Extreme Value Predictions from Data or Monte Carlo Simulations.

A. Naess\textsuperscript{1,2)}

1) Centre for Ships and Ocean Structures
2) Department of Mathematical Sciences
Norwegian University of Science and Technology
Introduction

✓ Task: Prediction of Extreme Values from data.
Introduction

✓ Task: Prediction of Extreme Values from data.
✓ Typical situation: Long series of observed or simulated data, or blocks of data.
Introduction

✓ Task: Prediction of Extreme Values from data.
✓ Typical situation: Long series of observed or simulated data, or blocks of data.
✓ Common procedure: Identify block extremes. Assume that these are Gumbel distributed. Fit a straight line to the data on ’Gumbel paper’ by e.g. maximum likelihood.
Introduction

✓ Task: Prediction of Extreme Values from data.

✓ Typical situation: Long series of observed or simulated data, or blocks of data.

✓ Common procedure: Identify block extremes. Assume that these are Gumbel distributed. Fit a straight line to the data on ’Gumbel paper’ by e.g. maximum likelihood.

✓ Weakness: How good is the Gumbel assumption? Hard to say.
The Extreme Value Distribution

\( X(t) = 'nice' \) stationary stochastic process with \( E[X(t)] = 0 \).

\[ M_X(T) = \sup \{ X(t) ; t \in [0, T] \} \]

Goal: Determine

\[ F_{M_X(T)}(a) = Prob\{M(T) \leq a\} \]
The Extreme Value Distribution

\[ X(t) = 'nice' \text{ stationary stochastic process with } E[X(t)] = 0. \]

\[ M_X(T) = \sup \{ X(t) ; t \in [0, T] \} \]

Goal: Determine

\[ F_{M_X(T)}(a) = \text{Prob}\{M(T) \leq a\} \]

Let \( N^+(a, T) = \text{number of } a\text{-upcrossings in } [0, T] \)

\[ \mathcal{E} = \{M_X(T) \leq a\} \quad \iff \quad \mathcal{E} = \{X(0) \leq a \text{ and } N^+(a, T) = 0\} \]

Hence

\[ \text{Prob}\{\mathcal{E}\} \xrightarrow{a \to \infty} \text{Prob}\{N^+(a, T) = 0\} \]
The Extreme Value Distribution

Assumption: Upcrossings of high levels are statistically independent events. \(\rightarrow\) The number of upcrossings in a time interval are Poisson distributed.
The Extreme Value Distribution

Assumption: Upcrossings of high levels are statistically independent events. ⇒ The number of upcrossings in a time interval are Poisson distributed.

Hence

\[ \text{Prob}\{\mathcal{E}\} \approx \text{Prob}\{N^+(a, T) = 0\} = \exp\{-E[N^+(a, T)]\} \]

\[ E[N^+(a, T)] = \nu^+_X(a) T \]

\[ \nu^+_X(a) = \text{expected number of } a\text{-upcrossings per unit of time.} \]
The Extreme Value Distribution

Assumption: Upcrossings of high levels are statistically independent events. \( \implies \) The number of upcrossings in a time interval are Poisson distributed.

Hence

\[
Prob\{\mathcal{E}\} \approx Prob\{N^+(a, T) = 0\} = \exp\{-E[N^+(a, T)]\}
\]

\[
E[N^+(a, T)] = \nu_X^+(a) T
\]

\( \nu_X^+(a) = \) expected number of \( a \)-upcrossings per unit of time.

The Rice-formula

\[
\nu_X^+(a) = \int_0^\infty sf_X \dot{X}(a, s) ds = E[\dot{X}^+ | X = a] \cdot f_X(a)
\]

where \( E[\dot{X}^+ | X = a] = E[\dot{X}^+] \) if \( X(t) \) and \( \dot{X}(t) \) are independent.
The Extreme Value Distribution

Assume that the PDF $f_X(x)$ of $X(t)$ is given as

$$f_X(x) = A \exp\{-\alpha(x)\}$$

where $A$ is a suitable constant and $\alpha(x)$ is a well-behaved function that is strictly increasing for increasing $x$ for $x \geq x_0$ for some $x_0 \ (> 0)$. 

