Achieving Optimal Clustering in Gaussian Mixture Models with Anisotropic Covariance Structures

Xin Chen
Princeton University
xc5557@princeton.edu

Anderson Ye Zhang
University of Pennsylvania
ayz@wharton.upenn.edu

Abstract

We study clustering under anisotropic Gaussian Mixture Models (GMMs), where covariance matrices from different clusters are unknown and are not necessarily the identical matrix. We analyze two anisotropic scenarios: homogeneous, with identical covariance matrices, and heterogeneous, with distinct matrices per cluster. For these models, we derive minimax lower bounds that illustrate the critical influence of covariance structures on clustering accuracy. To solve the clustering problem, we propose a variant of Lloyd's algorithm, adapted to estimate and utilize covariance information iteratively. We prove that the adjusted algorithm not only adheres to the minimax optimality but also converges within logarithmic iterations, bridging the gap between theoretical robustness and practical efficiency.

1 Introduction

Clustering is a fundamentally important task in statistics and machine learning [7, 2]. The most widely recognized and extensively studied model for clustering is the Gaussian Mixture Model (GMM) [16, 18], which is formulated as

$$Y_j = \theta^*_{z^*_j} + \epsilon_j, \text{ where } \epsilon_j \overset{ind}{\sim} \mathcal{N}(0, \Sigma^*_{z^*_j}), \ \forall j \in [n].$$

Here $Y = (Y_1, \ldots, Y_n)$ are the observations with n being the sample size. We define the set $[n] := \{1, 2, \ldots, n\}$. Assume k as the known number of clusters. Let $\{\theta^*_a\}_{a\in[k]}$ represent the unknown centers, and Σ^*_a denote the corresponding unknown covariance matrices. Define $z^* \in [k]^n$ as the cluster structure, where for each index $j \in [n]$, the value of z^*_j specifies the cluster to which the j-th data point is assigned. The goal is to recover z^* from Y. For any estimator $\hat{z}$, its clustering performance is measured by the misclustering error rate $h(\hat{z}, z^*)$, which will be introduced later in (2).

There has been increasing interests in theoretical and algorithmic analysis of the clustering under the GMM. In the scenario where the GMM is isotropic, meaning all the covariance matrices $\{\Sigma^*_a\}_{a\in[k]}$ are equivalent to the identity matrix, [14] obtained the minimax rate for the clustering which takes the form of $\exp(-(1+o(1))(\min_{a\neq b} \|\theta^*_a - \theta^*_b\|)^2/8)$, with respect to the misclustering error rate. A diverse range of methods have been explored in the context of the isotropic setting. Among these, Lloyd's algorithm [12] emerges as a particularly distinguished clustering algorithm, renowned for its extensive success in a myriad of disciplines. [14, 8] established computational and statistical guarantees for the Lloyd's algorithm. Specifically, they showed it achieves the minimax optimal rates after a few iterations provided with some decent initialization. Another popular approach to clustering especially for high dimensional data is the spectral clustering [20, 17, 19], which is an umbrella term for clustering after a dimension reduction through a spectral decomposition. [13] proves the spectral clustering also achieves the optimality under the isotropic GMM. Semidefinite programming (SDP) is also used for clustering by exploiting its low-rank structure, and its statistical properties has been studied in several literature, for example, [5].

Preprint. Under review.

Despite the numerous compelling findings, the bulk of existing research primarily centers on isotropic GMM. The understanding of clustering in an anisotropic context, where the covariance matrices are not constrained to being identity matrices, remains relatively limited. Some papers, such as [14, 5, 15, 1, 9, 22], present results applicable to sub-Gaussian mixture models, wherein the errors ϵ_j are assumed to follow some sub-Gaussian distributions with the variance proxy σ^2. At first glance, it might appear that these results encompass the anisotropic case, as distributions of the form $\{\mathcal{N}(0, \Sigma_a^*)\}_{a\in[k]}$ are indeed sub-Gaussian distributions. However, from a minimax perspective, the least favorable scenario among all sub-Gaussian distributions with variance proxy σ^2—and thus the most challenging for clustering—is when the errors are distributed as $\mathcal{N}(0, \sigma^2 I)$. Therefore, the minimax rates for clustering under the sub-Gaussian mixture model is essentially the one under the isotropic GMM, and methods such as the Lloyd's algorithm that requires no covariance matrix information can be rate-optimal. As a result, the aforementioned results are all essentially for isotropic GMMs.

A few papers have explored the direction of clustering under anisotropic GMMs. [3] gives a polynomial-time clustering algorithm that provably works well when the Gaussian distributions are well separated by hyperplanes. Their idea is further developed in [10] which allows the Gaussians to be overlapped with each other but only for two-cluster cases. [21] proposes another method for clustering under a balanced mixture of two elliptical distributions. They give a provable upper bound of their clustering performance with respect to an excess risk. Nevertheless, it remains unknown what is the fundamental limit of clustering under the anisotropic GMMs and whether there is any polynomial-time procedure that achieves it.

