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 identity 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 achieves the minimax optimality but also converges within a logarithmic number of iterations, thus bridging the gap between theoretical guarantees 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) [17, 19], 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 is 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 assignment vector, where for each index $j \in [n]$, the value of z^*_j specifies which cluster the j-th data point is assigned to. 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 interest in theoretical and algorithmic analysis of clustering under GMMs. In a scenario where a GMM is isotropic, meaning that all covariance matrices $\{\Sigma^*_a\}_{a\in[k]}$ are equal to the identity matrix, [15] obtained the minimax rate for 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 has been explored in the context of the isotropic setting. Among these, Lloyd's algorithm [13] stands out as a particularly effective clustering algorithm, renowned for its extensive success in a myriad of disciplines. [15, 8] establish 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 [21, 18, 20], which is an umbrella term for clustering after a dimension reduction through a spectral decomposition. [14] proves the spectral clustering also achieves the optimality under the isotropic GMM. Semidefinite programming (SDP)

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

is also used for clustering by exploiting its low-rank structure, and its statistical properties have been studied in literature, for example, [5].

Despite the numerous compelling findings, most existing research primarily focuses on isotropic GMMs. The understanding of clustering in an anisotropic context, where the covariance matrices are not constrained to be identity matrices, remains relatively limited. Some studies, including [15, 5, 16, 1, 9, 24], present results for 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 rate for clustering under the sub-Gaussian mixture model essentially equals the one under the isotropic GMM, and methods like Lloyd's algorithm, which require no covariance matrix information, can be rate-optimal. As a result, the aforementioned findings primarily pertain to isotropic GMMs.

A few studies have explored the direction of clustering under anisotropic GMMs. [3] presents a polynomial-time clustering algorithm that provably performs well when Gaussian distributions are well-separated by hyperplanes. This idea is further developed in [11], which extends the approach to allow overlapping Gaussians, albeit only in two-cluster scenarios. [22] proposes a novel method for clustering under a balanced mixture of two elliptical distributions. They establish a provable upper bound on their clustering performance. Nevertheless, the fundamental limit of clustering under anisotropic GMMs, and whether a polynomial-time procedure can achieve it, remains unknown.

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

- Our first contribution is on the minimax rates. We obtain minimax lower bounds for clustering under anisotropic GMMs with respect to the misclustering error rate. We show they take the form of

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

  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 depend not only on the centers but also on the covariance matrices. This is different from the isotropic case, whose signal-to-noise ratio is $\min_{a\neq b} \|\theta_a^* - \theta_b^*\|$. Our results precisely capture the role that covariance matrices play in the clustering problem. This shows that covariance matrices impact the fundamental limits of the clustering problem through complex interactions with the centers, especially in Model 2. We obtain the minimax lower bounds by drawing connections with Linear Discriminant Analysis (LDA) [6] and 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. Lloyd's algorithm, developed for the isotropic case, is no longer optimal as it only considers distances among centers [3]. We study an *adjusted Lloyd's algorithm* which estimates the covariance matrices in each iteration and adjusts the clusters accordingly. It can also be seen as a hard EM algorithm [4]. Here, we modify the E-step of the soft EM by implementing a maximization step that directly assigns data points to clusters, rather than calculating probabilities. As an iterative algorithm, we demonstrate that it achieves the minimax lower bound within $\log n$ iterations. This offers both statistical and computational guarantees, serving as valuable guidance for practitioners. Specifically, if we let $z^{(t)}$ denote the output of the algorithm after t iterations, it holds with high probability that

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

  for all $t \geq \log n$. The algorithm can be initialized using popular methods like spectral clustering or Lloyd's algorithm. In our numerical studies, we demonstrate that the proposed

algorithm significantly improves over the two aforementioned methods under anisotropic GMMs, and matches the optimal exponent specified in the minimax lower bound.

