Data-Driven Confidence Intervals with Optimal Rates for the Mean of Heavy-Tailed Distributions

Ambrus Tamás Szabolcs Szentpéteri Balázs Csanád Csáji

Institute for Computer Science and Control (SZTAKI), Hungarian Research Network (HUN-REN), Budapest, Hungary
Department of Probability Theory and Statistics, Institute of Mathematics, Eötvös Loránd University (ELTE), Budapest, Hungary

Overview

- *Uncertainty quantification* (UQ) for the mean of symmetric variables.
- *Non-asymptotically exact confidence intervals* (CIs) for the *true* expected value.
- The method combines *resampling* with *median-of-means* (MoM) estimates.
- *Distribution-free* guarantees are presented.
- *Exact* (user-chosen) coverage probabilities.
- *Optimal bounds* for the sizes of the CIs under heavy-tailed moment conditions.
- *Data-driven method;* the construction needs no information about the moments.

Main Assumptions

A1 $Y_1, \dots, Y_n$ *are i.i.d. in* $\mathbb{R}$ $(\mathcal{D}_0 \doteq \{Y_i\}_{i=1}^n)$.

A2 Q_Y *is symmetric about* μ.

A3 $\mathbb{E}\big[\,|Y - \mathbb{E}Y|^{1+a}\big] = M < \infty$ *for* $a \in (0, 1]$.

Preliminaries

Let k be an (odd) integer, $n = k\tilde{n}$ and the median-of-means (MoM) estimate defined by

$$\hat{\mu}(\mathcal{D}_0) \doteq \operatorname{med}\left(\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} Y_i, \dots, \frac{1}{\tilde{n}}\sum_{i=(k-1)\tilde{n}+1}^{k\tilde{n}} Y_i\right).$$

The main advantage of the MoM estimate $\hat{\mu}$ is:

Median-of-Means Estimate

Assume $A1$, $A2$ and $A3$. Let $\delta \in (0,1)$, $k = \lceil 8\ln(2/\delta)\rceil$ and $n = \tilde{n}k$, then

$$\mathbb{P}\left(\,|\hat{\mu} - \mu| > 8\left(\frac{12M^{1/a}\ln(1/\delta)}{n}\right)^{\frac{a}{1+a}}\right) \leq \delta.$$

Moreover, the rate is optimal w.r.t. δ and n.

Resampled Median-of-Means

Let us consider the following hypotheses:

$$H_0:\ \mu = \theta \quad \text{and} \quad H_1:\ \mu \neq \theta.$$

Let p denote the significance level and r, m be integers s.t. $p = r/m$. Let $\{\alpha_{i,j}\}$ be i.i.d. Rademacher variables for $i \in [n]$ and $j \in [m-1]$.

Hypothesis Test (for $\mu = \theta$)

1. Construct $m-1$ alternative datasets
$$\mathcal{D}_j(\theta) \doteq \{\alpha_{1,j}(Y_1 - \theta) + \theta, \dots, \alpha_{n,j}(Y_n - \theta) + \theta\}$$
for $j \in [m-1]$ and $\mathcal{D}_0(\theta) \doteq \mathcal{D}_0$.
2. Compute the reference variables
$$S_j(\theta) \doteq |\,\hat{\mu}(\mathcal{D}_j(\theta)) - \theta\,| \quad \text{for} \quad j \in [m-1]_0.$$
3. Compute the rank $\mathcal{R}(\theta)$ according to
$$\mathcal{R}(\theta) \doteq 1 + \sum_{i=1}^{m-1} \mathbb{I}\big(S_0(\theta) \succ_\pi S_j(\theta)\big),$$
4. Reject H_0 if and only if $\mathcal{R}(\theta) > m - r$.

Key Observations

If $\theta = \mu$, then $\{S_j(\theta)\}_{j=0}^{m-1}$ are exchangeable, however, if θ is "far" from μ, then $S_0(\theta)$ should be greater than $S_j(\theta)$.

The proposed hypothesis test provides *exact, non-asymptotic* and *distribution-free* guarantees:

Type I Error Probability

Assume A1 and A2, then for every $1 \leq r < m$

$$\mathbb{P}\,(\,\mathcal{R}(\mu) > m - r\,) = r/m.$$

The hypothesis test is completely *data-driven,* i.e., the algorithm does not need moment information.

Confidence Intervals

We include those parameters in the confidence set that are accepted by the presented hypothesis test.

$$\Theta_n \doteq \{\,\theta : \mathcal{R}(\theta) \leq m - r\,\}$$

Note that we do not need to generate random signs for each θ, the same set of signs can be used. Under A1 and A2, Θ_n is an *exact* confidence set for μ, i.e.,

$$\mathbb{P}\big(\,\mu \in \Theta_n\,\big) = 1 - r/m.$$

An important consequence of the construction is that Θ_n admits the special form of

$$\Theta_n = \bigcup_{\substack{J \subseteq [m-1],\\ |J| = m-r}} \bigcap_{j \in J} \{\,\theta : S_0(\theta) \prec_\pi S_j(\theta)\,\},$$

thus Θ_n is an interval that contains $\hat{\mu}(\mathcal{D}_0)$, if $\Theta_n \neq \emptyset$. We can compute the endpoints efficiently. An illustrative example of the CI construction for $m = 2$ and $k = 3$ is presented in Figure 1.

