Concentration inequalities for order statistics

Using the entropy method and Rényi's representation

Maud Thomas[1]
in collaboration with Stéphane Boucheron[1]

[1]LPMA Université Paris-Diderot

High Dimensional Probability VII
Cargèse, May 26 - 30, 2014

Background: order statistics

- Sample: $X_1, \ldots, X_n \sim_{\text{i.i.d.}} F$.

Order statistics

$X_{(1)} \geqslant \ldots \geqslant X_{(n)}$ non-increasing rearrangement of $X_1, \ldots, X_n$.

- $X_{(1)}$: sample maximum.
- $X_{(n/2)}$: sample median.
- $\forall k$, $\mathbb{P}\{X_{(k)} \leqslant t\} = \sum_{i=k}^{n} \binom{n}{i} F^i(t)(1 - F(t))^{n-i}$.
- Classical statistic theory and Extreme Value Theory provide:
  - Asymptotic distributions.
  - Convergence of moments.

Goal

derive simple, non-asymptotic variance/tail bounds for order statistics.

Background: concentration

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.

Example: Gaussian concentration

- X a standard Gaussian vector and $Z = f(X)$.
- Poincaré's inequality: $\mathrm{Var}[Z] \leqslant \mathbb{E}\|\nabla f\|^2$.
- Gross logarithmic Sobolev inequality: $\mathrm{Ent}[Z^2] \leqslant 2\mathbb{E}\|\nabla f\|^2$.
- Cirelson's inequality: $\mathbb{P}\{Z \geqslant \mathbb{E}Z + t\} \leqslant \exp(-t^2/(2L^2))$ if $\|\nabla f\| \leqslant L$.

Gaussian case and the Poincaré's inequality

- $f(X_1, \ldots, X_n) = X_{(k)}$
  the rank k order statistic of a sample is a simple function of n independent random variables.
- $||\nabla f|| = 1$.
- X_i are standard Gaussian.
  - Poincaré's inequality $\Rightarrow \mathrm{Var}[X_{(k)}] \leqslant 1$.
  - Extreme Value Theory $\Rightarrow \mathrm{Var}[X_{(1)}] = O(1/\log n)$.
  - Classical statistic theory $\Rightarrow \mathrm{Var}[X_{(n/2)}] = O(1/n)$.

We do not understand (clearly)

in which way order statistics are a smooth function of the sample.

Order statistics and spacings

Proposition (Boucheron, T. (2012))

For all $0 < k \leqslant n/2$

- $$\mathrm{Var}[X_{(k)}] \leqslant k\mathbb{E}\left[\left(X_{(k)} - X_{(k+1)}\right)^2\right] = k\mathbb{E}[\Delta_k{}^2]\,.$$
- *For all* $\lambda \in \mathbb{R}$,

$$\begin{aligned}
\mathrm{Ent}\left[e^{\lambda X_{(k)}}\right] &:= \lambda\mathbb{E}[X_{(k)}e^{\lambda X_{(k)}}] - \mathbb{E}[e^{\lambda X_{(k)}}]\log\mathbb{E}[e^{\lambda X_{(k)}}] \\
&\leqslant k\mathbb{E}\left[e^{\lambda X_{(k+1)}}\psi(\lambda(X_{(k)} - X_{(k+1)}))\right] \\
&= k\mathbb{E}\left[e^{\lambda X_{(k+1)}}\psi(\lambda\Delta_k)\right]
\end{aligned}$$

with $\psi(x) = 1 + (x-1)e^x$.

Remarks

- $V_k := k\Delta_k^2$ is called the Efron-Stein estimate of the variance of $X_{(k)}$.
- Without any assumption such as:
  - F belongs to the max-domain of attraction of an extreme value distribution G, *i.e*

$$\lim_{n\to+\infty} F^n(a_n x + b_n) = G(x)$$

    for every continuity point x of G.
  - $(X_{(k)})$ is a sequence of
    extreme order statistics, if k fixed, $n \to \infty$;
    central order statistics, if $k/n \to p \in (0,1)$ while, $n \to \infty$;
    intermediate order statistics, if $k/n \to 0$, $k \to \infty$.

