Robust k-means clustering for distributions with two moments

Nikita Zhivotovskiy[1]

[1]Google Research, Zürich
zhivotovskiy@google.com

Joint with Y. Klochkov (Cambridge) and A. Kroshnin (HSE)

https://arxiv.org/abs/2002.02339

To appear in Ann. of Stat.

Statistical k-means/Quantization of probability measures

Question

Given some distribution P in $\mathbb{R}^d$ and $k \geq 1$ find $A^* \subset \mathbb{R}^d$ such that it minimizes the distortion

$$D(A) = \mathbb{E} \min_{a \in A} \|X - a\|_2^2 \quad \text{among} \quad A \subset \mathbb{R}^d, |A| = k.$$

Example: if $k = 1$ then $\mathbb{E}\, X$ minimizes $\mathbb{E}\, \|X - a\|_2^2$ among $a \in \mathbb{R}^d$.

In statistical k-means clustering we are given N independent observations sampled according to P:

$$X_1, \ldots, X_N$$

Statistical k-means

Based on independent observations

$$X_1, \ldots, X_N$$

construct $\hat{A} \subset \mathbb{R}^d$, $|\hat{A}| = k$ such that it is close A^*.

We consider the excess distortion:

$$D(\hat{A}) - D(A^*) = \mathbb{E} \min_{a \in \hat{A}} \|X - a\|_2^2 - \mathbb{E} \min_{a \in A^*} \|X - a\|_2^2$$

Statistical k-means

A natural way to construct $\hat{A}$ is to minimize the empirical analog of $D(A)$. Define the *emipically optimal quantizer* as

$$\hat{A} \in \underset{A \subset \mathbb{R}^d, |A|=k}{\arg\min} \sum_{i=1}^{N} \min_{a \in A} \|X_i - a\|^2.$$

Theorem: D. Pollard, (AoS, 1981)

Assume that P satisfies $\mathbb{E}\|X\|^2 < \infty$ then for any integer $k \geq 1$, the optimal quantizer A^* exists and it holds that

$$D(\hat{A}) - D(A^*) \xrightarrow{a.s.} 0, \quad as \quad N \to \infty.$$

Statistical k-means

We have

$$D(\hat{A}) - D(A^*) \xrightarrow{a.s.} 0, \quad \text{as} \quad N \to \infty.$$

Question

1. *What is the rate of convergence?*
2. *How to handle heavy tailed distributions?*
3. *Is the empirically optimal quantizer the best solution?*

So far, only the first question was answered under the restrictive assumption that $\mathrm{supp}(P)$ is bounded.

Landmark results answering the first question

Assume that $\|X\| \leq T$ almost surely.

Linder, Lugosi, Zeger (1994, IEEE Trans. on inf. theory)

$$\mathbb{E}\, D(\hat{A}) - D(A^*) = \tilde{O}\left(T^2\sqrt{\frac{kd}{N}}\right).$$

Biau, Devroye and Lugosi, (2008, IEEE Trans. on inf. theory), Maurer (2016, ALT)

$$\mathbb{E}\, D(\hat{A}) - D(A^*) = O\left(T^2\frac{k}{\sqrt{N}}\right).$$

Fefferman, Mitter, Narayanan (2016, JoAMS), Foster, Rakhlin (2019)

$$\mathbb{E}\, D(\hat{A}) - D(A^*) = \tilde{O}\left(T^2\sqrt{\frac{k}{N}}\right).$$

(Towards heavy-tails) The problem of mean estimation

Question

Given $X_1, \dots, X_N$ i.i.d. observations of a random variable X, estimate

$$\mu = \mathbb{E}\, X.$$

Our candidate is the sample mean:

$$\overline{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

For Gaussian (sub-Gaussian) distributions it holds that, with probability at least $1 - \delta$,

$$|\overline{X} - \mu| \lesssim \sigma \sqrt{\frac{\log \frac{1}{\delta}}{N}}, \quad \text{vs Chebyshev's} \quad |\overline{X} - \mu| \leq \sigma \sqrt{\frac{1/\delta}{N}}$$

where σ^2 is the variance/variance proxy.

Robust mean estimation

Question

Can we get the sub-Gaussian bound in general?

