Minimax Rates of Clustering Mixture Models and Stochastic Block Models

Maximilien Dreveton

EPFL

June 27, 2024

Table of contents

INTRODUCTION

TWO SIMILAR PROBLEMS

Clustering and **community detection**: tasks of grouping objects together.

- **Clustering**: objects are n points $X_1, \cdots, X_n$ in $\mathbb{R}^d$.
- **Community detection**: objects are the n vertices of a graph with adjacency matrix $A \in \{0,1\}^{n\times n}$.

Toy-models : $z \in [k]^n$ cluster labeling vector.

- (isotropic) **Gaussian mixture model** (GMM): $X_i \mid z_i \sim \mathrm{Nor}(\mu_{z_i}, \sigma^2 I_d)$;
- (homogeneous) **stochastic block model** (SBM): $A_{ij} \mid z_i, z_j \sim \begin{cases} \mathrm{Ber}(p) & \text{if } z_i = z_j, \\ \mathrm{Ber}(q) & \text{otherwise.} \end{cases}$ for $i < j$ and $A_{ij} = A_{ji}$.

Statistical problem : recover z (up to a permutation) based on the observation of X or A only (we also assume k is known).

Introduction

Minimax rates in these two problems (1)

For any $z \in [k]^n$, denote $n_a(z) = \sum_{u \in [n]} \mathbb{1}\{z_u = a\}$ the size of cluster $a \in [k]$. Let $\beta > 1$ and define

$$\mathcal{Z}_{n,k,\beta} \; = \; \left\{ z \in [k]^n \colon n_a(z) \in \left[\frac{n}{\beta k}, \beta \frac{n}{k} \right] \forall a \in [k] \right\}.$$

Let $\hat{z}$ 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*).

Aim: study the *expected loss* $\mathbb{E}\left[\mathrm{loss}(\hat{z}, z)\right]$ of an estimator $\hat{z}$ (where expectation is taken with respect to X or A being generated from GMM or SBM.

Intuitive: difficulty of estimating z is governed by :

- for GMM: separation between the centers $\mu_1, \cdots, \mu_k$ (assuming σ fixed);
- for SBM: difference between $\mathrm{Ber}(p)$ and $\mathrm{Ber}(q)$.

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,*

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

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,*

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

Theorem 2 (Zhang and Zhou, 2016: minimax rate in homogeneous SBM)

Let $I = \mathrm{Ren}_{1/2}(\mathrm{Ber}(p), \mathrm{Ber}(q)) = -2 \log\left(\sqrt{pq} + \sqrt{(1-p)(1-q)}\right)$ *the Rényi divergence of order 1/2 between two Bernoulli distributions. Suppose* $\beta \in (1, \sqrt{2})$, *and let* $q < p$. *If* $\frac{nI}{k \log k} \gg 1$ *we have*

$$\inf_{\hat{z}} \sup_{z \in \mathcal{Z}_{n,k,\beta}} \mathbb{E}_{G \sim \mathrm{SBM}(z,p,q)} [\mathrm{loss}(\hat{z}, z)] \asymp \begin{cases} \exp\left(-(1 + o(1)) \frac{nI}{2}\right) & \text{if } k = 2, \\ \exp\left(-(1 + o(1)) \frac{nI}{\beta k}\right) & \text{if } k \geq 3. \end{cases}$$

Rate optimal algorithms:

- GMM: Lloyd's algorithm (Lu & Zhou, 2016); spectral clustering (Löffler et al., 2021);
- SBM: MLE (Zhang & Zhou, 2016); two-stage algorithms (Gao et al., 2017); semidefinite programs (Fei & Chen, 2018); VEM (Zhang & Zhou, 2020); spectral clustering (Zhang, 2023).

