Minimax-Optimal Dimension-Reduced Clustering for High-Dimensional Nonspherical Mixtures

Yuqi Gu

yuqi.gu@columbia.edu
Department of Statistics, Columbia University

Workshop on Statistical Network Analysis and Beyond
Tokyo, Japan, June 2025

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

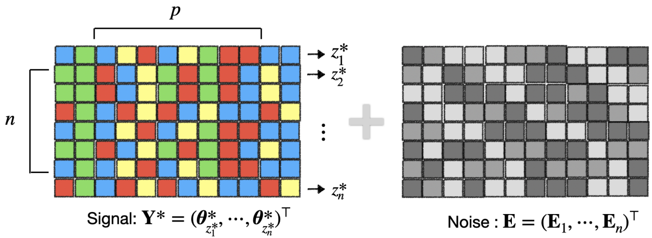
Joint Work with My PhD Student

Chengzhu Huang (Columbia Statistics)

Minimax-Optimal Dimension-Reduced Clustering for High-Dimensional Nonspherical Mixtures. *arXiv preprint* **arXiv:2502.02580.**

Chengzhu Huang and Yuqi Gu (2025+).

Clustering High-dimensional Data



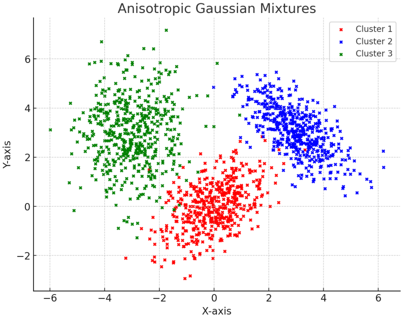
- Data: $\mathbf{Y}_{n \times p} = (\mathbf{y}_1, \cdots, \mathbf{y}_n)^\top$

$$\mathbf{y}_i = \boldsymbol{\theta}^*_{z_i^*} + \mathbf{E}_i \in \mathbb{R}^p, \quad i \in [n]$$

  cluster labels $z_i^* \in [K]$, centers $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_K^*$, mean-zero noise $\mathbf{E}_i \overset{\text{ind.}}{\sim} \mathcal{E}_{z_i^*}$

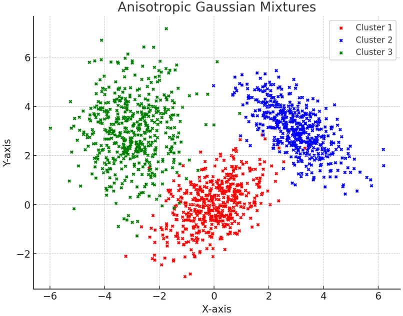
- Task: Recover the cluster labels $\mathbf{z}^* = (z_1^*, \cdots, z_n^*)$
- High-dimensional: $p \gg n$ may happen

Anisotropic/Nonspherical Mixtures



- Anisotropic/Nonspherical Mixtures: Noise is non-spherical in some clusters ($\mathrm{Cov}(\mathcal{E}_k) \neq \sigma^2 \mathbf{I}_p$)
- Widely observed in various real-world data

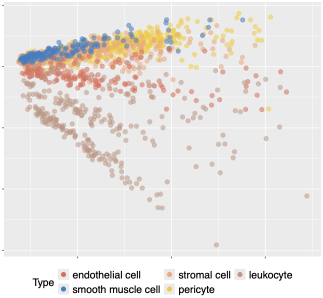
Anisotropic/Nonspherical Mixtures

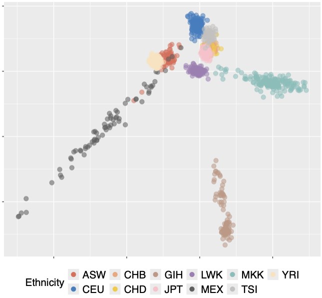


- Anisotropic/Nonspherical Mixtures: Noise is non-spherical in some clusters ($\mathrm{Cov}(\mathcal{E}_k) \neq \sigma^2 \mathbf{I}_p$)
- Widely observed in various real-world data

How to cluster adaptively and efficiently in high dimensions with $p \gg n$?

Examples of Nonspherical Mixtures

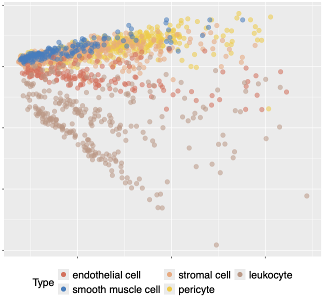


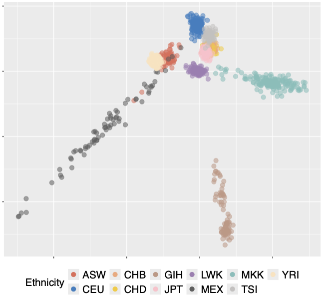


Visualizing 2-dim. (singular subspace) embeddings of high-dim. real data:

- Left: Single-cell sequencing data, with $n = 1604$ cells and $p = 19,298$ genes. Cell types are color-coded
- Right: HapMap data of human genetic variations, with $n = 1115$ and $p = 274,128$ SNPs. Ancestry groups are color-coded

Examples of Nonspherical Mixtures





Visualizing 2-dim. (singular subspace) embeddings of high-dim. real data:

- Left: Single-cell sequencing data, with $n = 1604$ cells and $p = 19,298$ genes. Cell types are color-coded
- Right: HapMap data of human genetic variations, with $n = 1115$ and $p = 274,128$ SNPs. Ancestry groups are color-coded

(Interpreted as degree-heterogeneous mixtures in [Lyu et al., 2025]. Also see e.g., [Jin, 2015, Ke and Jin, 2023], for degree corrected network models)

Exploit the Covariance Matrix

Meta Question

How to exploit covariance information to facilitate clustering?

Exploit the Covariance Matrix

Meta Question

How to exploit covariance information to facilitate clustering?

Key Challenges in High Dimensions:

- Estimating full-size $p \times p$ covariance matrices is not feasible
- Given partial information of the covariance, how to design clustering criterion?
- **Fundamental limit** and **efficient algorithm** for clustering in high-dim. nonspherical mixtures?

Brief Overview of Clustering Methods

For Gaussian mixtures:

- Spectral clustering ([Löffler et al., 2021], [Zhang and Zhou, 2024]): apply *K-Means* to $\mathbf{YV} \in \mathbb{R}^{n \times K}$, where $\mathbf{V} \in \mathbb{R}^{p \times K}$ are the top K right singular vectors of $\mathbf{Y}$

  $\Longrightarrow$ tailored to spherical noises, suboptimal for anisotropic noise

Brief Overview of Clustering Methods

For Gaussian mixtures:

- Spectral clustering ([Löffler et al., 2021], [Zhang and Zhou, 2024]): apply *K-Means* to $\mathbf{YV} \in \mathbb{R}^{n \times K}$, where $\mathbf{V} \in \mathbb{R}^{p \times K}$ are the top K right singular vectors of $\mathbf{Y}$

  $\Longrightarrow$ tailored to spherical noises, suboptimal for anisotropic noise

- EM-type algorithms ([Chen and Zhang, 2024], [Cai et al., 2019]): iteratively update the cluster labels and cluster centers & covariances

  $\Longrightarrow$ not apply to high-dim. $p \gtrsim n$ or requires specific parameters

Brief Overview of Clustering Methods

For Gaussian mixtures:

- Spectral clustering ([Löffler et al., 2021], [Zhang and Zhou, 2024]): apply *K-Means* to $\mathbf{YV} \in \mathbb{R}^{n \times K}$, where $\mathbf{V} \in \mathbb{R}^{p \times K}$ are the top K right singular vectors of $\mathbf{Y}$

  $\Longrightarrow$ tailored to spherical noises, suboptimal for anisotropic noise

- EM-type algorithms ([Chen and Zhang, 2024], [Cai et al., 2019]): iteratively update the cluster labels and cluster centers & covariances

  $\Longrightarrow$ not apply to high-dim. $p \gtrsim n$ or requires specific parameters

- Semi-definite Programming (SDP): convex relaxations for clustering ([Davis et al., 2025])

  $\Longrightarrow$ None adapt to $p \gtrsim n$ with general covariances

Brief Overview of Clustering Methods

For Gaussian mixtures:

- Spectral clustering ([Löffler et al., 2021], [Zhang and Zhou, 2024]): apply *K-Means* to $\mathbf{YV} \in \mathbb{R}^{n \times K}$, where $\mathbf{V} \in \mathbb{R}^{p \times K}$ are the top K right singular vectors of $\mathbf{Y}$

  $\Longrightarrow$ tailored to spherical noises, suboptimal for anisotropic noise

- EM-type algorithms ([Chen and Zhang, 2024], [Cai et al., 2019]): iteratively update the cluster labels and cluster centers & covariances

  $\Longrightarrow$ not apply to high-dim. $p \gtrsim n$ or requires specific parameters

- Semi-definite Programming (SDP): convex relaxations for clustering ([Davis et al., 2025])

  $\Longrightarrow$ None adapt to $p \gtrsim n$ with general covariances

For mixtures of other distributions/data types:

- Typically likelihood-based for specific models, also can struggle in high-dimensions
  (except [Tian et al., 2024], spectral clustering for high-dim. categorical data)

Brief Overview of Minimax Rates for Clustering

Assess the clustering by $h(\widehat{\mathbf{z}}, \mathbf{z}) = \min_{\phi \in \text{perm}(K)} \frac{1}{n} \sum_{i \in [n]} \mathbb{I}\{\widehat{z}_i \neq \phi(z_i)\}$.

