LIKELIHOOD ADJUSTED SEMIDEFINITE PROGRAMS FOR CLUSTERING HETEROGENEOUS DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Clustering is a widely deployed unsupervised learning tool. Model-based clustering is a flexible framework to tackle data heterogeneity when the clusters have different shapes. Likelihood-based inference for mixture distributions often involves non-convex and high-dimensional objective functions, imposing difficult computational and statistical challenges. The classic expectation-maximization (EM) algorithm is a computationally thrifty iterative method that maximizes a surrogate function minorizing the log-likelihood of observed data in each iteration, which however suffers from bad local maxima even in the special case of the standard Gaussian mixture model with common isotropic covariance matrices. On the other hand, recent studies reveal that the unique global solution of a semidefinite programming (SDP) relaxed K-means achieves the information-theoretically sharp threshold for perfectly recovering the cluster labels under the standard Gaussian mixture model. In this paper, we extend the SDP approach to a general setting by integrating cluster labels as model parameters and propose an iterative likelihood adjusted SDP (iLA-SDP) method that directly maximizes the *exact* observed likelihood in the presence of data heterogeneity. By lifting the cluster assignment to group-specific membership matrices, iLA-SDP avoids centroids estimation – a key feature that allows exact recovery under well-separateness of centroids without being trapped by their adversarial configurations. Thus iLA-SDP is less sensitive than EM to initialization and more stable on high-dimensional data. Our numeric experiments demonstrate that iLA-SDP can achieve lower mis-clustering errors over several widely used clustering methods including K-means, SDP and EM algorithms.

1 INTRODUCTION

Clustering analysis has been widely studied and regularly used in machine learning and its applications in network science (Girvan & Newman, 2002), computer vision (Shi & Malik, 2000; Joulin et al., 2010), manifold learning (Chen & Yang, 2021a) and bioinformatics (Karim et al., 2020). Perhaps by far the most popular clustering method is the K-means (MacQueen, 1967) partially because there are computationally convenient algorithms such as Lloyd's algorithm and K-means++ for heuristic approximation (Lloyd, 1982; Arthur & Vassilvitskii, 2007). Mathematically, K-means aims to find the optimal partition of data to minimize the total within-cluster squared Euclidean distances, which is equivalent to the maximum profile likelihood estimator under the standard Gaussian mixture model (GMM) with common isotropic covariance matrices (Chen & Yang, 2021b). Nevertheless, real data usually exhibit various degrees of heterogeneous features such as the cluster shapes may vary from component to component, which renders K-means as a sub-optimal clustering method.

Another popular clustering method is the classic expectation-maximization (EM) algorithm, which is a computationally thrifty method based on the idea of data augmentation to iteratively optimize the non-convex observed data likelihood (Dempster et al., 1977). Theoretical investigations reveal that the EM algorithm suffers from bad local maxima even in the one-dimensional standard GMM with well-separated cluster centers (Jin et al., 2016). Thus practically even when applied in highly favorable separation-to-noise ratio settings, careful initialization, often through multiple random initializations or a warm-start by another heuristic method such as hierarchical clustering (Fraley & Raftery, 2002), is the key for the EM algorithm to find the correct cluster labels and model parameters. With a reasonable initial start, the EM algorithm has been shown to achieve good statistical properties (Balakrishnan et al., 2017; Wu & Zhou, 2019).

In this paper, we consider the likelihood-based inference to tackle the problem of recovering cluster labels in the presence of data heterogeneity. Our motivation stems from the recent progress in understanding the computational and statistical limits for convex relaxation methods of the K-means clustering. Since K-means is a worst-case NP-hard problem (Aloise et al., 2009), various heuristic approximation algorithms such as Lloyd's algorithm (Lloyd, 1982; Lu & Zhou, 2016), and computationally tractable relaxations such as spectral clustering (Meila & Shi, 2001; Ng et al., 2001; Vempala & Wang, 2004; Achlioptas & McSherry, 2005; von Luxburg, 2007; von Luxburg et al., 2008) and semidefinite programs (SDP) (Peng & Wei, 2007; Mixon et al., 2016; Li et al., 2017; Fei & Chen, 2018; Chen & Yang, 2021a; Royer, 2017; Giraud & Verzelen, 2018; Bunea et al., 2016; Zhuang et al., 2022a), have been proposed in literature. Among the existing solutions, the SDP approach is particularly attractive in that it attains information-theoretically optimal threshold on centroid separations for exact recovery of cluster labels (Chen & Yang, 2021b).

Our contributions. We extend the SDP approach to a general setting with heterogeneous features by integrating cluster labels as model parameters (together with other component-specific parameters) and propose an *iterative likelihood adjusted SDP* (iLA-SDP) method that directly maximizes the *exact* observed data likelihood. Our idea is to tailor the strength of SDP relaxation of the K-means clustering method in the isotropic covariance case for likelihood-awareness inference. On one hand, iLA-SDP has a similar flavor as the EM algorithm by maximizing the likelihood function of the observed data. On the other hand, different from the EM framework, iLA-SDP treats the cluster labels as *unknown* parameters while profiles out the cluster centers (i.e., centroids), which brings several statistical and algorithmic advantages.

First in the arguably simplest one-dimensional GMM setting, EM is known to fail in certain configurations of centroids even when they are well-separated (Jin et al., 2016). In other words, EM is sensitive to initialization and model configuration. The main reason is due to the effort for estimating the cluster centers during the EM iterations. In iLA-SDP, cluster centers are regarded as nuisance parameters and profiled out to obtain a likelihood function in component-specific parameters including only the cluster covariance matrices. Thus iLA-SDP is more stable and performs empirically better than EM.

Second, cluster labels in EM are latent variables that are estimated by their posterior probabilities and the observed log-likelihood for component parameters and mixing weights are optimized through minorizing functions during iterations. In iLA-SDP, cluster labels are regarded as parameters optimized through the likelihood function jointly in the labels and covariance matrices. Thus iLA-SDP is a more direct approach than EM for taming the non-convexity in the observed log-likelihood objective and we prove that it perfectly recovers the true clustering structure if the clusters are well-separated under a lower bound without concerning the configurations of centroids.

The rest of the paper is organized as follows. In Section 2, we review some background on partition-based formulation for model-based clustering. In Section 3, we introduce the likelihood adjusted SDP for recovering the true partition structure and discuss its connection to the EM algorithm. In Section 4, we compare the performance of several widely used clustering methods on two real datasets.

