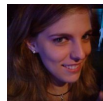


Some mathematics for k -means clustering

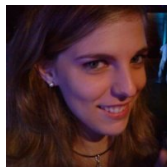
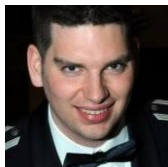
Rachel Ward

Berlin, December, 2015

Part 1: Joint work with Pranjal Awasthi, Afonso Bandeira, Moses Charikar, Ravi Krishnaswamy, and Soledad Villar



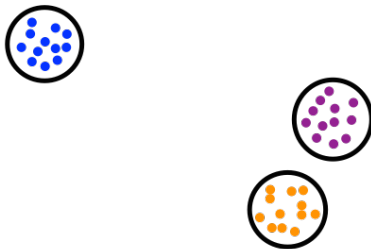
Part 2: Joint work with Dustin Mixon and Soledad Villar



The basic geometric clustering problem

Given a finite dataset $\mathcal{P} = \{x_1, x_2, \dots, x_N\}$, and target number of clusters k , find good partition so that data in any given partition are “similar”.

“Geometric” – assume points embedded in Hilbert space

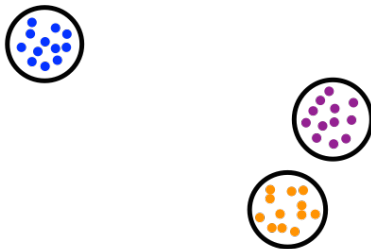


... Sometimes this is easy.

The basic geometric clustering problem

Given a finite dataset $\mathcal{P} = \{x_1, x_2, \dots, x_N\}$, and target number of clusters k , find good partition so that data in any given partition are “similar”.

“Geometric” – assume points embedded in Hilbert space



... Sometimes this is easy.

The basic geometric clustering problem



But often it is not so clear (especially with data in \mathbb{R}^d for d large) ...

k -means clustering

Most popular unsupervised clustering method. Points embedded in Euclidean space.



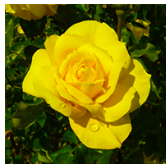
- ▶ x_1, x_2, \dots, x_N in \mathbb{R}^d , pairwise Euclidean distances are $\|x_i - x_j\|_2^2$.
- ▶ k -means optimization problem: among all k -partitions $C_1 \cup C_2 \cup \dots \cup C_k = \mathcal{P}$, find one that minimizes

$$\min_{C_1 \cup C_2 \cup \dots \cup C_k = \mathcal{P}} \sum_{i=1}^k \sum_{x \in C_i} \left\| x - \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j \right\|^2$$

- ▶ Works well for roughly spherical cluster shapes, uniform cluster sizes

k -means clustering

- ▶ Classic application: RGB Color quantization



- ▶ In general, as simple and (nearly) parameter-free pre-processing step for feature learning. These features then used for classification.

Lloyd's algorithm ('57) (a.k.a. "the" k -means algorithm)

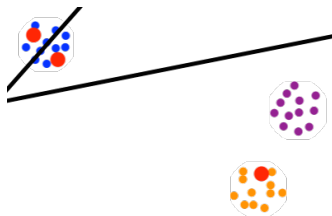
Simple algorithm for locally minimizing k -means objective;
responsible for popularity of k -means

$$\min_{C_1 \cup C_2 \cup \dots \cup C_k = \mathcal{P}} \sum_{i=1}^k \sum_{x \in C_i} \left\| x - \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j \right\|^2$$

- ▶ Initialize k "means" at random from among data points
- ▶ Iterate until convergence between (a) assigning each point to nearest mean and (b) computing new means as the average points of each cluster.
- ▶ Only guaranteed to converge to local minimizers (k -means is NP-hard)

Lloyd's algorithm ('57) (a.k.a. "the" k -means algorithm)

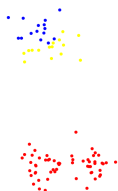
- ▶ Lloyd's method often converges to local minima



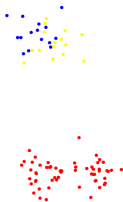
- ▶ [Arthur, Vassilvitskii '07] k -means++: Better initialization through non-uniform sampling, but still limited in high-dimension. Default in Matlab `kmeans()` algorithm
- ▶ [Kannan, Kumar '10] Initialize Lloyd's via spectral embedding.
- ▶ For these methods, no "certificate" of optimality

