Consistency and inconsistency in k-means clustering

Adam Quinn Jaffe

with Moïse Blanchard and Nikita Zhivotovskiy

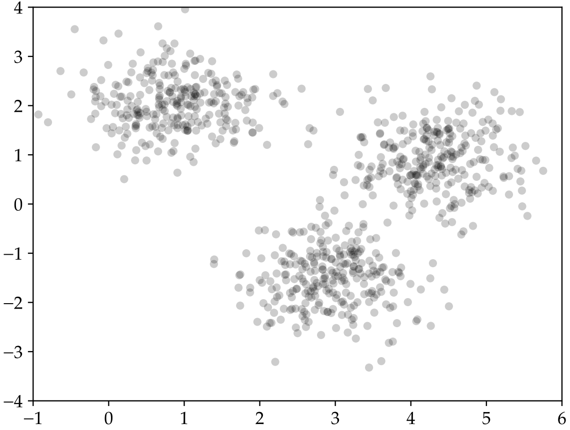
Fix $X_1, \ldots, X_n$ in $\mathbb{R}^m$ and $k \in \mathbb{N}$

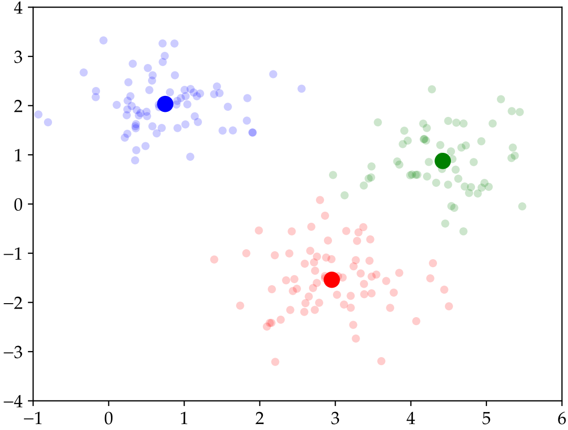
The *empirical k-means clustering problem* is

$$\begin{cases} \text{minimize} & \frac{1}{n}\sum_{i=1}^{n} \min_{a\in\mathcal{A}} \|a - X_i\|^2 \\ \text{over} & \mathcal{A} \subseteq \mathbb{R}^m \\ \text{with} & 1 \leq \#\mathcal{A} \leq k \end{cases}$$

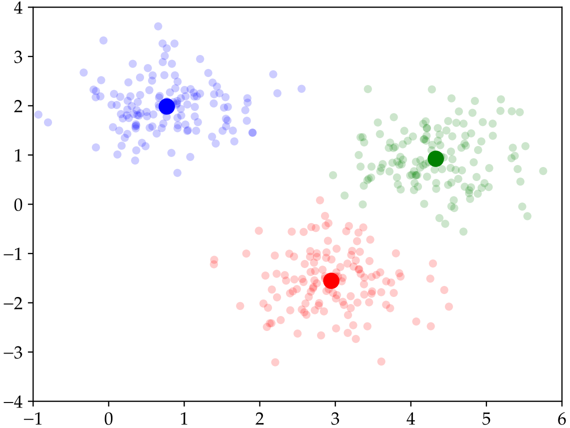
An optimal $\bar{\mathcal{A}}_n$ is called a set of *empirical k-means cluster centers*

Asymptotic theory for $\bar{\mathcal{A}}_n$ as $n \to \infty$ if $X_1, \ldots, X_n$ are i.i.d. samples?

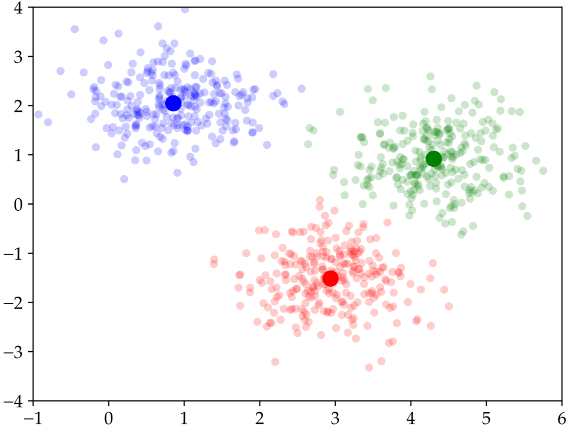

4
3
2
1
0
−1
−2
−3
−4
−1
0
1
2
3
4
5
6



4
3
2
1
0
−1
−2
−3
−4
−1
0
1
2
3
4
5
6



4
3
2
1
0
−1
−2
−3
−4
−1
0
1
2
3
4
5
6


4
3
2
1
0
−1
−2
−3
−4
−1
0
1
2
3
4
5
6

What is the corresponding population-level problem?