2 MODEL-BASED CLUSTERING: A PARTITION FORMULATION

We consider the model-based clustering problem. Suppose the data points $X_1, \ldots, X_n \in \mathbb{R}^p$ are independent random variables sampled from K-component Gaussian mixture model (GMM). Specifically, let $G_1^*, \ldots, G_K^*$ be the true partition of the index set $[n] := \{1, \ldots, n\}$ such that if $i \in G_k^*$, then

$$X_i = \mu_k + \epsilon_i, \tag{1}$$

where $\mu_k \in \mathbb{R}^p$ is the center of the k-th cluster and ϵ_i is an i.i.d. random noise term following the common distribution $N(0, \Sigma_k)$. Here we focus on the most general and realistic scenario where the within-cluster covariance matrices $\Sigma_1, \ldots, \Sigma_K$ are heterogeneous. In our formulation of the GMM, the true partition $(G_k^*)_{k=1}^K$ is treated as a *unknown* parameter in model (1), along with the component-wise parameters $(\mu_k, \Sigma_k)_{k=1}^K$. With this parameterization $(G_k, \mu_k, \Sigma_k)_{k=1}^K$, the log-likelihood function for observing the data $\mathbf{X} = \{X_1, \ldots, X_n\}$ is given by

$$\ell\big((G_k, \mu_k, \Sigma_k)_{k=1}^K \mid \mathbf{X}\big) = -\sum_{k=1}^K \frac{|G_k|}{2} \log(2\pi|\Sigma_k|) - \frac{1}{2}\sum_{k=1}^K \sum_{i \in G_k} (X_i - \mu_k)^T \Sigma_k^{-1} (X_i - \mu_k),$$

where $|G_k|$ is the cardinality of G_k and $|\Sigma_k|$ is the determinant of matrix Σ_k. Since we are primarily interested in recovering the clustering labels (or equivalently the assignment matrix, cf. Section 3.1 below) in the presence of cluster heterogeneity, we can first profile out the nuisance parameters μ_k in closed form and the resulting objective function as a profile log-likelihood for the remaining parameters (after dropping constants) is given by

$$\ell\big((G_k, \Sigma_k)_{k=1}^K \mid \mathbf{X}\big) = -\sum_{k=1}^K |G_k| \log(|\Sigma_k|) - \sum_{k=1}^K \sum_{i \in G_k} \|X_i\|^2_{\Sigma_k^{-1}} + \sum_{k=1}^K \frac{1}{|G_k|} \sum_{i,j \in G_k} \langle X_i, X_j \rangle_{\Sigma_k^{-1}}, \tag{2}$$

where $\langle v, u \rangle_\Sigma := v^T \Sigma u$ and $\|u\|^2_\Sigma := \langle u, u \rangle_\Sigma$ for any $u, v \in \mathbb{R}^p$ and $\Sigma \succ 0$. This leads us to a combinatorial optimization problem for the profile log-likelihood function

$$\max \Big\{ \ell\big((G_k, \Sigma_k)_{k=1}^K \mid \mathbf{X}\big) : \bigsqcup_{k=1}^K G_k = [n],\ \Sigma_k \succ 0 \Big\}, \tag{3}$$

where the disjoint union $\bigsqcup_{k=1}^K G_k = [n]$ means that $\bigcup_{k=1}^K G_k = [n]$ and $G_j \cap G_k = \emptyset$ if $j \neq k$. Note that the constrained optimization problem in (3) in the special case $\Sigma_1 = \cdots = \Sigma_k = \sigma^2 \mathrm{Id}_p$ reduces to the K-means clustering method, which is known to be worst-case NP-hard (Dasgupta, 2007; Mahajan et al., 2009). To overcome such computational difficulty, semidefinite program (SDP) relaxation is a tractable solution that achieves information-theoretically optimal exact recovery under the standard GMM with identical and isotropic covariance matrices (Chen & Yang, 2021a). Nevertheless, all existing formulations of various SDP relaxations of the standard GMM critically depend on the assumption that $\Sigma_1 = \cdots = \Sigma_k = \sigma^2 \mathrm{Id}_p$ with a *known* noise variance parameter σ^2 (Fei & Chen, 2018; Li et al., 2017; Peng & Wei, 2007; Chen & Yang, 2021a). This motivates us to seek alternative SDP formulations adjusting the (full) information coming from the likelihood function for the observed data $\mathbf{X}$.

3 Likelihood adjusted SDP for clustering heterogeneous data

In this section, we introduce the likelihood adjusted SDP (LA-SDP) for recovering the true partition structure $G_1^*, \ldots, G_K^*$ by applying convex relaxation to the profile log-likelihood function (3).

3.1 Oracle LA-SDP under known covariance matrices

In this subsection, we consider the oracle case where the covariance matrices $\Sigma_1, \ldots, \Sigma_K$ are known. Let us start with a well-studied SDP relaxation formulation (Peng & Wei, 2007) for approximating the combinatorial optimization problem of maximizing the profile log-likelihood function under the isotropic setting with known $\Sigma_1 = \ldots = \Sigma_K = \sigma^2 \mathrm{Id}_p$, which is known (Chen & Yang, 2021b) to attain the information-theoretically optimal threshold on centroid separations for exact recovery of cluster labels. Note that there is a one-to-one correspondence between any given partition $(G_k)_{k=1}^K$ of $[n]$ and a binary assignment matrix $H = (h_{ik}) \in \{0,1\}^{n \times K}$ (up to cluster labels permutation) such that $h_{ik} = 1$ if $i \in G_k$ and $h_{ik} = 0$ otherwise for $i \in [n]$ and $k \in [K]$. Because each row of H contains exactly one non-zero entry, the recovery of the true clustering structure (or its associated assignment matrix) by maximizing the profile log-likelihood function (after dropping constants) can be re-expressed as a (non-convex) mixed integer program:

$$\max_H \langle A, HBH^\top \rangle = \sum_{k=1}^K \frac{1}{|G_k|} \sum_{i,j \in G_k} \langle X_i, X_j \rangle, \quad \text{subject to } H \in \{0,1\}^{n \times K} \text{ and } H\mathbf{1}_K = \mathbf{1}_n, \tag{4}$$

where $A = X^\top X$ is the $n \times n$ similarity matrix, $\mathbf{1}_n$ denotes the n-dimensional vector of all ones, and B is the $K \times K$ diagonal matrix whose k-th diagonal component is $|G_k|^{-1} = \big(\sum_{i=1}^n h_{ik}\big)^{-1}$. Here, we have used the key identity $\sum_{k=1}^K w_k \sum_{i,j \in G_k} a_{ij} = \langle A, HBH^\top \rangle$ that holds for any diagonal matrix $B = \mathrm{diag}(w_1, \ldots, w_K)$ and similarity matrix $A = (a_{ij})_{i,j=1}^n$. Relaxing the above mixed integer program (4) by lifting the assignment matrix H into $Z = HBH^\top$, we arrive at its SDP relaxation as

