Modelling the distribution of univariate cluster maxima using multivariate extreme value methods

Jonathan Tawn

with

Emma Eastoe

Lancaster University

Based on Biometrika (2012) paper

Problem: What is the distribution of peak river flows?

Typically 30-50 years of river flow data but wish is estimate the level which occurs once on average in 100 years.

Standard Approach (peaks over threshold):

- Select high threshold u
- Identify independent clusters above u
- Focus on modelling only peak value Y per cluster
- Times of peaks occur as a Poisson process
- Peak sizes follow generalised Pareto distribution

Why do this? Is this the best method?

Set-up

- **Stationary series** $\{X_t\}$
- **Weak long-range dependence**
- **Marginal distribution function** F
- **Upper end point** x_F
- **Assume that there exists** $\phi_u > 0$ **such that for** $x > 0$

$$\lim_{u \to x_F} \Pr(\phi_u(X - u) > x \,|\, X > u) = [1 + \xi x]_+^{-1/\xi}$$

where ξ **is a shape parameter,** $y_+ = \max(y, 0)$

Generalised Pareto distribution (GPD)

- **For u close to x_F, motivates the asymptotic approximation for $x > 0$**

$$\Pr\{(X-u) > x \mid X > u\} = \left[1 + \frac{\xi x}{\sigma_u}\right]_+^{-1/\xi}$$

for $\sigma_u = \phi_u^{-1} > 0$

- **For large u**

$$\bar{F}(x) = p_u \left[1 + \frac{\xi(x-u)}{\sigma_u}\right]_+^{-1/\xi} \qquad x > u$$

where $p_u = \Pr(X > u) = \bar{F}(u)$

- **GPD tail for X**

GPD Extrapolation

For large u and $x > 0$

$$\Pr(X > x + u) = \left(1 + \xi \frac{x}{\sigma_u}\right)_+^{-1/\xi} \Pr(X > u)$$

We estimate $\Pr(X > u)$ empirically and use the formula for extrapolation

For an exponential tail $(\sigma_u = 1, \xi = 0)$ with $x > 0$

$$\Pr(X > x + u) = \exp(-x) \Pr(X > u)$$

Clusters and their Identification

- **Exceedances of u by $\{X_t\}$ occur in clusters: within cluster dependence, independence between clusters**
- **Use runs method to identify clusters: cluster terminates when $m-1$ consecutive values below u**
- **Leads to natural threshold-based extremal index (reciprocal mean cluster size) for threshold x of**

$$\theta(x, m) = \Pr\{\max(X_2, \ldots, X_m) < x \mid X_1 > x\}$$

Issues with dependence in cluster

- Need to account for dependence to derive distribution of block maximum
eg

$$\Pr(M_n < x) \approx \{F(x)\}^{n\theta(x,m)}$$

where $\theta(x, m)$ is threshold-based extremal index

- Ideal is to remove need to model dependence by selecting cluster maxima Y

Extremes of daily flows and peak flows

- X **daily flow**
- Y **peak daily flow**

$$\lim_{u \to x_*} \Pr\{\phi_u(X-u) > x \,|\, X > u\} = \lim_{u \to x_*} \Pr\{\phi_u(Y-u) > x \,|\, Y > u\}$$

Leadbetter (1991): Limiting asymptotic theory says both are GPD with the same parameters

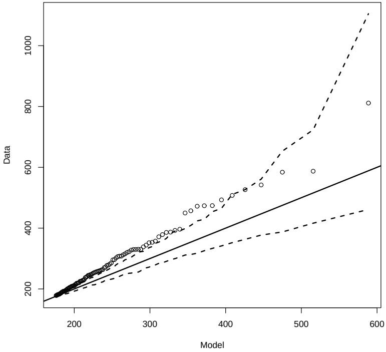
For non-limit threshold the two GPDs are different

River Lune at Caton (1979-2006, Winter daily data)
95% **threshold: 103 peaks, 251 exceedances,** $m = 12$

Parameter	X	Y
Scale	72 (60,92)	112 (89,153)
Shape	0.09 (-0.09,0.19)	0.00 (-0.31,0.12)
0.25 Quantile	21 (18,26)	32 (26,43)
0.5 Quantile	51 (44,63)	78 (63,98)
0.9 Quantile	184 (160,207)	257 (213,296)
0.99 Quantile	410 (318,485)	505 (362,618)

