MIXTURE OF COMMON SUBSPACE VIA RIEMANIANN GRADIENT: STATISTICAL AND COMPUTATIONAL LIMIT

Zhexu Jin, Joshua Agterberg

University of Illinois, Urbana Champaign

The Motivating Problem

Given L noisy symmetric low rank matrices (e.g. covariance matrices, networks), when can we cluster those matrices with the same subspace structure in a statistically consistent and computationally efficient manner?

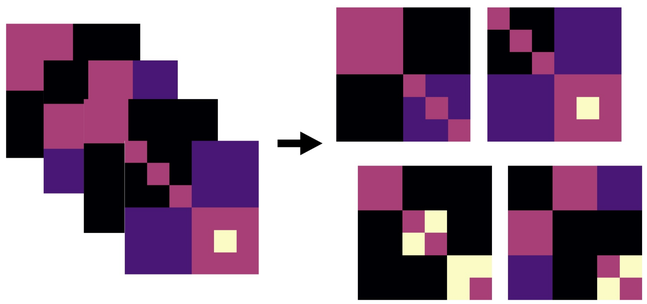

Fig. 1: Motivating Problem

Main Contributions

- Devise scalable initialization and riemannian gradient method for clustering mixture of common subspaces.
- Provide theoretical guarantee for clustering under a uniform signal to noise ratio condition across all layers.
- Derive computational and statistical lower bound for the uniform signal to noise ratio condition.

Mixture of Common Subspace Model

The model generalizes common subspace model [1]:

$$\boldsymbol{A}^{(l)} = \boldsymbol{S}^{(l)} + \boldsymbol{N}^{(l)} \qquad l \in [L]$$

Here

$$\boldsymbol{S}^{(l)} = \boldsymbol{U}^{(m)} \boldsymbol{R}^{(l)} \boldsymbol{U}^{(m)\top} \quad m \in [M] \quad m := \tau(l)$$

- signal strength λ:

$$\lambda := \sigma_{\min}(\boldsymbol{S}^{(l)}) = \sigma_{\min}(\boldsymbol{R}^{(l)})$$

- noise level σ :

$$\boldsymbol{N}_{ij}^{(l)} \begin{cases} \sim \mathrm{SG}(0, \sigma^2/2) & i > j \\ \sim \mathrm{SG}(0, \sigma^2) & i = j \\ = \boldsymbol{N}_{ji}^{(l)} & i < j \end{cases}$$

- subspace seperation:

$$\Delta := \min_{m \neq m'} \left\| \sin\Theta(\boldsymbol{U}^{(m)}, \boldsymbol{U}^{(m')}) \right\|$$

Initialization Algorithm

The unbiased estimator for $\boldsymbol{S}^{(l)2}$ is

$$\hat{\boldsymbol{S}^{(l)2}} := \boldsymbol{A}^{(l)2} - \frac{n+1}{2}\hat{\sigma}^2 \boldsymbol{I}_n$$

Algorithm 1 Clustering Initialization for Mixture of Common Subspace Model

Input: $\{\boldsymbol{A}^{(l)}\}_{l\in[L]}$, Matrix Cluster Number M

1: Compute $\widehat{\boldsymbol{\mathcal{U}}}$ as the top R eigenvectors of sum of squares $\boldsymbol{\mathcal{A}} := \sum_{l\in[L]} \boldsymbol{A}^{(l)2}$, denoted as $\widehat{\boldsymbol{\mathcal{U}}}$.

2: Estimate $\hat{\sigma}$ as (18)

3: For each $l \in [L]$, compute $\widehat{\boldsymbol{X}}^{(l)} = \widehat{\boldsymbol{\mathcal{U}}}^\top \widehat{\boldsymbol{S}^{(l)2}} \widehat{\boldsymbol{\mathcal{U}}} / \left\| \widehat{\boldsymbol{\mathcal{U}}}^\top \widehat{\boldsymbol{S}^{(l)2}} \widehat{\boldsymbol{\mathcal{U}}} \right\|_F$.

4: Run K-Means on $\{\widehat{\boldsymbol{X}}^{(l)}\}_{l\in[L]}$, output the cluster membership $\widehat{\tau}_0$.