Shrinkage Rate

Let us consider the diameter of Θ_n

$$d(\Theta_n) \doteq \sup\{\,|\theta_1 - \theta_2| : \theta_1, \theta_2 \in \Theta\,\}.$$

Our main *non-asymptotic* and *distribution-free* result about the shrinkage rate of Θ_n is:

Main Result

Assume $A1$, $A2$ and $A3$. Let $r < m$ be user-chosen integers, $\delta > 0$, $k = \lceil 8\ln(20(m-r)/\delta)\rceil$, then for $n \geq k(k + 8\ln(k))$ we have

$$\mathbb{P}\left(d(\Theta_n) > 8\left(\frac{12M^{1/a}\ln\left(\frac{10(m-r)}{\delta}\right)}{n}\right)^{\frac{a}{1+a}}\right) \leq \delta.$$

Multivariate Extension

A4 $Y - \mu \overset{d}{=} \mu - Y$, *for some vector* $\mu \in \mathbb{R}^q$.

A5 $\Sigma \doteq \mathbb{E}[(Y - \mathbb{E}Y)(Y - \mathbb{E}Y)^\mathrm{T}]$ *exists.*

For $\theta \in \mathbb{R}^q$ let us consider

$$H_0:\ \mu = \theta \quad \text{and} \quad H_1:\ \mu \neq \theta.$$

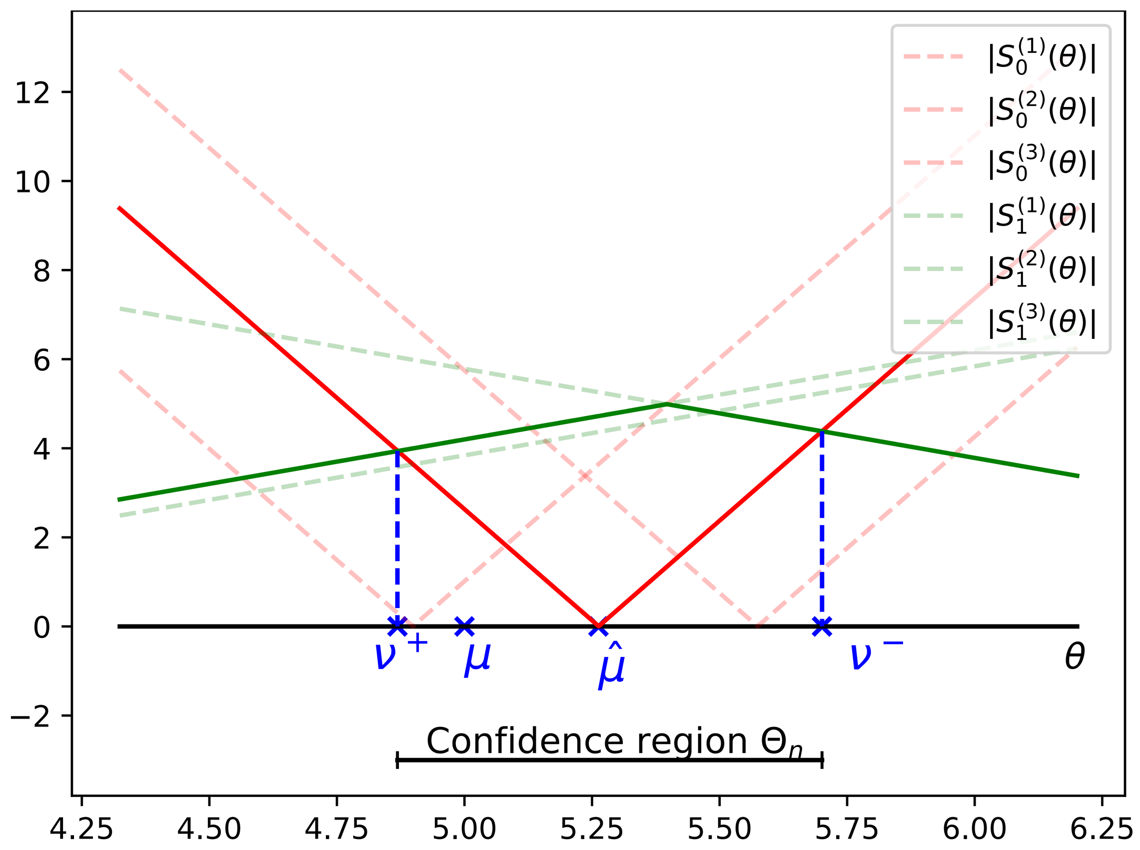

Figure 1: 50% confidence interval, $m = 2$, $n = 30$ and $k = 3$.