Each GPD fit seems fine from usual diagnostics

QQ plot for peaks under all exceedances fitted model



Limiting asymptotics are not appropriate at selected threshold

Complication: GPD diagnostics for Y do not pick up a problem

Link between distributions of X and Y

- $X \sim$ **GPD daily flow**
- Y **peak daily flow**

Rate of exceedance of peaks $\Pr(Y > u)$**, distribution of size of peaks:**

$$\Pr(Y - u > x \mid Y > u) = \frac{\theta(u + x, m)}{\theta(u, m)} \Pr(X - u > x \mid X > u)$$

where

$$\theta(x, m) = \Pr\{\max(X_2, \ldots, X_m) < x \mid X_1 > x\}$$

WHY? Link between distributions of X and Y

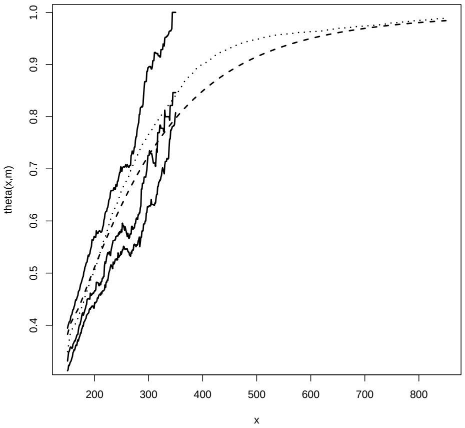
$$
\begin{aligned}
RHS &= \frac{\theta(u+x,m)}{\theta(u,m)}\Pr(X-u>x\,|\,X>u) \\
&= \frac{R(Y>u+x)}{R(X>u+x)}\,\frac{R(X>u)}{R(Y>u)}\,\frac{R(X>u+x)}{R(X>u)} \\
&= \frac{R(Y>u+x)}{R(Y>u)} \\
&= \Pr(Y-u>x\,|\,Y>u) \\
&= LHS
\end{aligned}
$$

Equality of distributions of X and Y

$$\Pr(Y - u > x \,|\, Y > u) = \frac{\theta(u + x, m)}{\theta(u, m)} \Pr(X - u > x \,|\, X > u)$$

The distributions of X and Y only agree when $\theta(u + x, m) = \theta(u, m)$ for all $x > 0$

Empirically estimated $\theta(x, m)$ for Lune data



Complication: no basis for extrapolation of plot beyond the data

New modelling strategy

For $x > 0$

$$\begin{aligned}\Pr(Y - u > x \mid Y > u) &= \frac{\theta(u+x, m)}{\theta(u, m)} \Pr(X - u > x \mid X > u) \\ &= \frac{\theta(u+x, m)}{\theta(u, m)} \left[1 + \frac{\xi x}{\sigma_u}\right]_+^{-1/\xi}\end{aligned}$$

- **Use ALL exceedances of** u **to fit GPD:** σ_u, ξ
- **Estimate** $\theta(u+x, m)$ **for** $x \geq 0$ **using ALL exceedances**
- **Need model for** $(X_2, \ldots, X_m) \mid X_1 > u$ **for large** u

Multivariate Extreme Values: Copulas

Model joint distribution function $F_{\mathbf{X}}$ of $\mathbf{X} = (X_1, \ldots, X_m)$

$$F_{\mathbf{X}}(x_1, \ldots, x_m) = C\{F(x_1), \ldots, F(x_m)\}$$

where

- F is the marginal distribution function for X_i constant over i due to stationarity
- C is the copula with uniform margins

Copulas with Gumbel margins

- By suitable transformation $\mathbf{X} \to \mathbf{S}$, C could have any marginal
- We take $\mathbf{S} = (S_1, \ldots, S_m)$ to have Gumbel marginals
- Now interested in

$$\begin{aligned} \theta(x, m) &= \Pr\{\max(S_2, \ldots, S_m) < t(x) \mid S_1 > t(x)\} \\ &= \sum_{B \in P(M)} (-1)^{|B|} \Pr\{S_j > t(x), j \in B \mid S_1 > t(x)\} \end{aligned}$$

where $t(x)$ is transform involving GPD from X to S and $P(M)$ is the power set of $\{2, \ldots, m\}$

Extremal Dependence

Pair (S_i, S_j)

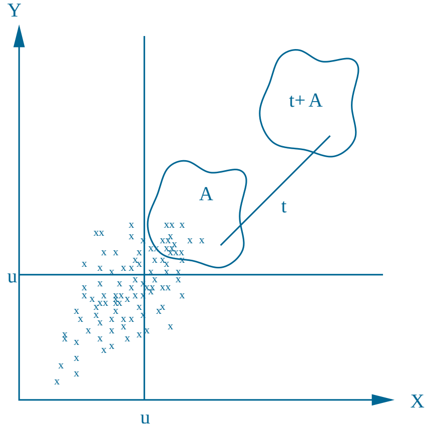
$$\chi_{ij} = \lim_{y \to \infty} \Pr(S_j > y \mid S_i > y)$$