Points drawn from Gaussian mixture model in \mathbb{R}^5 . Initialization for k -means++ via Matlab 2014b *kmeans()*, Seed 1

k -means ++



Spectral
initialization



k -means
Semidefinite
relaxation



Outline of Talk



- ▶ Part 1: Generative clustering models and exact recovery guarantees for SDP relaxation of k -means
- ▶ Part 2: Stability results for SDP relaxation of k -means

Generative models for clustering

[Nellore, W '2013]: Consider the “Stochastic ball model”:



- ▶ μ is isotropic probability measure in \mathbb{R}^d supported in a **unit** ball.
- ▶ Centers $c_1, c_2, \dots, c_k \in \mathbb{R}^d$ such that $\|c_i - c_j\|_2 > \Delta$.
- ▶ μ_j as translation of μ to c_j .
- ▶ Draw n points $x_{\ell,1}, x_{\ell,2}, \dots, x_{\ell,n}$ from μ_ℓ , $\ell = 1, \dots, k$. $N = nk$.
- ▶ $\sigma^2 = \mathbb{E}(\|x_{\ell,j} - c_\ell\|_2^2) \leq 1$.

$D \in \mathbb{R}^{N \times N}$ such that $D_{(\ell,i),(m,j)} = \|x_{(\ell,i)} - x_{(m,j)}\|_2^2$

Note: Unless Stochastic Block Model, edge weights here are not independent

Generative models for clustering

[Nellore, W '2013]: Consider the “Stochastic ball model”:



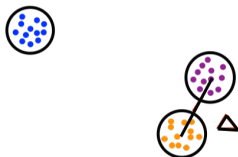
- ▶ μ is isotropic probability measure in \mathbb{R}^d supported in a **unit** ball.
- ▶ Centers $c_1, c_2, \dots, c_k \in \mathbb{R}^d$ such that $\|c_i - c_j\|_2 > \Delta$.
- ▶ μ_j as translation of μ to c_j .
- ▶ Draw n points $x_{\ell,1}, x_{\ell,2}, \dots, x_{\ell,n}$ from μ_ℓ , $\ell = 1, \dots, k$. $N = nk$.
- ▶ $\sigma^2 = \mathbb{E}(\|x_{\ell,j} - c_\ell\|_2^2) \leq 1$.

$D \in \mathbb{R}^{N \times N}$ such that $D_{(\ell,i),(m,j)} = \|x_{(\ell,i)} - x_{(m,j)}\|_2^2$

Note: Unless Stochastic Block Model, edge weights here are not independent

Stochastic ball model

Benchmark for “easy” clustering regime: $\Delta \geq 4$



Points within the same cluster are closer to each other than points in different clusters – simple thresholding of distance matrix.

Existing clustering guarantees in this regime: [Kumar, Kannan '10], [Elhamifar, Sapiro, Vidal '12], [Nellore, W. '13] $-\Delta = 3.75$

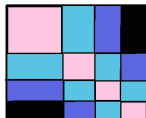
Generative models for clustering

Benchmark for “nontrivial” clustering case? $2 < \Delta < 4$



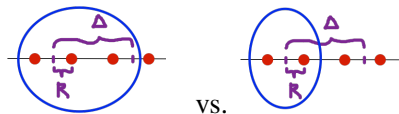
pairwise distance matrix D no longer looks too much like $\mathbb{E}[D]$,

$$\mathbb{E} \left[D_{(\ell,i),(m,j)} \right] = \|c_\ell - c_m\|_2^2 + 2\sigma^2$$



- ▶ Minimal number of points $n > d$ where d is ambient dimension
- ▶ Take care with distribution μ generating points

Subtleties in k -means objective



- ▶ In one dimension, k -means optimal solution ($k = 2$) switches at $\Delta = 2.75$
- ▶ [Iguchi, Mixon, Peterson, Villar '15] Similar phenomenon in 2D for distribution μ supported on boundary of ball, switch at $\Delta \approx 2.05$

k -means clustering



- Recall k -means optimization problem:

$$\min_{\mathcal{P}=C_1 \cup C_2 \cup \dots \cup C_k} \sum_{i=1}^k \sum_{x \in C_i} \left\| x - \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j \right\|^2$$

- Equivalent optimization problem:

$$\min_{\mathcal{P}=C_1 \cup C_2 \cup \dots \cup C_k} \sum_{i=1}^k \frac{1}{|C_i|} \sum_{x, y \in C_i} \|x - y\|^2$$

$$= \min_{\mathcal{P}=C_1 \cup C_2 \cup \dots \cup C_k} \sum_{\ell=1}^k \frac{1}{|C_\ell|} \sum_{(i,j) \in C_\ell} D_{i,j}$$

k -means clustering



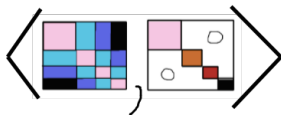
- Recall k -means optimization problem:

$$\min_{\mathcal{P}=C_1 \cup C_2 \cup \dots \cup C_k} \sum_{i=1}^k \sum_{x \in C_i} \left\| x - \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j \right\|^2$$

