A short note on the median-of-means estimator

Yen-Chi Chen
University of Washington
July 13, 2020

The main reference of this short note is Chapter 3 of

> [L2019] Lerasle, M. (2019). Lecture notes: Selected topics on robust statistical learning theory. arXiv preprint arXiv:1908.10761.

The median-of-means (MoM) is an old method but it has received a lot of attentions these day. The MoM estimator gives a nice concentration bound around the target parameter under mild conditions and it can be applied to an empirical process as well (it is called a median-of-means process in [L2019]).

1 MoM estimator of the mean

Consider a simple scenario where we observe $X_1, \cdots, X_n \sim F$. The goal is to estimator the mean of the underlying distribution $\mu_0 = \mathbb{E}(X_1) = \int x dF(x)$.

The MoM estimator works as follows. Assume that the sample size $n = K \cdot B$, where K is the number of subsamples and B is the size of each subsample. We first randomly split the data into K subsample and compute the mean using each subsample, which leads to estimators

$$\widehat{\mu}_1, \cdots, \widehat{\mu}_K$$

and each estimator is based on B observations. The MoM estimator is the median of all these estimator, i.e.,

$$\widehat{\mu}_{\mathsf{MoM}} = \mathsf{median}(\widehat{\mu}_1, \cdots, \widehat{\mu}_K).$$

Why is this estimator useful? It turns out that even with a very mild condition: $\mathsf{Var}(X_1) = \sigma^2 < \infty$, the MoM estimator has nice concentration inequality under 'finite sample' case.

Proposition 1 *Assume that* $\mathsf{Var}(X_1) < \infty$. *Then the MoM estimator has the following property:*

$$P(|\widehat{\mu}_{\mathsf{MoM}} - \mu_0| > \varepsilon) \leq e^{-2K\left(\frac{1}{2} - \frac{K}{n}\frac{\sigma^2}{\varepsilon^2}\right)^2} = e^{-2\frac{n}{B}\left(\frac{1}{2} - \frac{\sigma^2}{B\varepsilon^2}\right)^2}$$

for every $n = K \cdot B$.

Proposition 1 implies the following two forms, with different choices of ε. When we choose $\varepsilon = \sqrt{(2+\delta)K/n} = \sqrt{(2+\delta)/B}$, which leads to

$$P(|\widehat{\mu}_{\mathsf{MoM}} - \mu_0| > \sqrt{(2+\delta)K/n}) \leq e^{-K\frac{\delta^2}{2(2+\delta)^2}} \tag{1}$$

and when we further choose $\delta = 2$, it becomes

$$P(|\widehat{\mu}_{\mathsf{MoM}} - \mu_0| > \sqrt{4K/n}) \leq e^{-K/8}. \tag{2}$$

A powerful feature of Proposition 1 is that finite sample exponential concentration is not easy to obtain when we only assume variance exists. X_i's can be unbounded in this case (when X_i is bounded, Hoeffding's inequality implies such a concentration).

Now we prove Proposition 1. The proof is very simple and elegant.

Proof. First, observe that the event

$$\{|\widehat{\mu}_{\mathsf{MoM}} - \mu_0| > \varepsilon\}$$

implies that at least $K/2$ of $\widehat{\mu}_\ell$ has to be outside ε distance to μ_0. Namely,

$$\{|\widehat{\mu}_{\mathsf{MoM}} - \mu_0| > \varepsilon\} \subset \left\{ \sum_{\ell=1}^{K} I(|\widehat{\mu}_\ell - \mu_0| > \varepsilon) \geq \frac{K}{2} \right\}.$$