- **Asymptotic dependence** $\chi_{ij} > 0$
- **Asymptotic independence** $\chi_{ij} = 0$

Multivariate Regular Variation

Assuming a non-degenerate multivariate regular variation on a Gumbel marginal scale implies for all sets A in tail region

$$\Pr\{\mathbf{S} \in t + A\} \approx \exp(-t)\,\Pr\{\mathbf{S} \in A\}$$

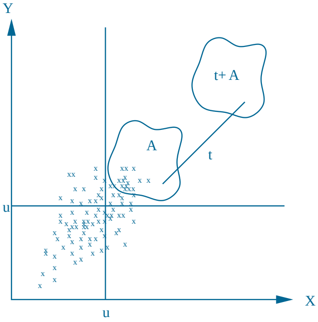


Hidden Regular Variation: Ledford and T. (1997, JRSS B)

Hidden regular variation on a Gumbel marginal scale implies for all sets A in tail region with ALL components large

$$\Pr\{\mathbf{S} \in t + A\} \approx \exp(-t/\eta_{\mathbf{S}})\,\Pr\{\mathbf{S} \in A\}$$

where $0 < \eta_{\mathbf{S}} \leq 1$



Ledford and Tawn: evaluation of $\theta(x, m)$

$$\begin{aligned}
\theta(x, m) &= \Pr\{\max(S_2, \ldots, S_m) < t(x) \mid S_1 > t(x)\} \\
&= \sum_{B \in P(M)} (-1)^{|B|} \Pr\{S_j > t(x), j \in B \mid S_1 > t(x)\} \\
&\approx \sum_{B \in P(M)} (-1)^{|B|} k_B \exp\{-t(x)[1/\eta_B - 1]\}
\end{aligned}$$

for large x

Asymptotic Dependence: a conditional viewpoint

If all variables are asymptotically dependent on S_1 then for $\mathbf{S} = (S_1, \mathbf{S}_{-1})$

$$\lim_{v \to \infty} \Pr\left(S_1 - v > s, \mathbf{S}_{-1} - S_1 < \mathbf{z} | S_1 > v\right) = \exp(-s) H(\mathbf{z})$$

with H non-degenerate and $s > 0$

If all components of $\mathbf{S}_{-1}$ are asymptotic independent on S_1 then H puts all mass at $-\infty$ for each component

Conditional Asymptotics:

Look for functions $\mathbf{a}$ and $\mathbf{b}$ such that

$$\lim_{v\to\infty} \Pr\left(S_1 - v > s \frac{\mathbf{S}_{-i} - \mathbf{a}(S_1)}{\mathbf{b}(S_1)} \leq \mathbf{z} \mid S_1 > v \right) = \exp(-s) G(\mathbf{z})$$

G is non-degenerate in each margin and $s > 0$

Note: limiting conditional independence
Applies for asymptotic dependence and asymptotic independence
Simple forms for $\mathbf{a}(s) = \alpha s$ and $\mathbf{b}(s) = s^{\beta}$ are sufficient in all theoretical examples

Conditional Method: Heffernan and T. (2004,JRSS B)

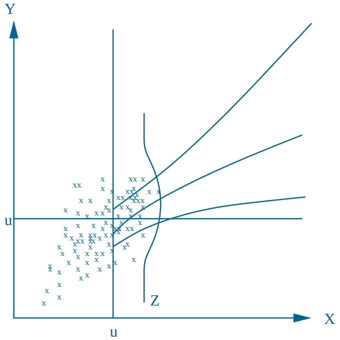
Given $S_1 = s > u$

$$\mathbf{S}_{-1} = \boldsymbol{\alpha} s + s^{\boldsymbol{\beta}} \mathbf{Z}$$

where $\mathbf{Z} \sim G$ **is independent of** S_1

$m-1$**-dimensional parameters** $-\mathbf{1} \leq \boldsymbol{\alpha} \leq \mathbf{1}$, $\boldsymbol{\beta} < \mathbf{1}$ **and additional constraints on** $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{Z}_{|i})$

Estimate G **nonparametrically**



Theoretical Examples

$$\mathbf{S}_{-1} = \boldsymbol{\alpha} S_1 + S_1^{\boldsymbol{\beta}} \mathbf{Z}$$

Asymptotic Dependence

$$\boldsymbol{\alpha} = \mathbf{1} \text{ and } \boldsymbol{\beta} = \mathbf{0}$$

