Modeling exceedances in extreme value theory:

foundations, regression, time series,

multivariate settings

Dani Gamerman

Departamento de Métodos Estatísticos - IM

Universidade Federal do Rio de Janeiro

VI COBAL, PUCP - Lima, 21 June 2019

Based on work with...

Cibele Behrens (CB)

Fernando Nascimento (FN)

Hedibert Lopes (HL)

Richard Davis (RD)

Manuele Leonelli (ML)

Content

1. Introduction

Precise knowledge and predicting capabilities for extremes are fundamental in many disciplines:

- Environmental sciences
- Finance and actuarial science
- Engineering and reliability

Standard statistical methods do not guarantee precise extrapolations towards the tail of the distribution where little, if no, data is available $\Longrightarrow$ **extreme value theory (EVT)**.

1) Block maxima

Let $X_1, \ldots, X_n$ be i.i.d and M_n their maximum.

If there exists sequences of constants $\{a_n \geq 0\}$ and $\{b_n\}$ such that

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) = G(x) \text{ and } G \text{ is non-degenerate}$$

then G is the d.f. of the generalized extreme value (GEV) distribution

$$G(x \mid \sigma, \xi) = \begin{cases} \exp\left\{-\left[1 + \xi\left(\frac{x}{\sigma}\right)\right]_+^{-\frac{1}{\xi}}\right\}, & \xi \neq 0; \\ \exp\left[-\exp\left(-\frac{x}{\sigma}\right)\right], & \xi = 0 \end{cases}$$

2) **Exceedances**

For X in the domain of attraction of the GEV distribution

$$\lim_{u \to x_F} \mathbb{P}(X > x + u \mid X > u) = 1 - G(x)$$

x_F is the upper limit of the support of X

G is the d.f. of the generalized Pareto distribution (GPD):

$$G(x \mid \sigma, \xi) = \begin{cases} 1 - \left[1 + \xi\left(\frac{x}{\sigma}\right)\right]_+^{-\frac{1}{\xi}}, & \xi \neq 0; \\ 1 - \exp\left(-\frac{x}{\sigma}\right), & \xi = 0 \end{cases}$$

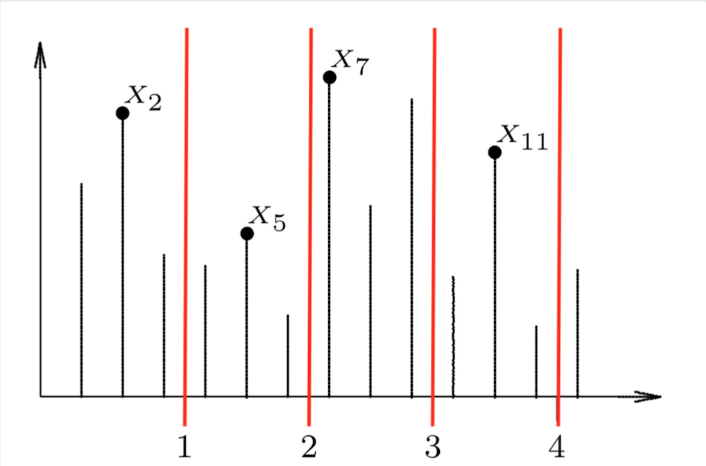
where $x > 0$, $\sigma > 0$, $[1 + \xi(\frac{x}{\sigma})] > 0$.

3 different extreme regimes:

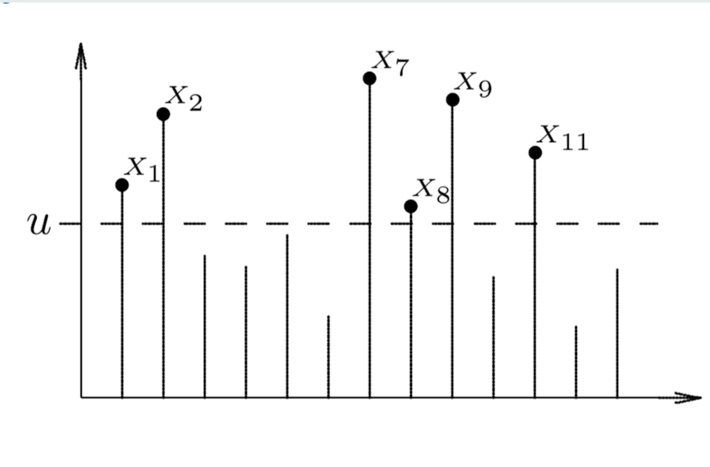
Frechet ($\xi > 0$); Gumbel ($\xi = 0$) and Weibull ($\xi < 0$; finite x_F)

