On Poisson approximation for local sequence alignments

Bojan Basrak, University of Zagreb

Self-Similarity, Long-Range Dependence and Extremes
BIRS-CMO 2018, Oaxaca, June 2018

based on the

joint work with Hrvoje Planinić

Scoring alignments

Let $A_1, \dots, A_n$ and $B_1, \dots, B_m$ be two independent iid sequences of letters (words) in finite alphabet $\mathcal{A}$ e.g. $\{A, C, T, G\}$ with distributions μ_A and μ_B respectively. For purpose of this lecture assume $m = n$.

Score function $s : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ measures similarity of letters:

$$\text{positive } s \ \leftrightarrow \text{ similar letters}\,.$$

For independent sequences we assume

$$\mathbb{E}s(A, B) < 0 \quad \text{and} \quad \mathbb{P}(s(A, B) > 0) > 0\,. \tag{1}$$

...GTAA**CTGAATCGCT**TATG...
...CACGGG**CTGATTCGCT**CG...

i

...GTAA**CTGAATCGCT**TATG...
...CACGGG**CTGATTCGCT**CG...

j

...GTAA**CTGAATCGCT**TATG...
...CGGG**CTGATTCGCT**CGAA...

For any $i, j \in \mathbb{Z}_+$ one compares sequences A and B up to these two positions by calculating

$$S_{i,j} = \left(\sup_{m<i,j} \sum_{0}^{m} s(A_{i-k}, B_{j-k}) \right)_+ .$$

It turns out one can ignore the edge effects and index the sequences over $\mathbb{Z}$, i.e. set

$$S_{i,j} = \left(\sup_{m} \sum_{0}^{m} s(A_{i-k}, B_{j-k}) \right)_+ .$$

Clearly array $(S_{i,j})$ is stationary and for $Z_{i,j} = s(A_i, B_j)$ on the **diagonal Lindley recursion** applies

$$S_{i,j} = (S_{i-1,j-1} + Z_{i,j})_+ . \tag{2}$$

It is easy to show that under (1) there exists a unique $\alpha^* > 0$ such that

$$\mathbb{E}e^{\alpha^* Z} = \mathbb{E}e^{\alpha^* s(A,B)} = 1$$

Thus if $Z = s(A, B)$ has **nonarithmetic distribution**, Cramér's arguments show that scores $(S_{i,j})$ form a stationary array with asymptotically **exponential tail**, i.e. as $u \to \infty$

$$\mathbb{P}(S > u) \sim Ce^{-\alpha^* u}\,.$$

In particular as $n \to \infty$ for any x

$$n^2 \mathbb{P}(S > \log n^{2/\alpha^*} + x) \to C e^{-\alpha^* x}$$

thus one may expect that $M_n = \sup_{i,j \leq n} S_{i,j}$ under some conditions satisfies

$$M_n - \log n^{2/\alpha^*} \xrightarrow{d} G \,, \tag{3}$$

where

$$G(x) = e^{-\vartheta C e^{-\alpha^* x}} \,,$$

for some $\vartheta \in [0,1]$.

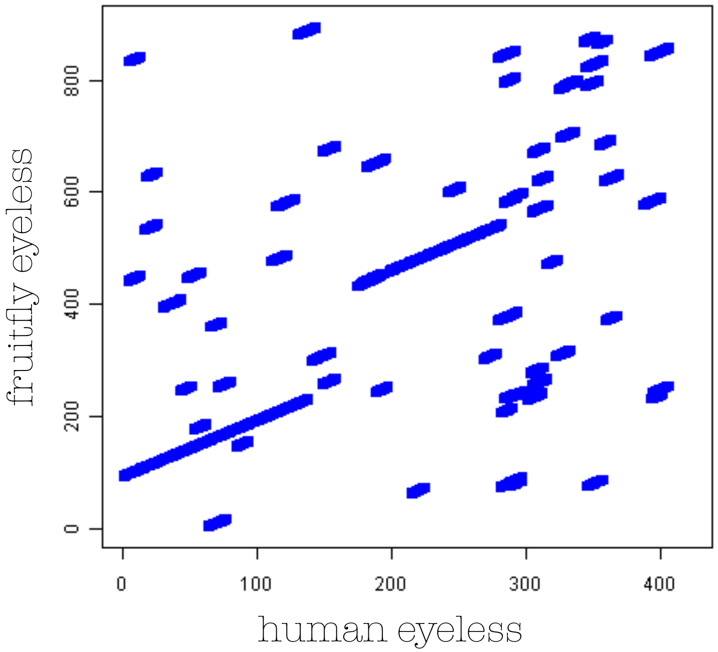
Actually, under conditions on s and μ_A, μ_B above, Dembo, Karlin and Zeitouni showed