$$\hat{Z} = \arg\max_{Z \in \mathbb{R}^{n \times n}} \langle A, Z \rangle, \quad \text{subject to } Z \succeq 0,\ \mathrm{tr}(Z) = K,\ Z\mathbf{1}_n = \mathbf{1}_n,\ Z \geqslant 0, \tag{5}$$

where $Z \geqslant 0$ means each entry $Z_{ij} \geqslant 0$ and $Z \succeq 0$ means the matrix Z is symmetric and positive semi-definite. This SDP formulation relaxes the integer constraint on H into two linear constraints $\mathrm{tr}(Z) = K$ and $Z \geqslant 0$ that are satisfied by any $Z = HBH^T$ as H ranges over feasible solutions of problem (4).

Now let us consider the general heterogeneous setting with (possibly) different and non-isotropic covariance matrices $\Sigma_1, \ldots, \Sigma_K$, and extend the SDP relaxation to this setting. Two technical difficulties arise by examining the previous argument. First, the first two terms in the profile log-likelihood function (3) are no longer independent of the assignment matrix, and is therefore not negligible. In particular, they also provide partial information about the cluster labels when the covariance matrices are different: $\|X_i\|^2_{\Sigma_k^{-1}}$ in the second term quantifies how well X_i aligns with the covariance matrix Σ_k encoding second-order information of the k-th cluster; while the first term plays the role of balancing the cluster sizes and favors assigning more points to clusters with smaller shapes (since density is expected to be high). Second, the similarity $\langle X_i, X_j \rangle_{\Sigma_k^{-1}}$ within cluster G_k in the third term now depends on k, making the key identity $\sum_{k=1}^K w_k \sum_{i,j \in G_k} a_{ij} = \langle A, HBH^\top \rangle$ for connecting the profile log-likelihood function with the objective function of the mixed integer program (4) no longer applicable.

To solve the two aforementioned difficulties, we propose to augment the single variable Z in the SDP relaxation (5) to K variables $(Z_k)_{k=1}^K$, where Z_k can be interpreted as the lifting of the k-th column H_k of the assignment matrix H via $Z_k = \frac{1}{|G_k|} H_k H_k^\top$, $|G_k| = \sum_{i=1}^n h_{ik} = H_k^\top \mathbf{1}_n$, that encodes the cluster membership associated with the k-th cluster. More specifically, by extending the key identity in the isotropic setting to $\sum_{k=1}^K w_k \sum_{i,j \in G_k} a_{ij}^{(k)} = \sum_{k=1}^K \langle A^{(k)}, H_k w_k H_k^\top \rangle$ for any weight vector $w = (w_k)_{k=1}^K$ and K similarity matrices $\big(A_k = (a_{ij}^{(k)})_{i,j=1}^n\big)_{k=1}^K$, we can analogously express the maximizing profile log-likelihood problem as the following (non-convex) mixed integer program:

$$\max_H \sum_{k=1}^K \langle A^{(k)}, H_k w_k H_k^\top \rangle, \quad \text{subject to } H_k \in \{0,1\}^{n \times 1} \text{ and } \sum_{k=1}^K H_k = \mathbf{1}_n, \tag{6}$$

where $w_k = |G_k|^{-1} = \big(\sum_{i=1}^n h_{ik}\big)^{-1}$, and the k-th cluster-specific similarity matrix $A^{(k)}$ is

$$A^{(k)} := -\log(|\Sigma_k|)\mathbf{1}_n\mathbf{1}_n^T - \frac{1}{2}\left[\mathrm{diag}(X^T\Sigma_k^{-1}X)\mathbf{1}_n^T + \mathbf{1}_n\mathrm{diag}(X^T\Sigma_k^{-1}X)^T\right] + X^T\Sigma_k^{-1}X. \tag{7}$$

Here, $\mathrm{diag}(A)$ stands for the column vector composed of all diagonal entries of a matrix A. Now by lifting H_k into $Z_k = H_k w_k H_k^\top$, we arrive at the following SDP relaxation for the profile log-likelihood objective function (2):

$$\begin{aligned} \big(\hat{Z}_1, \ldots, \hat{Z}_K\big) &= \underset{Z_1, \ldots Z_K \in \mathbb{R}^{n \times n}}{\arg\max} \sum_{k=1}^K \langle A_k, Z_k \rangle, \\ \text{subject to } Z_k \succeq 0,\ \sum_{k=1}^K \mathrm{tr}(Z_k) &= K,\ \big(\sum_{k=1}^K Z_k\big)\mathbf{1}_n = \mathbf{1}_n,\ Z_k \geqslant 0,\ \forall\, k \in [K], \end{aligned} \tag{8}$$

which relaxes the integer constraint on $H = (H_1, H_2, \cdots, H_k)$ into $(K+1)$ linear constraints $\sum_{k=1}^K \mathrm{tr}(Z_k) = K$ and $Z_k \geqslant 0$ for $k \in [K]$ that are satisfied by any $Z_k = H_k w_k H_k^\top$ as H ranges over feasible solutions of problem (6).

Since solving (8) requires the knowledge of the true covariance matrix for each component, we call the solution $(\hat{Z}_k)_{k=1}^K$ as the *oracle* likelihood adjusted SDP (LA-SDP) for estimating the cluster membership matrix of data points. In the special case of isotropic covariance matrices $\Sigma_1 = \cdots = \Sigma_K = \sigma^2 \mathrm{Id}_p$, Proposition 1 below shows that LA-SDP reduces to become equivalent to the previous SDP formulation (5).

Proposition 1 (**SDP relaxation for K-means is a special case of LA-SDP**). Suppose $\Sigma_k = \sigma^2 \mathrm{Id}_p$ for all $k \in [K]$. Let $\hat{Z}$ be the solution to (5) that achieves maximum M_1 and $\hat{Z}_k, k = 1, \ldots, K$, be the solution to (5) with maximum M_2. Then $M_1 = M_2$. And $\hat{Z} = \sum_{k=1}^K \hat{Z}_k$, if $\hat{Z}$ is unique in (5).