Brief Overview of Minimax Rates for Clustering

Assess the clustering by $h(\widehat{\mathbf{z}}, \mathbf{z}) = \min_{\phi \in \mathrm{perm}(K)} \frac{1}{n} \sum_{i\in[n]} \mathbb{I}\{\widehat{z}_i \neq \phi(z_i)\}$.

- **Isotropic Noise:** Let $\mathbf{E}_i \overset{\text{ind.}}{\sim} \mathcal{N}(0, \sigma^2 \mathbf{I}_p)$.

$$\inf_{\widehat{\mathbf{z}}} \sup_{\mathbf{z}^* \in \mathbf{\Theta}_z^*} \mathbb{E}\left[h(\widehat{\mathbf{z}}, \mathbf{z}^*)\right] \gtrsim \exp\left(-\frac{\triangle^2}{8\sigma^2}\right), \quad \text{by [Lu and Zhou, 2016]},$$

where $\triangle := \min_{k_1, k_2 \in [K]} \left\| \boldsymbol{\theta}_{k_1}^* - \boldsymbol{\theta}_{k_2}^* \right\|_2$.

Brief Overview of Minimax Rates for Clustering

Assess the clustering by $h(\widehat{\mathbf{z}}, \mathbf{z}) = \min_{\phi \in \text{perm}(K)} \frac{1}{n} \sum_{i\in[n]} \mathbb{I}\{\widehat{z}_i \neq \phi(z_i)\}$.

- **Isotropic Noise:** Let $\mathbf{E}_i \overset{\text{ind.}}{\sim} \mathcal{N}(0, \sigma^2 \mathbf{I}_p)$.

$$\inf_{\widehat{\mathbf{z}}} \sup_{\mathbf{z}^* \in \mathbf{\Theta}_z^*} \mathbb{E}\big[h(\widehat{\mathbf{z}}, \mathbf{z}^*)\big] \gtrsim \exp\left(-\frac{\triangle^2}{8\sigma^2}\right), \quad \text{by [Lu and Zhou, 2016]},$$

where $\triangle := \min_{k_1,k_2\in[K]} \left\|\boldsymbol{\theta}^*_{k_1} - \boldsymbol{\theta}^*_{k_2}\right\|_2$.

- **Anisotropic Noise with $p = O(1)$:** Let $\mathbf{E}_i \overset{\text{ind.}}{\sim} \mathcal{N}(0, \mathbf{\Sigma}_{z_i^*})$.

$$\inf_{\widehat{\mathbf{z}}} \sup_{\mathbf{z}^* \in \mathbf{\Theta}_z^*} \mathbb{E}\big[h(\widehat{\mathbf{z}}, \mathbf{z}^*)\big] \gtrsim \exp\Big(-\frac{{\text{SNR}^{\text{full}}}^2}{2}\Big), \quad \text{by [Chen and Zhang, 2024]},$$

where

$$\text{SNR}^{\text{full}} := \min_{k_1 \neq k_2 \in [K]} \min_{x\in\mathbb{R}^p} \Big\{\|{\mathbf{\Sigma}_{k_1}}^{-\frac{1}{2}}\mathbf{x}\|_2 : \underbrace{\phi_{\boldsymbol{\theta}^*_{k_1}, \mathbf{\Sigma}_{k_1}}(\mathbf{x})}_{\text{Gaussian pdf}} = \phi_{\boldsymbol{\theta}^*_{k_2}, \mathbf{\Sigma}_{k_2}}(\mathbf{x})\Big\}.$$

A Reduction from Clustering to Classification

From Clustering to Classification: Suppose that we are given the *true* centers and covariance matrices.

A Reduction from Clustering to Classification

From Clustering to Classification: Suppose that we are given the *true* centers and covariance matrices.

Q: The best way to classify?
A: Likelihood Ratio Estimator (by Neyman–Pearson Lemma).

Consider a two-component general Gaussian mixture model:

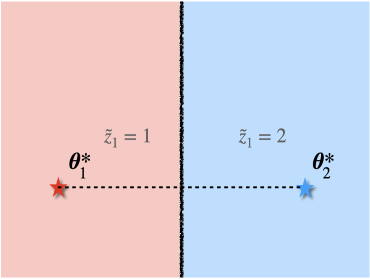
$$z_i^* \sim \frac{1}{2}\delta_1 + \frac{1}{2}\delta_2, \quad \mathbf{y}_i = \boldsymbol{\theta}^*_{z_i^*} + \mathbf{E}_i, \quad \mathbf{E}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{z_i^*}).$$

Likelihood Ratio Testing (LRT)-based estimator:

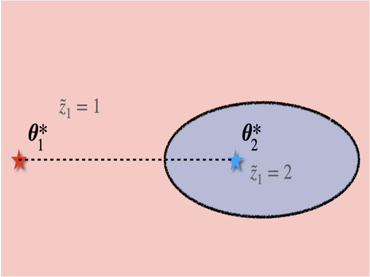
$$\widetilde{z}_i = \arg\max_{k \in \{1,2\}} \phi_{\boldsymbol{\theta}_k^*, \boldsymbol{\Sigma}_k}(\mathbf{y}_i).$$

Decision Boundary for Likelihood Ratio Testing (LRT)

- Case (a): Isotropic Noise ($\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \sigma^2 \mathbf{I}_p$)
- Case (b): Anisotropic Noise ($\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$)



(a) Isotropic Noise



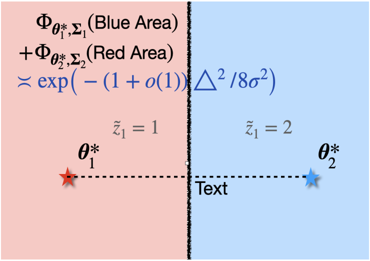
(b) Anisotropic Noise

Figure: Decision Boundary for LRT

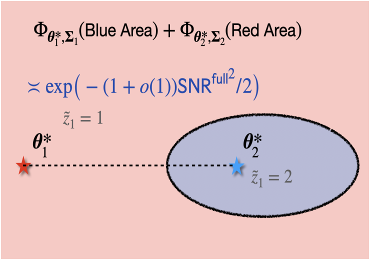
An Approach to Minimax Lower Bounds

A reduction from clustering to classification:

$$\inf_{\widehat{\mathbf{z}}} \sup_{\mathbf{z}^* \in \mathbf{\Theta}_z^*} \mathbb{E}\left[h(\widehat{\mathbf{z}}, \mathbf{z}^*)\right]$$
$$\gtrsim \Phi_{\boldsymbol{\theta}_1^*, \mathbf{\Sigma}_1}(\widetilde{z}_1 = 2) + \Phi_{\boldsymbol{\theta}_2^*, \mathbf{\Sigma}_2}(\widetilde{z}_1 = 1)$$
$$=: \mathcal{R}^{\mathrm{Bayes}}(\{\boldsymbol{\theta}_k^*\}_{k\in[2]}, \{\mathbf{\Sigma}_k\}_{k\in[2]})$$



(a) Isotropic Noise



(b) Anisotropic Noise ($p = O(1)$)

Remark: Throughout the discussion, let $\triangle$ or $\mathrm{SNR}^{\mathrm{full}}$ go to infinity

Rationale behind the Reduction

- Question: In which case is this reduction tight?
  Answer: In these cases where the information of centers $\{\boldsymbol{\theta}_k^*\}$ and covariance matrices $\{\boldsymbol{\Sigma}_k^*\}$ can be consistently estimated from data.

Rationale behind the Reduction

- Question: In which case is this reduction tight?
  Answer: In these cases where the information of centers $\{\boldsymbol{\theta}_k^*\}$ and covariance matrices $\{\boldsymbol{\Sigma}_k^*\}$ can be consistently estimated from data.
- Puzzling: Is this reduction still tight for anisotropic Gaussian mixtures when p is large?

