A Non Parametric test based on Extremal Process

Zaher KHRAIBANI

Seminar ECODEP

10 May 2023

Outline

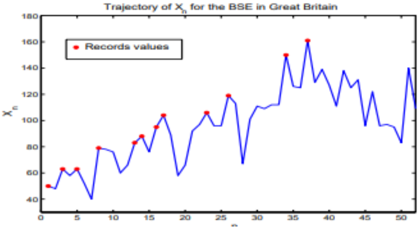
CY TECH

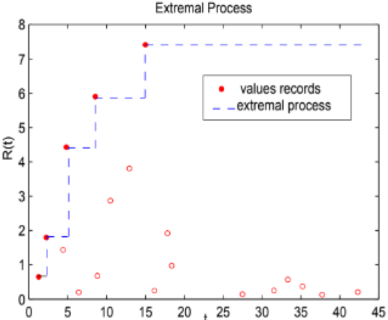
CERGY PARIS UNIVERSITÉ

Fondamental of Records Theory

1. Record theory began in 1952 (Chandler).
2. Record theory was applied in different domains (Sports, Climate Change, Economics, Hydrology, Seismology, Emidemiology, ...).
3. T denotes the current time and N_T is the number of records within the time-series $\{X_t, 1 \leq t \leq T\}$.
4. Exact distrbution for finite T versus classical Extreme Value Theory (EVT).
5. The record process represented the peak of the observed outbreak pattern of the epidemic.

EVT, Records Process, Extremal Process





Record process: Definition

- $\{R_n : n \geq 1\}$ and $\{L_n : n \geq 1\}$ are respectively the sequence of the record values and the record indices:

$$\begin{aligned} L_1 &= 1 \\ L_n &= \inf\{j > L_{n-1} : X_j > X_{L_{n-1}}\} \\ R_n &= X_{L_n} \end{aligned}$$

- N_n: total number of records among $\{X_1, ..., X_n\}$ with $N_1 = 1$:

$$N_n = \sum_{j=1}^{n} \delta_j;$$

where δ_j (indicator of record):

$$\delta_j = \begin{cases} 1, & \text{if } X_j > max(X_1, ..., X_{j-1}) \\ 0, & \text{elsewhere} \end{cases}$$

Number of records

The principal results of Record Theory for the i.i.d. case were produced over the period 1952-1983 (see Chandler (1953), Arnold (1998) and Nevzorov (2001)):

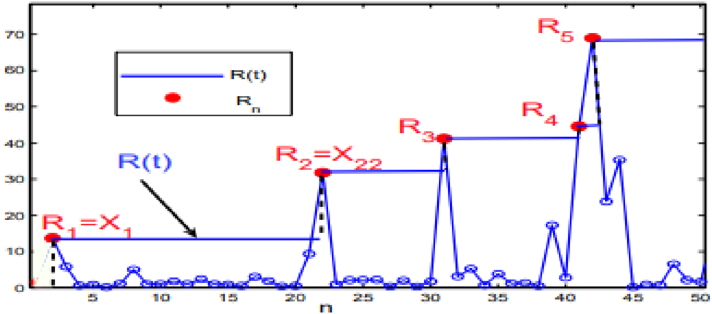
- $\{\delta_n\}_{n\geq 1}$ are independent with $\delta_n \sim \texttt{Bernoulli (1/n)}$
- The exact distribution of N_n is given by (Rényi 1962):

$$\mathbb{P}[N_n = m] = \frac{s(n,m)}{n!},\ 0 \leq m \leq n$$

$s(n,m)$: Stirling numbers of the first kind

Extremal process and records

- Extremal process: $R(t) = \{\bigvee X_k : T_k \leq t\} = \bigvee_{k=1}^{n(t)+1} X_k$; $n(t)$: number of occurrences until time t.
- $\{R(t)\} \Leftrightarrow \{\tau_n, R(\tau_n)\} = \{\tau_n, R_n\}$.
  τ_n : instant of the *nth* jump of $\{R(t)\}$, R_n: *nth* record.