Note that the SDP relaxed K-means in (5) is originally proposed in (Peng & Wei, 2007) and has been extensively studied in literature. In particular, it achieves the information-theoretical limit for exact recovery under the standard GMM (Chen & Yang, 2021a) and it is robust against outliers and adversarial attack (Fei & Chen, 2018). In the case of exact recovery where $\hat{Z} = Z^*$ and Z^* is the true cluster membership matrix such that $Z^*_{ij} = |G^*_k|^{-1}$ if $i, j \in G^*_k$ and $Z^*_{ij} = 0$ otherwise, then we can easily recover the true partition structure $G^*_1, \ldots, G^*_K$ or its associated assignment matrix from the block diagonal matrix $\hat{Z}$. Thus it is an interesting theoretical question of when the partition structure induced by $\hat{Z} = \sum_{k=1}^K \hat{Z}_k$ from the LA-SDP (see Figure 1 for an illustration) can achieve exact recovery. Theorem 2 below gives a lower bound of the separation signal-to-noise ratio for achieving exact recovery in the presence of data heterogeneity.

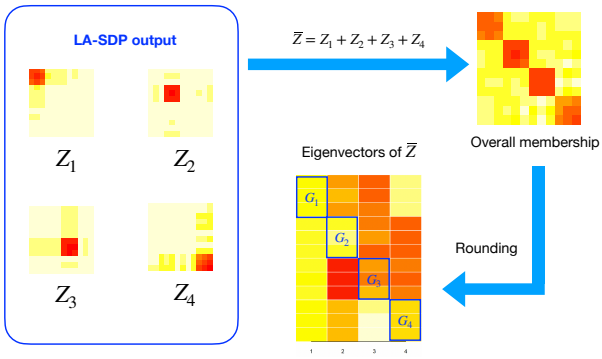

Figure 1: LA-SDP membership matrices to cluster labels via spectral rounding.

For each distinct pair $(k, l) \in [K]$, let $D_{(k,l)} := \frac{\sum_{i=1}^p (\lambda_i - \log(1+\lambda_i))}{p \max_i |\lambda_i|}$ characterize the closeness between Σ_k and Σ_l, where $\lambda_1, \ldots, \lambda_p$ enumerate all eigenvalues of $(\Sigma_l^{1/2}\Sigma_k^{-1}\Sigma_l^{1/2} - \mathrm{Id}_p)$. Let $\Delta^2 := \min_{k \neq l} \|\Sigma_k^{-1/2}(\mu_k - \mu_l)\|^2$ denote a covariance adjusted centroid separation, $n_k := |G^*_k|$ the size of true cluster G^*_k, $m = \min_{k \neq l} \frac{2 n_k n_l}{n_k + n_l}$ the least pairwise harmonic mean over cluster sizes, $\underline{n} = \min_k n_k$ the minimal cluster size, and $M := \max_{k \neq l} \|\Sigma_l^{1/2}\Sigma_k^{-1}\Sigma_l^{1/2}\|_{\mathrm{op}}$ (matrix operator norm).

Theorem 2 (**Exact recovery for LA-SDP**). Suppose there exist constants $\delta > 0$, $\beta \in (0, 1)$ and $\eta \in (0, 1)$ such that

$$\log n \geq \max\left\{\frac{(1-\beta)^2}{\beta^2}, \frac{(1-\beta)(1-\eta)K^2}{\beta^2 \max\{(M-1)^2, 1\}}\right\}\frac{C_1 n}{m}, \ \delta \leq \frac{\beta^2}{(1-\beta)^2}\frac{C_2 M^{1/2}}{K}, \ m \geq \frac{4(1+\delta)^2}{\delta^2}.$$

Then the LA-SDP achieves exact recovery, or $\hat{Z} = Z^*$, with probability at least $1 - C_7 K^3 n^{-\delta}$ if

$$\Delta^2 \geq (E_1 + E_2)\log n, \text{ and } \min_{k \neq l} D_{(k,l)} \geq C_3(1 + \log n/p + p/n), \tag{9}$$

where concrete expressions of E_1 and E_2 (depending on δ, β, η) are provided in Appendix A.3, and $C_1, \ldots, C_7$ are universal constants.

Our definition of the centroid separation Δ extends the separation-to-noise ratio (SNR) for the exact recovery under the isotropic covariance setting (Chen & Yang, 2021a) to the heterogeneous setting by taking into account the cluster shapes (i.e. second order information). From (18), we see that our theoretical centroid separation lower bound consists of two parts E_1 and E_2: E_1 reduced to the information-theoretically optimal threshold when $M = 1$, corresponding to same covariance matrices; E_2 tends to vanish for small M close to one and satisfying $M = 1 + o(1/\sqrt{n \log n})$ or remains as an extra term for large M. From our numerical results summarized in Figure 2, we can observe that our defined centroid separation Δ indeed captures the accuracy of cluster label recovery using LA-SDP—the mis-clustering error curves display almost identical patterns under different settings of the GMM. In comparison, the performance of the (original) SDP (5) and the K-means clustering method designed for the isotropic case become significantly worse as the condition number of the cluster covariance matrices increases. More details about implementation and model setups are provided in Appendix A.2.

3.2 Iterative LA-SDP under unknown covariance matrices: an alternating maximization algorithm

Since the oracle LA-SDP relies on the knowledge of covariance matrices $\Sigma_1, \ldots, \Sigma_K$, we propose a simple and practical data-driven algorithm for approximating LA-SDP when these covariance

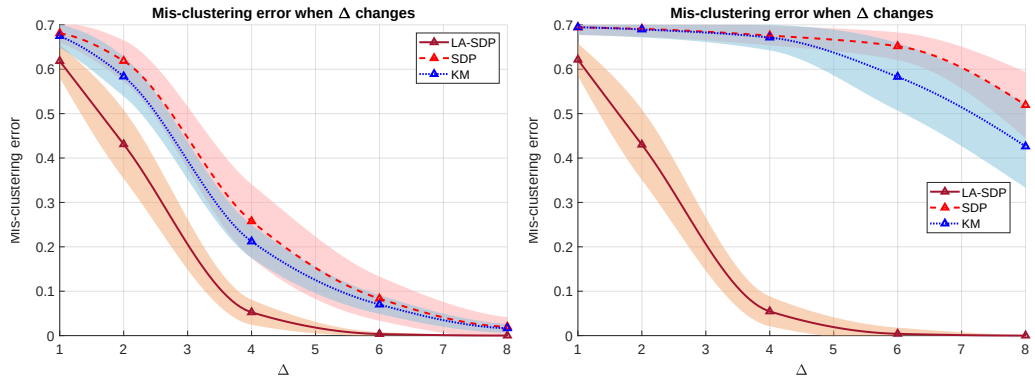

Figure 2: Mis-clustering error (with shaded error bars) vs centroid separation Δ under different conditional numbers of cluster covariance matrices $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_K$ ($M = 1$). The left (right) plot corresponds to a moderate (large) condition number of the common covariance matrix. Here, KM refers to K-means method; SDP refers to the original SDP (5).

matrices are unknown. The idea is to alternate between the SDP relaxation given a current estimate of the component covariance matrices and updating covariance matrices according to the maximum (penalized) likelihood given the new membership estimate. The next lemma gives a closed-form formula for updating covariance matrices given a current estimate of the assignments $Z_1, \ldots, Z_K$ based on their (unconstrained) MLEs on the observed data.