Rationale behind the Reduction

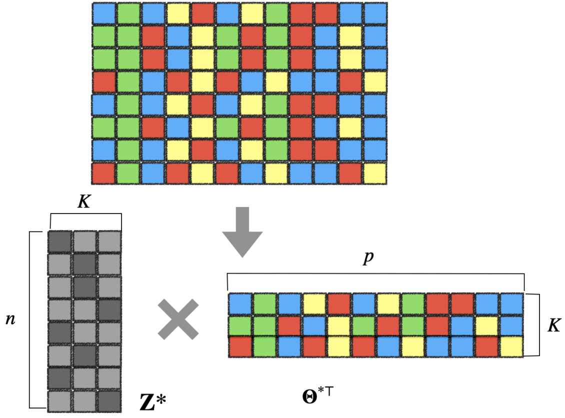
- Question: In which case is this reduction tight?
  Answer: In these cases where the information of centers $\{\boldsymbol{\theta}_k^*\}$ and covariance matrices $\{\boldsymbol{\Sigma}_k^*\}$ can be consistently estimated from data.
- Puzzling: Is this reduction still tight for anisotropic Gaussian mixtures when p is large?
- Observation: Unstructured covariance matrices are not recoverable when $p \gg n$, even when $\mathbf{z}^*$ is known.

Rationale behind the Reduction

- Question: In which case is this reduction tight?
  Answer: In these cases where the information of centers $\{\boldsymbol{\theta}_k^*\}$ and covariance matrices $\{\boldsymbol{\Sigma}_k^*\}$ can be consistently estimated from data.
- Puzzling: Is this reduction still tight for anisotropic Gaussian mixtures when p is large?
- Observation: Unstructured covariance matrices are not recoverable when $p \gg n$, even when $\mathbf{z}^*$ is known.

This work reveals:
$\mathcal{R}^{\text{Bayes}}$ isn't always achieavable. Instead, there exists a gap between the minimax rate and $\mathcal{R}^{\text{Bayes}}$, surprisingly related to an intriguing low-dimensional quantity $\text{SNR}^{\text{partial}}$ ($\ll \text{SNR}^{\text{full}}$).

A Subspace Viewpoint



A Subspace Viewpoint: Rank-K decomposition:

$$\mathbb{E}[\underbrace{\mathbf{Y}}_{n\times p}] = \mathbf{Y}^*_{n\times p} = \underbrace{\mathbf{Z}^*}_{n\times K}\underbrace{\mathbf{\Theta}^{*\top}}_{K\times p}$$

$\mathbf{V}^* \in \mathbb{R}^{p\times K}$: top-$K$ right singular vectors of $\mathbf{Y}^*$
$\mathbf{V} \in \mathbb{R}^{p\times K}$: top-$K$ right singular vectors of $\mathbf{Y} = \mathbf{Y}^* + \mathbf{E}$

New Minimax Lower Bound

Theorem (Informal Lower Bound)

If $\mathsf{SNR}^{\text{partial}} \to \infty$ *and* $p/n \to \infty$, *then*

$$\inf_{\widehat{\mathbf{z}}} \sup_{\mathbf{\Theta}_0} \mathbb{E}[h(\widehat{\mathbf{z}}, \mathbf{z}^*)] \gtrsim \exp\left(-(1+o(1))\frac{{\mathsf{SNR}^{\text{partial}}}^2}{2}\right),$$

where $\mathbf{\Theta}_0 := \underbrace{\widetilde{\mathbf{\Theta}}_0}_{\textit{centers and covariances}} \otimes \underbrace{\mathbf{\Theta}_z}_{\textit{assignments}}$ *and*

$$\mathsf{SNR}^{\text{partial}} := \min_{k_1, k_2 \in [K]} \min_{x \in \mathbb{R}^K} \left\{ \|(\mathbf{S}_k^*)^{-\frac{1}{2}} \mathbf{x}\|_2 : \underbrace{\phi_{\mathbf{w}_{k_1}^*, \mathbf{S}_{k_1}^*}(\mathbf{x})}_{K\text{-dim. pdf}} = \phi_{\mathbf{w}_{k_2}^*, \mathbf{S}_{k_2}^*}(\mathbf{x}) \right\},$$

$$\mathbf{w}_k^* = \mathbf{V}^{*\top} \boldsymbol{\theta}_k^* \in \mathbb{R}^K, \quad \mathbf{S}_k^* = \mathbf{V}^{*\top} \mathbf{\Sigma}_k \mathbf{V}^* \in \mathbb{R}^{K \times K}.$$

New Minimax Lower Bound

Theorem (Informal Lower Bound)

If $\mathsf{SNR}^{\text{partial}} \to \infty$ *and* $p/n \to \infty$, *then*

$$\inf_{\widehat{\mathbf{z}}} \sup_{\boldsymbol{\Theta}_0} \mathbb{E}[h(\widehat{\mathbf{z}}, \mathbf{z}^*)] \gtrsim \exp\left(-(1+o(1))\frac{{\mathsf{SNR}^{\text{partial}}}^2}{2}\right),$$

where $\boldsymbol{\Theta}_0 := \underbrace{\widetilde{\boldsymbol{\Theta}}_0}_{\textit{centers and covariances}} \otimes \underbrace{\boldsymbol{\Theta}_z}_{\textit{assignments}}$ *and*

$$\mathsf{SNR}^{\text{partial}} := \min_{k_1, k_2 \in [K]} \min_{x \in \mathbb{R}^K} \left\{ \|(\mathbf{S}_k^*)^{-\frac{1}{2}} \mathbf{x}\|_2 : \underbrace{\phi_{\mathbf{w}_{k_1}^*, \mathbf{S}_{k_1}^*}(\mathbf{x})}_{K\text{-dim. pdf}} = \phi_{\mathbf{w}_{k_2}^*, \mathbf{S}_{k_2}^*}(\mathbf{x}) \right\},$$

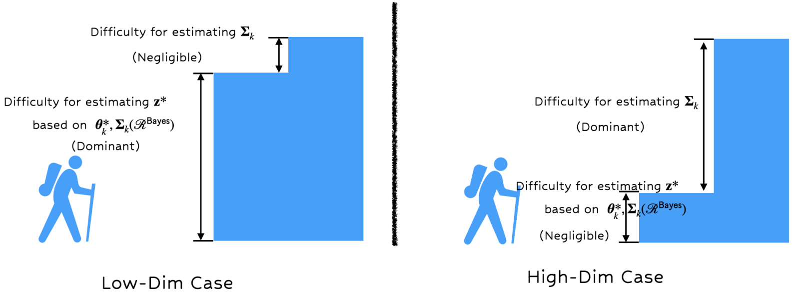
$$\mathbf{w}_k^* = \mathbf{V}^{*\top} \boldsymbol{\theta}_k^* \in \mathbb{R}^K, \quad \mathbf{S}_k^* = \mathbf{V}^{*\top} \boldsymbol{\Sigma}_k \mathbf{V}^* \in \mathbb{R}^{K \times K}.$$

Recall $\mathsf{SNR}^{\text{full}} := \min_{k_1 \neq k_2 \in [K]} \min_{x \in \mathbb{R}^p} \left\{ \|{\boldsymbol{\Sigma}_{k_1}}^{-\frac{1}{2}} \mathbf{x}\|_2 : \underbrace{\phi_{\boldsymbol{\theta}_{k_1}^*, \boldsymbol{\Sigma}_{k_1}}(\mathbf{x})}_{p\text{-dim. pdf}} = \phi_{\boldsymbol{\theta}_{k_2}^*, \boldsymbol{\Sigma}_{k_2}}(\mathbf{x}) \right\}.$

Implications

$$\mathcal{R}^{\text{Bayes}} = \exp\left(-(1+o(1))\frac{{\text{SNR}^{\text{full}}}^2}{2}\right) \ll \exp\left(-(1+o(1))\frac{{\text{SNR}^{\text{partial}}}^2}{2}\right)$$

$$\Longrightarrow \mathcal{R}^{\text{Bayes}} \text{ is not achievable.}$$



New Clustering Algorithm: COPO

$$\mathrm{SNR}^{\mathrm{partial}} := \min_{k_1 \neq k_2 \in [K]} \min_{x \in \mathbb{R}^K} \left\{ \|(\mathbf{S}_k^*)^{-\frac{1}{2}} \mathbf{x}\|_2 : \phi_{\mathbf{w}_{k_1}^*, \mathbf{S}_{k_1}^*}(\mathbf{x}) = \phi_{\mathbf{w}_{k_2}^*, \mathbf{S}_{k_2}^*}(\mathbf{x}) \right\}$$

only involves low-dimensional quantities

$$\mathbf{w}_k^* = \mathbf{V}^{*\top} \boldsymbol{\theta}_k^* \in \mathbb{R}^K, \quad \mathbf{S}_k^* = \mathbf{V}^{*\top} \boldsymbol{\Sigma}_k \mathbf{V}^* \in \mathbb{R}^{K \times K}.$$

