k-means clustering of Gaussian mixtures

Soledad Villar

University of Texas at Austin

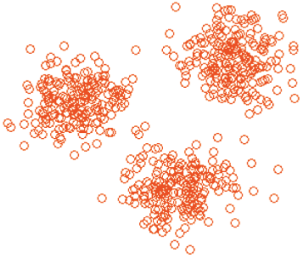
Based on work with:
Dustin Mixon (AFIT)
Rachel Ward (UT Austin)

SIAM Imaging

k-means SDP

k-means objective:

$$\sum_{t=1}^{k} \sum_{i \in C_t} \left\| x_i - \frac{1}{|C_t|} \sum_{j \in C_t} x_j \right\|^2$$

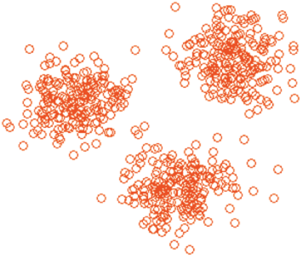


Peng, Wei, SIAM J. Optim., 2007

k-means SDP

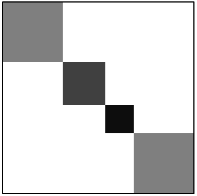
k-means objective:

$$\sum_{t=1}^{k} \sum_{i \in C_t} \left\| x_i - \frac{1}{|C_t|} \sum_{j \in C_t} x_j \right\|^2$$



SDP relaxation:

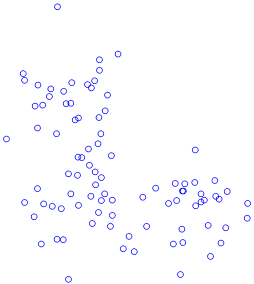
$$\begin{array}{ll} \text{minimize} & \mathrm{Tr}(D{\color{red}X}) \\ \text{subject to} & \mathrm{Tr}(X) = k \\ & X1 = 1 \\ & X \geq 0 \\ & X \succeq 0 \end{array}$$



Peng, Wei, SIAM J. Optim., 2007

What about outliers?

We exploited the SDP being tight.

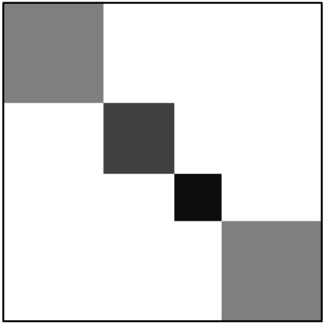


SDP guarantees for more realistic data?

The big idea

$X\mathbb{1} = \mathbb{1}$, $X \geq 0$ and $X^\top = X$, so X is doubly stochastic

$$X_{\text{opt}} \text{ integral} \implies PX_{\text{opt}} = \left[\; \underbrace{\hat{\gamma}_1 \cdots \hat{\gamma}_1}_{n_1 \text{ copies}} \quad \underbrace{\hat{\gamma}_2 \cdots \hat{\gamma}_2}_{n_2 \text{ copies}} \quad \cdots \quad \underbrace{\hat{\gamma}_k \cdots \hat{\gamma}_k}_{n_k \text{ copies}} \;\right]$$



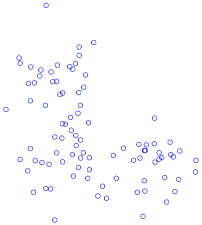
The big idea

$X\mathbb{1} = \mathbb{1}$, $X \geq 0$ and $X^\top = X$, so X is doubly stochastic

$$X_{\mathrm{opt}} \text{ integral} \Longrightarrow PX_{\mathrm{opt}} = \left[\underbrace{\hat{\gamma}_1 \cdots \hat{\gamma}_1}_{n_1 \text{ copies}} \quad \underbrace{\hat{\gamma}_2 \cdots \hat{\gamma}_2}_{n_2 \text{ copies}} \quad \cdots \quad \underbrace{\hat{\gamma}_k \cdots \hat{\gamma}_k}_{n_k \text{ copies}} \right]$$

What if X_{opt} is not integral?

Example: $P \mapsto PX_{\mathrm{opt}}$



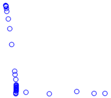
The big idea

$X1 = 1$, $X \geq 0$ and $X^\top = X$, so X is doubly stochastic

$$X_{\mathrm{opt}} \text{ integral} \implies PX_{\mathrm{opt}} = \Big[\, \underbrace{\hat{\gamma}_1 \cdots \hat{\gamma}_1}_{n_1 \text{ copies}} \;\; \underbrace{\hat{\gamma}_2 \cdots \hat{\gamma}_2}_{n_2 \text{ copies}} \;\; \cdots \;\; \underbrace{\hat{\gamma}_k \cdots \hat{\gamma}_k}_{n_k \text{ copies}} \,\Big]$$

What if X_{opt} is not integral?