Lemma 3 (**Updating formula for covariance matrices under alternating maximization**). For any feasible matrices $Z_1, \ldots, Z_K$ satisfying the constraints of (8),

$$\hat{\Sigma}_k := \frac{1}{\mathbf{1}_n^T Z_k \mathbf{1}_n} \sum_{i,j=1}^{n} \left[\frac{1}{2}(X_i X_i^T + X_j X_j^T) - X_i X_j^T \right] Z_{k,ij}, \quad k \in [K], \tag{10}$$

solve the following optimization problem

$$\hat{\Sigma}_1, \ldots, \hat{\Sigma}_K = \arg \max_{\Sigma_1, \ldots \Sigma_K \succeq 0} \sum_{k=1}^{K} \langle A_k, Z_k \rangle, \tag{11}$$

where recall that $A^{(k)} := A^{(k)}(\Sigma_k)$ is the Σ_k-dependent similarity matrix defined in (7).

Based on the lemma, we propose an *iterative LA-SDP* (iLA-SDP) by alternating maximization of the profile log-likelihood (3) for estimating the lifted cluster membership matrices $(Z_k)_{k=1}^K$ from LA-SDP (8) and the component covariance matrices $(\Sigma_k)_{k=1}^K$, as summarized in Algorithm 1. In the special case where the lifted membership matrix Z_k is of rank one, which holds for true lifted cluster membership matrices $(Z_k^*)_{k=1}^K$, the covariance matrices produced by iLA-SDP can be interpreted as within-cluster sample covariance matrices under soft clustering.

Proposition 4 (**Covariance estimation in iLA-SDP via soft clustering**). If $\text{rank}(Z_k) = 1$, then there exists weights $(w_{k,1}, \ldots, w_{k,n})$ such that these $\hat{\Sigma}_k$ in Lemma 3 can be written as

$$\hat{\Sigma}_k := \frac{1}{n_k} \sum_{i=1}^{n} w_{k,i} (X_i - \hat{\mu}_k)(X_i - \hat{\mu}_k)^\top, \text{ where } \hat{\mu}_k := \frac{1}{n_k} \sum_{i=1}^{n} w_{k,i} X_i \text{ and } n_k = \sum_{i=1}^{n} w_{k,i}. \tag{12}$$

It is further noted from the proof of Proposition 4 that when Z_k has rank one, the weights $w_{k,1}, \ldots, w_{k,n}$ are proportional to the leading non-zero eigenvector of Z_k. Thus the alternating maximization step (17) for updating the covariance matrices in iLA-SDP can be interpreted as a soft clustering technique that resembles the EM algorithm. Specifically, the E-step estimates the (hard) cluster label $Y_i \in \{0, 1\}^K$ associated with X_i by the posterior probabilities $\tau_{ik} := p(Y_{ik} \mid X_i, \hat{\theta}^{(t)})$ where $\hat{\theta}^{(t)} = (\hat{\pi}_k^{(t)}, \hat{\mu}_k^{(t)}, \hat{\Sigma}_k^{(t)})_{k=1}^K$ denotes the estimated GMM parameters at the t-th iteration in the EM. Then the M-step updates the parameters via $\hat{\pi}_k^{(t+1)} = m_k^{(t)}/n$ with $m_k^{(t)} = \sum_{i=1}^n \tau_{ik}^{(t)}$,

$$\hat{\mu}_k^{(t+1)} = \frac{1}{m_k^{(t)}} \sum_{i=1}^{n} \tau_{ik}^{(t)} X_i \text{ and } \hat{\Sigma}_k^{(t+1)} = \frac{1}{m_k^{(t)}} \sum_{i=1}^{n} \tau_{ik}^{(t)} (X_i - \hat{\mu}_k^{(t+1)})(X_i - \hat{\mu}_k^{(t+1)})^\top. \tag{13}$$

Note that (17) and (13) represent different weighting schemes in the soft clustering rule for obtaining an estimate for the cluster labels. In iLA-SDP, the weight $w_{k,i}$ for X_i belonging to component k is determined by the SDP in (8). Once the weights are calculated, remaining parameter updates in both iLA-SDP and EM boil down to simple averages with effective component sample sizes n_k and m_k, respectively. In Section 3.3 to follow, we provide deeper comparison between iLA-SDP and EM.

Remark 5. In Appendix A.1, we further propose two variations of iLA-SDP that can handle high-dimensional and large-size data with better computational and statistical efficiency. For high-dimensional data, we apply Fisher's LDA with an initial estimate of the cluster labels to find an optimal feature subspace that increases the SNR for better clustering, and for large-size data we combine the subsampling idea with iLA-SDP to reduce computational cost (Zhuang et al., 2022b).

Algorithm 1: The iterative likelihood adjusted SDP (iLA-SDP) algorithm

Input: Data matrix $X \in \mathbb{R}^{p\times n}$ containing n points. Initialization of assignments $G_1^{(0)}, \ldots, G_K^{(0)}$ or covariance matrices $\Sigma_1^{(0)}, \ldots, \Sigma_K^{(0)}$. The stopping criterion parameters ϵ, S.