Paper Organization. The remaining paper is organized as follows. In Section 2, we study Model 1 where the covariance matrices are unknown but homogeneous. In Section 3, we consider 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 procedure. In Section 4, we provide a numerical comparison with other popular methods. Proofs are included in the supplement.

Notation. For any matrix $X \in \mathbb{R}^{d\times d}$, we denote $\lambda_1(X)$ as its smallest eigenvalue and $\lambda_d(X)$ as its largest eigenvalue. In addition, we denote $\|X\|$ as its operator norm. For any two vectors u, v of the same dimension, we denote $\langle u, v\rangle = u^T v$ as their inner product. For any positive integer d, we denote I_d as the $d \times d$ identity matrix. We denote $\mathcal{N}(\mu, \Sigma)$ as the normal distribution with mean μ and covariance matrix Σ. We denote $\mathbb{I}\{\cdot\}$ as the indicator function. For two positive sequences $\{a_n\}$ and $\{b_n\}$, $a_n \preceq b_n$ and $a_n = O(b_n)$ both mean $a_n \le Cb_n$ for some constant $C > 0$ independent of n. We also write $a_n = o(b_n)$ or $\frac{b_n}{a_n} \to \infty$ when $\limsup_n \frac{a_n}{b_n} = 0$.

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 process can be displayed as follows:

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 *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^*. If Σ^* were known, then (1) can be converted into an isotropic GMM by a linear transformation $(\Sigma^*)^{-\frac{1}{2}} Y_j$. However, the unknown nature of Σ^* makes clustering under this model more challenging than under isotropic GMMs.

Loss Function. To measure the clustering performance, we consider the following loss function. 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 taken over all permutations of $[k]$ to address the identifiability issues of the labels $1, 2, \ldots, k$. The loss function measures the proportion of coordinates where z and z^* differ, modulo any permutation of label symbols. Thus, it is referred to as the misclustering error rate in this paper. Another loss that will be used is $\ell(z, z^*)$ defined as

$$\ell(z, z^*) = \sum_{j=1}^{n} \left\|\theta^*_{z_j} - \theta^*_{z^*_j}\right\|^2. \tag{3}$$

It measures the clustering performance of z considering the distances among the true centers. It is related to $h(z, z^*)$ as $h(z, z^*) \le \ell(z, z^*)/(n\Delta^2)$ and provides more information than $h(z, z^*)$. We will mainly use $\ell(z, z^*)$ in the technical analysis but will present results using $h(z, z^*)$ which is more interpretable.

Signal-to-noise Ratio. Define the signal-to-noise ratio

$$\text{SNR} = \min_{a,b\in[k]:a\neq b} \|(\theta^*_a - \theta^*_b)^T \Sigma^{*-\frac{1}{2}}\|, \tag{4}$$

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. We defer the geometric interpretation of SNR until after presenting Theorem 2.2.

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

$$\Delta = \min_{a,b\in[k]:a\neq b} \|\theta_a^* - \theta_b^*\|. \tag{5}$$

Then we can see SNR and Δ are of the same order if all eigenvalues of the covariance matrix Σ^* are assumed to be constants. If Σ^* is further assumed to be $\sigma^2 I_d$, then SNR equals Δ/σ. As a result, in [15, 8, 14] where the isotropic GMMs are studied, Δ/σ plays the role of signal-to-noise ratio and appears in their rates. Since (4) represents a direct generalization, we refer to it as the signal-to-noise ratio for Model 1.

2.2 Minimax Lower Bound

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

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

$$\inf_{\hat{z}} \sup_{z^*\in[k]^n} \mathbb{E}h(\hat{z}, z^*) \geq \exp\left(-(1+o(1))\frac{SNR^2}{8}\right). \tag{6}$$

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

Theorem 2.1 allows the cluster numbers k to grow with n and shows that SNR $\to \infty$ is a necessary condition to have a consistent clustering. If k is a constant, then SNR $\to \infty$ is also a sufficient condition. Theorem 2.1 holds for any arbitrary configurations of $\{\theta_a^*\}_{a\in[k]}$ and Σ^*, with the minimax lower bound depending on these through SNR. The parameter space is only for z^* while $\{\theta_a^*\}_{a\in[k]}$ and Σ^* are held fixed. Hence, (6) can be interpreted as a case-specific result, precisely capturing the explicit dependence of the minimax rates on $\{\theta_a^*\}_{a\in[k]}$ and Σ^*.