$$M_n / \log n^{2/\alpha^*} \overset{\text{a.s.}}{\to} \gamma \leq 1 \, .$$

Under (3) $\gamma = 1$, but this is not always the case — dependence is the key.

Dependence in such an array is indeed typically weak and one can often show that (3) holds i.e. the distribution of its maxima M_n over finite rectangle $\{1, \ldots, n\} \times \{1, \ldots, n\}$ after centering tends to the **Gumbel distribution** (cf. Karlin & Altschul, Dembo et al., Arratia et al., Neuhauser, Siegmund & Yakir, Hansen,...).

Thus extreme value theory is used to test unrelatedness in evolutionary biology $\longrightarrow$ still, biologist use more than just the maximum.



cf.A.Coghlan

dot plot, cf. Metzler et al.

...GTAA**CTGAATCGCT**TATG...

...CGGG**CTGATTCGCT**CGAA...

i

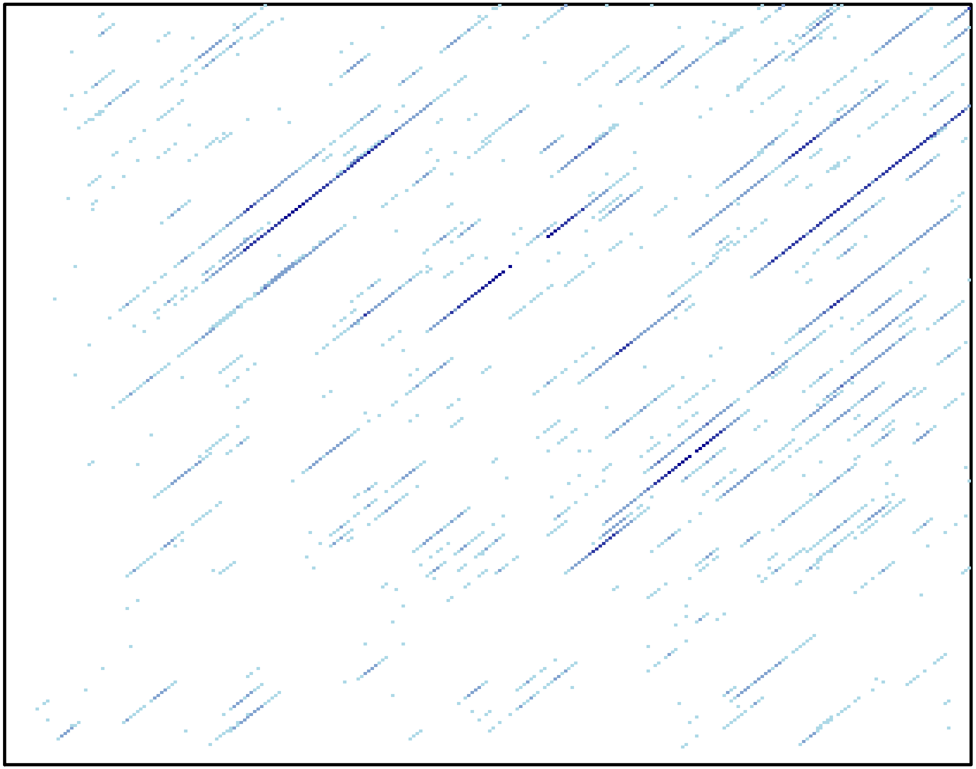
...GTAA**CTGAATCGCT**TATG...