Design of the Extremal Process

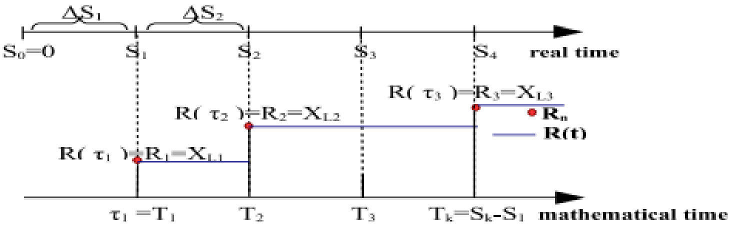


Figure 1: Extreme Process requires three steps to complete

1. $\{S_n\}_{n\geq 0}$: Occurrence time of an event (renewal process).
2. **Point Process**: (X_n, T_n); $X_n = (\Delta S_n)^{-1}$: time between two successive events.
3. $\{\Delta S_n\} \searrow \Leftrightarrow \{(\Delta S_n)^{-1}\} \nearrow$: `Emergent Event.`

Definition of the hypothesis test

1. Hypothesis test:
   H_0:$\{X_k\}$ i.i.d, $X_1 \sim F$, F continue.

   $$P(\Delta S_n \leq s) := E(s) = 1 - \exp(-\lambda s).$$

   H_1: $\{X_k\}$, are independent, $X_k \sim F_k$, where $\overline{F}_k = \overline{F}^{\rho_k}$, $\{\rho_k\}_k$ positive increasing sequence.

2. Statistic of test: N_n (number of record).

3. Error (α), Power $(1-\beta)$:

   $$\alpha = P_{H_0}(Reject\ H_0) = P_{H_0}(N_n \geq N_\alpha)$$

   $$1 - \beta = P_{H_1}(Accept\ H_1) = P_{H_1}(N_n \geq N_\alpha)$$

Distribution of N_n under H_0 et H_1

Proposition

$P_{H_0 \cup H_1}(N_n = m) = \frac{|s(n+1,m+1|\vec{u})|}{\prod_{j=1}^{n+1}(1+u_{j-1})}$ où $s(n+1, m+1|\vec{u})$ (generalized Stirling number of the first kind), $\vec{u} = (u_0, ..., u_n)$, $u_{j-1} = \frac{\sum_{k=1}^{j-1} \rho_k}{\rho_j}$, $j \geq 1$.

- Particulier case: $|s(n+1, m+1|\vec{u})| = s(n+1, m+1)$, si $\rho_k = \rho, \forall k$,
- $P_{H_0}(N_n = m) = \frac{s(n+1,m+1)}{(n+1)!}, 0 \leq m \leq n$, avec $\{s(n+1, m+1)\}$ Stirling number of the first kind.

Example

- $E(\Delta T_k) = (\lambda_k)^{-1}$, where $\lambda_k = \lambda.\rho_k = \lambda.a^k$, is the frequency of cases per unit time at time T_k , $a > 1$, the exponential growth of an infectious disease.

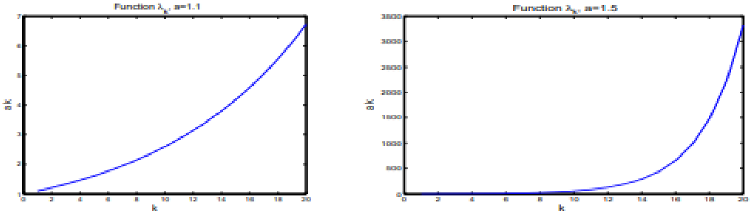


Figure 2: Fonction $\{\lambda_k\}$ where $\lambda_k = a^k$, for $a = (1.1, 1.5)$ and $\lambda = 1$