Theorem 2.1 is closely related to the LDA. If there are only two clusters with known centers and a covariance matrix, then estimating each z_j^* becomes exactly the task of the LDA: we aim to determine from which of two normal distributions, each with a different mean but the same covariance matrix, the observation Y_j is generated. In fact, this approach is also how Theorem 2.1 is proved: We first reduce the estimation problem of z^* to two-point hypothesis testing for each individual z_j^*. The error of these tests is analyzed in Lemma A.1 using the LDA, and we then aggregate all these testing errors together.

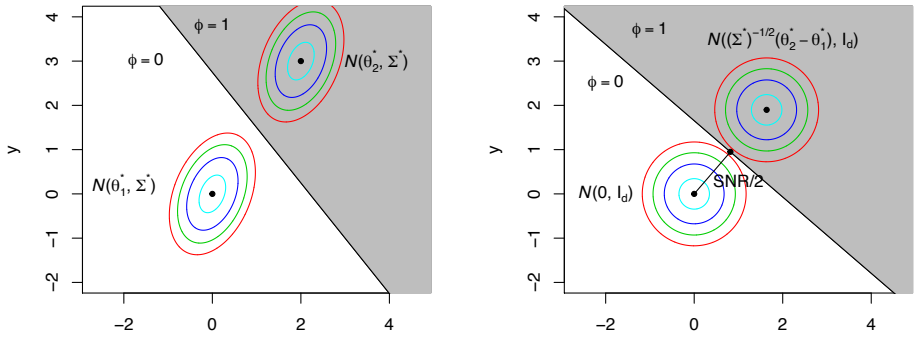

Figure 1: A geometric interpretation of SNR.

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 $\mathcal{N}(\theta_2^*, \Sigma^*)$ that X follows. The black line represents the optimal testing procedure ϕ displayed in Lemma A.1, dividing the space into two half-spaces. To calculate the testing error, we can make the 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 is $\|(\Sigma^*)^{-\frac{1}{2}}(\theta_2^* - \theta_1^*)\|$, and the distance from a center to the black curve is half of that. Then, the probability that $\mathcal{N}(0, I_d)$ falls within the grayed area equals $\exp(-(1+o(1))\|(\Sigma^*)^{-\frac{1}{2}}(\theta_2^* - \theta_1^*)\|^2/8)$, according to 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, taking into account the geometry of the covariance matrix. Since we have multiple clusters, SNR defined in (4) can be interpreted as the minimum effective distance among the centers $\{\theta_a^*\}_{a\in[k]}$, considering the anisotropic structure of Σ^*. This measure captures the intrinsic difficulty of the clustering problem.

2.3 Rate-Optimal Adaptive Procedure

In this section, we propose a computationally feasible and rate-optimal procedure for clustering under Model 1. Summarized in Algorithm 1, the proposed algorithm is a variant of Lloyd's algorithm. Starting with an initial setup, it iteratively updates the estimates of the centers $\{\theta_a^*\}_{a\in[k]}$ (in (7)), the covariance matrix Σ^* (in (8)), and the cluster assignment vector z^* (in (9)). 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 (8). Furthermore, (9) 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 differentiate clearly, 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. When applying Expectation Maximization (EM) to Model 1, the M step estimates the parameters $\{\theta_a^*\}_{a\in[k]}$ and Σ^*, while the E step estimates z^*. It turns out the updates on the parameters (7) - (8) are identical to those in the EM's M step. However, the update of z^* in Algorithm 1 differs from that in the EM. Instead of computing a conditional expectation typical of the E step, the algorithm performs maximization in (9). As a result, Algorithm 1 effectively consists solely of M steps for both parameters and z^*, characterizing it as a hard EM algorithm.