Define $Z_\ell = I(|\widehat{\mu}_\ell - \mu_0| > \varepsilon)$ and let $p_{\varepsilon,B} = \mathbb{E}(Z_\ell) = P(|\widehat{\mu}_\ell - \mu_0| > \varepsilon)$, then the above implies that

$$\begin{aligned} P(|\widehat{\mu}_{\mathsf{MoM}} - \mu_0| > \varepsilon) &\leq P\left(\sum_{\ell=1}^{K} Z_\ell \geq \frac{K}{2} \right) \\ &= P\left(\sum_{\ell=1}^{K} (Z_\ell - \mathbb{E}(Z_\ell)) \geq \frac{K}{2} - K p_{\varepsilon,B} \right) \\ &= P\left(\frac{1}{K} \sum_{\ell=1}^{K} (Z_\ell - \mathbb{E}(Z_\ell)) \geq \frac{1}{2} - p_{\varepsilon,B} \right). \end{aligned}$$

The key trick of the MoM estimator is that the random variable Z_ℓ is IID and is bounded. So by Hoeffding's inequality (one-sided),

$$P\left(\frac{1}{K} \sum_{\ell=1}^{K} (Z_\ell - \mathbb{E}(Z_\ell)) \geq t \right) \leq e^{-2Kt^2}.$$

As a result,

$$\begin{aligned} P(|\widehat{\mu}_{\mathsf{MoM}} - \mu_0| > \varepsilon) &\leq P\left(\frac{1}{K} \sum_{\ell=1}^{K} (Z_\ell - \mathbb{E}(Z_\ell)) \geq \frac{1}{2} - p_{\varepsilon,B} \right) \\ &\leq e^{-2K(\frac{1}{2} - p_{\varepsilon,B})^2}. \end{aligned}$$

To conclude that proof, note that the variance $\sigma^2 = \mathsf{Var}(X_1) < \infty$ and the Chebeshev's inequality implies

$$p_{\varepsilon,B} = P(|\widehat{\mu}_\ell - \mu_0| > \varepsilon) \leq \frac{\sigma^2}{B\varepsilon^2} = \frac{K}{n} \frac{\sigma^2}{\varepsilon^2}.$$

So the bound becomes

$$P(|\widehat{\mu}_{\mathsf{MoM}} - \mu_0| > \varepsilon) \leq e^{-2K(\frac{1}{2} - p_{\varepsilon,B})^2} \leq e^{-2K(\frac{1}{2} - \frac{K}{n}\frac{\sigma^2}{\varepsilon^2})^2}.$$

□

2 MoM process: example of the KDE

The MoM approach can also be applied to an empirical process. To motivate the problem, we consider the kernel density estimator (KDE). The observations $X_1, \cdots, X_n$ are IID from an unknown distribution with a PDF p. With a full-sample, the KDE is

$$\widehat{p}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right),$$

where K is a kernel function such as a Gaussian and $h > 0$ is the smoothing bandwidth.

Under the MoM procedure, we split the sample into K subsamples and compute the KDE within each sample, leading to

$$\widehat{p}_1(x), \cdots, \widehat{p}_K(x).$$

The MoM KDE is

$$\widehat{p}_{\mathsf{MoM}}(x) = \mathsf{median}(\widehat{p}_1(x), \cdots, \widehat{p}_K(x))$$

Note: this estimator may not be a density (it may not integrate to 1) but we can simply rescale it to resolve this issue.

Now we will show that suppose we have a bound of the form

$$P(\sup_x |\widehat{p}(x) - p(x)| > \varepsilon) \leq q_{n,\varepsilon}. \tag{3}$$

We then have a bound on $\widehat{p}_{\mathsf{MoM}}$ as

$$P\left(\sup_x |\widehat{p}_{\mathsf{MoM}}(x) - p(x)| > \varepsilon\right) \leq e^{-2K(\frac{1}{2} - q_{\varepsilon,B})^2}. \tag{4}$$

Derivation of equation (4). We use a similar derivation as the proof of Proposition 1. The following derivation is modified from

> [HBMV2020] Humbert, P., Bars, B. L., Minvielle, L., & Vayatis, N. (2020). Robust Kernel Density Estimation with Median-of-Means principle. arXiv preprint arXiv:2006.16590.

