High-Dimensional Sparse Clustering via Iterative Semidefinite Programming Relaxed K-Means

Presenter:Jongmin Mun

November 19, 2025

Sparse Clustering Problem

Goal: Partition unlabeled observations $\mathbf{X}_1, \dots, \mathbf{X}_n \in \mathbb{R}^p$ into clusters $G_1, \dots, G_K$ ($n \ll p$, known K)

Sparse Gaussian Mixture Model:

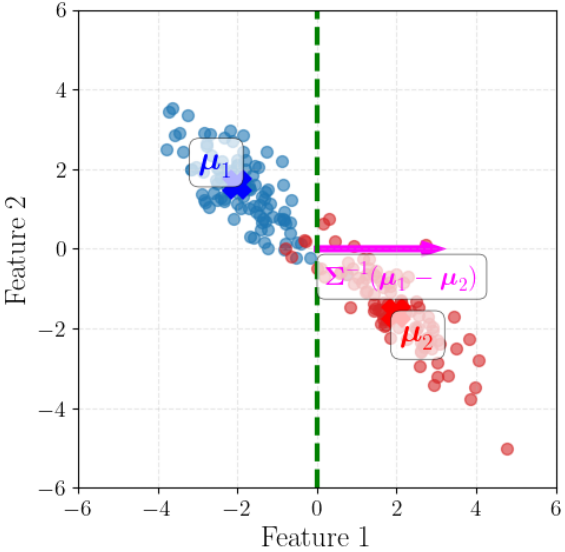
- Each observation $\mathbf{X}_i$ belongs to a cluster G_k with mean $\boldsymbol{\mu}_k \in \mathbb{R}^p$:

$$\mathbf{X}_i \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}) \quad \text{if} \quad \mathbf{X}_i \in G_k$$

- Fisher LDA boundary between G_k and G_ℓ: $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)$
- Sparsity assumption: there exist **signal features** and **noise features**

$$S_0 := \bigcup_{k \neq \ell} \operatorname{supp}(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)), \quad \{1, \dots, p\} \setminus S_0$$

Sparsity assumption

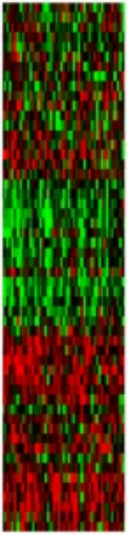


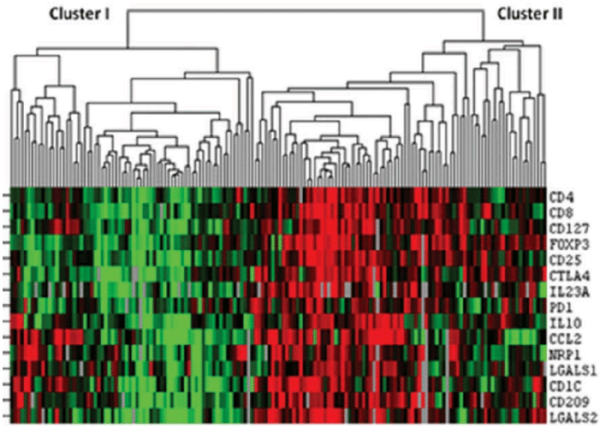
$\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \begin{pmatrix} -4 \\ 3.2 \end{pmatrix}$ is **not sparse**.

$$\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}^{-1} \begin{pmatrix} -4 \\ 3.2 \end{pmatrix}$$

$= \begin{pmatrix} -4 \\ 0 \end{pmatrix}$ is **sparse**.

