Leave-one-out Singular Subspace Perturbation Analysis for Spectral Clustering

Anderson Ye Zhang

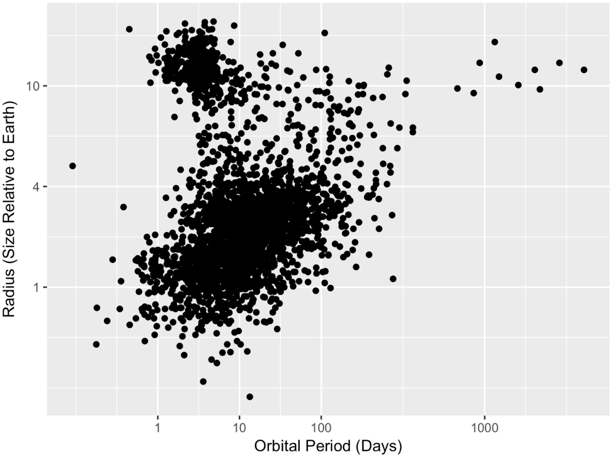
Department of Statistics and Data Science
University of Pennsylvania

Outline

- Spectral Clustering
- Existing Results
- A Novel Singular Subspace Perturbation Bound
- Spectral Clustering Revisit

Spectral Clustering

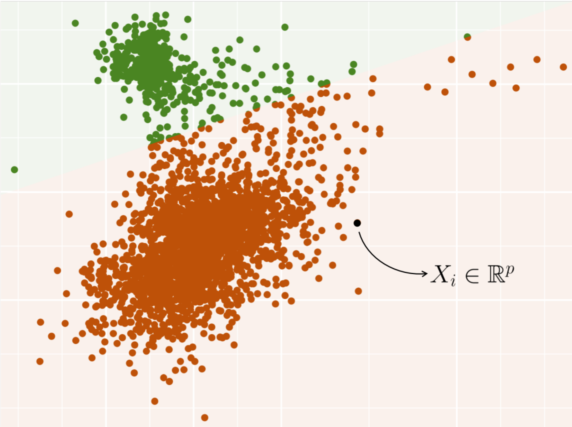
Clustering



Exoplanets: Orbital Period vs. Radius

Data Source: NASA Exoplanet Archive (https://exoplanetarchive.ipac.caltech.edu)

Clustering

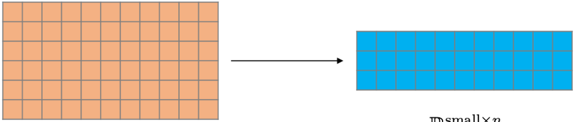


- Data matrix: $X = (X_1, X_2, \ldots, X_n) \in \mathbb{R}^{p \times n}$
- Perform clustering methods on $\{X_i\}_{i=1}^n$ (i.e., k-means)

Clustering

- When the dimension p is large, clustering directly on $\{X_i\}_{i=1}^n \in \mathbb{R}^p$ is computationally expensive.

- Natural idea: dimension reduction.



$\mathbb{R}^{\text{small}\times n}$

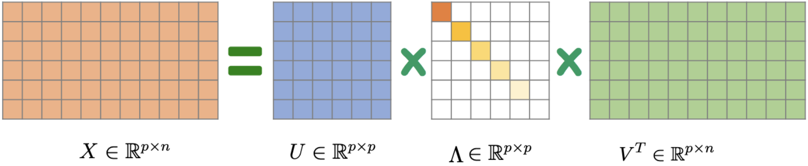
$X \in \mathbb{R}^{p\times n}$

- Spectral Clustering: Spectral Decomposition + Clustering

Spectral Clustering

Input: Data matrix $X \in \mathbb{R}^{p \times n}$, number of clusters k

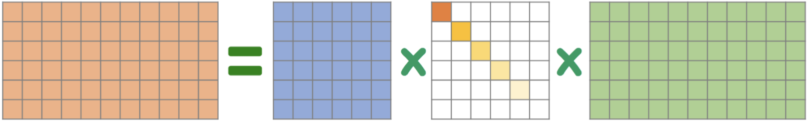
1. Perform SVD on X to have $X = \sum_{i=1}^{p} \lambda_i u_i v_i^T$.



Spectral Clustering

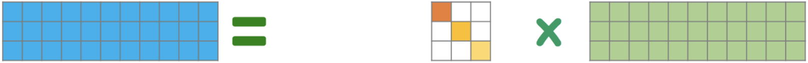
Input: Data matrix $X \in \mathbb{R}^{p \times n}$, number of clusters k

1. Perform SVD on X to have $X = \sum_{i=1}^{p} \lambda_i u_i v_i^T$.



$X \in \mathbb{R}^{p \times n}$ $\quad U \in \mathbb{R}^{p \times p}$ $\quad \Lambda \in \mathbb{R}^{p \times p}$ $\quad V^T \in \mathbb{R}^{p \times n}$

2. Let $\Lambda_k = \text{diag}\{\lambda_1, \ldots, \lambda_k\} \in \mathbb{R}^{k \times k}$ and $V_k = (v_1, \ldots, v_k)$ $\in \mathbb{R}^{n \times k}$. Define $X^{\text{low}} = \Lambda_k V_k^T \in \mathbb{R}^{k \times n}$.