The *population k-means clustering problem* is

$$\begin{cases} \text{minimize} & \mathbb{E}\left[\min_{a\in\mathcal{A}} \|a - X\|^2\right] \\ \text{over} & \mathcal{A} \subseteq \mathbb{R}^m \\ \text{with} & 1 \le \#\mathcal{A} \le k \end{cases}$$

An optimal $\mathcal{A}$ is called a set of *population k-means cluster centers*

Write $\mathrm{d_H}(\mathcal{A}, \mathcal{B})$ for the *Hausdorff distance between* non-empty finite subsets of $\mathbb{R}^m$, i.e. the smallest distance for any matching from $\mathcal{A}$ to $\mathcal{B}$.

Theorem (Pollard 1981)

If $\mathbb{E}\|X\|^2 < \infty$, *then* $\mathrm{d_H}(\bar{\mathcal{A}}_n, \mathcal{A}) \to 0$ *almost surely.*

Many other results studying convergence of the optimal distortion (Bartlett-Linder-Lugosi 1998, Biau-Devroye-Lugosi 2008, Klochkov-Kroshnin-Zhivotovskiy 2021) but not the cluster centers

Want further results for some modern problems...

I. **Introduction**
II. **Consistency – Adaptivity & Geometry**
III. **Inconsistency – Heavy Tails**
IV. **Future Work**

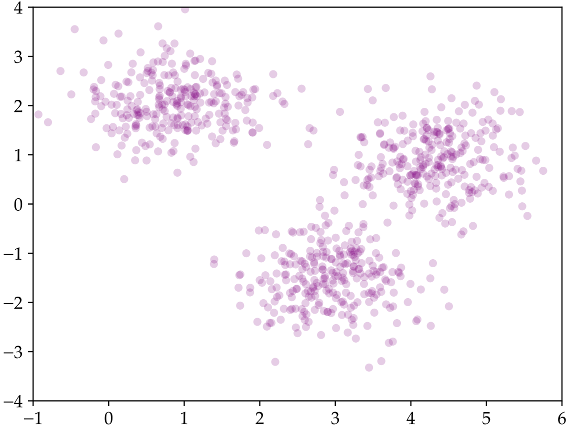
II. Consistency – Adaptivity & Geometry

Adaptivity.

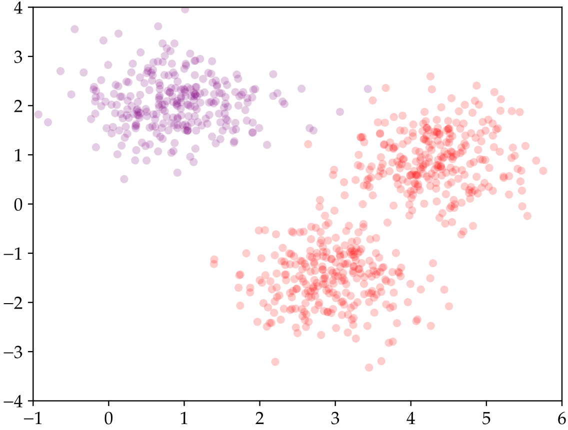
Pollard assumes $n \to \infty$ while k is fixed

In practice k is selected from the data!

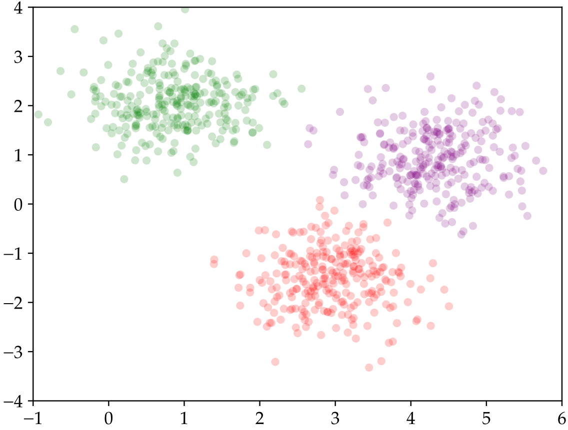
Asymptotic theory when k is selected from the "elbow method"?

