Universal lower bounds and optimal rates: Achieving minimax clustering error in sub-exponential mixture models

Maximilien Dreveton Alperen Gözeten Matthias Grossglauser Patrick Thiran

EPFL

July 1, 2024

Introduction

Clustering tasks of grouping n data points $X_1, \cdots, X_n$ in $\mathbb{R}^d$ into k clusters.

Mixture model

- $z \in [k]^n$ cluster labeling vector, family $\mathcal{F} = \{f_1, \cdots, f_k\}$ of pdf
- $\forall i \in [n]:\ X_i \mid z_i \sim f_{z_i}$

Statistical problem : recover z (up to a permutation) based on the observation of X only (we also assume k is known). Let $\hat{z} = \hat{z}(X)$ be an estimator of z. We define the *loss* of $\hat{z}$ as

$$\mathrm{loss}(z, \hat{z}) = \min_{\tau \in \mathrm{Sym}(k)} \frac{1}{n} \sum_{u=1}^{n} \mathbb{1}\{z_u \neq \tau(\hat{z}_u)\},$$

where $\mathrm{Sym}(k)$ is the group of permutations of $[k]$ (we can only recover the *partition*, not the *labels*).

Minimax rate:

$$\inf_{\hat{z}} \sup_{z \in [k]^n} \mathbb{E}_{X \sim \mathrm{MM}(z, f_1, \cdots, f_k)} \left[\mathrm{loss}(\hat{z}, z)\right]$$

Table of contents

Isotropic Gaussian mixture models (GMM): $X_i \mid z_i \sim \mathrm{Nor}(\mu_{z_i}, \sigma^2 I_d)$

Theorem 1 (Lu and Zhou, 2016: minimax rate in isotropic GMM)

Let $\Delta = \min_{a \neq b} \|\mu_a - \mu_b\|_2$. *Suppose* $\frac{\Delta}{\sigma \log(k)} \gg 1$. *Then,*

$$\inf_{\hat{z}} \sup_{z \in \mathcal{Z}_{n,k,\beta}} \mathbb{E}_{X \sim \mathrm{GMM}(z, \mu_1, \cdots, \mu_k)} [\mathrm{loss}(\hat{z}, z)] \asymp \exp\left(-(1 + o(1)) \frac{\Delta^2}{8\sigma^2} \right).$$

If $\frac{\Delta}{\sigma} + \log(k) = O(1)$, *then* $\inf_{\hat{z}} \sup_{z \in \mathcal{Z}_{n,k,\beta}} \mathbb{E}_{X \sim \mathrm{GMM}(z, \mu_1, \cdots, \mu_k)} [\mathrm{loss}(\hat{z}, z)] \geq c$ *for some constant* $c > 0$.

Rate optimal algorithms: Lloyd's algorithm (Lu & Zhou, 2016); spectral clustering (Löffler, Zhang & Zhou, 2021) (assuming $d \lesssim n$).

Recall: $\inf_{\hat{z}} \sup_{z \in \mathcal{Z}_{n,k,\beta}} \mathbb{E}_{X \sim \mathrm{GMM}(z,\mu_1,\cdots,\mu_k)} [\mathrm{loss}(\hat{z}, z)] \asymp e^{-(1+o(1))\frac{\mathrm{SNR}^2}{8}}$ where $\mathrm{SNR} = \frac{\min_{a \neq b} \|\mu_a - \mu_b\|}{\sigma}$.

GMM with Homogeneous Covariance Matrices: $X_i \mid z_i \sim \mathrm{Nor}(\mu_{z_i}, \Sigma)$

Then $\Sigma^{-1/2} X_i \sim \mathrm{Nor}(\Sigma^{-1/2}\mu_{z_i}, I_d)$, and the SNR exponent in the minimax rate becomes:

$$\min_{a \neq b} \|\Sigma^{-1/2}(\mu_a - \mu_b)\|_2 = \min_{a \neq b} \|\mu_a - \mu_b\|_\Sigma \quad \text{(Mahalanobis distance).}$$

Minimax rates in Gaussian mixture models

From isotropic to anisotropic GMM

Recall: $\inf_{\hat{z}} \sup_{z \in \mathcal{Z}_{n,k,\beta}} \mathbb{E}_{X \sim \mathrm{GMM}(z,\mu_1,\cdots,\mu_k)} [\mathrm{loss}(\hat{z}, z)] \asymp e^{-(1+o(1))\frac{\mathrm{SNR}^2}{8}}$ where $\mathrm{SNR} = \frac{\min_{a \neq b} \|\mu_a - \mu_b\|}{\sigma}$.