"Extreme Value Predictions from Data or Monte Carlo Simulations. – p. 5/21"
Assume that the PDF $f_X(x)$ of $X(t)$ is given as

$$f_X(x) = A \exp \{-\alpha(x)\}$$

where $A$ is a suitable constant and $\alpha(x)$ is a well-behaved function that is strictly increasing for increasing $x$ for $x \geq x_0$ for some $x_0 (> 0)$. For the processes of concern to us here, we may assume that

$$\nu_X^+(x) = \nu \exp \{-\alpha(x) + \delta(x)\}$$

$\nu = a$ positive constant.

$\delta(x)$: $|\delta(x)|$ is of much slower increase than $\alpha(x)$ as $x \to \infty$. 

The Extreme Value Distribution
The Extreme Value Distribution

Assume that the PDF $f_X(x)$ of $X(t)$ is given as

$$f_X(x) = A \exp \{-\alpha(x)\}$$

where $A$ is a suitable constant and $\alpha(x)$ is a well-behaved function that is strictly increasing for increasing $x$ for $x \geq x_0$ for some $x_0 (> 0)$.

For the processes of concern to us here, we may assume that

$$\nu^+_X(x) = \nu \exp \{-\alpha(x) + \delta(x)\}$$

$\nu = a$ positive constant.

$\delta(x): |\delta(x)|$ is of much slower increase than $\alpha(x)$ as $x \to \infty$.

$$F_{M_X(T)}(x) = \text{Prob} \{M_X(T) \leq x\} = \exp \{-\nu^+_X(x)T\}$$

$$= \exp \{-\exp \{-\alpha(x) + \delta(x) + \ln(\nu T)\}\}$$
The Extreme Value Distribution

From general theory, $F_{M_X(T)}(x)$ can be expected to approach a Gumbel distribution when $T$ increases to large values. That is,

$$F_{M_X(T)}(x) \rightarrow \exp\{ - \exp\{ -\alpha_T (x - b_T) \} \}, \ T \rightarrow \infty$$

for suitable parameters $a_T > 0$ and $b_T$. 
The Extreme Value Distribution

From general theory, $F_{M_X(T)}(x)$ can be expected to approach a Gumbel distribution when $T$ increases to large values. That is,

$$F_{M_X(T)}(x) \to \exp\left\{-\exp\{-\alpha_T(x - b_T)\}\right\}, \quad T \to \infty$$

for suitable parameters $a_T > 0$ and $b_T$.

Problem: Almost no useful information on the rate of convergence is available.
From general theory, $F_{M_X(T)}(x)$ can be expected to approach a Gumbel distribution when $T$ increases to large values. That is,

$$F_{M_X(T)}(x) \rightarrow \exp\{-\exp\{-\alpha_T(x - b_T)\}\}, \quad T \rightarrow \infty$$

for suitable parameters $a_T > 0$ and $b_T$.

Problem: Almost no useful information on the rate of convergence is available.

The convergence of $F_{M_X(T)}(x)$ towards a Gumbel distribution can be very slow.
The Extreme Value Distribution

From general theory, $F_{M_X(T)}(x)$ can be expected to approach a Gumbel distribution when $T$ increases to large values. That is,

$$F_{M_X(T)}(x) \rightarrow \exp\{- \exp\{-\alpha_T(x - b_T)\}\}, \quad T \rightarrow \infty$$

for suitable parameters $a_T > 0$ and $b_T$.

Problem: Almost no useful information on the rate of convergence is available.

The convergence of $F_{M_X(T)}(x)$ towards a Gumbel distribution can be very slow.

Still, this asymptotic result is used extensively in practical extreme value analyses to justify the application of a Gumbel distribution for extrapolating to long return period design values from a limited number of extreme value observations.
The Extreme Value Distribution

Introduce a new stochastic process $Y(t)$ defined as follows:

$$Y(t) = \beta[X(t)]$$

where $\beta(\cdot) =$ strictly increasing function, $\beta(x) = \alpha(x)$ for $x \geq x_0$. 
Introduce a new stochastic process $Y(t)$ defined as follows:

$$Y(t) = \beta[X(t)]$$

where $\beta(\cdot)$ = strictly increasing function, $\beta(x) = \alpha(x)$ for $x \geq x_0$.

