Risk and Environmental Sustainability

Linda Mhalla

1 Introduction **2**

1.1 Motivation **3**

1.2 Revision **20**

1.3 Example **34**

2 Statistics of Extremes **47**

2.1 Basic Methods for Maxima **48**
- Maxima 49
- Extremal Types Theorem 51

2.2 Basic Methods for Exceedances **67**
- Exceedance Theorem 68
- Statistical implications of stability 71

2.3 Targets of Inference **81**

3 Poisson Process **87**
- Improved inferences 88
- Eskdalemuir rainfall 89

3.1 Point Processes **93**

3.2 Poisson Processes **99**

3.3 Poisson Modelling of Extremes **106**

4 Complications **116**

4.1 Introduction **117**

4.2 Nonstationarity **119**

4.3 Dependence **146**

5 Multivariate Extremes **166**

5.1 Introduction **167**

5.2 Copulas **174**

5.3 Multivariate Models for Extremes **182**

5.4 Statistical Modelling **193**

1 Introduction

1.1 Motivation

Sustainability?

- ☐ How robust are human activities to environmental hazards in a changing world?
  - Sea level change?
  - Earthquakes, tsunamis, major windstorms?
  - Increases in air and water temperatures?
  - Changes to permafrost?
  - Changes in rainfall patterns — droughts and floods?
  - . . .
- ☐ Some examples, among many . . .

Irma, September 2017

Bondo, August 2017

Fukushima, March 2011

Nuclear power safety

☐ Fukushima ⇒ nuclear power safety concerns worldwide

☐ Swiss nuclear regulator asked for (re-)assessment of vulnerability of the four nuclear plants to

- high and low air temperatures
- high and low river water temperatures
- high winds (and tornados)
- intense rainfall, snowload, lightning strikes,
- earthquakes and any tsunamis are dealt with separately!

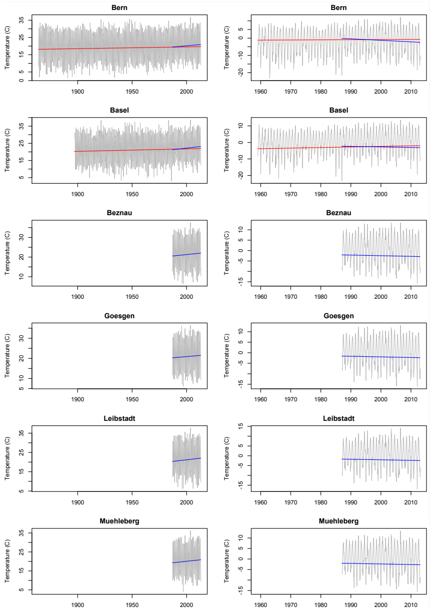
☐ Task: estimate quantiles for probabilities 10^{-4} per year (and 10^{-7} for high winds), and give their uncertainties

- based on 25 years of data or so at the plants themselves, and (at very most, and only for comparison) 150 years of data nearby

Swiss nuclear plants

Muhleberg

Air temperature maxima and minima



Tanaguarena, 1999

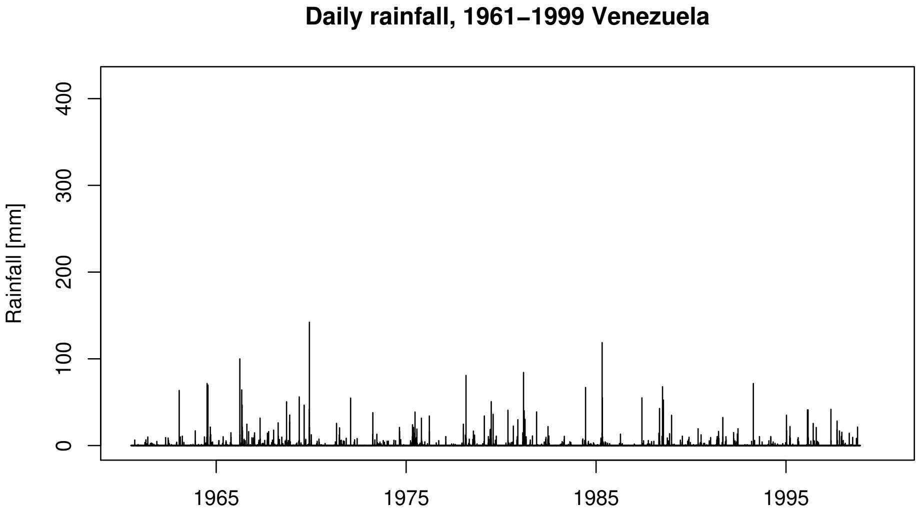
☐ Following two weeks of intermittent rainfall, torrential rainfall on 14–16 December 1999 spawned landsides throughout the upper watersheds of the Cerro Grande River near the coast of Venezuela.

☐ Mud floods, debris flows and flood surges then destroyed much of Tanaguarena and other coastal tourist towns. Perhaps 30,000 people died.

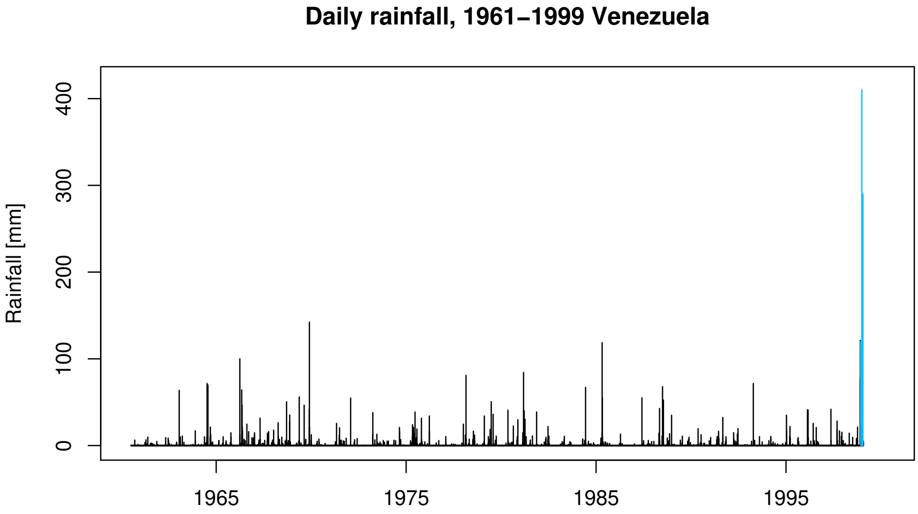
☐ The data are from the airport at Maiquetia: the estimated recurrence time for the three-day rainfall is between 250 years and 6 million years!

☐ Similar events, fortunately with less loss of life, have occurred nearby.

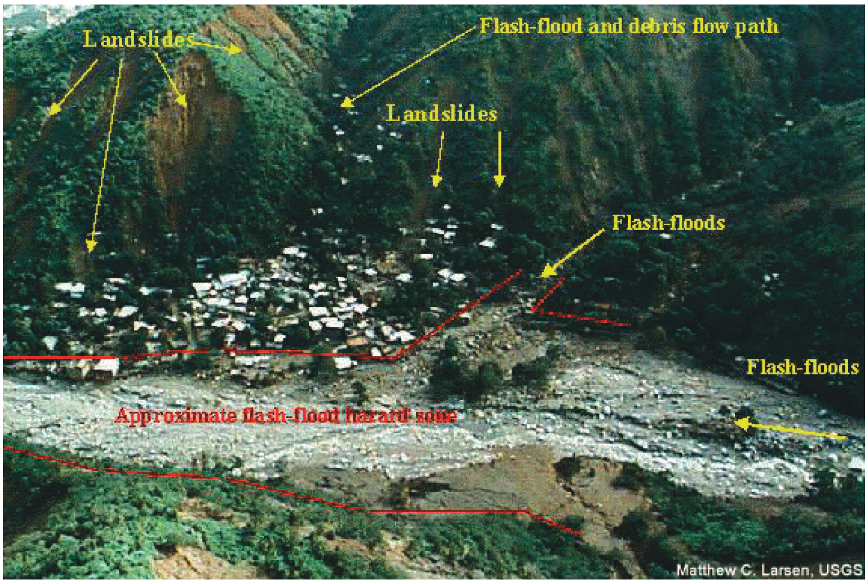
Rainfall at Maiquetia



Rainfall at Maiquetia

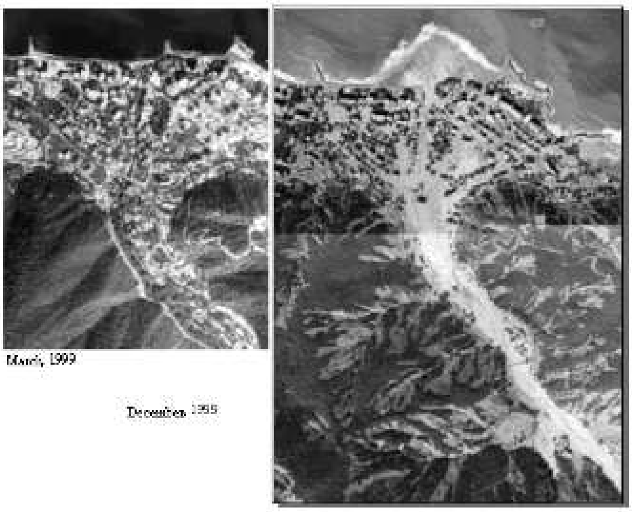


Tanaguarena



Cerro Grande rivermouth

Comparison of Cerro Grande fan before and after the Dec. 1999 flood disaster.



Risk

- From the Oxford English Dictionary:

  > (Exposure to) the possibility of loss, injury, or other adverse or unwelcome circumstance; a chance or situation involving such a possibility.

- Risk R can be expressed as

  $$R = (A, C, U, P, K),$$

  where

  A is an event that might occur,
  C is the consequences of the event,
  U is an assessment of uncertainties,
  P is a knowledge-based probability of the event
  K is the background knowledge that U and P are based on.

- The consequences C are highly situation-specific, so we focus on methods for estimating the risks based on data.
- This course mostly concerns the estimation of the probabilities P of rare events A based on data K that leads to a robust assessment of their uncertainties U.