Split N points into l blocks of size m. The median of means estimator:

$$\mathrm{MOM}(X) = \mathrm{Median}\left(\frac{1}{m}\sum_{i=1}^{m} X_i, \ldots, \frac{1}{m}\sum_{i=(l-1)m+1}^{lm} X_i\right).$$

Robust mean estimation

Split N points into l blocks of size m. The median of means estimator:

$$\mathrm{MOM}(X) = \mathrm{Median}\left(\frac{1}{m}\sum_{i=1}^{m} X_i, \ldots, \frac{1}{m}\sum_{i=(l-1)m+1}^{lm} X_i\right).$$

Theorem: Nemirovsky, Yudin, 1983

Fix the number of blocks $l = 8\log\frac{1}{\delta} \geq 1$. *Assume* $\mathbb{E}(X-\mu)^2 = \sigma^2 < \infty$. *With probability at least* $1-\delta$,

$$|MOM(X) - \mu| \lesssim \sigma\sqrt{\frac{\log\frac{1}{\delta}}{N}},$$

Higher dimensions

Let X be a random vector in $\mathbb{R}^d$ with $\mu = \mathbb{E}\, X$. We observe $X_1, \ldots, X_N$ independent copies of X.

The covariance matrix $\Sigma = \mathbb{E}(X - \mu)(X - \mu)^T$ exists and unknown.

Question

Is it still possible to get the same behaviour as in the Gaussian case?

For reference: the performance of the sample mean in the Gaussian case. With probability at least $1 - \delta$,

$$\|\overline{X} - \mu\|_2 \lesssim \sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\frac{\|\Sigma\|_{op} \log \frac{1}{\delta}}{N}}.$$

Closer look at mean estimation

Theorem: Lugosi, Mendelson (AoS, 2019)

For any random vector X such that $\mathbb{E}\|X\|^2 < \infty$ there is an estimator $\hat{\mu}$ satisfying

$$\|\hat{\mu} - \mu\|_2 \lesssim \sqrt{\frac{Tr(\Sigma)}{N}} + \sqrt{\frac{\|\Sigma\|_{op} \log \frac{1}{\delta}}{N}},$$

This can be rewritten in the excess distortion form:

$$\mathbb{E}\|\hat{\mu} - X\|_2^2 - \mathbb{E}\|\mu - X\|_2^2 \lesssim \frac{\mathrm{Tr}(\Sigma)}{N} + \frac{\|\Sigma\|_{op} \log \frac{1}{\delta}}{N}.$$

Observe that $\mathbb{E}\|X - \mu\|^2 = \mathrm{Tr}(\Sigma)$.

What about k-means?

Recall $D(A) = \mathbb{E} \min_{a \in A} \|X - a\|^2$. For $k = 1$,

$$\underbrace{\mathbb{E}\,\|\hat{\mu} - X\|_2^2}_{D(\hat{A})} - \underbrace{\mathbb{E}\,\|\mu - X\|_2^2}_{D(A^*)} \lesssim \frac{\mathrm{Tr}(\Sigma)}{N} + \frac{\|\Sigma\|_{op} \log \frac{1}{\delta}}{N}.$$

1. The dependence on N is $O\left(\frac{1}{N}\right)$.
2. Even if $\|X\| < T$ almost surely the bound scales as $\mathbb{E}\,\|X - \mu\|^2$ (not as T^2).
3. It only requires the existence of two moments, that is $\mathbb{E}\,\|X\|^2 < \infty$.
4. It has the logarithmic dependence on the confidence, $\log \frac{1}{\delta}$.

What about k-means?

$$\underbrace{\mathbb{E}\left\|\hat{\mu}-X\right\|_2^2}_{D(\hat{A})}-\underbrace{\mathbb{E}\left\|\mu-X\right\|_2^2}_{D(A^*)}\lesssim\frac{\mathrm{Tr}(\Sigma)}{N}+\frac{\|\Sigma\|_{op}\log\frac{1}{\delta}}{N}.$$

1. $\Omega\left(\frac{1}{\sqrt{N}}\right)$ bound for $k\geq 2$ even if $\|X\|\leq 1$ by Antos (2005, IEEE Trans. on inf. theory),
2. Recall: best known bounds for k-means (under $\|X\|<T$)

$$\mathbb{E}\,D(\hat{A})-D(A^*)=O\left(T^2\sqrt{\frac{kd}{N}}\right),\ \mathbb{E}\,D(\hat{A})-D(A^*)=\tilde{O}\left(T^2\sqrt{\frac{k}{N}}\right).$$

What about k-means?