Output: Estimated matrix cluster membership $\widehat{\tau}_0$

Riemannian Gradient Algorithm

Given a fixed $\hat{\tau}$, we minimize $f(\boldsymbol{U}^{(m)})$ for each m via Riemannian Gradient Descent on Stiefel manifold:

$$f(\boldsymbol{U}^{(m)}) := \sum_{l\in\hat{\tau}^{-1}(m)} \left\| \boldsymbol{U}^{(m)}\boldsymbol{U}^{(m)\top}\boldsymbol{A}^{(l)}\boldsymbol{U}^{(m)}\boldsymbol{U}^{(m)\top} - \boldsymbol{A}^{(l)} \right\|_F^2$$

$$\mathrm{grad}\, f(\boldsymbol{U}^{(m)}) := 4 \sum_{l\in\hat{\tau}^{-1}(m)} \left(\boldsymbol{U}^{(m)}\boldsymbol{U}^{(m)\top} - \boldsymbol{I} \right) \boldsymbol{A}^{(l)}\boldsymbol{U}^{(m)}\boldsymbol{U}^{(m)\top}\boldsymbol{A}^{(l)}\boldsymbol{U}^{(m)}$$

We optimize the nonconvex optimization above using an idea close to EM:

Algorithm 2 Riemannian Gradient Method for Subspace Refinement and Reclustering

Input: $\{\boldsymbol{A}^{(l)}\}_{l\in[L]}$, $\widehat{\tau}_0, r_m$

1: **while** $\frac{1}{L} h_c(\widehat{\tau}_k, \widehat{\tau}_{k+1}) \geq \epsilon_1$ and $k \leq K$ **do**

2: **for** $m \in [M]$ **do**

3: compute $\widehat{\boldsymbol{U}}_0^{(m)}$ as the top r_m eigenvectors of $\sum_{l\in\widehat{\tau}^{-1}(m)} \boldsymbol{A}^{(l)2}$.

4: **while** $\left\| \frac{1}{L_m} \mathrm{grad}\, f(\widehat{\boldsymbol{U}}_t^{(m)}) \right\|_F \geq \epsilon_2$ and $t \leq T$ **do**

5: $\widehat{\boldsymbol{U}}_{t+.5}^{(m)} = \widehat{\boldsymbol{U}}_t^{(m)} - \frac{\eta}{L} \mathrm{grad}\, f(\widehat{\boldsymbol{U}}_t^{(m)})$

6: $\widehat{\boldsymbol{U}}_{t+1}^{(m)} = \boldsymbol{Q}_{0.5}\boldsymbol{R}_{0.5}^\top$ where $\boldsymbol{Q}_{0.5}\boldsymbol{\Sigma}_{0.5}\boldsymbol{R}_{0.5}^\top = \widehat{\boldsymbol{U}}_{t+.5}^{(m)}$ is the singular value decomposition of $\boldsymbol{U}_{0.5}^{(m)}$.

7: Update $\widehat{\tau}_{k+1}$ such that $\widehat{\tau}_{k+1}(l) = m$ iff $m = \arg\min_{m\in[M]} f_l(\widehat{\boldsymbol{U}}_t^{(m)})$

Output: Estimated subspace $\left\{\widehat{\boldsymbol{U}}_t^{(m)}\right\}_{m\in[M]}$, and cluster $\widehat{\tau}_t$

Minimax Lower Bound

Define the loss function and parameter set

$$h_c(\hat{\tau}, \tau) := \inf_{\pi} \sum_{l\in[L]} \mathbb{I}\{\hat{\tau}(l) \neq \pi \circ \tau(l)\}. \quad \Theta(\lambda, n, L, r, M)$$

Theorem 1. *If we have when* $\frac{\lambda^2\Delta^2}{\sigma^2}(\log(\frac{\alpha L\sigma}{M\lambda\Delta}))^{-1} \to \infty$ *as* $n \to \infty$*, then it holds that*

$$\inf_{\hat{z}} \sup_{\theta\in\Theta} \mathbb{E}_{\mathbb{P}_\theta} \frac{1}{L} h_c(\hat{\tau}, \tau^*) \geq \exp\left(-(1+o(1)) \frac{\lambda^2\Delta^2}{8\sigma^2} \right).$$

Identifiablility Condition

Quotient by rotation of $\boldsymbol{U}^{(m)}$, permutation of τ, and $\left\| \boldsymbol{U}^{(m)\top}\boldsymbol{U}^{(m')} \right\|_F^2 < r$.