Table of contents

Table of contents

FROM ISOTROPIC TO ANISOTROPIC GMM

MINIMAX RATES: NEW SNRS

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

From isotropic to anisotropic GMM

Minimax rates: new SNRs

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

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

From isotropic to anisotropic GMM

Minimax rates: new SNRs

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

GMM with Homogeneous Covariance Matrices: $X_i \mid z_i \sim \text{Nor}(\mu_{z_i}, \Sigma)$
Then $\Sigma^{-1/2} X_i \sim \text{Nor}(\Sigma^{-1/2}\mu_{z_i}, I_d)$, 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: $X_i \mid z_i \sim \text{Nor}(\mu_{z_i}, \Sigma_{z_i})$
Chen and Zhang, 2021 show that the SNR should be replaced by $\min_{a \neq b} \text{SNR}'_{a,b}$

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

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

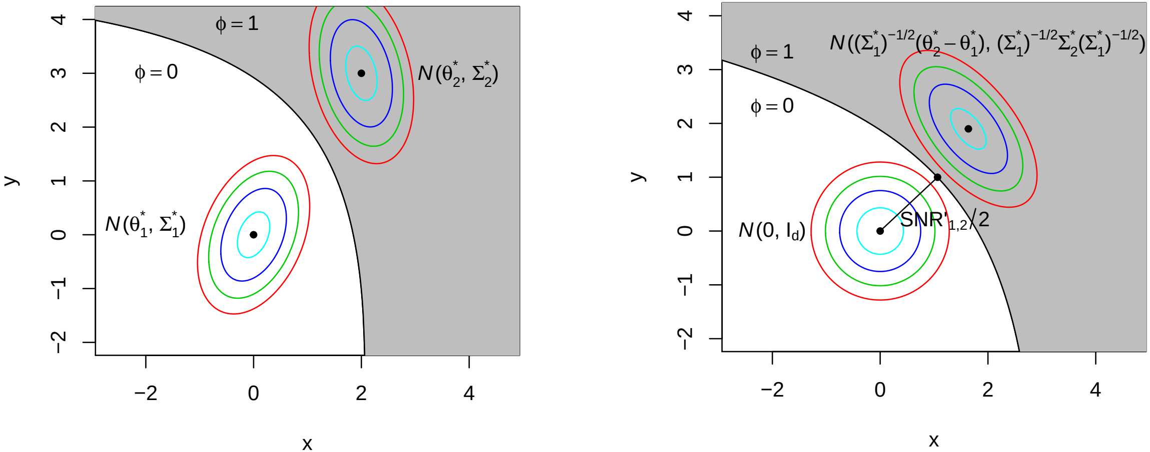
GMM with inhomogeneous Covariance Matrices: $X_i \mid z_i \sim \mathrm{Nor}(\mu_{z_i}, \Sigma_{z_i})$

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

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

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

From isotropic to anisotropic GMM



(a) Original Gaussians with optimal decision boundary

(b) After transformation $Y = \Sigma_1^{-1/2}(Y - \theta_1)$.

Figure. A geometric interpretation of SNR'_{12} (taken from Chen and Zhang, 2021).

From isotropic to anisotropic GMM

Where does this come from?

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.
Geometric interpretation: $\approx$ okay

From isotropic to anisotropic GMM

Another interpretation

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

Remark

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

Table of contents

Minimax rates in mixture models

(Non-gaussian) mixture models

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

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

Theorem 3 (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 $(X_1, \cdots, X_n) \in \mathcal{X}^n$, family $\mathcal{F} = \{f_1, \cdots, f_k\}$ of pdfs, number of clusters k.
Output: Predicted clusters $\hat{z} \in [k]^n$.
For $i = 1, \cdots, n$ let $\hat{z}_i^{(t)} = \arg\max_{a \in [k]} \log f(X_i)$.
Return: $\hat{z} = \hat{z}^{(t_{\max})}$.

Table of contents

Laplace mixture model

Algorithm 2: 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 $\hat{z}^{(0)} \in [k]^n$.

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

1. For $a = 1, \cdots, k$, let $\hat{\theta}_a^{(t)} = \hat{\theta}\left(\{X_i : \hat{z}_i^{(t-1)} = a\}\right)$ 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: $\hat{z} = \hat{z}^{(t_{\max})}$.

Laplace mixture model : $X_{i\ell} = \mu_{z_i \ell} + \sigma_{z_i \ell} \epsilon_{i\ell}$ where $\epsilon_{i\ell} \sim \mathrm{Lap}(0, 1)$.