GMM with Homogeneous Covariance Matrices: $X_i \mid z_i \sim \mathrm{Nor}(\mu_{z_i}, \Sigma)$
Then $\Sigma^{-1/2} X_i \sim \mathrm{Nor}(\Sigma^{-1/2}\mu_{z_i}, I_d)$, and the SNR exponent in the minimax rate becomes:

$$\min_{a \neq b} \|\Sigma^{-1/2}(\mu_a - \mu_b)\|_2 = \min_{a \neq b} \|\mu_a - \mu_b\|_\Sigma \quad \text{(Mahalanobis distance).}$$

GMM with inhomogeneous Covariance Matrices: $X_i \mid z_i \sim \mathrm{Nor}(\mu_{z_i}, \Sigma_{z_i})$
Chen and Zhang, 2021 show that the SNR should be replaced by $\min_{a \neq b} \mathrm{SNR}'_{a,b}$

Minimax rates in Gaussian mixture models

From isotropic to anisotropic GMM

Recall: $\inf_{\hat{z}} \sup_{z \in \mathcal{Z}_{n,k,\beta}} \mathbb{E}_{X \sim \mathrm{GMM}(z,\mu_1,\cdots,\mu_k)} [\mathrm{loss}(\hat{z}, z)] \asymp e^{-(1+o(1))\frac{\mathrm{SNR}^2}{8}}$ where $\mathrm{SNR} = \frac{\min_{a \neq b} \|\mu_a - \mu_b\|}{\sigma}$.

GMM with Homogeneous Covariance Matrices: $X_i \mid z_i \sim \mathrm{Nor}(\mu_{z_i}, \Sigma)$

Then $\Sigma^{-1/2} X_i \sim \mathrm{Nor}(\Sigma^{-1/2}\mu_{z_i}, I_d)$, and the SNR exponent in the minimax rate becomes:

$$\min_{a \neq b} \|\Sigma^{-1/2}(\mu_a - \mu_b)\|_2 = \min_{a \neq b} \|\mu_a - \mu_b\|_\Sigma \quad \text{(Mahalanobis distance).}$$

GMM with inhomogeneous Covariance Matrices: $X_i \mid z_i \sim \mathrm{Nor}(\mu_{z_i}, \Sigma_{z_i})$

Chen and Zhang, 2021 show that the SNR should be replaced by $\min_{a \neq b} \mathrm{SNR}'_{a,b}$ where

$$\mathrm{SNR}'_{a \neq b} = 2 \min_{x \in \mathcal{B}_{ab}} \|x\|$$

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

From isotropic to anisotropic GMM

Where does this come from?

Main idea: for *each* data point X_i, we test $X_i \sim \mathrm{Nor}(\mu_1, \Sigma_1)$ versus $X_i \sim \mathrm{Nor}(\mu_2, \Sigma_2)$.

Lemma 1 (Testing Error for Quadratic Discriminant Analysis (Chen & Zhang, 2021))

Consider two hypotheses $H_0 : Y \sim \mathrm{Nor}(\mu_1, \Sigma_1)$ and $H_1 : Y \sim \mathrm{Nor}(\mu_2, \Sigma_2)$. Define a testing procedure

$$\phi(x) = \mathbb{1}\{\log f_1(x) < \log f_2(x)\} = \mathbb{1}\left\{\log|\Sigma_1| + (x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) \geq \log|\Sigma_2| + (x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right\}.$$