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

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:

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

Update the covariance matrix:

$$\Sigma^{(t)} = \frac{\sum_{a\in[k]}\sum_{j\in[n]}(Y_j - \theta_a^{(t)})(Y_j - \theta_a^{(t)})^T \mathbb{I}\left\{z^{(t-1)} = a\right\}}{n}. \tag{8}$$

Update the cluster assignment vector:

$$z_j^{(t)} = \operatorname*{argmin}_{a\in[k]} (Y_j - \theta_a^{(t)})^T (\Sigma^{(t)})^{-1} (Y_j - \theta_a^{(t)}), \quad \forall j \in [n]. \tag{9}$$

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 $\exp\big(-(1+o(1))\mathrm{SNR}^2/8\big)$ which matches 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, as commonly assumed in clustering literature [15, 8, 14]. The theorem also shows that the number of iterations needed to achieve the optimal rate is at most $\log n$, providing implementation guidance to practitioners.

Theorem 2.2. *Assume $k = O(1)$, $d = O(\sqrt{n})$ 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 SNR $\to \infty$ and $\lambda_d(\Sigma^*)/\lambda_1(\Sigma^*) = O(1)$. For Algorithm 1, suppose $z^{(0)}$ satisfies $\ell(z^{(0)}, z^*) = o(n)$ with probability at least $1 - \eta$. Then with probability at least*

$1-\eta-n^{-1}-\exp(-SNR)$, *we have*

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

We make the following remarks on the assumptions of Theorem 2.2. When k is constant, the assumption that SNR $\to \infty$ is a necessary condition for consistent recovery of z^*, as outlined in the minimax lower bound presented in Theorem 2.1. The assumption on Σ^* ensures that the covariance matrix is well-conditioned. The dimensionality d is assumed to be $O(\sqrt{n})$, a stronger assumption than in [15, 8, 14], where $d = O(n)$ is sufficient. This is because, unlike these studies, our work requires estimating the covariance matrix Σ^* and controlling the estimation error $\|\Sigma^{(t)} - \Sigma^*\|$.

Theorem 2.2 needs a decent initialization $z^{(0)}$ in the sense that it is sufficiently close to the ground truth such that $\ell(z^{(0)}, z^*) = o(n)$. This is because our theoretical analysis requires the initialization being within a specific proximity to the true parameters. The requirement $\ell(z^{(0)}, z^*) = o(n)$ can be fulfilled by simple procedures. An example is the vanilla Lloyd's algorithm whose performance is studied in [15, 8]. Though [15, 8] are for isotropic GMMs, their results can be extended to sub-Gaussian mixture models with nearly identical proof. Since ϵ_j are sub-Gaussian random variables with proxy variance $\lambda_d(\Sigma^*)$, [8] implies the vanilla Lloyd's algorithm output $\hat{z}$ satisfies $\ell(\hat{z}, z^*) \leq n\exp(-(1+o(1))\Delta^2/(8\lambda_d(\Sigma^*)))$ with probability at least $1-\exp(-\Delta/\sqrt{\lambda_d(\Sigma^*)}) - n^{-1}$, under the assumption that $\Delta^2/(k^2(kd/n+1)\lambda_d(\Sigma^*)) \to \infty$. Then we have $\ell(\hat{z}, z^*) = o(n)$ with high probability under the assumptions of Theorem 2.2, and hence it can be used as an initialization for the algorithm.

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 follows,

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 refer to this as *Model 2* throughout the paper to distinguish it from Model 1, as discussed in Section 2. The key difference between (10) and (1) is that here we have distinct covariance matrices $\{\Sigma^*_a\}_{a\in[k]}$ for each cluster, instead of a single shared Σ^*. We use the same loss function as defined in (2).