Proof

Efron-Stein inequality (Efron, Stein (1981))

Let $f : \mathbb{R}^n \to \mathbb{R}$ be measurable, and let $Z = f(X_1, \ldots, X_n)$.
Let $Z_i = f_i(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ where $f_i : \mathbb{R}^{n-1} \to \mathbb{R}$ is an arbitrary measurable function.
Suppose Z is square-integrable, then:

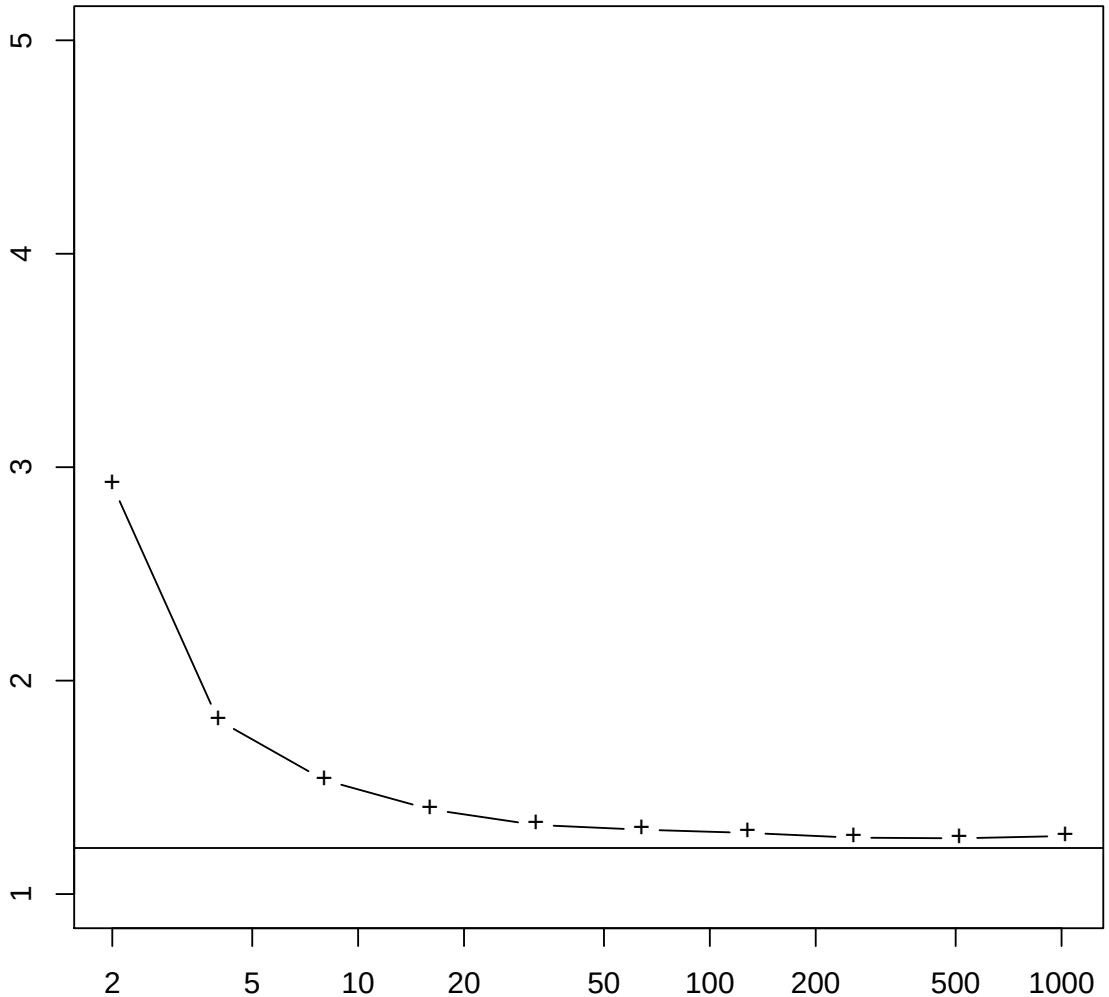
$$\mathrm{Var}[Z] \leqslant \mathbb{E}\left[\sum_{i=1}^{n} (Z - Z_i)^2\right] .$$