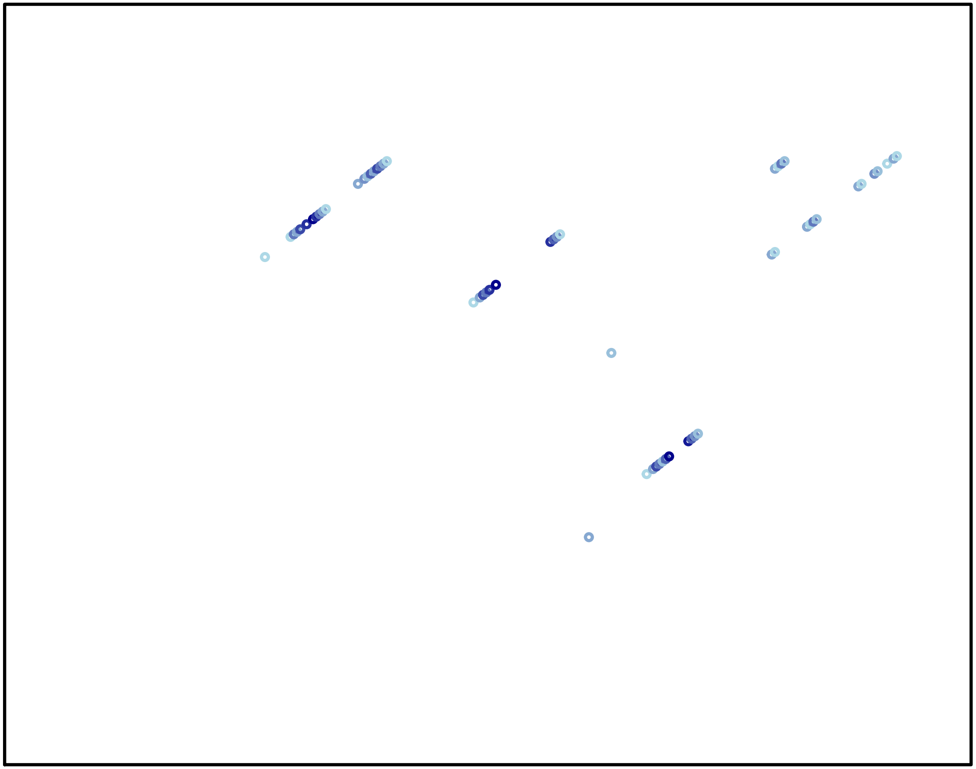
...CGGG**CTGATTCGCTC**GAA...