(Assignments to covariance matrices) If we have the initialization of assignments, let $\Sigma_k^{(0)} := |G_k^{(0)}|^{-1} \sum_{i\in G_k^{(0)}} (X_i - \bar{X}_k)(X_i - \bar{X}_k)^T$ to be the sample covariance of each cluster $k \in [K]$, where $\bar{X}_k := |G_k^{(0)}|^{-1} \sum_{i\in G_k^{(0)}} X_i$.

for $s = 1, \ldots, S$ **do**

- (Adjusted-SDP) Solve the Adjusted-SDP in (8) using X and $\Sigma_1^{(s-1)}, \ldots, \Sigma_K^{(s-1)}$ to get solution $Z_1^{(s)}, \ldots, Z_K^{(s)}$.
- Compute the sum $\bar{Z}^{(s)} := \sum_{k=1}^K Z_k^{(s)}$ and the relative norm $r^{(s)} := \|\bar{Z}^{(s)} - \bar{Z}^{(s-1)}\|_F / \|\bar{Z}^{(s-1)}\|_F$ for $s \geq 2$. We will break the loop if $r^{(s)} < \epsilon$.
- (Assignments to covariance matrices) Use formula in Lemma 3 to get covariance matrices $\Sigma_1^{(s)}, \ldots, \Sigma_K^{(s)}$ from $Z_1^{(s)}, \ldots, Z_K^{(s)}$.

Perform the spectral decomposition of $\bar{Z}^{(S)}$ and take the top K eigenvectors $(\hat{u}_1, \ldots, \hat{u}_K)$.

Run K-means clustering on $(\hat{u}_1, \ldots, \hat{u}_K)$ and extract the cluster labels $\hat{G}_1, \ldots, \hat{G}_K$ as a partition estimate for $[n]$.

Output: A partition estimate $\hat{G}_1, \ldots, \hat{G}_K$ for $[n]$.

3.3 Connections between iLA-SDP and EM algorithms

It is interesting to observe that our proposed iLA-SDP algorithm is closely connected to the classic EM algorithm, which approximates the maximum likelihood estimation (MLE) of the observed data in statistical models with latent variables (Dempster et al., 1977). The key idea of EM algorithm in the model-based clustering context is *data augmentation* where the latent variables represent the cluster labels. More specifically, for each data point $X_i \in \mathbb{R}^p$, we associate with an unobserved one-hot encoded cluster label $Y_i := \{Y_{i1}, \ldots, Y_{iK}\} \in \{0,1\}^K$. Then the EM algorithm aims to iteratively maximize the *expected log-likelihood of the complete data* $(X_i, Y_i)_{i=1}^n$ given by

$$\theta^{(t+1)} = \arg\max_{\theta} \Big\{ Q(\theta \mid \theta^{(t)}) := \mathbb{E}_{\mathbf{Y}\sim q(\cdot\mid \mathbf{X}, \theta^{(t)})}[\ell_c(\theta \mid \mathbf{X}, \mathbf{Y})] \Big\}, \tag{14}$$

where $\theta = ((\pi_k, \mu_k, \Sigma_k)_{k=1}^K)$ contains parameters in the GMM, $(\pi_k)_{k=1}^K$ are the weight parameters such that $\pi_k \geqslant 0$ and $\sum_{k=1}^K \pi_k = 1$, and the complete log-likelihood function is

$$\ell_c(\theta \mid \mathbf{X}, \mathbf{Y}) := p(\mathbf{X}, \mathbf{Y} \mid \theta) = -\frac{1}{2} \sum_{i=1}^n \sum_{k=1}^K Y_{ik} \left[\log(2\pi|\Sigma_k|) - (X_i - \mu_k)^\top \Sigma_k^{-1} (X_i - \mu_k) \right].$$

Alternatively, the EM algorithm (14) can be interpreted as minorize-maximization (MM) that maximizes a best lower bound for the log-likelihood of the observed data

$$\ell(\theta \mid \mathbf{X}) := \log p(\mathbf{X} \mid \theta) = \log \sum_{\mathbf{Y}} p(\mathbf{X}, \mathbf{Y} \mid \theta) \geqslant \sum_{\mathbf{Y}} q(\mathbf{Y} \mid \mathbf{X}) \log \frac{p(\mathbf{X}, \mathbf{Y} \mid \theta)}{q(\mathbf{Y} \mid \mathbf{X})} =: \mathcal{L}(q, \theta)$$

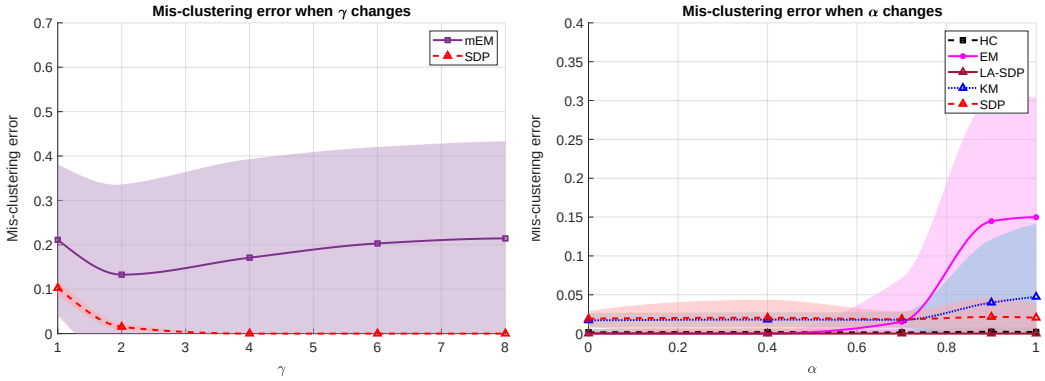

Figure 3: Mis-clustering error (with shaded error bars) vs γ (captures the signal strength of GMM) and α (perturbation percentage of initialization). mEM (SDP) refers to the reduced version of EM (LA-SDP) where we consider covariance matrices as fixed and equal to identity. The first plot compares the performance of mEM and SDP when separation is large with random initialization; the second plot compares all methods when we enlarge the perturbation percentage α applied to the random initialization from hierarchical clustering (HC).