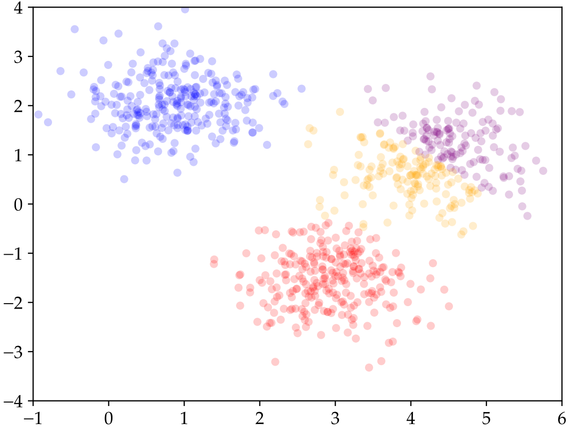


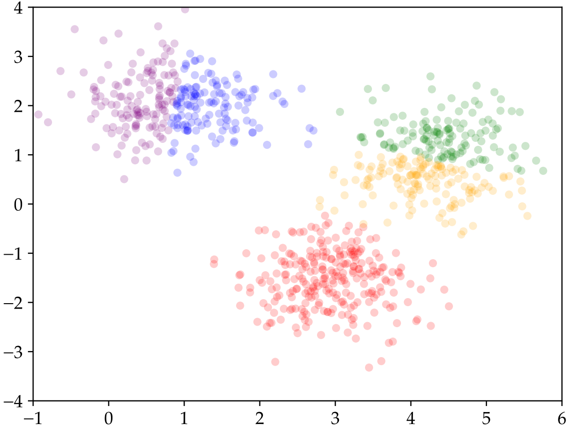
4
3
2
1
0
−1
−2
−3
−4
−1
0
1
2
3
4
5
6

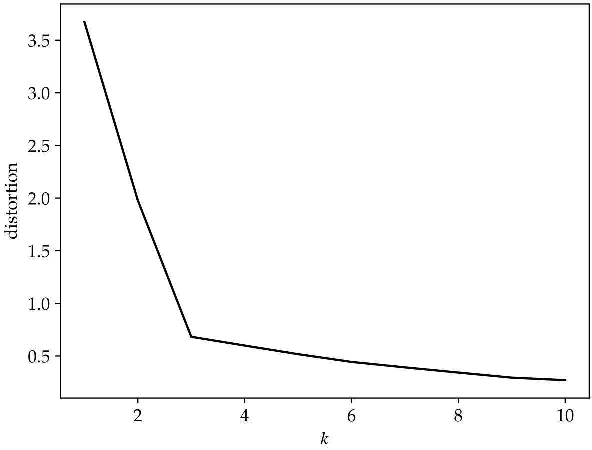


4
3
2
1
0
−1
−2
−3
−4
−1
0
1
2
3
4
5
6


4
3
2
1
0
−1
−2
−3
−4
−1
0
1
2
3
4
5
6


4
3
2
1
0
−1
−2
−3
−4
−1
0
1
2
3
4
5
6


4
3
2
1
0
−1
−2
−3
−4
−1
0
1
2
3
4
5
6



3.5
3.0
2.5
2.0
1.5
1.0
0.5
distortion
2
4
6
8
10
k

Fix $X_1, \ldots, X_n$ and define the *distortion* of k as

$$m(k) := \min_{\substack{\mathcal{A} \subseteq \mathbb{R}^m \\ 1 \leq \# \mathcal{A} \leq k}} \frac{1}{n} \sum_{i=1}^{n} \min_{a \in \mathcal{A}} \|a - X_i\|^2.$$

Fix $X_1, \ldots, X_n$ and define the *distortion* of k as

$$m(k) := \min_{\substack{\mathcal{A} \subseteq \mathbb{R}^m \\ 1 \leq \#\mathcal{A} \leq k}} \frac{1}{n} \sum_{i=1}^{n} \min_{a \in \mathcal{A}} \|a - X_i\|^2.$$

Then set:

$$k(X_1, \ldots, X_n) := \arg\max_{k \geq 2} \Big(m(k+1) + m(k-1) - 2m(k) \Big)$$

The *empirical elbow-method k-means clustering problem* is

$$\begin{cases} \text{minimize} & \frac{1}{n}\sum_{i=1}^{n} \min_{a\in\mathcal{A}} \|a - X_i\|^2 \\ \text{over} & \mathcal{A} \subseteq \mathbb{R}^m \\ \text{with} & 1 \leq \#\mathcal{A} \leq k(X_1, \dots, X_n) \end{cases}$$

The *empirical elbow-method k-means clustering problem* is

$$\begin{cases} \text{minimize} & \frac{1}{n}\sum_{i=1}^{n} \min_{a\in\mathcal{A}} \|a - X_i\|^2 \\ \text{over} & \mathcal{A} \subseteq \mathbb{R}^m \\ \text{with} & 1 \leq \#\mathcal{A} \leq k(X_1, \ldots, X_n) \end{cases}$$

Let $\bar{\mathcal{A}}_n$ denote an optimizer.