Example: disease subtype discovery from gene expression





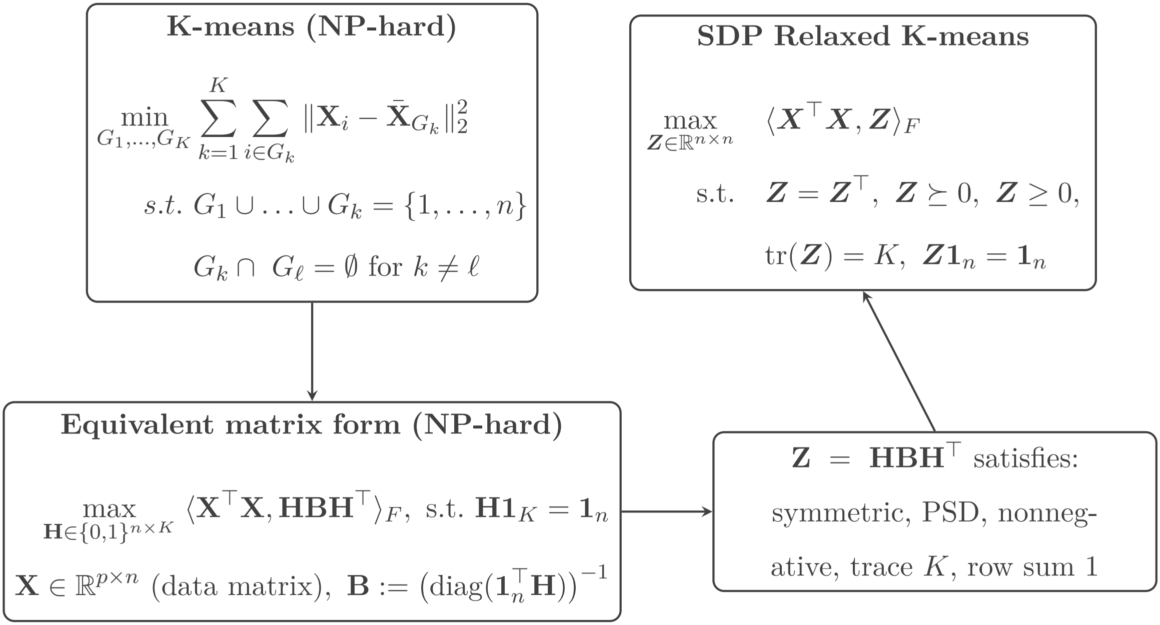
Example: disease subtype discovery from gene expression

	Our method	IF-PCA	SKM	SAS
Leukemia	**0.93**	0.84	0.79	0.87

- Leukemia dataset: $n = 45$, $p = 3871$, $K = 2$
- Cluster data with labels hidden, evaluate accuracy with true labels
- Baselines: IF-PCA (feature selection $\rightarrow$ clustering), SKM and SAS (iteratively alternate over feature selection and clustering)

Our approach

- **SDP K-means (Peng and Wei, 2005)**
  - Avoids explicit cluster-center estimation
  - Minimax optimal in fixed-dimension, non-sparse regimes **(Chen et al., 2021)**
- **Our contributions**
  - Motivating theory: Extend the analysis of Chen et al. 2021 to sparse setting to study the role of sparsity and feature selection on SDP K-means
  - Extend SDP K-means into an iterative, sparsity-aware algorithm for known covariance setting
  - Extend our algorithm into unknown covariance setting, using the high-dimensional precision matrix estimation tool


K-means (NP-hard)
$\min_{G_1,\dots,G_K} \sum_{k=1}^{K} \sum_{i \in G_k} \|\mathbf{X}_i - \bar{\mathbf{X}}_{G_k}\|_2^2$
$s.t.\ G_1 \cup \ldots \cup G_k = \{1, \ldots, n\}$
$G_k \cap\ G_\ell = \emptyset$ for $k \neq \ell$
SDP Relaxed K-means
$\max_{\boldsymbol{Z} \in \mathbb{R}^{n \times n}} \quad \langle \boldsymbol{X}^\top \boldsymbol{X}, \boldsymbol{Z} \rangle_F$
s.t. $\boldsymbol{Z} = \boldsymbol{Z}^\top,\ \boldsymbol{Z} \succeq 0,\ \boldsymbol{Z} \geq 0,$
$\mathrm{tr}(\boldsymbol{Z}) = K,\ \boldsymbol{Z}\mathbf{1}_n = \mathbf{1}_n$
Equivalent matrix form (NP-hard)
$\max_{\mathbf{H} \in \{0,1\}^{n \times K}} \langle \mathbf{X}^\top \mathbf{X}, \mathbf{H}\mathbf{B}\mathbf{H}^\top \rangle_F,$ s.t. $\mathbf{H}\mathbf{1}_K = \mathbf{1}_n$
$\mathbf{X} \in \mathbb{R}^{p \times n}$ (data matrix), $\mathbf{B} := \left(\mathrm{diag}(\mathbf{1}_n^\top \mathbf{H})\right)^{-1}$
$\mathbf{Z} = \mathbf{H}\mathbf{B}\mathbf{H}^\top$ satisfies:
symmetric, PSD, nonneg-
ative, trace K, row sum 1