- Equivalent optimization problem:

$$\begin{aligned} & \min_{\mathcal{P}=C_1 \cup C_2 \cup \dots \cup C_k} \sum_{i=1}^k \frac{1}{|C_i|} \sum_{x, y \in C_i} \|x - y\|^2 \\ &= \min_{\mathcal{P}=C_1 \cup C_2 \cup \dots \cup C_k} \sum_{\ell=1}^k \frac{1}{|C_\ell|} \sum_{(i, j) \in C_\ell} D_{i, j} \end{aligned}$$

k -means clustering



... equivalent to:

$$\min_{Z \in \mathbb{R}^{N \times N}} \langle D, Z \rangle$$

subject to $\{Rank(Z) = k, \lambda_1(Z) = \dots = \lambda_k(Z) = 1, Z\mathbf{1} = \mathbf{1}, Z \geq 0\}$

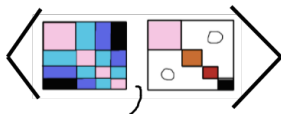
Spectral clustering relaxation:

$$\min_{Z \in \mathbb{R}^{N \times N}} \langle D, Z \rangle$$

subject to $\{Rank(Z) = k, \lambda_1(Z) = \dots = \lambda_k(Z) = 1, Z\mathbf{1} \neq \mathbf{1}, Z \geq 0\}$

Spectral clustering: Get top k eigenvectors, followed by clustering on reduced space

k -means clustering



... equivalent to:

$$\min_{Z \in \mathbb{R}^{N \times N}} \langle D, Z \rangle$$

subject to $\{Rank(Z) = k, \lambda_1(Z) = \dots = \lambda_k(Z) = 1, Z\mathbf{1} = \mathbf{1}, Z \geq 0\}$

Spectral clustering relaxation:

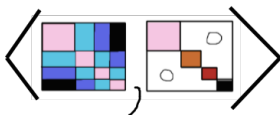
$$\min_{Z \in \mathbb{R}^{N \times N}} \langle D, Z \rangle$$

subject to $\{Rank(Z) = k, \lambda_1(Z) = \dots = \lambda_k(Z) = 1, Z\mathbf{1} = \mathbf{1}, Z \geq 0\}$

Spectral clustering: Get top k eigenvectors, followed by clustering on reduced space

Our approach: Semidefinite relaxation for k -means

[Peng, Wei '05] Proposed k -means semidefinite relaxation:



$$\min \langle D, Z \rangle$$

$$\text{subject to } \{ \text{Tr}(\mathbf{Z}) = \mathbf{k}, \mathbf{Z} \succeq 0, \mathbf{Z} \mathbf{1} = \mathbf{1}, \mathbf{Z} \succeq 0 \}$$

Note: Only parameter in SDP is k , the number of clusters, even though generative model assumes equal num. points n in each cluster

k -means SDP – recovery guarantees

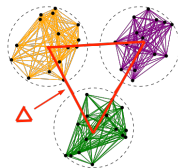
- ▶ μ is isotropic probability measure in \mathbb{R}^d supported in a **unit** ball.
- ▶ Centers $c_1, c_2, \dots, c_k \in \mathbb{R}^d$ such that $\|c_i - c_j\|_2 > \Delta$.
- ▶ μ_j as translation of μ to c_j . $\sigma^2 = \mathbb{E}(\|x_{\ell,j} - c_{\ell}\|_2^2) \leq 1$.

Theorem (with A., B., C., K., V. '14)

Suppose

$$\Delta \geq \sqrt{\frac{8\sigma^2}{d}} + 8$$

Then k -means SDP recovers clusters as unique optimal solution with probability $\geq 1 - 2dk \exp\left(-\frac{cn}{\log^2(n)d}\right)$.