In this paper, we will investigate the clustering task under two anisotropic GMMs. Model 1 is when the covariance matrices are all equal to each other (i.e., homogeneous), equal to some unknown matrix Σ^*. Model 2 is more flexible, where the covariance matrices are unknown and are not necessarily equal to each other (i.e., heterogeneous). The contribution of this paper is two-fold, summarized as follows:

- Our first contribution is on the minimax rates. We obtain the minimax lower bound for clustering under the anisotropic GMM with respect to the misclustering error rate. We show it takes the form of

$$\inf_{\hat{z}} \sup_{z^*} \mathbb{E} h(z, z^*) \geq \exp\left(-(1+o(1))\frac{(\text{signal-to-noise ratio})^2}{8}\right),$$

  where the signal-to-noise ratio under Model 1 is equal to $\min_{a,b\in[k]:a\neq b} \|(\theta_a^* - \theta_b^*)^T {\Sigma^*}^{-\frac{1}{2}}\|$. The signal-to-noise ratio for Model 2 is more intricate and will be introduced in Section 3. For both models, we can see the minimax rates depends not only on the centers but also the covariance matrices. This is different from the istropic case whose signal-to-noise ratio is $\min_{a\neq b} \|\theta_a^* - \theta_b^*\|$. Our results precisely captures the role the covariance matrices played in the clustering problem. It shows covariance matrices impact the fundamental limits of the clustering problem through complicated interacting with the centers especially in the Model 2. The minimax lower bounds are obtained by connections with the Linear Discriminant Analysis (LDA) [6] and the Quadratic Discriminant Analysis (QDA).
- Our second and more important contribution is on the computational side. We propose a computationally feasible procedure and rate-optimal algorithm for the anisotropic GMM. The Lloyd's algorithm is no longer optimal as it is developed under the isotropic case and only considers the distances among the centers [3]. We studies an *adjusted Lloyd's algorithm* which estimates the covariance matrices in each iteration and recovers the clusters adjusted to the covariance matrix information. It can also be seen as a hard EM algorithm [4]. As an iterative algorithm, we show that it obtains the minimax lower bound within $\log n$ iterations. This offers both a statistical and computational guarantee, serving as valuable guidance for practitioners. Specifically, if we let $z^{(t)}$ denote the output of the algorithm after t iterations, we have with high probability,

$$h(z^{(t)}, z^*) \leq \exp\left(-(1+o(1))\frac{(\text{signal-to-noise ratio})^2}{8}\right),$$

  holds for all $t \geq \log n$. The algorithm can be initialized by popular methods such as the spectral clustering or the Lloyd's algorithm. In numeric studies, we show the proposed

algorithm improves greatly from the two aforementioned methods under the anisotropic GMM, and matches with the optimal exponent given in the minimax lower bound.

Paper Organization. The remaining paper is organized as follows. In Section 2, we study the Model 1 where the covariance matrices are unknown but homogeneous. In Section 3, we consider the Model 2 where covariance matrices are unknown and heterogeneous. For both cases, we establish the minimax lower bound for the clustering and propose a computationally feasible and rate-optimal procedures. In Section 4, we provide a numeric comparison with other popular methods. Proofs are included in the supplement.

Notation. For any matrix $X \in \mathbb{R}^{d\times d}$, we denote $\lambda_1(X)$ to be its smallest eigenvalue and $\lambda_d(X)$ to be its largest eigenvalue. In addition, we denote $\|X\|$ to be its operator norm. For any two vectors u, v of the same dimension, we denote $\langle u, v\rangle = u^T v$ to be its inner product. For any positive integer d, we denote I_d to be the $d \times d$ identity matrix. We denote $\mathcal{N}(\mu, \Sigma)$ to be the normal distribution with mean μ and covariance matrix Σ.

2 GMM with Unknown but Homogeneous Covariance Matrices

2.1 Model

We first consider the GMM where the covariance matrices of different clusters are unknown but are assumed to be equal to each other. Then the data generating progress can be displayed as follow:

Model 1:
$$Y_j = \theta^*_{z^*_j} + \epsilon_j, \text{ where } \epsilon_j \overset{ind}{\sim} \mathcal{N}(0, \Sigma^*), \forall j \in [n]. \tag{1}$$

Throughout the paper, we call it the *Model 1* for simplicity and to distinguish it from a different and more complicated one that will be introduced in Section 3. The goal is to recover the underlying cluster assignment vector z^*.

Loss Function. To measure the clustering performance, we consider the misclustering error rate defined as follows. For any $z, z^* \in [k]^n$, we define

$$h(z, z^*) = \min_{\psi\in\Psi} \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}\left\{\psi(z_j) \neq z^*_j\right\}, \tag{2}$$

where $\Psi = \{\psi : \psi \text{ is a bijection from } [k] \text{ to } [k]\}$. Here the minimum is over all the permutations over $[k]$ due to the identifiability issue of the labels $1, 2, \ldots, k$.

Signal-to-noise Ratio. Define the signal-no-noise ratio

$$\mathsf{SNR} = \min_{a,b\in[k]:a\neq b} \|(\theta^*_a - \theta^*_b)^T \Sigma^{*-\frac{1}{2}}\|, \tag{3}$$

which is a function of all the centers $\{\theta^*_a\}_{a\in[k]}$ and the covariance matrix Σ^*. As we will show later in Theorem 2.1, SNR captures the difficulty of the clustering problem and determines the minimax rate. For the geometric interpretation of SNR, we defer it after presenting Theorem 2.2.

A quantity closely related to SNR is the minimum distance among the centers. Define Δ as

$$\Delta = \min_{a,b\in[k]:a\neq b} \|\theta^*_a - \theta^*_b\|. \tag{4}$$

Then we can see SNR and Δ are in the same order if all eigenvalues of the covariance matrix Σ^* is assumed to be constants. If Σ^* is further assumed to be an identical matrix, then we have SNR equal to Δ. As a result, in [14, 8, 13] where the isotropic GMMs are studied, Δ plays the role of signal-to-noise ratio and appears in their rates.

2.2 Minimax Lower Bound

We first establish the minimax lower bound for the clustering problem under the Model 1.