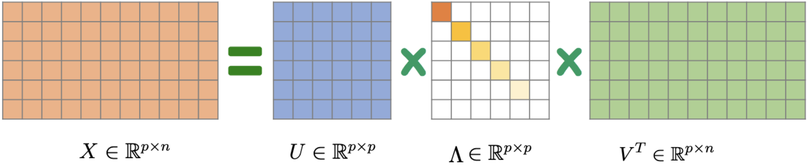


$X^{\text{low}} = \Lambda_k V_k^T \in \mathbb{R}^{k \times n}$ $\quad \Lambda_k \in \mathbb{R}^{k \times k}$ $\quad V_k^T \in \mathbb{R}^{k \times n}$

Spectral Clustering

Input: Data matrix $X \in \mathbb{R}^{p\times n}$, number of clusters k

1. Perform SVD on X to have $X = \sum_{i=1}^{p} \lambda_i u_i v_i^T$.



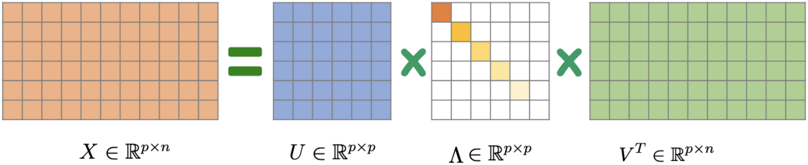
2. Let $\Lambda_k = \mathrm{diag}\{\lambda_1, \ldots, \lambda_k\} \in \mathbb{R}^{k\times k}$ and $V_k = (v_1, \ldots, v_k)$ $\in \mathbb{R}^{n\times k}$. Define $X^{\text{low}} = \Lambda_k V_k^T \in \mathbb{R}^{k\times n}$.

3. Apply k-means on columns of X^{low}.

Spectral Clustering

Input: Data matrix $X \in \mathbb{R}^{p \times n}$, number of clusters k

1. Perform SVD on X to have $X = \sum_{i=1}^{p} \lambda_i u_i v_i^T$.



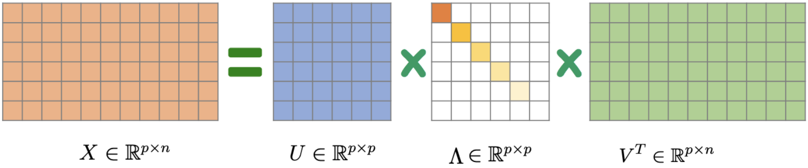
2. Let $\Lambda_k = \text{diag}\{\lambda_1, \ldots, \lambda_k\} \in \mathbb{R}^{k \times k}$ and $V_k = (v_1, \ldots, v_k)$ $\in \mathbb{R}^{n \times k}$. Define $X^{\text{low}} = \Lambda_k V_k^T \in \mathbb{R}^{k \times n}$.

Remark: Singular vectors are weighted as they are not equally important.

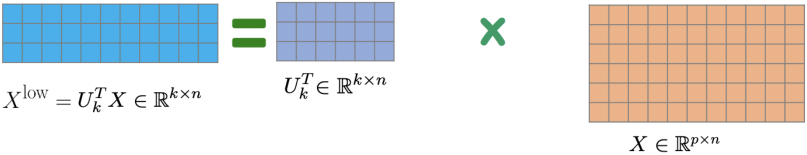
Equivalent Representation

Input: Data matrix $X \in \mathbb{R}^{p \times n}$, number of clusters k

1. Perform SVD on X to have $X = \sum_{i=1}^{p} \lambda_i u_i v_i^T$.



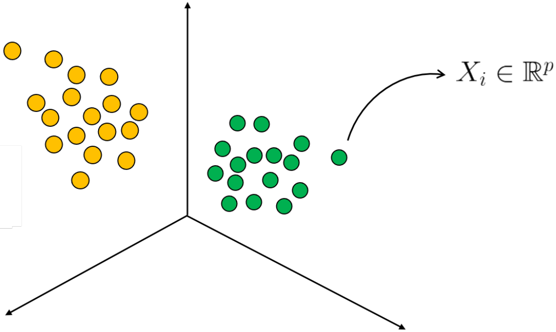
2. Let $U_k = (u_1, \ldots, u_k) \in \mathbb{R}^{n \times k}$. Then $X^{\text{low}} = U_k^T X \in \mathbb{R}^{k \times n}$.