Proof: construct dual certificate matrix, PSD, orthogonal to rank- k matrix with entries $\|x_i - c_j\|_2^2$, satisfies dual constraints bound largest eigenvalue of residual “noise” matrix [Vershynin '10]

k -means SDP – recovery guarantees

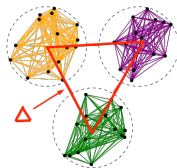
- ▶ μ is isotropic probability measure in \mathbb{R}^d supported in a **unit** ball.
- ▶ Centers $c_1, c_2, \dots, c_k \in \mathbb{R}^d$ such that $\|c_i - c_j\|_2 > \Delta$.
- ▶ μ_j as translation of μ to c_j . $\sigma^2 = \mathbb{E}(\|x_{\ell,j} - c_{\ell}\|_2^2) \leq 1$.

Theorem (with A., B., C., K., V. '14)

Suppose

$$\Delta \geq \sqrt{\frac{8\sigma^2}{d}} + 8$$

Then k -means SDP recovers clusters as unique optimal solution with probability $\geq 1 - 2dk \exp\left(-\frac{cn}{\log^2(n)d}\right)$.



Proof: construct dual certificate matrix, PSD, orthogonal to rank- k matrix with entries $\|x_i - c_j\|_2^2$, satisfies dual constraints bound largest eigenvalue of residual “noise” matrix [Vershynin '10]

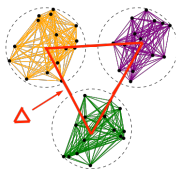
k -means SDP – cluster recovery guarantees

Theorem (with A., B., C., K., V. '14)

Suppose

$$\Delta \geq \sqrt{\frac{8\sigma^2}{d}} + 8$$

Then k -means SDP recovers clusters as unique optimal solution with probability $\geq 1 - 2dk \exp\left(-\frac{cn}{\log^2(n)d}\right)$.



- ▶ In fact, deterministic dual certificate sufficient condition. The “stochastic ball model” satisfies conditions with high probability.
- ▶ [Iguchi, Mixon, Peterson, Villar '15]: Recovery also for $\Delta \geq 2\sigma \frac{\sqrt{k}}{d}$, constructing different dual certificate

Inspirations

- ▶ [Candes, Romberg, Tao '04; Donoho '04] Compressive sensing
- ▶ Matrix factorizations
 - ▶ [Recht, Fazel, Parrilo '10] Low-rank matrix recovery
 - ▶ [Chandrasekaran, Sanghavi, Parrilo, Willsky '09] Robust PCA
 - ▶ [Bittorf, Recht, Re, Tropp '12] Nonnegative matrix factorization
 - ▶ [Oymak, Hassibi, Jalali, Chen, Sanghavi, Xu, Fazel, Ames, Mossel, Neeman, Sly, Abbe, Bandeira, ...] **community detection, stochastic block model**
 - ▶ Many more...

Stability of k -means SDP

Stability of k -means SDP



Recall SDP:

$$\min_{Z \in \mathbb{R}^{N \times N}} \langle D, Z \rangle$$

subject to $\{Rank(Z) = k, \lambda_1(Z) = \dots = \lambda_k(Z) = 1, Z\mathbf{1} = \mathbf{1}, Z \geq 0\}$

- ▶ For data $X = [x_1, x_2, \dots, x_N]$ “close” to being separated in k clusters, SDP solution $XZ_{opt} = [\hat{c}_1, \hat{c}_2, \dots, \hat{c}_N]$ should be “close” to a cluster solution XZ_C
- ▶ “Clustering is only hard when data does not fit the clustering model”

Stability of k -means SDP



Gaussian mixture model with “even” weights:

- ▶ centers $c_1, c_2, \dots, c_k \in \mathbb{R}^d$
- ▶ For each $t \in \{1, 2, \dots, k\}$, draw $x_{t,1}, x_{t,2}, \dots, x_{t,n}$ i.i.d. from $\mathcal{N}(\gamma_t, \sigma^2 I)$, $N = nk$ points total.
- ▶ $\Delta = \min_{a \neq b} \|c_a - c_b\|_2$.
- ▶ Want stability results in regime $\Delta = C\sigma$ for small $C > 1$
- ▶ Note: now $\mathbb{E}\|x_{t,j} - c_t\|^2 = d\sigma^2$

Stability of k -means SDP



Gaussian mixture model with “even” weights:

- ▶ centers $c_1, c_2, \dots, c_k \in \mathbb{R}^d$
- ▶ For each $t \in \{1, 2, \dots, k\}$, draw $x_{t,1}, x_{t,2}, \dots, x_{t,n}$ i.i.d. from $\mathcal{N}(\gamma_t, \sigma^2 I)$, $N = nk$ points total.
- ▶ $\Delta = \min_{a \neq b} \|c_a - c_b\|_2$.
- ▶ Want stability results in regime $\Delta = C\sigma$ for small $C > 1$
- ▶ Note: now $\mathbb{E}\|x_{t,j} - c_t\|^2 = d\sigma^2$