Theorem 2.1. *Under the assumption* $\frac{\mathsf{SNR}}{\log k} \to \infty$, *we have*

$$\inf_{\hat{z}} \sup_{z^* \in [k]^n} \mathbb{E} h(z, z^*) \geq \exp\left(-(1+o(1))\frac{\mathsf{SNR}^2}{8}\right). \tag{5}$$

If $\mathsf{SNR} = O(1)$ *instead, we have* $\inf_{\hat{z}} \sup_{z^* \in [k]^n} \mathbb{E} h(z, z^*) \geq c$ *for some constant* $c > 0$.

Theorem 2.1 allows the cluster numbers k grow with n and shows that $\mathsf{SNR} \to \infty$ is a necessary condition to have a consistent clustering. If k is a constant, then the condition can be reduced to $\mathsf{SNR} \to \infty$. Theorem 2.1 holds any arbitrary $\{\theta_a^*\}_{a\in[k]}$ and Σ^*, and the minimax lower bound depend on them through SNR. The parameter space is only for z^* while $\{\theta_a^*\}_{a\in[k]}$ and Σ^* are fixed. Hence, (5) can be interpreted as a pointwise result, and it captures precisely the explicit dependence of the minimaxity on $\{\theta_a^*\}_{a\in[k]}$ and Σ^*.

Theorem 2.1 is closely related to the Linear Discriminant Analysis (LDA). If there are only two clusters, and if the centers and the covariance matrices are known, then estimating each z_j^* is exactly the task of LDA: we want to figure out which normal distribution the observation Y_j is generated from, where the two normal distributions have different means but the same covariance matrix. In fact, this is also how Theorem 2.1 is proved: we will first reduce the estimation problem of z^* into two-point hypothesis testing problems for each individual z_j^*, the error of which is given in Lemma A.1 by the analysis of LDA (we will introduce next), and then aggregate all the testing errors together.

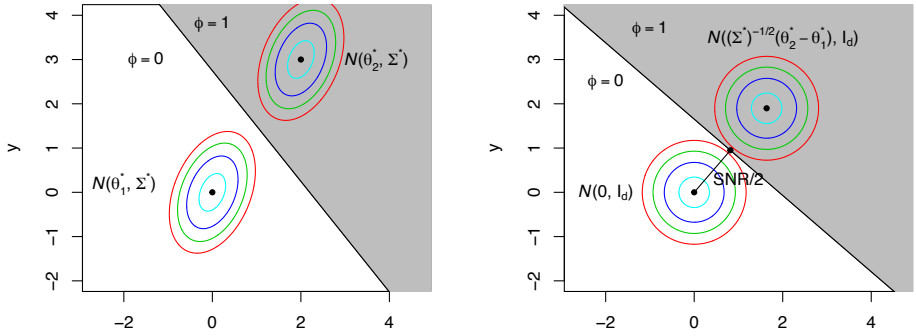

Figure 1: A geometric interpretation of SNR.

With the help of Lemma A.1, we have a geometric interpretation of SNR. In the left panel of Figure 1, we have two normal distributions $\mathcal{N}(\theta_1^*, \Sigma^*)$ and $\mathcal{N}(\theta_2^*, \Sigma^*)$ where X follows from. The black line represents the optimal testing procedure ϕ displayed in Lemma A.1 that divides the space into two half spaces. To calculate the testing error, we can make a transformation $X' = \Sigma^{-\frac{1}{2}}(X - \theta_1^*)$ so that the two normal distributions become isotropic: $\mathcal{N}(0, I_d)$ and $\mathcal{N}((\Sigma^*)^{-\frac{1}{2}}(\theta_2^* - \theta_1^*), I_d)$ as displayed in the right panel. Then the distance between the two centers are $\|(\Sigma^*)^{-\frac{1}{2}}(\theta_2^* - \theta_1^*)\|$, and the distance between a center and the black curve is a half of it. Then $\mathbb{P}_{H_0}(\hat{\phi} = 1)$ is equal to the probability of $\mathcal{N}(0, I_d)$ in the grayed area, which is equal to $\exp(-(1+o(1))\|(\Sigma^*)^{-\frac{1}{2}}(\theta_2^* - \theta_1^*)\|^2/8)$ by Gaussian tail probability. As a result, $\|(\Sigma^*)^{-\frac{1}{2}}(\theta_2^* - \theta_1^*)\|$ is the effective distance between the two centers of $\mathcal{N}(\theta_1^*, \Sigma^*)$ and $\mathcal{N}(\theta_2^*, \Sigma^*)$ for the clustering problem, considering the geometry of the covariance matrix. Since we have multiple clusters, SNR defined in (3) can be intepreted as the minimum effective distances among the centers $\{\theta_a^*\}_{a\in[k]}$ considering the anisotropic structure of Σ^* and it captures the intrinsic difficulty of the clustering problem.

2.3 Rate-Optimal Adaptive Procedure

In this section, we will propose a computational feasible and rate-optimal procedure for clustering under Model 1. Summarized in Algorithm 1, the proposed algorithm is a variant of the Lloyd algorithm. Starting from some initialization, it updates the estimation of the centers $\{\theta_a^*\}_{a\in[k]}$ (in (6)), the covariance matrix Σ^* (in (7)), and the cluster assignment vector z^* (in (8)) iteratively. This algorithm differs from Lloyd's algorithm in that the latter is designed for isotropic GMMs and does not

incorporate the covariance matrix update outlined in equation (7). Furthermore, in (8) it updates the estimation of z_j^* using $\operatorname{argmin}_{a\in[k]}(Y_j-\theta_a^{(t)})^T(Y_j-\theta_a^{(t)})$ instead. To clearly differentiate, we refer to the classic form as the *vanilla Lloyd's algorithm* and our modified version, which accommodates the unknown and anisotropic covariance matrix, as the *adjusted Lloyd's algorithm*.