Novelty : (sub-)exponential tails instead of sub-gaussian.

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 4 (Dreveton, Gözeten, Grossglauser, Thiran, 2024)

Suppose $d = \Theta(1)$, $\sigma_{a\ell} = \Theta(1)$ *and* $\|\mu_a - \mu_b\|_1 = \Theta(\rho_n)$ *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:

- $\mathrm{loss}(z, \hat{z}^{(0)}) \ll 1/(k\rho_n)$ can be attained by spectral clustering.
- If $\sigma_{a\ell} = \sigma_{b\ell}$, 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

Table of contents

Stochastic Block Model (SBM)

Original definition vs Modern definition

Definition 3. Let $p(\boldsymbol{x})$ be the probability function for a stochastic multigraph, and let $\{B_1, \ldots, B_t\}$ be a partition of the nodes into mutually exclusive and exhaustive subsets called node-blocks. We say that $p(\boldsymbol{x})$ is a stochastic blockmodel with respect to the partition $\{B_1, \ldots, B_t\}$ if and only if

(1) the random vectors $\boldsymbol{X}_{ij}$ are statistically independent; and

(2) for any nodes $i \neq j$ and $i' \neq j'$, if i and i' are in the same node-block and j and j' are in the same node-block, then the random vectors $\boldsymbol{X}_{ij}$ and $\boldsymbol{X}_{i'j'}$ are identically distributed.

Figure. Original definition of a SBM by (Holland et al., 1983).

Stochastic Block Model

New definition

- 'Modern' definition of SBM restricts interactions (edges) to belong to $\{0, 1\}$;
- Here: generalisation of the 'old' definition. Interactions take value in a space $\mathcal{S}$.
  - Examples: multiplex networks ($\mathcal{S} = \{0, 1\}^M$), weighted networks ($\mathcal{S} = \mathbb{R}_+$), signed networks ($\mathcal{S} = \{0, -, +\}$), censored networks ($\mathcal{S} = \{\mathrm{unobserved}, \mathrm{observed\&present}, \mathrm{observed\&absent}\}$).

SBM with edge covariates : Let f and g be two pdf on $\mathcal{S}$. Conditionally on z, we observe $A \in \mathcal{S}^{n \times n}$ such that $A_{ij} = A_{ji}$ is sampled from f if $z_i = z_j$, and from g otherwise. We note $A \sim \mathrm{SBM}(z, f, g)$.

Example: 'modern' SBM has $\mathcal{S} = \{0, 1\}$, $f = \mathrm{Ber}(p)$ and $g = \mathrm{Ber}(q)$.

Define the *Rényi divergence* of order $1/2$ between f and g as

$$\mathrm{Ren}_{1/2}(f, g) \;=\; -2 \log \int \sqrt{\frac{df}{d\mu}} \sqrt{\frac{dg}{d\mu}} d\mu,$$

where μ is an arbitrary measure which dominates f and g.

Table of contents

Theorem 5 (Avrachenkov et al., 2020)

Suppose $\beta \in (1, \sqrt{2})$, and let $I = \mathrm{Ren}_{1/2}(f, g)$. If $\frac{nI}{k \log k} \gg 1$, we have

$$\inf_{\hat{z}} \sup_{z \in \mathcal{Z}_\beta} \mathbb{E}_{X \sim \mathrm{SBM}(z,f,g)} \left[\mathrm{loss}(\hat{z}, z) \right] \asymp \begin{cases} \exp\left(-(1+o(1)) \frac{nI}{2}\right) & \text{if } k = 2, \\ \exp\left(-(1+o(1)) \frac{nI}{\beta k}\right) & \text{if } k \geq 3. \end{cases}$$

Furthermore, if $\frac{nI}{k} = O(1)$ then $\inf_{\hat{z}} \sup_{z \in Z_\beta} \mathbb{E}\left[\mathrm{loss}(\hat{z}, z)\right] \geq c$ for some constant $c > 0$.

