Non-asymptotic bounds for probability weighted moment estimators

joint work with A. Ben-Hamou (LPSM, SU) and P. Naveau (LSCE, CNRS)

Supported by ANR T-REX

Maud Thomas

Sorbonne Université
Laboratoire de Probabilités, Statistique et Modélisation, UMR CNRS 8001

January 17, 2022

Probability weighted moments

- Motivated by hydrologists and applied statisticians (Hosking and Wallis 1987; Landwehr et al. 1979)
- Explicit and simple expressions can be found for classical Extreme Value distributions (Katz et al. 2002)
- Let X be a real-valued random variable with c.d.f. F ($\overline{F} = 1 - F$), such that $\mathbb{E}|X| < +\infty$
- For integers $r, s \geq 0$, the probability weighted moments (PWM) are defined as

$$\mathbb{E}\left[XF(X)^r\overline{F}(X)^s\right] \tag{1}$$

- Generalizations have been proposed: (Diebolt, et al. 2003, 2004, 2007) and (Guillou et al., 2009)
  → For example, (Diebolt et al., 2007) considered $\mathbb{E}[X\omega(F(X))$ where ω is a weight function
  → Choice of ω = trade-off between complexity and moment existence
  → If $\mathbb{E}[|X|] < +\infty$, the choice $\omega(u) = u^r$ offers simplicity
  → Choice that we opt in this work

Link with order statistics

- $(X_1, \ldots, X_p)$ i.i.d. p-sample with c.d.f F
- $X_{(1:p)} \leq \cdots \leq X_{(p:p)}$ corresponding order statistics
- For $1 \leq q \leq p$,

(2) $$\alpha_{q:p} = \mathbb{E}\left[X_{(q:p)}\right] = q\binom{p}{q}\mathbb{E}\left[XF(X)^{q-1}\overline{F}(X)^{p-q}\right]$$

and equivalently

$$\mathbb{E}\left[XF(X)^{r}\overline{F}(X)^{s}\right] = \frac{\mathbb{E}\left[X_{(r+1:s+r+1)}\right]}{(r+1)\binom{s+r+1}{r+1}}$$

PWM estimators

- From (1), replacing F by its empirical estimator $\mathbb{F}_n$, PWM estimators can be built
  → Let $X_1, \ldots, X_n \sim F$

$$\frac{1}{n}\sum_{i=1}^{n} X_i \mathbb{F}_n(X_i)^r \overline{\mathbb{F}}_n(X_i)^s$$

- From (2), estimate $\alpha_{q:p}$ by

$$T_{q:p} = \frac{1}{\binom{n}{p}} \sum_{1 \le i_1 < \ldots < i_p \le n} S_q(X_{i_1}, \ldots, X_{i_p})$$

where $S_q(X_{i_1}, \ldots, X_{i_p})$ corresponds to the qth order statistics in $(X_{i_1}, \ldots, X_{i_p})$ (Landwehr et al. 1979)
  → Deduce an estimator for the PWM
  → Strong link with U-statistics (Hoeffding, 1948, 1963)
  → In case of no ties

$$T_{q:p} = \sum_{k=1}^{n} a_k X_{(k:n)}\,, \text{with } a_k = \frac{\binom{k-1}{q-1}\binom{n-k}{p-q}}{\binom{n}{p}}$$

PWM estimators

- From (1), replacing F by its empirical estimator $\mathbb{F}_n$, PWM estimators can be built

  → Let $X_1, \ldots, X_n \sim F$

$$\frac{1}{n}\sum_{i=1}^{n} X_i \mathbb{F}_n(X_i)^r \overline{\mathbb{F}}_n(X_i)^s$$

- From (2), estimate $\alpha_{q:p}$ by

$$T_{q:p} = \frac{1}{\binom{n}{p}} \sum_{1 \le i_1 < \ldots < i_p \le n} S_q(X_{i_1}, \ldots, X_{i_p})$$

  where $S_q(X_{i_1}, \ldots, X_{i_p})$ corresponds to the qth order statistics in $(X_{i_1}, \ldots, X_{i_p})$ (Landwehr et al. 1979)

  → Deduce an estimator for the PWM

  → Strong link with U-statistics (Hoeffding, 1948, 1963)

  → In case of no ties

$$T_{q:p} = \sum_{k=1}^{n} a_k X_{(k:n)}\,, \text{with } a_k = \frac{\binom{k-1}{q-1}\binom{n-k}{p-q}}{\binom{n}{p}}$$

Can we obtain finite-sample results with large probability for PWM estimators under minimal moment conditions?