Modified logarithmic Sobolev inequality (Wu(2000); Massart (2000))

Let $\tau(x) = e^x - x - 1$. With the same notations, for any $\lambda \in \mathbb{R}$,

$$\mathrm{Ent}\left[e^{\lambda Z}\right] \leqslant \mathbb{E}\left[\sum_{i=1}^{n} e^{\lambda Z} \tau\left(-\lambda(Z - Z_i)\right)\right] .$$

Graphical assessment



- Ratio between the Efron-Stein estimate and the variance of the maximum of n independent Gaussian random variables.
- $n = 2^p$ for $p = 1, \ldots, 10$.
- The asymptote is the line $y = 12/\pi^2 \approx 1.22$.

Rényi's representation

- The order statistics of an exponential sample are distributed as partial sums of independent exponentially distributed random variables.

Rényi's representation (Rényi (1953))

Let $Y_{(1)} \geqslant Y_{(2)} \geqslant \ldots \geqslant Y_{(n)}$ be the order statistics of an independent sample of the standard exponential distribution, then

$$\left(Y_{(n)}, \ldots, Y_{(i)}, \ldots, Y_{(1)}\right) \sim \left(\frac{E_n}{n}, \ldots, \sum_{k=i}^{n} \frac{E_k}{k}, \ldots, \sum_{k=1}^{n} \frac{E_k}{k}\right)$$

where $E_1, \ldots, E_n$ are i.i.d standard exponential random variables.

Quantile transformation

Definition (Quantile function)

$$F^{\leftarrow}(p) = \inf\{x : F(x) \geqslant p\}, p \in (0,1) \ .$$

Notation

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

Representation for order statistics

If $Y_{(1)} \geqslant \ldots \geqslant Y_{(n)}$ are the order statistics of an exponential sample, then

$$(U \circ \exp)(Y_{(1)}) \geqslant \ldots \geqslant (U \circ \exp)(Y_{(n)})$$

are distributed as the order statistics of a sample drawn according to F.

Hazard rate, spacings and order statistics

Definition (Hazard rate)

The hazard rate h of a differentiable distribution function F is defined as:

$$h = F'/\overline{F} = F'/(1-F) \ .$$

Lemma

The distribution function F has non-decreasing hazard rate h, iff $U \circ \exp$ is concave. Indeed,

$$(U \circ \exp)' = \frac{1}{h\,(U \circ \exp)} \ .$$

- If the distribution is log-concave, then the associated hazard rate is non-decreasing.

Variance bound for order statistics when the hazard rate is non-decreasing

- Recall: $V_k = k\Delta_k^2$.

Proposition (Boucheron, T. (2012))

If F has non-decreasing hazard rate h, then for $1 \leqslant k \leqslant n/2$,

$$\mathrm{Var}\left[X_{(k)}\right] \leqslant \mathbb{E}V_k \leqslant \frac{2}{k}\mathbb{E}\left[\left(\frac{1}{h(X_{(\boldsymbol{k+1})})}\right)^2\right].$$

Towards an exponential Efron-Stein inequality

Definition (Exponential Efron-Stein inequality)

Let $Z = f(X_1, \ldots, X_n)$ where $X_1, \ldots, X_n$ are independent random variables and V its Efron-Stein estimate of the variance of Z.
Z satisfies an exponential Efron-Stein inequality if for all θ, $\lambda > 0$ such that $\lambda\theta < 1$ and $\mathbb{E}\left[e^{\lambda V/\theta}\right] < \infty$:

$$\log \mathbb{E}\left[e^{\lambda(Z-\mathbb{E}Z)}\right] \leqslant \frac{\lambda\theta}{1-\lambda\theta} \log \mathbb{E}\left[e^{\lambda V/\theta}\right] .$$

Problem

For an exponential sample, $\mathbb{E}\left[e^{\lambda V/\theta}\right] = \infty$.