Graphical representation

Block maxima



Exceedances



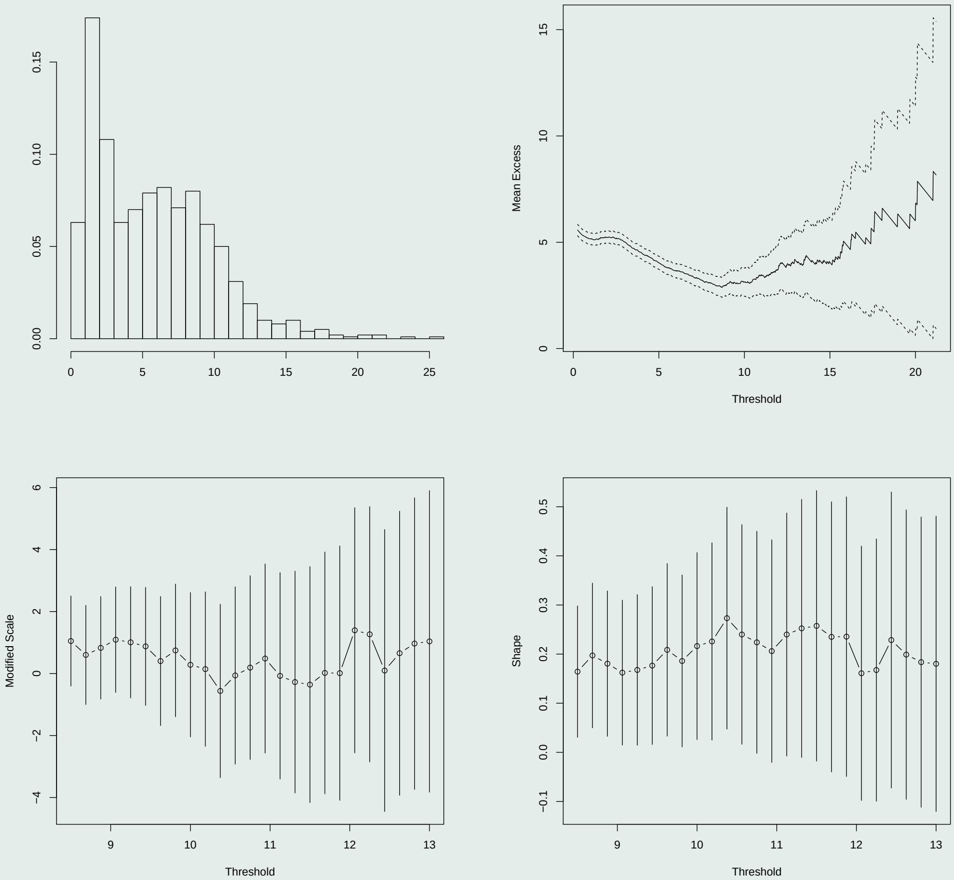
This talk concentrates on exceedances

1.2. Standard approach for inference

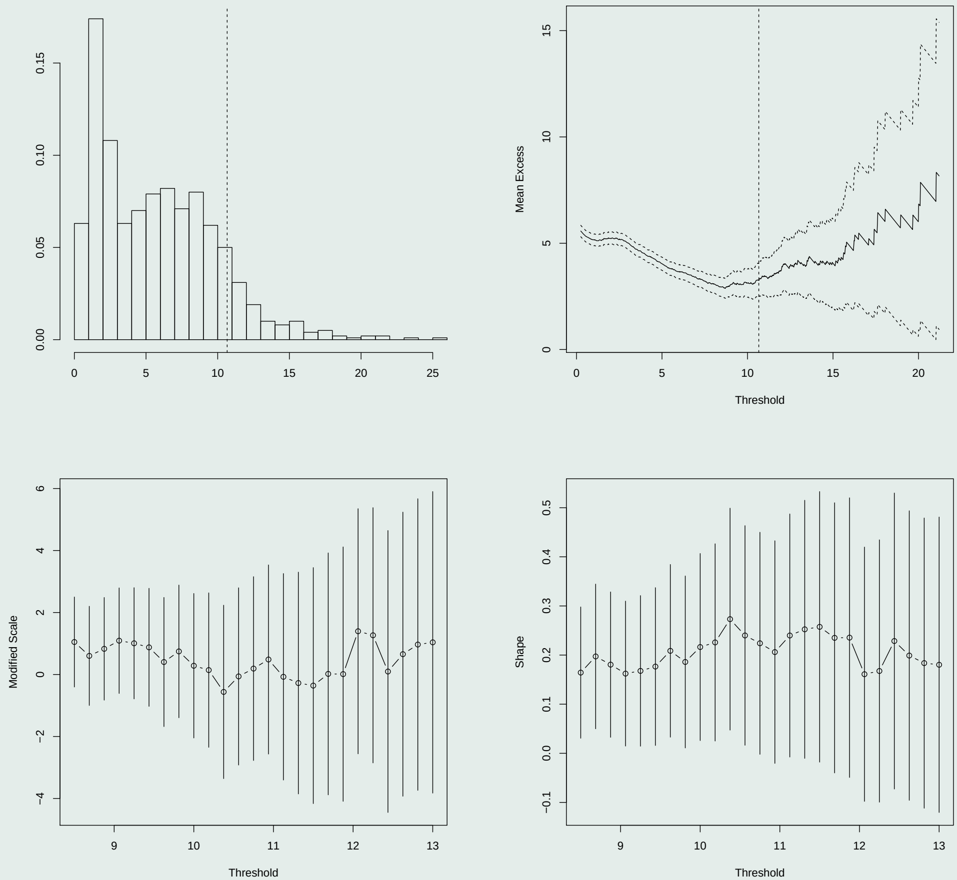
- Pre-set the threshold and use only the data beyond it to estimate GPD
- Questions: what is its value? where does tail begin?
- Pickands (1975) suggests threshold as large as possible
- Too high threshold: few data points $\rightarrow$ unreliable tail inference
- Too low threshold: too far from GPD $\rightarrow$ biased tail inference
- Graphical techniques were introduced to set the threshold

