Power k-Means Clustering (Poster #96)

Jason Xu‡ and **Kenneth Lange***

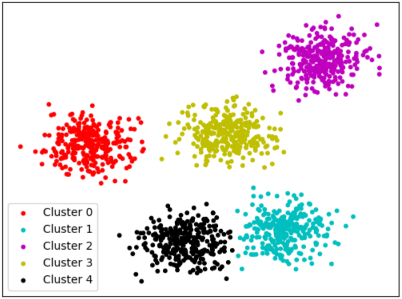
‡Department of Statistical Science, Duke University
*Departments of Biomathematics, Statistics, Human Genetics, UCLA

Thirty-sixth International Conference on Machine Learning
June 13, 2019, Long Beach, CA

Partitional clustering and k-means

- Given a representation of n observations and a measure of similarity, seek an optimal partition $\boldsymbol{C} = \{C_1, \ldots, C_k\}$ into k groups
- $\boldsymbol{X} \in \mathbb{R}^{d \times n}$ denotes n datapoints, $\boldsymbol{\theta} \in \mathbb{R}^{d \times k}$ represent k centers
- k-means: assign each observation to the cluster represented by the nearest center, minimizing within-cluster variance

$$\operatorname*{argmin}_{\boldsymbol{C}} \sum_{j=1}^{k} \sum_{\boldsymbol{x} \in C_j} \|\boldsymbol{x} - \boldsymbol{\theta}_j\|^2 \quad = \quad \operatorname*{argmin}_{\boldsymbol{C}} \sum_{j=1}^{k} |C_j| \operatorname{Var}(C_j)$$



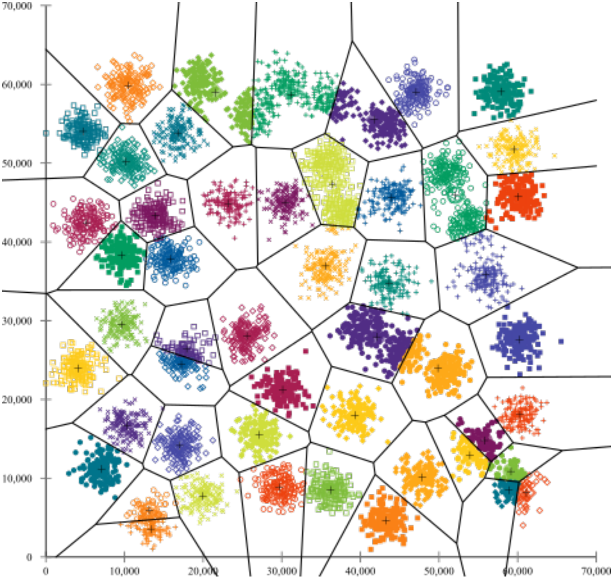
Lloyd's algorithm (1957)

Greedy approach: seeks local minimizer of k-means objective, rewritten

$$\sum_{i=1}^{n} \min_{1 \leq j \leq k} \|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|^2 := f_{-\infty}(\boldsymbol{\theta})$$

1. Update label assignments: $C_j^{(m)} = \{\boldsymbol{x}_i : \boldsymbol{\theta}_j^{(m)} \text{ is closest center}\}$
2. Recompute centers by averaging: $\boldsymbol{\theta}_j^{(m+1)} = \frac{1}{|C_j^{(m)}|} \sum_{\boldsymbol{x}_i \in C_j^{(m)}} \boldsymbol{x}_i$

Simple yet effective, remains most widely used clustering algorithm



Issues even when implicit assumptions are met

Drawbacks of Lloyd's algorithm

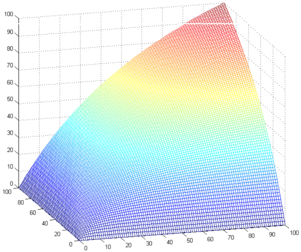
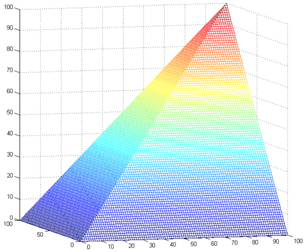
Even in ideal settings, Lloyd's algorithm is prone to local minima

- Sensitive to initialization, gets trapped in poor solutions, worsens in high dimensions
- Objective is non-smooth, highly non-convex
- "External" improvements: good initialization schemes (k-means++)

Goal: an "internal" improvement that retains the simplicity of Lloyd's algorithm, and seeks to optimize the same measure of quality