$\Longrightarrow$ This motivates us to propose a novel clustering method

Idea of **Co**variance **Pro**jected Spectral Clustering (COPO):

- Replace $\mathbf{V}^*_{p \times K}$ with $\mathbf{V}_{p \times K}$ (empirical top-K right singular subspace of data $\mathbf{Y}$);
- Iteratively update the estimates for $\mathbf{w}_k^*$ (projected centers) and $\mathbf{S}_k^*$ (projected covariances)

Algorithm 1: Covariance Projected Spectral Clustering (COPO)

Input: Data matrix $\mathbf{Y} \in \mathbb{R}^{n \times p}$, number of clusters K, an initial cluster estimate $\widehat{\mathbf{z}}^{(0)}$
Output: Cluster assignment vector $\widehat{\mathbf{z}} \in [K]^n$

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

For each $k \in [K]$, update the cluster centers by

$$\widehat{\theta}_k^{(t)} = \frac{\sum_{i \in [n]} \mathbf{1}\left\{\widehat{z}_i^{(t-1)} = k\right\} \mathbf{y}_i}{\sum_{i \in [n]} \mathbf{1}\left\{\widehat{z}_i^{(t-1)} = k\right\}},$$

and update the projected covariance matrices by

$$\widehat{\mathbf{S}}_k^{(t)} := \frac{\sum_{i \in c_k} \mathbf{V}^\top (\mathbf{y}_i - \widehat{\theta}_k^{(t)})^\top (\mathbf{y}_i - \widehat{\theta}_k^{(t)}) \mathbf{V}}{\sum_{i \in [n]} \mathbf{1}\left\{\widehat{z}_i^{(t-1)} = k\right\}} \qquad \text{(size } K \times K\text{)}$$

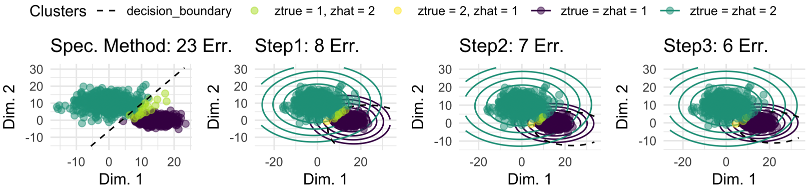
Update the cluster labels[a] for $i \in [n]$ by comparing the Mahalanobis distance in $\mathbb{R}^K$:

$$\widehat{z}_i^{(t)} = \arg\min_{k \in [K]} \underbrace{\left[(\mathbf{y}_i - \widehat{\boldsymbol{\theta}}_k^{(t)})^\top \mathbf{V}\right]}_{1 \times K} \underbrace{\widehat{\mathbf{S}}_k^{(t)\,-1}}_{K \times K} \underbrace{\left[\mathbf{V}^\top (\mathbf{y}_i - \widehat{\boldsymbol{\theta}}_k^{(t)})\right]}_{K \times 1}.$$

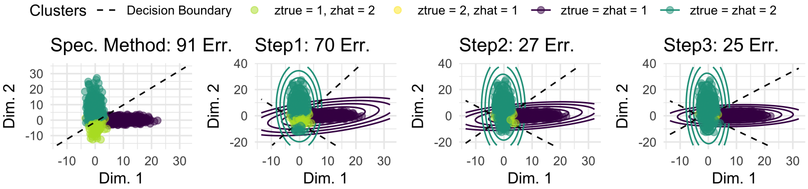
end

[a] We drop the log term from log-likelihood of normal distribution

Numerical Examples



(a) A Case with Elliptical Decision Boundaries



(b) A Case with Hyperbolic Decision Boundaries

Figure: Spectral clustering [Löffler et al., 2021] and our method in the subspace spanned by the top-2 empirical singular vectors. Data from a 2-component Gaussian mixture with $n = 500$ and $p = 1000$.

Non-Gaussian Mixture Models?

For mixtures of non-Gaussian distributions, Gaussian EM algorithm should not be directly applied.

But how about after projection?

Non-Gaussian Mixture Models?

For mixtures of non-Gaussian distributions, Gaussian EM algorithm should not be directly applied.

But how about after projection?

Inferential Results in Singular Subspace Perturbation Theory[a]

$$\mathbf{U}_{i,:}\mathbf{O} - \mathbf{U}^*_{i,:} \Rightarrow \mathcal{N}\big(\mathbf{0}, \mathbf{D}^{*-1}\underbrace{\mathbf{V}^{*\top}\mathbf{\Sigma}_{z_i^*}\mathbf{V}^*}_{=:\mathbf{S}_{z_i^*}\ (K\times K)}\mathbf{D}^{*-1}\big)$$

even when $\mathbf{Y}_i$ itself is not Gaussian!
Here $\mathbf{U}^* \in \mathbb{R}^{p\times K}$ are the left singular vectors of $\mathbf{Y}^*$, and $\mathbf{D}^*$ is a diagonal matrix with K singular values of $\mathbf{Y}^*$

[a] [Yan et al., 2024, Agterberg et al., 2022, Xia, 2021]

Non-Gaussian Mixture Models?

For mixtures of non-Gaussian distributions, Gaussian EM algorithm should not be directly applied.

But how about after projection?

Inferential Results in Singular Subspace Perturbation Theory[a]

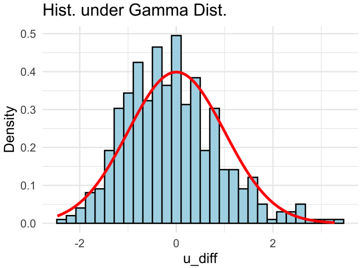
$$\mathbf{U}_{i,:}\mathbf{O} - \mathbf{U}^*_{i,:} \Rightarrow \mathcal{N}\Big(\mathbf{0}, \mathbf{D}^{*-1}\underbrace{\mathbf{V}^{*\top}\boldsymbol{\Sigma}_{z_i^*}\mathbf{V}^*}_{=:\mathbf{S}_{z_i^*}\ (K\times K)}\mathbf{D}^{*-1}\Big)$$

even when $\mathbf{Y}_i$ itself is not Gaussian!
Here $\mathbf{U}^* \in \mathbb{R}^{p\times K}$ are the left singular vectors of $\mathbf{Y}^*$, and $\mathbf{D}^*$ is a diagonal matrix with K singular values of $\mathbf{Y}^*$

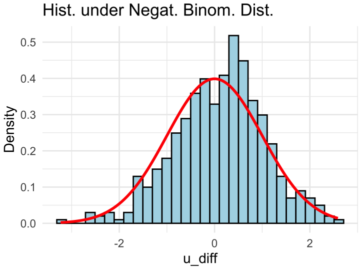
[a] [Yan et al., 2024, Agterberg et al., 2022, Xia, 2021]

Justifies the use of LRT-based estimation for projected data!

Example of Non-Gaussian Noise



(a) Gamma Distribution



(b) Negative Binomial Distribution

Figure: Histogram of scaled $(\mathbf{U}\mathbf{R}_{\mathbf{U}} - \mathbf{U}^*)_{1,1}$.

Main Noise Assumptions

Assumptions on Gaussian Noise with Arbitrary Dependence

- $\mathbf{E}_i \overset{\text{ind.}}{\sim} \mathcal{N}(\boldsymbol{\theta}^*_{z_i^*}, \boldsymbol{\Sigma}_{z_i^*})$;
- $\max_{i\in[n],j\in[p]} \mathrm{Var}(E_{i,j}) \le \sigma^2$.

Main Noise Assumptions

Assumptions on Gaussian Noise with Arbitrary Dependence

- $\mathbf{E}_i \overset{\text{ind.}}{\sim} \mathcal{N}(\boldsymbol{\theta}^*_{z_i^*}, \boldsymbol{\Sigma}_{z_i^*})$;
- $\max_{i\in[n],j\in[p]} \mathrm{Var}(E_{i,j}) \le \sigma^2$.