  Example: MRL plot - exceedance means increase linearly

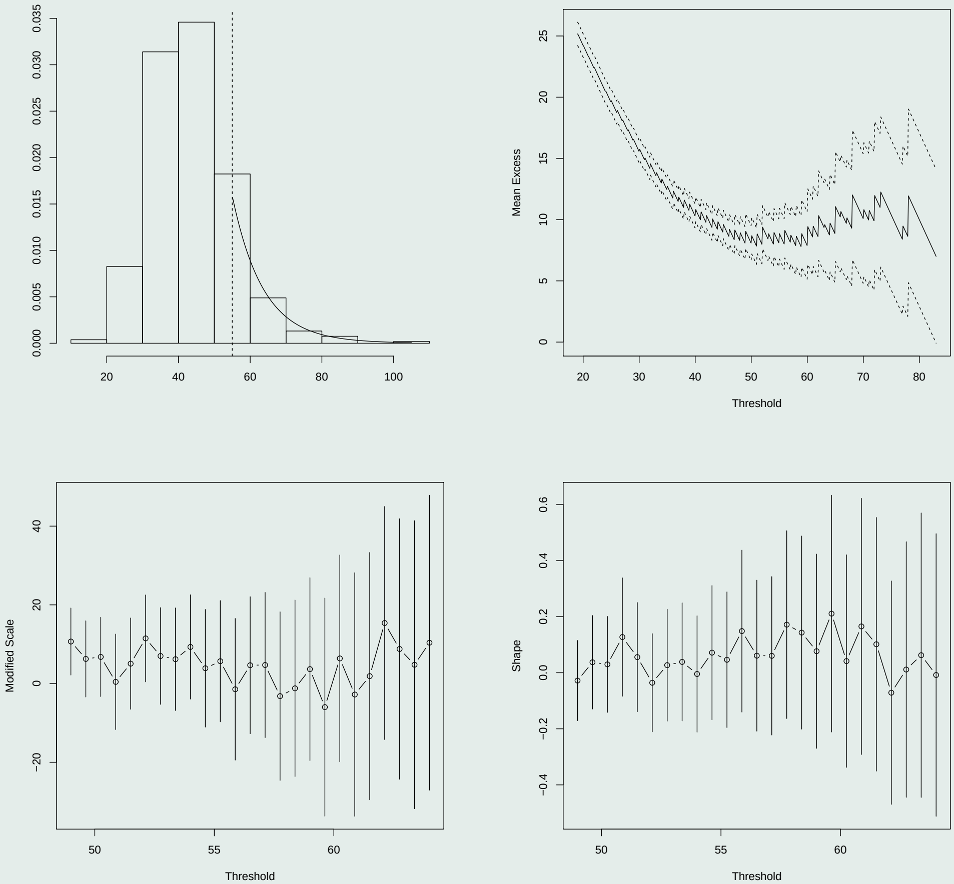
Threshold determination: simulated data



Threshold determination: simulated data



Threshold determination: Leeds NO_2 data



Alternative approaches

- Standard approach discards most of the data
- Relies heavily on graphical and unstable tools
- It makes sense to use all data instead of only extreme data
- This can be achieved in many ways but should:

  1) be as flexible as possible in the bulk (outside the tail)

  2) not pre-set threshold

A bit of history

- Frigessi et al. (2002): Mixture of Weibull for bulk and GPD for tail, with data dependent weights
- Bermudez et al. (2003): estimates bulk of the data based on the data frequency
- Tancredi et al. (2003): Mixture of uniforms for bulk and estimates number of observations in tail
- CB, HL & DG (2004): Gamma for bulk and GPD for tail. The threshold is a parameter to be estimated
- McDonald et al. (2011): mixture of normals for bulk and GPD for tail

2. Univariate model: MGPD

Introduced by FN, DG & HL (2012):

$$f(x \mid \phi, \psi) = \begin{cases} h(x \mid \phi), & x \le u \\ [1 - H(u \mid \phi)]g(x - u \mid \psi), & x > u \end{cases}$$

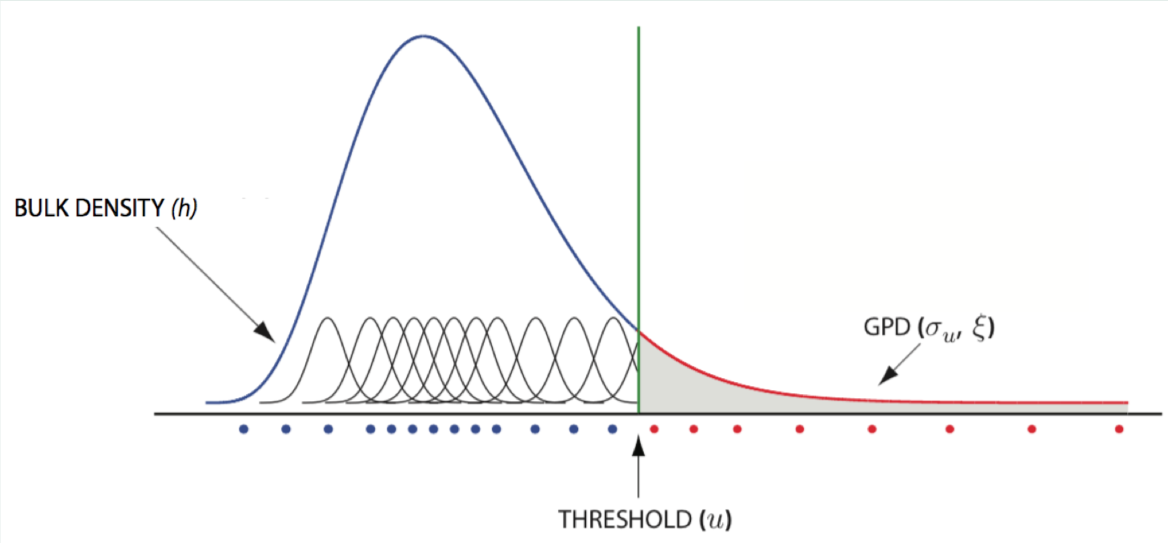
g, G : GPD density, d.f.

h, H : mixture of Gamma densities, d.f.'s (non-parametric flavour)

ϕ : Gamma parameters

ψ : GPD parameters

Graphical representation



Continuity constraints at threshold could be imposed but are not needed

Quantiles

Main interest of EVT: higher quantiles (beyond observed data)

The p-quantile q of mixture of Gammas (h) is given by

$$p = H(q \mid \phi) = \sum_{j=1}^{k} p_j \int_0^q f_{G,j}(x \mid \phi) dx.$$

The quantiles must be computed numerically

In $MGPD$ model, the higher quantiles (beyond threshold) are

$$q = \frac{((1-p^*)^{-\xi} - 1)\sigma}{\xi}, \text{ where } p^* = \frac{p - H(u \mid \phi)}{1 - H(u \mid \phi)}.$$

Inference for MGPD

Bayesian approach is used

Priors must be carefully devised: threshold and identifiability

Castellanos and Cabras (2007): reference prior for GPD parameters

Posterior distribution is way too complicated

$\rightarrow$ no analytic results can be extracted

$\rightarrow$ Block MCMC is used

Higher quantile estimation: simulation results

	u=6			u=9			u=12		
Quantile	T	$MGPD$	POT	T	$MGPD$	POT	T	$MGPD$	POT
0.99	20.06	23.13	22.07	21.56	20.48	20.21	17.55	17.77	17.11
0.999	65.21	53.19	42.68	51.49	41.44	38.06	37.30	31.59	28.54
0.99999	419.44	314.54	130.58	319.43	191.20	116.41	211.45	319.09	72.86

T-True quantile, POT- based on using DIP to determine the threshold.