3. Apply k-means on columns $\{U_k^T X_i\}_{i=1}^{n} \in \mathbb{R}^k$.

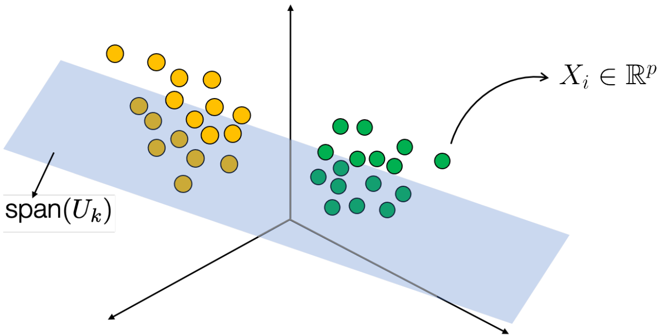
Projection

$$X_i \in \mathbb{R}^p \to U_k^T X_i \in \mathbb{R}^k$$



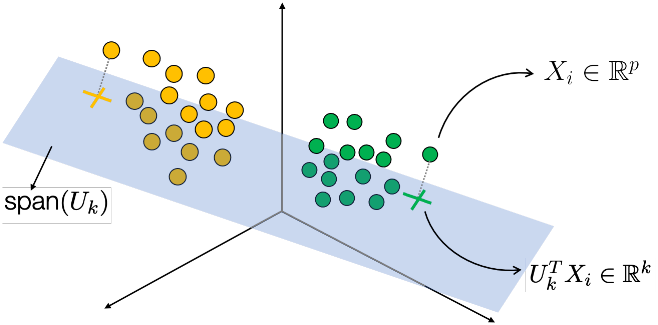
Projection

$$X_i \in \mathbb{R}^p \to U_k^T X_i \in \mathbb{R}^k$$



Projection

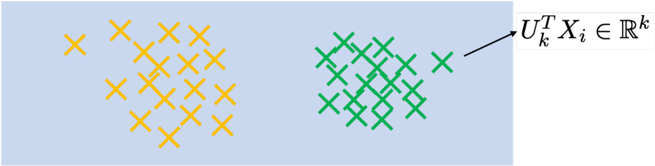
$$X_i \in \mathbb{R}^p \to U_k^T X_i \in \mathbb{R}^k$$



Spectral Clustering

Input: Data matrix $X \in \mathbb{R}^{p \times n}$, number of clusters k

1. Perform SVD on X to have $X = \sum_{i=1}^{p} \lambda_i u_i v_i^T$.
2. Let $U_k = (u_1, \ldots, u_k) \in \mathbb{R}^{n \times k}$.

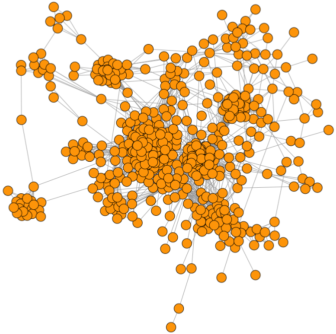


3. Apply k-means on $\{U_k^T X_i\}_{i=1}^{n}$.

Spectral Clustering

- is computationally appealing
- often has remarkably good performance
- has been widely used in various problems

Single Cell Analysis



Networks

Q: Why does spectral clustering work?

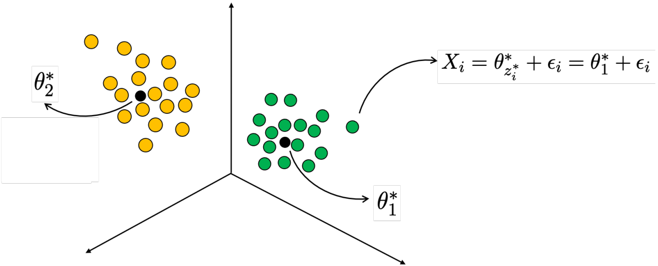
Existing Results

Mixture Models

- $X = (X_1, \ldots, X_n) \in \mathbb{R}^{p \times n}$
- k centers $\theta_1^*, \ldots, \theta_k^* \in \mathbb{R}^p$
- $z^* \in [k]^n$: underlying true cluster assignment vector
- The observations $\{X_i\}_{i \in [n]}$ are generated as follows:

$$X_i = \theta^*_{z_i^*} + \epsilon_i,$$

  where $\{\epsilon_i\}_{i=1}$ are noises.
- Goal: Recover the cluster assignment z^*



Mixture Models

For each $i \in [n]$, $X_i = \theta^*_{z_i^*} + \epsilon_i$

Matrix Form / Low-rank Structure:

$$\begin{aligned} X &= (X_1, \ldots, X_n) \\ &= (\theta^*_{z_1^*}, \ldots, \theta^*_{z_n^*}) && + (\epsilon_1, \ldots, \epsilon_n) \\ &=: \Theta^* \text{ (signal matrix)} && + E \text{ (noise matrix)} \end{aligned}$$

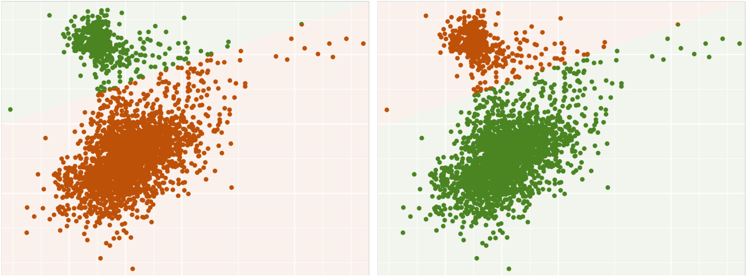
where $\Theta^* \in \mathbb{R}^{p \times n}$ is rank-k as it has k unique columns.

