# A convex clustering formulation using the similarity matrix

Yutong Wang[1] Laura Balzano[1] Clayton Scott[1] Venkatesh Saligrama[2]

[1]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor.

[2]Department of Electrical and Computer Engineering, Boston University.

## Background and problem

We consider the problem of clustering. Suppose our data is $p$-dimensional and suppose there are $K$ clusters. For each $i = 1, \dots, K$, let $n_i$ be the number of elements in cluster $i$. Let $n = n_1 + \cdots + n_K$.

### Generative model

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 & \cdots & \underbrace{\mathbf{X}_i}_{n_i \text{ columns}} & \cdots \mathbf{X}_K \end{bmatrix} \in \mathbb{R}^{p \times n}$$

sampled i.i.d from $\mathcal{D}_i$ a distribution on $\mathbb{R}^p$

with mean $\mu_i = \mathbb{E}_{X \sim \mathcal{D}_i}[X]$ and $v_i = \mathbb{E}_{X \sim \mathcal{D}_i}\|X\|^2 < \infty$

Note: We assume that the data matrix is partitioned as above to simplify presentation. Our method allows the input data matrix $\mathbf{X}$ with shuffled column.

Let $\mathbf{1}_{m \times n}$ denote the $m \times n$ matrix of all ones. Consider the $n \times n$ matrix

$$\mathbf{Z}_{\mathcal{P}} = \begin{bmatrix} n_1^{-1}\mathbf{1}_{n_1 \times n_1} & & & 0 \\ & n_2^{-1}\mathbf{1}_{n_2 \times n_2} & & \\ & & \cdots & \\ 0 & & & n_K^{-1}\mathbf{1}_{n_K \times n_K} \end{bmatrix} \in \mathbb{R}^{n \times n}$$

which reveals the cluster membership of columns of $\mathbf{X}$ since

$$\|\mathbf{Z}_{\mathcal{P}}(:, j) - \mathbf{Z}_{\mathcal{P}}(:, \ell)\|_1 = \begin{cases} 2 & : \ \mathbf{X}(:, j) \text{ and } \mathbf{X}(:, \ell) \text{ from same cluster} \\ 0 & : \ \text{otherwise.} \end{cases}$$

Thus, Kmeans clustering with $\ell_1$-distance on $\mathbf{Z}_{\mathcal{P}}$ yields the correct clustering of $\mathbf{X}$. However, we don't have access to $\mathbf{Z}_{\mathcal{P}}$. Our work focuses on recovering $\mathbf{Z}_{\mathcal{P}}$ using the *similarity matrix* $\mathbf{X}^T\mathbf{X}$ of $\mathbf{X}$. We consider the following optimization problem:

$$\mathbf{Z}^\star_{\mathbf{X}} \in \underset{\mathbf{Z} \in \mathbb{R}^{n \times n}}{\arg\max} \quad n^{-1} \operatorname{Tr}(\mathbf{Z}\mathbf{X}^T\mathbf{X}) \tag{1}$$

$$s.t. \quad \|\mathbf{Z}\|_* \le K, \tag{2}$$

$$\|\mathbf{Z}\|_{op} \le 1 \tag{3}$$

Empirically, we often observe that $\mathbf{Z}^\star_{\mathbf{X}} \approx \mathbf{Z}_{\mathcal{P}}$. In this work, we study the population version of (1):

$$\mathbf{Z}^\star = \underset{\mathbf{Z} \in \mathbb{R}^{n \times n}}{\arg\max} \quad n^{-1} \operatorname{Tr}(\mathbf{Z}\mathbb{E}_{\mathbf{X}}[\mathbf{X}^T\mathbf{X}]) \tag{4}$$

$$s.t. \quad \|\mathbf{Z}\|_* \le K, \tag{5}$$

$$\|\mathbf{Z}\|_{op} \le 1 \tag{6}$$

We show that under certain separation condition, $\mathbf{Z}^\star = \mathbf{Z}_{\mathcal{P}}$.

## Results

Our first theorem gives a sufficient criterion for $\mathbf{Z}^\star = \mathbf{Z}_{\mathcal{P}}$:

### Theorem

Let $\mathbf{V} = \operatorname{diag}(v_1, \dots, v_K)$, $\mathbf{M} = (\mu_i^T\mu_j)_{ij}$ and $\mathbf{N} = \operatorname{diag}(n_1, \dots, n_K)$. If

$$(K-1)^{(K-1)/2} \frac{\det(\mathbf{V}) + \det(\mathbf{N})\det(\mathbf{M})}{\left(\|\mathbf{V}\|_F + \|\mathbf{N}\|_F\|\mathbf{M}\|_F\right)^{(K-1)}} \ge \max\{v_1, \dots, v_K\} \tag{7}$$

then $\mathbf{Z}_{\mathcal{P}}$ is the unique maximizer of (4).