True cluster in combinatorial problem is block-diagonal

Let $G_1^*, \ldots, G_K^*$ be true clusters.

The corresponding decision variable is:

$$\mathbf{Z}^* = \mathbf{H}^*\mathbf{B}^*\mathbf{H}^{*\top} =$$

$$\begin{bmatrix} \frac{1}{|G_1^*|}\mathbf{1}_{|G_1^*|\times|G_1^*|} & 0 & \cdots & 0 \\ 0 & \frac{1}{|G_2^*|}\mathbf{1}_{|G_2^*|\times|G_2^*|} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{|G_K^*|}\mathbf{1}_{|G_K^*|\times|G_K^*|} \end{bmatrix}$$

In practice:

Run spectral clustering on $\hat{\mathbf{Z}}$

Optimal point of SDP problem:

$\hat{\mathbf{Z}}$ = Continuous matrix s.t. Symmetric, PSD, nonnegative, trace K, row sum 1

Theory (Chen et al. 2021):

If $\min_{1\le k\neq \ell\le K}\left\|\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_k\right\|_2^2$ is large enough, with high probability, $\hat{\mathbf{Z}}$ is exactly $\mathbf{Z}^*$

For any feature subset $S \subseteq [p]$,

Let $\hat{\mathbf{Z}}(S)$ denote the solution of the SDP corresponding to S:

$$\max_{\mathbf{Z} \in \mathbb{R}^{n \times n}} \quad \langle \mathbf{X}_{S,\cdot}^{\top} \mathbf{X}_{S,\cdot}, \mathbf{Z} \rangle_F \quad \text{s.t.} \quad \mathbf{Z} = \mathbf{Z}^{\top},\ \mathbf{Z} \succeq 0,\ \operatorname{tr}(\mathbf{Z}) = K, \mathbf{Z}\mathbf{1}_n = \mathbf{1}_n,\ \mathbf{Z} \geq 0.$$

Collection of strong signal feature subsets:

$$\mathcal{S} := \left\{ S \subset [p] : \min_{1 \leq k \neq \ell \leq K} \left\| (\boldsymbol{\mu}_\ell - \boldsymbol{\mu}_k)_{S \cap S_0} \right\|_2^2 \gtrsim \left(\log n + \frac{|S| \log p}{n} + \sqrt{\frac{|S| \log p}{n}} \right) \right\}$$

Theorem (Uniform recovery of restricted SDPs)

Assume $\boldsymbol{\Sigma} = \mathbf{I}_p$. Then for any distribution instance in our model,

$$\mathbb{P}\left(\hat{\mathbf{Z}}(S) = \mathbf{Z}^*,\ \forall S \in \mathcal{S} \right) \gtrsim 1 - \frac{K}{n}.$$

Theorem (Tightness of the required separation)

Assume $\mathbf{\Sigma} = \mathbf{I}_p$. There exists a distribution instance in our model such that

1. $\min_{1\le k\ne\ell\le K} \left\|(\boldsymbol{\mu}_\ell^* - \boldsymbol{\mu}_k^*)_{S_0}\right\|_2^2 = C\log n$, *where C is a constant,*
2. *For any clustering method f,* $\mathbb{P}\big(f(\mathbf{X}_{S_0,\cdot}) \ne \{G_1^*, \dots, G_K^*\}\big) \gtrsim 1 - \frac{1}{n}$