Algorithm 1 can also be interpreted as a hard EM algorithm. If we apply the Expectation Maximization (EM) for the Model 1, we will have E step for estimating parameters $\{\theta_a^*\}_{a\in[k]}$ and Σ^* and M step for estimating z^*. It turns out the updates on the parameters (6) - (7) are exactly the same as the updates of EM (M step). However, the update on z^* in Algorithm 1 is different from that in the EM. Instead of taking a conditional expectation (E step), we also take a maximization in (8). As a result, Algorithm 1 consists solely of M steps for both the parameters and z^*, which is known as a hard EM algorithm.

Algorithm 1: Adjusted Lloyd's Algorithm for Model 1 (1).

Input: Data Y, number of clusters k, an initialization $z^{(0)}$, number of iterations T.
Output: $z^{(T)}$

for $t=1,\dots,T$ **do**

Update the centers:

$$\theta_a^{(t)}=\frac{\sum_{j\in[n]}Y_j\mathbb{I}\left\{z^{(t-1)}=a\right\}}{\sum_{j\in[n]}\mathbb{I}\left\{z^{(t-1)}=a\right\}},\quad \forall a\in[k]. \tag{6}$$

Update the covariance matrix:

$$\Sigma^{(t)}=\frac{\sum_{a\in[k]}\sum_{j\in[n]}(Y_j-\theta_a^{(t)})(Y_j-\theta_a^{(t)})^T\mathbb{I}\left\{z^{(t-1)}=a\right\}}{n}. \tag{7}$$

Update the cluster estimations:

$$z^{(t)}=\underset{a\in[k]}{\operatorname{argmin}}(Y_j-\theta_a^{(t)})^T(\Sigma^{(t)})^{-1}(Y_j-\theta_a^{(t)}),\quad j\in[n]. \tag{8}$$

In Theorem 2.2, we give a computational and statistical guarantee of the proposed Algorithm 1. We show that starting from a decent initialization, within $\log n$ iterations, Algorithm 1 achieves the error rate $\exp\left(-(1+o(1))\mathsf{SNR}^2/8\right)$ which matches with the minimax lower bound given in Theorem 2.1. As a result, Algorithm 1 is a rate-optimal procedure. In addition, the algorithm is fully adaptive to the unknown $\{\theta_a^*\}_{a\in[k]}$ and Σ^*. The sole piece of information presumed to be known is k, the number of clusters, a common assumption evidenced in clustering literature [14, 8, 13]. The theorem also shows that the number of iterations to achieve the optimal rate is at most $\log n$, which provides an implementation guidance to practitioners.

Theorem 2.2. *Assume $d=O(\sqrt{n})$ and $\min_{a\in k}\sum_{j=1}^n\mathbb{I}\{z_j^*=a\}\geq\frac{\alpha n}{k}$ for some constant $\alpha>0$. Assume $\frac{\mathsf{SNR}}{\log k}\to\infty$ and $\lambda_d(\Sigma^*)/\lambda_1(\Sigma^*)=O(1)$. For Algorithm 1, suppose $z^{(0)}$ satisfies $h(z^{(0)},z^*)=o(k^{-1})$ with probability at least $1-\eta$. Then with probability at least $1-\eta-n^{-1}-\exp(-\mathsf{SNR})$, we have*

$$h(z^{(t)},z^*)\leq\exp\left(-(1+o(1))\frac{\mathsf{SNR}^2}{8}\right),\quad \textit{for all } t\geq\log n.$$

We have remarks on the assumptions of Theorem 2.2. We allow the number of cluster k grow with n. When k is a constant, the assumption on $\mathsf{SNR}\to\infty$ is the necessary condition to have a consistent recovery of z^* according to the minimax lower bound presented in Theorem 2.1. The assumption on Σ^* is to make sure the covariance matrix is well-conditioned. The dimensionality d is assumed to be $O(\sqrt{n})$, an assumption that is stronger than that in [14, 8, 13] which only needs $d=O(n)$. This is due to that compared to these papers, we need to estimate the covariance matrix Σ^* and to have a control on the estimation error $\|\Sigma^{(t)}-\Sigma^*\|$.

The requirement for the initialization $h(z^{(0)}, z^*) = o(k^{-1})$ can be fulfilled by simple procedures. A popular choice is the spectral clustering. For instance, we can use a variant of spectral clustering studied in [13]. Since Model 1 can be seen as a sub-Gaussian mixture model, in their Proposition D.1, they show the spectral clustering output $\hat{z}^{\text{spectral}}$ achieves

$$h(\hat{z}^{\text{spectral}}, z^*) = O\left(\frac{k}{\mathsf{SNR}^2}\right), \tag{9}$$

with probability at least $1 - \exp(-0.08)$, under the same assumption as in Theorem 2.2. Then it can be used as an initialization in Theorem 2.2 if we pose a slightly stronger assumption $\mathsf{SNR}/k \to \infty$. As a result, we immediately have the following corollary.