Remarks

- Assumes f and g are known by the algorithm
- If f and g are unknown: results in (Xu et al., 2020) but with many additional technical conditions

Questions

- Here the Rényi-divergence is the key quantity. Why?
- Difference between 2 and more than 2 clusters

Minimax rate in homogeneous SBM

Two versus more than two communities

Two communities

- If the two communities are of different sizes (for example $n_1 > n_2$), then nodes in the community 1 have a higher expected degree than nodes in the community 2
- Hence the worst setting is when the two communities are of the same size
- The $n/2$ in the exponential error rate $e^{-(1+o(1))\frac{n}{2}I}$ represent the community sizes

Three (or more) communities

- One could think that having $k = 3$ communities of size n/k would be the worse, leading to an error rate of $e^{-(1+o(1))\frac{n}{k}I}$
- But, the worst case is two small communities of size $\frac{n}{\beta k}$ and one big of size $n - 2\frac{n}{\beta k}$. This leads to the minimax rate of $e^{-(1+o(1))\frac{n}{\beta k}I}$

Minimax rates in homogeneous SBM

Why Rényi divergence? (1)

Setting: $n+1$ nodes, two communities of sizes $n/2$ and $n/2+1$; f and g denote the pdf for intra- and inter-cluster interactions.
Nodes $1, \cdots, n/2$ in community 1; nodes $n/2+1, \cdots, n$ in community 2. The last node $n+1$ belongs either to community 1 or 2.

Fundamental Testing Problem: A genie gives you $z = (\underbrace{1, \cdots, 1}_{n/2}, \underbrace{2, \cdots, 2}_{n/2}, ?)$. You have to find z_{n+1}.

Denote $X = (A_{n+1,1}, A_{n+1,2}, \cdots, A_{n+1,n}) \in \mathcal{S}^n$ (X_j denotes interaction between nodes $n+1$ and j) and the two hypothesis:

$$H_1 : z_{n+1} = 1 \quad \text{vs} \quad H_2 : z_{n+1} = 2.$$

Under H_1: $X \sim f^{\otimes n/2} \otimes g^{\otimes n/2} =: h_1$,
Under H_2: $X \sim g^{\otimes n/2} \otimes f^{\otimes n/2} =: h_2$.

$$\text{MLE: } \phi_{\mathrm{MLE}}(X) = \begin{cases} H_1 & \text{if } h_1(X) > h_2(X) \\ H_2 & \text{if } h_1(X) \leq h_2(X). \end{cases}$$

Guarantee of MLE? Classic Chernoff–Stein theory of hypothesis testing applies for f and g independent of n (Cover & Thomas, 1999). But generalisation is possible.

WHY RÉNYI DIVERGENCE? (2)

Under H_1: $X \sim f^{\otimes n/2} \otimes g^{\otimes n/2} =: h_1$,
Under H_2: $X \sim g^{\otimes n/2} \otimes f^{\otimes n/2} =: h_2$.

MLE: $\phi_{\mathrm{MLE}}(X) = \begin{cases} H_1 & \text{if } h_1(X) > h_2(X) \\ H_2 & \text{if } h_1(X) \le h_2(X). \end{cases}$

Let $\mathrm{Ren}_t(f,g) = -(1-t)^{-1} \log \int f^t(x) g^{1-t}(x) dx$ be the Rényi divergence of order t between two pdf f and g, and define the *Chernoff information*

$$\mathrm{Chernoff}(h_1, h_2) = \sup_{t \in (0,1)} (1-t) \mathrm{Ren}_t(h_1, h_2).$$

Lemma 2 (Gao, Ma, Zhang, Zhou, 2018; Dreveton et al., 2024)

The worst-case error of $\phi \colon X \mapsto \phi(X) \in \{H_1, H_2\}$ *is* $r(\phi) = \max\left\{\mathbb{P}_{H_1}(\phi(X) = H_2) ; \mathbb{P}_{H_2}(\phi(X) = H_1)\right\}$.
We have $\inf_\phi r(\phi) = r(\phi_{\mathrm{MLE}})$. *Moreover, if* $\mathrm{Chernoff}(h_1, h_2) \gg 1$ *we have*

$$r(\phi_{\mathrm{MLE}}) = e^{-(1+o(1)) \mathrm{Chernoff}(h_1, h_2)}.$$

Minimax rates in homogeneous SBM

Why Rényi divergence? (3)

Under H_1: $X \sim f^{\otimes n/2} \otimes g^{\otimes n/2} =: h_1$,
Under H_2: $X \sim g^{\otimes n/2} \otimes f^{\otimes n/2} =: h_2$.

