{\sqrt{d}} w_i$, 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 [4]. Determining whether $(\xi^*_k)_{k=1}^K$ is actually the solution of the population problem (1) is, itself, a nontrivial task and the answer could 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 Supplementary Material. Let $\sigma_{\max} = \max_i \sigma_i$ and set $\alpha_i^2 := \mathbb{E}\|\frac{1}{\sqrt{d}} w_i\|_2^2$ 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 $\bar{\alpha}_n^2 \le C\sigma_{\max}^2$. Then, one can show that

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

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 $1 - p_n$,

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

where δ is as in (3) and $c > 2$. In a general problem, $\bar{\alpha}_n$, $\sigma_{\max}$ and δ all could 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 $\sigma_{\max} = \max_{1\le i\le n} \|\Sigma_i\|$ where $\|\cdot\|$ denotes the operator norm, and $\bar{\alpha}_n^2 := \frac{1}{n}\sum_{i=1}^n \mathrm{tr}(\Sigma_i)/d$. □

For both Theorems 1 and 2, it is possible to relax part (b) of Assumption 1 to the modified version (b') discussed earlier.

2.2 Distance to fake centers

We now extend Theorem 3, to allow for "fake" centers $\{\widetilde{\xi}_\ell\}_{\ell=1}^L$ and 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*

$$\Big(\frac{1}{n}\sum_{i=1}^n \|x_i - \widetilde{\xi}_{\widetilde{z}_i}\|^p\Big)^{1/p} \leq \varepsilon, \qquad \min_{\ell_1 \neq \ell_2,\, \widetilde{\omega}(\ell_1) \neq \widetilde{\omega}(\ell_2)} \|\widetilde{\xi}_{\ell_1} - \widetilde{\xi}_{\ell_2}\| \geq \delta \tag{6}$$

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

$$\frac{\delta}{\varepsilon} > \frac{(1+\kappa)c}{\pi_{\min}^{1/p}} \qquad \Longrightarrow \qquad \mathrm{Miss}(z, \widehat{z}) < K(1+\kappa)^p c^p \Big(\frac{\varepsilon}{\delta}\Big)^p. \tag{7}$$

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. 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 an application of Theorem 3 in a setting 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 1. *Assume that $\{x_i\}_{i=1}^n$ are independent draws from a K-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 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 highly nontrivial to achieve it when $L_n/n \to 0$. This is especially the case since one can achieve quite complex cluster configurations with the model in Proposition 1, for some of which applying k-means with K clusters will have a misclassification rate bounded away from zero. Figure 1 and 3 show examples of the mixture-of-curves model. There are two clusters, a uniform distribution on a circle in $\mathbb{R}^3$ and a Gaussian distribution on a line passing through the center of the circle. The centers of these two clusters coincide at zero, causing any k-means algorithm applied with $K = 2$ to incur a substantial error.

Various extensions of Proposition 1 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 1 and w_i are sub-Gaussian noise vectors as in Example 2. Then, under the same assumptions as in Proposition 1, 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.

3.2 Mixture of higher-order submanifolds

A version of Proposition 1 can be stated for a higher-dimensional version of the mixture of curves models, 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 $t, s \in \mathbb{R}$ such that $\|t - s\| \leq \frac{1}{\rho}$.

Proposition 2. *Assume that $\{x_i\}_{i=1}^n$ are independent draws from a K-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 2 to this setting in the Supplementary Material.

3.3 Numerical experiments

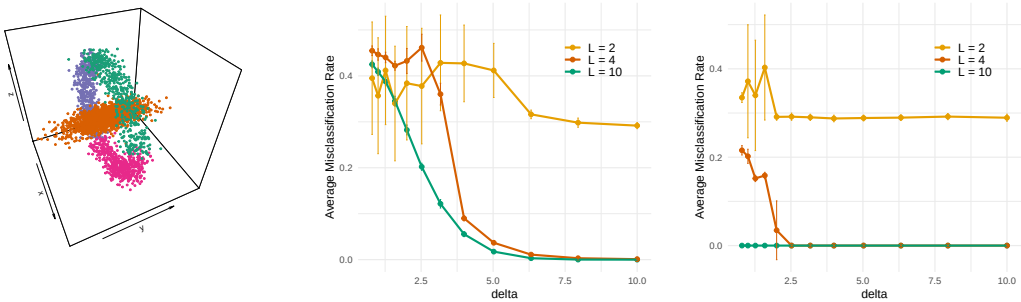
We first consider the (noiseless) line-circle model in $\mathbb{R}^3$, an example of mixture-of-curves. This has two clusters: (1) The uniform distribution on the circumference of a circle in the xz-plane, centred at the origin, and (2) the standard Gaussian distribution on the y axis. The minimum separation 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. 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 the purpose of 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 when $L = 2$, when L increases, 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 model. Both show that 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 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 $\Sigma_2 = \mathrm{diag}(\sigma^2, 0)$. Here, we set $\sigma = 5$ and $n = 3000$ data points are equally sampled 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 as a general Gaussian mixture. Figure 2 shows the average missclassification rate over 32 repetitions for different L. The results are consistent with Theorem 3. The plots show that as δ increases, the misclassification rate decreases.