Then $\inf_{\hat\phi}(\mathbb{P}_{H_0}(\hat\phi = 1) + \mathbb{P}_{H_1}(\hat\phi = 0)) = \mathbb{P}_{H_0}(\phi = 1) + \mathbb{P}_{H_1}(\phi = 0)$ *(Neyman-Pearson).*
If $\min\{\mathrm{SNR}'_{1,2}, \mathrm{SNR}'_{2,1}\} \to \infty$*, we have*

$$\mathbb{P}_{H_0}(\phi = 1) + \mathbb{P}_{H_1}(\phi = 0) \asymp e^{-(1+o(1))\frac{\left(\min\{\mathrm{SNR}'_{1,2}, \mathrm{SNR}'_{2,1}\}\right)^2}{8}}.$$

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

Proof: complicated computations.
Geometric interpretation: $\approx$ okay

Table of contents

From isotropic to anisotropic GMM

Hypothesis testing: standard setting

Let $\mathcal{Y} = (Y_1, \cdots, Y_n)$ and test $H_0 : \mathcal{Y} \sim f^{\otimes n}$ versus $H_1 : \mathcal{Y} \sim g^{\otimes n}$.
If $f \neq g$ are *independent* of n, we have

$$\inf_{\hat{\phi}} (\mathbb{P}_{H_0}(\hat{\phi} = 1) + \mathbb{P}_{H_1}(\hat{\phi} = 0)) \asymp e^{-(1+o(1))\, n \, \mathrm{Chernoff}(f,g)}$$

where we define the *Chernoff information* as

$$\mathrm{Chernoff}(f, g) = -\log \inf_{t \in (0,1)} \int f^t(x) g^{1-t}(x) dx.$$

(Note: $\mathrm{Chernoff}(f^{\otimes n}, g^{\otimes n}) = n\, \mathrm{Chernoff}(f, g)$).
Key observation: $\mathbb{P}_{H_1}\left(\log \frac{f}{g}(x) > 0\right) = \mathbb{P}\left(e^{t \log \frac{f}{g}(x)} > 1\right) \leq \mathbb{E}_g\left[e^{t \log \frac{f}{g}}\right] = \int f^t g^{1-t} \leq e^{-\mathrm{Chernoff}(f,g)}$.

Chernoff information between Gaussians

- $\mathrm{Chernoff}\left(\mathrm{Nor}(\mu_1, \sigma^2 I_d), \mathrm{Nor}(\mu_2, \sigma^2 I_d)\right) = \frac{\|\mu_1 - \mu_2\|_2^2}{8\sigma^2}$
- $\mathrm{Chernoff}\left(\mathrm{Nor}(\mu_1, \Sigma), \mathrm{Nor}(\mu_2, \Sigma)\right) = \frac{1}{8}\|\Sigma^{-1/2}(\mu_1 - \mu_2)\|_2^2$
- $\mathrm{Chernoff}\left(\mathrm{Nor}(\mu_1, \Sigma_1), \mathrm{Nor}(\mu_2, \Sigma_2)\right)$ still complicated

Provide another interpretation of SNRs.

Minimax rates in mixture models

Chernoff information

Mixture model (MM): $X_i \mid z_i \sim f_{z_i}$ where $\mathcal{F} = \{f_1, \cdots, f_k\}$ is a family of pdf.
Define

$$\mathrm{Chernoff}(\mathcal{F}) \; = \; \min_{1 \le a \neq b \le k} \mathrm{Chernoff}(f_a, f_b).$$

Theorem 2 (Dreveton, Gözeten, Grossglauser, Thiran, 2024)

Suppose $\mathrm{Chernoff}(\mathcal{F}) \gg \log k$. *Then,*

$$\min_{\hat{z}} \max_{z \in \mathcal{Z}_{n,\beta}} \mathbb{E}_{X \sim \mathrm{MM}(f_1, \cdots, f_k)} \left[\mathrm{loss}(z, \hat{z}) \right] \; = \; e^{-(1+o(1))\mathrm{Chernoff}(\mathcal{F})}$$

Algorithm 1: Clustering mixture models (known pdf).

Input: Set of n data points $(X_1, \cdots, X_n) \in \mathcal{X}^n$, number of clusters k, family $\mathcal{F} = \{f_1, \cdots, f_k\}$ of pdfs.
Output: Predicted clusters $\hat{z} \in [k]^n$.
For $i = 1, \cdots, n$ let $\hat{z}_i^{(t)} = \arg\max_{a \in [k]} \log f_a(X_i)$.
Return: $\hat{z} = \hat{z}^{(t_{\max})}$.