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

$$B_{a,b} = \Bigg\{ x \in \mathbb{R}^d : x^T \Sigma_a^{*\frac{1}{2}} \Sigma_b^{*-1} (\theta^*_a - \theta^*_b) + \frac{1}{2} x^T \left(\Sigma_a^{*\frac{1}{2}} \Sigma_b^{*-1} \Sigma_a^{*\frac{1}{2}} - I_d \right) x \\ \leq -\frac{1}{2}(\theta^*_a - \theta^*_b)^T \Sigma_b^{*-1} (\theta^*_a - \theta^*_b) + \frac{1}{2}\log|\Sigma^*_a| - \frac{1}{2}\log|\Sigma^*_b| \Bigg\}.$$

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

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

The form of SNR$'$ is closely connected to the testing error of the 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 $\text{SNR}'_{a,b} = \|(\theta^*_a - \theta^*_b)^T \Sigma^{*-\frac{1}{2}}\|$ for any $a, b \in [k]$ such that $a \neq b$. Hence, SNR$'$ = SNR and Model 2

effectively reduces to Model 1. (2) When $\Sigma_a^* = \sigma_a^2 I_d$ for any $a \in [k]$ where $\sigma_1, \ldots, \sigma_k > 0$ are large constants, we have $\text{SNR}'_{a,b}, \text{SNR}'_{b,a}$ 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 Model 2.

Theorem 3.1. *Assume* $d = O(1)$ *and* $\max_{a,b\in[k]} \lambda_d(\Sigma_a^*)/\lambda_1(\Sigma_b^*) = O(1)$. *Under the assumption* $\frac{SNR'}{\sqrt{\log k}} \to \infty$, *we have*

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

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

Although the statement of Theorem 3.1 appears similar to that of Theorem 2.1, the two minimax lower bounds differ due to the varying dependencies of the centers and covariance matrices on SNR' versus SNR. Using the same argument as in Section 2.2, the minimax lower bound established in Theorem 3.1 closely relates to the QDA between two normal distributions with different means and different covariance matrices.

Lemma 3.1 (Testing Error for the QDA)**.** *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*

$$\phi = \mathbb{I}\left\{\log|\Sigma_1^*| + (x-\theta_1^*)^T(\Sigma_1^*)^{-1}(x-\theta_1^*) \geq \log|\Sigma_2^*| + (x-\theta_2^*)^T(\Sigma_2^*)^{-1}(x-\theta_2^*)\right\}.$$

Assume $d = O(1)$ *and* $\max_{a,b\in\{1,2\}} \lambda_d(\Sigma_a^*)/\lambda_1(\Sigma_b^*) = O(1)$. *If* $\min\left\{SNR'_{1,2}, SNR'_{2,1}\right\} \to \infty$, *we have*

$$\inf_{\hat{\phi}}(\mathbb{P}_{\mathbb{H}_0}(\hat{\phi}=1) + \mathbb{P}_{\mathbb{H}_1}(\hat{\phi}=0)) \geq \exp\left(-(1+o(1))\frac{\min\left\{SNR'_{1,2}, SNR'_{2,1}\right\}^2}{8}\right).$$

Otherwise, $\inf_{\hat{\phi}}(\mathbb{P}_{\mathbb{H}_0}(\hat{\phi}=1) + \mathbb{P}_{\mathbb{H}_1}(\hat{\phi}=0)) \geq c$ *for some constant* $c > 0$.