Population-level problem is analogous, let $\mathcal{A}$ denote optimizer

Theorem (AQJ 2025)

If $\mathbb{E}\|X\|^2 < \infty$, *then* $\mathrm{d_H}(\bar{\mathcal{A}}_n, \mathcal{A}) \to 0$ *almost surely.*

Similar result for k-medoids (Kaufman-Rousseeuw 1987), where k is fixed but *domain* is chosen from the data.

Geometry.

The field of *distributional data analysis* concerns statistical inference where the data are probability distributions themselves

Applications in demography (Chen-Lin-Müller 2021), econometrics (Gunsilius 2023), Bayesian statistics (Srivastava-Cevher-Dinh-Dunson 2015), etc.

Define the *Wasserstein distance* $W_2(\mu, \nu)$ via

$$W_2^2(\mu, \nu) := \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^m \times \mathbb{R}^m} \|x - y\|^2 \mathrm{d}\pi(x, y)$$

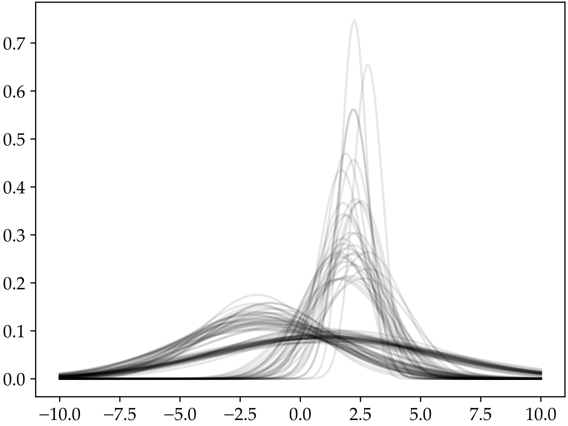
Fix $\mu_1, \ldots, \mu_n$ probability measures on $\mathbb{R}^m$ and $k \in \mathbb{N}$.

The *empirical k-barycenters clustering problem* is

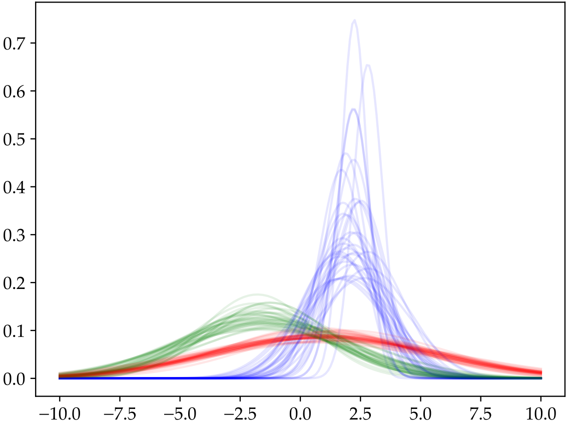
$$\begin{cases} \text{minimize} & \frac{1}{n}\sum_{i=1}^{n} \min_{\nu \in \mathcal{A}} W_2^2(\mu_i, \nu) \\ \text{over} & \mathcal{A} \subseteq \mathcal{P}_2(\mathbb{R}^m) \\ \text{with} & 1 \leq \#\mathcal{A} \leq k \end{cases}$$

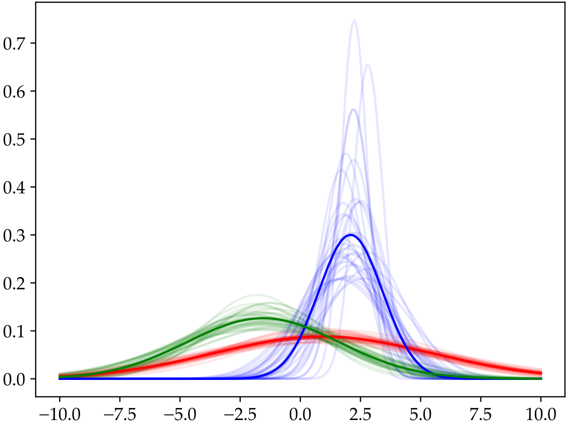
An optimal $\bar{\mathcal{A}}_n$ is called a set of *empirical k-barycenters cluster centers*

Population-level problem is analogous, let $\mathcal{A}$ denote optimizer.