If we consider uniform recovery problem restricted to moderately sized subsets satisfying $|S| \lesssim (n\log n)/\log p$:

$$\log n + \frac{|S|\log p}{n} + \sqrt{\frac{|S|\log p}{n}} \asymp \log n.$$

Then SDP K-means is optimal in terms of required separation

Intuition from the theorem

- Simple scenario: $K = 2$, $(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*)_{S_0} = \mu_0 \mathbf{1}_{|S_0|}$, and $|S| \lesssim (n \log n)/\log p$
- Exact recovery for all S is possible if and only if

$$|S \cap S_0|\mu_0^2 \gtrsim \log n + \frac{|S|\log p}{n} + \sqrt{\frac{|S|\log p}{n}}.$$

Insights:

1. Features should be chosen based on $\min_{1\le k\neq \ell\le K}\left\|(\boldsymbol{\mu}_\ell^* - \boldsymbol{\mu}_k^*)_{S\cap S_0}\right\|_2^2$
2. Mild under- or over-selection is acceptable; severe misselection is harmful
3. Once the algorithm reaches a high-signal subset S, reliable clustering follows

Our Iterative Method under identity covariance

Initialize

Initial clusters

$\hat{G}_1^{(0)}$ and $\hat{G}_2^{(0)}$

Feature Selection

Given: $\hat{G}_1^{(t-1)}$ and $\hat{G}_2^{(t-1)}$

estimate $\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*$ by $\bar{\mathbf{X}}_{\hat{G}_1^{(t-1)}} - \bar{\mathbf{X}}_{\hat{G}_2^{(t-1)}}$

$$\hat{S}^t := \left\{ j \in [p] : \left| \left(\bar{\mathbf{X}}_{\hat{G}_1^{(t-1)}} - \bar{\mathbf{X}}_{\hat{G}_2^{(t-1)}} \right)_j \right| > \sqrt{\frac{2n\log(2p)}{|\hat{G}_1^{(t-1)}||\hat{G}_2^{(t-1)}|}} \right\}.$$

Until convergence

Cluster Update

Given: selected features $\hat{S}^{(t)}$

Solve $\max_{\boldsymbol{Z}\in\mathbb{R}^{n\times n}} \langle \boldsymbol{X}_{\hat{S}^{(t)},\cdot} \boldsymbol{X}_{\hat{S}^{(t)},\cdot}^\top, \boldsymbol{Z} \rangle$

$$\text{s.t. } \boldsymbol{Z}^\top = \boldsymbol{Z}, \boldsymbol{Z} \succeq 0, \operatorname{tr}(\boldsymbol{Z}) = K, \boldsymbol{Z}\mathbf{1}_n = \mathbf{1}_n, \boldsymbol{Z} \geq 0.$$

Return $\hat{G}_1^{(t)}$ and $\hat{G}_2^{(t)}$

Extension to Unknown Common Covariance

Assumptions:

- Each row of $\mathbf{\Sigma}^{-1}$ has at most J nonzero off-diagonal entries
- There exists a small subset of relevant features $S_0 \subseteq \{1, \ldots, p\}$ with $|S_0| \ll p$, such that

$$S_0 := \bigcup_{k \neq \ell} \operatorname{supp}\Big(\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_\ell)\Big) \subset [p].$$

Feature selection step: Replace $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ estimation with

$$\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

Unknown Covariance: Clustering Step

SDP K-Means for general $\boldsymbol{\Sigma}$ (Zhuang et al 2023), without sparsity:

$$\max_{\boldsymbol{Z}\in\mathbb{R}^{n\times n}} \quad \left\langle (\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^\top \boldsymbol{\Sigma}(\boldsymbol{\Sigma}^{-1}\boldsymbol{X}),\ \boldsymbol{Z} \right\rangle$$

$$\text{s.t.} \quad \boldsymbol{Z}^\top = \boldsymbol{Z}, \boldsymbol{Z} \succeq 0, \operatorname{tr}(\boldsymbol{Z}) = K, \boldsymbol{Z}\mathbf{1}_n = \mathbf{1}_n, \boldsymbol{Z} \geq 0.$$