Corollary 2.1. *Assume $d = O(\sqrt{n})$ and $\min_{a\in k}\sum_{j=1}^{n}\mathbb{I}\{z_j^* = a\} \geq \frac{\alpha n}{k}$ for some constant $\alpha > 0$. Assume $\frac{\mathsf{SNR}}{k} \to \infty$ and $\lambda_d(\Sigma^*)/\lambda_1(\Sigma^*) = O(1)$. Using the spectral clustering $\hat{z}^{\text{spectral}}$ as the initialization $z^{(0)}$ in Algorithm 1, we have with probability at least $1 - \exp(-0.08) - n^{-1} - \exp(-\mathsf{SNR})$,*

$$h(z^{(t)}, z^*) \leq \exp\left(-(1+o(1))\frac{\mathsf{SNR}^2}{8}\right), \quad \textit{for all } t \geq \log n.$$

3 GMM with Unknown and Heterogeneous Covariance Matrices

3.1 Model

In this section, we study the GMM where the covariance matrices of each cluster are unknown and not necessarily equal to each other. The data generation process can be displayed as follow,

Model 2: $$Y_j = \theta^*_{z_j^*} + \epsilon_j, \text{ where } \epsilon_j \overset{ind}{\sim} \mathcal{N}(0, \Sigma^*_{z_j^*}), \forall j \in [n]. \tag{10}$$

We call it *Model 2* throughout the paper to distinguish from Model 1 studied in Section 2. The difference between (10) and (1) is that we now have $\{\Sigma_a^*\}_{a\in[k]}$ instead of a shared Σ^*. We consider the same loss function defined as in (2).

Signal-to-noise Ratio. The signal-to-noise ratio for the Model 2 is defined as follows. We use the notation SNR' to distinguish it from the SNR for the Model 1. Compared to SNR, SNR' is much more complicated and does not have an explicit formula. We first define a space $\mathcal{B}_{a,b} \in \mathbb{R}^d$ for any $a, b \in [k]$ such that $a \neq b$:

$$\mathcal{B}_{a,b} = \Bigg\{ x \in \mathbb{R}^d : x^T \Sigma_a^{*\frac{1}{2}} \Sigma_b^{*-1}(\theta_a^* - \theta_b^*) + \frac{1}{2} x^T \left(\Sigma_a^{*\frac{1}{2}} \Sigma_b^{*-1} \Sigma_a^{*\frac{1}{2}} - I_d\right) x \\ \leq -\frac{1}{2}(\theta_a^* - \theta_b^*)^T \Sigma_b^{*-1}(\theta_a^* - \theta_b^*) + \frac{1}{2}\log|\Sigma_a^*| - \frac{1}{2}\log|\Sigma_b^*| \Bigg\}.$$

We then define $\mathsf{SNR}'_{a,b} = 2\min_{x\in\mathcal{B}_{a,b}} \|x\|$ and

$$\mathsf{SNR}' = \min_{a,b\in[n]:a\neq b} \mathsf{SNR}'_{a,b}. \tag{11}$$

The form of SNR' is closely connected to the testing error of the Quadratic Discriminant Analysis (QDA), which we will give in Lemma 3.1. The interpretation of the SNR', particularly from a geometric perspective, will be deferred until after the presentation of Lemma 3.1. Here let us consider a few special cases where we are able to simplify SNR': (1) When $\Sigma_a^* = \Sigma^*$ for all $a \in [k]$, by simple algebra, we have $\mathsf{SNR}'_{a,b} = \|(\theta_a^* - \theta_b^*)^T \Sigma^{*-\frac{1}{2}}\|$ for any $a, b \in [k]$ such that $a \neq b$. Hence, $\mathsf{SNR}' = \mathsf{SNR}$ and Model 2 is reduced to Model 1. (2) When $\Sigma^* = \sigma_a^2 I_d$ for any $a \in [k]$ where $\sigma_1, \ldots, \sigma_k > 0$ are large constants, we have $\mathsf{SNR}'_{a,b}, \mathsf{SNR}'_{b,a}$ both close to $2\|\theta_a^* - \theta_b^*\|/(\sigma_a + \sigma_b)$. From these examples, we can see SNR' is determined by both the centers $\{\theta_a^*\}_{a\in[k]}$ and the covariance matrices $\{\Sigma_a^*\}_{a\in[k]}$.

3.2 Minimax Lower Bound

We first establish the minimax lower bound for the clustering problem under the Model 2.

Theorem 3.1. *Under the assumption* $\frac{\mathsf{SNR}'}{\log k} \to \infty$, *we have*

$$\inf_{\hat{z}} \sup_{z^* \in [k]^n} \mathbb{E} h(z, z^*) \geq \exp\left(-(1+o(1))\frac{\mathsf{SNR}'^2}{8}\right).$$

If $\mathsf{SNR}' = O(1)$ *instead, we have* $\inf_{\hat{z}} \sup_{z^* \in [k]^n} \mathbb{E} h(z, z^*) \geq c$ *for some constant* $c > 0$.

Despite that the statement of Theorem 3.1 looks similar to that of Theorem 2.1, the two minimax lower bounds are different from each other due to the discrepancy in the dependence of the centers and the covariance matrices in SNR' and SNR. By the same argument as in Section 2.2, the minimax lower bound established in Theorem 3.1 is closely related to the Quadratic Discriminant Analysis (QDA) between two normal distributions with different means and different covariance matrices.