Mixture Models

- Loss $\ell(\hat{z}, z^*)$: the proportion of data points misclustered, considering all label permutations:

$$\ell(\hat{z}, z^*) = \frac{1}{n} \min_{\phi \in \Phi} \sum_{i \in [n]} \mathbb{I}\left\{\phi\left(\hat{z}_i\right) \neq z_i^*\right\},$$

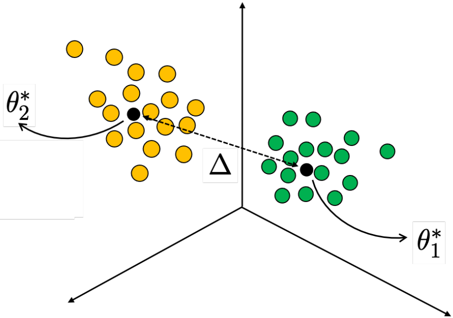
where $\Phi = \{\phi : \text{ bijection from } [k] \text{ to } [k]\}$.



Mixture Models

Signal Strength Δ: the minimum distance among centers:

$$\Delta = \min_{j,l \in [k]: j \neq l} \left\| \theta_j^* - \theta_l^* \right\| .$$



Assumptions

For simplicity, in this talk we assume:

- The number of clusters $k = O(1)$
- The cluster sizes are all in the same order
- The dimension $p \lesssim n$

Polynomial Error Rate

Recall the noise matrix $E = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$.

Proposition

We have

$$\ell(\hat{z}, z^*) \leq C \frac{\|E\|^2}{n\Delta^2}.$$

Remarks:

- Deterministic result. No assumption on the distribution of the noises $\{\epsilon_i\}$.
- No spectral gap condition on the signal matrix Θ^*.

Polynomial Error Rate

Proposition

We have

$$\ell(\hat{z}, z^*) \leq C \frac{\|E\|^2}{n\Delta^2}.$$

Consequence: For isotropic Gaussian mixtures where $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 I_p)$, then whp.

$$\ell(\hat{z}, z^*) \leq C \frac{\sigma^2}{\Delta^2}.$$

$\frac{\Delta^2}{\sigma^2}$: signal-to-noise ratio

Polynomial Error Rate

Consequence: For isotropic Gaussian mixtures where $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 I_p)$, then whp.

$$\ell(\hat{z}, z^*) \leq C \frac{\sigma^2}{\Delta^2}.$$

Minimax Rate for Clustering: If we consider all possible clustering methods, how small the clustering error can be?

$$\exp\left(-(1+o(1))\frac{1}{8}\frac{\Delta^2}{\sigma^2}\right)$$

In literature, spectral clustering is often used as an initialization for sophisticated algorithms to achieve the minimax rate.

Puzzling: But numerically such improvement is often marginal.

Q: Can we obtain a sharp upper bound for spectral clustering?

A Novel Singular Subspace Perturbation Bound

Classical Perturbation Theory

Two matrices $M, Y \in \mathbb{R}^{p \times n}$ where Y is a perturbation of M:

$$Y = M + E.$$

SVD:

$$M = \sum_{j \in [p \wedge n]} \sigma_j u_j v_j^T \text{ and } Y = \sum_{j \in [p \wedge n]} \hat{\sigma}_j \hat{u}_j \hat{v}_j^T,$$

where $\sigma_1 \geq \ldots \geq \sigma_{p \wedge n}$ and $\hat{\sigma}_1 \geq \ldots \geq \hat{\sigma}_{p \wedge n}$.