Assumptions on General Noise with Block Dependence

- There exists a partition $S_1, S_2, \ldots, S_l$ of $[p]$ with $|S_l| \le m$ for $l \in [l]$ s.t. $\{\mathbf{E}_{i,S_l}\}_{l=1}^{l}$ are mutually independent for $i \in [n]$ and $l \in [l]$.
- $\exists$ a random matrix $\mathbf{E}' = (E'_{i,j}) \in \mathbb{R}^{n\times p}$ obeying the same dependence structure s.t. for any $i \in [n], j \in [p]$, it holds that $\left\|E'_{i,j}\right\|_\infty \le B$, $\mathbb{E}[E'_{i,j}] = 0$, $\left\|\mathrm{Cov}(\mathbf{E}'_{i,:})\right\| \lesssim \|\mathrm{Cov}(\mathbf{E}_{i,:})\|$, and $E_{i,j} = E'_{i,j}$ w.h.p..

Main Noise Assumptions

Assumptions on Gaussian Noise with Arbitrary Dependence

- $\mathbf{E}_i \overset{\text{ind.}}{\sim} \mathcal{N}(\boldsymbol{\theta}^*_{z_i^*}, \boldsymbol{\Sigma}_{z_i^*})$;
- $\max_{i\in[n],j\in[p]} \mathrm{Var}(E_{i,j}) \le \sigma^2$.

Assumptions on General Noise with Block Dependence

- There exists a partition $S_1, S_2, \ldots, S_l$ of $[p]$ with $|S_l| \le m$ for $l \in [l]$ s.t. $\{\mathbf{E}_{i,S_l}\}_{l=1}^{l}$ are mutually independent for $i \in [n]$ and $l \in [l]$.
- $\exists$ a random matrix $\mathbf{E}' = (E'_{i,j}) \in \mathbb{R}^{n\times p}$ obeying the same dependence structure s.t. for any $i \in [n], j \in [p]$, it holds that $\left\|E'_{i,j}\right\|_\infty \le B$, $\mathbb{E}[E'_{i,j}] = 0$, $\left\|\mathrm{Cov}(\mathbf{E}'_{i,:})\right\| \lesssim \|\mathrm{Cov}(\mathbf{E}_{i,:})\|$, and $E_{i,j} = E'_{i,j}$ w.h.p..

Common Assumption: The smallest singular values of $\mathbf{V}^{*\top}\boldsymbol{\Sigma}_k\mathbf{V}^*$ for $k \in [K]$ are lower bounded

Motivation for Local Dependence

Local Dependence: American National Election Survey (ANES)

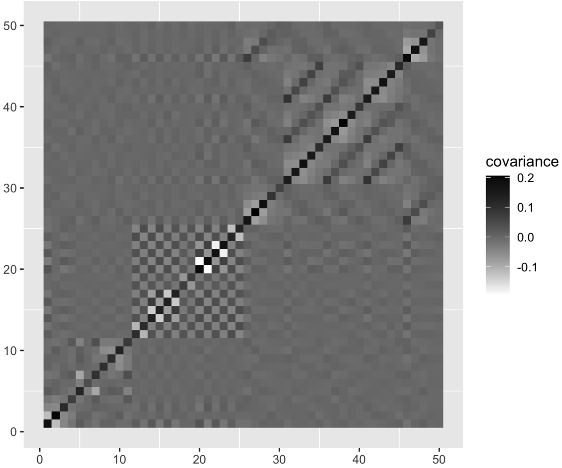


Figure: Approximate noise covariance matrix for a subset of survey items in ANES.

Upper Bound

Theorem (Informal Upper Bound)

Assume SNR $\gg \sqrt{\log\log(n \vee p)}$ *and a reasonable initialization. Then for all* $t \geq c \log n$*:*

1. *If* $\mathrm{SNR} \leq 2\sqrt{\log n}$*, then*

$$\mathbb{E}[h(\widehat{\mathbf{z}}^{(t)}, \mathbf{z}^*)] \lesssim \exp\left(-(1+o(1))\frac{{\mathrm{SNR}^{\mathrm{partial}}}^2}{2}\right).$$

2. *If* $\mathrm{SNR} \geq (\sqrt{2}+\epsilon)\sqrt{\log n}$ *with an arbitrary positive number* ϵ*, then* $h(\widehat{\mathbf{z}}^{(t)}, \mathbf{z}^*) = 0$ *with probability* $1 - o(1)$*.*

Techniques:

- Universality on matrix concentration [Bandeira et al., 2023][Brailovskaya and van Handel, 2022]
- Leave-one-out perturbation analysis [Zhang and Zhou, 2024]
- Delicate analysis under local dependence

Remarks on the Upper Bound

- Optimality. COPO is minimax optimal under general anisotropic Gaussian mixtures when $p \gg n$

Remarks on the Upper Bound

- Optimality. COPO is minimax optimal under general anisotropic Gaussian mixtures when $p \gg n$
- Covering Weak Signal Strength. Allow $\mathrm{SNR}^{\mathrm{partial}}$ growing slightly exceeding $\sqrt{\log\log(n \vee p)}$

Remarks on the Upper Bound

- Optimality. COPO is minimax optimal under general anisotropic Gaussian mixtures when $p \gg n$
- Covering Weak Signal Strength. Allow $\mathrm{SNR}^{\mathrm{partial}}$ growing slightly exceeding $\sqrt{\log\log(n \vee p)}$
- Computational Efficiency of COPO. The time costs consist of
  - performing the top-K SVD on $\mathbf{Y}$, which is $O(npK)$
  - iterative averaging over the projected centers space $\mathbb{R}^K$ and the projected covariance matrix space $\mathbb{R}^{K\times K}$ in $O(\log n)$ iterations

Remarks on the Upper Bound

- Optimality. COPO is minimax optimal under general anisotropic Gaussian mixtures when $p \gg n$
- Covering Weak Signal Strength. Allow $\mathrm{SNR}^{\mathrm{partial}}$ growing slightly exceeding $\sqrt{\log\log(n \vee p)}$
- Computational Efficiency of COPO. The time costs consist of
  - performing the top-K SVD on $\mathbf{Y}$, which is $O(npK)$
  - iterative averaging over the projected centers space $\mathbb{R}^K$ and the projected covariance matrix space $\mathbb{R}^{K\times K}$ in $O(\log n)$ iterations
- Block Size. The block size m can scale as the order $O(p^a)$ with $a \in (0,1)$, corresponding to severely dependent noise matrix entries

Remarks on the Upper Bound

- Optimality. COPO is minimax optimal under general anisotropic Gaussian mixtures when $p \gg n$
- Covering Weak Signal Strength. Allow $\mathrm{SNR}^{\mathrm{partial}}$ growing slightly exceeding $\sqrt{\log\log(n \vee p)}$
- Computational Efficiency of COPO. The time costs consist of
  - performing the top-K SVD on $\mathbf{Y}$, which is $O(npK)$
  - iterative averaging over the projected centers space $\mathbb{R}^K$ and the projected covariance matrix space $\mathbb{R}^{K\times K}$ in $O(\log n)$ iterations
- Block Size. The block size m can scale as the order $O(p^a)$ with $a \in (0,1)$, corresponding to severely dependent noise matrix entries
- Covering Sub-Gaussian/Sub-exponential mixtures with arbitrary local dependence: high-dim. count data, discrete data, skewed data

Simulation: Gaussian Mixtures

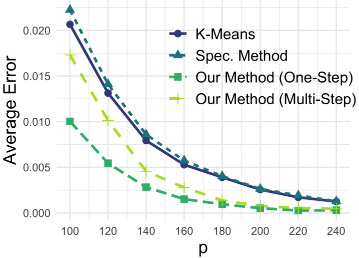
n	p	Spec. err.	COPO err.	COPO time	EM err. (%Suc.)	EM time
500	40	0.436	0.441	0.056	0.005 (97.0%)	2.2
500	80	0.412	0.418	0.057	0.057 (94.5%)	12.5
500	120	0.374	0.376	0.062	0.190 (88.0%)	32.7
500	160	0.342	0.335	0.059	0.322 (65.0%)	22.0
500	200	0.302	0.275	0.063	0.299 (40.5%)	24.4
500	500	0.127	0.085	0.075	–	–
500	1000	0.041	0.032	0.096	–	–
500	1500	0.015	0.012	0.112	–	–
500	2000	0.005	0.005	0.124	–	–
500	5000	0.000	0.000	0.206	–	–

Table: Clustering error rates and computation times for Gaussian mixtures. The unit of time is seconds. The (%Suc.) means the proportion of simulation trials in which the EM algorithm runs without failures.