Environmental sustainability

- Climate change, loss of biodiversity, population growth . . . all threaten our future.
- Change to average conditions are important — world GDP is estimated to drop by 12% for each 1°C of warming (WEF) — but many immediate impacts come from increases in the sizes and occurrence of (previously) rare events:
  - heat waves are dangerous for vulnerable human populations and can impact on food security;
  - hurricanes, typhoons and other major storms can have massive impacts on habitations and consequently on insurance premiums;
  - heavy rainfall leading to widespread flooding can make homes uninhabitable for months and lead to drastic reductions in their value;
  - wildfires can devastate large areas even in first world countries (e.g., Los Angeles earlier this year);
  - et cetera . . .
- Economic sustainability (major financial crashes, food prices, . . .) also involve (formerly) rare events.
- Many such events are **compound**, i.e., depend on a rare combination of several variables.

Plan

- ☐ Many risky situations can be formulated in terms of the Poisson process, which is a basic stochastic model for point events — analogous to the Gaussian distribution in modelling continuous random variables.
- ☐ Draft plan . . .
  - Today: motivation, basics of statistical modelling, Poisson process
  - Weeks 2–6: Modelling rare events (extreme-value statistics)
  - Weeks 7–8: Point process and Poisson process
  - Weeks 9–10: Multivariate (compound) rare events
  - Weeks 11–14: Probabilistic forecasting
- ☐ Much of the course will use the contents of Coles (2001) *An Introduction to the Statistical Modeling of Extreme Values*, Springer.

1.2 Revision

Statistical models

- ☐ A **statistical model** is a set of probability distributions used to
  - describe the variation in (future or existing) data;
  - help understand underlying phenomena;
  - predict future data and answer 'what if' questions;
  - give a realistic assessment of the uncertainty of inferences.
- ☐ We suppose that observed data y are a realisation of a random variable Y from the model, so y might have been different.
- ☐ A model is **parametric** if the distributions can be indexed by a finite parameter vector θ; otherwise it is **nonparametric**.
  - $y_1, \ldots, y_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$, with $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$, is a parametric model;
  - $y_1, \ldots, y_n \overset{\text{iid}}{\sim} F$, with F unknown, is a nonparametric model.
- ☐ In this course almost all the models will be parametric, and key steps are
  - formulation of appropriate models;
  - inference on the parameters, usually by likelihood methods.

Likelihood

- ☐ Let y be a data set, assumed to be the realisation of a random variable Y from a parametric model $f_Y(y;\theta)$, where the unknown parameter θ lies in a **parameter space** $\Theta \subset \mathbb{R}^p$.
- ☐ The **likelihood** (for θ based on y) and the corresponding **log likelihood** are

$$L(\theta) = L(\theta; y) = f_Y(y;\theta), \quad \ell(\theta) = \log L(\theta), \quad \theta \in \Theta.$$

- ☐ The **maximum likelihood estimate** (MLE) $\widehat{\theta}$ satisfies $\ell(\widehat{\theta}) \geq \ell(\theta)$, for all $\theta \in \Theta$.
- ☐ Often $\widehat{\theta}$ is unique and in many cases it satisfies the **score (or likelihood) equation**

$$\frac{\partial \ell(\theta)}{\partial \theta} = 0,$$

which is interpreted as a vector equation of dimension $p \times 1$ if θ is a $p \times 1$ vector.

- ☐ The **observed information** and **expected (Fisher) information** are defined as

$$\jmath(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^{\mathrm{T}}}, \quad \imath(\theta) = \mathrm{E}\left\{\jmath(\theta)\right\};$$

these are $p \times p$ matrices if θ has dimension p. The information matrices encode the curvature of the log likelihood and provide information about the variability of $\widehat{\theta}$.

Log likelihood

- ☐ For both theoretical and numerical reasons we prefer to work with the log likelihood.
- ☐ If the data are a random sample, i.e., $y_1, \ldots, y_n \overset{\text{iid}}{\sim} f(y;\theta)$, then

$$L(\theta) = f(y;\theta) = f(y_1, \ldots, y_n; \theta) = \prod_{j=1}^{n} f(y_j;\theta), \quad \theta \in \Theta,$$

so

$$\ell(\theta) = \log L(\theta) = \sum_{j=1}^{n} \log f(y_j;\theta), \quad \theta \in \Theta.$$

- ☐ If the data are independent but not identically distributed, with $y_j \sim f_j(y_j;\theta)$, then

$$\ell(\theta) = \sum_{j=1}^{n} \log f_j(y_j;\theta), \quad \theta \in \Theta.$$

- ☐ If the data are dependent and ordered in time, then we can write

$$\ell(\theta) = \log f(y_1;\theta) + \sum_{j=2}^{n} \log f(y_j \mid y_1, \ldots, y_{j-1};\theta), \quad \theta \in \Theta.$$

- ☐ In each case the information matrices are sums and (under mild conditions) are of order n.

Maximum likelihood estimator

☐ In large samples from a **regular model** in which the true parameter is $\theta^0_{p\times 1}$, the maximum likelihood estimator $\widehat{\theta}$ has an approximate normal distribution,

$$\widehat{\theta} \dot{\sim} \mathcal{N}_p\left\{\theta^0, \jmath(\widehat{\theta})^{-1}\right\},$$

so we can compute an approximate $(1-2\alpha)$ confidence interval for the rth parameter θ^0_r as

$$\widehat{\theta}_r \pm z_\alpha v_{rr}^{1/2},$$

where v_{rr} is the rth diagonal element of the matrix $\jmath(\widehat{\theta})^{-1}$.

☐ This approximation also holds under weaker conditions, for non-identically distributed and dependent data.

☐ This is easily implemented:

- we (carefully!) code the negative log likelihood $-\ell(\theta)$;
- we minimise $-\ell(\theta)$ numerically, ensuring that the routine returns $\widehat{\theta}$ and the Hessian matrix $\jmath(\widehat{\theta}) = -\partial^2\ell(\theta)/\partial\theta\partial\theta^{\mathrm{T}}|_{\theta=\widehat{\theta}}$
- we compute $\jmath(\widehat{\theta})^{-1}$, and use the square roots of its diagonal elements, $v_{11}^{1/2}, \ldots, v_{pp}^{1/2}$, as standard errors for the corresponding elements of $\widehat{\theta}$.

Delta method and transformations

The asymptotic normality result can be used to derive standard errors for other quantities of interest. If $\widehat{\phi} = g(\widehat{\theta})$, where $g : \mathbb{R}^p \to \mathbb{R}^k$ for $k \leq p$ is a differentiable function of θ non-vanishing at θ^0 then

$$\widehat{\phi} \dot{\sim} \mathcal{N}_k(\phi^0, \nabla\phi^{\mathrm{T}}\jmath(\widehat{\theta})^{-1}\nabla\phi),$$

where

$$\nabla\phi = \left[\partial\phi/\partial\theta_1, \ldots, \partial\phi/\partial\theta_p\right]^{\mathrm{T}}.$$

The variance matrix and the jacobian are evaluated at the maximum likelihood estimate $\widehat{\theta}$.

Maximum likelihood estimator

Example 1 (Weibull model)
Let $(y_1, \ldots, y_n)$ *be a data set, assumed to be a realisation of a Weibull random variable with scale* $\lambda > 0$ *and shape* $\alpha > 0$. *The corresponding density is*

$$f_Y(y; \lambda, \alpha) = \tfrac{\alpha}{\lambda^\alpha} y^{\alpha-1} \exp\left\{-(y/\lambda)^\alpha\right\}, \quad x \geq 0, \lambda > 0, \alpha > 0.$$

The Weibull distribution includes the exponential as special case when $\alpha = 1$. *The log likelihood for the Weibull(*λ, α*) model is*

$$\ell(\lambda, \alpha) = n\ln(\alpha) - n\alpha\ln(\lambda) + (\alpha - 1)\sum_{i=1}^n \ln y_i - \lambda^{-\alpha}\sum_{i=1}^n y_i^\alpha.$$

The gradient of this function is easily obtained by differentiation

$$\tfrac{\partial\ell(\lambda,\alpha)}{\partial\lambda} = -\tfrac{n\alpha}{\lambda} + \alpha\lambda^{-\alpha-1}\sum_{i=1}^n y_i^\alpha,$$

$$\tfrac{\partial\ell(\lambda,\alpha)}{\partial\alpha} = \tfrac{n}{\alpha} - n\ln(\lambda) + \sum_{i=1}^n \ln y_i - \sum_{i=1}^n \left(\tfrac{y_i}{\lambda}\right)^\alpha \times \ln\left(\tfrac{y_i}{\lambda}\right).$$

R demo to follow

Likelihood ratio statistic

- ☐ Suppose that likelihood inference for model A is OK, so $\widehat{\theta}_A \dot{\sim} \mathcal{N}\{\theta_A, \jmath_A(\widehat{\theta}_A)^{-1}\}$.
- ☐ Model $f_B(y)$ is **nested** within model $f_A(y)$ if A reduces to B on restricting some parameters:
  - for example, $f_B \equiv \mathcal{N}(0, \sigma^2)$ is nested within $f_A \equiv \mathcal{N}(\mu, \sigma^2)$, because B is obtained by setting $\mu = 0$ in A;
  - the maximised log likelihoods satisfy $\widehat{\ell}_A \geq \widehat{\ell}_B$, because the maximisation for A is over a larger set than for B.
- ☐ The **deviance** for model A is defined to be $D_A = \text{const} - 2\widehat{\ell}_A$, and then $D_B > D_A$.
- ☐ The **likelihood ratio statistic** for comparing A and B is

$$W = 2(\widehat{\ell}_A - \widehat{\ell}_B) = D_B - D_A.$$

- ☐ If model B is true and the models have p_A and p_B parameters, then

$$W \dot{\sim} \chi^2_{p_A - p_B}.$$

- ☐ The deviance is often used to compare models, and so is the **Akaike information criterion**

$$\text{AIC} = 2p_A - 2\widehat{\ell}_A,$$

with smaller values of both D_A and AIC being preferred.