Unknown Covariance: Clustering Step

Given selected features $\hat{S}^{(t)}$, we solve SDP with sub-matrices:

$$\max_{\boldsymbol{Z}\in\mathbb{R}^{n\times n}} \quad \left\langle \left(\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{\top}_{\hat{S}^{(t)},\cdot} \boldsymbol{\Sigma}_{\hat{S}^{(t)},\hat{S}^{(t)}} \left(\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)_{\hat{S}^{(t)},\cdot},\ \boldsymbol{Z} \right\rangle$$

$$\text{s.t.} \quad \boldsymbol{Z}^{\top} = \boldsymbol{Z}, \boldsymbol{Z} \succeq 0, \mathrm{tr}(\boldsymbol{Z}) = K, \boldsymbol{Z}\mathbf{1}_n = \mathbf{1}_n, \boldsymbol{Z} \geq 0$$

What We Need to Estimate

For both the feature selection and clustering steps, the key quantity required for the extension is

$$\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \quad \text{and} \quad \boldsymbol{\Sigma}^{-1}\boldsymbol{X}.$$

We *do not* need to explicitly estimate the full precision matrix $\boldsymbol{\Sigma}^{-1}$.

Our approach: nodewise regression

We adapt the **Innovated Scalable Efficient Estimation (ISEE; Fan and Lv 2016)**

Idea: Partition the feature indices $[p]$ into disjoint subsets $A_1, A_2, \ldots, A_m$. For each subset A, estimate

$$\boldsymbol{\Sigma}^{-1}\mathbf{X} = \begin{array}{|c|} \hline (\boldsymbol{\Sigma}^{-1}\mathbf{X})_{A,\cdot} \\ \hline \cdots \\ \hline \cdots \\ \hline \cdots \\ \hline \end{array} \qquad \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k = \begin{array}{|c|} \hline (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k)_{A,\cdot} \\ \hline \cdots \\ \hline \cdots \\ \hline \cdots \\ \hline \end{array}$$

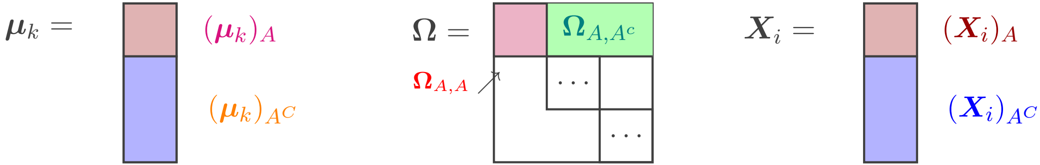
by nodewise regression.

Nodewise regression

By multivariate Gaussian assumption $\boldsymbol{X}_i \overset{iid}{\sim} N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$,

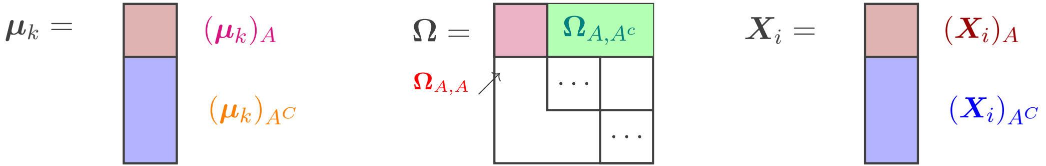
$$\underbrace{(\boldsymbol{X}_i)_A}_{\text{response}} = \underbrace{(\boldsymbol{\mu}_k)_A + \boldsymbol{\Omega}_{A,A}^{-1}\boldsymbol{\Omega}_{A,A^c}(\boldsymbol{\mu}_k)_{A^c}}_{\text{intercept}} - \underbrace{\boldsymbol{\Omega}_{A,A}^{-1}\boldsymbol{\Omega}_{A,A^c}}_{\text{slope}} \boldsymbol{X}_{A^c,i}$$

$$+ \underbrace{\boldsymbol{E}_{A,i}}_{\text{residual}}, \text{ where } \boldsymbol{E}_{A,i} \sim N(\mathbf{0}, \boldsymbol{\Omega}_{A,A}^{-1}).$$