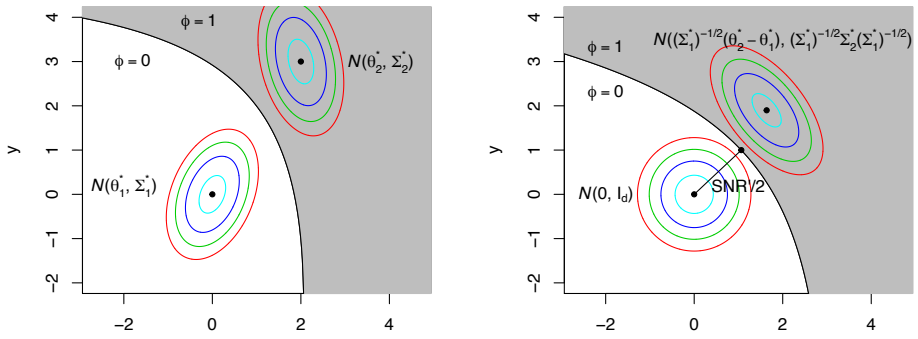

Figure 2: A geometric interpretation of SNR'.

Lemma 3.1 provides a geometric interpretation of SNR'. In the left panel of Figure 2, we have two normal distributions $\mathcal{N}(\theta_1^*, \Sigma_1^*)$ and $\mathcal{N}(\theta_2^*, \Sigma_2^*)$ from which X can be generated, and the black curve represents the optimal testing procedure ϕ, as detailed in Lemma 3.1. Since Σ_1^* is not necessarily equal to Σ_2^*, the black curve is not necessarily a straight line. If $\mathbb{H}_0$ is true, the probability that X is incorrectly classified occurs when X falls into the gray area, represented by $\mathbb{P}_{\mathbb{H}_0}(\phi = 1)$. To calculate this, we transform X to $X' = (\Sigma_1^*)^{-\frac{1}{2}}(X - \theta_1^*)$, standardizing the first distribution. Then, as 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 $\mathbb{I}\{X' \in B_{1,2}\}$. As a result, in the right panel of Figure 2, $B_{1,2}$ represents the space colored by gray, and the black curve is

its boundary. Then $\mathbb{P}_{\mathbb{H}_0}(\phi = 1)$ is equal to $\mathbb{P}(\mathcal{N}(0, I_d) \in B_{1,2})$. Under the assumption $d = O(1)$ and $\max_{a,b\in\{1,2\}} \lambda_d(\Sigma_a^*)/\lambda_1(\Sigma_b^*) = O(1)$, in Lemma C.10, we can show $\mathbb{P}(\mathcal{N}(0, I_d) \in B_{1,2}) = \exp(-(1 + o(1))\text{SNR}_{1,2}^{'2}/8)$. As a result, 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 propose a computationally feasible and rate-optimal procedure for clustering under Model 2. Similar to Algorithm 1, Algorithm 2 is a variant of Lloyd's algorithm, adjusted to accommodate 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 instead of a common one.

Algorithm 2: Adjusted Lloyd's Algorithm for Model 2.

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 assignment vector:

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

In Theorem 3.2, we give a computational and statistical guarantee for Algorithm 2. We demonstrate that, with proper initialization, Algorithm 2 achieves the minimax lower bound within $\log n$ iterations. The assumptions needed in Theorem 3.2 are similar to those in Theorem 2.2, except that we require stronger assumptions on the dimensionality d since now we have k (instead of one) covariance matrices to be estimated. In addition, by assuming $\max_{a,b\in[k]} \lambda_d(\Sigma_a^*)/\lambda_1(\Sigma_b^*) = O(1)$, we ensure not only that each of the k covariance matrices is well-conditioned but also that they are comparable to one another.

Theorem 3.2. *Assume $k, d = O(1)$ 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 $\text{SNR}' \to \infty$ and $\max_{a,b\in[k]} \lambda_d(\Sigma_a^*)/\lambda_1(\Sigma_b^*) = O(1)$. For Algorithm 2, suppose $z^{(0)}$ satisfies $\ell(z^{(0)}, z^*) = o(n)$ with probability at least $1 - \eta$. Then with probability at least $1 - \eta - 5n^{-1} - \exp(-\text{SNR}')$, we have*

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

The vanilla Lloyd's algorithm can be used as the initialization for Algorithm 2. This is because Model 2 is also a sub-Gaussian mixture model. By the same argument as in Section 2.3, the output of the vanilla Lloyd's algorithm $\hat{z}$ satisfies $\ell(\hat{z}, z^*) = o(n)$ with high probability under the assumptions of Theorem 3.2.