Profile log likelihood

- ☐ Split $\theta = (\psi, \lambda)$ into a **parameter of interest** ψ and a **nuisance parameter** λ that are variation independent, i.e., $(\psi, \lambda) \in \Theta_\psi \times \Theta_\lambda$, and write the overall MLE as $\widehat{\theta} = (\widehat{\psi}, \widehat{\lambda})$.
- ☐ A $(1-2\alpha)$ confidence region for ψ can be based on the **profile log likelihood**

$$\ell_{\mathrm{p}}(\psi) = \max_{\lambda \in \Theta_\lambda} \ell(\psi, \lambda) = \ell(\psi, \widehat{\lambda}_\psi),$$

and is

$$\left\{\psi \in \Theta_\psi : 2\{\ell(\widehat{\psi}, \widehat{\lambda}) - \ell(\psi, \widehat{\lambda}_\psi)\} \le \chi^2_{\dim\psi}(1-2\alpha)\right\}.$$

- ☐ When ψ is scalar, this yields

$$\left\{\psi \in \Theta_\psi : \ell(\psi, \widehat{\lambda}_\psi) \ge \ell(\widehat{\psi}, \widehat{\lambda}) - \tfrac{1}{2}\chi^2_1(1-2\alpha)\right\},$$

and $\chi^2_1(0.95) = 3.84$, $\chi^2_1(0.99) = 6.63$ and $\chi^2_1(0.999) = 10.83$.
- ☐ Such intervals are preferable to the standard interval $\widehat{\psi} \pm z_\alpha v_{\psi\psi}^{1/2}$ when the distribution of $\widehat{\psi}$ is asymmetric, but require more computation, since they involve many maximisations of ℓ.

Profile log likelihood: back to Example 1

Consider the shape parameter $\psi \equiv \alpha$ as parameter of interest, and the scale λ as nuisance parameter. Using the gradient,

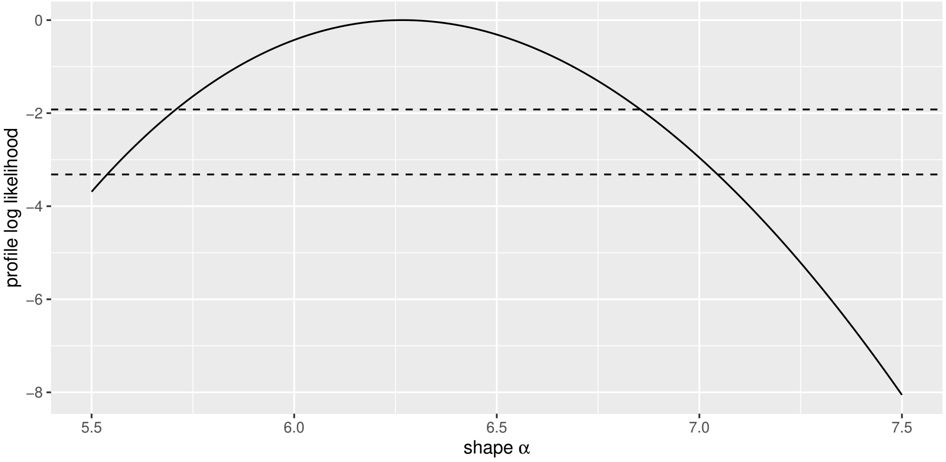
$$\frac{\partial \ell(\lambda, \alpha)}{\partial \lambda} = -\frac{n\alpha}{\lambda} + \alpha\lambda^{-\alpha-1}\sum_{i=1}^n y_i^\alpha$$

we find that the value of the scale that maximises the log likelihood for given α is

$$\widehat{\lambda}_\alpha = \left(\frac{1}{n}\sum_{i=1}^n y_i^\alpha\right)^{1/\alpha}.$$

and plugging in this value gives a function of α alone, thereby also reducing the optimisation problem for the Weibull to a line search along $\ell_{\mathrm{p}}(\alpha)$.

Profile log likelihood: back to Example 1



Regular models

The above approximate distributions hold under **regularity conditions**:

(C1) the true value θ^0 of θ is interior to the parameter space $\Theta \subset \mathbb{R}^p$ for some fixed p;

(C2) the densities defined by any two distinct values of θ are different;

(C3) there is a neighbourhood $\mathcal{N}$ of θ^0 within which the first three derivatives of ℓ with respect to θ exist almost surely, and for $r, s, t = 1, \dots, d$ satisfy

$$|\partial^3 \log f(Y;\theta)/\partial\theta_r\partial\theta_s\partial\theta_t| < m(Y),$$

with $\mathrm{E}_g\{m(Y)\} < \infty$; and

(C4) the first two **Bartlett identities** hold within $\mathcal{N}$, i.e., for $\theta \in \mathcal{N}$,

$$\begin{aligned} 0 &= \nabla_\theta \int f(y;\theta)\,\mathrm{d}y = \int \nabla_\theta \log f(y;\theta) \times f(y;\theta)\,\mathrm{d}y, \\ 0 &= \nabla_\theta^2 \int f(y;\theta)\,\mathrm{d}y \\ &= \int \nabla_\theta^2 \log f(y;\theta) \times f(y;\theta)\,\mathrm{d}y + \int \nabla_\theta \log f(y;\theta)\, \nabla_\theta^{\mathrm{T}} \log f(y;\theta) \times f(y;\theta)\,\mathrm{d}y, \end{aligned}$$

where $\nabla_\theta \cdot = \partial \cdot /\partial\theta$ and $\nabla_\theta^2 \cdot = \partial^2 \cdot /\partial\theta\partial\theta^{\mathrm{T}}$.

Regularity conditions

- ☐ These conditions are sufficient (but not necessary) to prove theorems giving the limiting distributions for $\widehat{\theta}$ and W as the sample size (or more generally some measure of the information in the data) goes to infinity.
- ☐ Why they are needed:
  - (C1) ensures that $\widehat{\theta}$ can be 'on all sides' of θ^0 in the limit — if it fails, then any limiting distribution cannot be normal;
  - (C2) is essential for consistency, otherwise $\widehat{\theta}$ might not converge to a unique limit;
  - (C3) is needed to bound terms of a Taylor series — can be replaced by other conditions; and
  - (C4) ensures that $\widehat{\theta}$ is consistent for θ^0 and that the asymptotic variance of $\widehat{\theta}$ is the inverse Fisher information $\imath(\theta^0)^{-1}$.
- ☐ In some of the models arising later, (C4) may fail (or be close to failing), because the support of the data depends on a parameter.

Model checking

- ☐ To check whether an assumed model for data is suitable we often use graphs, because
  - they show the data directly;
  - unexpected features may be visible.
- ☐ If the data are assumed to be a random sample $y_1, \ldots, y_n \overset{\text{iid}}{\sim} F$, with order statistics

$$y_{(1)} \le y_{(2)} \le \cdots \le y_{(n)},$$

then a **quantile-quantile plot (Q-Q plot)** shows

$$(F^{-1}\{1/(n+1)\}, y_{(1)}), \ldots, (F^{-1}\{n/(n+1)\}, y_{(n)})$$

where $F^{-1}\{1/(n+1)\}, \ldots, F^{-1}\{n/(n+1)\}$ are called the **plotting positions** for F.

- ☐ Ideally this plot
  - should be a straight line if the assumption is correct;
  - shows model failure as systematic curvature;
  - shows outliers as isolated points,

  but variation can be expected even if the assumption is correct!
- ☐ In practice F is often unknown and must be replaced by an estimate $\widehat{F}$.

1.3 Example

Poisson process in the line

- ☐ A simple model for times of events (earthquakes, typhoons, heatwaves, ...).
- ☐ Write $N(\mathcal{A})$ for the number of events in a set $\mathcal{A} \subset [0, t_0]$, where t_0 is fixed and known.
  - let $N(w, w+t)$ denote the number of events in $(w, w+t]$, and set
  - $N(t) = N(0, t)$, $t > 0$.
- ☐ Let $\dot{\mu}(t)$ be a non-negative **intensity function** giving the rate of events around t (picture!), and whose integral $\mu(0, t_0) = \int_0^{t_0} \dot{\mu}(t)\,\mathrm{d}t < \infty$, and suppose that
  - events in disjoint subsets of $[0, t_0]$ are independent, i.e., $N(\mathcal{A}_1)$ is independent of $N(\mathcal{A}_2)$ whenever $\mathcal{A}_1 \cap \mathcal{A}_2 = \emptyset$;
  - $\mathrm{P}\{N(t, t+\delta t) = 0\} = 1 - \dot{\mu}(t)\delta t + o(\delta t)$ for small δt; and
  - $\mathrm{P}\{N(t, t+\delta t) = 1\} = \dot{\mu}(t)\delta t + o(\delta t)$ for small δt.
- ☐ The last two properties imply that

$$\mathrm{P}\{N(t, t+\delta t) > 1\} = o(\delta t) \to 0, \quad \delta t \to 0,$$

so the process is **orderly**: multiple occurrences at the same time cannot occur.

Poisson process in the line, II

- ☐ Under these assumptions,
  - the **void probability** of the set $(w, w+t]$ is

$$\mathrm{P}\left\{N(w, w+t) = 0\right\} = \exp\left\{-\mu(w, w+t)\right\},$$

  - the random **waiting time** T from w to the next event has PDF

$$f_T(t) = \dot{\mu}(w+t)\exp\left\{-\mu(w, w+t)\right\}, \quad t > 0,$$

  i.e., $\mu(w, w+T) \sim \exp(1)$, independent of waiting times on other intervals;
  - the joint density of events at $0 < t_1 < \cdots < t_n < t_0$ is

$$\exp\left\{-\mu(0, t_0)\right\} \prod_{j=1}^{n} \dot{\mu}(t_j), \quad 0 < t_1 < \cdots < t_n < t_0,$$

  - and $N(0, t_0) \sim \mathrm{Poiss}\{\mu(0, t_0)\}$.
- ☐ Hence if the sets $\mathcal{A}_1, \mathcal{A}_2, \ldots$ are disjoint, the corresponding numbers of events satisfy

$$N(\mathcal{A}_j) \overset{\text{ind}}{\sim} \mathrm{Poiss}\{\mu(\mathcal{A}_j)\}.$$