Nodewise regression

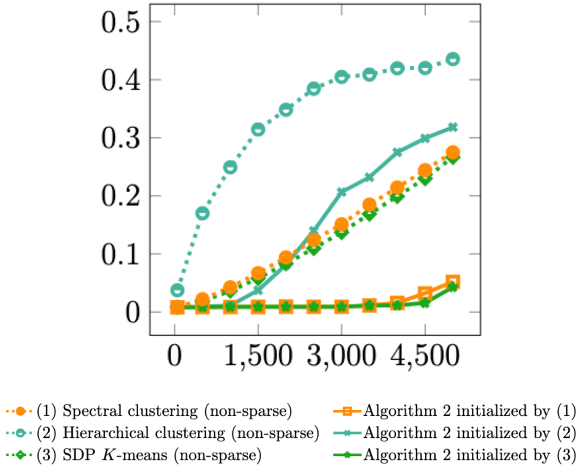
- $(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k)_A = \underbrace{\boldsymbol{\Omega}_{A,A}}_{\text{Cov(residual)}} \underbrace{\left((\boldsymbol{\mu}_k)_A + \boldsymbol{\Omega}_{A,A}^{-1}\boldsymbol{\Omega}_{A,A^c}(\boldsymbol{\mu}_k)_{A^c}\right)}_{\text{intercept}}$
- $(\boldsymbol{\Sigma}^{-1}\mathbf{X}_i)_A = (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k)_A + \underbrace{\boldsymbol{\Omega}_{A,A}}_{\text{Cov(residual)}} \underbrace{\mathbf{E}_{A,i}}_{\text{residual}}.$



Our Iterative Method under sparse precision matrix

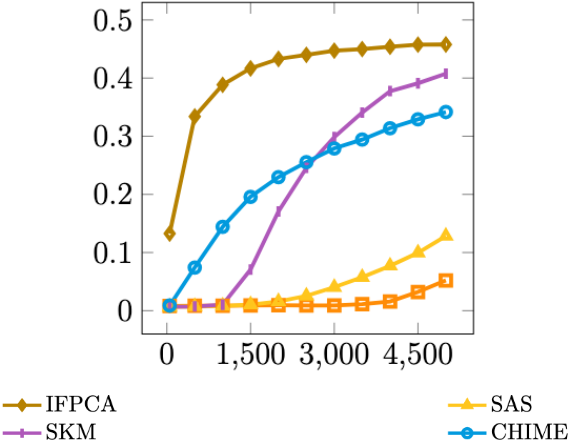
1. **Initialize:** Obtain initial cluster assignments $\hat{G}_1^0, \hat{G}_2^0 \subset [n]$.
2. **Iterate for** $t = 0, 1, 2, \ldots,$ until convergence:
   2.1 **ISEE subroutine**: Given $\hat{G}_1^{(t)}$ and $\hat{G}_2^{(t)}$, estimate $\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*)$, $\mathbf{\Sigma}^{-1}\mathbf{X}$,
   2.2 **Feature selection:** Let $\hat{S}^{t+1}$ be features where estimated $\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*)$ vector has large magnitude.
   Threshold defined by ℓ_2 convergence rate of ISEE
   2.3 **Cluster update:** Run SDP-relaxed K-means on the selected features $\tilde{\boldsymbol{X}}_{\hat{S}^{t+1},\cdot}$. $\mathbf{\Sigma}_{\hat{S}^{t+1},\hat{S}^{t+1}}$ to estimate new clusters $\hat{G}_1^{t+1}, \hat{G}_2^{t+1}$.

Simulations: identity covariance



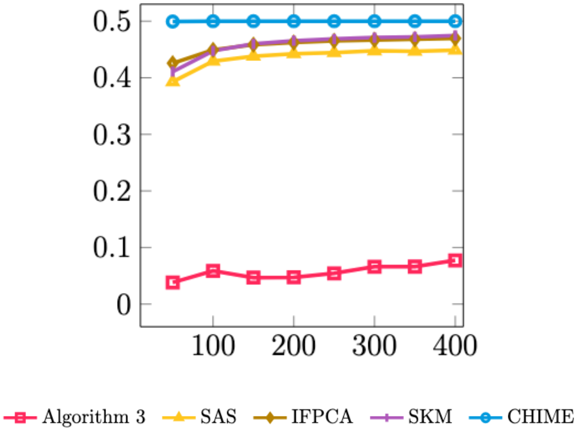
- X-axis: Dimension p increases while $|S_0| = 10$ and $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2 = 5^2$ are fixed
- Y-axis: mis-clustering rate
- Our method improves upon the non–sparsity-aware baseline.
- Dependence on initial clustering is mild.