Left Singular Subspaces:

$$U_r = (u_1, \ldots, u_r) \text{ and } \hat{U}_r = (\hat{u}_1, \ldots, \hat{u}_r).$$

Classical Perturbation Theory

By Wedin's Theorem: if $\sigma_r - \sigma_{r+1} \geq 2\left\|(I - U_r U_r^T)E\right\|_F$, then

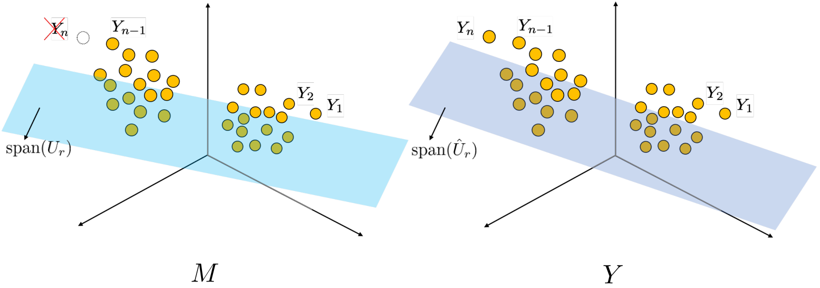
$$\|\hat{U}_r \hat{U}_r^T - U_r U_r^T\|_\mathsf{F} \leq \frac{2\sqrt{2}\left\|(I - U_r U_r^T)E\right\|_F}{\sigma_r - \sigma_{r+1}}.$$

However, this bound is tight in the worst case and sub-optimal in many settings.

Leave-one-column-out Perturbation

Consider

$$M = (Y_1, Y_2, \ldots, Y_{n-1}, 0) \text{ and } Y = (Y_1, Y_2, \ldots, Y_{n-1}, Y_n).$$



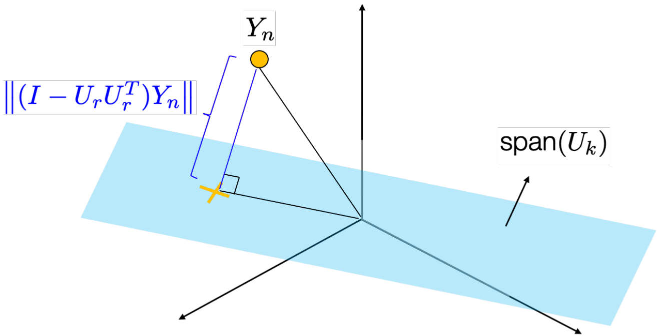
Leave-one-column-out Perturbation

Consider

$$M = (Y_1, Y_2, \ldots, Y_{n-1}, 0) \text{ and } Y = (Y_1, Y_2, \ldots, Y_{n-1}, Y_n).$$

Wedin's Theorem: if $\sigma_r - \sigma_{r+1} \geq 2 \left\| (I - U_r U_r^T) Y_n \right\|$, then

$$\|\hat{U}_r \hat{U}_r^T - U_r U_r^T\|_{\mathsf{F}} \leq \frac{2\sqrt{2} \left\| (I - U_r U_r^T) Y_n \right\|}{\sigma_r - \sigma_{r+1}}.$$



Leave-one-column-out Perturbation

Consider

$$M = (Y_1, Y_2, \dots, Y_{n-1}, 0) \text{ and } Y = (Y_1, Y_2, \dots, Y_{n-1}, Y_n).$$

Wedin's Theorem: if $\sigma_r - \sigma_{r+1} \geq 2 \left\| (I - U_r U_r^T) Y_n \right\|$, then

$$\|\hat{U}_r \hat{U}_r^T - U_r U_r^T\|_{\mathsf{F}} \leq \frac{2\sqrt{2} \left\| (I - U_r U_r^T) Y_n \right\|}{\sigma_r - \sigma_{r+1}}.$$

Theorem (Z., Zhou. 2022)

If $\sigma_r - \sigma_{r+1} \geq 2 \left\| (I - U_r U_r^T) Y_n \right\|$, then

$$\|\hat{U}_r \hat{U}_r^T - U_r U_r^T\|_{\mathsf{F}} \leq \frac{2\sqrt{2} \left\| (I - U_r U_r^T) Y_n \right\|}{\sigma_r - \sigma_{r+1}} \times 2 \sqrt{\sum_{j=1}^{r} \left(\frac{u_j^T Y_n}{\sigma_j} \right)^2}$$

Leave-one-column-out Perturbation

Consider

$$M = (Y_1, Y_2, \ldots, Y_{n-1}, 0) \text{ and } Y = (Y_1, Y_2, \ldots, Y_{n-1}, Y_n).$$

Wedin's Theorem: if $\sigma_r - \sigma_{r+1} \geq 2 \left\| (I - U_r U_r^T) Y_n \right\|$, then

$$\|\hat{U}_r \hat{U}_r^T - U_r U_r^T\|_{\mathsf{F}} \leq \frac{2\sqrt{2} \left\| (I - U_r U_r^T) Y_n \right\|}{\sigma_r - \sigma_{r+1}}.$$