0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0


0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

Theorem (AQJ 2025)

If $\mathbb{E}W_2^2(\mu, \delta_0) < \infty$, *then* $\mathrm{d_H}(\bar{\mathcal{A}}_n, \mathcal{A}) \to 0$ *almost surely.*

Extends results for Wassestein barycenters (Le Gouic-Loubes 2017)

Similar results for other metric spaces $(\mathcal{X}, d)$, extending results from functional data analysis (Thorpe-Theil-Johansen-Cade 2015)

III. Inconsistency – Heavy Tails

Pollard shows consistency when $\mathbb{E}\|X\|^2 < \infty$.

When $k = 1$ the optimal cluster center is just the mean, so SLLN implies consistency when $\mathbb{E}\|X\| < \infty$.

Do we have consistency for general $k \geq 2$ and $\mathbb{E}\|X\| < \infty$?

Suppose $X_1, \ldots, X_n$ i.i.d. with symmetric Pareto(2) distribution

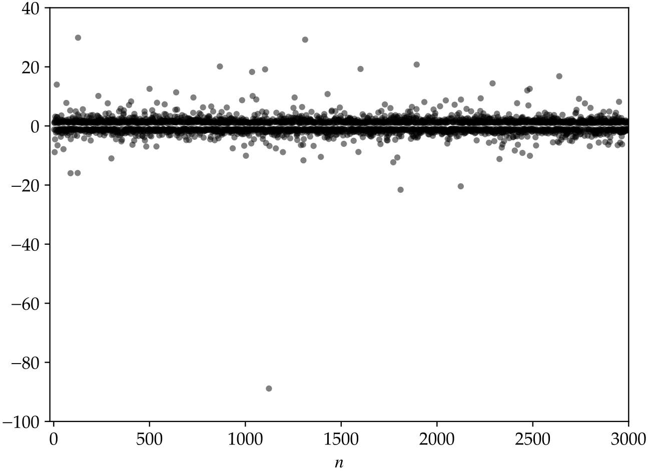
$$\mathbb{P}(|X| > t) = \frac{1}{t^2}.$$

Note $\mathbb{E}|X|^2 = \infty$ but $\mathbb{E}|X|^p < \infty$ for all $1 \leq p < 2$.

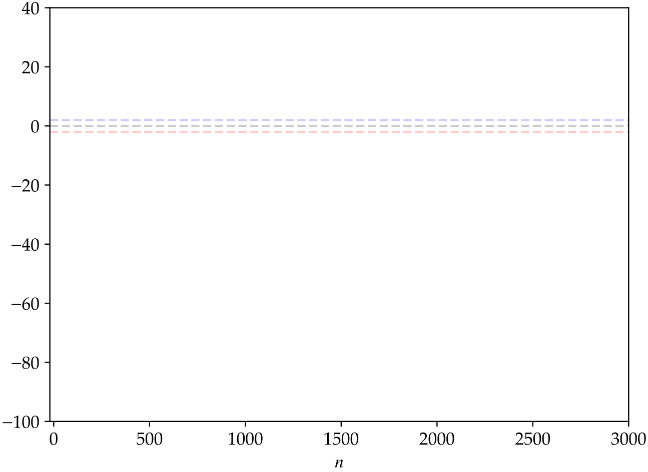
Consider k-means clustering for $k = 2$

Unique set of population 2-means cluster centers is $\mathcal{A} = \{-2, 2\}$.

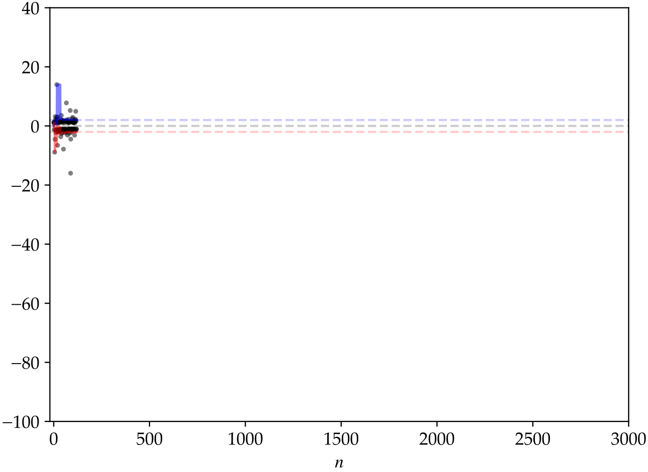
Consistency of empirical k-means cluster centers?