Distribition of N_n under H_0, H_1

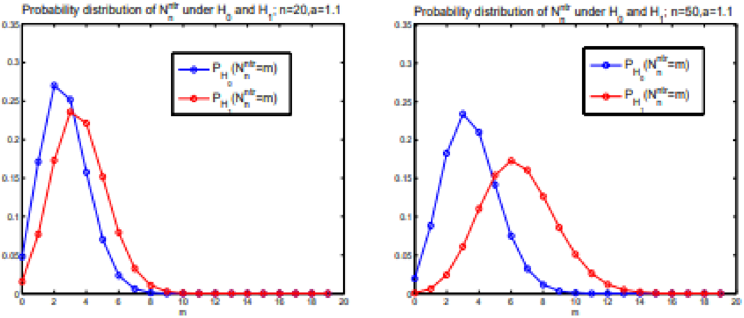
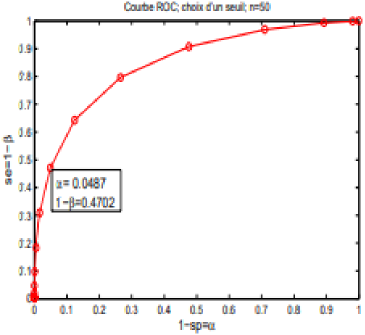
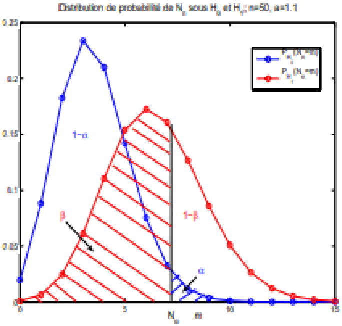


Figure 3: Distribution of N_n under H_0 and $H_1(a = 1.1)$ for $n = 20, 50$

Mode increases relatively more rapidly under H_1 than under $H_0 \Longrightarrow 1 - \beta \nearrow$.

Determination of α and $1-\beta$ ($a = 1.1$)

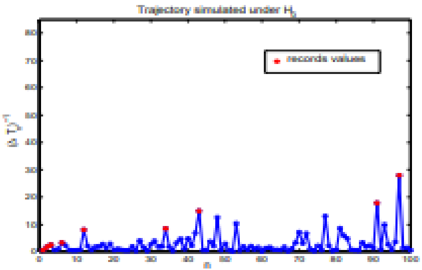


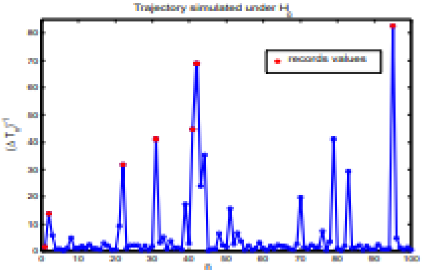


- Optimize the choice of $(\alpha, 1-\beta)$ such that they are neighbors to (0, 1).
- $1-\beta \nearrow$ with n for α given.

Test on simulated trajectories under H_0

- Under H_0: $\{(\Delta T_k)^{-1}\}_{k\leq n}$, i.i.d, $\Delta T_1 \sim exp(1)$.



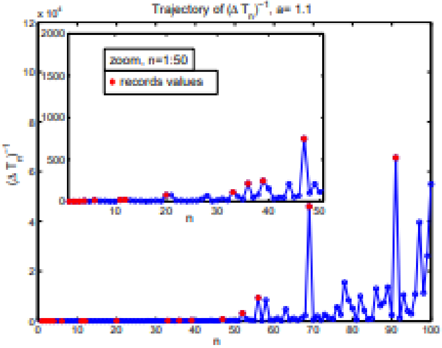


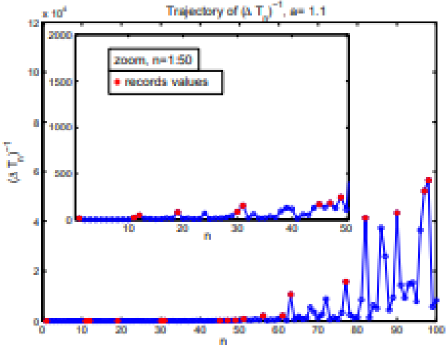
n	10	20	50	100
N_n	3	4	6	8^*

H_0 is rejected for $n = 100$ with $\alpha = 0.0489$.

Test on simulated trajectories under H_1

$$P_{H_{0,1}}(X_k \leq x) := P(\Delta T_k \geq x^{-1}) = exp(-\lambda a^k x^{-1}), a = 1.1$$





- Traj1: H_0 is rejected for $n \geq 20$
- Traj2: H_0 is rejected for $n \geq 50$

n	10	20	30	40	50	100
Traj1	4	7^{**}	$7^{*}*$	10^{***}	11^{***}	15^{***}
Traj2	0	3	4	5	8^{**}	16^{***}

Table 1: Number of observed records

n	20	30	40	50	100
N_α	6	7	8	8	9
α	0.0312	0.0199	0.0103	0.0162	0.0183
$1-\beta$	0.1259	0.156	0.1725	0.3095	0.7878

