A Conditional Approach to Modelling Multivariate Extremes

Janet E. Heffernan
Lancaster University & J. Heffernan Consulting

INTERFACE 2008      RISK : Reality

May 21 – 24, 2008
Durham, NC
Looking at Tails: why we need asymptotics

• Statistical maxim: **extrapolate at your peril!**

• Risk assessment: **need to understand tails!**

How can we reconcile these opposing requirements?
Looking at Tails: why we need asymptotics

- Statistical maxim: extrapolate at your peril!
- Risk assessment: need to understand tails!

How can we reconcile these opposing requirements?

Asymptotic models for tail behaviour:

- componentwise maxima
- point process: simultaneous excess of high marginal thresholds
Inadequacies of previous methods

Componentwise maxima / Multivariate Regular Variation

- rich class of models for Asymptotically Dependent variables
- models Asymptotically Independent variables as exactly independent
Inadequacies of previous methods

Componentwise maxima / Multivariate Regular Variation

- rich class of models for Asymptotically Dependent variables
- models Asymptotically Independent variables as exactly independent

Ledford and Tawn / Hidden Regular Variation

- richer class of models for Asymptotically Independent variables
- only allows investigation of joint tail
Conditional extreme value models

Wish list:

- rich class of models for *Asymptotically Dependent/Independent* and *Negatively* dependent variables
- multivariate (not just bivariate)
- examination of tail region where “action” is
Conditional extreme value models

Wish list:

• rich class of models for Asymptotically Dependent/Independent and Negatively dependent variables

• multivariate (not just bivariate)

• examination of tail region where “action” is!

Approach using Conditional Asymptotics:

1. condition on one variable being extreme

2. examine behaviour of remaining variables conditional on having an extreme component

Characterising the tail of a multivariate distribution

- $X = (X_1, \ldots, X_d)$; $X_{-i}$ is $X$ excluding $X_i$.

- $n$ i.i.d. observations of $X$ from unknown distribution $F$

Three stages of understanding required:

$\Pr(X_i > u)$ assessed empirically;

$X_i \mid X_i > u$ needs a marginal model;

$X_{-i} \mid X_i = x_i$ for $x_i > u$ additionally needs a dependence model.

We look at all three characteristics for each conditioning variable $i$: 
Marginal model:

Semi-parametric:

- asymptotic justification for the generalised Pareto distribution (GPD) as a conditional model above a threshold

\[
\Pr(X_i > u + x \mid X_i > u) \sim \text{GPD}(x)
\]

where \(x > 0\) and \(u\) is a high threshold for variable \(X_i\).

- Below the threshold, we use the empirical distribution \(\tilde{F}_{X_i}(x)\).
Transforming data to known margins

Leeds U.K. air quality data:

We transform variable $X$ to have standard Gumbel marginal distributions.
Conditional asymptotics for multivariate extremes

For $d$ dimensional $\mathbf{Y}$, we consider for each $i = 1, \ldots, d$

$$
P (\mathbf{Y}_{-i} < z_{-i} \mid Y_i = y_i)
$$

as $y_i \to \infty$.

(Here, $\mathbf{Y}$ has Gumbel margins.)
Asymptotic assumptions

We assume that there exist vector normalising functions

- \( a_{i}(y_{i}) : \mathbb{R} \rightarrow \mathbb{R}^{(d-1)} \)
- \( b_{i}(y_{i}) : \mathbb{R} \rightarrow \mathbb{R}^{(d-1)} \)

such that for all fixed \( z_{-i} \) and any sequence \( y_{i} \rightarrow \infty \)

\[
\lim_{y_{i} \rightarrow \infty} \Pr \left\{ \frac{Y_{-i} - a_{i}(y_{i})}{b_{i}(y_{i})} \leq z_{-i} \mid Y_{i} = y_{i} \right\} = G_{i}(z_{-i})
\]

where \( G_{i} \) has non-degenerate margins.

Operations involving vectors are componentwise.
Form of normalising functions

Analogous to parameters of GEV in univariate EVT:

\( a_{j|i}(y_i) \) satisfies:

\[
\lim_{y_i \to \infty} \Pr\{Y_j > a_{j|i}(y_i) \mid Y_i = y_i\} > 0
\]

\( b_{j|i}(y_i) \) is then given by:

\[
b_{j|i}(y_i) = h_{j|i} \{a_{j|i}(y_i) \mid Y_i = y_i\}^{-1}
\]

where \( h_{j|i} \) is the conditional hazard function for \( Y_j \mid Y_i = y_i \).
What do the functions look like?

