On the number of maximum random vectors

A phase transition, correlation inequality and a CLT

Or Zuk

Department of Statistics and Data Science
The Hebrew University of Jerusalem

31 May, 2022

Problem formulation

- Consider n random vectors in $X_1, ..., X_n \in \mathbb{R}^k$ with i.i.d. coordinates, $X_{ij} \overset{i.i.d.}{\sim} F$ for $i = 1, .., n\,; j = 1, .., k$.

Problem formulation

- Consider n random vectors in $X_1, ..., X_n \in \mathbb{R}^k$ with i.i.d. coordinates, $X_{ij} \overset{i.i.d.}{\sim} F$ for $i = 1, .., n \,; j = 1, .., k$.
- We say that a vector X_i dominates a vector X_j, if it is not smaller than X_j in all coordinates, i.e. $X_i \succeq X_j$ if $X_{il} \geq X_{jl}$, $\forall l = 1, .., k$.

Problem formulation

- Consider n random vectors in $X_1, ..., X_n \in \mathbb{R}^k$ with i.i.d. coordinates, $X_{ij} \stackrel{i.i.d.}{\sim} F$ for $i = 1, .., n\,; j = 1, .., k$.
- We say that a vector X_i dominates a vector X_j, if it is not smaller than X_j in all coordinates, i.e. $X_i \succeq X_j$ if $X_{il} \geq X_{jl}$, $\forall l = 1, .., k$.
- We say that a vector X_i is a maximum, if no other vector X_j dominates it. Let $\mathcal{M}_{k,n} \subset [n] \equiv \{1, .., n\}$ be the index set of the maximum vectors, and let $M_{k,n} = |\mathcal{M}_{k,n}|$ be the *number* of maximum vectors.

Problem formulation

- Consider n random vectors in $X_1, ..., X_n \in \mathbb{R}^k$ with i.i.d. coordinates, $X_{ij} \overset{i.i.d.}{\sim} F$ for $i = 1, .., n\,; j = 1, .., k$.
- We say that a vector X_i dominates a vector X_j, if it is not smaller than X_j in all coordinates, i.e. $X_i \succeq X_j$ if $X_{il} \geq X_{jl}$, $\forall l = 1, .., k$.
- We say that a vector X_i is a maximum, if no other vector X_j dominates it. Let $\mathcal{M}_{k,n} \subset [n] \equiv \{1, .., n\}$ be the index set of the maximum vectors, and let $M_{k,n} = |\mathcal{M}_{k,n}|$ be the *number* of maximum vectors.
- **Qu:** What can we say about the distribution of $M_{k,n}$? (moments, bounds, asymptotic results ..)

- The problem was studied extensively for *continuous* F.
  W.l.o.g. we can assume that $F = U[0,1]$.
- Define the normalized expectation: $p_{k,n} \equiv \frac{E[M_{k,n}]}{n} = P(1 \in \mathcal{M}_{k,n})$.
- A combinatorial result for the expectation: (see e.g. [BDHT05]):

$$p_{k,n} = \sum_{u=1}^{n} \binom{n-1}{u-1} \frac{(-1)^{u-1}}{u^k} .$$

- A recurrence relation:

$$p_{1,n} = \frac{1}{n} \quad ; \quad p_{k,n} = \frac{1}{n} \sum_{u=1}^{n} p_{k-1,u}, \forall k > 1.$$

  Hence, $\forall k > 1$: $p_{k,n} = \frac{1}{n} \sum_{u \in \mathcal{U}_{k,n}} \frac{1}{u_1 u_2 \ldots u_{k-1}}$, where

$$\mathcal{U}_{k,n} \equiv \left\{ u = (u_1, \ldots, u_{k-1}) \in \mathbb{Z}^{k-1} \; ; \; 1 \leq u_1 \leq u_2 \leq \ldots \leq u_{k-1} \leq n \right\} .$$