Lemma 3.1 (Testing Error for Quadratic Discriminant Analysis)**.** *Consider two hypotheses* $\mathbb{H}_0 : X \sim \mathcal{N}(\theta_1^*, \Sigma_1^*)$ *and* $\mathbb{H}_1 : X \sim \mathcal{N}(\theta_2^*, \Sigma_2^*)$. *Define a testing procedure*

$$\phi = \mathbb{I}\left\{\log|\Sigma_1^*| + (x-\theta_1^*)^T\Sigma_1^*(x-\theta_1^*) \geq \log|\Sigma_2^*| + (x-\theta_2^*)^T\Sigma_2^*(x-\theta_2^*)\right\}.$$

Then we have $\inf_{\hat{\phi}}(\mathbb{P}_{\mathbb{H}_0}(\hat{\phi} = 1) + \mathbb{P}_{\mathbb{H}_1}(\hat{\phi} = 0)) = \mathbb{P}_{\mathbb{H}_0}(\phi = 1) + \mathbb{P}_{\mathbb{H}_1}(\phi = 0)$. *If* $\min\left\{\mathsf{SNR}'_{1,2}, \mathsf{SNR}'_{2,1}\right\} \to \infty$, *we have*

$$\inf_{\hat{\phi}}(\mathbb{P}_{H_0}(\hat{\phi} = 1) + \mathbb{P}_{H_1}(\hat{\phi} = 0)) \geq \exp\left(-(1+o(1))\frac{\min\left\{\mathsf{SNR}'_{1,2}, \mathsf{SNR}'_{2,1}\right\}^2}{8}\right).$$

Otherwise, $\inf_{\hat{\phi}}(\mathbb{P}_{H_0}(\hat{\phi} = 1) + \mathbb{P}_{H_1}(\hat{\phi} = 0)) \geq c$ *for some constant* $c > 0$.

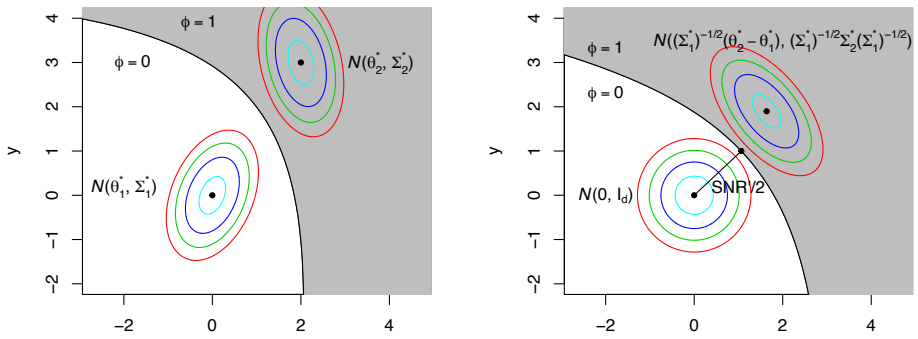

Figure 2: A geometric interpretation of SNR'.

From Lemma 3.1, we can have a geometric interpretation of SNR'. In the left panel of Figure 2, we have two normal distribution $\mathcal{N}(\theta_1^*, \Sigma_1^*)$ and $\mathcal{N}(\theta_2^*, \Sigma_2^*)$ where X can be generated from and the black curve represents the optimal testing procedure ϕ displayed in Lemma 3.1. Since Σ_1^* is not necessarily equal to Σ_2^*, the black curve is not necessarily a straight line. If $\mathbb{P}_{\mathbb{H}_0}(\phi = 1)$, then the probability for X to be wrong classified is when X falls in the gray area, which is $\mathbb{P}_{H_0}(\hat{\phi} = 1)$. To calculate it, we can make a transformation $X' = (\Sigma_1^*)^{-\frac{1}{2}}(X - \theta_1^*)$. Then displayed in the right panel of Figure 2, the two distributions become $\mathcal{N}(0, I_d)$ and $\mathcal{N}((\Sigma_1^*)^{-\frac{1}{2}}(\theta_2^* - \theta_1^*), (\Sigma_1^*)^{-\frac{1}{2}}\Sigma_2^*(\Sigma_1^*)^{-\frac{1}{2}})$, and the optimal testing procedure ϕ becomes $\phi' = \mathbb{I}\{X' \in \mathcal{B}_{1,2}\}$. As a result, in the right panel of Figure 2, $\mathcal{B}_{1,2}$ represents the space colored by gray, and the black curve is its boundary. Then $\mathbb{P}_{H_0}(\hat{\phi} = 1)$ is equal to $\mathbb{P}(\mathcal{N}(0, I_d) \in \mathcal{B}_{1,2})$, which can be shown to be determined by the minimum distance between the center of $\mathcal{N}(0, I_d)$ and the space $\mathcal{B}_{1,2}$. Denote the minimum distance by $\mathsf{SNR}'_{1,2}/2$, by Lemmas C.8 and Lemma C.9, we can show $\mathbb{P}(\mathcal{N}(0, I_d) \in \mathcal{B}_{1,2}) = \exp(-(1+o(1))\mathsf{SNR}'^2_{1,2}/8)$. As a result, the SNR' can be interpreted as the minimum effective distance among the centers $\{\theta_a^*\}_{a\in[k]}$ considering the anisotropic and heterogeneous structure of $\{\Sigma_a^*\}_{a\in[k]}$ and it captures the intrinsic difficulty of the clustering problem under Model 2.

3.3 Optimal Adaptive Procedure

In this section, we will propose a computational feasible and rate-opitmal procedure for clustering under Model 2. Similar to the Algorithm 1, the proposed Algorithm 2 can be seen as a variant of the Lloyd's algorithm that is adjusted to the unknown and heterogeneous covariance matrices. It can also be interpreted as a hard EM algorithm under Model 2. Algorithm 2 differs from Algorithm 1 in (13) and (14), as now there are k covariance matrices to be estimated.

Algorithm 2: Adjusted Lloyd's Algorithm for Model 2 (10).