Solution: annealing along a continuum of smooth surfaces via majorization-minimization

A geometric approach: k-harmonic means (2001)

$$H(x_1, \ldots, x_k) = \left(\frac{1}{k}\sum_{j=1}^{k} x_j^{-1}\right)^{-1} \text{ as a proxy for } \min(x_1, \ldots, x_k)$$

Zhang et al. propose instead minimizing the criterion

$$\sum_{i=1}^{n} \Big(\frac{1}{k}\sum_{j=1}^{k} \|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|^{-2}\Big)^{-1} \; := \; f_{-1}(\boldsymbol{\theta})$$

A member of the power means family

Class of *power means*: $M_s(\boldsymbol{z}) = \left(\frac{1}{k}\sum_{i=1}^{k} z_i^s\right)^{\frac{1}{s}}$ for $z_i \in (0, \infty)$

- $s = 1$ yields arithmetic mean, $s = -1$ yields harmonic mean, etc
- Continuous, symmetric, homogeneous, strictly increasing
- Will be useful to generalize the good intuition behind KHM

Classical mathematical results $\Rightarrow$ nice algorithmic properties

1. Well-known $\lim_{s \to -\infty} M_s(z_1, \ldots, z_k) = \min\{z_1, \ldots, z_k\}$
2. Power mean inequality $M_s(z_1, \ldots, z_k) \leq M_t(z_1, \ldots, z_k), \quad s \leq t$

From power means to clustering criteria

Recall $M_s(\boldsymbol{z}) = \left(\frac{1}{k}\sum_{i=1}^{k} z_i^s\right)^{\frac{1}{s}}$

$$f_{-1}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \Big(\frac{1}{k}\sum_{j=1}^{k} \|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|^{-2}\Big)^{-1} \qquad \text{(KHM)}$$

- substitute $z_j = \|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|^2$ into $M_{-1}(\boldsymbol{z})$, sum over i

$$f_{-\infty}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \min_{1 \le j \le k} \|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|^2 \qquad \text{(k-means)}$$

- the same, substituting instead into "$M_{-\infty}(\boldsymbol{z})$"

What about all the other power means?

A continuum of smoother objectives

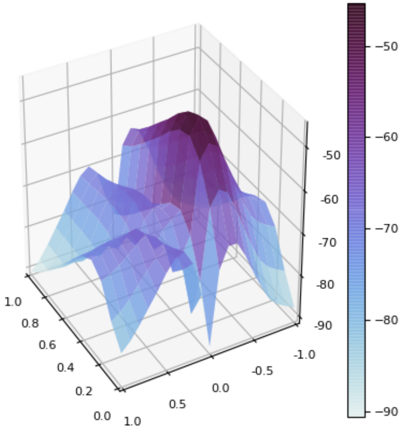
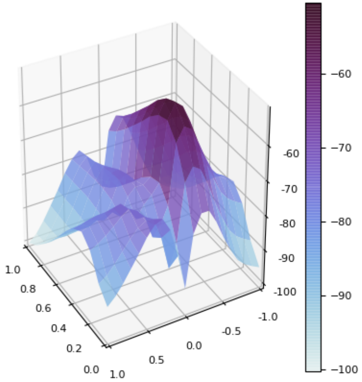


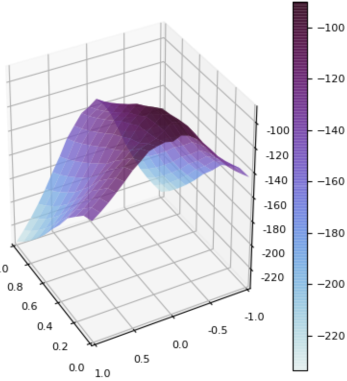
Figure: A cross-section of the k-means objective $-f_{-\infty}(\boldsymbol{\theta})$ with $k = 3$ clusters in dimension $d = 1$. Third center is fixed at its true value.

A continuum of smoother objectives

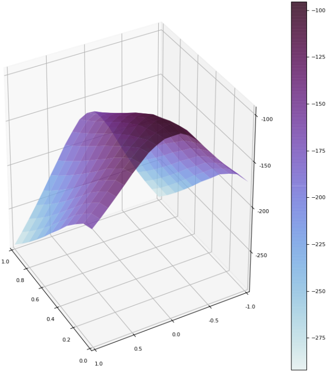


(a) $s = -10.0$ (b) $s = -1.0$ (KHM)

A continuum of smoother objectives



(c) $s = -0.2$



(d) $s = 0.3$

Gradually approaching the k-means criterion

Proposition: *For any* $\{s^{(m)}\} \to -\infty$, $\lim_{m\to\infty} \min_{\boldsymbol{\theta}} f_{s^{(m)}}(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} f_{-\infty}(\boldsymbol{\theta})$.