MLE: $\phi_{\mathrm{MLE}}(X) = \begin{cases} H_1 & \text{if } h_1(X) > h_2(X) \\ H_2 & \text{if } h_1(X) \leq h_2(X). \end{cases}$

Final ingredient:

$$
\begin{aligned}
\mathrm{Chernoff}(h_1, h_2) &= \sup_{t \in (0,1)} (1-t)\mathrm{Ren}_t(\underbrace{f^{\otimes n/2} \otimes g^{\otimes n/2}}_{h_1}, \underbrace{g^{\otimes n/2} \otimes f^{\otimes n/2}}_{h_2}) \\
&= \sup_{t \in (0,1)} (1-t) \left[\sum_{i=1}^{n/2} \mathrm{Ren}_t(f,g) + \sum_{i=n/2+1}^{n} \mathrm{Ren}_t(g,f) \right] \quad \text{(linearity of Rényi divergence)} \\
&= \frac{n}{2} \sup_{t \in (0,1)} \left\{ (1-t)\mathrm{Ren}_t(f,g) + t\,\mathrm{Ren}_{1-t}(f,g) \right\} \quad \text{using } (1-t)\mathrm{Ren}_t(f,g) = t\,\mathrm{Ren}_{1-t}(g,f) \\
&= \frac{n}{2} \mathrm{Ren}_{1/2}(f,g).
\end{aligned}
$$

Zero-inflated distribution : Suppose that the distributions f and g can be written as follows

$$f(x) \;=\; (1 - a\rho_n)\delta_0(x) + a\rho_n\tilde{f}(x) \quad \text{and} \quad g(x) \;=\; (1 - b\rho_n)\delta_0(x) + b\rho_n\tilde{g}(x), \tag{3.1}$$

When $\rho_n \ll 1$, the Rényi divergence $I = \mathrm{Ren}_{1/2}(f, g)$ between such zero-inflated distributions equals

$$I \;=\; (1 + o(1))\rho_n \left[\left(\sqrt{a} - \sqrt{b} \right)^2 + 2\sqrt{ab}\, \mathrm{Hel}^2(\tilde{f}, \tilde{g}) \right], \tag{3.2}$$

where $\mathrm{Hel}^2(\tilde{f}, \tilde{g}) \in [0, 1]$ is the *Hellinger divergence* defined by

$$\mathrm{Hel}^2(f, g) \;=\; \frac{1}{2} \int \left(\sqrt{\frac{df}{d\mu}} - \sqrt{\frac{dg}{d\mu}} \right)^2 d\mu.$$

Zero-inflated distribution : Suppose that the distributions f and g can be written as follows

$$f(x) \;=\; (1 - a\rho_n)\delta_0(x) + a\rho_n \tilde{f}(x) \quad \text{and} \quad g(x) \;=\; (1 - b\rho_n)\delta_0(x) + b\rho_n \tilde{g}(x), \tag{3.1}$$

When $\rho_n \ll 1$, the Rényi divergence $I = \mathrm{Ren}_{1/2}(f, g)$ between such zero-inflated distributions equals

$$I \;=\; (1 + o(1))\rho_n \left[\left(\sqrt{a} - \sqrt{b} \right)^2 + 2\sqrt{ab}\, \mathrm{Hel}^2(\tilde{f}, \tilde{g}) \right], \tag{3.2}$$

where $\mathrm{Hel}^2(\tilde{f}, \tilde{g}) \in [0, 1]$ is the *Hellinger divergence* defined by

$$\mathrm{Hel}^2(f, g) \;=\; \frac{1}{2} \int \left(\sqrt{\frac{df}{d\mu}} - \sqrt{\frac{dg}{d\mu}} \right)^2 d\mu.$$