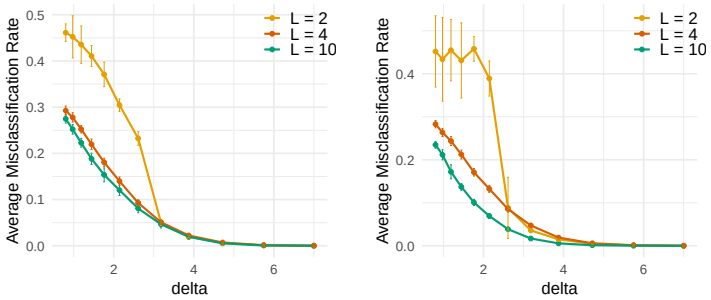
4 Proofs

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



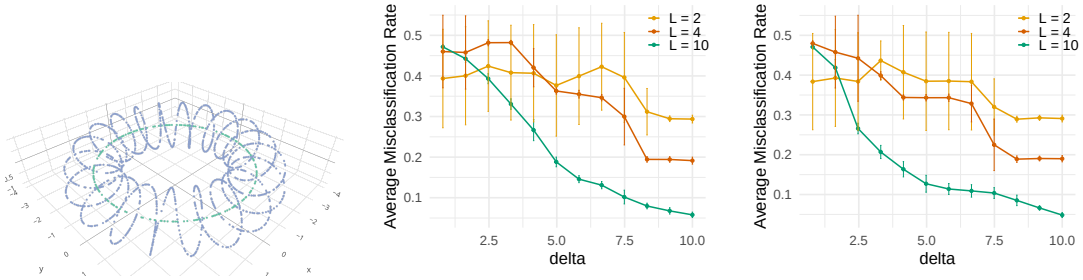
(a) Noisy ($L = 4, \delta = 3$) (b) Noisy (c) Noiseless

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.



(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 noisy and noiseless versions of the model.



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

Figure 3: Circle-torus model: The (generalized) misclassification rate versus δ, the small radius of the torus, in the noisy and noiseless versions of the model.

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, 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 $\xi \in \mathcal{X}^L$,

$$Q(\xi; \mu^*)^p = \sum_{k=1}^K \pi_k \min_{1 \le \ell \le L} \|\xi_k^* - \xi_\ell\|^p = \frac{1}{n}\sum_{i=1}^n \min_{1 \le \ell \le L} \|\xi_{z_i}^* - \xi_\ell\|^p. \tag{8}$$

The proofs relies on the following three lemmas whose proofs are deferred to the Supplementary Material:

Lemma 1. *If* $(\frac{1}{n}\sum_{i=1}^n \|x_i - \xi^*_{z_i}\|^p)^{1/p} \le \varepsilon$*, 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 $\delta > 2\gamma + 4\varepsilon$ valid. Finally, the result follows from Lemma 3. □

Proof of Theorem 2. The argument is similar to one that has appeared in recent literature [12, 9]. 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 1. 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 $\mathcal{T} = \{\tau_1, \ldots, \tau_{L'}\}$ 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 (6) 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 (7) 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 (7) is satisfied and we have $\mathrm{Miss}(z, \widehat{z}) \le K(c_2/c_1)^2/n$. Hence, as long as $c_1 > \sqrt{K}c_2$, we will have $\mathrm{Miss}(z, \widehat{z}) < 1/n$ 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 $c_1 = \sqrt{K}c_2 + \rho$ satisfy 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. □

References

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

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

[3] 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.

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

[5] 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.

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

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

[8] 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.

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

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

[11] 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.

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

[13] 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.

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

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

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

[17] James MacQueen et al. 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.

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

[19] 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.

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

[21] 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.

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

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

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

[25] 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.

[26] 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]