Simulations: identity covariance



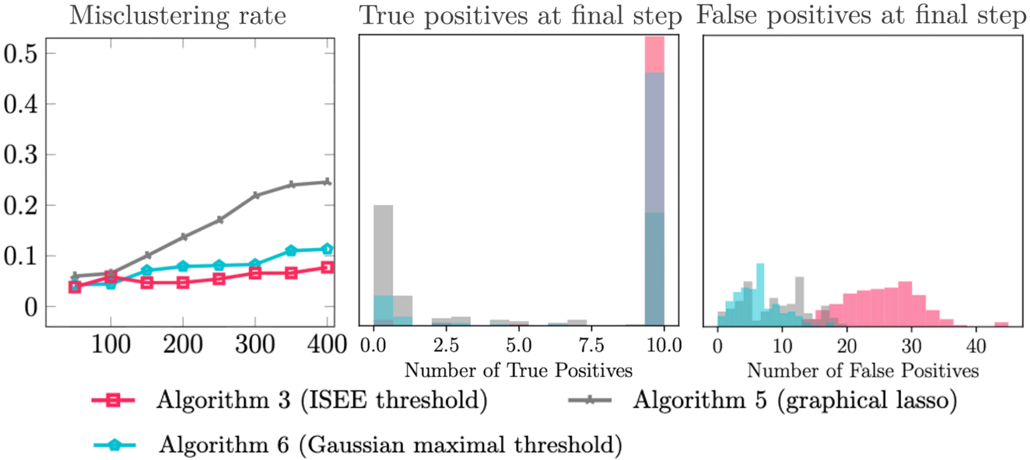
- X-axis: Dimension p increases while $|S_0| = 10$ and $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2 = 5^2$ are fixed
- Y-axis: mis-clustering rate
- Our method outperforms existing two-step and iterative methods

Simulations: effectiveness of ISEE and threshold



- Σ^{-1}: chain graph with correlation 0.45
- X-axis: Dimension p increases while $|S_0| = 10$, $\|\Sigma^{-1}(\mu_1 - \mu_2)\|_2^2 = 4^2$ fixed
- Y-axis: mis-clustering rate
- Our method outperforms existing two-step and iterative methods

Simulations: sparse precision matrix



Real Data Analysis

	Our method	IFPCA	SKM	SAS
Leukemia	**0.93**	0.84	0.79	0.87
MNIST	**0.94**	0.61	0.57	0.56

- Leukemia: $n = 45$, $p = 3871$
- MNIST: $n = 1000$, $p = 784$, digits 1 and 7
- Evaluation metric: clustering accuracy

Conclusion

Sparsity-aware iterative clustering combining convex relaxation, feature selection, and high-dimensional precision estimation.

- Theory-guided:
  - SDP K-means achieves simultaneous exact recovery on feature subsets with sufficient signal; signal requirement is optimal under mild assumptions.
  - Mild under- or over-selection is acceptable; aggressive misselection is harmful.
- Algorithm highlights:
  - Alternates between feature selection (via estimated Fisher LDA) and clustering (SDP K-means).
  - Nodewise regression (ISEE) avoids full precision matrix estimation.

Future Extensions

- Key insight: Mild under- or over-selection is tolerable, but aggressive misselection is harmful
- Current limitation: Past clustering and feature selection results are not explicitly utilized
- Proposed improvements: Introduce explicit exploration and memory
  - Use Thompson sampling for randomized feature selection
  - Update Beta distributions to retain memory of past results
  - Random draws from the Beta distributions encourage exploration

Prelimary simulation for Thompson sampling approach