- Choosing one instance (i.e. f_{-1}) as proxy may not always be a good idea, now interpreted as early stopping along solution path
- Starting at $s^{(0)} < 1$, gradually decreasing $s \to -\infty$ can be understood as a form of annealing

Toward an iterative solution: majorization-minimization

A surrogate $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m)$ is said to *majorize* the function $f(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_m$ if

$$\begin{aligned} f(\boldsymbol{\theta}_m) &= g(\boldsymbol{\theta}_m \mid \boldsymbol{\theta}_m) \qquad \text{tangency at } \boldsymbol{\theta}_m \\ f(\boldsymbol{\theta}) &\leq g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m) \qquad \text{domination for all } \boldsymbol{\theta}. \end{aligned}$$

MM algorithm: iterates $\boldsymbol{\theta}_{m+1} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_n)$

- Example: Expectation-Maximization (EM) is an example of MM
- Lloyd's algorithm can be considered EM for Gaussian mixtures with limiting $\sigma^2 \to 0$

Illustration of MM algorithm

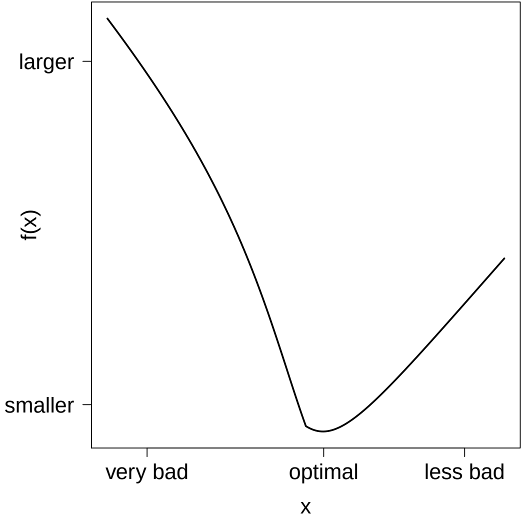


Illustration of MM algorithm

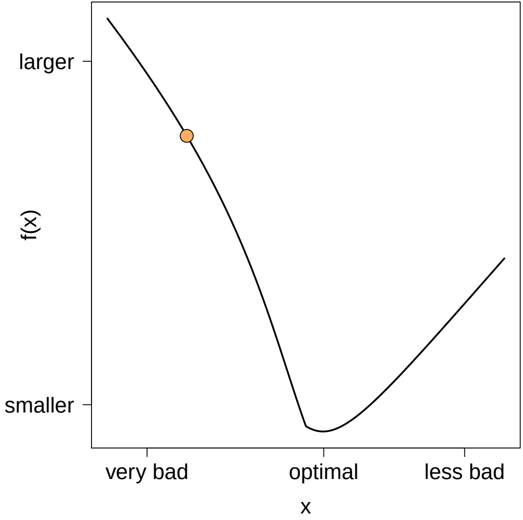