Note: Poisson process in the line

☐ To find the probability of no events in $(w, w+t]$ we divide it into k subintervals of length $\delta t = t/k$, and then let $\delta t \to 0$. Then as events in disjoint intervals are independent,

$$\begin{aligned} \mathrm{P}\left\{N(w, w+t)=0\right\} &= \prod_{i=0}^{k-1} \mathrm{P}\left[N\left\{w+i\delta t, w+(i+1)\delta t\right\}=0\right] \\ &= \prod_{i=0}^{k-1}\left\{1-\dot{\mu}(w+i\delta t)\delta t+o(\delta t)\right\} \end{aligned}$$

has negative logarithm

$$-\sum_{i=0}^{k-1}\log\left\{1-\dot{\mu}(w+i\delta t)\delta t+o(\delta t)\right\} = \sum_{i=0}^{k-1}\dot{\mu}(w+i\delta t)\delta t+o(k\delta t) \to \int_w^{w+t}\dot{\mu}(u)\,\mathrm{d}u = \mu(w, w+t),$$

where the limit follows because as $\delta t \to 0$ with t fixed, $o(k\delta t) = k\,o(\delta t)/\delta t \to 0$. Hence

$$\mathrm{P}\left\{N(w, w+t)=0\right\} = \exp\left\{-\mu(w, w+t)\right\}, \quad t>0.$$

☐ The time T after w to the next event exceeds t if and only if $N(w, w+t)=0$, so

$$\mathrm{P}(T>t) = \mathrm{P}\left\{N(w, w+t)=0\right\} = \exp\left\{-\mu(w, w+t)\right\},$$

and thus T has PDF

$$f_T(t) = -\frac{\mathrm{dP}\left\{N(w, w+t)=0\right\}}{\mathrm{d}t} = \dot{\mu}(w+t)\exp\left\{-\mu(w, w+t)\right\}. \quad t>0.$$

Put another way, $\mu(w, w+T) \sim \exp(1)$.

☐ If events in $(0, t_0]$ have been observed at times $t_1, \ldots, t_n$, where $0<t_1<\cdots<t_n<t_0$, then, as events in disjoint sets are independent, the joint probability density of the data is

$$\dot{\mu}(t_1)e^{-\mu(0,t_1)} \times \dot{\mu}(t_2)e^{-\mu(t_1,t_2)} \times \cdots \times \dot{\mu}(t_n)e^{-\mu(t_{n-1},t_n)} \times e^{-\mu(t_n,t_0)},$$

where the final term is the probability of no events in $(t_n, t_0]$. This joint density reduces to

$$\exp\left\{-\mu(0, t_0)\right\}\prod_{j=1}^{n}\dot{\mu}(t_j), \quad 0<t_1<\cdots<t_n<t_0. \qquad (1)$$

note 1 of slide 36

Poisson process in the line, III

- Without further assumptions on μ, the Poisson process is a nonparametric model.
- The simplest parametric version is the **homogeneous Poisson process**, with $\dot\mu(t) \equiv \dot\mu$ a positive constant, under which the times between events are independent with PDF

$$f_T(t) = \dot\mu(t)\exp\left\{-\mu(w, t+w)\right\} = \dot\mu\exp\left(-\dot\mu t\right), \quad t > 0,$$

  i.e., the intervals $T_1, \dots, T_n \overset{\text{iid}}{\sim} \exp(\dot\mu)$.
- A simple parametric model for trend might set

$$\dot\mu(t) = \exp\left(\beta_0 + \beta_1 t\right), \quad \beta_0, \beta_1 \in \mathbb{R},$$

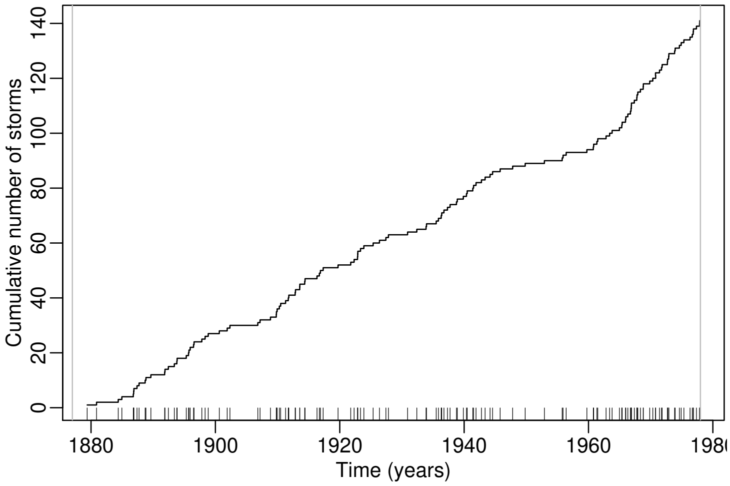
  which reduces to the homogeneous model when $\beta_1 = 0$.
- In principle we could model more complex trends by replacing $\beta_0 + \beta_1 t$ by linear combinations of basis functions,

$$\beta_0 + \beta_1 b_1(t) + \cdots + \beta_p b_p(t),$$

  e.g., using trigonometric functions for seasonality, but we must compute $\int_0^{t_0} \dot\mu(t)\,\mathrm{d}t$.
- Such models are linear exponential families, so theory from the second year applies ...

Cyclones

Times of major cyclonic storms striking the Bay of Bengal from 1877–1977; jittered vertically for visualisation (Mooley, 1981, *Sankhyā*). In November 1970, Cyclone Bhola, the deadliest storm in world history, occurred in the Bay of Bengal and killed around half a million people. It brought a storm surge estimated at 10.4m to the coast.



Cyclones II

- The storm times don't look very even, but perhaps that's just randomness ...
- Take $[0, t_0] \equiv$ [1 January 1877, 31 December 1977], so the t_j are measured in years after the start of 1877 and run up to $t_0 = 101$.
- Under the simplest possible model, the data are a homogeneous Poisson process with $n = 141$ events in $[0, 101]$. Then $\mu(t) = \dot{\mu}t$, so (writing $\lambda = \dot{\mu}$ for simpler notation) the likelihood is

  $$L(\lambda) = f(t_1, \dots, t_n; \lambda) = \exp\{-\mu(0, t_0)\} \prod_{j=1}^{n} \dot{\mu}(t_j) = \exp(-t_0\lambda)\lambda^n,$$

  giving maximised log likelihood, MLE and corresponding observed information

  $$\ell(\widehat{\lambda}) = -93.96,, \quad \widehat{\lambda} = n/t_0 = 141/101 \doteq 1.4 \text{ events/year}, \quad \jmath(\widehat{\lambda}) = n/\widehat{\lambda}^2 = t_0^2/n \doteq 72.3,$$

  and the approximate 95% confidence interval based on $\widehat{\lambda}$ has limits

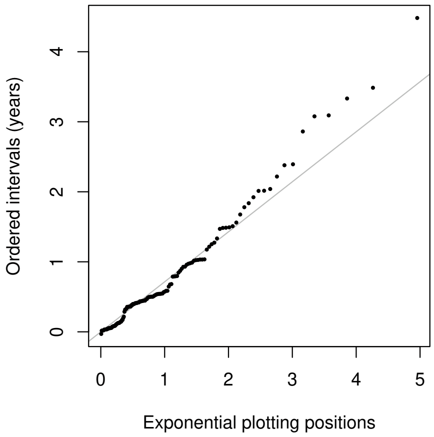
  $$\widehat{\lambda} \pm 1.96\jmath(\widehat{\lambda})^{-1/2} \doteq (1.17, 1.63) \text{ events/year}.$$

Cyclones III

```
# numerical optimisation of the likelihood of the simplest model
# bengal has data in units of years
load("bengal.dat")  # data available on Moodle page
bengal # look at event times

nlogL <- function(lambda, t, t0=101)
{ # negative log likelihood
  t0*lambda - log(lambda)*length(t)
}
(fit <- optim(par=c(1.2), fn=nlogL, hessian=T, t=bengal-1877, t0=101))
$par
[1] 1.395937
$value
[1] 93.95685
...
$hessian
      [,1]
[1,] 72.35818
```

Cyclones V

☐ Under this model, and setting $t_0 = 101$, the intervals $t_1 - t_0, t_2 - t_1, \dots t_n - t_{n-1} \overset{\text{iid}}{\sim} \exp(\lambda)$, so a QQ-plot of these intervals against exponential plotting positions should be a straight line.



The grey line corresponds to $y = x/\widehat{\lambda}$.

Cyclones IV

☐ The QQplot shows departures from the exponential distribution (the larger values are systematically too big), so the basic model may be too simple.

☐ Let $\dot{\mu}(t) = \lambda \exp(t\beta)$, so $\mu(0, t_0) = \lambda(e^{t_0\beta} - 1)/\beta$, where $\beta > 0$ would lead to increases in the rate, and conversely.

☐ The code on the next slide fits this model and computes the standard errors, giving

$$\ell(\widehat{\lambda}, \widehat{\beta}) = -89.65, \quad \widehat{\lambda} = 0.88_{0.17}, \quad \widehat{\beta} = 0.0086_{0.0030}.$$

☐ The likelihood ratio statistic for comparing the models is

$$2\{-89.65 - (-93.96)\} = 8.62 \mathrel{\dot\sim} \chi^2_{2-1},$$

which gives (approximate) significance level 0.0034, fairly strong evidence of an increase in numbers of cyclones.

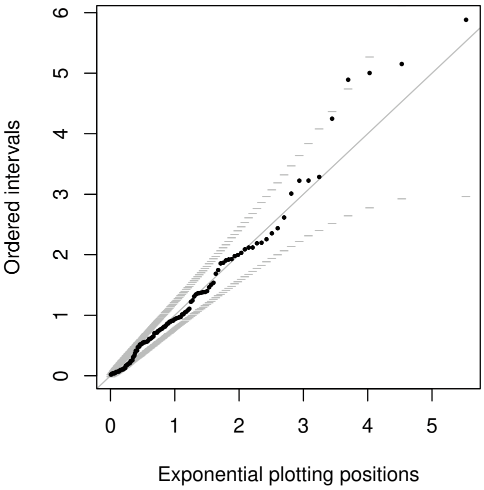
☐ Looking at the original data, we might query this model of smooth increase. As $\mu(w, w+T) \sim \exp(1)$, we could try a QQplot of

$$\widehat{\mu}(t_{j-1}, t_j) = \widehat{\lambda}\left(e^{\widehat{\beta}t_j} - e^{\widehat{\beta}t_{j-1}}\right)/\widehat{\beta}, \quad j = 1, \dots, n.$$

Cyclones V

```
# comparison of homogeneous model with log-linear trend
# bengal has data in units of years

nlogL <- function(th, t, t0=101)
{ # negative log likelihood
  int <- th[1]*(exp(t0*th[2])-1)/th[2]
  int - sum( log(th[1]) + t*th[2] )
}

(fit <- optim(par=c(1.4,0.1), fn=nlogL, hessian=T, t=bengal-1877, t0=101 ))
$par
[1] 0.881341438 0.008567385

$value
[1] 89.64509

...

$hessian
          [,1]        [,2]
[1,]  181.5231   9273.082
[2,] 9273.0820 588285.232

(se <- sqrt(diag(solve(fit$hessian))))
[1] 0.168186263 0.002954354
```

Cyclones VI



The grey line corresponds to $y = x$, and the grey minus signs show the 95% ranges for individual order statistics.