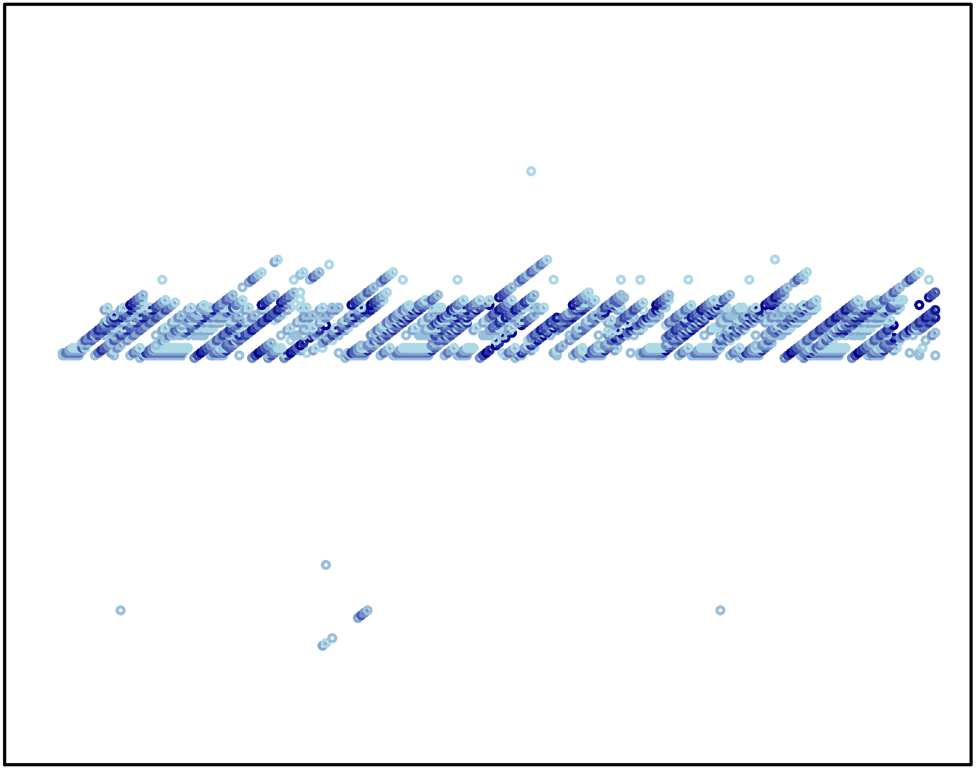
$j+1$

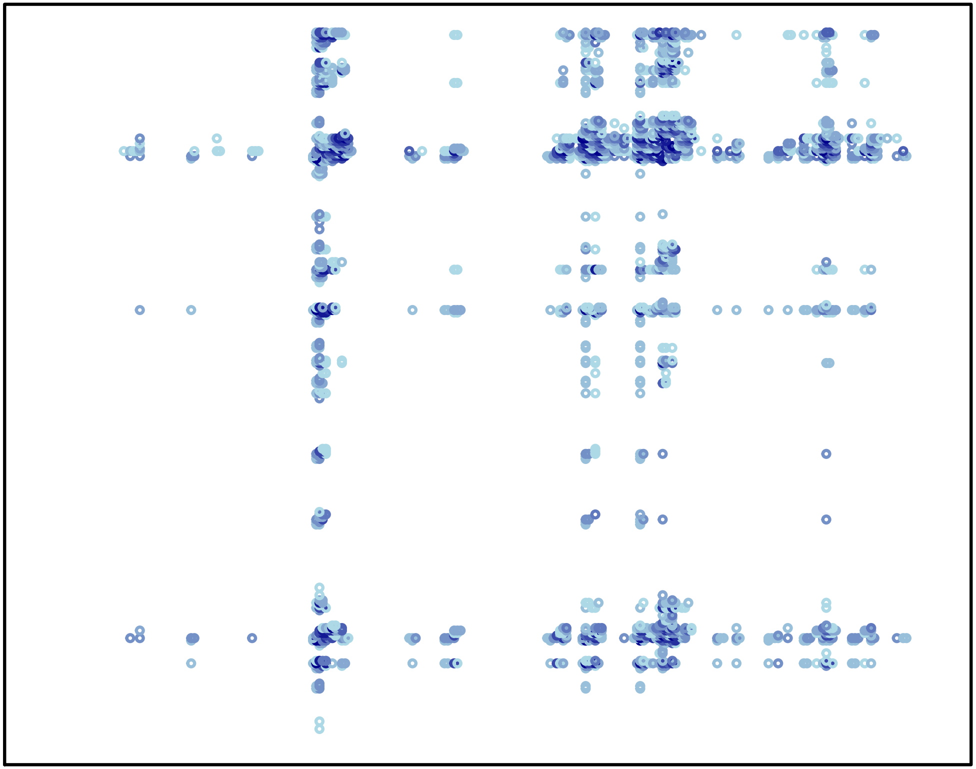
...GTAA**CTGAATCGCT**TATG...

...GGG**CTGATTCGCTC**GAAG...









Recall Kulback-Leibler **divergence** between probability measures ν and μ on finite set $\mathcal{A}_0$ is defined by

$$H(\nu \mid \mu) = \sum_{a' \in \mathcal{A}'} \log \nu(a') \frac{\nu(a')}{\mu(a')} \, .$$

On $\mathcal{A}_0 = \mathcal{A} \times \mathcal{A}$ if μ is the law of the pair (A, B) with marginals μ_A and μ_B, the quantity

$$H(\mu \mid \mu_A \times \mu_B) \geq 0$$

is called mutual information, it equals 0 iff $\mu = \mu_A \times \mu_B$.