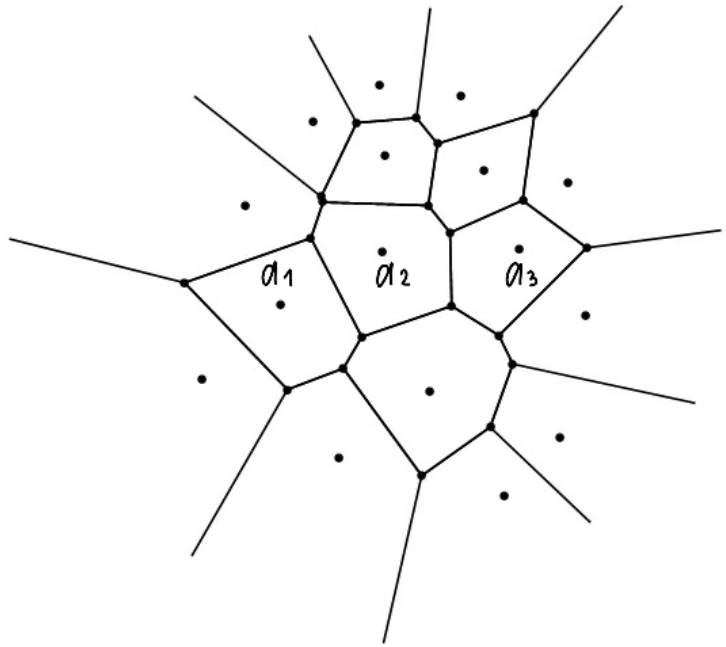
$$\underbrace{\mathbb{E}\,\|\hat{\mu} - X\|_2^2}_{D(\hat{A})} - \underbrace{\mathbb{E}\,\|\mu - X\|_2^2}_{D(A^*)} \lesssim \frac{\mathrm{Tr}(\Sigma)}{N} + \frac{\|\Sigma\|_{op} \log \frac{1}{\delta}}{N}.$$

$\mathbb{E}\,\|X - \mu\|^2$ scaling is not enough in the upper bound for $k \geq 2$.

Consider $k = 2$. We have $D(A^*) = 0$, $\mathbb{E}\,X^2 = 1$, and with constant probability $\hat{A} = \{0\}$. We have $D(\hat{A}) - D(A^*) = \Theta(1)$ with const. pr.

Voronoi diagrams and clustering

Consider $A^* = (a_1, \ldots, a_k)$ that minimizes $\mathbb{E} \min_{a \in A} \|X - a\|^2$.



Let $p_{\min}$ minimize the probability mass of a cluster of A^*.

Classical result: if $\mathbb{E} \|X\|^2 < \infty$ and $|\mathrm{supp}(P)| \geq k$ then $p_{\min} > 0$.

We have an invisible cluster if $p_{\min} \lesssim \frac{1}{N}$. A good guess $p_{\min} \gg \frac{1}{N}$.

Main Theorem

Theorem

Assume that P satisfies $\mathbb{E}\|X\|^2 < \infty$. Fix $\delta \in (0,1)$. There is an estimator $\hat{A}_{\delta,p_{\min}}$ that depends on $p_{\min}$ and δ such that, with probability at least $1-\delta$,

$$D(\hat{A}_{\delta,p_{\min}}) - D(A^*) \lesssim \mathbb{E}\|X-\mu\|^2 \left((\log N)^2 \sqrt{\frac{k}{Np_{\min}}} + \sqrt{\frac{\log \frac{1}{\delta}}{Np_{\min}}} \right)$$

- Proportional to $\mathbb{E}\|X-\mu\|^2$ – OK even in the bounded case.
- Dimension free and works in a separable Hilbert space.
- Exponential confidence $\log \frac{1}{\delta}$.