Observed tightness of SDP

points in \mathbb{R}^5 – projected to first 2 coordinates

Stability of k -means SDP

$$\min \langle D, Z \rangle$$

$$\text{subject to } \{ \text{Tr}(\mathbf{Z}) = \mathbf{k}, Z \geq 0, Z1 = 1, Z \geq 0 \}$$

Theorem (with D. Mixon and S. Villar, 2016)

Consider $N = nk$ points $x_{j,\ell}$ generated via Gaussian mixture model with centers c_1, c_2, \dots, c_k . Then with probability $\geq 1 - \eta$, the SDP optimal centers $[\hat{c}_{1,1}, \hat{c}_{1,2}, \dots, \hat{c}_{j,\ell}, \dots, \hat{c}_{k,n}]$ satisfy

$$\frac{1}{N} \sum_{j=1}^k \sum_{\ell=1}^n \|\hat{c}_{j,\ell} - c_j\|_2^2 \leq \frac{C(k\sigma^2 + \log(1/\eta))}{\Delta^2}$$

where C is not too big.

- ▶ Since $\mathbb{E}[\|x_{j,\ell} - c_j\|_2^2] = d\sigma^2$, noise reduction in expectation
- ▶ Apply Markov's inequality to get rounding scheme

Stability of k -means SDP

$$\min \langle D, Z \rangle$$

$$\text{subject to } \{ \text{Tr}(\mathbf{Z}) = \mathbf{k}, Z \geq 0, Z1 = 1, Z \geq 0 \}$$

Theorem (with D. Mixon and S. Villar, 2016)

Consider $N = nk$ points $x_{j,\ell}$ generated via Gaussian mixture model with centers c_1, c_2, \dots, c_k . Then with probability $\geq 1 - \eta$, the SDP optimal centers $[\hat{c}_{1,1}, \hat{c}_{1,2}, \dots, \hat{c}_{j,\ell}, \dots, \hat{c}_{k,n}]$ satisfy

$$\frac{1}{N} \sum_{j=1}^k \sum_{\ell=1}^n \|\hat{c}_{j,\ell} - c_j\|_2^2 \leq \frac{C(k\sigma^2 + \log(1/\eta))}{\Delta^2}$$

where C is not too big.

- ▶ Since $\mathbb{E}[\|x_{j,\ell} - c_j\|_2^2] = d\sigma^2$, noise reduction in expectation
- ▶ Apply Markov's inequality to get rounding scheme

Observed tightness of SDP

points in \mathbb{R}^5 – projected to first 2 coordinates

Observation: when not tight after one iteration, it is tight after two or three iterations: $[x_1, x_2, \dots, x_N] \rightarrow [\hat{c}_1, \hat{c}_2, \dots, \hat{c}_N] \rightarrow [\hat{c}'_1, \hat{c}'_2, \dots, \hat{c}'_N]$

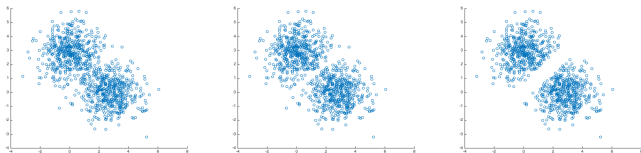
[Animation courtesy of Soledad Villar]

Summary



- ▶ We analyzed a convex relaxation of the k -means optimization problem, and showed that such an algorithm can recover global k -means optimal solutions if the underlying data can be partitioned in separated balls.
- ▶ In the same setting, popular heuristics like Lloyd's algorithm can get stuck in local optimal solutions
- ▶ We also showed that the k -means SDP is stable, providing noise reduction for Gaussian mixture models
- ▶ Philosophy: It is OK, and in fact better, that k -means SDP does not always return hard clusters. Denoising level indicates "clusterability" of data

Future directions



- ▶ SDP relaxation for k -means clustering is not fast – complexity scales at least N^6 where N is number of points. Fast solvers.
- ▶ Guarantees for kernel k -means for non-spherical data
- ▶ Make dual-certificate based clustering algorithms interactive (semi-supervised)

Thanks!

Mentioned papers:

1. *Relax, no need to round: integrality of clustering formulations* with P. Awasthi, A. Bandeira, M. Charikar, R. Krishnaswamy, and S. Villar. *ITCS*, 2015.
2. *Stability of an SDP relaxation of k -means*. D. Mixon, S. Villar, R. Ward. Preprint, 2016.