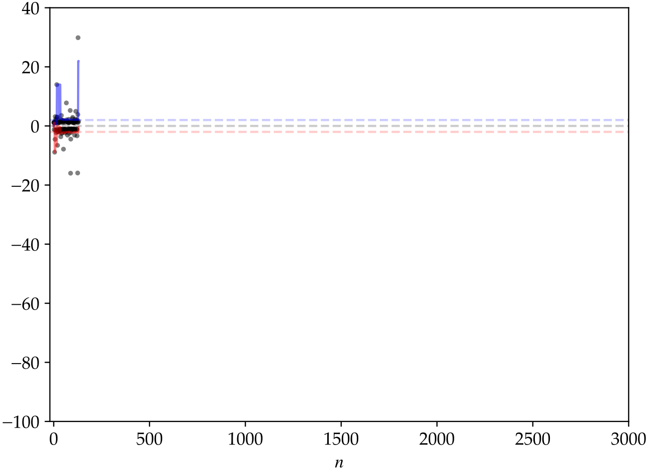
samples $X_1, \dots, X_n$,



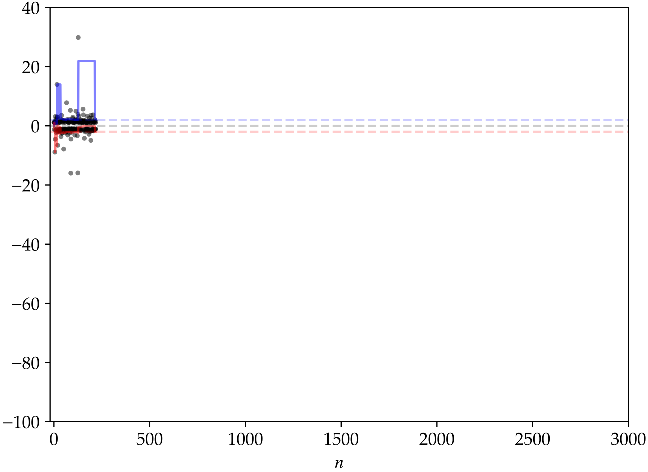
smaller cluster center, larger cluster center



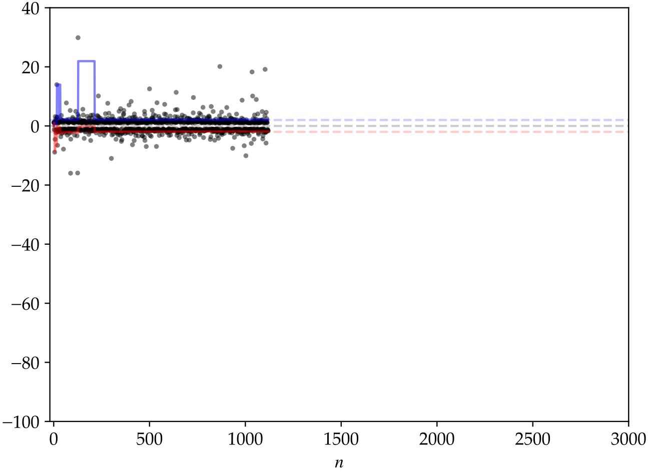
samples $X_1, \ldots, X_n$, smaller cluster center, larger cluster center



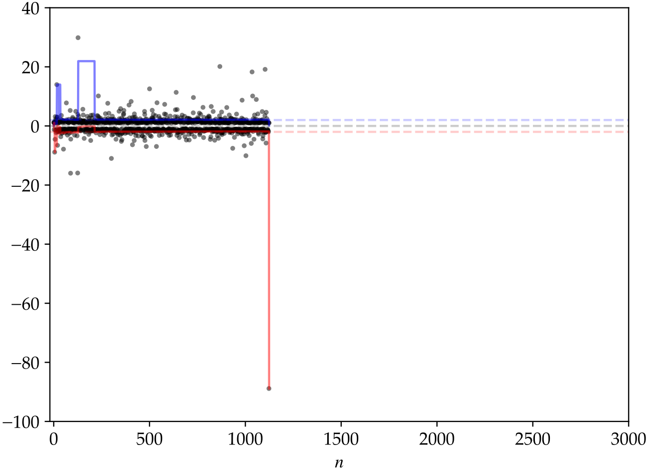
samples $X_1, \ldots, X_n$, smaller cluster center, larger cluster center