Illustration of MM algorithm

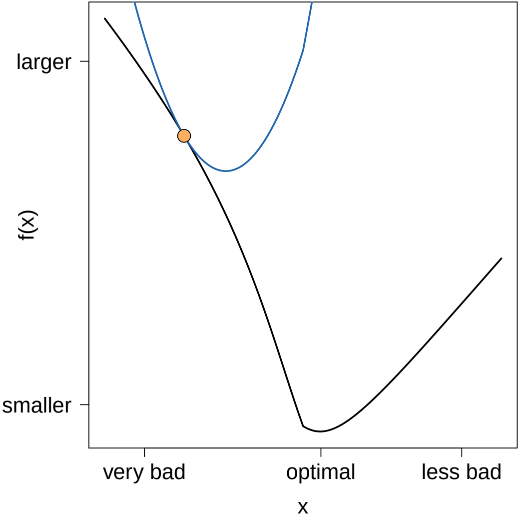


Illustration of MM algorithm

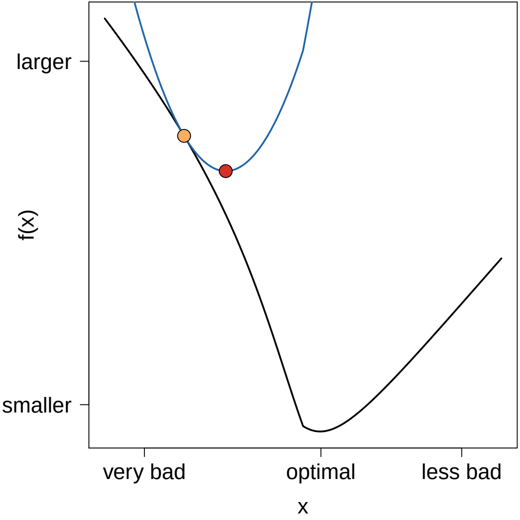


Illustration of MM algorithm

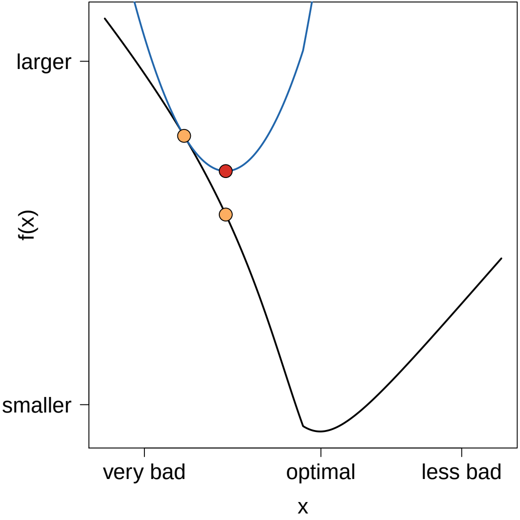


Illustration of MM algorithm

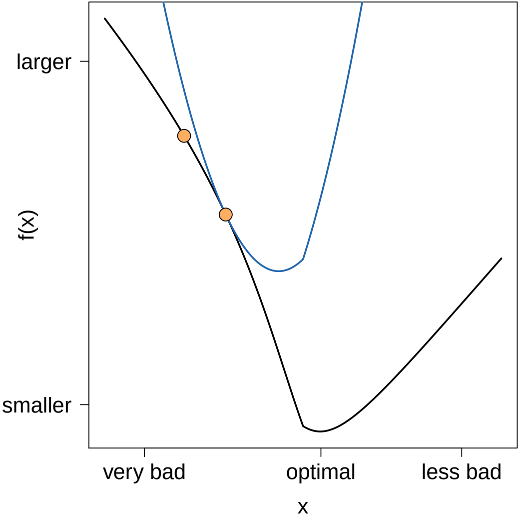


Illustration of MM algorithm

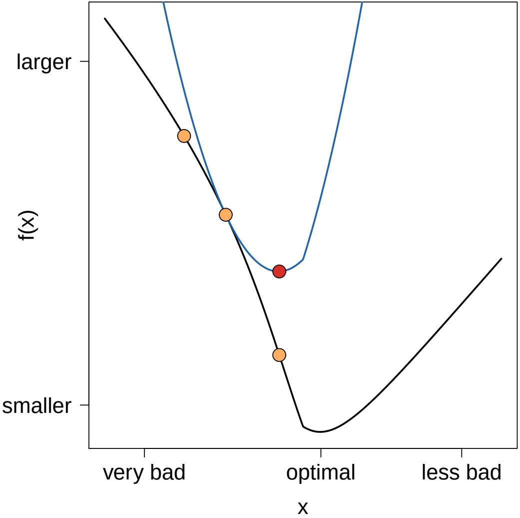


Illustration of MM algorithm

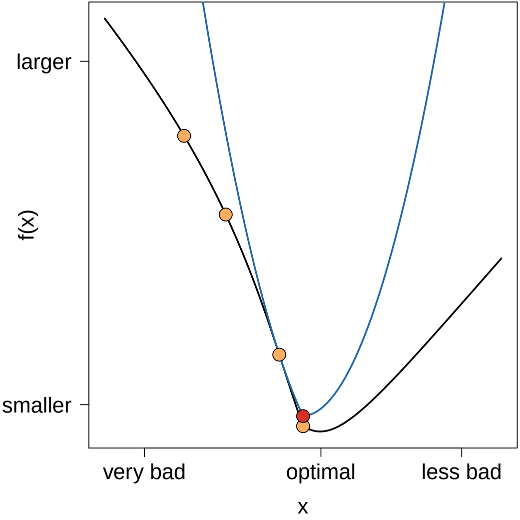
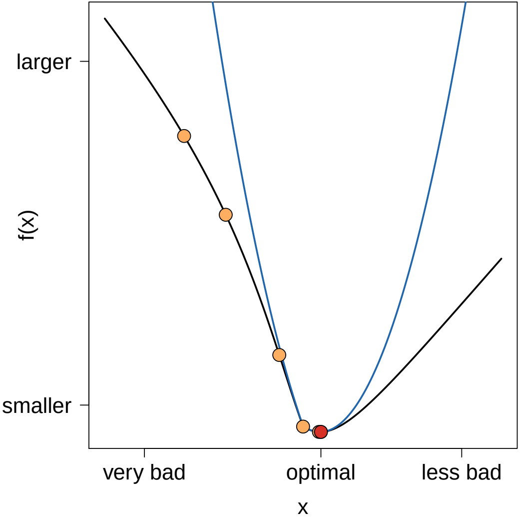


Illustration of MM algorithm



By all means, *k*-means

Algorithm 1 Power k-means algorithm pseudocode

Initialize $s^{(0)}, \boldsymbol{\theta}^{(0)}$; input data $\boldsymbol{x} \in \mathbb{R}^{n \times d}$, constant $\eta > 1$:
repeat
$$w_{ij}^{(m+1)} \leftarrow \left(\sum_{l=1}^{k} \|\boldsymbol{x}_i - \boldsymbol{\theta}_l^{(m)}\|^{2s} \right)^{\frac{1}{s}-1} \|\boldsymbol{x}_i - \boldsymbol{\theta}_j^{(m)}\|^{2(s-1)}$$
$$\boldsymbol{\theta}_j^{(m+1)} \leftarrow \Big(\sum_{i=1}^{n} w_{ij}^{(m+1)} \Big)^{-1} \Big(\sum_{i=1}^{n} w_{ij}^{(m+1)} \boldsymbol{x}_i \Big)$$
$$s^{(m+1)} \leftarrow \eta \cdot s^{(m)} \quad \text{(optional)}$$
until convergence

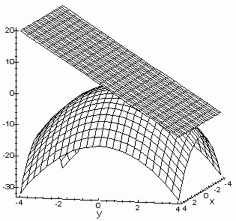