Corollary [Exact recovery in sparse homogeneous SBM with edge covariates]

Consider an SBM with same-size communities and edge covariate distributions given in (3.1), where $\tilde{f}, \tilde{g}$ are independent of n and $\rho_n = \log n / n$. Then, exact recovery is

- solvable if $\left(\sqrt{a} - \sqrt{b} \right)^2 + 2\sqrt{ab}\, \mathrm{Hel}^2(\tilde{f}, \tilde{g}) > k$;
- unsolvable if $\left(\sqrt{a} - \sqrt{b} \right)^2 + 2\sqrt{ab}\, \mathrm{Hel}^2(\tilde{f}, \tilde{g}) < k$.

$\mathrm{Hel}^2(\tilde{f}, \tilde{g})$ characterises the additional information gained by observing the edge covariates.

Table of contents

CONCLUSION

Summary:

1. Similarity in the analysis of minimax rates of mixture model and stochastic block models
2. Chernoff information is the key quantity

Possible extensions:

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

References I

Avrachenkov, K., Dreveton, M., & Leskelä, L. (2020). **Community recovery in non-binary and temporal stochastic block models.** *arXiv preprint arXiv:2008.04790*.

Chen, X., & Zhang, A. Y. (2021). **Optimal clustering in anisotropic gaussian mixture models.** *arXiv preprint arXiv:2101.05402*.

Cover, T., & Thomas, J. (1999). ***Elements of information theory.*** John Wiley & Sons.

Dreveton, M., Gözeten, A., Grossglauser, M., & Thiran, P. (2024). **Universal lower bounds and optimal rates: Achieving minimax clustering error in sub-exponential mixture models.** *arXiv preprint arXiv:2402.15432*.

Even, B., Giraud, C., & Verzelen, N. (2024). **Computation-information gap in high-dimensional clustering.** *arXiv preprint arXiv:2402.18378*.

Fei, Y., & Chen, Y. (2018). **Exponential error rates of sdp for block models: Beyond grothendieck's inequality.** *IEEE Transactions on Information Theory*, *65*(1), 551–571.

Gao, C., Ma, Z., Zhang, A. Y., & Zhou, H. H. (2017). **Achieving optimal misclassification proportion in stochastic block models.** *Journal of Machine Learning Research*, *18*(60), 1–45.

Holland, P. W., Laskey, K. B., & Leinhardt, S. (1983). **Stochastic blockmodels: First steps.** *Social networks*, *5*(2), 109–137.

Lelarge, M., & Miolane, L. (2019). **Asymptotic bayes risk for gaussian mixture in a semi-supervised setting.** *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 639–643.

References II

Löffler, M., Zhang, A. Y., & Zhou, H. H. (2021).**Optimality of spectral clustering in the gaussian mixture model.** *The Annals of Statistics*, *49*(5), 2506–2530.

Lu, Y., & Zhou, H. H. (2016).**Statistical and computational guarantees of Lloyd's algorithm and its variants.** *arXiv preprint arXiv:1612.02099*.

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

Tifrea, A., Yüce, G., Sanyal, A., & Yang, F. (2024).**Can semi-supervised learning use all the data effectively? A lower bound perspective.** *Advances in Neural Information Processing Systems*, *36*.

Xu, M., Jog, V., & Loh, P.-L. (2020).**Optimal rates for community estimation in the weighted stochastic block model.** *Annals of Statistics*, *48*(1), 183–204.

Zhang, A. Y., & Zhou, H. H. (2016).**Minimax rates of community detection in stochastic block models.** *The Annals of Statistics*, *44*(5), 2252–2280. https://doi.org/10.1214/15-AOS1428

Zhang, A. Y., & Zhou, H. H. (2020).**Theoretical and computational guarantees of mean field variational inference for community detection.** *The Annals of Statistics*, *48*(5), 2575–2598.

Zhang, A. Y. (2023).**Fundamental limits of spectral clustering in stochastic block models.** *arXiv preprint arXiv:2301.09289*.