Example: $P \mapsto PX_{\mathrm{opt}}$



Moral: PX_{opt} is a "denoised" version of P

How to explain denoising?

X_{plant} = planted clustering (integral)

Denoising $\longleftrightarrow$ small "mean squared error"

$$\mathrm{MSE} = \frac{1}{N} \sum_{t=1}^{k} \sum_{i=1}^{n} \| c_{t,i} - \hat{\gamma}_t \|^2$$
$$= \frac{1}{N} \| P X_{\mathrm{opt}} - P X_{\mathrm{plant}} \|_{\mathrm{F}}^2 \le \frac{1}{N} \|P\|_2^2 \|X_{\mathrm{opt}} - X_{\mathrm{plant}}\|_{\mathrm{F}}^2$$

Triangle: $\|P\|_2 \le \|\text{Gaussian centers}\|_2 + \|\text{Gaussian noise}\|_2$

Remaining task: Estimate $\|X_{\mathrm{opt}} - X_{\mathrm{plant}}\|_{\mathrm{F}}$

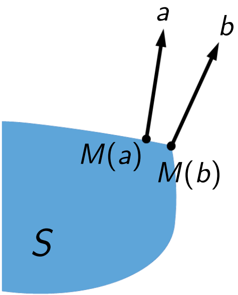
How to explain denoising?

$$M(v) := \arg\max_{x \in S} \langle x, v \rangle$$

$$a \approx_S b \implies M(a) \approx M(b)$$

Trick: Find R such that

- $M(R) = X_{\text{plant}}$
- $R \approx_S -D$



Guédon, Vershynin, `arXiv:1411.4686`

Mixon, Villar, Ward, `arXiv:1602.06612`

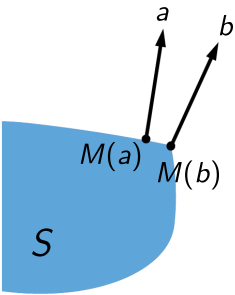
How to explain denoising?

$$M(v) := \arg\max_{x \in S} \langle x, v \rangle$$

$$a \approx_S b \implies M(a) \approx M(b)$$

Trick: Find R such that

- $M(R) = X_{\text{plant}}$
- $R \approx_S -D$



Theorem

$$\|X_{\text{opt}} - X_{\text{plant}}\|_F \leq \epsilon \text{ whp provided } \min_{i \neq j} \|\gamma_i - \gamma_j\| \gtrsim \frac{k\sigma}{\epsilon}.$$

Guédon, Vershynin, arXiv:1411.4686

Mixon, Villar, Ward, arXiv:1602.06612

Estimating Gaussian centers

After denoising, “round”:

for $i = 1 : k$
  $v_i \leftarrow$ denoised point with most neighbors
  delete denoised point and neighbors
endfor

Mixon, Villar, Ward, `arXiv:1602.06612`

Dasgupta, FOCS, 1999

Estimating Gaussian centers

After denoising, "round":

for $i = 1 : k$
  $v_i \leftarrow$ denoised point with most neighbors
  delete denoised point and neighbors
endfor

Theorem

$$\frac{1}{k}\sum_{i=1}^{k}\|v_i - \hat{\gamma}_i\|^2 \lesssim k^2\sigma^2 \text{ whp provided } \min_{i\neq j}\|\gamma_i - \gamma_j\| \gtrsim k\sigma.$$

(Trade Dasgupta's m-dependence for k-dependence)

How to remove the k-dependence in SNR and MSE?

Mixon, Villar, Ward, `arXiv:1602.06612`

Dasgupta, FOCS, 1999

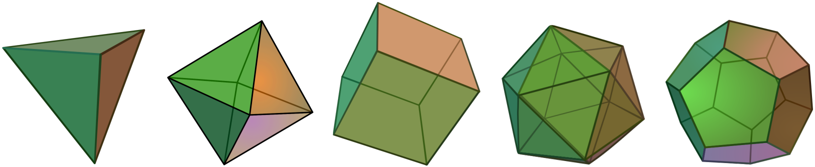
Fundamental limits of k-means clustering

We say $\Gamma \subseteq \mathbb{R}^m$ is a **stable isogon** if

- $|\Gamma| > 1$
- the symmetry group $G \leq O(m)$ acts transitively on Γ
- for each $\gamma \in \Gamma$, the stabilizer G_γ has the property that

$$\{x \in \mathbb{R}^m : Qx = x \ \ \forall Q \in G_\gamma\} = \operatorname{span}\{\gamma\}$$

