University of Colorado
Boulder

A Randomized Approach to Efficient Kernel Clustering

Farhad Pourkamali-Anaraki and Stephen Becker

December 8, 2016

GlobalSIP 2016: Symposium on Compressed Sensing, Deep Learning

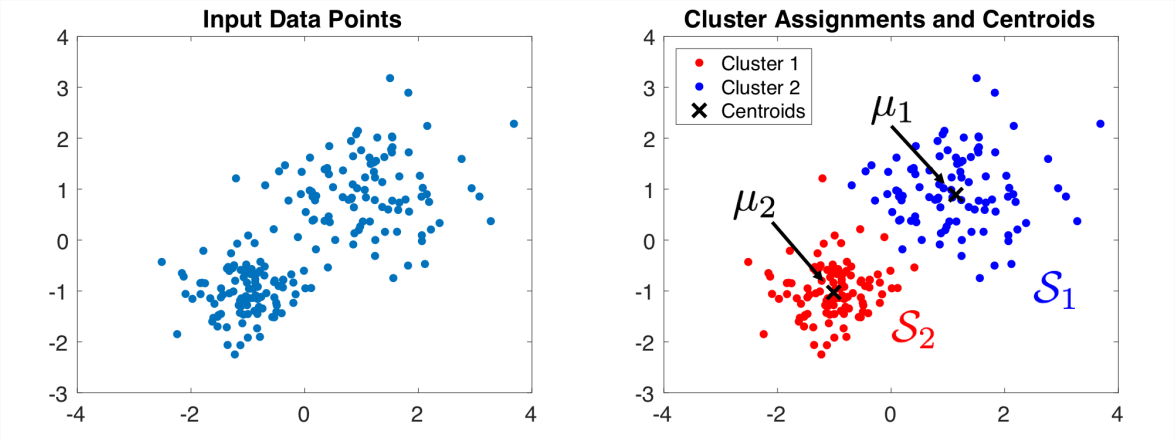
Outline

1. Kernel K-means Clustering
2. Our Randomized Method for Efficient Kernel K-means
3. Theoretical Analysis

Kernel K-means Clustering

K-means Clustering

- K-means Algorithm: Partitions $x_1, \ldots, x_n$ into K clusters



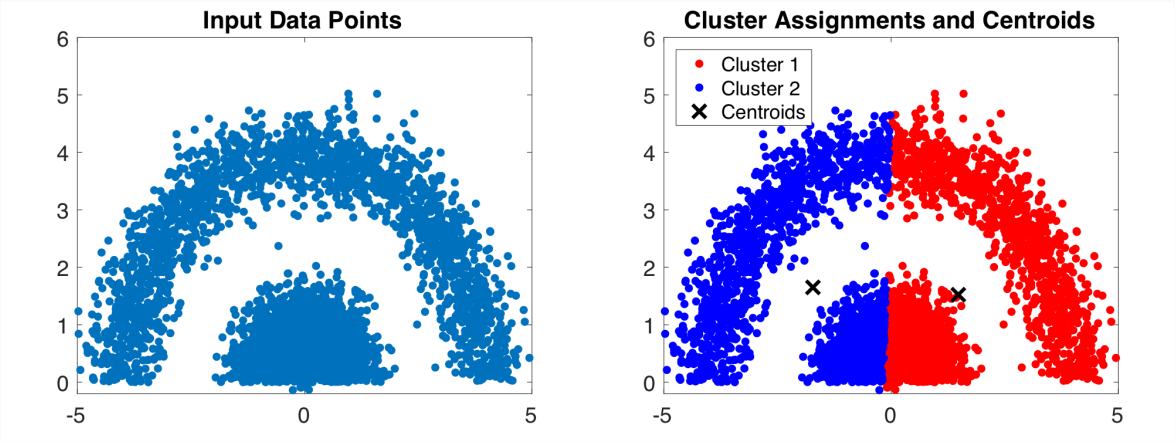
- K-means Objective: $t_{ik} \in \{0, 1\}$ binary indicator variable

$$\mathcal{F}(\mathcal{S}) = \sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik} \|x_i - \mu_k\|_2^2$$

- K-means works perfectly when clusters are linearly separable

Kernel Clustering

- K-means does not perform well on finding non-linearly separable clusters of varying densities and distributions



- Partitioning may be easier in a lifted space:

$$\text{non-linear mapping } \Phi : x_i \mapsto \underbrace{\Phi(x_i), \quad i = 1, \ldots, n}_{\text{linearly separable}}$$