- Asymptotics for fixed k, as $n \to \infty$ (see, e.g., [BNS66]):

$$p_{k,n} \sim \frac{\log^{k-1}(n)}{n(k-1)!} \quad \text{as} \quad n \to \infty .$$

  (Higher-order terms are also available, yielding an asymptotic expansion)

- $V_{k,n} \equiv Var(M_{k,n})$
- An approximate combinatorial formula for the variance is given in [BCHL98].
- An asymptotic result for the variance is also available, including asymptotic independence of the events $\{1 \in \mathcal{M}_{k,n}\}, \{2 \in \mathcal{M}_{k,n}\}$
- Asymptotic Normality of $M_{k,n}$ was established in [BDHT05].

1. Introduction
2. Related Literature
3. **The Distribution F**
   - A Phase Transition when $k, n \to \infty$
4. A Correlation Inequality
5. References

Weak vs. Strong Maxima

- We say that a vector X_i *strongly* dominates a vector X_j (denoted $X_i \succ X_j$), if X_i dominates X_j ($X_i \succeq X_j$), and in addition $\exists l \in [k]$ such that $X_{il} > X_{jl}$.
- We say that a vector X_i is a *weak* maximum, if no other vector X_j strongly dominates it. The previous definition refers to *weak* dominance and a *strong* maximum.
- Let $\mathcal{S}_{k,n} \subset \{1, .., n\}$ be the index set of the *strong* maximum vectors, and let $S_{k,n} = |\mathcal{S}_{k,n}|$ be the *number* of maximum vectors.
- We denote $q_{k,n} = \frac{E[S_{k,n}]}{n}$. For the continuous case, $q_{k,n} = p_{k,n}$.
  For general F: $q^{(F)}_{k,n} \leq p^{(F)}_{k,n}$ and the inequality may be strict.

Example: Binary F

General F functions are of interest for two reasons:

1. Ties may be prevalent for discrete or mixed distributions.
2. Even if the true underlying distributions are continuous, we may have a finite tolerance $\epsilon > 0$, and will not distinguish between vectors within this tolerance.

QU: Why the binary case?

Proposition: Let $p_{k,n}^{(F)}$ be defined as above for a general F, $p_{k,n}$ for the continuous case, and $p_{k,n}^{(p)}$ for the *Bernoulli*(p) case. Then,

1. $p_{k,n}^{(F)} \le p_{k,n}$.
2. $p_{k,n}^{(p)} \le p_{k,n}^{(F)}$ for every $p \in \{1 - F(x); x \in \mathbb{R}\}$.

Continuous vs. *Bernoulli*(p) Comparison

We derived exact combinatorial results and asymptotic results for fixed k as $n \to \infty$ for $p_{k,n}$ allowing a comparison between the continuous and binary cases:

$p_{k,n}$	Exact	$n \to \infty$
Continuous	$\sum_{u=1}^{n} \binom{n-1}{u-1} \frac{(-1)^{u-1}}{u^k}$	$\sim \frac{\log^{k-1}(n)}{n(k-1)!}$
Bernoulli(p) (strong)	$\sum_{i=0}^{k} \binom{k}{i} p^i (1-p)^{k-i} \left(1-p^i\right)^{n-1}$	$\sim p^k (1-p^k)^{n-1}$
Bernoulli(p) (weak)	$\sum_{i=0}^{k} \binom{k}{i} p^i (1-p)^{k-i} \left(1-p^i+p^i(1-p)^{k-i}\right)^{n-1}$	$\sim p^k$