The extreme value distribution $F_{MY(T)}(y)$ of $Y(t)$:

$$F_{MY(T)}(y) = \exp\{-\exp\{-\alpha[\beta^{-1}(y)] + \delta[\beta^{-1}(y)] + \ln(\nu T)\}\}$$
The Extreme Value Distribution

Introduce a new stochastic process \( Y(t) \) defined as follows:

\[
Y(t) = \beta[X(t)]
\]

where \( \beta(\cdot) = \text{strictly increasing function}, \beta(x) = \alpha(x) \) for \( x \geq x_0 \).

The extreme value distribution \( F_{MY(T)}(y) \) of \( Y(t) \):

\[
F_{MY(T)}(y) = \exp\{-\exp\{-\alpha[\beta^{-1}(y)] + \delta[\beta^{-1}(y)] + \ln(\nu T)\}\}
\]

For \( y \geq y_o = \alpha(x_o) \)

\[
F_{MY(T)}(y) = \exp\{-\exp\{-y + \delta[\beta^{-1}(y)] + \ln(\nu T)\}\}
\]
The Extreme Value Distribution

Introduce a new stochastic process $Y(t)$ defined as follows:

$$Y(t) = \beta[X(t)]$$

where $\beta(\cdot) = \text{strictly increasing function}$, $\beta(x) = \alpha(x)$ for $x \geq x_0$.

The extreme value distribution $F_{MY(T)}(y)$ of $Y(t)$:

$$F_{MY(T)}(y) = \exp\{-\exp\{-\alpha[\beta^{-1}(y)] + \delta[\beta^{-1}(y)] + \ln(\nu T)\}\}$$

For $y \geq y_o = \alpha(x_o)$

$$F_{MY(T)}(y) = \exp\{-\exp\{-y + \delta[\beta^{-1}(y)] + \ln(\nu T)\}\}$$

Previous comments $\Rightarrow \delta[\beta^{-1}(y)]/y$ negligible for relevant extreme values. Conclusion: The transformed process $Y(t)$ has an extreme value distribution that is close to the Gumbel distribution for large $T$. 
The POT Approach

The Peaks-Over-Threshold approach is based on utilizing all peak events of a given time series exceeding a specified threshold.
The POT Approach

The Peaks-Over-Threshold approach is based on utilizing all peak events of a given time series exceeding a specified threshold.

The word ‘event’ is used deliberately, because in many practical applications, one would consider peak values with a time separation less than a certain value to belong to the same peak event.
The POT Approach

By considering peak events instead of epoch maxima, the number of data-points for statistical processing may be increased considerably.
The POT Approach

By considering peak events instead of epoch maxima, the number of data-points for statistical processing may be increased considerably.

However, in this connection it is appropriate to recall the basic assumption of the POT method that the threshold is high relative to typical values.
The POT Approach

By considering peak events instead of epoch maxima, the number of data-points for statistical processing may be increased considerably.

However, in this connection it is appropriate to recall the basic assumption of the POT method that the threshold is high relative to typical values.

Excessive lowering of the threshold to obtain more data may lead to substantial bias.
The POT Approach

There are some basic concerns regarding the quality of the data that should be addressed before applying a POT analysis.
The POT Approach

There are some basic concerns regarding the quality of the data that should be addressed before applying a POT analysis.

One is related to trend and ’seasonal’ variation, i.e. non-stationarity.
The POT Approach

There are some basic concerns regarding the quality of the data that should be addressed before applying a POT analysis.

One is related to trend and ’seasonal’ variation, i.e. non-stationarity.

Another concern is related to the amount of data. To get reasonable results, there must be enough data to ensure that the threshold level is not excessively lowered during the analysis.
The POT Approach

The POT method is based on the generalized Pareto (GP) distribution in the following manner:
The POT Approach

The POT method is based on the generalized Pareto (GP) distribution in the following manner:

It has been shown that asymptotically, the excess values above a high level will follow a GP distribution if and only if the parent distribution belongs to the domain of attraction of one of the extreme value distributions.
The POT Approach

The POT method is based on the generalized Pareto (GP) distribution in the following manner:

It has been shown that asymptotically, the excess values above a high level will follow a GP distribution if and only if the parent distribution belongs to the domain of attraction of one of the extreme value distributions.