Probability measure $\mu = \mu_A \times \mu_B$ can be tilted to

$$\mu^*(a,b) = e^{\alpha^* s(a,b)}\mu(a,b) = e^{\alpha^* s(a,b)}\mu_A(a)\mu_B(b)\,.$$

Under measure μ^*

- ▶ pairs (A_i, B_i) are still iid, but with possible mutual dependence, thus in general $\mu^* \neq \mu_A^* \times \mu_B^*$.
- ▶ tilting gives positive drift to diagonal random walk in (2).

The case $\mu^* = \mu_A^* \times \mu_B^*$ is degenerate

$$H(\mu^* \mid \mu) = H(\mu_A^* \mid \mu_A) + H(\mu_B^*|\mu_B)$$

which further equals $2\max\{H(\mu_A^* \mid \mu_A), H(\mu_B^*|\mu_B)\}$ when everything symmetric.

It turns out that we need the following (cf. Dembo et al. 1994) assumption

$$H(\mu^* \mid \mu) > 2\max\{H(\mu_A^* \mid \mu_A), H(\mu_B^*|\mu_B)\}\,. \tag{4}$$

This excludes scoring functions of the form

$$s(a,b) = s_1(a) + s_2(b)\,.$$

and degenerate distributions for A or B.

planinić, b. (2018)

Theorem Under assumptions above on μ_A, μ_B and s on $[0,1]^2 \times \mathbb{R}$

$$N_n = \sum_{i,j} \delta_{\frac{i,j}{n}, S_{i,j} - \log n^{2/\alpha^*}} \xrightarrow{d} \sum_i \sum_j \delta_{(T_i, \hat{P}_i + \hat{Q}_{i,j})} .$$

Where

▷ $\sum_i \delta_{(T_i, \hat{P}_i)}$ is PRM(Leb$\times\nu$) where $\nu(x,\infty) = \vartheta C e^{-\alpha^* x}$ and

▷ $((\hat{Q}_{i,j}))$ is a sequence of iid random elements in $\mathbb{R}^{\mathbb{Z}}$ such that $\sup_j \hat{Q}_{i,j} = 0$ a.s.

Poisson approximation

Assume that an array of independent rv's $X_{n,i}$, $i \in I_n$ is such that for all (i.e. "bounded") sets of interest K ,

$$\lim_n \max_i \mathbb{P}(X_{n,i} \in K) = 0 \quad \text{and } \sum_i \mathbb{P}(X_{n,i} \in \ \cdot\ ) \overset{v}{\to} \nu\,.$$

Then

$$\sum_{I_n} \delta_{X_{n,i}} \ \overset{d}{\to}\ N \sim \mathsf{PRM}(\nu).$$

In special cases, Chen–Stein method gives a rate of convergence as well, and handles dependence even for point process convergence on compact spaces (cf. Arratia et al. and Barbour and Brown).

Recall the convergence of point processes in distribution (wrt vague topology) is equivalent to convergence of Laplace functionals, i.e.

$$N_n = \sum_{I_n} \delta_{X_{n,i}} \ \stackrel{d}{\rightarrow} \ N$$

iff

$$\mathbb{E}\exp(-N_n f) = \mathbb{E}\exp\left(-\sum_{I_n} f(X_{n,i})\right) \ \rightarrow \ \mathbb{E}\exp(-N f)$$

for all nonneg. continuous f with support in some bounded set K.