Statistical estimator

Fix $l \sim \log \frac{1}{\delta}$. Recall that for any function f,

$$\mathrm{MOM}(f(X)) = \mathrm{Median}\left(\frac{1}{m}\sum_{i=1}^{m} f(X_i), \ldots, \frac{1}{m}\sum_{i=(l-1)m+1}^{lm} f(X_i)\right).$$

$$\hat{A}_{\delta, p_{\min}} = \underset{\substack{A\subset\mathbb{R}^d, |A|=k \\ \text{s.t. each cluster of } A \text{ have } \geq Np_{\min}/2 \text{ points}}}{\arg\min} \mathrm{MOM}\left(\min_{a\in A} \|X-a\|^2\right),$$

Recall the definition of the empirically optimal quantizer

$$\hat{A} \in \underset{A\subset\mathbb{R}^d, |A|=k}{\arg\min} \sum_{i=1}^{N} \min_{a\in A} \|X_i - a\|^2.$$

Several words about the proof technique

1. Uniform results for MOM minimizers of the form:

$$\sup_{f\in\mathcal{F}}(\mathrm{MOM}(f)-\mathbb{E}f)\lesssim \mathbb{E}\sup_{f\in\mathcal{F}}(\mathrm{MOM}(f)-\mathbb{E}f)+\sqrt{\frac{\sup_{f\in\mathcal{F}}\mathrm{var}(f)\log\frac{1}{\delta}}{N}},$$

   like Talagrand's inequality for emp. processes w.o. $\|\mathcal{F}\|_\infty<\infty$.
2. Let $\hat{A}_{\delta,p_{\min}}=(a_1,\dots,a_k)$. With high probability,

$$\min_{i\le k}\|a_i\|\lesssim\sqrt{\mathbb{E}\|X\|^2},\quad\text{and}\quad\max_{i\le k}\|a_i\|\lesssim\sqrt{\frac{\mathbb{E}\|X\|^2}{p_{\min}}}.$$

3. Proof for finite d using chaining arguments.
4. Johnson-Lindenstrauss embedding to remove d (and to get extra log N-factors).

Lower bounds

In a special case of $\mathbb{R}^d$ we may remove $\log N$-factors and obtain:

$$D(\hat{A}_{\delta,p_{\min}}) - D(A^*) \lesssim \mathbb{E}\,\|X-\mu\|^2 \left(\sqrt{\frac{kd}{Np_{\min}}} + \sqrt{\frac{\log\frac{1}{\delta}}{Np_{\min}}}\right).$$

Theorem

Fix $k = 4$, $d = 1$, $\sigma > 0$ and $p_{\min} < 1/10$. For any empirically designed quantizer $\hat{A}_N$ there is a distribution with $\mathbb{E}\,X^2 = \sigma^2$ and $p_{\min}$ such that, with probability at least $\frac{1}{4}$,

$$D(\hat{A}_N, P) - D(A^*, P) \geq \frac{\mathbb{E}\,X^2}{80}\sqrt{\frac{1}{Np_{\min}}}.$$

Open questions

- More computational friendly algorithms.
- Our techniques recover $\tilde{O}\left(T^2\sqrt{\frac{k}{N}}\right)$ but in general $p_{\min} \leq \frac{1}{k}$.

  In the heavy-tailed scenarios our bound is at least linear in k.

$$\sqrt{\frac{k}{p_{\min} N}} \geq \frac{k}{\sqrt{N}}.$$

  What is the best possible dependence on k in the upper bounds?
- Fast rates: when is $O\left(\frac{1}{N}\right)$ possible in the heavy-tailed setup? Levrard (2014 AoS) analyzes this in Hilbert spaces, but in the bounded setup.

Informal conclusions

Robust statistical k-means clustering is possible for heavy-tailed distributions with the rate of convergence

$$D(\tilde{A}) - D(A^*) = \tilde{O}\left(\mathbb{E}\,\|X-\mu\|^2 \sqrt{\frac{k}{Np_{\min}}}\right).$$

Since $p_{\min} > 0$ we have $D(\tilde{A}) - D(A^*) \to 0$ as $N \to \infty$ which is a version of Pollard's strong consistency result.

If $p_{\min} = p_{\min}(N)$ then

$$Np_{\min} \to \infty \iff \text{No invisible clusters} \iff D(\hat{A}) - D(A^*) \to 0.$$