Comparison for fixed k and $n \to \infty$

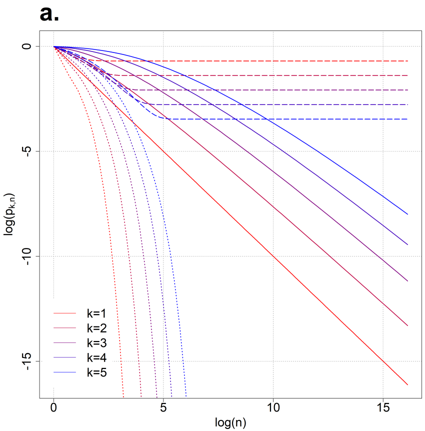

Figure 1: Value of $p_{k,n} = q_{k,n}$ (solid lines), $q^{(0.5)}_{k,n}$ (dashed lines) and $p^{(0.5)}_{k,n}$ (dotted lines) as a function of n, for $k = 1, .., 5$.

The γ functional of F

- For a distribution F, define $\gamma \equiv \gamma_F$ as follows:

$$\gamma \equiv \gamma_F \equiv -E_F \log \left[S(X)\right] \tag{1}$$

where $S(x) = P(X \geq x) = 1 - \lim_{\epsilon \searrow 0} F(x - \epsilon)$ is a (left-continuous) survival function.

- **Properties:** $\gamma \in (0, 1]$ for finite real-valued X.
  $\gamma = 1$ for any continuous F.
  $\gamma = -p \log(p)$ for the *Bernoulli*(p) distribution, maximized at $p = e^{-1}$ with the value $\gamma = e^{-1} \approx 0.368$.

Phase Transition Theorem

- **Theorem** 1: Let $k_1, k_2, \ldots$ be a sequence of positive integers

  (a) If

$$\liminf_{n\to\infty} \frac{k_n}{\log(n)} > \gamma^{-1},$$

  then

$$\mathbf{1}_{\mathcal{M}_{k_n,n}}(1) \xrightarrow{n\to\infty} 1 \quad , \quad P\text{-a.s.}$$

  (b) If

$$\limsup_{n\to\infty} \frac{k_n}{\log(n)} < \gamma^{-1},$$

  then

$$\mathbf{1}_{\mathcal{M}_{k_n,n}}(1) \xrightarrow{n\to\infty} 0 \quad , \quad P\text{-a.s.}$$

Remark: A related result was obtained in [Hwa04] using analytic technique.
Our proof uses probabilistic arguments (in particular an extreme-value Theorem from [Fer93]).

Asymptotics for fixed k, $n \to \infty$

Numeric results for the continuous case are consistent with the phase transition.

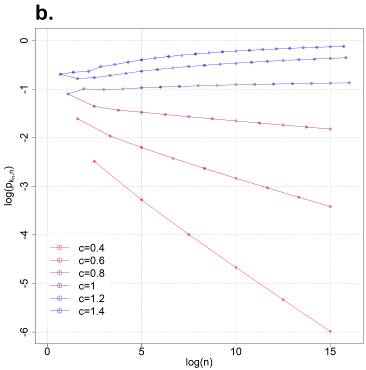

Figure 2: Value of $\log(p_{k_n,n})$ for the continuous case computed using the exact combinatorial formula (line-connected circles) for $k_n = \lfloor (c \log(n) \rfloor$ for n from 1 to 10^7 and k_n up to $\lfloor (c \log(10^7) \rfloor$ for each c.

An exact combinatorial formula for the variance

- Let $e_{k,n} \equiv= P(1, 2 \in \mathcal{M}_{k,n})$
- **For the continuous case:**

$$V_{k,n} = np_{k,n}(1 - p_{k,n}) + n(n-1)[e_{k,n} - {p_{k,n}}^2],$$

with

$$e_{k,n} = \sum_{\substack{a,b,c,d \in \mathbb{Z}_+: \\ a+b+c+d=n-2}} (-1)^{a+b} \binom{n-2}{a\ b\ c\ d} \frac{(a+b+2c+2)^k - (b+c+1)^k - (a+c+1)^k}{(a+c+1)^k (b+c+1)^k (a+b+c+2)^k}.$$