Arrays and neighbourhoods

$$\begin{aligned} I_n = {} & \text{ sequence of index sets, e.g. } I_n = \{1, \ldots, k_n\}, k_n \to \infty \\ X_{n,i} = {} & \text{ sequence of random elements in a Polish space } \mathbb{X}' \\ B_i = B_{n,i} = {} & \text{ predetermined neighbourhoods of each } i, i \in B_i \\ \sigma(I_n \setminus B_i) = \sigma - {} & \text{ algebra generated by } X_{n,i} \notin B_i \end{aligned}$$

Consider bounded $K \subseteq \mathbb{X} \subseteq \mathbb{X}'$ and

Measures of clustering: for fixed K and f nonneg. bounded continuous with support in K, i.e. $\in \mathcal{C}_K$

$$b_1 = \sum_{i \in I_n} \sum_{j \in B_i \setminus \{i\}} \mathbb{P}(X_{n,i} \in K) \cdot \mathbb{P}(X_{n,j} \in K)$$

$$b_2 = \sum_{i \in I_n} \sum_{j \in B_i \setminus \{i\}} \mathbb{P}(X_{n,i} \in K, X_{n,j} \in K)$$

$$b_3 = \sum_{i \in I_n} \mathbb{E} \left| \mathbb{E} \left[e^{-f(X_{n,i})} \mid \sigma(I_n \setminus B_i) \right] - \mathbb{E}[e^{-f(X_{n,i})}] \right| .$$

planinić, b. (2018)

Theorem If

$$\lim_n \max_i P(X_{n,i} \in K) = 0 \quad \text{and} \sum_{i \in I_n} \mathbb{P}(X_{n,i} \in \cdot) \xrightarrow{v} \lambda$$

and for all bounded K and $f \in \mathcal{C}_K$

$$b_1,\ b_2,\ b_3 \longrightarrow 0.$$

Then

$$N_n = \sum_{I_n} \delta_{X_{n,i}} \ \xrightarrow{d}\ N\,,$$

where N is PRM(λ).

planinić, b. (2018)

Corollary Suppose $I_n = \{1, \dots, k_n\}^d$, and $X_{n,i}$ are identically distributed, if

$$k_n^d \mathbb{P}(X_{n,1} \in \cdot) \xrightarrow{v} \nu$$

and for all bounded K and $f \in \mathcal{C}_R$

$$b_1, \; b_2, \; b_3 \longrightarrow 0.$$

Then

$$N_n = \sum_{I_n} \delta_{\frac{i}{k_n}, X_{n,i}} \xrightarrow{d} N \, ,$$

where N is PRM(Leb$\times\nu$).

Regularly varying arrays

Consider a stationary array of random variables $\boldsymbol{X} = (X_{\boldsymbol{i}} : \boldsymbol{i} \in \mathbb{Z}^d)$.

We observe $\boldsymbol{X}$ over large (and increasing) section of $\mathbb{Z}^d$, for instance a square

$$\{1, \ldots, n\} \times \{1, \ldots, n\}$$

Independent observations indexed over $\mathbb{Z}$

cf. Leadbetter-Rootzén, Resnick

Take a sequence (a_n) s.t.

$$nP(X_0/a_n \in \cdot) \xrightarrow{v} \mu\,,$$

for measure μ s.t. for $x > 0$

$$\mu(-\infty, -x) = qx^{-\alpha} \quad \text{and} \quad \mu(x, \infty) = px^{-\alpha}\,.$$

Theorem For iid X_t, regular variation is equivalent to

$$N_n = \sum_1^n \delta_{\frac{i}{n}, \frac{X_i}{a_n}} \xrightarrow{d} N = \sum_i \delta_{T_i, P_i}\,,$$

where N is PRM(Leb$\times\mu$).

Stationary array $\boldsymbol{X}$ is **regularly varying** with index $\alpha > 0$ if all of its fidi's are multivariate regularly varying with index α.