Table of contents

Laplace mixture model

Algorithm 2: Lloyd-type algorithm for clustering parametric mixture models.

Input: Set of n data points $(X_1, \cdots, X_n) \in \mathcal{X}^n$, parametric family $\mathcal{P}_\Theta = \{f_\theta, \theta \in \Theta\}$ of pdfs, number of clusters k, number of iteration $t_{\max}$, initial clustering $\hat{z}^{(0)} \in [k]^n$.

For $t = 1 \cdots t_{\max}$ **do**

1. For $a = 1, \cdots, k$, let $\hat{\theta}_a^{(t)} = \hat{\theta}\left(\{X_i \colon \hat{z}_i^{(t-1)} = a\}\right)$ be an estimate of θ_a;
2. For $i = 1, \cdots, n$ let $\hat{z}_i^{(t)} = \arg\max_{a \in [k]} \log f_{\hat{\theta}_a^{(t)}}(X_i)$.

Return: $\hat{z} = \hat{z}^{(t_{\max})}$.

Previous work: Show that Algorithm 2 attain the minimax rate in *sub-gaussian* mixture models
Our contribution: *sub-exponential* tails instead of sub-gaussian

Laplace mixture model: $\forall \ell \in [d] \colon X_{i\ell} = \mu_{z_i \ell} + \sigma_{z_i \ell} \epsilon_{i\ell}$ where $\epsilon_{i\ell} \sim \mathrm{Lap}(0, 1)$ (pdf $f(x) = \frac{1}{2} e^{-|x|}$).

Estimate mean and variance as:

$$\hat{\mu}(Y_1, \cdots, Y_m) = \frac{1}{m} \sum_{i=1}^{m} Y_i \quad \text{and} \quad \hat{\sigma}(Y_1, \cdots, Y_m) = \frac{1}{m} \sum_{i=1}^{m} |Y_i - \hat{\mu}(Y_1, \cdots, Y_m)|.$$

Theorem 3 (Dreveton, Gözeten, Grossglauser, Thiran, 2024)

Suppose $\sum_{i=1}^{n} \mathbb{1}\{z_i = a\} \geq \alpha n/k$ *for some constant* $\alpha > 0$, $d = \Theta(1)$, $\sigma_{a\ell} = \Theta(1)$ *and* $\|\mu_a - \mu_b\|_1 = \Theta(d\rho_n)$ *with* $\rho_n \gg \sqrt{k}$ *and* $\mathrm{loss}(z, \hat{z}^{(0)}) \ll 1/(k\rho_n)$. *Then, the output* $\hat{z}$ *of Algorithm 2 after* $\Omega(\log n)$ *iterations verifies*

$$\mathrm{loss}(z, \hat{z}) \leq e^{-(1+o(1))\mathrm{Chernoff}(\mathcal{F})}.$$

Remarks:

- We also show that $\mathrm{loss}(z, \hat{z}^{(0)}) \ll 1/(k\rho_n)$ can be attained by spectral clustering
- If $\sigma_{1\ell} = \sigma_{2\ell} = \cdots = \sigma_{k\ell}$, then $\mathrm{Chernoff}(\mathcal{F}) = \min_{1 \leq a \neq b \leq k} \|\Sigma^{-1}(\mu_a - \mu_b)\|_1$
- Similar results for other mixture models (such as exponential family mixtures) under sub-exponential assumptions

Table of contents

Conclusion

Summary:

1. Minimax rates in mixture models: Chernoff information is the key quantity
2. Lloyd-type algorithm attain the minimax rate when parameters (mean, variance) are unknown (in low dimension) and pdf have sub-exponential tails.

Possible extensions:

- Mixture models in high dimension ($d \gg n$): if parameter are unknown, minimax rates are different. Isotropic Gaussian done? (Ndaoud, 2022); (Even, Giraud & Verzelen, 2024)
- Mixture models with tails heavier than sub-exponential
- Robustness to perturbations: mixture + random noise, mixture + adversary, mixture + outliers
- (Semi)-supervised rates (Lelarge & Miolane, 2019; Tifrea et al., 2024)