Concentration of measure phenomenon

Concentration of measure phenomenon

Any function of many independent random variables that does not depend too much on any of them is concentrated around its mean value.

- Markov inequality: $X > 0$, $\mathbb{P}\{X > t\} \leq \mathbb{E}X/t$.
- Chebyschev inequality: $\mathbb{P}\{|X - \mathbb{E}X| > t\} \leq \operatorname{Var} X/t^2$
  → First step: variance bound
  → Second step: deviance inequality

- **Sub-Gaussian variable** X is sub-Gaussian on the left and on the right with variance factor v if

$$\mathbb{P}[|X - \mathbb{E}X| \geq t] \leq \exp\left(-\frac{t^2}{2v}\right)$$

- **Sub-gamma variable** X is sub-gamma on the left and on the right with variance factor v and scale factor if

$$\mathbb{P}[|X - \mathbb{E}X| \geq t] \leq \exp\left(-\frac{t^2}{2(v+ct)}\right)$$

  → Bernstein-type inequality

General representation of order statistics

Rényi representation(Rényi, 1953)

$$\left(Y_{(1:n)},\dots,Y_{(n-k+1:n)},\dots,Y_{(n:n)}\right) \sim \left(\frac{E_1}{n},\dots,\sum_{i=n-k+1}^{n}\frac{E_{n-i+1}}{i},\dots,\sum_{i=1}^{n}\frac{E_{n-i+1}}{i}\right)$$

where $E_1,\dots,E_n$ i.i.d. $\sim \mathcal{E}\mathrm{xp}(1)$

- Sum of independent random variables = good framework for concentration
- **U-function**: if $F^{\leftarrow}$ = quantile function, define

$$U(t) = F^{\leftarrow}(1-1/t), t \in (1,\infty)$$

Representation for order statistics

$$(X_{(1:n)},\dots,X_{(n:n)}) \overset{d}{=} \left(\left(U\circ\exp\right)(Y_{(1:n)}),\dots,\left(U\circ\exp\right)(Y_{(n:n)})\right)$$

$\Rightarrow$ Derive concentration inequalities for order statistics?

Concentration results for order statistics?

Poincaré inequality (Talagrand, 1991 ; Bobkov & Ledoux, 1997)

Let g a differentiable function on $\mathbb{R}^n$, and $Z = g(E_1, \ldots, E_n)$, then

$$\mathrm{Var}[Z] \leqslant 4\mathbb{E}\left[||\nabla g||^2\right]$$

Variance bound for order statistics

$$\mathrm{Var}[X_{(n-k+1:n)}] \leq 4\sum_{i=k}^{n} \frac{1}{i^2}\mathbb{E}\left[\frac{1}{h(X_{(n-k+1:n)})^2}\right] \leq \frac{4}{k}\left(1+\frac{1}{k}\right)\mathbb{E}\left[\frac{1}{h(X_{(n-k+1:n)})^2}\right]$$

Concentration results for order statistics?

Bernstein-type inequality of exponential vectors (Bobkov & Ledoux, 1997)

Let g a differentiable fucntion on $\mathbb{R}^n$, and $Z = g(E_1, \ldots, E_n)$. Assume that $\max_i \left|\partial_i g\right| < \infty$ and let $v = \sup ||\nabla g||^2$. Then, for all $0 < \delta < 1/2$,

$$\mathbb{P}\left\{ |Z - \mathbb{E}Z| \geqslant \sqrt{8v\ln(1/\delta)} + \max_i \left|\partial_i g\right| \ln(1/\delta) \right\} \leqslant 2\delta \ .$$

- h hazard rate associated with F defined as $h = f/(1-F)$

Order statistics are sub-gamma

If h is non-decreasing, then for $\delta > 0$

$$\mathbb{P}\left\{ \left|X_{(n-k+1:n)} - \mathbb{E}X_{(n-k+1:n)}\right| \leqslant \sqrt{\frac{8}{k}\left(1+\frac{1}{k}\right)\mathbb{E}\left[\frac{1}{h(X_{(n-k+1:n)})^2}\right]\ln(1/\delta)} + \frac{\ln(1/\delta)}{k \inf_x h(x)} \right\} \leqslant 1 - 2\delta \ .$$

Concentration results for PWM estimators?

