Practical Issues in Applications of Multivariate Extreme Values

Jonathan Tawn

with

Caroline Keef and Mark Latham

Lancaster, UK
Two Applications

- Sea-surge data
  Modelling of surge process over space for joint flood risk assessment for coastal sites and for offshore sites needed for insurance industry
Two Applications

• **Sea-surge data**
  Modelling of surge process over space for joint flood risk assessment for *coastal sites and for offshore sites* needed for *insurance industry*

• **River flow data**
  Modelling of river flow for network for joint flood risk assessment for *planning purposes and insurance*
Surge Data

Hindcast output from the CSX model, a 2d numerical surge model for the European Continental Shelf forced by DNMI pressure data for the period 1955-2000.

Data are: hourly maxima over 5-day blocks for 46 years at 259 sites.
Daily river flows for a network of sites in River Thames catchment in UK
Marginal Standardisation and Notation

\(X\): univariate variable of most interest
\(Y\): \(d\)-dimensional variable

Transform marginals to Gumbel distributions

\[\Pr(X > x) = \Pr(Y_i > x) \sim \exp(-x) \text{ as } x \to \infty \text{ for } i = 1, \ldots, d\]

Lack of Memory Property

\[\Pr(X > t + x) \sim \exp(-t) \Pr(X > x) \text{ for large } x\]

Allows focus on dependence structure
Standardisation for Surge Data

A large surge event on the Danish coast in original and transformed margins

![Surge Data Diagrams]
What is the Aim of Analysis?

- **Sea-surge data**
  
  Simulation of surge events large at a given location
  
  Estimation of spatial risk measure

  \[ E(\#\{Y > x\} \mid X > x) \]

  Dimension reduction for physical understanding

- **River flow data**

  Estimation of \( Pr(Y > x \mid X > x) \)
What is the Aim of Analysis?

- **Sea-surge data**
  Simulation of surge events large at a given location
  Estimation of spatial risk measure
  \[ E(\#\{Y > x\} \mid X > x) \]
  Dimension reduction for physical understanding

- **River flow data**
  Estimation of \[ \Pr(Y > x \mid X > x) \]
Schematic of Threshold Approach

Under assumption of asymptotic dependence

\[ \lim_{x \to \infty} \Pr(Y > x \mid X > x) > 0 \]

the following homogeneity property holds for all sets \( A \) extreme in at least one variable

\[ \Pr((X, Y) \in t + A) \approx \exp(-t) \Pr((X, Y) \in A) \]
Is Surge Process Asymptotically Dependent?

X: Danish Site

The diagrams illustrate the distribution of surge process values in the Danish site, with axes labeled as North and East. The color scale ranges from -2.2 to 8.2, indicating varying levels of surge process values in different directions.
Is Surge Process Asymptotically Dependent?

X: UK Site

![Image of graph showing surge process in North and East directions with data points and color scale]
Sites Significant on Testing for Asymptotic Dependence

X: Danish Site
Sites Significant on Testing for Asymptotic Dependence

X: UK Site
Problems for River Flow Application

Plot of data availability for Thames catchment sites

[Graph showing data availability for different sites over years]
Regression Interpretation of Threshold Method

For $X > u$

$$Y = X + Z$$

where $Z$ is independent of $X$

$$\hat{P}_r((X, Y) \in t + A) = \exp(-v) \int_v^\infty \frac{1}{m} \sum_{i=1}^m 1\{(x, x+z_i) \in t+A\} \exp(-x) \, dx$$
Heffernan and Tawn (2004, JRSS B)

For $X > u$

$$Y = aX + X^bZ$$

where $Z$ is independent of $X$

$d$-dimensional parameters $0 \leq a \leq 1$ and $b$

Nonparametric model for $Z$
Theoretical Examples

\[ Y = aX + X^bZ \]

Asymptotic Dependence

\[ a = 1 \text{ and } b = 0 \]

Asymptotic Independence with \( Y_j \)

\[ a_j < 1 \]

Multivariate Normal Copula

\[ a_j = \rho_j^2 \text{ and } b_j = \frac{1}{2} \text{ for } j = 1, \ldots, d \]
Estimates of $a$

$X$: Danish Site
Estimates of $a$

$X$: UK Site
Which Sites are Asymptotically Dependent?

Test $a_j = 1, b_j = 0$

$X$: Danish Site
Search for Parsimonious Model

Dimension of model parameters currently $259 \times 258 \times 2$

Dimension Reduction helpful/insightful
Search for Parsimonious Model

Dimension of model parameters currently $259 \times 258 \times 2$

Dimension Reduction helpful/insightful
How many sites do we need to condition on to get all sites asymptotically dependent on a conditioning site?
Search for Parsimonious Model

Dimension of model parameters currently $259 \times 258 \times 2$

Dimension Reduction helpful/insightful

How many sites do we need to condition on to get all sites asymptotically dependent on a conditioning site?
**Parsimonious Spatial Model**