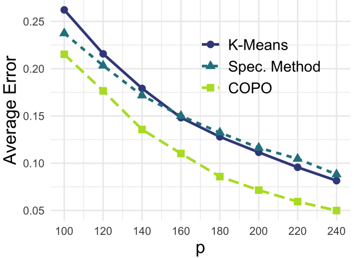
Simulation: Other Mixtures

- Mixtures of Ising Models: multivariate binary data, local dependence induced by graphical Ising models
- Multivariate Probit Mixtures: multivariate binary data, local dependence induced by dichotomizing underlying Gaussian variables
- Multivariate Gamma Mixtures: multivariate positive skewed continuous data
- Negative Binomial Mixtures: multivariate nonegative count data

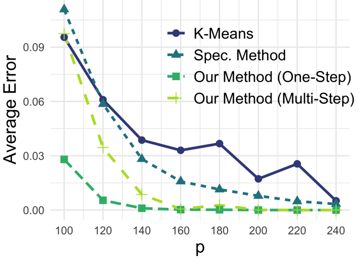
Simulation: Other Mixtures



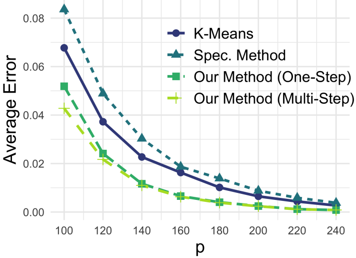
(a) Mixtures of Ising Models



(b) Multivariate Probit Mixtures



(c) Multivariate Gamma Mixtures

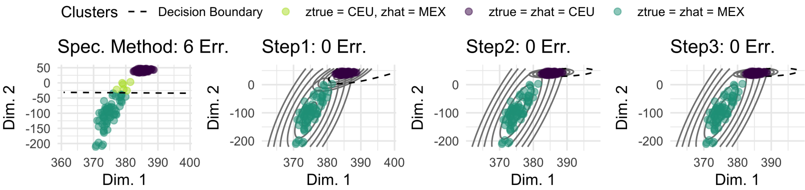


(d) Negative Binomial Mixtures

HapMap3 Data

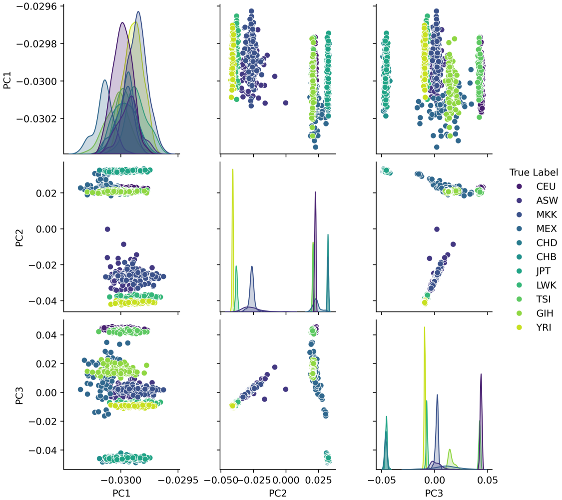
- $p > 2.7 \times 10^5$, $n = 1301$.
- On two subpopulations CEU (Utah residents with Northern and Western European ancestry) and MEX (Mexican ancestry):

  COPO achieves exact recovery, $h(\widehat{\mathbf{z}}^{\text{kmeans}}, \mathbf{z}^*) = 3.4\%$ and $h(\widehat{\mathbf{z}}^{\text{spectral}}, \mathbf{z}^*) = 2.6\%$.



HapMap3 Data

For full-size dataset, our method achieves an accuracy of 75.7%, outperforming the *K-means* (60.9%) and the spectral clustering (74.4%).



Summary

- A novel clustering algorithm for high-dimensional data: Covariance Projected Spectral Clustering (COPO)
- COPO projects p-dimensional data onto empirical top-K right singular subspace of $\mathbf{Y}$, and iteratively refines cluster assignments based on projected centers and projected covariance matrices
- A new minimax lower bound for clustering unveiling an intriguing informational dimension-reduction phenomenon
- COPO is optimal for general high-dim. Gaussian mixtures and strongly adaptive to a broad family of other mixture models

Huang and Gu (2025+). Minimax-Optimal Dimension-Reduced Clustering for High-Dimensional Nonspherical Mixtures. *arXiv preprint* **arXiv:2502.02580.**

References I

Agterberg, J., Lubberts, Z., and Priebe, C. E. (2022).
Entrywise estimation of singular vectors of low-rank matrices with heteroskedasticity and dependence.
IEEE Transactions on Information Theory, 68(7):4618–4650.

Bandeira, A. S., Boedihardjo, M. T., and van Handel, R. (2023).
Matrix concentration inequalities and free probability.
Inventiones Mathematicae, pages 1–69.

Brailovskaya, T. and van Handel, R. (2022).
Universality and sharp matrix concentration inequalities.
arXiv preprint arXiv:2201.05142.

Cai, T. T., Ma, J., and Zhang, L. (2019).
Chime: Clustering of high-dimensional gaussian mixtures with em algorithm and its optimality.
Annals of Statistics.

References II

Chen, X. and Zhang, A. Y. (2024).
Achieving optimal clustering in Gaussian mixture models with anisotropic covariance structures.
In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Davis, D., Diaz, M., and Wang, K. (2025).
Clustering a mixture of gaussians with unknown covariance.
Bernoulli, 31(3):2105–2126.

Gao, C. and Zhang, A. Y. (2022).
Iterative algorithm for discrete structure recovery.
The Annals of Statistics, 50(2):1066–1094.

Jin, J. (2015).
Fast community detection by score.
The Annals of Statistics, pages 57–89.

References III

Ke, Z. T. and Jin, J. (2023).
Special invited paper: The score normalization, especially for heterogeneous network and text data.
Stat, 12(1):e545.

Lu, Y. and Zhou, H. H. (2016).
Statistical and computational guarantees of lloyd's algorithm and its variants.
arXiv preprint arXiv: Arxiv-1612.02099.

Lyu, Z., Chen, L., and Gu, Y. (2025).
Degree-heterogeneous latent class analysis for high-dimensional discrete data.
Journal of the American Statistical Association, (just-accepted):1–25.

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

References IV

Tian, Z., Xu, J., and Tang, J. (2024).
Clustering high-dimensional noisy categorical data.
Journal of the American Statistical Association, 119(548):3008–3019.

Xia, D. (2021).
Normal approximation and confidence region of singular subspaces.
Electronic Journal of Statistics, 15(2):3798–3851.

Yan, Y., Chen, Y., and Fan, J. (2024).
Inference for heteroskedastic PCA with missing data.
The Annals of Statistics, 52(2):729–756.

Zhang, A. Y. and Zhou, H. H. (2024).
Leave-one-out singular subspace perturbation analysis for spectral clustering.
The Annals of Statistics, 52(5):2004–2033.

(Hard) EM Algorithm for Gaussian Mixtures

Classical Viewpoint: consider covariance matrices as part of the parameters.

EM Algorithm Given $\{w_{i,k}^{(t)}\}_{i\in[n],k\in[K]}$, $\{\boldsymbol{\theta}_k^{(t)}\}_{k\in[K]}$, $\{\boldsymbol{\Sigma}_k^{(t)}\}_{k\in[K]}$,

- **E-step**: Update the posterior: $w_{i,k}^{(t+1)} = \frac{\phi_{\boldsymbol{\theta}_k^{(t)},\boldsymbol{\Sigma}_k^{(t)}}(\mathbf{y}_i)}{\sum_{l\in[K]} w_{i,l}^{(t)}\phi_{\boldsymbol{\theta}_l^{(t)},\boldsymbol{\Sigma}_l^{(t)}}(\mathbf{y}_i)}$.
- **M-step**: Re-estimate the parameters:

$$\boldsymbol{\theta}_k^{(t+1)} = \frac{\sum_{i\in[n]} w_{i,k}^{(t+1)}\mathbf{y}_i}{\sum_{i\in[n]} w_{i,k}^{(t+1)}}, \quad \boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_{i\in[n]} w_{i,k}^{(t+1)}(\mathbf{y}_i - \boldsymbol{\theta}_k^{(t+1)})(\mathbf{y}_i - \boldsymbol{\theta}_k^{(t+1)})^\top}{\sum_{i\in[n]} w_{i,k}^{(t+1)}}.$$

Recursively update until convergence. Then the estimation is given by $\widehat{z}_i \coloneqq \arg\max_{k\in[K]} w_{i,k}^{(t)}$.

Hard EM

Hard EM: Update assignment recursively: $\mathbf{w}_{i,k}^{(t+1)} = 1_{\{k=\arg\max_{l\in[K]} \phi_{\boldsymbol{\theta}_k^{(t)},\boldsymbol{\Sigma}_k^{(t)}}(\mathbf{y}_i)\}}$.