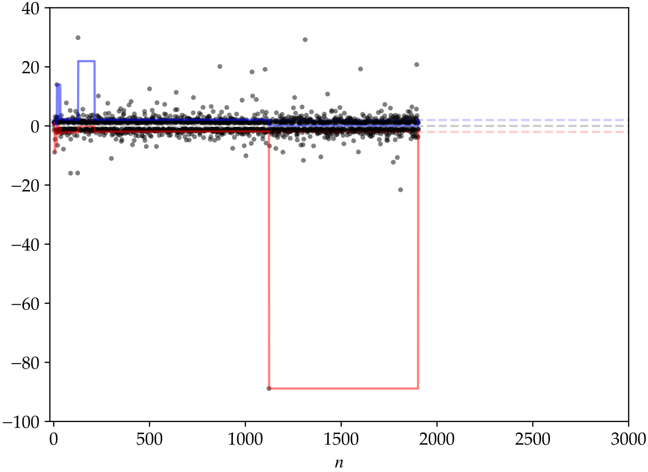
samples $X_1, \ldots, X_n$, smaller cluster center, larger cluster center



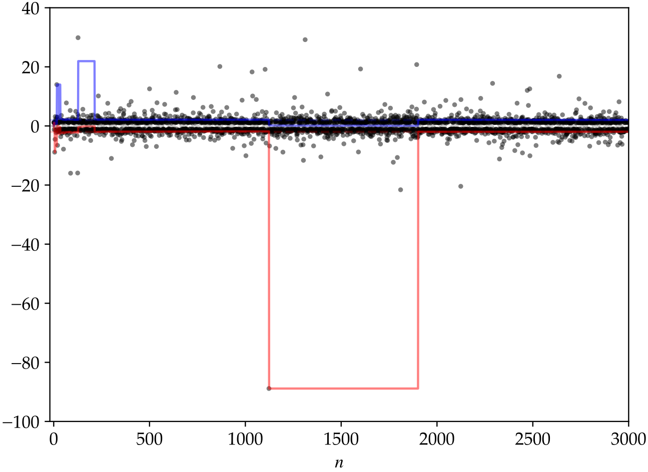
samples $X_1, \ldots, X_n$, smaller cluster center, larger cluster center



samples $X_1, \ldots, X_n$, smaller cluster center, larger cluster center



samples $X_1, \ldots, X_n$, smaller cluster center, larger cluster center



samples $X_1, \ldots, X_n$, smaller cluster center, larger cluster center

Let $\bar{\mathcal{A}}_n = \{a_n, b_n\} \subseteq \mathbb{R}$ with $a_n < b_n$ denote the empirical k-means cluster centers for the Pareto(2) distribution when $k = 2$.

Let $\bar{\mathcal{A}}_n = \{a_n, b_n\} \subseteq \mathbb{R}$ with $a_n < b_n$ denote the empirical k-means cluster centers for the Pareto(2) distribution when $k = 2$.

Theorem (Blanchard-AQJ-Zhivotovskiy 2025)

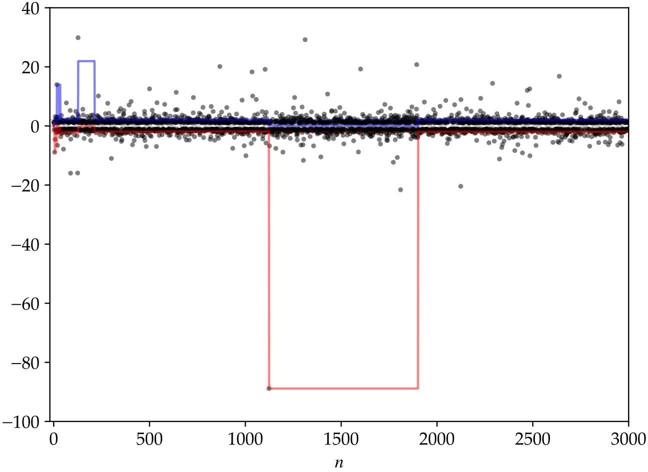
There exists a universal constant $c > 0$ such that the cases

$$a_n \leq -c\frac{\sqrt{n}}{\log n} \qquad \text{and} \qquad b_n \geq c\frac{\sqrt{n}}{\log n}$$

both occur infinitely often almost surely.

Consequently, $\limsup_{n\to\infty} \mathrm{d_H}(\bar{\mathcal{A}}_n, \mathcal{A}) = \infty$ almost surely.

So k-means clustering can be inconsisent even when $\mathbb{E}\|X\| < \infty$



samples $X_1, \ldots, X_n$, smaller cluster center, larger cluster center

Explanation: extreme outcomes create clusters with few points!