We examined a wide range of multivariate distributions and found

- \( a_{i}(y) \) and \( b_{i}(y) \) fall in the parametric family:

\[
\begin{align*}
a_{i}(y) &= a_{i}y + I\{a_{i}=0, b_{i}<0\}(c_{i} - d_{i} \log y) \\
b_{i}(y) &= y^{b_{i}}
\end{align*}
\]

- There is no general form for \( G_{i} \) or for its marginal distributions \( G_{j|i} \).

- \( a_{i}, b_{i}, c_{i} \) and \( d_{i} \) are vector constants and \( I \) is an indicator function

- \( 0 \leq a_{j|i} \leq 1, -\infty < b_{j|i} < 1, -\infty < c_{j|i} < \infty, 0 \leq d_{j|i} \leq 1 \) for all \( j \neq i \).
Model assumptions

Asymptotic structure assumed to hold exactly above high threshold

\[
\Pr \left\{ \frac{Y_{-i} - a_{i}(y_{i})}{b_{i}(y_{i})} \leq z_{-i} \mid Y_i = y_i \right\} = G_{i}(z_{-i})
\]

for \( y_{i} > u_{Y} \).

Then

\[
Z_{-i} = \frac{Y_{-i} - a_{i}(y_{i})}{b_{i}(y_{i})}, \text{ for } y_{i} > u_{Y},
\]

are assumed to follow \( G_{i} \) and be independent of \( Y_{i} \).
Model assumptions

Asymptotic structure assumed to hold exactly above high threshold

\[
Pr \left\{ \frac{Y_{-i} - a_{i}(y_{i})}{b_{i}(y_{i})} \leq z_{-i} \mid Y_{i} = y_{i} \right\} = G_{i}(z_{-i})
\]

for \( y_{i} > u_{Y} \).

Then

\[
Z_{-i} = \frac{Y_{-i} - a_{i}(y_{i})}{b_{i}(y_{i})}, \text{ for } y_{i} > u_{Y},
\]

are assumed to follow \( G_{i} \) and be independent of \( Y_{i} \).

Model is semiparametric:

- **parametric** forms for \( a_{i}(y) \) and \( b_{i}(y) \);
- a **nonparametric** model for \( G_{i} \).
Air quality data — fitted model

\[ Y_j | Y_i = y_i : Y_j = \hat{a}_{j|i}(y_i) + \hat{b}_{j|i}(y_i)Z_{j|i} \]

for \( y_i > u_Y \).

Fitted models (Gumbel scale):
Air quality data — extrapolation

\( \text{PM}_{10} \) and NO:

Data, original margins

Data, Gumbel margins

Monte Carlo sample, Gumbel

Monte Carlo sample, original
Combining conditional models

What about self-consistency on $\{Y : Y_i > u_Y, Y_j > u_Y\}$?

We combine estimates from both models by partitioning:

$$
\Pr(Y \in C) = \Pr(Y \in C_1) + \Pr(Y \in C_2)
$$
Air quality — estimating functionals

Example: estimate return levels $v_p$ satisfying

$$\Pr \left( \sum_{i=1}^{m} Y_i > v_p \right) = p$$

for a given $p$. 

---

**Graphs**

- **O3, NO2**
  - **summer**
  - **winter**

- **NO2, SO2, PM10**
  - **perfect dependence and independence**
Conditional Approach – Conclusions

Advantages:

- accommodates any functional of multivariate extreme events
- handles asymptotic dependence and asymptotic independence including negative dependence
- is truly multivariate – not just bivariate
- has straightforward regression interpretation
Conditional Approach – Conclusions

Advantages:

- accommodates any functional of multivariate extreme events
- handles asymptotic dependence and asymptotic independence including negative dependence
- is truly multivariate – not just bivariate
- has straightforward regression interpretation

Difficulties:

- Consistency of different conditional models genuine concern
- Semi-parametric nature makes inference non-standard
Recent Applications and Developments

Applications of conditional modelling approach


Recent Applications and Developments

Applications of conditional modelling approach


Theoretical developments

References