↳ Find another decoupling inequality.

↳ Negative Association.

Decoupling inequality: negative association

Negative association

X and Y are negatively associated if for any non-decreasing functions f, g

$$\mathbb{E}[f(X)g(Y)] \leqslant \mathbb{E}[f(X)]\,\mathbb{E}[g(Y)] .$$

Lemma

If the distribution function F has non-decreasing hazard rate, then $X_{(k+1)}$ and $\Delta_k = X_{(k)} - X_{(k+1)}$ are negatively associated.

Exponential Efron-Stein inequality for order statistics

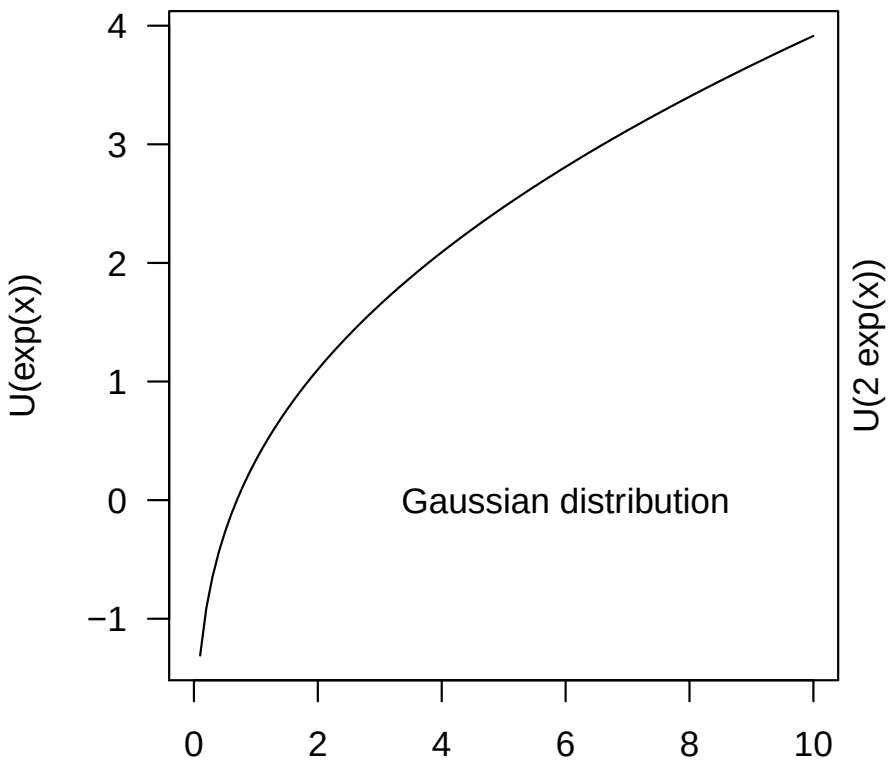
Proposition (Boucheron, T. (2012))