- **Idea:** Thanks to Rényi representation consider $T_{q:p}$ also a function of independent exponential variables

$T_{q:p}$ is sub-gamma

Assume h is non-decreasing. Then, $T_{q:p} - \mathbb{E}\left[T_{q:p}\right]$ is sub-gamma on the right tail with variance factor

$$v = 4 \sum_{j=1}^{n} \mathbb{E}\left[\left(\frac{1}{j} \sum_{k=j}^{n} \frac{a_k}{h(X_{(n-k+1:n)})}\right)^2\right]$$

and scale factor

$$c = \frac{1}{\inf h} \max_{1 \le j \le n} \frac{1}{j} \sum_{k=j}^{n} a_k \,.$$

Concentration results for PWM estimators?

- **Idea:** Thanks to Rényi representation consider $T_{q:p}$ also a function of independent exponential variables

$T_{q:p}$ is sub-gamma

Assume h is non-decreasing. Then, $T_{q:p} - \mathbb{E}\left[T_{q:p}\right]$ is sub-gamma on the right tail with variance factor

$$v = 4\sum_{j=1}^{n} \mathbb{E}\left[\left(\frac{1}{j}\sum_{k=j}^{n} \frac{a_k}{h(X_{(n-k+1:n)})}\right)^2\right]$$

and scale factor

$$c = \frac{1}{\inf h} \max_{1\le j\le n} \frac{1}{j}\sum_{k=j}^{n} a_k\,.$$

- These results are true under the assumption h is non-decreasing
- This assumption imposes that F has finite exponential moments
- Assumption not satisfied for heavy-tailed distributions

Our goal

Can we obtain finite-sample results with large probability for PWM estimators under minimal moment conditions?

- Classical issue in concentration theory
  $\hookrightarrow$ Sub-Gaussian inequalities for the empirical mean available under the assumption of exponential moments
  $\hookrightarrow$ *Median-of-means* technique (Catoni (2012), Devroye et al. (2016) and (Lerasle, 2019))

Our contribution

Use the median-of-means methodology to design a sub-Gaussian estimator under the only assumption that X has a finite second moment.

A new *median-of-means* estimator for $\alpha_{q:p} = \mathbb{E}\left[X_{q:p}\right]$

1. Choose a probability level $\delta \in [2\exp(1 - \frac{n}{p+1}), 1[$.
2. Divide the sample $(X_1, \ldots, X_n)$ into $m = \lceil \ln(2/\delta) \rceil$ blocks (i.e. disjoint subsets) $B_1, \ldots, B_m$, each of size
$$|B_j| \geq \left\lfloor \frac{n}{m} \right\rfloor \geq p+1\,.$$
3. Within each block j, we construct the U-statistic estimator
$$\widehat{\alpha}_{q:p}^{(j)} = \frac{1}{\binom{|B_j|}{p}} \sum_{\substack{A \subset B_j \\ |A|=p}} S_q\left(X_i,\, i \in A\right)$$
4. Compute the median among blocks, i.e.
$$\widehat{\alpha}_{q:p} = \mathsf{median}\left(\widehat{\alpha}_{q:p}^{(1)}, \ldots, \widehat{\alpha}_{q:p}^{(m)}\right)$$

Result

Proposition

The *median-of-means* estimator $\widehat{\alpha}_{q:p}$ satisfies

$$\mathbb{P}\left(\left|\widehat{\alpha}_{q:p} - \alpha_{q:p}\right| \geq 2e\sqrt{\frac{2v_{q:p}\left(1+\ln(2/\delta)\right)}{n}}\right) \geq 1-\delta\,,$$

where $v_{q:p} = p\mathrm{Var}\left(X_{(q:p)}\right)$.

- Beyond the i.i.d. case
  → The result remains true when the sample $(X_1,\ldots,X_n)$ satisfies a negative dependence assumption known as conditional negative association.

Why does this work?

Sketch of proof

- Take $t = 2e\sqrt{\frac{2v_{q:p}(1+\ln(2/\delta))}{n}}$.
- By definition of the median, both the number of

$$\#\{j\colon \hat{\alpha}_{q:p}^{(j)} \geq \hat{\alpha}_{q:p}\} = \#\{j\colon \hat{\alpha}_{q:p}^{(j)} \leq \hat{\alpha}_{q:p}\} \geq m/2$$

- Let $Y_j = \mathbb{1}_{\{\hat{\alpha}_{q,p}^{(j)} - \alpha_{q,p} > t\}}$
- So

$$\mathbb{P}\left(\hat{\alpha}_{q:p} - \alpha_{q:p} > t\right) \leq \mathbb{P}\left(\sum_{j=1}^{m} Y_j \geq m/2\right),$$