- H_1 is accepted at least from $n = 100$ because $1-\beta \nearrow$ when $n \nearrow$. ($1-\beta$ small for $n \leq 50$ due to a slow emergence)

Extremal Process: Definition

The process $\mathcal{R} : [0, \infty) \longrightarrow [0, \infty)$ is a stochastic process having the two following properties:

- The trajectories of $R(t)$ can be derived from the point process $\mathcal{N} = \{(T_k, X_k)\}_{k \geq 1}$ and its trajectories are **RCLL**.
- For $0 = t_0 < t_1 < ... < t_m$,$\exists \{\text{U}_k\}_{0 \leq k \leq m}$ non-negative such that :

$$(\mathcal{R}(0), \mathcal{R}(t_1), ..., \mathcal{R}(t_m)) \stackrel{\text{d}}{=} (\mathcal{U}_0, \mathcal{U}_0 \vee \mathcal{U}_1, ..., \mathcal{U}_0 \vee ... \vee \mathcal{U}_m).$$

$\mathcal{R}(t)$ is G*-extremal* if:

$$F_{t_1,...,t_n}(x_1, ..., x_n) = G^{t_1}(x_1).G^{t_2 - t_1}(x_2)...G^{t_n - t_{n-1}}(x_n),$$

with $G^t(x) := [G(x)]^t$ and $G(x) = P(\mathcal{R}(t) \leq x)$

- Max-increments:

$$\mathcal{U}(s,t] := \bigvee_{k=n(s)+2}^{n(t)+1} X_k = \bigvee_{T_k \in (s,t]} X_k, 0 \leq s < t.$$

- Classical Approach: $\{T_k\}, \{X_k\}$ are **independent**:

  $X_k = \Psi_k^{-1}$, where $\{\Psi_k\}$ iid, same distribution of $\{\Delta T_k\}$ but independent of $\{\Delta T_k\} \Longrightarrow \mathcal{U}(r,s]$ and $\mathcal{U}(s,t]$ are independent, $0 \leq r \leq s \leq t$.

- New Approach:$\{T_k\}, \{X_k\}$ are **dependent**:

  $\Longrightarrow \mathcal{U}(r,s]$ and $\mathcal{U}(s,t]$ are dependent.

Distribution of $\mathcal{R}(t)$

- Classical cas: $\{T_k\}$ and $\{X_k\}$ are Independent:

Proposition

$\mathcal{R}(t)$ is a generalized G-extremal process; for

$$0 < x_1 < ... < x_n, \; 0 = t_1 < t_2 < ... < t_n :$$

$$F_{t_1,...,t_n}(x_1, ... x_n) = \Phi_1(x_1) G^{t_2 - t_1}(x_2) ... G^{t_n - t_{n-1}}(x_n).$$

- $\{T_k\}$ and $\{X_k\}$ are Dependent:

Proposition

$\mathcal{R}(t)$ is a generalized extremal process:

$$P_t(x) := P(\bigvee_{k=2}^{n(t)+1} X_k \leq x) = e^{-t} \sum_{m=0}^{[xt]} \frac{t^m}{m!}\Big(1 - \frac{m}{xt}\Big)^m$$

with $P_t(0) = e^{-t}$ and $P_0(x) = 1$.

Comparison Distribution

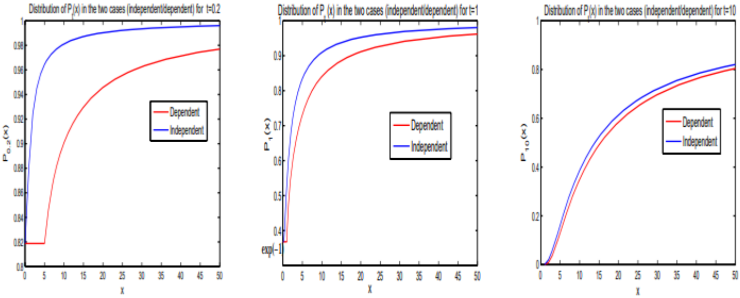


Figure 4: $\{P_t(x)\}_x$ for $t = 1$, $t = 0.2$ and $t = 10$ in the two cases (Dependent and Independent)

Hypothesis Test

Consequently, let $R_\alpha^{indep.}$ the quantile at the level α in the independent setting, that is $P^{indep.}(\mathcal{R}(t) > R_\alpha^{indep.}) = \alpha)$, and similarly in the dependent setting, then $R_\alpha^{indep.} < R_\alpha^{dep.}$, so if we do not reject H_0 in the independent setting because the observed $\mathcal{R}(t)$ is less than $R_\alpha^{indep.}$, then we will also do not reject H_0 in the dependent setting.