- **For the binary case:** replace above $p_{k,n}$ by $p^{(p)}_{k,n}$ and $e_{k,n}$ by $e^{(p)}_{k,n}$, with

$$e^{(p)}_{k,n} = \sum_{\substack{a,d \geq 0\ ;\ b,c \geq 1: \\ a+b+c+d=k}} \binom{k}{a\ b\ c\ d} \left[1 - p^d(p^b + p^c - p^{b+c})\right]^{n-2}.$$

- **Test your intuition:** Are the events $\{i \in \mathcal{M}_{k,n}\}$ positively or negatively correlated? (i.e. what is the sign of $e_{k,n} - {p_{k,n}}^2$?)

The Correlations' Sign

Answer: It depends! (on k and n)

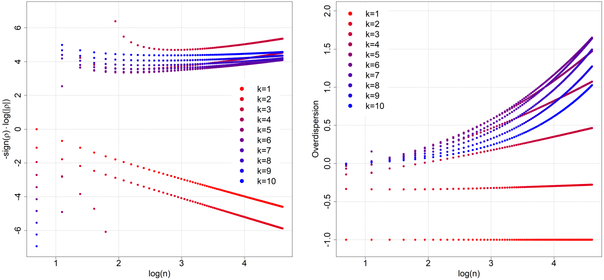

Figure 3: **Left:** $-\text{sgn}(\rho_{k,n}) \cdot \log(|\rho_{k,n}|)$ for $k = 1, .., 10$ and $n = 2, 3, .., 100$. Positive (negative) values corresponding to positive (negative) correlations. **Right:** The over-dispersion $\frac{Var(Z_{k,n})}{np_{k,n}(1-p_{k,n})} - 1 = (n-1)\rho_{k.n}$ for the same values of k and n.

Manuscript: `https://arxiv.org/abs/2112.15534` **[JZ21]**

The XYZ Inequality

- **Theorem:** (XYZ-inequality, [She82]) Let $X_i \overset{i.i.d.}{\sim} U[0,1]$, $i = 1, .., n \geq 3$. Let $E_{ij} \equiv \{X_i < X_j\}, \forall i \neq j$. Let $T \subset [n] \times [n]$ be a set of (ordered) pairs (i,j) and define the event $\Gamma \equiv \Gamma_T = \bigcap_{(i,j) \in T} E_{ij}$. Then:

$$P(E_{12}|\Gamma) \leq P(E_{12}|\Gamma, E_{13}).$$

- While intuitive, many natural generalizations fail and there are known counter-examples. For example, the above inequality may not hold when conditioning further on E_{43}, and there are known counter-examples satisfying:

$$P(E_{12}|\Gamma) > P(E_{12}|\Gamma, E_{13} \cap E_{43}).$$

Similarly, if we replace E_{12} by intersection of events like $E_{12} \cap E_{14}$.

A correlation inequality for random variables in a matrix

- **Theorem:** Let $X_{ij} \overset{i.i.d.}{\sim} F$ be continuous random variables.
Let $V_{ij} \equiv \{X_i \prec X_j\} = \bigcap_{l=1}^{k} \{X_{il} < X_{jl}\}$. Then:

$$P\left(\bigcap_{j=3}^{n} \overline{V_{2j}} \middle| \bigcap_{j=3}^{n} \overline{V_{1j}}\right) \leq P\left(\bigcap_{j=3}^{n} \overline{V_{2j}} \middle| \bigcap_{j=3}^{n} V_{j1}\right). \tag{2}$$

- **Remark:** The matrix structure of the X_{ij} in the inequality above is quite specific, and is used in the proof.
An open problem: For $X_i \overset{i.i.d.}{\sim} U[0,1]$, for what families $\{F_t, G_s \in [n] \times [n]\}$ can we generalize the inequality to:

$$P(\bigcap_{t=1}^{T} \bigcup_{(i,j)\in F_t} E_{ij} | \bigcap_{s=1}^{S} \bigcup_{(i,j)\in G_s} E_{ij}) \leq P(\bigcap_{t=1}^{T} \bigcup_{(i,j)\in F_t} E_{ij} | \bigcap_{s=1}^{S} \bigcap_{(i,j)\in G_s} E_{ij}).$$