- Using a Chernoff bound,

$$\mathbb{P}\left(\sum_{j=1}^{m} Y_j \geq m/2\right) \leq e^{-\sup_{\lambda \geq 0}\left\{\frac{\lambda m}{2} - \ln \mathbb{E}\left[e^{\lambda \sum Y_j}\right]\right\}}. \tag{3}$$

Link with Extreme Value Theory (EVT)

- Explicit and simple expressions can be found for classical EVT distributions (Katz et al. 2002)
- $Y_1, Y_2, \ldots, Y_n$ be i.i.d. $\sim F$
- Under certain conditions, if there exist $a_n > 0$ and b_n such that

$$\frac{\max(Y_1, \ldots, Y_n) - b_n}{a_n} \xrightarrow[n\to\infty]{\mathscr{L}} G(x)$$

with G non degenerated then G belongs to the family of Generalized Extreme Value (GEV) distributions, that is

$$G_\xi(x) = \exp(-(1+\xi x)^{-1/\xi}) \quad \xi \in \mathbb{R},\ 1+\xi x > 0\,.$$

- ξ = shape parameter
  → reflects the behavior of the tail of F
  → the larger ξ, the heavier the tail
  → $\mathbb{E}[|X|^k] = \infty$ for $k \geq 1/\xi$

Block maxima method

- Observations: $Y_1, \ldots, Y_m, Y_{m+1}, \ldots, Y_n$.
- Group data into k blocks of size m
- For each block i, compute the maxima X_i.

$$\underbrace{Y_1, \ldots, Y_m}_{X_1}, \quad \underbrace{Y_{m+1}, \ldots, Y_{2m}}_{X_2}, \quad \cdots \quad \underbrace{Y_{n(k-1)+1}, \ldots, Y_n}_{X_k}$$

- Sample of k i.i.d. maxima $(X_1, \ldots, X_k)$.
- Fit a GEV
- Note that $X_i \sim F^m$.

PWM estimator for ξ

- Classic estimator $\widehat{\xi}$ is solution of the equation

$$\frac{3^{\widehat{\xi}}-1}{2^{\widehat{\xi}}-1}=\frac{3\beta_2-\beta_0}{2\beta_1-\beta_0}$$

  → No explicit solution

- Asymptotic properties widely studied when $\xi < 1/2$ (Diebolt et al. 2008, Ferreira and de Haan, 2015)

- Finite sample results for PWM estimators seems to be sparse
  → Furrer and Naveau (2017) derived explicit variance expressions for finite samples in the case of a Generalized Pareto tail.

A new *median-of-means* estimator for ξ

- Let $\beta_{r,m} = \mathbb{E}[XF^{mr}(X)] = \frac{1}{rm+1}\alpha_{rm+1:rm+1}$
- Let

$$\xi_m = \frac{1}{\log 2}\log\frac{4\beta_{3,m} - 2\beta_{1,m}}{2\beta_{1,m} - \beta_{0,m}}$$

  $\rightarrow$ if $F = G_\xi$ then $\xi_m = \xi$.

- Estimator for ξ

$$\widehat{\xi}_m = \frac{1}{\log 2}\log\frac{4\widehat{\beta}_{3,m} - 2\widehat{\beta}_{1,m}}{2\widehat{\beta}_{1,m} - \widehat{\beta}_{0,m}}$$

  where $\widehat{\beta}_{m,r} = \widehat{\alpha}_{rm+1:rk+1}/(rm+1)$

Concentration inequality for $\widehat{\xi}_m$

With probability $\geq 1-3\delta$,

$$|\widehat{\xi}_{k,m} - \xi_m| \leq \frac{1}{\ln 2} \frac{4e(7m+3)(2^{\xi_m}+1)}{4\beta_3 - 2\beta_1 - 4e(7m+3)2^{\xi_m}\sqrt{\frac{2v_m(1+\ln(2/\delta))}{k}}} \sqrt{\frac{2v_m(1+\ln(2/\delta))}{k}}$$

where $v_m = \max(v_0, v_1, v_3)$.

- $\widehat{\xi}_{k,m}$ is sub-Gaussian
- YET, we do not have a variance bound...
- Misspecification bias: $|\xi_m - \xi|$
  $\rightarrow$ Controlled by second order regular variation assumptions, that we do not want to impose here.

Thank you for your attention!