- Kernel Trick:

$$\underbrace{\langle \Phi(x_i), \Phi(x_j) \rangle}_{\text{inner products}} = \underbrace{\kappa(x_i, x_j)}_{\text{kernel function}} , \quad \forall i, j \in \{1, \ldots, n\}$$

Kernel K-means

- Kernel K-means Objective:

$$\mathcal{L}(\mathcal{S}) = \sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik} \underbrace{\|\Phi(x_i) - \mu_k\|_2^2}_{\langle \Phi(x_i) - \mu_k, \Phi(x_i) - \mu_k \rangle}$$

- Kernel K-means requires access to the full kernel matrix:

$$\mathbf{K} = \begin{bmatrix} \kappa(x_1, x_1) & \dots & \kappa(x_1, x_n) \\ \vdots & \ddots & \\ \kappa(x_n, x_1) & \dots & \kappa(x_n, x_n) \end{bmatrix}$$

- Memory: $O(n^2)$

not scalable for large data sets

Our Randomized Method for Efficient Kernel K-means

Low-Rank Approximation of Kernel Matrices

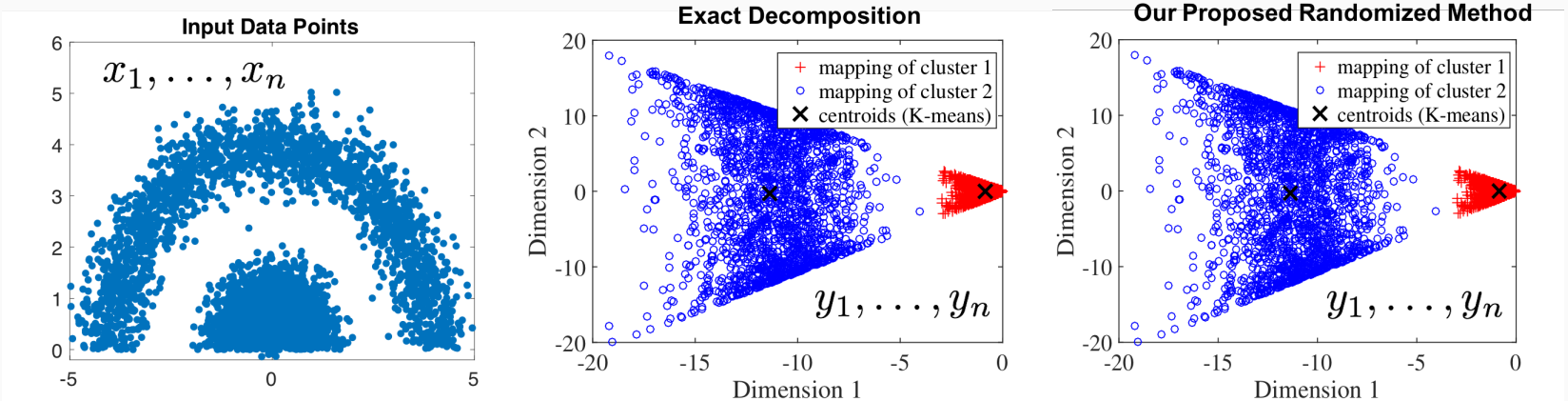
- Eigenvalue Decomposition:

$$\mathbf{K} \approx \mathbf{U}_r \Lambda_r \mathbf{U}_r^T = \left(\mathbf{U}_r \Lambda_r^{1/2}\right)\left(\Lambda_r^{1/2} \mathbf{U}_r^T\right) = \underbrace{\mathbf{Y}^T \mathbf{Y}}_{\text{linearization of } \mathbf{K}}, \quad \mathbf{Y} \in \mathbb{R}^{r \times n}$$

  - $\Lambda_r \in \mathbb{R}^{r \times r}$ and $\mathbf{U}_r \in \mathbb{R}^{n \times r}$: top n eigenvalues and eigenvectors
  - Requires r passes over $\mathbf{K}$

- Perform standard K-means on $\mathbf{Y} = [y_1, \ldots, y_n]$ in $\mathbb{R}^r$
  - Memory: $O(nr)$ vs. $O(n^2)$

- Our Method: Single pass over $\mathbf{K}$ to compute $y_1, \ldots, y_n$



$$\kappa(x_i, x_j) = \langle x_i, x_j \rangle^2, \quad r = 2$$

One-Pass Kernel K-means

1: $r' \leftarrow r + l, \quad \mathbf{R} \in \mathbb{R}^{n \times r'}$: random sampling matrix