Note that the event

$$\left\{\sup_x |\widehat{p}_{\mathsf{MoM}}(x) - p(x)| > \varepsilon\right\} \subset \left\{\sup_x \sum_{\ell=1}^{K} I(|\widehat{p}_\ell(x) - p(x)| > \varepsilon) > \frac{K}{2}\right\}.$$

Using the fact that

$$|\widehat{p}_\ell(x) - p(x)| \leq \sup_x |\widehat{p}_\ell(x) - p(x)|,$$

we have

$$I(|\widehat{p}_\ell(x) - p(x)| > \varepsilon) \leq I(\sup_x |\widehat{p}_\ell(x) - p(x)| > \varepsilon)$$

which implies

$$\sup_x \sum_{\ell=1}^{K} I(|\widehat{p}_\ell(x) - p(x)| > \varepsilon) \leq \sum_{\ell=1^K} I(\sup_x |\widehat{p}_\ell(x) - p(x)| > \varepsilon).$$

As a result,

$$\begin{aligned}\left\{\sup_x |\widehat{p}_{\mathsf{MoM}}(x) - p(x)| > \varepsilon\right\} &\subset \left\{\sup_x \sum_{\ell=1}^{K} I(|\widehat{p}_\ell(x) - p(x)| > \varepsilon) > \frac{K}{2}\right\} \\ &\subset \left\{\sum_{\ell=1}^{K} I(\sup_x |\widehat{p}_\ell(x) - p(x)| > \varepsilon) > \frac{K}{2}\right\}\end{aligned}$$

Define $Y_\ell = I(\sup_x |\widehat{p}_\ell(x) - p(x)| > \varepsilon)$ and denote $E(Y_\ell) = P(\sup_x |\widehat{p}_\ell(x) - p(x)| > \varepsilon) = q_{\varepsilon,B}$. Then we conclude that

$$\begin{aligned}P\left(\sup_x |\widehat{p}_{\mathsf{MoM}}(x) - p(x)| > \varepsilon\right) &\leq P\left(\sum_{\ell=1}^{K} Y_\ell \geq \frac{K}{2}\right) \\ &= P\left(\sum_{\ell=1}^{K} (Y_\ell - \mathbb{E}(Y_\ell)) \geq \frac{K}{2} - K q_{\varepsilon,B}\right) \\ &= P\left(\frac{1}{K}\sum_{\ell=1}^{K} (Y_\ell - \mathbb{E}(Y_\ell)) \geq \frac{1}{2} - q_{\varepsilon,B}\right) \\ &\leq e^{-2K(\frac{1}{2} - q_{\varepsilon,B})^2},\end{aligned}$$

which is equation (4).

Concrete rate under KDE. Suppose the PDF p belongs to a β-Hölder class, then we have the following concentration under suitable conditions of K and h:

$$P\left(\sup_x |\widehat{p}(x) - p(x)| > C_1\sqrt{\frac{\delta|\log h|}{nh^d}} + C_2 h^\beta\right) \leq e^{-\delta},$$

where C_1 and C_2 are some constants depending on the Hölder class and the kernel function; see Lemma 1 of [HBMV2020]. By choosing $\varepsilon = C_1\sqrt{\frac{\log t|\log h|}{nh^d}} + C_2 h^\beta$ with $t \geq 1$ and use the fact that B plays the same role as n, we obtain

$$q_{\varepsilon,B} \leq \frac{1}{t},$$

which leads to

$$P\left(\sup_x |\widehat{p}_{\mathsf{MoM}}(x) - p(x)| > C_1\sqrt{\frac{\log t|\log h|}{Bh^d}} + C_2 h^\beta\right) \leq e^{-2K(\frac{1}{2} - \frac{1}{t})^2}.$$

When $t = 4$ and use the fact that $B = n/K$, we further obtain

$$P\left(\sup_x |\widehat{p}_{\mathsf{MoM}}(x) - p(x)| > C_1\sqrt{\frac{K\log 4|\log h|}{nh^d}} + C_2 h^\beta\right) \leq e^{-K/8}.$$

After some algebra, we conclude that

$$\widehat{p}_{\mathsf{MoM}}(x) - p(x) = O(h^{\beta}) + O_P\left(\sqrt{\frac{K\log h}{nh^d}}\right).$$

If we set the total number of subsample K to be fixed and let $n \to \infty$, the convergence rate is the same as the original KDE.