Summary: $MGPD$ quantiles closer to true in 8 out of 9 simulations

Higher quantile estimation: real data results

	Espiritu Santo, Puerto Rico (in ft^3/s)		
$Prob$	E	$MGPD$	MG_k
0.95	798	793.29	842.7
0.99	1360	1426.04	1398.8
0.999	2600	2677.56	2197.0
0.9999	N/A	4612.30	3014.0
	Barcelos, Portugal (in mm)		
0.95	73.1	74.54	74.71
0.99	99.4	101.73	104.09
0.999	117.5	137.84	151.50
0.9999	143.5	171.41	233.00

$MGPD$ closer to empirical than MG_k in 6 out of 7 situations

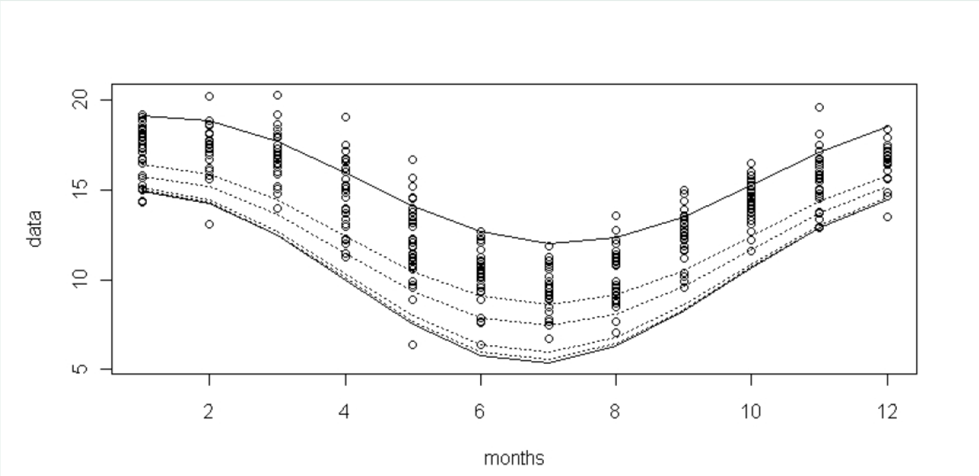
Regression (FN, DG & HL, 2011)

Auxiliary variables $(x_1, ..., x_p)$ may help explaining extreme behaviour

$\rightarrow$ regression in the form $g(u, \sigma, \xi) = x'\beta$

Cabras et al. (2011): regress x on orthogonal σ and $\nu = \sigma(1 + \xi)$

Application: monthly minima of cities in state of Rio de Janeiro



Full: minimum; Dashed: 5%, 1%, 0.01% and 0.00001% quantiles.

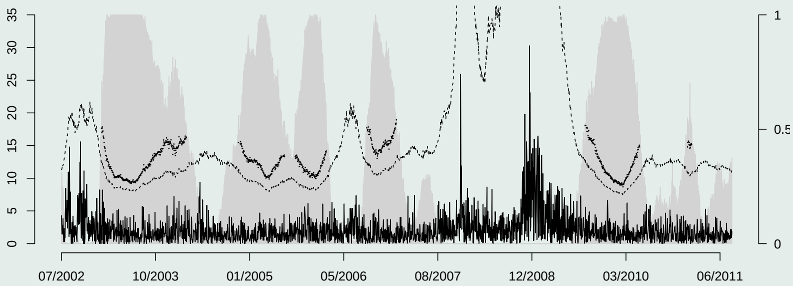
Time Series (FN, DG & HL, 2016)

EVT frequently applied to time series setting, typically not acknowledged

Possibility: $(u, \sigma, \xi) \to (u_t, \sigma_t, \xi_t)$

Our proposal: dynamic model for temporal variation of (u_t, σ_t, ξ_t)

Application: return of Petrobras stocks 2000-2014



Absolute returns, 99.9999% quantiles and maximum (if median $\xi < 0$)

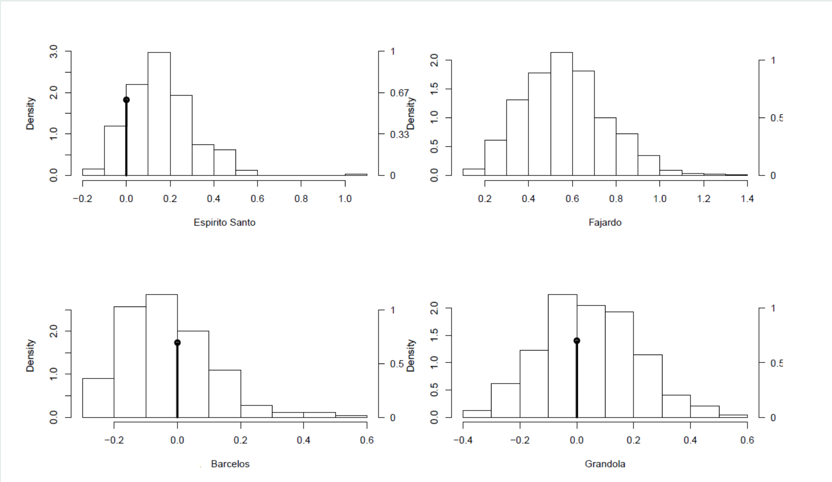
Grey area $= P(\text{finite maximum at } t \mid x) = P(\text{Weibull regime at } t \mid x), \forall t$

Regime identification (FN, DG & RD, 2016)

So far, shape ξ assumed to vary continuously

Identification of 3 regimes $\rightarrow$ probability mass at $\xi = 0$ (Gumbel)

Applications: Puerto Rico river flows and Portugal rainfalls



$P(Gumbel \mid x)$: Esp. Santo $=$ 0.61; Barcelos $=$ 0.69; Grandola $=$ 0.70

Quantiles are similar, but mixture models add regime identification

3. Multivariate extreme model (ML & DG, 2019)

Univariate setting: limiting distribution of block maxima is GEV

This distribution has known density expression.