2: $\mathbf{W} \in \mathbb{R}^{n \times r'} \leftarrow (\mathbf{R}^T \mathbf{HDK})^T$

- Preconditioning: $\mathbf{K} \mapsto \mathbf{HDK}$
- $\mathbf{H} \in \mathbb{R}^{n \times n}$: Hadamard matrix
- $\mathbf{D} \in \mathbb{R}^{n \times n}$: stochastic diagonal matrix with entries $\{\pm 1\}$

3: find an orthonormal matrix $\mathbf{Q} \in \mathbb{R}^{n \times r}$ by QR decomposition

4: solve $\mathbf{B}(\mathbf{Q}^T \mathbf{\Omega}) = (\mathbf{Q}^T \mathbf{W}), \quad \mathbf{\Omega} = \mathbf{DHR}$

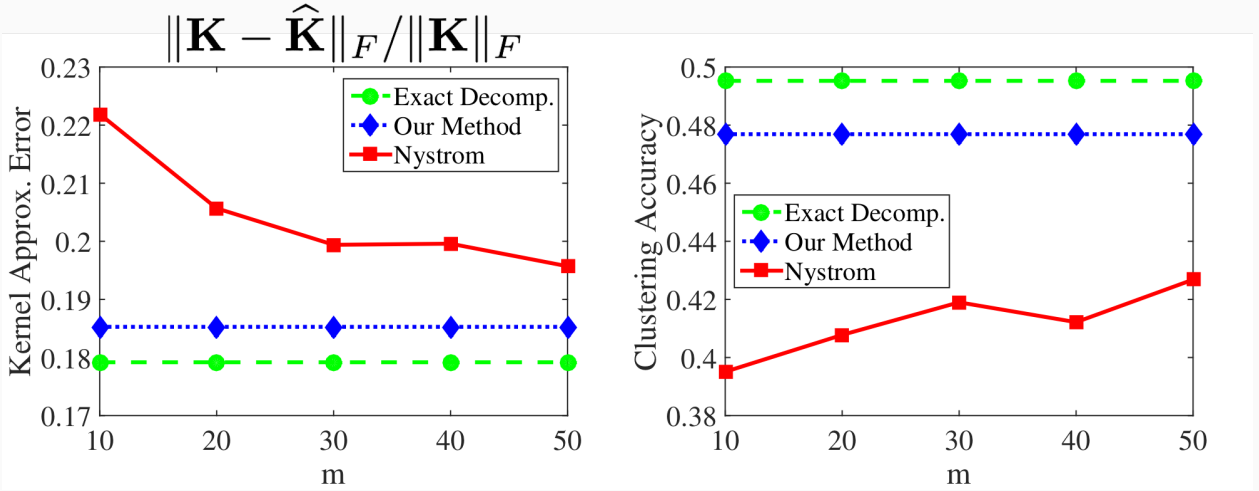
5: $\mathbf{B} = \mathbf{V \Sigma V}^T$

6: $\mathbf{Y} = \mathbf{\Sigma}^{1/2} \mathbf{V}^T \mathbf{Q}^T \in \mathbb{R}^{r \times n}$

7: perform standard K-means on $\mathbf{Y} = [y_1, \ldots, y_n]$

Experimental Evaluation

- Our proposed method ($r = 2$, $l = 5$) vs. Nyström method
- Nyström method: sampling m columns of $\mathbf{K}$ using uniform sampling
- "Image segmentation" data set with $n = 2310$ and $K = 7$



Sampling $r' = 7$ rows of the preconditioned $\mathbf{K}$ leads to a more accurate decomposition than sampling $m = 50 \approx 7r'$ columns of $\mathbf{K}$

Theoretical Analysis

Theorem

- Consider the kernel k-means objective function: $\mathcal{L}(\mathcal{S})$
- $\mathcal{S}^*$: optimal solution of Kernel K-means using the full kernel matrix
- $\widehat{\mathcal{S}}$: optimal solution of the approximate Kernel K-means

$$\mathbf{K} = \widehat{\mathbf{K}} + \underbrace{\mathbf{E}}_{\text{error}}$$

- Then, we have:

$$\mathcal{L}(\widehat{\mathcal{S}}) - \mathcal{L}(\mathcal{S}^*) \leq 2\|\mathbf{E}\|_*$$

where $\|\mathbf{E}\|_*$ represents the trace norm.

The optimal objective value under the low-rank approximation is not far from the true objective value