Or equivalently if there exists a **tail process/array** such that as $x \rightarrow \infty$

$$\left(\frac{X_{\boldsymbol{t}}}{x}\right)_{\boldsymbol{t}\in\mathbb{Z}^d}\Bigg|\, |X_{\mathbf{0}}| > x \xrightarrow{d} (Y_{\boldsymbol{t}})_{\boldsymbol{t}\in\mathbb{Z}^d}$$

Clearly

$$|Y_{\mathbf{0}}| \sim \text{Pareto}(\alpha)\,.$$

On $\mathbb{Z}^d$ we will consider the **lexicographic order**

e.g. for $d = 2$ and $\boldsymbol{i} = (i_1, i_2), \boldsymbol{j} = (j_1, j_2) \in \mathbb{Z}^2$

$$\boldsymbol{i} \leq \boldsymbol{j} \Longleftrightarrow i_1 < j_1 \text{ or } (i_1 = j_1 \text{ and } i_2 \leq j_2).$$

For $m = 1, 2, \ldots$ introduce **rectangular section** of the array

$$\boldsymbol{X}_m = (X_{\boldsymbol{k}} : \boldsymbol{k} = (k_1, \ldots, k_d)\,, 1 \leq k_i \leq m).$$

Restricting dependence

Take $a_n \to \infty$ such that

$$n^d \mathbb{P}(X_{\mathbf{0}}/a_n \in \cdot) \xrightarrow{v} \mu.$$

And assume for some $r_n \to \infty$ and $r_n/n \to 0$

$$\lim_{m\to\infty} \limsup_{n\to\infty} \mathbb{P}\left(\bigvee_{m \le \|\boldsymbol{i}\| \le r_n} |X_{\boldsymbol{i}}| > a_n u \,\middle|\, |X_{\mathbf{0}}| > a_n u \right) = 0\,, \quad u > 0\,. \qquad \text{(AC)}$$

(AC) implies

$$Y_{\boldsymbol{i}} \overset{\text{a.s.}}{\to} 0, \quad \text{as } \|\boldsymbol{i}\| \to \infty\,, \quad \text{and} \quad \vartheta = \mathbb{P}(\sup_{\boldsymbol{j}<0} |Y_{\boldsymbol{j}}| \le 1) > 0.$$

Dependent observations indexed over $\mathbb{Z}$

Building on Davis & Resnick, Davis & Hsing, Davis & Mikosch,... Planinić, Soulier, B.(2018), Tafro, B.(2016) & Krizmanić, Segers, B. (2012) prove convergence of point processes.

Blocks of data

Consider a block

$$\frac{\boldsymbol{X}_{r_n}}{a_n}$$

as an element of

$$\tilde{l}_0 = \{\boldsymbol{x} = (x_{\boldsymbol{i}})_{\boldsymbol{i} \in \mathbb{Z}^d} : \lim_{|\boldsymbol{i}| \to \infty} x_{\boldsymbol{i}} = 0\} / \sim$$

where we set $\boldsymbol{x} \sim \boldsymbol{y}$ if they are equal up to a shift. With sup norm $\tilde{l}_0$ is a separable complete metric space.

Limits of blocks

planinić, soulier, b.

Lemma

Under assumption (AC) as $n \to \infty$

$$\left(\frac{\boldsymbol{X}_{r_n}}{a_n} \,\middle|\, M_{r_n} > a_n\right) \Rightarrow \left(Y_{\boldsymbol{i}},\ \boldsymbol{i} \in \mathbb{Z}^d \,\middle|\, \sup_{\boldsymbol{j}<0} |Y_{\boldsymbol{j}}| \leq 1\right)$$

in $\tilde{l}_0$.

Note, conditionally on $\sup_{\boldsymbol{j}<\boldsymbol{0}} |Y_{\boldsymbol{j}}| \leq 1$, random variable $L_Y = \sup_{\boldsymbol{i}\in\mathbb{Z}^d} |Y_{\boldsymbol{i}}|$ and random cluster

$$\boldsymbol{Q} = (Q_{\boldsymbol{i}})_{\boldsymbol{i}} = (Y_{\boldsymbol{i}}/L_Y)_{\boldsymbol{i}}$$

are independent.