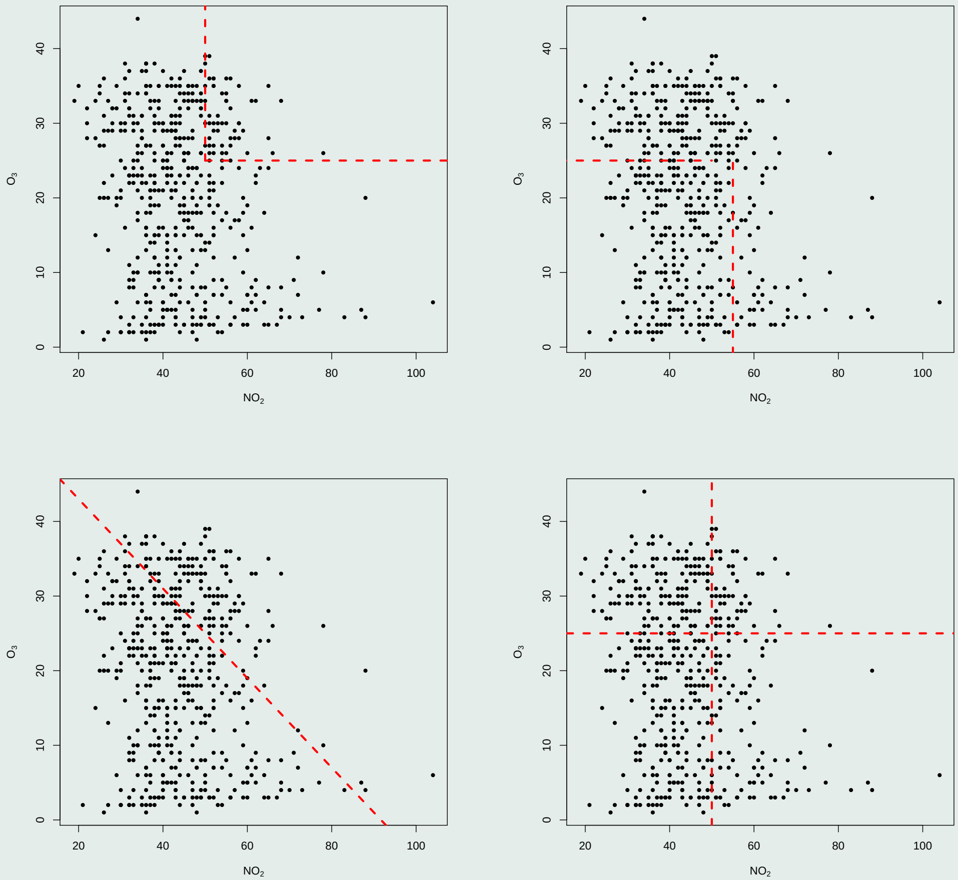
Multivariate setting: GEV requires exponent or spectral measure.

These are typically not known and a number of options were proposed

Data above threshold is assumed to be extreme and used for inference

- Parametric:
  - for the exponent measure *(simpler but less flexible) Coles and Tawn 1991, 1994; Jaruskova 2009; Joe 1990*
  - for the spectral measure *(computationally more intensive) Ballani and Schlather 2011; Boldi and Davison 2007; Cooley et al. 2010*
- Nonparametric: for the spectral measure (*Einmahl and Segers, 2009; Guillotte et al. 2011*).
- Other theoretical justifications (*Bortot et al. 2000; Heffernan and Tawn 2004; Ramos and Ledford 2009; De Carvalho and Davison, 2014; Wadsworth et al, 2017*).

Which observations are extreme?



Asymptotic independence

Coefficient of asymptotic dependence

$$\chi = \lim_{u \to 1} \chi(u) \text{ where } \chi(u) = P(F_1(X_1) > u \mid F_2(X_2) > u).$$

for $X_i \sim F_i$, for $i = 1, 2$.

$\chi = 0 \;\Rightarrow$ asymptotic independence

$\chi \in (0, 1] \;\Rightarrow$ asymptotic dependence

Example: $X_1, X_2 \sim \mathcal{N}$, cor$(X_1, X_2) = \rho \neq 0$, then

$$\lim_{u \to 1} P(F_1(X_1) > u \mid F_2(X_2) > u) = 0.$$

Thus, normal distributions are asymptotic independent

Multivariate dependence assessed via pairs of r.v.

Bivariate GEV: $\chi = 0 \quad \Leftrightarrow \quad X_1$ and X_2 are independent.

Because of this deficiency, models based on different theoretical justifications have started to appear (*Heffernan and Tawn, 2004; Ramos and Ledford, 2009*.

Coefficient of subasymptotic dependence

$$\bar{\chi} = \lim_{u \to 1} \bar{\chi}(u) \text{ where } \bar{\chi}(u) = \frac{2 \log P(F_1(X_1) > u)}{\log P(F_1(X_1) > u, F_2(X_2) > u)} - 1$$

$\bar{\chi} = 1 \;\Rightarrow$ asymptotic dependence

$\bar{\chi} \in (-1, 1) \;\Rightarrow$ asymptotic independence

Copulae

A copula C is a flexible tool to construct multivariate distributions with given margins. Let $X_1, \ldots, X_d$ be r.v.s with d.f.s $F_1, \ldots, F_d$.

A **copula** C is a function $C : [0,1]^d \to [0,1]$ s.t.

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d))$$

- Sklar's theorem guarantees there always exists one such copula;
- C is a d.f. in $[0,1]$ itself;
- separate marginal and dependence modelling.

$$f(x_1, \ldots, x_d) = c(F_1(x_1), \ldots, F_d(x_d)) f_1(x_1) \cdots f_d(x_d).$$

Elliptical copulae

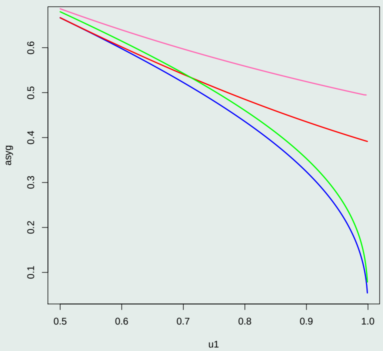
C is often (a mixture of) elliptical distributions: (skew-)normal, (skew-)T.