Cyclones VII

- ☐ Visual guidance about 'acceptable' variation is useful . . .
- ☐ The new QQplot is better but even if they are individually (mostly) inside the limits,
  - – the largest intervals still seem too long, and
  - – the smallest intervals seem too short?
- ☐ The original data variation looks more like a change in slope around 1960 than a smooth increase in rate . . . .
- ☐ Maybe we could explain this variation by allowing
  - – seasonality?
  - – (random?) changes in the rate?
  - – external climatic factors such as the El Niño-Southern Oscillation (ENSO)?
- ☐ The latter would be preferable — if we could predict how climate change would influence the ENSO, we could then make an educated guess about the likely future frequencies of cyclones . . . If ENSO does influence the occurrence of cyclones, then it would be a so-called **causal variable**, unlike time, or pure randomness.

Summary

- ☐ The course will mostly concern statistical modelling for rare events that could have big impacts.
- ☐ We've now:
  - – seen some basic modelling ideas that will be used repeatedly;
  - – met the simplest Poisson process for the occurrence of random point events;
  - – applied that model to a small dataset.
- ☐ The Poisson process is a key ingredient in rare event modelling, so next week we shall look at it in more generality.

2 Statistics of Extremes

slide 47

2.1 Basic Methods for Maxima

slide 48

Probability framework for maxima

☐ Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$ and define the maximum $M_n = \max(X_1, \ldots, X_n)$, giving

$$\begin{aligned} \mathrm{P}(M_n \leq x) &= \mathrm{P}(X_1 \leq x, \ldots, X_n \leq x) \\ &= \mathrm{P}(X_1 \leq x) \times \cdots \times \mathrm{P}(X_n \leq x) \\ &= F(x)^n. \end{aligned}$$

☐ F is unknown, so approximate F^n by some limit distribution, but as $n \to \infty$,

$$F(x)^n \to \begin{cases} 0, & F(x) < 1, \\ 1, & F(x) = 1, \end{cases}$$

so $M_n \xrightarrow{D} x^*$, where $x^* = \sup\{x : F(x) < 1\}$ is the upper support point of F. Not a useful limit.

☐ All the argument below applies equally to minima, because

$$\min(X_1, \ldots, X_n) = -\max(-X_1, \ldots, -X_n).$$

Our general discussion is for maxima, and we make this transformation without comment when we model minima.

slide 49

Central limit theorem

☐ Recall that to get convergence of a sum $\sum_{j=1}^n X_j$ in the central limit theorem, we consider the centred and scaled quantities

$$Z_n = \frac{\sum_{j=1}^n X_j - b_n}{a_n},$$

and with the choices $b_n = n\mu$ and $a_n = n^{1/2}\sigma$, where μ and σ are the mean and standard deviation of X_j, assumed finite, we then obtain

$$Z_n \xrightarrow{D} Z \sim \mathcal{N}(0, 1), \quad n \to \infty.$$

☐ This suggests studying the convergence of the centered and scaled quantities

$$\frac{M_n - b_n}{a_n},$$

for suitable series $\{b_n\} \subset \mathbb{R}$ and $\{a_n\} \subset \mathbb{R}_+$.

slide 50

Extremal Types Theorem

Theorem 2 (Extremal types) *Let* $M_n = \max(X_1, \ldots, X_n)$ *be the maximum of a random sample* $X_1, \ldots, X_n$. *If sequences of real numbers* $\{a_n\} > 0$ *and* $\{b_n\}$ *can be chosen so that the centred and scaled sample maximum,* $(M_n - b_n)/a_n$, *has a non-degenerate limiting distribution* G, *then this must be the* generalized extreme-value distribution (GEV),

$$G(x) = \begin{cases} \exp\left[-\{1 + \xi(x-\eta)/\tau\}_+^{-1/\xi}\right], & \xi \neq 0, \\ \exp\left[-\exp\{-(x-\eta)/\tau\}\right], & \xi = 0, \end{cases} \quad x \in \mathbb{R}, \tag{2}$$

where $a_+ = \max(a, 0)$ *for any real* a, *and with* $\xi, \eta \in \mathbb{R}$ *and* $\tau > 0$. *Put another way,* $(M_n - b_n)/a_n \xrightarrow{D} Z$ *as* $n \to \infty$, *where* Z *has distribution function* G.

☐ The 'types', which arise for $\xi = 0$, $\xi > 0$ and $\xi < 0$, are now usually subsumed into (2), and are discussed below.

☐ This theorem provides a single distribution for maxima, and is in some ways stronger than the CLT, since we only assume that linear rescaling can result in a non-degenerate distribution, without other assumptions on F.

Minima

☐ In general discussion we consider maxima and large values — what about minima and small values?

☐ As

$$Y = \min(X_1, \ldots, X_m) = -\max(-X_1, \ldots, -X_m) = -Y^-,$$

say, then

$$\tilde{G}(y) \approx \mathrm{P}(Y \leq y) = \mathrm{P}(Y^- \geq -y) \approx 1 - G(-y),$$

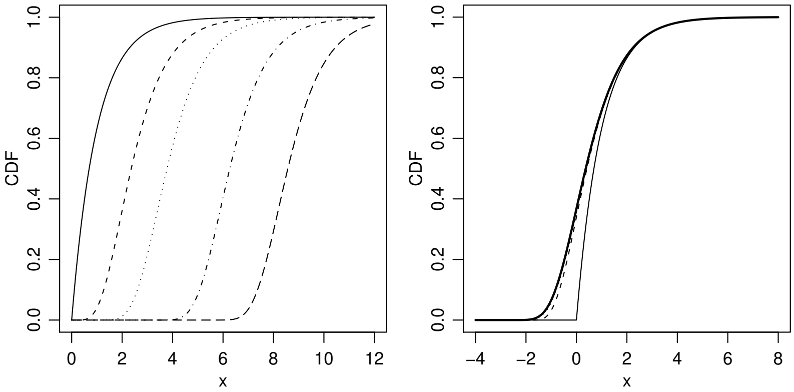
where G is the GEV approximation for $\max(-X_1, \ldots, -X_m)$. Hence

$$\tilde{G}(y; \tilde{\eta}, \tilde{\tau}, \tilde{\xi}) = 1 - G(-y; -\eta, \tau, \xi),$$

where G is estimated from the negative minima.

Examples

Example 3 *Find sequences* $\{a_m\}$ *and* $\{b_m\}$ *such that maxima of independent variables from the (a) uniform, (b) exponential, and (c) Pareto distributions have non-degenerate limiting distributions.*



Distributions of maxima (left) and renormalized maxima (right) of $m = 1$, 7, 30, 365, 3650 standard exponential variables (from left to right), with limiting Gumbel distribution (heavy).

Note I to Example 3

☐ Note that

$$\mathrm{P}\left\{(M-b_m)/a_m \le y\right\} = \mathrm{P}\left\{M \le b_m + a_m y\right\} = F^m(b_m + a_m y),$$

and we need to choose a_m and b_m such that this has a limit as $m \to \infty$. We saw from Theorem 2 that a limit $G(y) = \exp\{-\Lambda(y)\}$, so it is equivalent to identify Λ.

☐ (a) In the uniform case, $F(x) = x$ for $x \in [0,1]$. Provided $0 \le b_m + a_m y \le 1$, we therefore have

$$F(b_m + a_m y)^m = (b_m + a_m y)^m,$$

so if we set $b_m = 1$, $a_m = 1/m$ and $-m \le y \le 0$, we have $(b_m + a_m y)^m \to e^y$. Hence

$$\Lambda(y) = \begin{cases} -y, & y \le 0, \\ 0, & y > 0, \end{cases}$$

i.e., $\Lambda(y) = (-y)_+$, where $a_+ = \max(a, 0)$ for real a. Clearly Λ is decreasing on $(-\infty, 0)$. Hence

$$G(y) = \exp\{-\Lambda(y)\} = \begin{cases} e^y, & y \le 0, \\ 1, & y > 0, \end{cases}$$

which is the distribution function of $-W$, where $W \sim \exp(1)$. It is straightforward to check that this G is (2) with $\eta = 1$, $\tau = 1$ and $\xi = -1$.

☐ (b) In the exponential case, $F(x) = 1 - \exp(-x)$ for $x > 0$. Provided $b_m + a_m y > 0$,

$$F(b_m + a_m y)^m = \left[1 - \exp\left\{-(b_m + a_m y)\right\}\right]^m,$$

so if we set $b_m = \log m$ and $a_m = 1$, and if $y > -\log m$,

$$G(y) = \lim_{m\to\infty} F(b_m + a_m y)^m = \lim_{m\to\infty} \left(1 - \frac{e^{-y}}{m}\right)^m = \exp\left(-e^{-y}\right), \quad y \in \mathbb{R},$$

which is (2) with $\eta = 0$, $\tau = 1$ and $\xi = 0$. Here $\Lambda(y) = e^{-y}$ with support in $\mathbb{R}$.