Computational Lower Bound

Under the low degree likelihood ratio framework. Consider the following test

$$H_0^{(L)} : \boldsymbol{A}^{(l)} = \epsilon^{(l)}\lambda \boldsymbol{u}\boldsymbol{u}^\top + \boldsymbol{Z} \quad \text{vs} \quad H_1^{(L)} : \boldsymbol{A}^{(l)} = \epsilon^{(l)}\lambda \left(b^{(l)} \boldsymbol{u}_\delta \boldsymbol{u}_\delta^\top + (1-b^{(l)}) \boldsymbol{u}\boldsymbol{u}^\top \right) \tag{1}$$

Theorem 2. *For the hypothesis testing above, if* $\lambda = o(\frac{n^{1/2}}{L^{1/4}\delta})$*, then* $\left\| \mathcal{L}_L^{\leq D} \right\|^2 = 1 + o(1)$*. (Suggests exponential computational complexity under this reigime)*

Initialization Guarantee

Theorem 3. *If dispersion condition is met and signal strength condition:*

$$\frac{\lambda}{\sigma} \gtrsim \left(\frac{M^{1/4}\kappa r^{1/2}}{\alpha^{1/4}} \vee \frac{M^{3/8}\kappa^{1/2}(rR)^{1/4}}{\alpha^{3/8} s^{1/4}} \right) \frac{n^{1/2}}{L^{1/4}\Delta}$$

With probability at least $1 - \exp(-cn)$*:*

$$\frac{1}{L} h_c(\tau^*, \hat{\tau}) \lesssim \frac{\alpha}{M} \frac{\Delta^2}{r\kappa^4} \lesssim 1$$

Simulation

For $M = 2$, we sample $\left\| \boldsymbol{U}_1^\top \boldsymbol{U}_2 \right\|_F^2 = 1.5$, $n = 200$, $r = 4$ and repeatedly cluster for 20 times

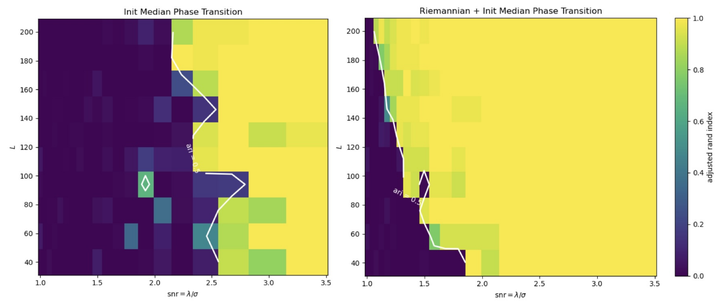

Fig. 2: Phase Transition for Initialization and Riemaniann Gradient Step

Comparison on Real Data

We consider Brain Heart Lung (BHL) gene profile dataset as analyzed by [2]. We cluster the covariance matrix $\boldsymbol{S} = \boldsymbol{M}^\top\boldsymbol{M}$. Hence, $n = 1124$, $L = 27$ and $r = 1$.

	GMCS (Our)	lr-Lloyd	DEEM	K-means	SKM	DTC	TBM	EM
Cluster Error	**3.70**	**3.70**	7.41	11.11	11.11	18.52	11.11	11.11

Table 1: GMCS: Gradient Mixture of Common Subspace; SKM: sparse K-means ; DTC: dynamic tensor clustering ; TBM: tensor block model (TBM) ; EM: standard EM;

[1] J. Arroyo, A. Athreya, J. Cape, G. Chen, C. E. Priebe, and J. T. Vogelstein, "Inference for multiple heterogeneous networks with a common invariant subspace," *Journal of Machine Learning Research*, vol. 22, no. 142, pp. 1–49, 2021.

[2] Z. Lyu and D. Xia, "Optimal clustering by Lloyd's algorithm for low-rank mixture model," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 88, no. 1, pp. 43–65, 2026.