Definition and Notation

- $N(t) = N(0,t] = \sum_{j=2}^{n(t)+1} \delta_j$: number of nontrivial records among $\{X_1, X_2, ..., X_{n(t)+1}\}$. Or equivalently, the number of jumps of $\mathcal{R}(.)$ in $(0,t]$.

- $\{\delta_j\}_j$ are independent with $P(\delta_j = 1) = j^{-1}$

- $\{\delta_j = 1, n(t) = n\}_{j \leq n+1}$ are not independent implying that $\{N(t)\}$ and $\{N_n\}$ depend on $n(.)$.

- $N(s,t] = \sum_{j=n(s)+2}^{n(t)+1} \delta_j, 0 \leq s < t$

Record indicator distribution

Recall: $\{\delta_j\}_j$ are independent (classical case), but $\{\delta_j = 1, n(t) = n\}_{j \leq n+1}$ are not independent.

Proposition

For $j \geq 2$,

$$P(\delta_j = 1, n(t) \geq j-1) = \sum_{n \geq j-1} P(\delta_j = 1, n(t) = n)$$

$$= [-(j-1)]^{n-1} e^{-t} \sum_{n \geq j-1} \left[\sum_{l=1}^{n-1} \frac{[-t(j-1)^{-1}]^l}{l!} + \left(1 - e^{-t(j-1)^{-1}}\right) \right]$$

Lemma

$\{N_n\}$ depends on $\{n(t)\}$: $P(N_n = m | n(t) = n) := P(N_n = m)$

Distribution of $N(t)$

Proposition

The increments of $\{N(0,t]\}$ *are non-independent and non-homogenous.*

Proposition

$$
\begin{aligned}
P(N(0,t]=0) &= \int_{x>0} P_t(x)\,d\Phi_1(x) \\
&= e^{-t}\Big[1-\sum_{m\geq 1}(-m)^m \sum_{k=m+1}^{\infty}\frac{(-m^{-1}t)^k}{k!}\Big]
\end{aligned}
$$

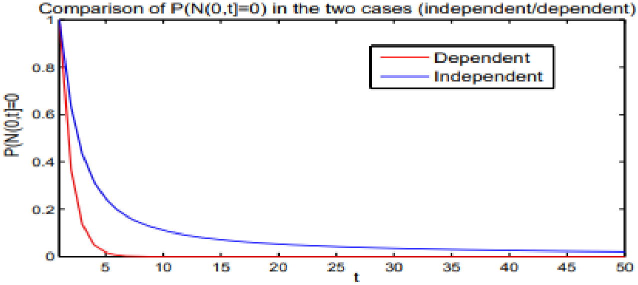


Figure 5: $\{P(N(0,t]=0)\}_t$ in the dependent and independent cases

$\{N(0,t]\}$ is greater in the dependent case than in the independent case. This result is coherent with Figure 2.

Distribution of $N(t)$

Proposition

Assume that, for $N \geq 1$ *and* $m_0 = 0 < m_1 < ... < m_N$ *,*
$A(m_1, ..., m_N) = \{\{0 < x_{m_{k-1}+1} < x_{m_{k+1}}, t_{m_{k-1}+1} + (m_k - m_{k-1} - 1)x^{-1}_{m_{k-1}+1} < t_{m_k} \leq t_{m_{k+1}} - x^{-1}_{m_{k+1}+1}, t_{m_{k+1}} = t_{m_k} + x^{-1}_{m_{k+1}}\}_{k=1}^{N}, t_1 = 0, t_{m_{N+1}} \leq t\}$*. Then*

$$P(N(0,t] = N) = \sum_{m_0=0<m_1<m_2<...<m_N} \int \cdots \int_{A(m_1,...,m_N)} P_{t-t_{m_N+1}}(x_{m_N+1}) \Pi_{k=1}^{N} [d\Phi_1(x_{m_k+1}) \times dE_1^{*(m_k - m_{k-1} - 1)}(t_{m_k} - t_{m_{k-1}+1}) \tilde{P}_{t_{m_k} - t_{m_{k-1}+1} | t_{m_k} - t_{m_{k-1}+1}}(x_{m_{k-1}+1})] \times d\Phi_1(x_1).$$