Input: Data Y, number of clusters k, an initialization $z^{(0)}$, number of iterations T.
Output: $z^{(T)}$

for $t = 1, \ldots, T$ **do**

Update the centers:

$$\theta_a^{(t)} = \frac{\sum_{j\in[n]} Y_j \mathbb{I}\left\{z^{(t-1)} = a\right\}}{\sum_{j\in[n]} \mathbb{I}\left\{z^{(t-1)} = a\right\}}, \quad \forall a \in [k]. \tag{12}$$

Update the covariance matrices:

$$\Sigma_a^{(t)} = \frac{\sum_{j\in[n]} (Y_j - \theta_a^{(t)})(Y_j - \theta_a^{(t)})^T \mathbb{I}\left\{z^{(t-1)} = a\right\}}{\sum_{j\in[n]} \mathbb{I}\left\{z^{(t-1)} = a\right\}}, \quad \forall a \in [k]. \tag{13}$$

Update the cluster estimations:

$$z^{(t)} = \underset{a\in[k]}{\operatorname{argmin}} (Y_j - \theta_a^{(t)})^T (\Sigma_a^{(t)})^{-1} (Y_j - \theta_a^{(t)}) + \log|\Sigma_a^{(t)}|, \quad j \in [n]. \tag{14}$$

In Theorem 3.2, we give a computational and statistical guarantee of the proposed Algorithm 2. We show that provided with some decent initialization, Algorithm 2 is able to achieve the minimax lower bound within $\log n$ iterations. The assumptions needed in Theorem 3.2 are similar to those in Theorem 3.2, except that we require stronger assumptions on k and the dimensionality d since now we have k (instead of one) covariance matrices to be estimated. In addition, by $\max_{a,b\in[k]} \lambda_d(\Sigma_a^*)/\lambda_1(\Sigma_b^*) = O(1)$ we not only assume each of the k covariance matrices is well-conditioned, but also assume they are comparable to each other.

Theorem 3.2. *Assume* $d = O(1)$ *and* $\min_{a\in k} \sum_{j=1}^n \mathbb{I}\{z_j^* = a\} \geq \frac{\alpha n}{k}$ *for some constant* $\alpha > 0$. *Assume* $k = O(1)$, $\mathit{SNR}' \to \infty$ *and* $\max_{a,b\in[k]} \lambda_d(\Sigma_a^*)/\lambda_1(\Sigma_b^*) = O(1)$. *For Algorithm 2, suppose* $z^{(0)}$ *satisfies* $h(z^{(0)}, z^*) = o(1)$ *with probability at least* $1 - \eta$. *Then with probability at least* $1 - \eta - n^{-1} - \exp(-\mathit{SNR}')$, *we have*

$$h(z^{(t)}, z^*) \leq \exp\left(-(1 + o(1))\frac{\mathit{SNR}'^2}{8}\right), \quad \textit{for all } t \geq \log n.$$

Given the assumption that $\max_{a,b\in[k]} \lambda_d(\Sigma_a^*)/\lambda_1(\Sigma_b^*) = O(1)$, Model 2 also qualifies as a sub-Gaussian mixture model. Consequently, when spectral clustering is used as the initialization for Algorithm 2, the equation (9) remains applicable. This leads us to the following corollary.

Corollary 3.1. *Assume* $d = O(1)$ *and* $\min_{a\in k} \sum_{j=1}^n \mathbb{I}\{z_j^* = a\} \geq \frac{\alpha n}{k}$ *for some constant* $\alpha > 0$. *Assume* $k = O(1)$, $\mathit{SNR}' \to \infty$ *and* $\lambda_d(\Sigma^*)/\lambda_1(\Sigma^*) = O(1)$. *Using the spectral clustering* $\hat{z}^{\text{spectral}}$ *as the initialization* $z^{(0)}$ *in Algorithm 2, we have with probability at least* $1 - \exp(-0.08) - n^{-1} - \exp(-\mathit{SNR})$,

$$h(z^{(t)}, z^*) \leq \exp\left(-(1 + o(1))\frac{\mathit{SNR}'^2}{8}\right), \quad \textit{for all } t \geq \log n.$$

4 Numerical Studies

In this section, we compare the performance of the proposed methods with other popular clustering methods on synthetic datasets under different settings.

Model 1. The first simulation is designed for the GMM with unknown but homogeneous covariance matrices (i.e., Model 1). We independently generate $n = 1200$ samples with dimension $d = 50$ from $k = 30$ clusters. Each cluster has 40 samples. We set $\Sigma^* = U^T \Lambda U$, where Λ is a 50×50 diagonal matrix with diagonal elements selected from 0.5 to 8 with equal space and U is a randomly generated orthogonal matrix. The centers $\{\theta_a^*\}_{a\in[n]}$ are orthogonal or each other with $\|\theta_1^*\| = \ldots = \|\theta_{30}^*\| = 9$. We consider four popular clustering methods: (1) the spectral clustering method in [13] (denoted as "spectral"), (2) the vanilla Lloyd's algorithm in [14] (denoted as "vanilla Lloyd"), (3) the proposed Algorithm 1 initialized by the spectral clustering (denoted as "spectral + Alg 1"), and (4) Algorithm 1 initialized by the vanilla Lloyd (denoted as "vanilla Lloyd + Alg 1"). The comparison is presented in left panel of Figure 3.