for any posterior distribution $q(\mathbf{Y} \,|\, \mathbf{X})$. Under this perspective, the EM algorithm can be expressed as an *alternating maximization* algorithm on $\mathcal{L}(q, \theta)$ between E-step $q^{(t+1)} = \arg\max_q \mathcal{L}(q, \theta^{(t)})$ and M-step $\theta^{(t+1)} = \arg\max_\theta \mathcal{L}(q^{(t+1)}, \theta)$. Thus, give any $q(\mathbf{Y} \,|\, \mathbf{X})$, the M-step maximizes the expected complete log-likelihood as a surrogate function that minorizes $\ell(\theta \,|\, \mathbf{X})$ because $\mathcal{L}(q, \theta) = \mathbb{E}_{\mathbf{Y} \sim q(\cdot \,|\, \mathbf{X})}[\ell_c(\theta \,|\, \mathbf{X}, \mathbf{Y})] - H(q(\mathbf{Y} \,|\, \mathbf{X}))$ where $H(q)$ denotes the relative entropy of distribution q, while given the current parameter estimate $\theta^{(t)}$, the E-step is maximized at $q^{(t+1)}(\mathbf{Y} \,|\, \mathbf{X}) = p(\mathbf{Y} \,|\, \mathbf{X}, \theta^{(t)})$ because

$$\ell(\theta^{(t)} \mid \mathbf{X}) \geqslant \mathcal{L}(p(\mathbf{Y} \mid \mathbf{X}, \theta^{(t)}), \theta^{(t)}) = \sum_{\mathbf{Y}} p(\mathbf{Y} \mid \mathbf{X}, \theta^{(t)}) \log p(\mathbf{X} \mid \theta^{(t)}) = \ell(\theta^{(t)} \mid \mathbf{X}),$$

where the first inequality is actually an equality at $p(\mathbf{Y} \,|\, \mathbf{X}, \theta^{(t)})$. Even though the EM and iLA-SDP are both alternating maximization algorithms aiming to solve the MLE for the observed data log-likelihood and both can be viewed as soft clustering methods (cf. Proposition 4), there are several important differences we would like to highlight.

First, cluster labels are (random) latent variables and they are estimated via posterior probabilities in the EM algorithm, while the labels are treated as unknown parameters in iLA-SDP that are estimated via direct maximization of the observed data likelihood.

Second, the EM algorithm is a special case of the minorization-maximization (MM) algorithm (Hunter & Lange, 2000) by iteratively performing the coordinate ascent on the expected complete data log-likelihood as a minorizing surrogate function, while our iLA-SDP is *exact* in the sense that it directly optimizes the observed data log-likelihood via a convex relaxation formulation. Thus iLA-SDP is a more direct approach than EM for tackling the non-convex observed log-likelihood objective and it is principled to perfectly recover the true clustering structure if the clusters are well-separated under an SNR lower bound in Theorem 2. As in the EM algorithm, iLA-SDP monotonically maximizes the observed data log-likelihood over iterations; cf. Figure 6 in Appendix.

Third, the EM algorithm in each iteration must estimate the cluster center parameters $(\mu_k)_{k=1}^K$, while our iLA-SDP profiles out the effect of centroid estimation and leverages only pairwise Mahalanobis distances between data to accommodate the heterogeneity of cluster shapes. Partly because the error in estimating the centroids propagates to other parameters, EM is more sensitive to initialization with inaccurate labels and the centroid configurations even in the standard GMM (Jin et al., 2016), and iLA-SDP behaves better than EM, an observation we empirically verify in our simulation experiments; cf. Figure 3 for comparison between iLA-SPD and EM algorithms. From the first plot we can observe that LA-SDP with isotropic known covariance matrices, which reduces to the K-means SDP in (5), performs stable and achieves exact recovery when the separation is large. However, EM fails with random initialization in this adversarial centroids configuration. Moreover, from the second figure we can see that LA-SDP is fairly stable with perturbation of initialization if the separation is large while EM can go worse as the perturbation percentage of initialization α approaches 1, i.e., all the labels

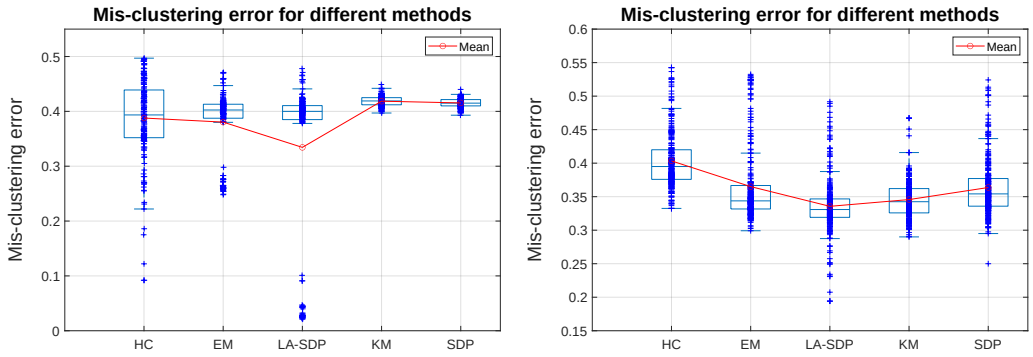

Figure 4: Box plots of mis-clustering error (with means) with different methods. The left (right) plot summarizes the results for the banknote authentication dataset (landsat satellite dataset).

are selected randomly. In other words, EM is more sensitive to initialization and LA-SDP is more stable if the signal is strong. More details of the settings in Figure 3 can be found in Appendix A.2.

4 REAL-DATA APPLICATIONS

In this section, we test the performance of iLA-SDP against several widely used clustering methods on two real datasets from the UCI machine learning repository.

Banknote authentication dataset. We first look at the performances of our methods for a banknote authentication dataset where the separation of clusters are not large. The images were taken from genuine and forged banknote-like specimens, where the features were extracted by Wavelet Transform tool. It contains 1372 samples and $p = 4$ attributes with $K = 2$ clusters. We choose total $n = 1000$ samples randomly and equally from two clusters to make cluster sizes balanced among total 200 replicates. HC is used as initialization for EM, KM and LA-SDP. If the initialization for the assignments $G_1^{(0)},\ G_2^{(0)}$ is highly unbalanced, i.e., $\|G_1^{(0)}\|/\|G_2^{(0)}\| > 4$ if $\|G_1^{(0)}\| > \|G_2^{(0)}\|$, then the covariance matrices of two clusters should differ significantly and we calculate the covariance estimation for unconstrained optimization problem; Otherwise we will calculate the estimation of the covariance matrices through graphical lasso with parameter $\lambda = 2$ since the similarity shows that we could reduce the estimation of parameters. Then we run Algorithm 1 to get the results for LA-SDP. The comparison of those four methods can be found from the left plot in Figure 4, where we can observe that the separation between two clusters is not well in the sense that the medians of all methods are similar. Nevertheless, LA-SDP can achieve better performance than EM especially when the subsets of data we drew are informative. The reason LA-SDP has lower mean is that sometimes LA-SDP can achieve nearly exact recovery in the sense that there are 36 out of 200 times when the mis-clustering error for LA-SDP are below 0.05, while EM can only hit around 0.25. This indicates that there are more chances for LA-SDP to sort two clusters with fairly good performance even the separation is problematic.