- Same $\mathcal{O}(nkd)$ time complexity as Lloyd; one additional parameter $s^{(0)}$

Proposition: *For any decreasing sequence $s^{(m)} \leq 1$, the iterates $\boldsymbol{\theta}^{(m)}$ produced by Algorithm 1 generates a decreasing sequence of objective values $f_{s^{(m)}}(\boldsymbol{\theta}^{(m)})$ bounded below by 0. As a consequence, the sequence of objective values converges.*

The shape of power means to come

Gradient has a nice form: $\displaystyle \frac{\partial}{\partial z_j} M_s(z_1, \ldots, z_k) = \left(\frac{1}{k}\sum_{i=1}^{k} z_i^s\right)^{\frac{1}{s}-1} \frac{1}{k} z_j^{s-1}$

Quadratic form of Hessian (not shown) shows that $M_s(\boldsymbol{z})$ is concave for $s \leq 1$



This means that whenever $s \leq 1$, the following inequality holds:

$$M_s(z_1, \ldots, z_k) \quad \leq \quad M_s(z_1^{(m)}, \ldots, z_k^{(m)}) + \sum_{j=1}^{k} \frac{\partial}{\partial z_j} M_s(z_1^{(m)}, \ldots, z_k^{(m)})(z_j - z_j^{(m)})$$

Minimizing power means objectives

Let $w_{ij}^{(m)} = \frac{\partial}{\partial \boldsymbol{\theta}_j} M_s(\|\boldsymbol{x}_i - \boldsymbol{\theta}_1^{(m)}\|^2, \ldots, \|\boldsymbol{x}_i - \boldsymbol{\theta}_k^{(m)}\|^2)$ for a given value $\boldsymbol{\theta}^{(m)}$