Corollary

If $\sigma_r - \sigma_{r+1} \geq 2 \left\| (I - U_r U_r^T) Y_n \right\|$, then

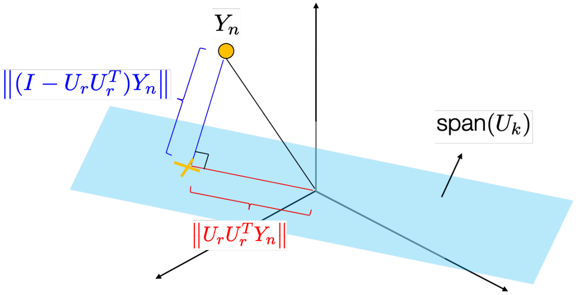
$$\|\hat{U}_r \hat{U}_r^T - U_r U_r^T\|_{\mathsf{F}} \leq \frac{2\sqrt{2} \left\| (I - U_r U_r^T) Y_n \right\|}{\sigma_r - \sigma_{r+1}} \times \frac{2 \left\| U_r U_r^T Y_n \right\|}{\sigma_r}$$

Leave-one-column-out Perturbation

Corollary

If $\sigma_r - \sigma_{r+1} \geq 2 \left\| (I - U_r U_r^T) Y_n \right\|$, then

$$\|\hat{U}_r \hat{U}_r^T - U_r U_r^T\|_{\mathsf{F}} \leq \frac{2\sqrt{2}\left\|(I - U_r U_r^T) Y_n\right\|}{\sigma_r - \sigma_{r+1}} \times \frac{2\left\|U_r U_r^T Y_n\right\|}{\sigma_r}$$



Remark: Its a deterministic result.

Spectral Clustering Revisit

Sub-Gaussian Mixtures

Theorem (Z., Zhou. 2022)

Assume $\epsilon_i \sim \mathrm{SG}_p(\sigma^2)$ independently and

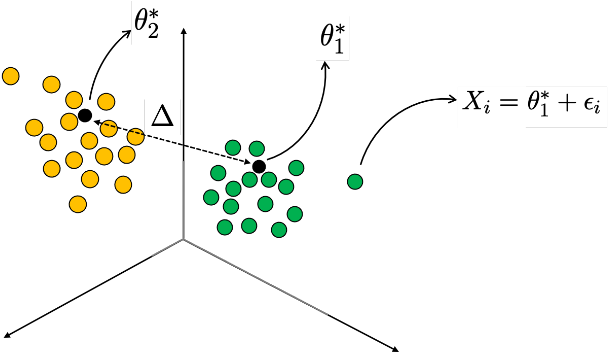
the kth largest singular value of the signal matrix Θ^* is $\geq C\sqrt{n}\sigma$.

With high probability we have

$$\ell(\hat{z}, z^*) \leq \exp\left(-(1-o(1))\frac{\Delta^2}{8\sigma^2}\right).$$

Entrywise Error Analysis

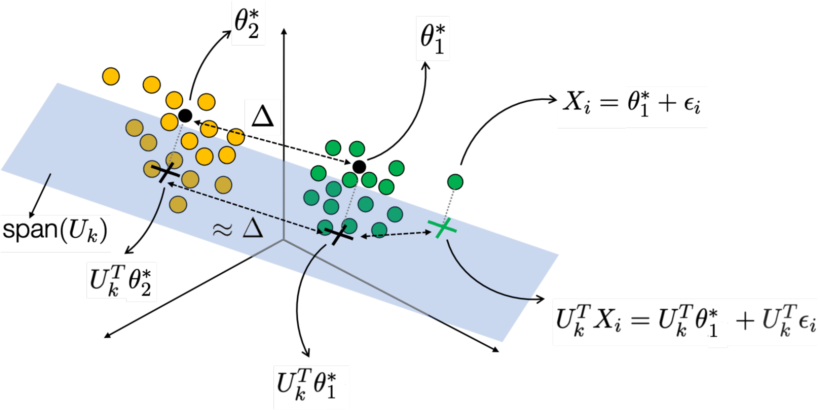
Consider X_i such that $z_i^* = 1$. Is X_i mis-clustered in the spectral clustering?



$$\mathbb{I}\left\{X_i \text{ is mis-clustered}\right\} \leq \mathbb{I}\left\{(1-o(1))\frac{\Delta}{2} \leq \left\|U_k^T \epsilon_i\right\|\right\}$$

Entrywise Error Analysis

Consider X_i such that $z_i^* = 1$. Is X_i mis-clustered in the spectral clustering?



$$\mathbb{I}\left\{X_i \text{ is mis-clustered}\right\} \leq \mathbb{I}\left\{(1-o(1))\frac{\Delta}{2} \leq \left\|U_k^T \epsilon_i\right\|\right\}$$

Entrywise Error Analysis

Since $X_i = \theta^*_{z_i^*} + \epsilon_i$, we have $U_k^T X_i = U_k^T \theta^*_{z_i^*} + U_k^T \epsilon_i$. We can show

$$\mathbb{I}\{X_i \text{ is mis-clustered}\} \leq \mathbb{I}\left\{(1-o(1))\frac{\Delta}{2} \leq \left\|U_k^T \epsilon_i\right\|\right\}$$
$$= \mathbb{I}\left\{(1-o(1))\frac{\Delta}{2} \leq \left\|U_k U_k^T \epsilon_i\right\|\right\}.$$