Example: Platonic solids



Broome, Waldron, preprint, 2010

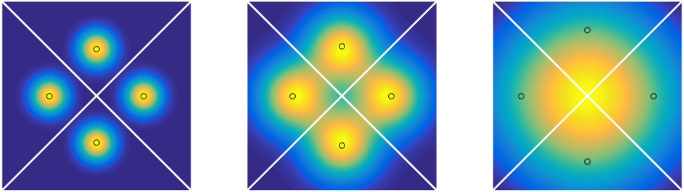
Fundamental limits of k-means clustering

Given $\Gamma \subseteq \mathbb{R}^m$, consider the Voronoi cells $\{V_\gamma\}_{\gamma\in\Gamma}$

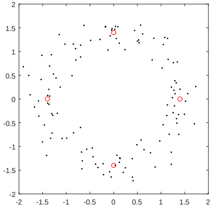
$\mathcal{D}$ = mixture of Gaussians centered at Γ

Define the **Voronoi means** by

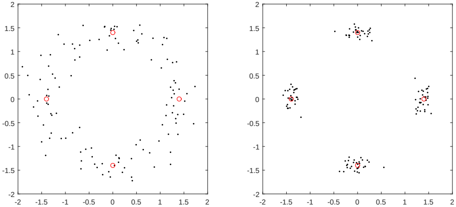
$$\mu_\gamma := \mathop{\mathbb{E}}_{X\sim\mathcal{D}}\left[X \middle| X \in V_\gamma\right]$$



Fundamental limits of k-means clustering

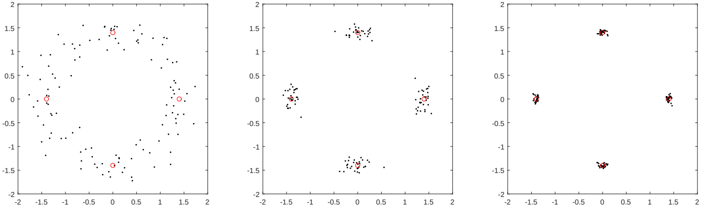


Fundamental limits of k-means clustering



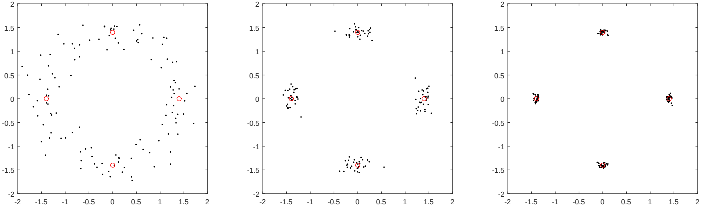
Mixon, Villar, Ward, arXiv:1602.06612

Fundamental limits of k-means clustering



Mixon, Villar, Ward, arXiv:1602.06612

Fundamental limits of k-means clustering



Voronoi Means Conjecture

Draw N points from a balanced mixture of spherical Gaussians of equal variance centered at points in a stable isogon. Then the k-means-optimal centroids converge in probability to the Voronoi means as $N \to \infty$.

Mixon, Villar, Ward, `arXiv:1602.06612`

Fundamental limits of k-means clustering

Γ = standard orthoplex in first $k/2$ dimensions of $\mathbb{R}^m$

$\mathcal{D}$ = balanced Gaussian mixture with centers Γ, covariance $\sigma^2 I$

Theorem

For every $\sigma > 0$, either

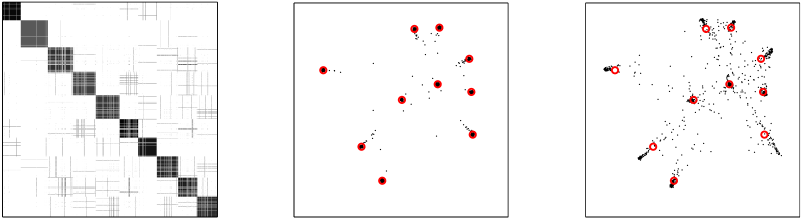
$$\min_{\substack{\gamma,\gamma' \in \Gamma \\ \gamma \neq \gamma'}} \|\gamma - \gamma'\| \gtrsim \sigma\sqrt{\log k} \quad \text{or} \quad \min_{\gamma \in \Gamma} \|\mu_\gamma - \gamma\| \gtrsim \sigma\sqrt{\log k}$$

Moral: If VMC, then either SNR or MSE exhibits k-dependence

Mixon, Villar, Ward, arXiv:1602.06612

Numerical experiment on MNIST dataset

1. Train a simple (one layer) neural network using TensorFlow.
2. Use it to map 1000 testing digits to feature space.
3. Run SDP denoising.
4. Find clusters using rounding scheme.



Questions?

Clustering subgaussian mixtures by semidefinite programming
D. G. Mixon, S. Villar, R. Ward
arXiv:1602.06612