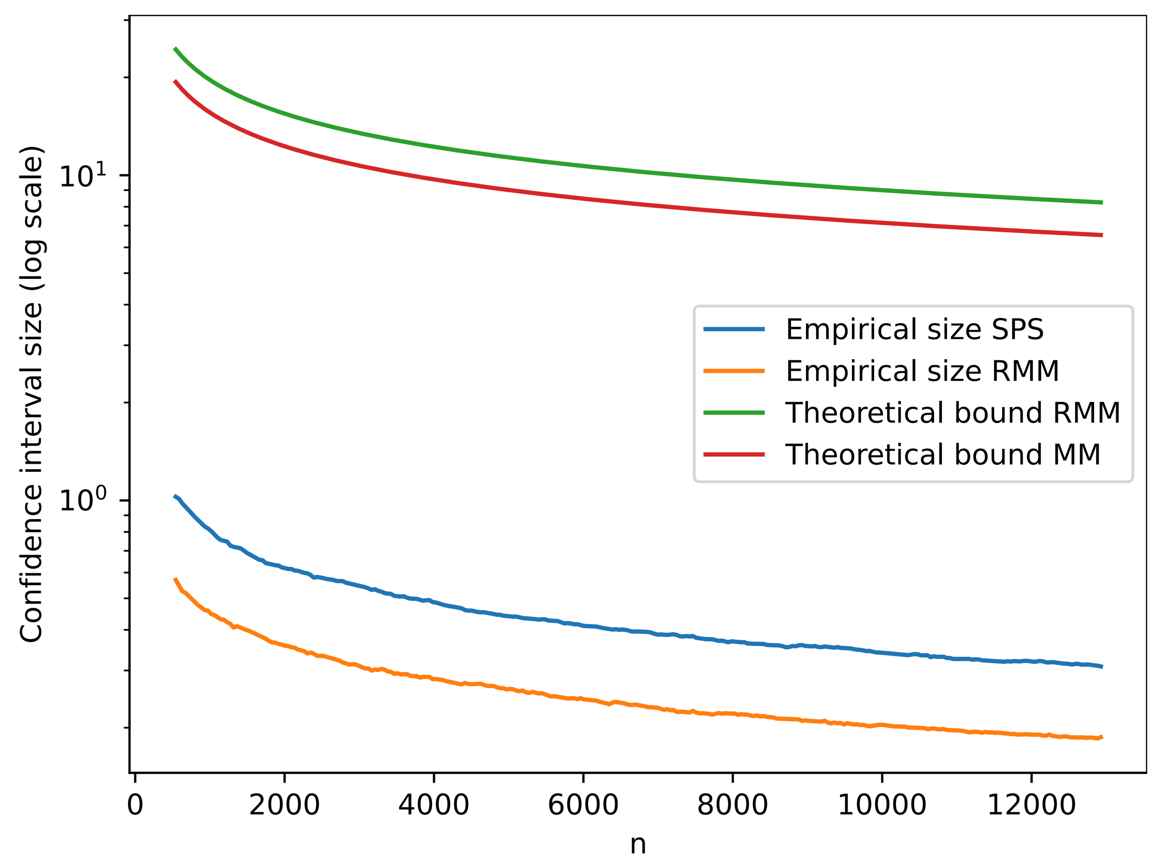

Figure 2: Comparison of confidence interval sizes.

Let $\tilde{\mu}$ be a subgaussian mean estimator, i.e., there exist $c_1, c_2, c_3 > 0$ such that for every $\delta > 0$:

$$\mathbb{P}\left(\|\tilde{\mu} - \mu\| > \sqrt{\frac{c_1\operatorname{Tr}(\Sigma)}{n}} + \sqrt{\frac{c_2\lambda^*\ln(c_3/\delta)}{n}}\right) \leq \delta.$$

The MoM tournament estimator and its polynomial time relaxation are in the subgaussian regime. The multivariate alternative datasets are

$$\mathcal{D}_j(\theta) \doteq \{\alpha_{i,j}\mathbb{1} \odot (Y_1 - \theta) + \theta\}_{i=1}^n,$$

where $\odot$ denotes the Hadamard (element-wise) product and the reference variable functions

$$S_j(\theta) \doteq \|\,\tilde{\mu}(\mathcal{D}_j(\theta)) - \theta\,\| \quad \text{for } j \in [m-1]_0.$$

Type II Error Rate

Assume A1, A4, and A5 Let $\delta > 0$ and $r < m$ be user-chosen integers. For $\theta \neq \mu$ if

$$\sqrt{\frac{c_1(T + \Delta^2)}{n}} + \sqrt{\frac{c_2(\lambda^* + \Delta^2)\ln\left(\frac{c_3(m-r)}{\delta}\right)}{n}} < \frac{\Delta}{2},$$

with $T = \operatorname{Tr}(\Sigma)$ holds for $\Delta \doteq \|\theta - \mu\|$, then

$$\mathbb{P}(\,\mathcal{R}(\theta) > m - r\,) \geq 1 - \delta.$$

Future research directions include applications to robust optimization and multiarmed bandits.

Contact Information

Email: tamasamb@sztaki.hu (Ambrus Tamás)
Website: http://www.sztaki.hu/~tamas-ambrus/