Asymptotic Independence with S_j (independence)

$$\alpha_j < 1 \qquad (\alpha_j = 0, \beta_j = 0)$$

Positive (negative) extremal dependence with S_j

$$0 < \alpha_j < 1 \qquad (-1 < \alpha_j < 0)$$

Multivariate Normal Copula

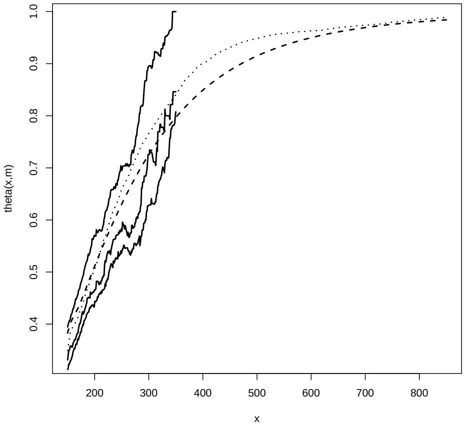
$$\alpha_j = \mathbf{sign}(\rho_{1j})\rho_{1j}^2 \text{ and } \beta_j = \frac{1}{2} \text{ for } j = 2, \ldots, m$$

Heffernan and Tawn: evaluation of $\theta(x, m)$

$$
\begin{aligned}
\theta(x, m) &= \Pr\{\max(X_2, \ldots, X_m) < x \mid X_1 > x\} \\
&= \Pr\{\max(S_2, \ldots, S_m) < t(x) \mid S_1 > t(x)\}
\end{aligned}
$$

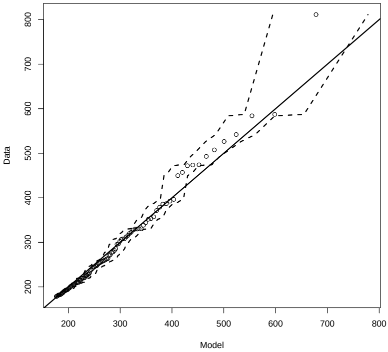
- **Simulate $S_1 | S_1 > t(x)$, Exponential**
- **Simulate $\mathbf{Z}$ independently of S_1**
- $\mathbf{S}_{-1} = \boldsymbol{\alpha} S_1 + S_1^{\boldsymbol{\beta}} \mathbf{Z}$
- **Count proportion with** $\max(S_2, \ldots, S_m) < t(x)$

Model-based estimate of $\theta(x, m)$ for Lune data



Dashed: Heffernan and Tawn conditional approach (44 parameters)
Dotted: Ledford and Tawn joint tail approach (4094 parameters)

Fit of new distribution for Lune data



Assess performance using simulation study

X_t marginally Exponential
Dependence 1st order Markov
50 years data

- Process 1 - Gaussian copula
- Process 2 - Inverted BEV copula - logistic
- Process 3 - BEV copula - logistic

Quantiles: relative bias (std dev) ($\times 10^3$)

$u = 90\%$ quantile

Excess Quantile	POT	New LT	New HT
0.99	-20 (10)	-6 (1)	5 (2)
0.9999	-90 (30)	-9 (1)	-9 (1)
0.99	-60 (20)	-10 (3)	-6 (4)
0.9999	-300 (40)	-10 (2)	-9 (2)
0.99	30 (60)	20 (30)	30 (20)
0.9999	-200 (120)	10 (20)	20 (10)

Efficiency gains at $u = 90\% : \times 10, \times 20, \times 10$
Efficiency gains at $u = 95\% : \times 2, \times 10, \times 10$
Efficiency would be much better if no bias in GPD estimation of X tail

Benefits of new approach: stationary case

- Greater theoretical justification for thresholds used in practice
- Uses more data, all values in clusters are used
- Improves quantile estimation particularly for long return periods
- Substantial efficiency gains: reduces both variance and bias relative to peaks over threshold method
  - benefit reduces as threshold increases
- Minimal differences between LT v HT: latter much easier though
- Extension to other cluster functionals is easy (for HT)

Benefits of new approach: uncertainty of m

POT:

Vary m
new cluster maxima data for each m
re-fit GPD
potential for inconsistencies over m

New Method:

Vary m
Only $\theta(x, m)$ term varies in its evaluation
Model parameters remain same

$$\Pr(Y - u > x \mid Y > u) = \frac{\theta(u + x, m)}{\theta(u, m)} \left[1 + \frac{\xi x}{\sigma_u}\right]_+^{-1/\xi}$$

Benefits of new approach: non-stationary case

Non-stationarity can occur marginally or in dependence structure:

- **POT methods cannot distinguish between these**
- **New approach captures marginal changes in GPD part and dependence changes in $\theta(x, m)$**

$$\Pr(Y - u > x \mid Y > u) = \frac{\theta(u + x, m)}{\theta(u, m)} \left[1 + \frac{\xi x}{\sigma_u}\right]_+^{-1/\xi}$$