Distribution of $\mathcal{R}(t)$

Proposition

The increments $U(0,s]$ and $U(s,t]$ are independent and homogeneous.

$$
\begin{aligned}
P(U(s,t] \leq x) &:= P(\vee_{k=n(s)+2}^{n(t)+1} X_k \leq x) \\
&= \sum_{m \geq 0} \Phi_1^m(x) P(n(s,t] = m) \\
&= \exp(-(t-s)[1-\Phi_1(x)]) := G^{t-s}(x)
\end{aligned}
$$

where $G(x) = \exp(-1 + \Phi_1(x))$.

Distribution of $\mathcal{R}(t)$

Proposition

$\{R(t)\}$ *is defined by* $R(t) := \vee_{k=1}^{n(t)+1} X_k$, *which is generated by the point process* $\mathcal{N} = \{(T_k, X_k)\}$ *where the components of* N *are independent is a generalized* $G -$ *extreme process, i.e. for* $0 < x_1 < x_2 .... < x_n$ *and* $0 = t_1 < t_2 ... < t_n$, $F_{t_1 ..., t_n}(x_1, ..., x_n) = \Phi_1(x_1) G^{t_2 - t_1}(x_2) ... G^{t_n - t_{n-1}}(xn)$.

Probability distribution of $N_{n(t)}$

Proposition

The random measure $N(0,t]$ has non-independent and non-homogeneous increments.

Proposition

The probability distribution of $N_{n(t)}$ is equal to :

$$
\begin{aligned}
P(N(t)=m) := P(N_{n(t)}=m) &= \sum_{n \geq m} \frac{|s(n+1,m+1)|}{(n+1)!} P(n(t)=n) \\
&= e^{-t} \sum_{n \geq m} \frac{|s(n+1,m+1)|}{(n+1)!} \frac{t^n}{n!}
\end{aligned}
$$

Lema

$P(N(t) = 0) = E(G^t(X_1))$, *and in the case* $\overline{E}_1(t) = e^{-t}$,

$$E(G^t(X_1)) = (1 - \exp(-t))/t.$$

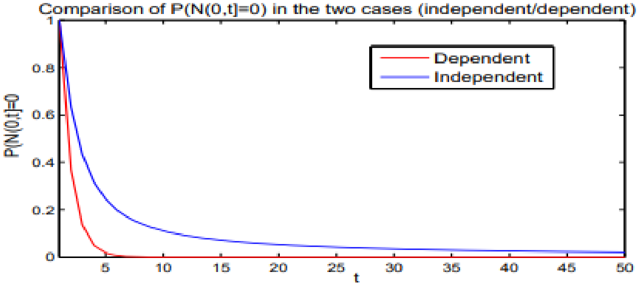


Figure 6: $\{P(N(0, t] = 0)\}_t$ in the dependent and independent cases

Statistical test

A statistical test based on the distribution of the number of observed records in an interval $(0, t]$;

$$P_{H_0}(N(t) \geq N_{t,\alpha}) = \alpha$$

could be taken. Furthermore it would be feasible to compare the test statistic $P_{H_0}(\mathcal{R}(t) \geq R_{t,\alpha})$ with $P_{H_0}(R_n \geq R_\alpha)$ and $P_{H_0}(N(t) \geq N_{t,\alpha})$ with $P_{H_0}(N_n \geq N_\alpha)$.

Characteristics of records

1. Robustness in the case of independent radom variables .
2. Exact distribution a n finite compared to the classical extreme value theory (EVT).Not unreasonable to model the X_n by the distributions (GEV).Gumbel,Weibull, Frechet
3. Nevzorov 2014, use it to construct an test detecting the outliers in a "normal" dataset. record process represents the maximum observed trend of such a phenomena.
4. Khraibani 2014, non-parametric test based on the number of observed records.