The assumption of a Poisson process model for the exceedance times combined with GP distributed excesses lead to the generalized extreme value (GEV) distribution for corresponding extremes.
The Generalized Pareto Distribution

The expression for the GP distribution is

\[ G(y) = G(y; a, c) = \text{Prob}[Y \leq y] = 1 - \left(1 + c \frac{y}{a}\right)^{-1/c} \]

Here \( a > 0 \) is a scale parameter and \( c \) determines the shape of the distribution. \((z)_+ = \max(0, z)\).
The Generalized Pareto Distribution

The expression for the GP distribution is

\[ G(y) = G(y; a, c) = \text{Prob}[Y \leq y] = 1 - \left(1 + c \frac{y}{a}\right)^{-1/c} \]

Here \( a > 0 \) is a scale parameter and \( c \) determines the shape of the distribution. \((z)_+ = \max(0, z)\).

\( G(y) \) represents the conditional CDF of the excess \( Y = X - u \), given that \( X > u \) for \( u \) sufficiently large.
The Generalized Pareto Distribution

The expression for the GP distribution is

\[ G(y) = G(y; a, c) = \text{Prob}[Y \leq y] = 1 - \left(1 + c \frac{y}{a}\right)^{-1/c} \]

Here \( a > 0 \) is a scale parameter and \( c \) determines the shape of the distribution. \((z)_+ = \max(0, z)\).

\( G(y) \) represents the conditional CDF of the excess \( Y = X - u \), given that \( X > u \) for \( u \) sufficiently large.

The cases \( c > 0 \), \( c = 0 \) and \( c < 0 \) correspond to Fréchet (Type II), Gumbel (Type I), and reverse Weibull (Type III) domains of attraction, respectively.
The expression for the GP distribution is

\[ G(y) = G(y; a, c) = \text{Prob}[Y \leq y] = 1 - \left(1 + c\frac{y}{a}\right)^{-1/c} \]

Here \( a > 0 \) is a scale parameter and \( c \) determines the shape of the distribution. \((z)_+ = \max(0, z)\).

\( G(y) \) represents the conditional CDF of the excess \( Y = X - u \), given that \( X > u \) for \( u \) sufficiently large.

The cases \( c > 0 \), \( c = 0 \) and \( c < 0 \) correspond to Fréchet (Type II), Gumbel (Type I), and reverse Weibull (Type III) domains of attraction, respectively.

For \( c = 0 \), the expression \( \left(1 + c\frac{y}{a}\right)^{-1/c} \) is understood as \( \exp(-y/a) \).
The Generalized Extreme Value Distribution

Analysis based on the maximum observation during specified periods of time, like one year, assumes that the resulting set of data are independent and identically distributed (iid) and follow a generalized extreme value (GEV) distribution for maxima with CDF

$$GEV(x; a, b, c) = \exp \left\{ - \left( 1 + c \frac{x - b}{a} \right)^{-1/c} \right\}$$

$a = \text{scale parameter},$ $b = \text{location parameter},$ $c = \text{shape parameter}.$
The Generalized Extreme Value Distribution

Analysis based on the maximum observation during specified periods of time, like one year, assumes that the resulting set of data are independent and identically distributed (iid) and follow a generalized extreme value (GEV) distribution for maxima with CDF

$$GEV(x; a, b, c) = \exp \left\{ - \left( 1 + c \frac{x - b}{a} \right)^{-1/c} \right\}$$

$a = \text{scale parameter}, \ b = \text{location parameter}, \ c = \text{shape parameter}$.

The cases $c > 0, \ c = 0$ and $c < 0$ correspond to Fréchet (Type II), Gumbel (Type I), and reverse Weibull (Type III) extreme value distributions, respectively.
The Generalized Extreme Value Distribution

c = 0 is again interpreted as a limiting case

\[ GEV(x; a, b, 0) = \exp \left\{ -e^{-(x-b)/a} \right\} \]

which is the standard expression for the Gumbel distribution.
The Generalized Extreme Value Distribution

\( c = 0 \) is again interpreted as a limiting case

\[
GEV(x; a, b, 0) = \exp \left\{ -e^{-(x-b)/a} \right\}
\]

which is the standard expression for the Gumbel distribution.