We conclude this section with a time complexity analysis of Algorithm 2. Compared to the vanilla Lloyd's algorithm, our method introduces additional computational overhead due to the need for

computing the inverse and determinant of covariance matrices. Specifically, the time complexity of Algorithm 2 is $O(nkd^3T)$. In contrast, the vanilla Lloyd's algorithm has a lower time complexity of $O(nkdT)$. The increase in complexity stems from matrix operations in d dimensions, as both matrix inversion and determinant computation scale as $O(d^3)$.

4 Numerical Studies

In this section, we compare the performance of the proposed methods with other popular clustering methods on synthetic and real 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 to each other with $\|\theta_1^*\| = \ldots = \|\theta_{30}^*\| = 9$. We consider four popular clustering methods: (1) the spectral clustering method in [14] (denoted as "spectral"), (2) the vanilla Lloyd's algorithm in [15] (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 the 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 $n = 1200$, $k = 2$, and $d = 9$. We set $\Sigma_1^* = I_d$ and $\Sigma_2^* = \Lambda_2$, a diagonal matrix where the first diagonal entry is 0.5 and the remaining entries are 5. We set the cluster sizes to be 900 and 300, respectively. To simplify the calculation of SNR$'$, we set $\theta_1^* = 0$ and $\theta_2^* = 5e_1$, with e_1 being the vector that has a 1 in its first entry and 0s elsewhere. The comparison is presented in the right panel of Figure 3.

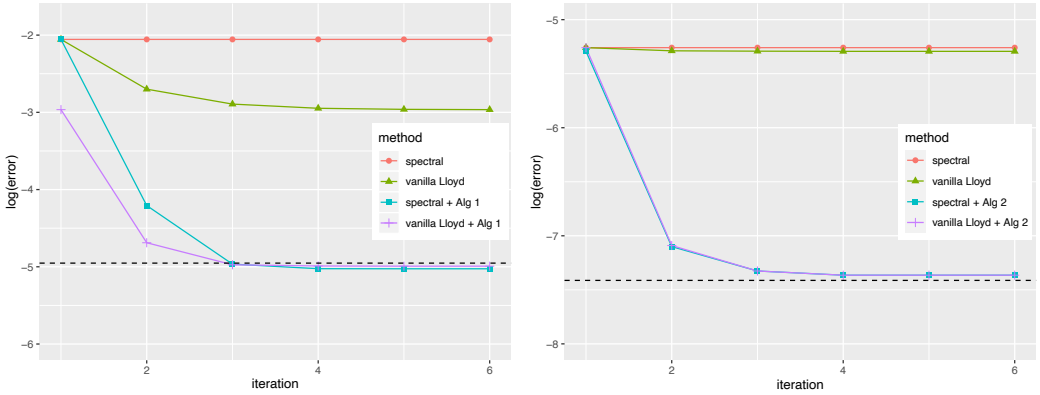

Figure 3: Left: Performance of Algorithm 1 compared with other methods under Model 1. Right: Performance of Algorithm 2 compared with other methods under Model 2.

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 panels represent the optimal exponents $-\text{SNR}^2/8$ and $-\text{SNR}'^2/8$ of the minimax bounds, respectively. It is observed that both Algorithm 1 and Algorithm 2 meet these benchmarks after three iterations. This justifies the conclusion that both algorithms are rate-optimal.