Need to decouple the dependence between $U_k U_k^T$ and ϵ_i. Recall $X = (X_1, \ldots, X_n)$. Define its leave-ith-column-out counterpart

$$X_{-i} = (X_1, \ldots, X_{i-1}, 0, X_{i+1}, \ldots, X_n)$$

and let $U_{k,-i} = (u_{1,-i}, \ldots, u_{k,-i})$ be its leading k left singular subspace.

$$\leq \mathbb{I}\left\{(1-o(1))\frac{\Delta}{2} \leq \left\|U_{k,-i} U_{k,-i}^T \epsilon_i\right\| + \left\|\left(U_k U_k^T - U_{k,-i} U_{k,-i}^T\right) \epsilon_i\right\|\right\}.$$

Entrywise Error Analysis

When $\epsilon_i \overset{iid}{\sim} \mathsf{SG}_d(\sigma^2)$, under the aforementioned singular gap condition, using our novel perturbation bound, we have

$$\begin{aligned}\left\|\left(U_k U_k^T - U_{k,-i} U_{k,-i}^T\right) \epsilon_i\right\| &\le \|U_k U_k^T - U_{k,-i} U_{k,-i}^T\|_{\mathsf{F}} \|\epsilon_i\| \\ &\le o(\Delta + \left\|U_{k,-i} U_{k,-i}^T \epsilon_i\right\|).\end{aligned}$$

Then

$$\begin{aligned}&\mathbb{I}\left\{X_i \text{ is mis-clustered}\right\} \\ &\le \mathbb{I}\left\{(1-o(1))\frac{\Delta}{2} \le \left\|U_{k,-i} U_{k,-i}^T \epsilon_i\right\| + \left\|\left(U_k U_k^T - U_{k,-i} U_{k,-i}^T\right) \epsilon_i\right\|\right\} \\ &\le \mathbb{I}\left\{(1-o(1))\frac{\Delta}{2} \le \left\|U_{k,-i} U_{k,-i}^T \epsilon_i\right\|\right\}.\end{aligned}$$

We have

$$\mathbb{P}\left((1-o(1))\frac{\Delta}{2} \le \left\|U_{k,-i} U_{k,-i}^T \epsilon_i\right\|\right) \le \exp\left(-(1-o(1))\frac{\Delta^2}{8\sigma^2}\right).$$

Sub-Gaussian Mixtures

Theorem (Z., Zhou. 2022)

Assume $\epsilon_i \overset{iid}{\sim} \mathrm{SG}_p(\sigma^2)$ and

the kth largest singular value of the signal matrix Θ^* is $\geq C\sqrt{n}\sigma$.

With high probability we have

$$\ell(\hat{z}, z^*) \leq \exp\left(-(1 - o(1))\frac{\Delta^2}{8\sigma^2}\right).$$

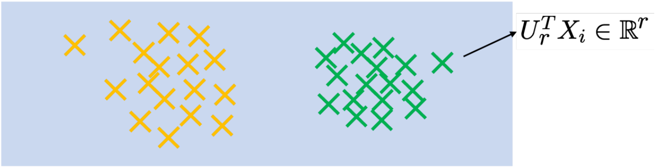
Q: Can we get rid of the spectral gap condition?

Spectral Clustering w. Adaptive Dimension Reduction

Idea: Use $r \leq k$ singular subspace $U_r \in \mathbb{R}^{p \times r}$ instead of $U_k \in \mathbb{R}^{p \times k}$.

Input: Data matrix $X \in \mathbb{R}^{p \times n}$, number of clusters k

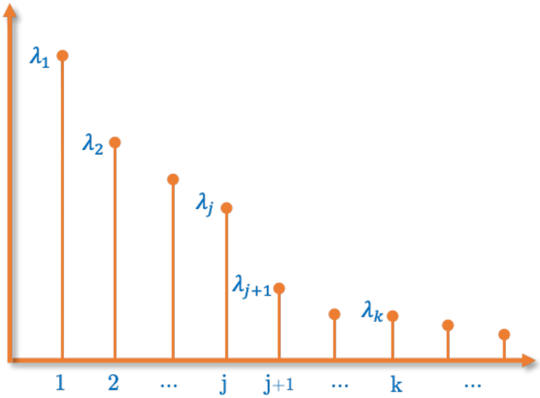
1. Perform SVD on X to have $X = \sum_{i=1}^{p} \lambda_i u_i v_i^T$.
2. Choose $r = \max\{j \leq k : \lambda_j - \lambda_{j+1} \geq \rho\sqrt{n}\sigma\}$.
3. Let $U_r = (u_1, \ldots, u_r) \in \mathbb{R}^{n \times r}$.