The \( c \)-parameter in the GP distribution representing the asymptotic
distribution of exceedances, is the same as the \( c \)-parameter in the
corresponding GEV distribution.
The de Haan Estimators

\[ n = \text{the total number of data points or peaks.} \quad k = \text{the number of observations above the threshold } u. \]
$n = \text{the total number of data points or peaks. } k = \text{the number of observations above the threshold } u.$

The highest, second highest, ..., $k$-th highest, $(k + 1)$-th highest variates are denoted by $X_{n,n}, X_{n-1,n}, ..., X_{n-(k+1),n}, X_{n-k,n} = u$, respectively.
The de Haan Estimators

$n = \text{the total number of data points or peaks.} \; k = \text{the number of observations above the threshold } u.$

The highest, second highest, ..., $k$-th highest, $(k+1)$-th highest variates are denoted by $X_{n,n}$, $X_{n-1,n}$, ..., $X_{n-(k+1),n}$, $X_{n-k,n} = u$, respectively.

The parameter estimators are based on the quantities

$$M_n^{(r)} = \frac{1}{k} \sum_{i=0}^{k-1} \{\ln(X_{n-i,n}) - \ln(X_{n-k,n})\}^r$$

$$= \frac{1}{k} \sum_{i=0}^{k-1} \{\ln(X_{n-i,n}) - \ln(u)\}^r$$

defined for $r = 1, 2$. 
The de Haan Estimators

The estimator $\hat{c}$ of $c$:

$$\hat{c} = M_n^{(1)} + 1 - \frac{1}{2} \left\{ 1 - \frac{(M_n^{(1)})^2}{M_n^{(2)}} \right\}^{-1}$$

$\hat{c} \rightarrow c$ as $n \rightarrow \infty \ (i.p.)$
The de Haan Estimators

The estimator $\hat{c}$ of $c$:

$$\hat{c} = M_n^{(1)} + 1 - \frac{1}{2} \left\{ 1 - \frac{(M_n^{(1)})^2}{M_n^{(2)}} \right\}^{-1}$$

$\hat{c} \rightarrow c$ as $n \rightarrow \infty$ (i.p.)

The estimator $\hat{a}$ of $a$:

$$\hat{a} = \rho X_{n-k,n} M_n^{(1)} = \rho u M_n^{(1)}$$

where $\rho = 1$ if $\hat{c} \geq 0$, while $\rho = 1 - \hat{c}$ if $\hat{c} < 0$. 
The Moment Estimators

In terms of the mean value $E(Y)$ and the standard deviation $s(Y)$ of the exceedance variate $Y$, it can be shown that

$$a = \frac{1}{2} E(Y) \{1 + [E(Y)/s(Y)]^2\}$$

and

$$c = \frac{1}{2} \{1 - [E(Y)/s(Y)]^2\}$$

which provide the moment estimators for $a$ and $c$. 
The c0 Estimators

Assume, as previously, that

\[ f_X(x) = A \exp \{-\alpha(x)\} \]

Then the data obtained by transforming the initial data using the function \( \alpha(x) \) are Gumbel distributed with good approximation. That is, the transformed data will have \( c \approx 0 \).
The c0 Estimators

Assume, as previously, that

\[ f_X(x) = A \exp \{-\alpha(x)\} \]

Then the data obtained by transforming the initial data using the function \( \alpha(x) \) are Gumbel distributed with good approximation. That is, the transformed data will have \( c \approx 0 \).

The c0 estimators apply to the transformed data, for which only the \( a \)-parameter has to be estimated.
The c0 Estimators

Assume, as previously, that

\[ f_X(x) = A \exp\{-\alpha(x)\} \]

Then the data obtained by transforming the initial data using the function \( \alpha(x) \) are Gumbel distributed with good approximation. That is, the transformed data will have \( c \approx 0 \).

The \( c0 \) estimators apply to the transformed data, for which only the \( a \)-parameter has to be estimated.

This leads to two \( c0 \) estimators: One based on the de Haan estimator for \( a \). The other is based on the moment estimator for \( a \).
Estimation of Wind Speed

The long term distribution of wind speed is well described by the distribution

\[ F_X(x) = 1 - \exp \left\{ -\gamma x^\theta \right\} \]

where \( 1.5 \lesssim \theta \lesssim 2.5 \).
Estimation of Wind Speed

The long term distribution of wind speed is well described by the distribution

\[ F_X(x) = 1 - \exp \left\{ -\gamma x^\theta \right\} \]

where \(1.5 \lesssim \theta \lesssim 2.5\).

As a representative value: \(\theta = 2.0\)
Utsira

Extreme Value Predictions from Data or Monte Carlo Simulations. – p. 20/21
Ferder