☐ (c) In the Pareto case, $F(x) = 1 - x^{-\alpha}$ for $x > 1$ and $\alpha > 0$. Provided $b_m + a_m y > 1$, we have

$$F(b_m + a_m y)^m = \left\{1 - (b_m + a_m y)^{-\alpha}\right\}^m$$

so if we set $b_m = 0$ and $a_m = m^{1/\alpha}$, and if $y > m^{-1/\alpha}$, we have

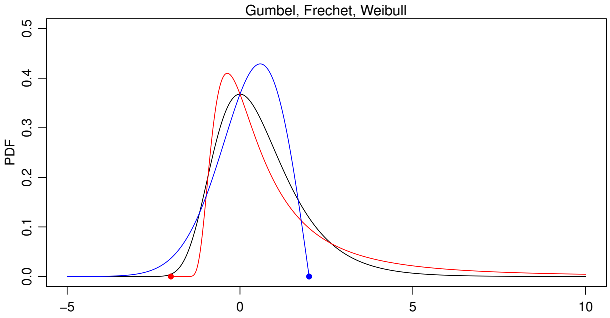
$$G(y) = \lim_{m\to\infty} F(b_m + a_m y)^m = \lim_{m\to\infty} \left(1 - \frac{y^{-\alpha}}{m}\right)^m = \exp\left(-y^{-\alpha}\right), \quad y \ge 0,$$

which is (2) with $\eta = 1$, $\tau = 1/\alpha$ and $\xi = 1/\alpha$. In this case

$$\Lambda(y) = \begin{cases} \infty, & y \le 0, \\ y^{-\alpha}, & y > 0. \end{cases}$$

☐ Note that we have not shown that the three limits above are the only ones possible, just that we can choose a_m and b_m to obtain these limits.

GEV and ‘three types’



☐ ξ is a shape parameter determining the rate of tail decay, with:

- $\xi > 0$ giving the heavy-tailed **Fréchet (Type II)** density with support $(\eta - \tau/\xi, \infty)$;
- $\xi = 0$ giving the light-tailed **Gumbel (Type I)** density, with support $\mathbb{R}$;
- $\xi < 0$ giving the short-tailed **(reverse) Weibull (Type III)** density, with support $(-\infty, \eta - \tau/\xi)$.

☐ The usual Weibull distribution gives a model for minima.

☐ η and τ are location and scale parameters (not so crucial as the shape parameter ξ).

Properties of the GEV

☐ **Support**: If $\xi > 0$ then $Y > \eta - \tau/\xi$, and if $\xi < 0$ then $Y < \eta - \tau/\xi$.

☐ **Moments**: $\mathrm{E}(Y^r)$ exists only if $\xi < 1/r$, so the mean exists only if $\xi < 1$, the variance only if $\xi < 1/2$, etc. In applications (particularly in finance) some moments may not exist.

☐ **Quantiles**: solve $G(y) = p$ for $0 < p < 1$, but usually we use the **return levels** given by solving $G(y_p) = 1 - p$ (next slide) — so y_p is the $(1-p)$ quantile (careful!)

☐ **Maximum likelihood estimation**: is regular only if $\xi > -1/2$. Not usually a problem in applications.

☐ **Max-stability**: if $Y_1, \ldots, Y_T \overset{\text{iid}}{\sim}$ GEV(η, τ, ξ) then $\max(Y_1, \ldots, Y_T) \sim$ GEV(η_T, τ_T, ξ_T), i.e.,

$$G(y; \eta, \tau, \xi)^T = G(y; \eta_T, \tau_T, \xi_T)$$

where

$$\eta_T = \begin{cases} \eta + \tau(T^\xi - 1)/\xi, & \xi \neq 0, \\ \eta + \tau \log T, & \xi = 0, \end{cases} \qquad \tau_T = \tau T^\xi, \quad \xi_T = \xi,$$

so the distribution type and shape parameter are unchanged by taking maxima.

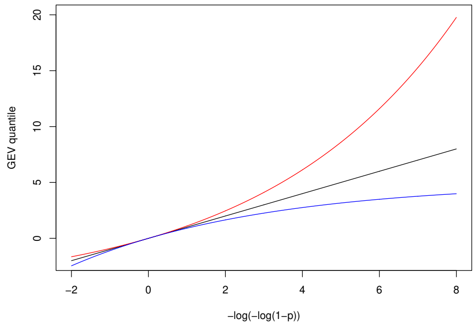
☐ In fact the GEV is the only max-stable class of distributions.

Quantiles and return levels

- Define the **return level** associated to the **return period** $T = 1/p$ (blocks) as

$$y_p = \eta + \tau \frac{\{-\log(1-p)\}^{-\xi} - 1}{\xi}, \quad 0 < p < 1,$$

  i.e., the solution to $G(y_p) = 1 - p = 1 - 1/T$.
- Informally, y_p is the level expected to be exceeded once every T blocks.
- The plot below compares the quantiles for $\xi = -0.2$ (blue) and $\xi = 0.2$ (red) with the Gumbel quantiles (black).



Statistical approach

- Assume **background data** $x_1, x_2, \ldots$ are IID realisations from some continuous distribution F to which the GEV approximation applies.
- Take maxima $y = \max(x_1, \ldots, x_m)$ of blocks of size m from the background data.
  - for environmental time series, typically $m \approx 365$ for annual maxima, $m \approx 30$ for monthly maxima, ...
  - in finance, typically $m = 250$ for annual maxima, $m = 20$ for monthly maxima, ...
- Suppose the resulting series of maxima $y_1, \ldots, y_n$ are IID GEV(η, τ, ξ).

Exploratory plot for maxima

☐ For GEV with $\xi \neq 0$,

$$F^{-1}\{j/(n+1)\} = G^{-1}\{j/(n+1)\} = \eta + \frac{\tau}{\xi}\left[\{-\log(j/(n+1))\}^{-\xi} - 1\right], \quad j = 1, \dots, n,$$

which involves all three parameters (which must therefore be estimated!).

☐ By taking $\eta = 0$, $\tau = 1$, and the limit as $\xi \to 0$, we get the **Gumbel plotting positions**

$$-\log\left[-\log\{j/(n+1)\}\right], \quad j = 1, \dots, n.$$

☐ Plot ordered block maxima $y_{(1)} \leq \cdots \leq y_{(n)}$ against Gumbel plotting positions.

☐ After allowing for noise,
- convex shape suggests $\xi > 0$,
- straight line suggests $\xi \approx 0$,
- concave shape suggests $\xi < 0$.

☐ Outliers, heavy rounding or other issues with data should be visible.

☐ Comparison of these plots for different block sizes may also suggest a minimum block size for the GEV to apply.

Estimation

☐ Mostly we use maximum likelihood estimation according to the recipe on slide 24.

☐ This has theoretical and practical advantages:
- it is efficient (has the smallest possible variance) in large samples — in regular situations;
- likelihood ratio tests are generally fairly powerful;
- there's a simple recipe to follow — write down the likelihood and maximise it — which works in many situations;
- lots of code already exists and can be readily applied. Hooray!

☐ Other methods of estimation are also used:
- method of moments estimation to get initial values for maximising a likelihood;
- probability-weighted (or L-) moments estimation is widely used in hydrology and some other domains, because it can beat ML estimation in small samples;
- in more complex problems the likelihood can be awkward, and then other methods must be used.

Moment estimation

☐ Define moments for random variable X as $\mu'_r = \mathrm{E}(X^r)$ for $r = 1, \dots$ (if μ'_r finite).

☐ If X depends on $p \times 1$ parameter vector θ, then $\mu'_r = \mu'_r(\theta)$, and we estimate θ by solving the equations

$$\mu'_r(\theta) = n^{-1} \sum_j X_j^r, \quad r = 1, \dots, p.$$

☐ Moment estimators usually simple but inefficient (variance larger than for competing approaches)

☐ For GEV, μ'_r exists only if $\xi r < 1$, so must have $\xi < 1/3$ to estimate all three parameters, and $\xi < 1/6$ for them to have finite variances. Much too restrictive for use in practice.

☐ Useful for finding starting-values for ML estimation.

L-moment estimation

☐ Define **probability-weighted moments** as $\mu'_{r,s,t} = \mathrm{E}[X^r F(X)^s \{1 - F(X)\}^t]$ for $r, s, t = 0, 1, 2, \dots$, or equivalently

$$\mu'_{r,s,t} = \int_0^1 x_p^r p^s (1-p)^t dp, \quad \text{where} \quad F(x_p) = p;$$

ordinary moments have $s = t = 0$.

☐ Use $\mu'_{1,s,0}$ for $s = 0, 1, \dots$ to fit GEV and $\mu'_{1,0,t}$ with $t = 0, 1$ to fit GPD.

☐ In practice estimate the **L-moments** $\lambda_1, \lambda_2, \dots$, linear combinations of the μ's, by expressions like

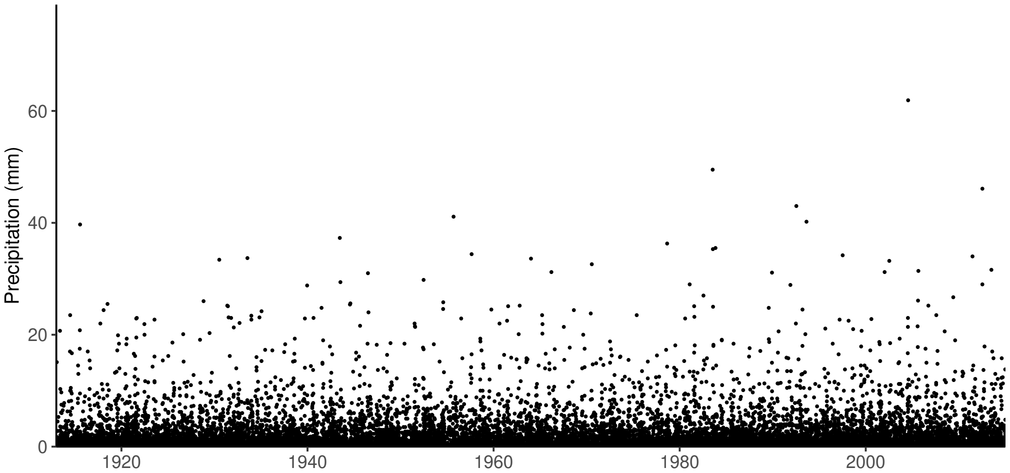
$$\widehat{\lambda}_1 = \frac{1}{\binom{n}{1}} \sum_{j=1}^n X_{(j)}, \quad \widehat{\lambda}_2 = \frac{1}{2\binom{n}{2}} \sum_{j=1}^n \left\{ \binom{j-1}{1} - \binom{n-j}{1} \right\} X_{(j)}, \quad \dots,$$

☐ L-moment estimators of η, τ and ξ based on $\widehat{\lambda}_1$, $\widehat{\lambda}_2$ and $\widehat{\lambda}_3$ are linear in the observations, so are more robust than the ordinary moment estimators.