Landsat satellite dataset. This database was generated from landsat Multi-Spectral Scanner image data. The test set includes 2000 satellite images, 6 different clusters with 36 attributes ($36 = 4$ spectral bands $\times$ 9 pixels in neighbourhood). Every attribute is an integer from 0 to 255 indicating the color for certain pixel. We performed 4 methods on the transformed dataset with total 300 replicates. For each repetition, we draw total $n = 1200$ samples randomly and equally from clusters to make cluster sizes balanced. And for each attribute, we scale its range to $[0, 1]$ and then take the function $f(x) = \log(1/x - 1)$ entry-wise to transform the range to $\mathbb{R}_+$. Then, we run Algorithm 3 on the transformed dataset $\tilde{X}$ to get the results for LA-SDP with $\epsilon = 10^{-2},\ p_0 = K = 6$, and $R = 50$. From the results we can see that our method LA-SDP performs the best for all four methods. Especially, LA-SDP out-performs EM since the initialization (HC) is rough, which results in both biases of the estimations of group means and covariance matrices for EM while LA-SDP only uses the group covariance matrices as its initialization. Here we can see that KM algorithm performs relatively good since after performing LDA, the transformed data are normalized so that the "within-cluster" covariance matrices are isotropic corresponding to HC, which meets the assumptions for KM and SDP.

REFERENCES

Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In Peter Auer and Ron Meir (eds.), *Learning Theory*, pp. 458–469, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-31892-7.

Daniel Aloise, Amit Deshpande, Pierre Hansen, and Preyas Popat. Np-hardness of euclidean sum-of-squares clustering. *Machine learning*, 75(2):245–248, 2009.

David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pp. 1027–1035, USA, 2007. Society for Industrial and Applied Mathematics. ISBN 9780898716245.

Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77 – 120, 2017. doi: 10.1214/16-AOS1435. URL https://doi.org/10.1214/16-AOS1435.

Florentina Bunea, Christophe Giraud, Martin Royer, and Nicolas Verzelen. PECOK: a convex optimization approach to variable clustering. *arXiv:1606.05100*, 2016.

Xiaohui Chen and Yun Yang. Diffusion k-means clustering on manifolds: Provable exact recovery via semidefinite relaxations. *Applied and Computational Harmonic Analysis*, 52:303–347, 2021a. ISSN 1063-5203.

Xiaohui Chen and Yun Yang. Cutoff for exact recovery of gaussian mixture models. *IEEE Transactions on Information Theory*, 67(6):4223–4238, 2021b.

Sanjoy Dasgupta. The hardness of k-means clustering. *Technical Report CS2007-0890, University of California, San Diego*, 2007.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. doi: https://doi.org/10.1111/j.2517-6161.1977.tb01600.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1977.tb01600.x.

Yingjie Fei and Yudong Chen. Hidden integrality of sdp relaxation for sub-gaussian mixture models. *arXiv:1803.06510*, 2018.

Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002. doi: 10.1198/016214502760047131. URL https://doi.org/10.1198/016214502760047131.

Christophe Giraud and Nicolas Verzelen. Partial recovery bounds for clustering with the relaxed $k$means. *arXiv:1807.07547*, 2018.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002. doi: 10.1073/pnas.122653799. URL https://www.pnas.org/doi/abs/10.1073/pnas.122653799.

David R. Hunter and Kenneth Lange. Quantile regression via an mm algorithm. *Journal of Computational and Graphical Statistics*, 9(1):60–77, 2000. doi: 10.1080/10618600.2000.10474866.

Chi Jin, Yuchen Zhang, Sivaraman Balakrishnan, Martin J. Wainwright, and Michael I. Jordan. Local maxima in the likelihood of gaussian mixture models: Structural results and algorithmic consequences. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 4123–4131, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Armand Joulin, Francis Bach, and Jean Ponce. Discriminative clustering for image co-segmentation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1943–1950, June 2010. doi: 10.1109/CVPR.2010.5539868.

Md Rezaul Karim, Oya Beyan, Achille Zappa, Ivan G Costa, Dietrich Rebholz-Schuhmann, Michael Cochez, and Stefan Decker. Deep learning-based clustering approaches for bioinformatics. *Briefings in Bioinformatics*, 22(1):393–415, 02 2020. ISSN 1477-4054. doi: 10.1093/bib/bbz170. URL https://doi.org/10.1093/bib/bbz170.

Xiaodong Li, Yang Li, Shuyang Ling, Thomas Stohmer, and Ke Wei. When do birds of a feather flock together? k-means, proximity, and conic programming. *arXiv:1710.06008*, 2017.

Stuart Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28: 129–137, 1982.

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

J.B. MacQueen. Some methods for classification and analysis of multivariate observations. *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability*, pp. 281–297, 1967.

Meena Mahajan, Prajakta Nimbhorkar, and Kasturi Varadarajan. The planar k-means problem is np-hard. In Sandip Das and Ryuhei Uehara (eds.), *WALCOM: Algorithms and Computation*, pp. 274–285, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. ISBN 978-3-642-00202-1.

Marina Meila and Jianbo Shi. Learning segmentation by random walks. In *In Advances in Neural Information Processing Systems*, pp. 873–879. MIT Press, 2001.

Dustin G Mixon, Soledad Villar, and Rachel Ward. Clustering subgaussian mixtures by semidefinite programming. *arXiv:1602.06612v2*, 2016.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pp. 849–856. MIT Press, 2001.

Jiming Peng and Yu Wei. Approximating k-means-type clustering via semidefinite programming. *SIAM J. OPTIM*, 18(1):186–205, 2007.

Martin Royer. Adaptive clustering through semidefinite programming. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 1795–1803. Curran Associates, Inc., 2017.

Jianbo Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, Aug 2000. ISSN 1939-3539. doi: 10.1109/34.868688.

Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *J. Comput. Syst. Sci*, 68:2004, 2004.

Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

Ulrike von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *Annals of Statistics*, 36(2):555–586, 2008.

Yihong Wu and Harrison H. Zhou. Randomly initialized em algorithm for two-component gaussian mixture achieves near optimality in $o(\sqrt{n})$ iterations, 2019. URL https://arxiv.org/abs/1908.10935.

Yubo Zhuang, Xiaohui Chen, and Yun Yang. Wasserstein k-means for clustering probability distributions. In *Proceedings of Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS)*, 2022a.

Yubo Zhuang, Xiaohui Chen, and Yun Yang. Sketch-and-lift: scalable subsampled semidefinite program for k-means clustering. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151 of *Proceedings of Machine Learning Research*, pp. 9214–9246. PMLR, 28–30 Mar 2022b.