Roughly speaking, we have with high probability:

$$\left\{ \frac{X_{(n)}}{X_{(n-1)}} = \tilde{\Omega}(\sqrt{n}) \right\} \subseteq \left\{ \text{there exists a cluster with } \tilde{O}(1) \text{ samples} \right\}$$

Relationship between cluster imbalance and statistical rates of convergence (Klochkov-Kroshnin-Zhivotovskiy 2021)

Positive results when $\mathbb{E}\|X\| < \infty$?

For $a \in \bar{\mathcal{A}}_n$, define the Voronoi regions

$$\bar{\mathcal{V}}_n(a) := \{X_i : \|a - X_i\| \leq \|a' - X_i\| \text{ for all } a' \in \bar{\mathcal{A}}_n\}$$

Elementary bound:

$$\|a\| \leq \frac{\sum_{i=1}^n \|X_i\|}{\#\bar{\mathcal{V}}_n(a)} \text{ for all } a \in \bar{\mathcal{A}}_n$$

If $\min_{a \in \bar{\mathcal{A}}_n} \#\bar{\mathcal{V}}_n(a) = \Omega(n^{-1})$ then $\{\bar{\mathcal{A}}_n\}_{n \in \mathbb{N}}$ is uniformly bounded.

Pollard: If $\mathbb{E}\|X\|^2 < \infty$, then $\min_{a \in \bar{\mathcal{A}}_n} \#\bar{\mathcal{V}}_n(a) = \Omega(n^{-1})$ almost surely.

Consider *constrained k-means clustering* where each cluster is required to contain at least $\gamma_n \in \mathbb{N}$ many samples.

Some existing work on computational and methodological considerations (Ng 2000, Cuturi-Doucet 2014), but no statistical theory.

Some consistency results:

- Easy: If $\mathbb{E}\|X\| < \infty$ and $\gamma_n \geq \alpha n$ for small enough $0 \leq \alpha < 1$, then constrained k-means clustering is consistent.
- Harder: If $\mathbb{E}\|X\| < \infty$ and $\gamma_n \geq (\log n)^4$, then constrained k-means clustering converges to k'-means clustering for some $1 \leq k' \leq k$.

IV. Future Work

Consistency:

- Computational considerations for k-barycenters clustering?
- Relative efficiency for convex surrogates of k-means clustering?

Inconsistency:

- Finer understanding of the Pareto(2) example?
- Precise cutoffs for γ_n for consistency of balanced clustering?

Methodology:

- Other ways to impose balance constraints?
- How to interpret imbalance in practice?

Thank you!

References

P. L. Bartlett, T. Linder, and G. Lugosi. The minimax distortion redundancy in empirical quantizer design. *IEEE Trans. Inform. Theory*, 44(5):1802-1813, 1998

G. Biau, L. Devroye, and G. Lugosi. On the performance of clustering in Hilbert spaces. *IEEE Trans. Inform. Theory*, 54:781-790, 2008

Y. Chen, Z. Lin, Z. and H.-G. Müller. Wasserstein Regression. *J. Amer. Statist. Assoc.,* 118(542), 869-882, 2021

J. A. Cuesta-Albertos, A. Gordaliza, and C. Matrán. Trimmed k-means: an attempt to robustify quantizers. *Ann. Statist.*, 25(2):553-576, 1997.

M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. *ICML*, 32(2):685-693, 2014

S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions.* Lecture Notes in Mathematics (Springer Berlin), 2007.

F. Gunsilius. Distributional synthetic controls. *Econometrica.* 91(3):1105-1117, 2023.

L. Kaufman and P. J. Rousseeuw. Clustering by means of medoids. In *Statistical Data Analysis Based on the L^1-Norm and Related Methods*, 405-416, 1987.

Y. Klochkov, A. Kroshnin, and N. Zhivotovskiy. Robust k-means clustering for distributions with two moments. *Ann. Statist.,* 49(4):2206-2230, 2021.

T. Le Gouic and J.-M. Loubes. Existence and consistency of Wasserstein barycenters. *Probab. Theory Related Fields,* 168:901-917, 2017.

M. K. Ng. A note on constrained k-means algorithms. *Pattern Recognit.*, 33(3):515-519, 2000.

D. Pollard. Strong consistency of k-means clustering. *Ann. Statist.*, 9(1):135-140, 1981.

S. Srivastava, V. Cevher, Q. Dinh, and D. Dunson. WASP: Scalable Bayes via barycenters of subset posteriors. *AISTATS,* 38:912-920, 2015.

M. Thorpe, F. Theil, A. M. Johansen, and N. Cade. Convergence of the k-means minimization problem using Γ-convergence. *SIAM J. Appl. Math.*, 75:2444-2474, 2015.