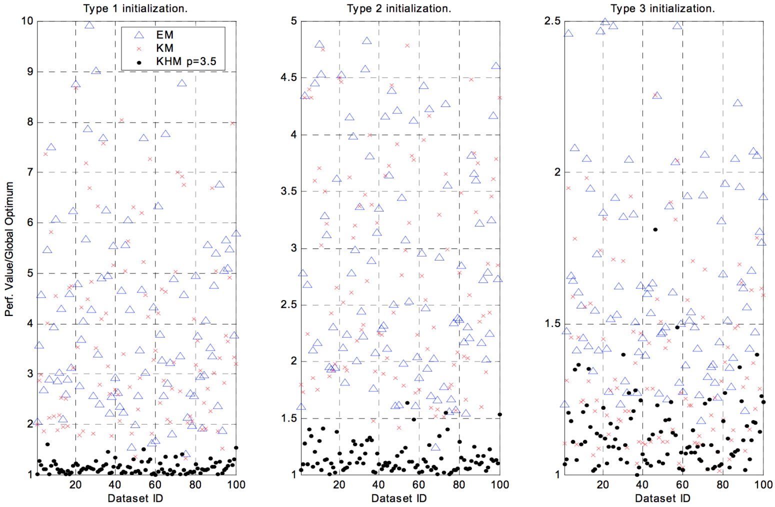
$$f_s(\boldsymbol{\theta}) = \sum_{i=1}^{n} M_s(\boldsymbol{\theta}; \boldsymbol{x}_i) \leq \overbrace{\sum_{i=1}^{n} \left(M_s(\theta^{(m)}; \boldsymbol{x}_i) + \sum_{j=1}^{k} w_{ij}^{(m)} \|\boldsymbol{x}_i - \theta_j^{(m)}\| \right)}^{C^{(m)}}$$

$$+ \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij}^{(m)} \|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|^2 := g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m)})$$

Unlike objective $f_s(\boldsymbol{\theta})$, the right-hand side $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m)})$ is easy to minimize!

$$\mathbf{0} = -2 \sum_{i=1}^{n} w_{ij}^{(m)} (\boldsymbol{x}_i - \boldsymbol{\theta}_j), \qquad \hat{\boldsymbol{\theta}}_j = \frac{1}{\sum_{i=1}^{n} w_{ij}^{(m)}} \sum_{i=1}^{n} w_{ij}^{(m)} \boldsymbol{x}_i.$$

Analogous experiment in KHM paper when $d = 2$



Performance comparison

Table: Variation of information under k-means++ initialization

	$d = 2$	$d = 5$	$d = 10$	$d = 20$	$d = 50$	$d = 100$	$d = 200$
Lloyd	0.637	0.261	0.234	0.223	0.199	0.206	0.183
KHM	0.651	0.328	0.339	0.319	0.263	0.280	0.231
s^0=-1	(**0.593**)	(**0.199**)	0.133	0.136	0.084	0.087	0.069
-3	0.593	0.226	(**0.111**)	(**0.069**)	(**0.022**)	(**0.027**)	0.026
-9	0.608	0.252	0.199	0.169	0.078	0.036	(**0.026**)
-18	0.615	0.259	0.218	0.208	0.140	0.101	0.077

Power k-means performs best for *all* choices of $s^{(0)}$ under good seedings!

Performance comparison

Table: Root k-means quality ratio with k-means++ initialization

	$d = 2$	$d = 5$	$d = 10$	$d = 20$	$d = 50$	$d = 100$	$d = 200$
Lloyd	1.036	1.236	1.363	1.411	1.476	1.492	1.481
KHM	1.044	1.290	1.473	1.504	1.556	1.586	1.556
s^0=-1	**(1.029)**	**(1.164)**	1.185	1.221	1.178	1.181	1.149
-3	1.030	1.187	**(1.155)**	**(1.110)**	**(1.044)**	**(1.054)**	**(1.059)**
-9	1.032	1.220	1.293	1.296	1.192	1.086	1.069
-18	1.034	1.228	1.328	1.370	1.351	1.254	1.203

Other measures such as adjusted Rand index convey the same trends

Closing remarks

- KHM degrades rapidly as d increases, and its benefits become less noticeable even in the plane with the availability of good seedings
- Power k-means succeeds in settings where Lloyd's and KHM break down, despite "ideal" setting
- Speed: power k-means takes ≈ 50 iterations ($\approx$ 20 seconds) on MNIST with $n = 60\,000, d = 784$
- Convergence rates $\Rightarrow$ optimal annealing schedules, choices of $s^{(0)}$?
- Bregman and other non-Euclidean extensions

Thank you!

Poster #96

jason.q.xu@duke.edu // jasonxu90.github.io