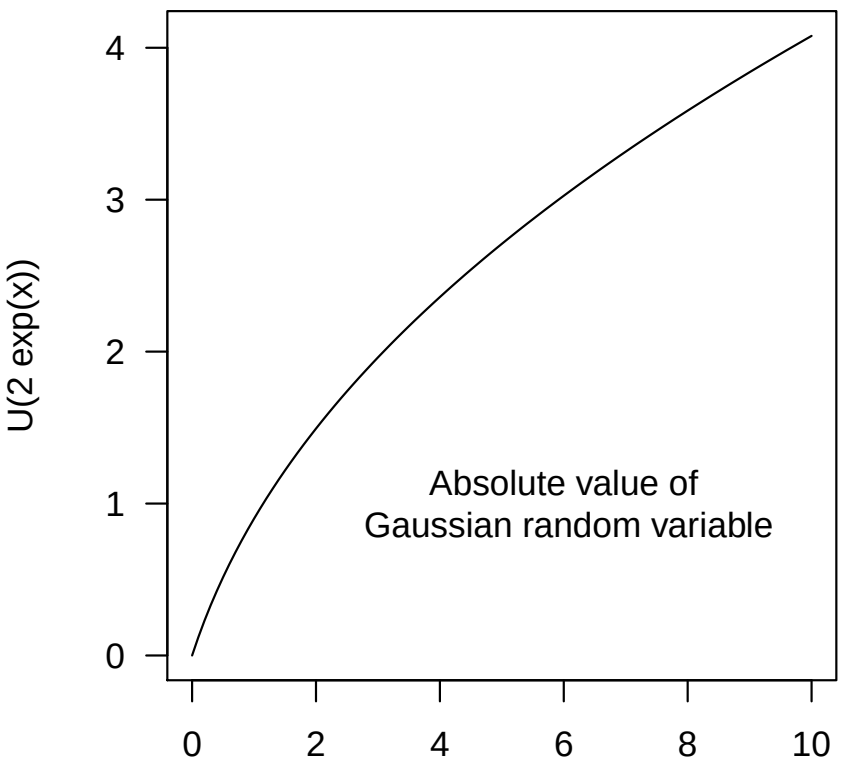
If F has non-decreasing hazard rate h, then for $\lambda \geqslant 0$, and $1 \leqslant k \leqslant n/2$,

$$\begin{aligned} \log \mathbb{E} e^{\lambda(X_{(k)} - \mathbb{E}X_{(k)})} &\leqslant \lambda \frac{k}{2} \mathbb{E}\left[\Delta_k \left(e^{\lambda \Delta_k} - 1\right)\right] \\ &= \lambda \frac{k}{2} \mathbb{E}\left[\sqrt{\frac{V_k}{k}} \left(e^{\lambda \sqrt{V_k/k}} - 1\right)\right]. \end{aligned}$$

Gaussian hazard rate

$U(t) = \Phi^{\leftarrow}(1 - 1/t)$ for $t > 1$.





Variance of absolute values of Gaussian random variables

Proposition (Boucheron, T. (2012))

Let $n \geqslant 3$, let $X_{(k)}$ be the rank k order statistic of absolute values of n standard independent Gaussian random variables,

$$\mathrm{Var}[X_{(k)}] \leqslant \frac{1}{k \log 2} \frac{8}{\log\left(\frac{2n}{k}\right) - \log(1 + \frac{4}{k} \log\log\left(\frac{2n}{k}\right))} \ .$$

- For the maximum ($k = 1$), the bound becomes:

$$\frac{1}{\log 2} \frac{8}{\log 2n - \log(1 + 4 \log\log 2n)} \ .$$

- M_n : maximum of n standard Gaussian r.v
  - Chatterjee (Talagrand L1-L2 inequality): $\mathrm{Var}[M_n] \leqslant \frac{1}{1+\log n}$.
  - Nourdin (Ornstein-Uhlenbeck process): $\mathrm{Var}[M_n] \leqslant \frac{2}{\log n}$.

Bernstein inequality for the maximum of absolute values of Gaussian random variables

Theorem (Boucheron, T. (2012))

For n such that the solution v_n of equation

$$16/x + \log(1 + 2/x + 4\log(4/x)) = \log(2n)$$

is smaller than 1,
for all $0 \leqslant \lambda < \frac{1}{\sqrt{v_n}}$,

$$\log \mathbb{E} e^{\lambda(X_{(1)} - \mathbb{E}X_{(1)})} \leqslant \frac{v_n \lambda^2}{2(1 - \sqrt{v_n}\lambda)} \ .$$

Thank you for your attention !