Inhomoegeneous cov. matrices with $p = O(1)$: the hard EM is proved to be minimax-optimal [Chen and Zhang, 2024].

Existing Methods

- Iterative methods directly on p-dimensional data (EM algorithm, Lloyd algorithm) is computationally expensive **for large** p.
- Spectral Methods: Efficient, Statistically Optimal under simple Isotropic (spherical) Gaussian Mixtures.

Related Existing Methods

- Get the top-K SVD $(\mathbf{U}, \mathbf{D}, \mathbf{V})$ of $\mathbf{R}$ and perform K-means for $\mathbf{UD}$ (*Weighted Spectral Clustering*) [Zhang and Zhou, 2024].
- For fixed-p Gaussian mixtures, [Chen and Zhang, 2024] uses $p \times p$ covariance matrix to adjust Lloyd algorithm

Motivation for Our Method

Drawbacks of existing methods:

1. Cov. matrices $\Sigma_{z_i^*} := \mathrm{Cov}(\mathbf{E}_i)$ $(p \times p)$ are **not full-rank**
2. No consistent estimator for Σ_k when $n \asymp p$.

Singular Subspace Perturbation Theory

$$\mathbf{U}_{i,:}\mathbf{O} - \mathbf{U}^*_{i,:} \Rightarrow \mathcal{N}\big(\mathbf{0}, \mathbf{D}^{*-1} \underbrace{\mathbf{V}^{*\top} \boldsymbol{\Sigma}_{z_i^*} \mathbf{V}^*}_{=:\mathbf{S}_{z_i^*}\ (K \times K)} \mathbf{D}^{*-1}\big)$$

even when R itself is not Gaussian!

Key: Directly motivate our new method of projection + covariance adjustment

Our Proposal

Algorithm 2: Covariance Projected Spectral Clustering

Input: Data matrix $\mathbf{R} \in \mathbb{R}^{n\times p}$, number of clusters K, an initial cluster estimate $\widehat{\mathbf{z}}^{(0)}$
Output: Cluster assignment vector $\widehat{\mathbf{z}} \in [K]^n$

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

For each $k \in [K]$, estimate the centers θ_k^* by $\widehat{\theta}_k^{(t)} = \frac{\sum_{i\in[n]} \mathbf{1}\left\{\widehat{z}_i^{(t-1)}=k\right\}\mathbf{R}_i}{\sum_{i\in[n]} \mathbf{1}\left\{\widehat{z}_i^{(t-1)}=k\right\}}$ and estimate the projected covariance matrix by

$$\widehat{\mathbf{S}}_k^{(t)} := \frac{\sum_{i\in c_k} \mathbf{V}^\top(\mathbf{R}_i - \widehat{\theta}_k^{(t)})^\top(\mathbf{R}_i - \widehat{\theta}_k^{(t)})\mathbf{V}}{\sum_{i\in[n]} \mathbf{1}\left\{\widehat{z}_i^{(t-1)} = k\right\}} \qquad \text{(size } K \times K\text{)}$$

Estimate the cluster memberships:

$$\widehat{z}_i^{(t)} = \min_{k\in[K]} \underbrace{(\mathbf{R}_i - \widehat{\theta}_k^{(t)})^\top \mathbf{V}\widehat{\mathbf{S}}_k^{(t)\,-1}\mathbf{V}^\top(\mathbf{R}_i - \widehat{\theta}_k^{(t)})}_{\approx(\mathbf{UO}-\mathbf{U}^*)_{i,:}\mathrm{Cov}(\mathbf{UO}-\mathbf{U}^*)^{-1}(\mathbf{UO}-\mathbf{U}^*)_{i,:}^\top} + \log|\widehat{\mathbf{S}}_k^{(t)}|.$$

end

Project the high-dim. $\mathbf{R}_i$ to the space spanned by the cluster centers – We don't deal with $p \times p$ cov. mat. anymore!

Upper Bound

Theorem (Informal Upper Bound)

We assume that SNR $\to \infty$ *and the initialization* $\widehat{\mathbf{z}}^{(0)}$ *satisfies* $h(\widehat{\mathbf{z}}^{(0)}, \mathbf{z}^*) \leq c\frac{1}{K\log(n)}$ *with probability at least* $1-\eta$. *Then for all* $t \geq \log n$, *it holds with probability at least* $1-\eta-Cn^{-1}$ *that*

$$h(\widehat{\mathbf{z}}^{(t)}, \mathbf{z}^*) \leq \exp\left(-(1+o(1))\frac{\mathrm{SNR}^2}{2}\right).$$

where $h(\widehat{\mathbf{z}}, \mathbf{z}) = \min_{\phi\in\mathrm{perm}(K)} \frac{1}{n}\sum_{i\in[n]} \mathbb{I}\{\widehat{z}_i \neq \phi(z_i)\}$.

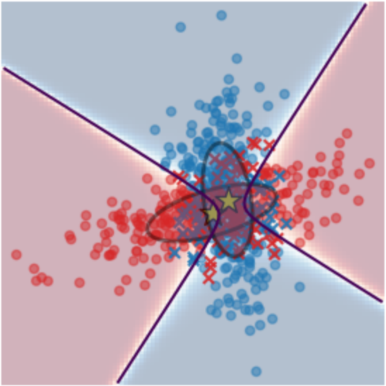
- Required Technique: Universality on Matrix Concentration [Bandeira et al., 2023][Brailovskaya and van Handel, 2022].

Decision Boundary

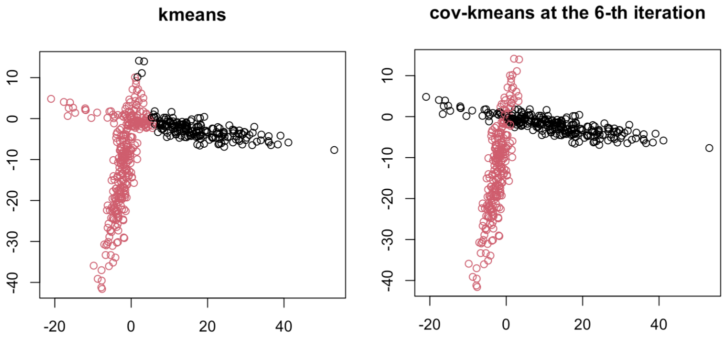
- Let $\mathbf{S}_k = \mathbf{V}^{*\top}\Sigma_k\mathbf{V}^*$. SNR (Signal-Noise-Ratio) is defined as

$$\text{SNR} := \min_{k_1 \neq k_2 \in [K]} \min_{\mathbf{x} \in \mathcal{B}_{k_1,k_2}} \left\| \mathbf{S}_{k_1}^{-\frac{1}{2}} (\mathbf{x} - \mathbf{V}^{*\top}\theta_{k_1}^*) \right\|_2$$

- $\mathcal{B}_{k_1,k_2}$ is the decision boundary between two Gaussians with $\mathcal{N}(\mathbf{V}^{*\top}\boldsymbol{\theta}_{k_1}^*, \mathbf{S}_{k_1})$ and $\mathcal{N}(\mathbf{V}^{*\top}\boldsymbol{\theta}_{k_2}^*, \mathbf{S}_{k_2})$.



Simulation Example



(a) Spec. Method (b) Cov. Proj. Spec. Method

Figure: Comparison Example

Why Performing Projection?

$$\Sigma_k = \mathbf{V}^*\mathbf{S}_k\mathbf{V}^{*\top} + \mathbf{V}_\perp^*\mathbf{V}_\perp^{*\top}\Sigma_k\mathbf{V}_\perp^*\mathbf{V}_\perp^{*\top}$$
$$+\mathbf{V}^*\mathbf{V}^{*\top}\Sigma_k\mathbf{V}_\perp^*\mathbf{V}_\perp^{*\top} + \mathbf{V}_\perp^*\mathbf{V}_\perp^{*\top}\Sigma_k\mathbf{V}^*\mathbf{V}^{*\top}$$

Question: Why are we only interested in $\mathbf{S}_k$?

Reasons:

1. For some discrete cases, $\mathbf{S}_k$ is enough. (Lower Bound 1)
2. For Gaussian mixtures with $p \asymp n$, the info. in the perpendicular space (in red) can not be consistently estimated. (Lower Bound 2)

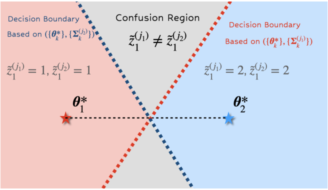
Insights into Barrier of Covariance Estimation

Clustering error is represented by whether the first example is correctly clustered. Imagine we have the access to $\mathbf{Y}$, $\{z_i^*\}_{i=2}^n$.