Model 2. We also compare the performances of four methods (spectral, vanilla Lloyd, spectral + Alg 2, and vanilla Lloyd + Alg 2) for the GMM with unknown and heterogeneous covariance matrices (i.e., Model 2). In this case, we take $n = 1200$, $k = 3$ and $d = 5$. We set $\Sigma_1^* = I$, $\Sigma_2^* = \Lambda_2$ which is a 5×5 diagonal matrix with elements generated from 0.5 to 8 with equal space and $\Sigma_3^* = U^T \Lambda_3 U$, where Λ_3 is a diagonal matrix with elements selected uniformly from 0.5 to 2 and U is a randomly generated orthogonal matrix. To simplify the calculation of SNR', we take θ_1^* as a randomly selected unit vector, $\theta_2^* = \theta_1^* + 5e_1$ with e_1 denoting the vector with a 1 in the first coordinate and 0's elsewhere and $\theta_3^* = \theta_2^* + v_1$ with v_1 randomly selected satisfying $\|v_1\| = 10$. The comparison is presented in the right panel of Figure 3.

In Figure 3, the x-axis is the number of iterations and the y-axis is the logarithm of the misclustering error rate, i.e., $\log(h)$. Each of the curves plotted is an average of 100 independent trials. We can see both Algorithm 1 and Algorithm 2 outperform the spectral clustering and the vanilla Lloyd's algorithm significantly. Additionally, the dashed lines in the left and right panel represent the optimal exponents $-\mathsf{SNR}^2/8$ and $-\mathsf{SNR}'^2/8$ of the minimax bound respectively. Observably, both Algorithm 1 and Algorithm 2 attain these benchmarks after three iterations. This justifies the conclusion that both algorithms are rate-optimal.

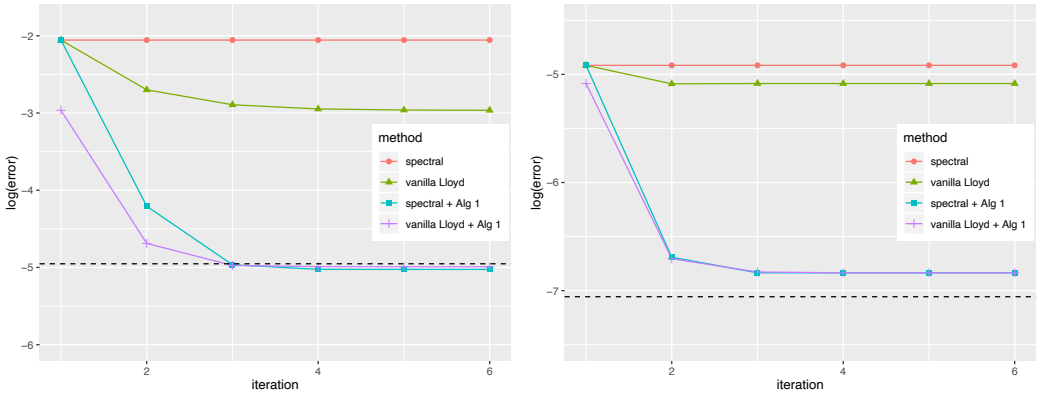

Figure 3: Left: Performance of the Algorithm 1 compared with other methods under Model 1. Right: Performance of the Algorithm 2 compared with other methods under Model 2.

5 Conclusion

This paper focuses on clustering methods and theory for anisotropic GMMs, presenting new minimax bounds and an adjusted Lloyd's algorithm tailored for varying covariance structures. Our theoretical and empirical analyses demonstrate the algorithm's ability to achieve optimality within a logarithmic number of iterations. Future work will explore broader covariance structures and refine these methods to further bridge theoretical robustness with computational feasibility in complex clustering scenarios.

References

[1] Emmanuel Abbe, Jianqing Fan, and Kaizheng Wang. An ℓ_p theory of pca and spectral clustering. *The Annals of Statistics*, 50(4):2359–2385, 2022.

[2] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[3] S Charles Brubaker and Santosh S Vempala. Isotropic pca and affine-invariant clustering. In *Building Bridges*, pages 241–281. Springer, 2008.

[4] Arthur P Dempster, Nan M Laird, and Donald 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.

[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] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[7] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[8] Chao Gao and Anderson Y Zhang. Iterative algorithm for discrete structure recovery. *The Annals of Statistics*, 50(2):1066–1094, 2022.

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

[10] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 553–562, 2010.

[11] Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, pages 1302–1338, 2000.

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

[13] Matthias Löffler, Anderson Y Zhang, and Harrison H Zhou. Optimality of spectral clustering in the gaussian mixture model. *The Annals of Statistics*, 49(5):2506–2530, 2021.

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

[15] Mohamed Ndaoud. Sharp optimal recovery in the two component gaussian mixture model. *The Annals of Statistics*, 50(4):2096–2126, 2022.

[16] Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

[17] Daniel A Spielman and Shang-Hua Teng. Spectral partitioning works: Planar graphs and finite element meshes. In *Proceedings of 37th Conference on Foundations of Computer Science*, pages 96–105. IEEE, 1996.

[18] D Michael Titterington, Adrian FM Smith, and Udi E Makov. *Statistical analysis of finite mixture distributions*. Wiley,, 1985.

[19] S. Vempala and G. Wang. A spectral algorithm for learning mixture models. *J. Comput. Syst. Sci.*, 68(4):841–860, 2004.

[20] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[21] Kaizheng Wang, Yuling Yan, and Mateo Díaz. Efficient clustering for stretched mixtures: Landscape and optimality. *Advances in Neural Information Processing Systems*, 33:21309–21320, 2020.

[22] Anderson Y Zhang and Harrison H Zhou. Leave-one-out singular subspace perturbation analysis for spectral clustering. *arXiv preprint arXiv:2205.14855*, 2022.