We also proved a result that computes the maximizer of (1) using SVD:

### Proposition

Let $\mathbf{Q} \in \mathbb{R}^{n \times n}$ be a symmetric PSD matrix with SVD $\mathbf{Q} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$. Let $\mathbf{\Sigma} = \operatorname{diag}(\sigma_1, \dots, \sigma_n)$ be such that $\sigma_i \ge \sigma_{i+1}$. Let $K$ be a positive integer with $K \le n$ and suppose that $\sigma_K > \sigma_{K+1}$. Then

$$\max_{\mathbf{Z} \in \mathbb{R}^{n \times n}} \operatorname{Tr}(\mathbf{Z}\mathbf{Q}) \quad s.t. \ \|\mathbf{Z}\|_* \le K, \quad \|\mathbf{Z}\|_{op} \le 1 \tag{8}$$

has a unique optimizer given by

$$\mathbf{Z}^\star = \mathbf{U} \operatorname{diag}(\underbrace{1, \dots, 1}_{K \text{ copies}}, 0, \dots, 0)\mathbf{U}^T$$

The main idea behind the proof of the proposition is to relate (8) to the following

$$\max_{(s_1, \dots, s_n) \in \mathbb{R}^n} \sum_{i=1}^n \sigma_i s_i \quad s.t. \ \sum_i |s_i| \le K, \quad |s_i| \le 1 \tag{9}$$

The consequence of the proposition is that the objective value of (8) is the $K$-th Ky-Fan norm of $Q$ (sum of the $K$ largest singular values).

## Algorithms

**Algorithm 1** SVD-based convex clustering

**Input:** data matrix $\mathbf{X}$, number of clusters $K$

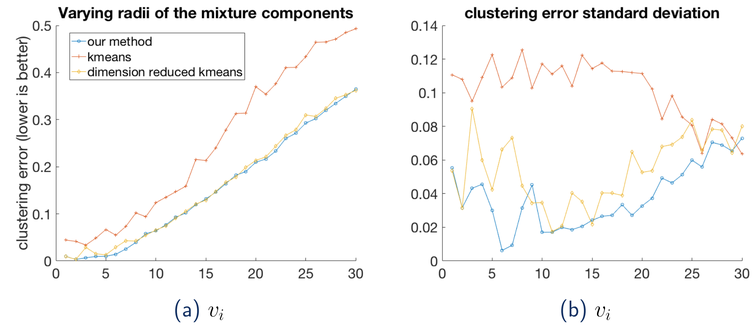
1: Compute SVD of $\mathbf{X}^T\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$ so that the diagonal of $\mathbf{\Sigma}$ is non-increasing.

2: $\mathbf{Z}^\star_{\mathbf{X}} \leftarrow \mathbf{U} \operatorname{diag}(\underbrace{1, \dots, 1}_{K \text{ copies}}, 0, \dots, 0)\mathbf{U}^T$.

**Output:** Clustering of the columns of $\mathbf{Z}^\star_{\mathbf{X}}$ by Kmeans with $\ell_1$-distance.

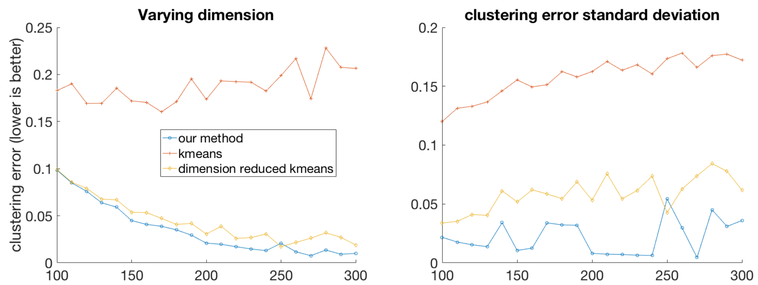
## Performance

For simulations, we used $K = 3$ spherical Gaussian mixture models, i.e., $\mathcal{D}_i = \mathcal{N}(\mu_i, (v_i/p)\mathbb{I})$. Hence, $v_i = \mathbb{E}_{X \sim \mathcal{D}_i}\|X\|^2$ is proportional to the radius of the $i$-th Gaussian mixture component for $p$ fixed. We compare our method against (1) *kmeans* applied directly on the data matrix $\mathbf{X}$, and (2) *dimension reduced kmeans*, i.e., kmeans applied to the first $K$ principle components of $\mathbf{X}$.




(a) $v_i$ (b) $v_i$

Figure: Above, we varied the radii of the Gaussian mixture components.




(a) Dimension $p$ (b) Dimension $p$

Figure: Above, we increased the dimension $p$ on the $x$-axis while keeping $v_i$ constant.

We note that our method performs on average as well as dimension reduced kmeans. However, our method had smaller standard deviation in the clustering error.

## Acknowledgements

ELECTRICAL ENGINEERING & COMPUTER SCIENCE
UNIVERSITY OF MICHIGAN

MIDAS
MICHIGAN INSTITUTE FOR DATA SCIENCE
UNIVERSITY OF MICHIGAN