Asymptotic behaviour:

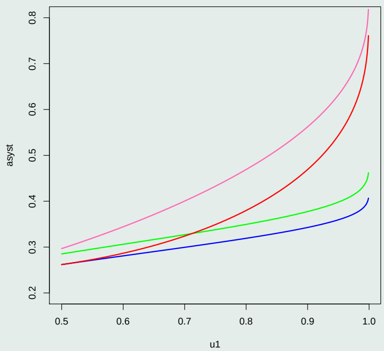
(skew-)normal - asymptotic independence ($\chi(u) \to 0$, $\bar{\chi}(u) \to (-1, 1)$)

(skew-)T - asymptotic dependence ($\chi(u) \to (0, 1)$, $\bar{\chi}(u) \to 1$)

$\chi(u)$



$\bar{\chi}(u)$



Our approach

We propose a new approach for multivariate extremes that

- marginally utilize flexible extreme mixture models - MGPD
- exploit the flexibility of copulae to model dependence
- assess extreme dependence from the chosen copula
- formally utilize all data available

Joint multivariate modelling

Mixture of elliptic copulae with MGPD margins

$$f(x \mid \cdot) = \sum_{i=1}^{r} \omega_i c_i(F_1(x_1), \ldots, F_d(x_d)) f_1(x_1) \cdots f_d(x_d),$$

where f_i is MGPD, c_i is a copula density and $\sum_{i=1}^{r} \omega_i = 1$, $\omega_i \geq 0$.

So for example if Gaussian

$$f(x \mid \cdot) = \sum_{i=1}^{r} \omega_i c_i^{\text{gauss}}(F_1(x_1), \ldots, F_d(x_d)) f_1(x_1) \cdots f_d(x_d)$$

where $c_i^{\text{gauss}}(u_1, \ldots, u_d) = \mid R_i \mid^{-1/2} \exp\left(-\frac{1}{2} y^{\mathrm{T}} (R_i^{-1} - I_d) y\right)$, with $y^{\mathrm{T}} =$ $(\Phi^{-1}(u_1)), \ldots, \Phi^{-1}(u_d))$.

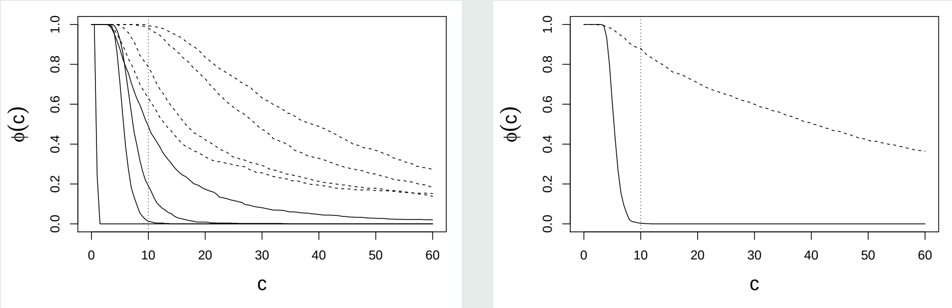
Ascertainment of asymptotic independence

Few proposals separate extreme dependence from extreme independence

Our proposal: use $\phi(c) = P(v > c \mid x)$, where $v =$ dof of T copula

Ideally, $\phi > 0.5$ indicates asymptotic independence

simulated data real data



Asymptotic (in)dependent data: solid (broken) line

$c = 10$ seems to provide a reasonable choice

Simulation study - 1000 observations, 8 models

1) Asymptotically independent models

2G - Mixture of 2 Gaussian copulae with MGPD margins

SN - Skew Normal copula with MGPD margins

MO - Morgenstern copula with lognormal-GPD margins

BL - Bilogistic copula with lognormal margins

2) Asymptotically dependent models

2T - Mixture of 2 T-copulae with MGPD margins

SN - Skew-T copula with MGPD margins

AL - Asymmetric logistic copula with lognormal-GPD margins

CA - Cauchy copula with lognormal margins

Summary of estimation: asymptotic independent data

	2G	SN	MO	BL
d.o.f.	16.5 (5.8,141.5)	28.9 (10.2,135.8)	38.9 (13.0,154.3)	13.0 (4.0,157.9)
ϕ	0.787	0.983	0.995	0.631
δ_{95}	0.42 (0.31,0.53)	0.38 (0.27,0.49)	0.36 (0.21,0.51)	0.18 (0,0.65)

- number of dof large, as expected with asymptotic independent data

- δ_{95} - asymptotic indicator (Huser & Wadsworth, 2018), threshold 0.95

  $\delta > (<) 0.5 \rightarrow$ asymptotic (in)dependence

  choice of threshold values did not matter here

- ϕ seems to behave well wrt δ

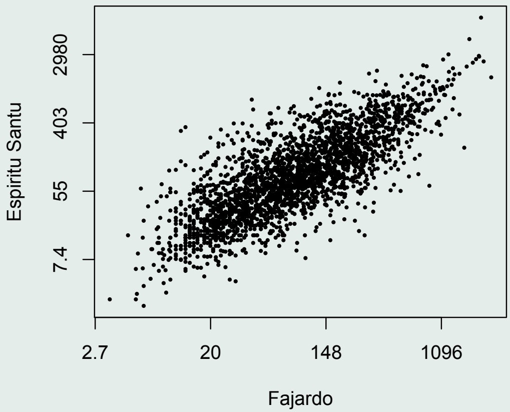
Summary of estimation: asymptotic dependent data

	2T	ST	AL	CA
d.o.f.	9.8 (3.6,51.9)	5.6 (3.9,9.3)	7.3 (4.4,16.0)	0.9 (0.8,1.1)
ϕ	0.490	0.013	0.191	0
δ_{95}	0.48 (0.40,0.57)	0.48 (0.42,0.55)	0.13 (0,0.66)	0.60 (0.53,0.69)