Asymptotic Independence

The correlation inequality yields the following asymptotic independence result:

Theorem: If $k > 1$, then

$$e_{k,n} \sim p_{k,n}^2 \sim \left[\frac{\log^{k-1}(n)}{n(k-1)!}\right]^2 \quad \text{as} \quad n \to \infty$$

and hence

$$\rho_{k,n} \equiv \mathrm{Corr}(Z_1, Z_2) = o\left[\frac{\log^{k-1}(n)}{n}\right] \quad \text{as} \quad n \to \infty\,.$$

where $Z_i \equiv Z_i^{k,n}$ is defined as the indicator r.v. of the event $\{i \in \mathcal{M}_{k,n}\}$.
This can be used further to prove a CLT for weakly correlated triangular arrays.

A Central Limit Theorem for Partial Sums

1. For any $k \in \mathbb{N}$,

$$\frac{1}{np_{k,n}} \sum_{i=1}^{n} Z_i^{k,n} \to 1 \quad \text{as} \quad n \to \infty \quad \text{in} \quad L^2(P).$$

2. Let $(m_n)_{n=1}^{\infty}$ be a sequence of positive integers such that $\lim_{n\to\infty} \frac{n}{m_n} = \alpha \in (0,1)$ and assume that $k > 1$. In addition, for any $n \geq 1$ and $1 \leq i \leq m_n$ denote $U_{ni} \equiv \frac{Z_{ni} - p_{k,m_n}}{\sqrt{p_{k,m_n}(1-p_{k,m_n})}}$, where $Z_{ni} \equiv Z_i^{k,m_n}$. Then,

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} U_{ni} - \frac{1}{m_n}\sum_{i=1}^{m_n} U_{ni}\right) \xrightarrow{d} \mathcal{N}(0, 1-\alpha) \quad \text{as } n \to \infty.$$

3. $\forall k > 1, \exists (m_n)_{n=1}^{\infty}$ such that $n \ll m_n \ll n \log^k n$ and:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_{ni} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{as } n \to \infty. \tag{3}$$

Manuscript: [JZ22]

[BCHL98] Zhi-Dong Bai, Chern-Ching Chao, Hsien-Kuei Hwang, and Wen-Qi Liang.
On the variance of the number of maxima in random vectors and its applications.
The Annals of Applied Probability, 8(3):886–895, 1998.

[BDHT05] Zhi-Dong Bai, Luc Devroye, Hsien-Kuei Hwang, and Tsung-Hsi Tsai.
Maxima in hypercubes.
Random Structures & Algorithms, 27(3):290–309, 2005.

[BNS66] Ole Barndorff-Nielsen and Milton Sobel.
On the distribution of the number of admissible points in a vector random sample.
Theory of Probability & Its Applications, 11(2):249–269, 1966.

[Fer93] Thomas S Ferguson.
On the asymptotic distribution of max and mex.
Statistical Papers, 34(1):97–111, 1993.

[Hwa04] Hsien-Kuei Hwang.
Phase changes in random recursive structures and algorithms.
In *Probability, Finance and Insurance*, pages 82–97. World Scientific, 2004.

[JZ21] Royi Jacobovic and Or Zuk.
A phase transition for the probability of being a maximum among random vectors with general iid coordinates.
arXiv preprint arXiv:2112.15534, 2021.

[JZ22] Royi Jacobovic and Or Zuk.
A correlation inequality for random points in a hypercube and a related limit theorem.
preprint, 2022.

[She82] Larry A Shepp.
The xyz conjecture and the fkg inequality.
The Annals of Probability, pages 824–827, 1982.

Thank You