Robustness of MoM KDE. Although MoM KDE does not improve the convergence rate (in fact, it may have a slower rate depending on the choice of K), it improves the robustness of the density estimator against outliers. In [HBMV2020], the authors showed that as long as we have outliers that are strictly less than $K/2$ points, the MoM KDE will not be affected by these outliers. One intuition is that when the number of outliers is less than $K/2$, then we have at least $K/2$ subsamples that are not affected by the outliers, which occupy more than half of the subsamples.

3 MoM process: general case

To conclude this note, we briefly describe a general MoM process. Suppose that the parameter of interest is $\mu(f) = \mathbb{E}(f(X_1))$ for $f \in \mathcal{F}$. Then a full-sample estimator is

$$\widehat{\mu}(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

for each $f \in \mathcal{F}$. Using the MoM procedure, we obtain K subsample estimators ($K = nB$)

$$\widehat{\mu}_1(f), \cdots, \widehat{\mu}_K(f).$$

The MoM estimator is then

$$\widehat{\mu}_{\mathsf{MoM}}(f) = \mathsf{median}(\widehat{\mu}_1(f), \cdots, \widehat{\mu}_K(f)), \quad f \in \mathcal{F}.$$

In reality, we are often interested in the L_∞ loss of the estimator, i.e.,

$$\sup_{f\in\mathcal{F}} |\widehat{\mu}_{\mathsf{MoM}}(f) - \mu(f)|.$$

Suppose that for a full-sample estimator $\widehat{\mu}(f)$, we have

$$P(\sup_{f\in\mathcal{F}} |\widehat{\mu}(f) - \mu(f)| > \varepsilon) \le \eta_{\varepsilon,n},$$

or equivalently,

$$P(\sup_{f\in\mathcal{F}} |\widehat{\mu}(f) - \mu(f)| > \phi_{\delta,n}) \le \frac{1}{\delta}, \tag{5}$$

for some $\delta > 2$.

Using the same derivation as equation (4), one can show that equation (5) implies

$$P\left(\sup_{f\in\mathcal{F}} |\widehat{\mu}_{\mathsf{MoM}}(f) - \mu(f)| > \phi_{\delta,n/K}\right) \le e^{-2K(\frac{1}{2}-\frac{1}{\delta})^2}. \tag{6}$$

4 Remarks.

Finite-sample exponential concentration without sub-Gaussian. A feature of an MoM estimator is that it establishes a finite-sample exponential concentration without assuming sub-Gaussian or bounded condition. In Proposition 1, all we need is the existence of the second moment, which is a very mild condition. We allow for a heavy tail and the estimator can still be concentrating rapidly. The key is the use of median, which makes the estimator more robust.

Optimal rate without sub-Gaussian assumption in high-dimensions. Following the previous point on the exponential concentration, one can show that the MoM estimator of a multivariate sample mean in $\mathbb{R}^d$ can achieve a rate of $O_P\left(\sqrt{\frac{\log d}{n}}\right)$ without assuming that X is sub-Gaussian. All we need is moment conditions with at least $2\log d$ moments exists. See Theorem 48 and Remark 49 in [L2019].

Two major benefits of using an MoM estimator. To sum up, there are two major benefits of using an MoM estimator. First, it allows a heavy tail distribution–we can still obtain an exponential concentration with only moment conditions. Second, it allows outliers in the data. As is argued in the MoM KDE case, an MoM estimator is often more robust to outliers. This feature is not limited to the KDE case, the robustness against outliers also applies to other scenarios.