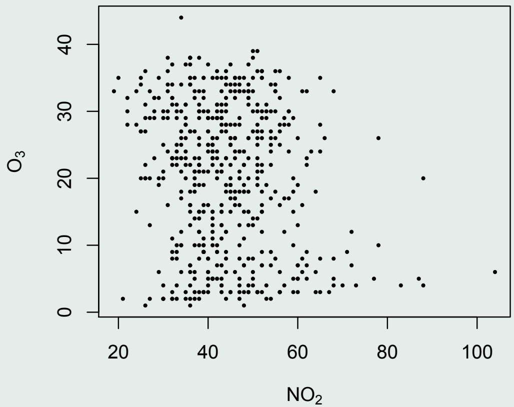
- number of dof not large, as expected with asymptotic dependent data
- ϕ behaves very well (and ok for 2T copula with dof=7)
- ϕ behaves better than δ

Applications

Puerto Rico rivers



Leeds pollutants



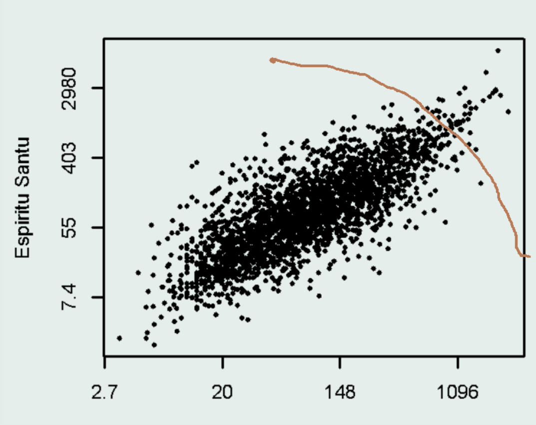
Puerto Rico rivers: 2492 observations

Leeds pollutants: 532 observations

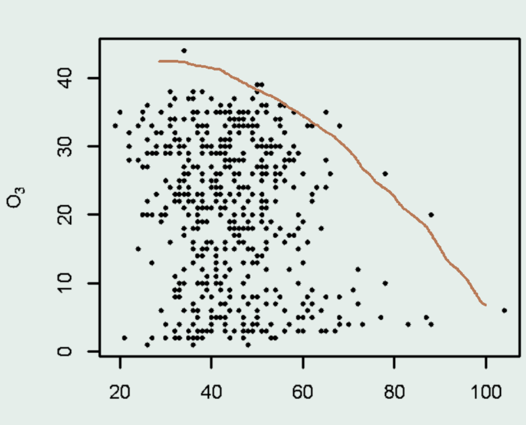
1000 and 100 observations retained for predictions only

Applications

Puerto Rico rivers



Leeds pollutants



Puerto Rico rivers: 2492 observations, asymptotic dependence

Leeds pollutants: 532 observations, asymptotic independence

1000 and 100 observations retained for predictions only

Results: predictions of the 99.5% quantile

	Empirical	Marginal	Joint	POT 90	POT 95	POT 97.5
Fajardo	[1710,1800]	1900	1865	1881	1940	1943
		(1554,2544)	(1564,2289)	(1583,2409)	(1582,2692)	(1636,2524)
Espiritu Santo	[1350,1380]	1463	1388	1465	1450	1445
		(1215,1886)	(1210,1663)	(1237,1896)	(1235,1869)	(1251,1791)

Empirical quantiles obtained from **test** dataset

POT - Peaks over threshold method

Summary: Joint > Marginal MGPD > POT

Results: exceedance probabilities $P(X_1 > x_1, X_2 > x_2)$

Puerto Rico rivers

(x_1, x_2)	(720,730)	(900,780)	(1300,1100)
Emp. Pred.	0.015	0.010	0.005
T	0.0175	0.0115	0.0044
EVD 90	0.0209	0.0141	0.0057
EVD 95	0.0214	0.0145	0.0058
EVD 97.5	0.0211	0.0154	0.0064
Bortot 90	0.0186	0.0122	0.0046
Bortot 95	0.0205	0.0135	0.0050
Bortot 97.5	0.0216	0.0153	0.0060
Ramos 90	0.0203	0.0135	0.0054
Ramos 95	0.0201	0.0136	0.0054
Ramos 97.5	0.0207	0.0149	0.0062

Leeds pollutants

(x_1, x_2)	(55,32)	(58,33)
Emp. Pred.	0.020	0.010
G	0.0188	0.0104
EVD 90	0.0549	0.0405
EVD 95	0.0854	0.0607
EVD 97.5	0.0875	0.0635
Bortot 90	0.0161	0.0085
Bortot 95	0.0133	0.0071
Bortot 97.5	0.0099	0.0050
Ramos 90	0.0114	0.0052
Ramos 95	0.0122	0.0049
Ramos 97.5	0.0093	0.0034

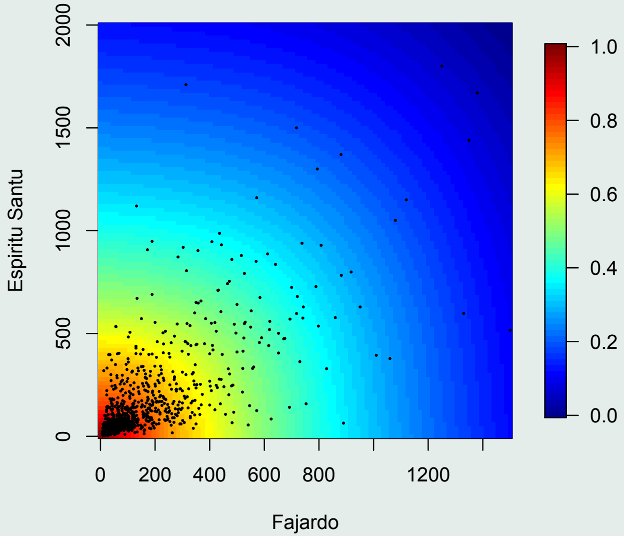
Empirical probabilities obtained from test dataset

EVD - R package EVD (Stephenson, 2002); Bortot - Bortot et al (2000);