Insights into Barrier of Covariance Estimation

Clustering error is represented by whether the first example is correctly clustered. Imagine we have the access to $\mathbf{Y}$, $\{z_i^*\}_{i=2}^n$.

We can find M ϵ-packing-like parameter tuples with the same centers and different covariances: $\{(\{\boldsymbol{\theta}_k^*\}_{k\in[2]}, \{\boldsymbol{\Sigma}_k^{(j)}\}_{k\in[2]})\}_{j\in[M]}$.
$\Longrightarrow$ M different likelihood ratio estimators $\{\widetilde{z}^{(j)}\}$, each corresponding to a decision boundary.



large p
$\Rightarrow$ large M
$\overset{\text{Fano}}{\Rightarrow}$ $p_e > \frac{1}{2}$ (in multiple testing)
$\Rightarrow$ Unable to distinguish $j \in [M]$
$\Rightarrow$ Error must occurs in confusion region
$\Rightarrow$ misclust. prob. $\geq \exp(-(1+o(1))\frac{\mathrm{SNR}^{\mathrm{partial}\,2}}{2})$

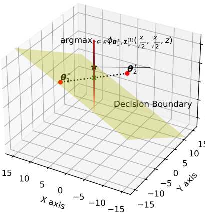
An Illustrative Example in $\mathbb{R}^3$

Two-Component Mixtures in $\mathbb{R}^3$ ($p = 3$, $K = 2$) with two sets of para. $\{\boldsymbol{\theta}_k^*, \boldsymbol{\Sigma}_k^{(1)}\}_{k\in[2]}$ and $\{\boldsymbol{\theta}_k^*, \boldsymbol{\Sigma}_k^{(2)}\}_{k\in[2]}$:

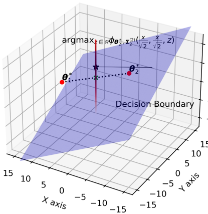
$$\boldsymbol{\theta}_1^* = (x, 0, 0)^\top, \quad \boldsymbol{\theta}_2^* = (0, x, 0)^\top,$$

$$\boldsymbol{\Sigma}_1^{(1)} = \boldsymbol{\Sigma}_2^{(1)} = \begin{pmatrix} 1 & 0 & c \\ 0 & 1 & -c \\ c & -c & 1 \end{pmatrix}, \quad \boldsymbol{\Sigma}_1^{(2)} = \boldsymbol{\Sigma}_2^{(2)} = \begin{pmatrix} 1 & 0 & -c \\ 0 & 1 & c \\ -c & c & 1 \end{pmatrix}$$

Submatrix in $\mathbb{R}^{(p-K)\times K}$ represents the complexity of covariance matrix



(a) Case 1



(b) Case 2

Price to Pay for Misspecifying the Covariance Matrix

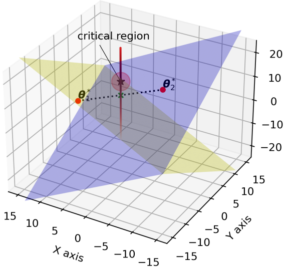
What if we misspecify Case 1 as Case 2? i.e., what is the outcome of using the wrong decision boundary?

Price to Pay for Misspecifying the Covariance Matrix

What if we misspecify Case 1 as Case 2? i.e., what is the outcome of using the wrong decision boundary?

Consider classification task

- Decision Boundaries ϕ_1^*, ϕ_2^*
- Case 1: optimal risk attained by ϕ_1^*
- When wrongly using ϕ_2^*:



$$\mathbb{P}_{\text{case1}}[\phi_2^* \neq z^*] = \text{optimal risk} + \text{constant} \times \underbrace{\text{density of crit. reg.}}_{\asymp \exp(-(1+o(1))\frac{{\text{SNR}^{\text{partial}}}^2}{2}) \gg \text{optimal risk}}$$

What happens in High Dimensions

To apply minimax framework, we need exponentially many hard-to-distinguish cases (to translate it into a multiple testing problem).

What happens in High Dimensions

To apply minimax framework, we need exponentially many hard-to-distinguish cases (to translate it into a multiple testing problem).

- The previous example in $\mathbb{R}^3$ only represents two-case testing problem

What happens in High Dimensions

To apply minimax framework, we need exponentially many hard-to-distinguish cases (to translate it into a multiple testing problem).

- The previous example in $\mathbb{R}^3$ only represents two-case testing problem
- Note that the corr. submat. can be represented by a vec. in $\mathbb{S}^{p-K-1}$.
- By the existence of an almost orthogonal vector set on $\mathbb{S}^{p-K-1}$, we can construct exponentially many hard-to-distinguish cases with similar *critical region* among every pair :)
- The density within each critical region is approximately $\exp(-(1+o(1))\frac{\mathrm{SNR}^{\mathrm{partial}^2}}{2})$!

What happens in High Dimensions

To apply minimax framework, we need exponentially many hard-to-distinguish cases (to translate it into a multiple testing problem).

- The previous example in $\mathbb{R}^3$ only represents two-case testing problem
- Note that the corr. submat. can be represented by a vec. in $\mathbb{S}^{p-K-1}$.
- By the existence of an almost orthogonal vector set on $\mathbb{S}^{p-K-1}$, we can construct exponentially many hard-to-distinguish cases with similar *critical region* among every pair :)
- The density within each critical region is approximately $\exp(-(1+o(1))\frac{\mathrm{SNR}^{\mathrm{partial}^2}}{2})$!

It hints that

impossibility of distinguishing hard cases

$$\Rightarrow \text{a raise of risk by } \exp(-(1+o(1))\frac{\mathrm{SNR}^{\mathrm{partial}^2}}{2})$$

Proof Overview: Reduction Framework

Step 1: Reduction from Minimax Risk to Local Risk.

$$\inf_{\widehat{\mathbf{z}}} \sup_{(\mathbf{z}^*,(\boldsymbol{\theta}_1^*,\boldsymbol{\theta}_2^*,\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2))} \mathbb{E}[h(\widehat{\mathbf{z}},\mathbf{z}^*)] \gtrsim \inf_{\widehat{z}_1} \text{Classify. Err. of the first sample}$$

Proof Overview: Reduction Framework

Step 1: Reduction from Minimax Risk to Local Risk.

$$\inf_{\widehat{\mathbf{z}}} \sup_{(\mathbf{z}^*, (\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2))} \mathbb{E}[h(\widehat{\mathbf{z}}, \mathbf{z}^*)] \gtrsim \inf_{\widehat{z}_1} \text{Classify. Err. of the first sample}$$

Step 2: Reduction from Local Risk to Discrepency between two LRTs.

Given an ϵ-packing-like parameter tuple collection $\{(\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\Sigma}_1^{(j)}, \boldsymbol{\Sigma}_2^{(j)})\}_{j \in [M]}$, we have

$$\inf_{\widehat{z}_1} \text{Classify. Err.} \gtrsim \min_{j_1 \neq j_2 \in [M]} \text{diff. between } \phi_{j_1}^* \text{ and } \phi_{j_2}^*,$$

where ϕ_j^* is the LRT for the j-th parameter $\{\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\Sigma}_1^{(j)}, \boldsymbol{\Sigma}_2^{(j)}\}$.

A Glimpse at Proof Techniques

Consider a weighted misclustering error instead

$$l(\mathbf{z}, \mathbf{z}^*) := \sum_{i \in n} \left\langle \mathbf{V}^{*\top}(\boldsymbol{\theta}^*_{z_i} - \boldsymbol{\theta}^*_{z_i^*}),\ \mathbf{S}^{*\,-1}_{z_i}\mathbf{V}^{*\top}(\boldsymbol{\theta}^*_{z_i} - \boldsymbol{\theta}^*_{z_i^*}) \right\rangle \mathbb{1}_{\{z_i \neq z_i^*\}}.$$

One-Step Analysis [Gao and Zhang, 2022, Chen and Zhang, 2024]

$$l(\widehat{\mathbf{z}}^{(t)}, \mathbf{z}^*) \leq \underbrace{\xi_{\text{oracle}}}_{\text{oracle error}} + \underbrace{\frac{1}{4} l(\widehat{\mathbf{z}}^{(t-1)}, \mathbf{z}^*)}_{\text{remnant effect from the last step}},$$

where ξ_{oracle} represents the weighted misclustering error given the true centers and projected covariance matrices

Consequence: after $O(\log n)$ steps, $l(\widehat{\mathbf{z}}^{(t)}, \mathbf{z}^*)$ is on the same order as ξ_{oracle}, which is $\exp(-(1+o(1))\mathrm{SNR}^{\text{partial}^2}/2)$.