Partition \((X, Y) = (X_C, Y_C)\) where

- \(X_C\) the six conditioning sites
- \(Y_C\) the remaining sites

Then

\[
[X_C, Y_C] = [X_C][Y_C \mid X_C]
\]

where \([X_C]\) is low dimensional, and

\([Y_C \mid X_C]\) is simpler due to asymptotic dependence property

Extremes for \([Y_C]\) only arise when \([X_C]\) is extreme in at least one component
Spatial Risk Measure

\[ E(\#\{Y > x\} \mid X > x) \text{ where } x \text{ is the 97\% quantile} \]

Comparison of empirical, global model, parsimonious model
Extrapolation of Spatial Risk Measure

\[ E(\#\{Y > x\} \mid X > x) \text{ where } x \text{ is the 97\% and 99.9\% quantiles for global model} \]
Simulated Fields on Original Scale

Exceeds 1000 year level on Danish coast site
Simulated Fields on Original Scale

Exceeds 1000 year level on UK coast site
Handling Missing Data for River Flows

Partition $Y = (Y_M, Y_O)$ where $Y_M$ missing; $Y_O$ observed
Also $Z = (Z_M, Z_O)$

Then need to model $[Z_M \mid Z_O]$

Approach is:
Handling Missing Data for River Flows

Partition $Y = (Y_M, Y_O)$ where $Y_M$ missing; $Y_O$ observed
Also $Z = (Z_M, Z_O)$

Then need to model $[Z_M \mid Z_O]$

Approach is:
- Transform margins

$$Z^N = T(Z) = \Phi^{-1}(\hat{F}(Z))$$
Handling Missing Data for River Flows

Partition $\mathbf{Y} = (\mathbf{Y}_M, \mathbf{Y}_O)$ where $\mathbf{Y}_M$ missing; $\mathbf{Y}_O$ observed
Also $\mathbf{Z} = (\mathbf{Z}_M, \mathbf{Z}_O)$

Then need to model $[\mathbf{Z}_M \mid \mathbf{Z}_O]$

Approach is:
- Transform margins

\[
\mathbf{Z}^N = T(\mathbf{Z}) = \Phi^{-1}(\hat{F}(\mathbf{Z}))
\]

- Model dependence by MVN copula

\[
\begin{pmatrix}
\mathbf{Z}_M^N \\
\mathbf{Z}_O^N
\end{pmatrix} \sim \text{MVN}
\begin{pmatrix}
0 \\
0
\end{pmatrix},
\begin{pmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & \Sigma_{22}
\end{pmatrix}
\]
Handling Missing Data for River Flows

Partition \( Y = (Y_M, Y_O) \) where \( Y_M \) missing; \( Y_O \) observed
Also \( Z = (Z_M, Z_O) \)

Then need to model \( [Z_M \mid Z_O] \)

Approach is:
- Transform margins

\[
Z^N = T(Z) = \Phi^{-1}(\hat{F}(Z))
\]

- Model dependence by MVN copula

\[
\begin{pmatrix}
Z^N_M \\
Z^N_O
\end{pmatrix}
\sim \text{MVN}
\begin{pmatrix}
0 \\
0
\end{pmatrix},
\begin{pmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & \Sigma_{22}
\end{pmatrix}
\]

- Take a sample from this conditional distribution

\[
[\hat{Z}_M^N \mid Z_O^N]
\]
Handling Missing Data for River Flows

Partition \( Y = (Y_M, Y_O) \) where \( Y_M \) missing; \( Y_O \) observed
Also \( Z = (Z_M, Z_O) \)

Then need to model \([Z_M \mid Z_O]\)

Approach is:
- Transform margins

\[
Z^N = T(Z) = \Phi^{-1}(\hat{F}(Z))
\]

- Model dependence by MVN copula

\[
\begin{pmatrix}
Z^N_M \\
Z^N_O
\end{pmatrix}
\sim
\text{MVN}
\begin{pmatrix}
0 \\
0
\end{pmatrix},
\begin{pmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & \Sigma_{22}
\end{pmatrix}
\]

- Take a sample from this conditional distribution

\[
[\hat{Z}^N_M \mid Z^N_O]
\]

- Back transform sample and downweight values in sample

\[
\hat{Z}_M = T^{-1}(\hat{Z}^N_M)
\]
Example of Handling Missing Data

Joint distribution model for $\mathbf{Z} = (Z_1, Z_2, Z_3)$ with infilled sample to replace missing $Z_3$ values
Extrapolation with Missing Data

Recall conditional model is for $X > u$

$$Y = aX + X^bZ$$

Extrapolation: simulate $X > v$ and independently simulate $Z$
then join as above to give $Y$
Simulation Study to Assess Infill Method

Consider 3 different patterns of missingness with

\[ X : \text{Full data}; \ Y_1 : 50\%; \ Y_2 : 90\%; \ Y_3 : 80\%; \]

9 true distributions of \( Z \)

Methods:
Use overlapping data only ★
Infill method ○

Compare Estimators of:

\[ P_i = \Pr(Y_i > x \mid X > x) \text{ for } i = 1, 2, 3 \]

by RMSE efficiency relative to the Full Data case
Efficiency Results for Handling Missing Data

Results for $P_1, P_2, P_3$ respectively

The infill method does well!