Real Data. To further demonstrate the effectiveness of our methods, we conduct experiments using the Fashion-MNIST dataset [23]. In the first analysis, we use a total of 12000 28×28 grayscale images, consisting of 6,000 images each from the T-shirt/top class and the Trouser class. The left

panel of Figure 4 gives a visualization of the data points using their first two principal components, showing the anisotropic and heterogeneous covariance structures. Since a large number of pixels have zero across most images, we apply PCA to reduce dimensionality from 784 to 50 by retaining the top 50 principal components. Our Algorithm 2 achieves a misclustering error of 5.71%, outperforming the vanilla Lloyd's algorithm, which has an error of 8.24%. In the second analysis, we incorporate an additional class, the Bag class, increasing the total to 18,000 images across three classes. Following the same preprocessing steps, the visualization of the dataset's structure in the right panel of Figure 4 again confirms the presence of anisotropic and heterogeneous covariances. Here, Algorithm 2 achieves an error of 3.97%, an improvement over the 5.67% error rate observed with the vanilla Lloyd's algorithm.

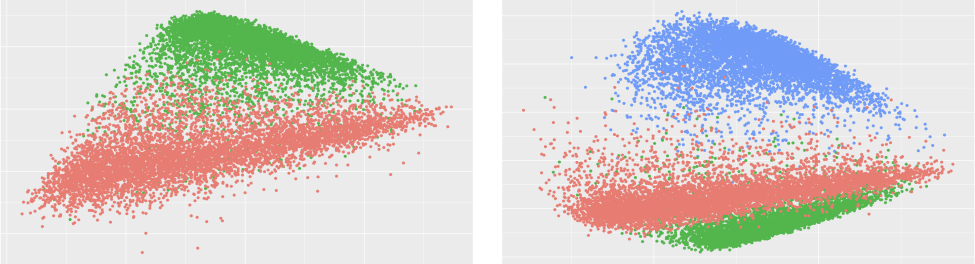

Figure 4: Visualization of the Fashion-MNIST dataset using the first two principal components. The data points are color-coded to indicate class membership: Red represents the T-shirt/top class, green denotes the Trouser class, and blue signifies the Bag class. This illustration shows the existence of anisotropic and heterogeneous covariance structures.

5 Conclusion

This paper focuses on clustering methods and theory for GMMs, with anisotropic covariance structures, 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. Despite these advances, our results have some limitations that are worth addressing in future work:

1. **High-Dimensional Settings:** Current results are restricted to dimensions d growing at a rate slower than n, specifically $d = O(\sqrt{n})$ as stated in Theorem 2.2. Theorem 3.2 further requires a stronger assumption $d = O(1)$. These constraints stem from technical challenges in estimating covariance matrices accurately and in controlling matrix determinant. Adopting more sophisticated analytical tools could potentially relax these bounds to $d = O(n)$. In scenarios where d exceeds n, the misclustering error deviates from the simpler exponential decay observed under isotropic GMMs, as shown in [16]. This suggests that our model might also exhibit similar complexities, warranting further exploration into the technique used in [16] for potential extensions.
2. **Ill-Conditioned Covariance Structures:** Our analysis relies on the assumption of well-conditioned covariance matrices, where $\max_{a,b\in[k]} \lambda_d(\Sigma_a^*)/\lambda_1(\Sigma_b^*) = O(1)$. This condition is crucial for the current analytical framework, as it helps manage the estimation errors of covariance matrices and their inverses. While more advanced techniques may allow for a relaxation of this assumption, handling ill-conditioned or degenerate covariance matrices remains challenging, particularly due to the difficulty of working with matrix inverses in such cases. While minimax lower bounds suggest that clustering is still possible even when the covariance matrix is degenerate, it raises computational challenges for our current algorithms. This highlights the need for developing new algorithms that can function effectively under less restrictive conditions.

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] Daniel Hsu, Sham Kakade, Tong Zhang, et al. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012.

[11] 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.

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

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

[14] 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.

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

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

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

[18] 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.

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

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

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

[22] 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.

[23] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

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