☐ Have good small-sample properties, but don't generalise to complex settings.

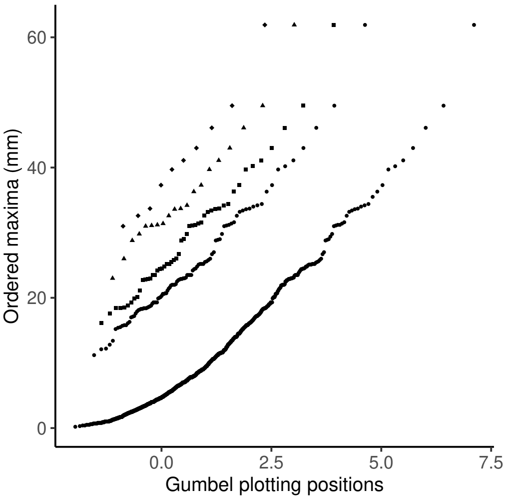
Abisko daily rainfall data

☐ Daily precipitation in Abisko, in northern Sweden, 1913–2014. The largest value is 61.9 mm, but many values are zero and most of the positive values are quite small.



Abisko block maxima

☐ Gumbel QQplot of maxima for blocks of lengths (from bottom) one month and one, two, five and ten years.



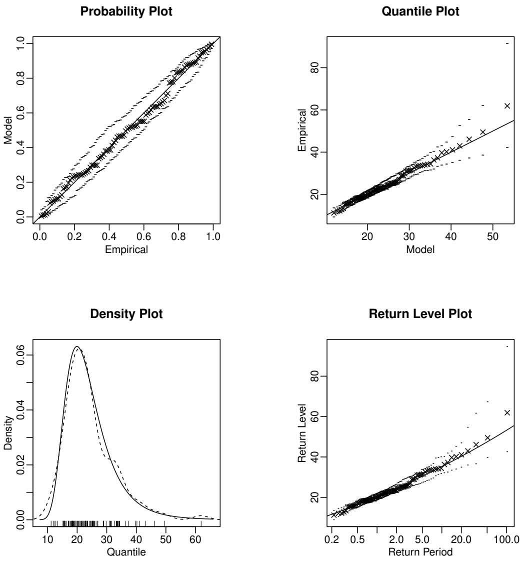
Abisko annual maxima

☐ QQplot suggests stability from one year onwards, with slight convexity, so let's fit the GEV to annual maxima:

```
library(evd)
(fit <- fgev(year.max))

Call: fgev(x = year.max)
Deviance: 691.9509

Estimates
     loc     scale     shape
20.40530   5.84596   0.08353

Standard Errors
    loc    scale    shape
0.64854  0.48317  0.07193

Optimization Information
  Convergence: successful
  Function Evaluations: 27
  Gradient Evaluations: 7
```

Abisko annual maxima

☐ Let's check the fit using `plot(fit)`:



Commentary

☐ These (horrible!) plots use the fitted GEV CDF $\widehat{G} \equiv G(\cdot; \widehat{\eta}, \widehat{\tau}, \widehat{\xi})$ and are the

- **probability plot** showing $\left\{(j/(n+1), \widehat{G}(y_{(j)})) : j = 1, \dots, n\right\}$, which should be a straight line of unit gradient if $\widehat{G}$ is a good fit;
- **quantile plot** showing $\left\{(\widehat{G}^{-1}\{j/(n+1)\}, y_{(j)}) : j = 1, \dots, n\right\}$, which should be a straight line of unit gradient if $\widehat{G}$ is a good fit;
- **return level plot** showing (solid line) $(-1/\log(1-p), \widehat{G}^{-1}(1-p))$, for $0 < p < 1$, and the points $\{(-1/\log\{j/(n+1)\}, y_{(j)}) : j = 1, \dots, n\}$, which should lie on the line if $\widehat{G}$ is a good fit;
- **density plot** showing a kernel density estimate based on $y_1, \dots, y_n$ (shown by the rug) and the fitted GEV density.

☐ Some of the plots have pointwise 95% limits for individual points.

☐ They show essentially the same information but on different scales to highlight different aspects of the fit.

☐ In this case the fit seems reasonable.

2.2 Basic Methods for Exceedances

Exceedance Theorem

Theorem 4 (Exceedance) *Let X be a random variable having distribution function F, and suppose that a function $c(u)$ can be chosen so that the limiting distribution of $(X-u)/c(u)$, conditional on $X > u$, is non-degenerate as u approaches the upper support value $x^* = \sup\{x : F(x) < 1\}$ of X. If such a limiting distribution exists, it must be of* generalized Pareto *form, i.e.,*

$$H(x) = \begin{cases} 1 - (1 + \xi x/\sigma)_+^{-1/\xi} & \xi \neq 0, \\ 1 - \exp(-x/\sigma), & \xi = 0, \end{cases} \quad x > 0, \tag{3}$$

where $\xi \in \mathbb{R}$ and $\sigma > 0$. Expression (3) *is commonly known as the* generalized Pareto distribution (GPD).

☐ There is a close connection with the ETT, which applies for maxima under the same conditions as the ET applies for exceedances, and with the same ξ.

☐ The GPD is a natural model for exceedances over high thresholds (and under low ones, using $1 - H(-x)$).

Example 5 *Find a limiting distribution for threshold exceedances for $Z \sim N(0,1)$. Recall that $1 - \Phi(z) \sim \phi(z)/z$ as $z \to \infty$.*

Note to Example 5

☐ Here $x^* = \infty$ and for large z we have $1 - \Phi(z) \sim \phi(z)/z$.

☐ By analogy with renormalising maxima we aim to find a function $c_u > 0$ such that

$$\lim_{u\to\infty} \mathrm{P}\{(Z-u)/c_u > x \mid Z > u\}$$

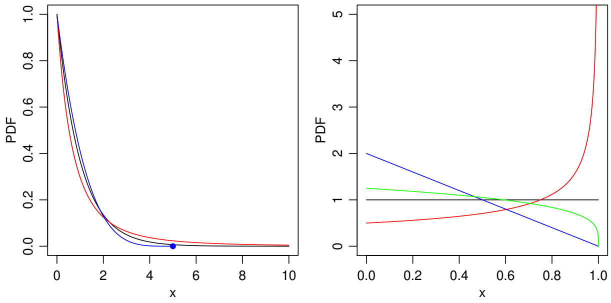
is non-degenerate. The hint gives that for fixed $x > 0$ and large u,

$$\begin{aligned}
\mathrm{P}\{(Z-u)/c_u > x \mid Z > u\} &= \frac{\mathrm{P}(Z > u + c_u x)}{\mathrm{P}(Z > u)} \\
&= \frac{1 - \Phi(u + c_u x)}{1 - \Phi(u)} \\
&\sim \frac{\phi(u + c_u x)/(u + c_u x)}{\phi(u)/u} \\
&= \frac{u}{u + c_u x} \exp\{u^2/2 - (u + c_u x)^2/2\} \\
&= \frac{1}{1 + c_u x/u} \exp(-c_u u x - c_u^2 x^2/2),
\end{aligned}$$

so if we choose $c_u = 1/u$ then the ratio tends to unity and the exponent tends to $-x$, i.e., the limiting distribution for an appropriately rescaled exceedance is standard exponential.

☐ If we had chosen $c_u = 1/(\sigma u)$ for any fixed $\sigma > 0$ we would have an exponential limit, with mean σ, as in (3), so we can think of the parameter σ as arising because we don't know the ideal scaling function.

Generalized Pareto distribution



☐ A flexible distribution whose density can take a variety of shapes.

☐ Left: exponential density ($\xi = 0$, black), heavy-tailed density ($\xi = 0.5$, red) and light-tailed density ($\xi = -0.2$, blue, with upper terminal shown); all have $\sigma = 1$.

☐ Right: densities with negative shape parameter and upper terminal at $x = 1$, with $\xi = -1$ (black), $\xi = -2$ (red), $\xi = -0.5$ (blue) and $\xi = -0.8$ (green).

Stability and threshold choice

☐ This approach relying on exceedances is termed the **peaks over threshold (POT)** approach. It is easy to explain and understand but requires the choice of a threshold u.

- u too low will lead to bias (model inappropriate) and u too high will increase variance (too few exceedances).

☐ If $X \sim \mathrm{GPD}(\sigma, \xi)$, then $X - u \mid X > u \sim \mathrm{GPD}(\sigma + \xi u, \xi)$, and this implies that

$$\mathrm{E}(X - u \mid X > u) = \frac{\sigma + \xi u}{1 - \xi}, \quad \xi < 1,$$

so a **mean excess plot (or mean residual life plot)** of

$$\frac{\sum_j (x_j - u) I(x_j > u)}{\sum_j I(x_j > u)} \quad \text{against} \quad u,$$

should be approximately straight with slope $\xi/(1-\xi)$ above $u_{\min}$.

☐ Can also test for equal shape parameters above u (Northrop–Coleman test).

Statistical implications of stability

☐ If $\mathcal{A}_u = [u, \infty)$ and u is sufficiently large that the GP approximation holds, we may write

$$\begin{aligned} \mathrm{P}(X \in \mathcal{A}_v \mid X \in \mathcal{A}_u) &= \mathrm{P}\left\{X > u + (v-u) \mid X > u\right\} \\ &\doteq \{1 + \xi(v-u)/\sigma_u\}_+^{-1/\xi}, \quad v > u, \end{aligned}$$

giving the estimate

$$\mathrm{P}(X \in \mathcal{A}_v) \doteq \mathrm{P}(X \in \mathcal{A}_u) \times \{1 + \xi(v-u)/\sigma_u\}_+^{-1/\xi} \doteq \frac{n_u}{n} \times \left\{1 + \widehat{\xi}(v-u)/\widehat{\sigma}_u\right\}_+^{-1/\xi},$$

where n_u is the number of exceedances of u.

☐ These calculations presuppose that the stability relations apply to exceedances of u, i.e., that the corresponding data are already in the asymptotic regime, which is generally hard to verify due to a lack of data.