4. Apply k-means on $\{U_r^T X_i\}_{i=1}^{n}$.

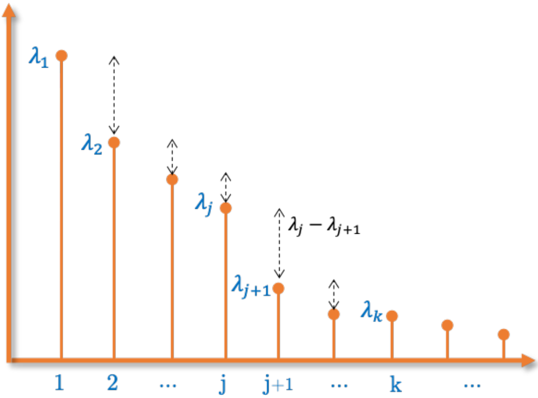
Adaptive Dimension Reduction

$$r = \max\{j \le k : \lambda_j - \lambda_{j+1} \ge \rho\sqrt{n}\sigma\}$$



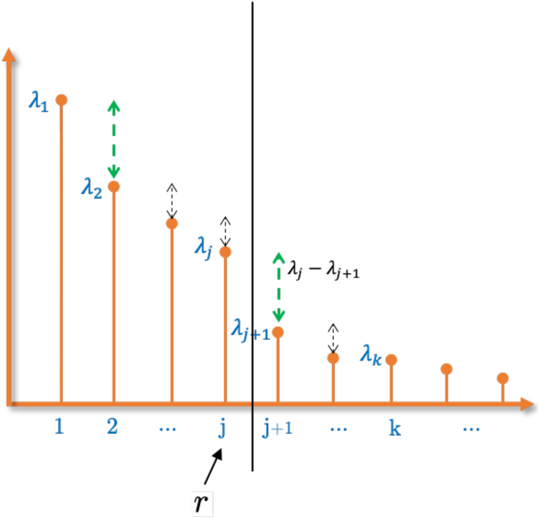
Adaptive Dimension Reduction

$$r = \max\{j \le k : \lambda_j - \lambda_{j+1} \ge \rho\sqrt{n}\sigma\}$$



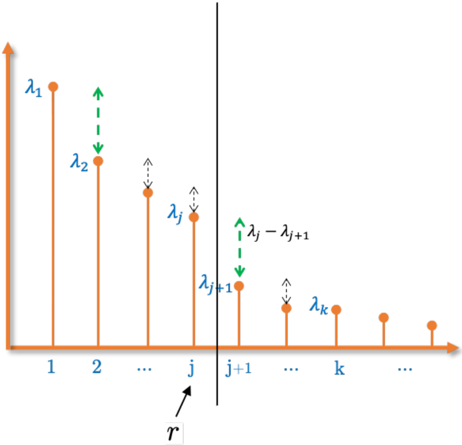
Adaptive Dimension Reduction

$$r = \max\{j \le k : \lambda_j - \lambda_{j+1} \ge \rho\sqrt{n}\sigma\}$$



Adaptive Dimension Reduction

$$r = \max\{j \le k : \lambda_j - \lambda_{j+1} \ge \rho\sqrt{n}\sigma\}$$



In this way, the spectral gap $\lambda_r - \lambda_{r+1}$ is large and λ_{r+1} is small. Hence, we can guarantee $u_1, \ldots, u_r$ are all important and $u_{r+1}, u_{r+2}, \ldots$ are all less important.

Spectral Clustering w. Adaptive Dimension Reduction

Idea: Use $r \leq k$ singular subspace $U_r \in \mathbb{R}^{p \times r}$ instead of $U_k \in \mathbb{R}^{p \times k}$.

Theorem (Z., Zhou. 2022)

Assume $\epsilon_i \overset{iid}{\sim} \mathsf{SG}_p(\sigma^2)$. With high probability we have

$$\ell(\hat{z}, z^*) \leq \exp\left(-(1 - o(1))\frac{\Delta^2}{8\sigma^2}\right),$$

if we select the reduced dimension r adaptively.

Special Case: Isotropic Gaussian Mixtures

Theorem (Z., Zhou. 2022)

Assume $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2 I_p)$. With high probability we have

$$\ell(\hat{z}, z^*) \leq \exp\left(-(1 - o(1))\frac{\Delta^2}{8\sigma^2}\right).$$

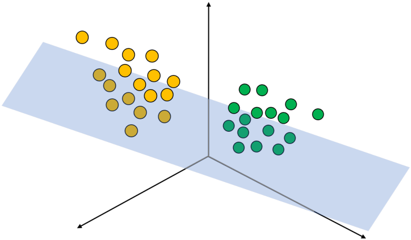
Remarks:

- Spectral clustering is optimal under the isotropic Gaussian mixture model as it achieve the minimax rate for clustering:

$$\exp\left(-(1 + o(1))\frac{\Delta^2}{8\sigma^2}\right)$$

- No spectral gap condition on Θ^*.

Summary



- Leave-one-out singular subspace perturbation
- Spectral clustering for sub-Gaussian mixtures

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

Thank You