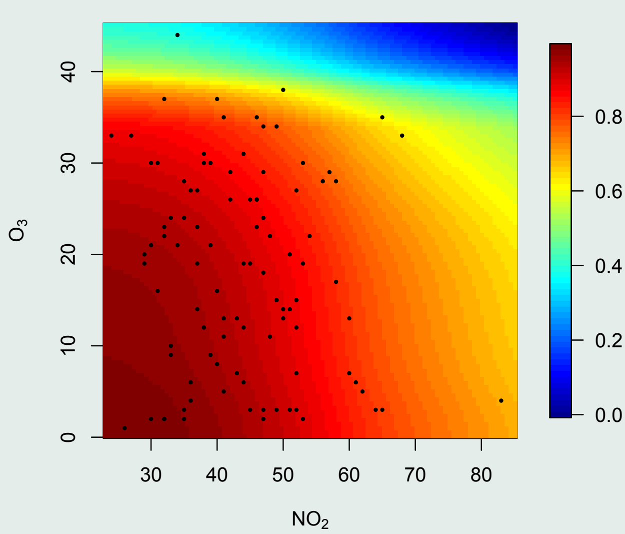
Ramos - Ramos & Ledford (2009)

Summary: Our > Bortot > Ramos > EVD

Maps of the predictive probabilities of joint exceedances



Puerto Rico rivers



Leeds pollutants

Predictive probabilities based on fitted dataset

Dots represent the **test** dataset

Results: asymptotic dependence

Puerto Rico rivers

d.o.f.	ϕ	δ_{90}	δ_{95}	$\delta_{97.5}$
5.3	0.003	0.63	0.43	0.47
(3.8,7.9)		(0.59,0.67)	(0.28,0.58)	(0.36,0.58)

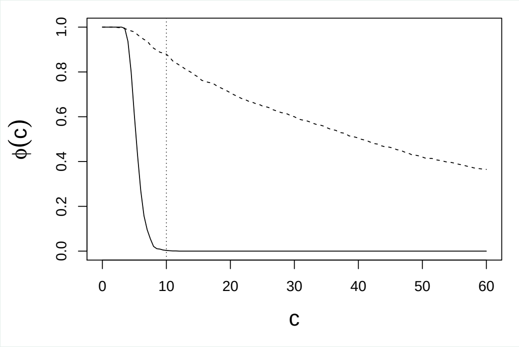
Leeds pollutants

d.o.f.	ϕ	δ_{80}
26.2	0.93	0.14
(7.7,133.2)		(0.02,0.26)

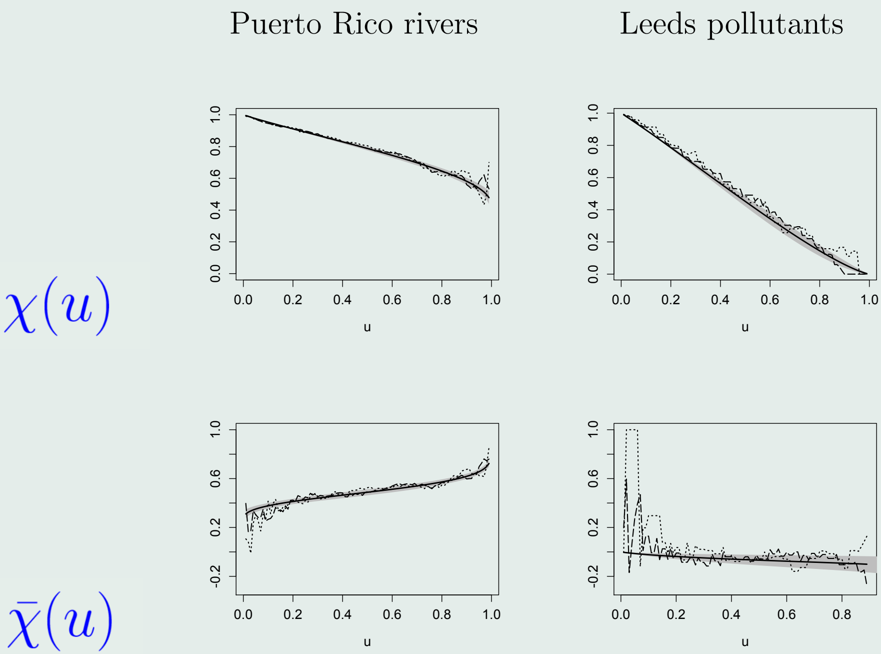
small (large) dof for Puerto Rico (Leeds) confirm visual inspection

ϕ is very decided (also, confirms visual inspection of data)

δ seems undecided for Puerto Rico



Coefficients of asymptotic dependence



Confirming asymptotic (in)dependence in Puerto Rico (Leeds)

Does bulk bias the estimation of tail?

Posterior mean (and 95% C.I.) of the dof of the T model and ϕ estimated using only extremes

	Mean	95% Int.	ϕ
Puerto Rico	9.89	(2.70,45.53)	0.25
Leeds	21.57	(2.74,107.89)	0.55

Posterior means (and 95% C.I.) for χ (Puerto Rico) and $\bar{\chi}$ (Leeds).

	Puerto Rico rivers: χ
Full dataset	0.45 (0.39,0.50)
Extreme points	0.43 (0.35,0.51)

	Leeds pollutants: $\bar{\chi}$
Full dataset	-0.13 (-0.21,-0.04)
Extreme points	-0.23 (-0.48,0.08)

Summary: Bulk did not bias results; only decreased uncertainty

4. Conclusion

- Our approach is flexible, uses the full data information and does not underestimate uncertainty

- Many extensions beyond bivariate case are available

  Vine copulae may be a possibility

- Modeling dependence separately for bulk and tail

  Main concern is the computational effort

- Regression, time series, etc can be brought to multivariate scenery

Main references

MGPD: FN, DG & HL (2012).

A semiparametric Bayesian approach to extreme value estimation

Statistics & Computing, 22, 661-675.

Regression: FN, DG & HL (2011), **EES**.

Time Series: FN, DG & HL (2016), **Test**.

Regime Identification: FN, DG & RD (2016), **BJPS**.

Multivariate extremes: ML & DG (2019).

Semiparametric bivariate modelling with flexible extremal dependence

Statistics & Computing, to appear (available online).

Gracias!

dani@im.ufrj.br

www.statpop.com.br