Then, for some homogeneous measure $\nu = \nu_{\vartheta,\alpha,\boldsymbol{Q}}$ on $\tilde{l}_0$, (AC) implies

$$k_n^d \mathbb{P}\left(\frac{\boldsymbol{X}_{r_n}}{a_n} \in \ \cdot\ \right) \xrightarrow{v} \nu\,.$$

Convergence theorem

for blocks of data

Set $k_n = \lfloor n/r_n \rfloor$ and for $\boldsymbol{i} \in K_n = \{1, \ldots, k_n\}^d$ introduce rectangular section of the array

$$\boldsymbol{X}_{n,\boldsymbol{i}} = (X_{\boldsymbol{k}} : \boldsymbol{k} \in ((\boldsymbol{i}-1)r_n, \boldsymbol{i}r_n]).$$

Consider the point process of clusters, on $[0,1]^d \times \tilde{l}_0$ defined by

$$N''_n = \sum_{\boldsymbol{i} \in K_n} \delta_{\frac{\boldsymbol{i}}{k_n}, \frac{\boldsymbol{X}_{n,\boldsymbol{i}}}{a_n}} \, .$$

planinić, b. (2018)

Theorem Suppose (AC) holds if for all bounded K and $f \in \mathcal{C}_R$

$$b_1,\ b_2,\ b_3 \longrightarrow 0.$$

Then

$$N_n'' \xrightarrow{d} N'' = \sum_i \delta_{(T_i, P_i \cdot \boldsymbol{Q}_i)} \,.$$

and N is PRM(Leb×ν).

In particular

$$\mathbb{P}\left(\frac{M_n}{a_n} \leq x\right) \xrightarrow{d} e^{-\vartheta x^{-\alpha}}$$

Probabilistic model for the dot-plots

Step 1 transformation $S_{ij} \mapsto X_{ij} = e^{S_{ij}}$ gives a stationary regularly varying field.

Step 2 (X_{ij}) satisfies (AC) except in degenerate case $s(a,b) = s_1(a) + s_2(b)$.

Step 3 Blocking observations (X_{ij}) into $r_n \times r_n$ squares is justified.

Step 4 Under (4) one can show

$$b_1,\ b_2,\ b_3 \longrightarrow 0.$$

for all bounded K in space $\tilde{l}_0$ and corresponding f's.

¡ Muchas gracias !

Approximation with m–dependent arrays

Assume there exists a sequence of m–dependent regularly varying arrays $(X_{\boldsymbol{i}}^{(m)})$ such that

$$n^d \mathbb{P}(|X^{(m)}| > a_n) \to d^{(m)} > 0 \,.$$

Then on $\tilde{l}_0 \setminus \{\mathbf{0}\}$

$$k_n^d \mathbb{P}\left(\frac{\boldsymbol{X}_{r_n}^{(m)}}{a_n} \in \ \cdot\ \right) \xrightarrow{v} \nu^{(m)} \,.$$

Assume further

(i) measures $\nu^{(m)}$ converge in vague-topology to a nonzero measure ν .

(ii) for any $u > 0$

$$\lim_{m\to\infty} \limsup_{n\to\infty} \mathbb{P}(\max_{|\boldsymbol{i}|\le n} |X_{\boldsymbol{i}}^{(m)} - X_{\boldsymbol{i}}| > a_n u) = 0 \ .$$

Theorem Under assumptions above, as $n \to \infty$,

$$N''_n \xrightarrow{d} N'' = \sum_i \delta_{(T_i, P_i \cdot \boldsymbol{Q}_i)} \,.$$

Example

Spatial moving average process, assume $\xi_{\boldsymbol{i}}$ are iid RegVar(α) and for some $(c_{\boldsymbol{t}})$

$$X_{\boldsymbol{t}} = \sum_{\mathbb{Z}^d} c_{\boldsymbol{i}} \xi_{\boldsymbol{t}-\boldsymbol{i}} \, ,$$

e.g.

$$X_{t,s} = \sum_{i,j} c_{i,j} \xi_{t-i,s-j} \, .$$

is regularly varying if $\sum_{\boldsymbol{j}\in\mathbb{Z}^2} |c_{\boldsymbol{j}}|^{\delta} < \infty$ for some $\delta < \alpha \wedge 1$ (Davis and Resnick 1985). Appropriate approximation is of course

$$X^{(m)}_{t,s} = \sum_{|i|,|j|<m/2} c_{i,j} \xi_{t-i,s-j} \, .$$

Thank you very much!