Relation between GPD and GEV

☐ We have

$$\begin{aligned}
\mathrm{P}\left(\frac{M_n - b_n}{a_n} \le x\right) = F^n(b_n + a_n x) &= \left[1 - \frac{n\{1 - F(b_n + a_n x)\}}{n}\right]^n, \\
&= \left\{1 - \frac{\Lambda_n(x)}{n}\right\}^n,
\end{aligned}$$

where

$$\Lambda_n(x) = n\{1 - F(b_n + a_n x)\}, \quad x \in \mathbb{R}.$$

☐ Since $(1 + d_n/n)^n \to e^d$ for $\{d_n\}$ iff $d_n \to d$, we see that $(M_n - b_n)/a_n$ has a non-degenerate limiting distribution $\exp\{-\Lambda(x)\}$ iff

$$\lim_{n\to\infty} \Lambda_n(x) = \Lambda(x) \text{ exists for all } x \in \mathbb{R},$$

and $\Lambda(x)$ must be decreasing (and must take at least three values).

☐ If F is continuous, and we choose b_n such that $1 - F(b_n) = 1/n$, then for $x \ge 0$,

$$\Lambda_n(x) = n\{1 - F(b_n + a_n x)\} = \frac{1 - F(b_n + a_n x)}{1 - F(b_n)} = \mathrm{P}(X > b_n + a_n x \mid X > b_n),$$

so $\mathrm{P}(X > b_n + a_n x \mid X > b_n) \to \Lambda(x)$ is equivalent to convergence of $(M_n - b_n)/a_n$ to a non-degenerate limiting random variable.

Limit Λ

Lemma 6 *The only possible non-degenerate limit for $\Lambda_n(x)$ as $n \to \infty$ is*

$$\Lambda(x) = \begin{cases} \left(1 + \xi \frac{x-\eta}{\tau}\right)_+^{-1/\xi}, & \xi \ne 0, \\ \exp\left(-\frac{x-\eta}{\tau}\right), & \xi = 0. \end{cases}$$

☐ The location and scale parameters η and τ are not needed for the limit result, but are needed for statistical applications in which F is unknown.

Later we'll need

$$-\dot{\Lambda}(x) = -\frac{\mathrm{d}\Lambda(x)}{\mathrm{d}x} = \begin{cases} \tau^{-1}\left\{1 + \xi(x-\eta)/\tau\right\}_+^{-1/\xi-1}, & \xi \ne 0, \\ \tau^{-1}\exp\left\{-(x-\eta)/\tau\right\}, & \xi = 0, \end{cases}$$

which is non-negative because $\Lambda(x)$ is decreasing,

Proof of Lemma 6

☐ If the support of f is a finite union of separated intervals, then as $t \to \infty$ we need only consider the right-most interval. If the support is an infinite union, then $r'(u)$ cannot have a limit.

☐ For $t \geq 1$, set $b_t = F^{-1}(1 - 1/t) \in \mathcal{I}$ and $a_t = r(b_t) > 0$. If $x > 0$ satisfies $b_t + a_t x \in \mathcal{I}$, then

$$\begin{aligned} -\log t\{1 - F(b_t + a_t x)\} &= -\log\{1 - F(b_t + a_t x)\} - [-\log\{1 - F(b_t)\}] & (4)\\ &= \mathcal{H}(b_t + a_t x) - \mathcal{H}(b_t) & (5)\\ &= a_t \int_0^x \mathcal{H}'(b_t + a_t u)\,\mathrm{d}u \\ &= a_t \int_0^x \frac{\mathrm{d}u}{r(b_t + a_t u)}. \end{aligned}$$

Taylor's theorem implies that for each u there exists $s \equiv s(u) \in (0, u)$ such that

$$r(b_t + a_t u) = r(b_t) + a_t u r'\{b_t + a_t s(u)\}, \qquad (6)$$

so, as $r(b_t) = a_t$,

$$\frac{r(b_t + a_t u)}{a_t} = \frac{r(b_t) + a_t u r'(b_t + a_t s)}{a_t} = 1 + u r'(b_t + a_t s), \quad 0 < u < x. \qquad (7)$$

If $r'\{b_t + a_t s(u)\}$ did not depend on u, then (4) would exactly equal

$$a_t \int_0^x \frac{\mathrm{d}u}{r(b_t + a_t u)} = \int_0^x \frac{\mathrm{d}u}{1 + u r'(b_t + a_t s)} = \xi^{-1}\log(1 + \xi x),$$

with $\xi \equiv \xi(x, t) = r'(b_t + a_t s)$ depending on t and x, and the integral equalling x if $\xi = 0$.

☐ To show that such a ξ exists, we define

$$I(s) = \int_0^x \frac{\mathrm{d}u}{1 + u r'(b_t + a_t s)},$$

and note that since r and r' are continuous, the function $s(u)$ defined implicitly by (6) is continuous for $u \in (0, x)$. Hence there exist $s_1, s_2 \in [0, x]$ such that

$$r'(b_t + a_t s_1) \leq r'\{b_t + a_t s(u)\} \leq r'(b_t + a_t s_2), \quad 0 \leq u \leq x,$$

and thus that

$$\frac{1}{1 + u r'(b_t + a_t s_2)} \leq \frac{1}{1 + u r'\{b_t + a_t s(u)\}} \leq \frac{1}{1 + u r'(b_t + a_t s_1)}, \quad 0 \leq u \leq x, \qquad (8)$$

yielding

$$I(s_2) \leq \int_0^x \frac{\mathrm{d}u}{1 + u r'\{b_t + a_t s(u)\}} \leq I(s_1).$$

Now $I(s)$ is continuous, so the intermediate value theorem implies that there exists some $s^* \equiv s^*(x, t)$ such that $I(s^*)$ equals the integral. Hence

$$\int_0^x \frac{a_t}{r(b_t + a_t u)}\,\mathrm{d}u = \int_0^x \frac{\mathrm{d}u}{1 + u r'(b_t + a_t s^*)} = \frac{1}{\xi(t, x)}\log\{1 + \xi(t, x) x\}, \qquad (9)$$

with $\xi(t, x) = r'\{b_t + a_t s^*(x, t)\}$, as required. The integral equals x if $\xi(t, x) = 0$.

☐ The above argument also applies if $x < 0$ is such that $b_t + a_t x \in \mathcal{I}$, with $x < u < 0$, etc.

note 1 of slide 73

Proof of Lemma 6, II

☐ To show that $\xi = \lim_{t\to\infty} \xi(t,x) = \lim_{t\to\infty} r'\{b_t + a_t s^*(x,t)\}$ does not depend on x, we show that as $t \to \infty$ or equivalently $b_t \to x^*$, $b_t + a_t x \to x^*$, which implies that $b_t + a_t s^*(x,t) \to x^*$.

☐ First suppose that $x^* = \infty$, and write

$$b_t + a_t x = b_t(1 + x a_t/b_t),$$

so b_t and $b_t + a_t x$ have the same limit if a_t/b_t is bounded as $t \to \infty$. But for any fixed $u < b_t$ and some u_1 between u and b_t,

$$\frac{a_t}{b_t} = \frac{r(b_t)}{b_t} = \frac{r(u) + (b_t - u) r'(u_1)}{b_t}$$

is bounded because $r'(u_1) \to \xi$ as $t \to \infty$. Hence as $b_t \to \infty$, $b_t + a_t x \to \infty$ also.

☐ When $x^* < \infty$, we must have $F(x) \to 1$ as $x \to x^*$, since otherwise there is positive mass on x^* and the limiting distribution will be degenerate, which is not allowed by hypothesis. Hence

$$-\log\{1 - F(x)\} = \int_x^{x_*} \frac{\mathrm{d}u}{r(u)} \to \infty, \quad x \to x^*,$$

and this implies that $r(x^*) = 0$, since if $r(x^*) > 0$, the assumptions that $r(x)$ is positive and continuous in $\mathcal{I}$ would imply that the integral is finite. Hence l'Hôpital's rule gives

$$\xi = \lim_{b_t \to x^*} r'(b_t) = \lim_{b_t \to x^*} \frac{r(x^*) - r(b_t)}{x^* - b_t} = \lim_{b_t \to x^*} \frac{0 - a_t}{x^* - b_t} = -\lim_{b_t \to x^*} \frac{a_t}{x^* - b_t}. \qquad (10)$$

Notice that if $x^* < \infty$, then $\xi \le 0$, since the rightmost expression in (10) cannot be positive. Now

$$x^* - b_t - a_t x = (x^* - b_t)\left(1 - \frac{a_t x}{x^* - b_t}\right),$$

where the second term on the right-hand side converges to $1 + \xi x$ as $b_t \to x^*$, and hence is bounded, implying that the entire expression tends to zero.

☐ Hence if x is such that $\lim_{t\to\infty} 1 + a_t x/b_t = 1 + \xi x > 0$, then

$$\lim_{t\to\infty} \xi(t,x) = \lim_{t\to\infty} r'\{b_t + a_t s^*(x,t)\} = \lim_{t\to\infty} r'(b_t) = \xi$$

does not depend on x, and

$$t\{1 - F(b_t + a_t x)\} \to \begin{cases} (1 + \xi x)_+^{-1/\xi}, & \xi \ne 0, \\ \exp(-x), & \xi = 0, \end{cases}$$

for both positive and negative x; here the $(\cdot)_+$ gives a formula valid for all $x \in \mathbb{R}$.

☐ To see that the limit is unique up to location and scale, note that the value of ξ depends on the limit for $r'(x)$ as $x \to x^*$. This does not depend on location and scale changes, since replacing x and x^* by $y = \eta + \tau x$ and $y^* = \eta + \tau x^*$ simply puts $r'(x) = r'\{(y - \eta)/\tau\} \to r'\{(y^* - \eta)/\tau\} = r'(x^*)$ (exercise!), and only leads to location and scale changes in Λ, yielding

$$\Lambda(x) = \Lambda\{(y - \eta)/\tau\},$$

which is the general form given.

note 2 of slide 73

Exceedance theorem

☐ Theorem 4 follows directly from Lemma 6. When $x > 0$, the probability that a rescaled version of X exceeds $x + u$, conditional on it exceeding u, i.e.,

$$\mathrm{P}\left\{(X - b_n)/a_n > x + u \mid (X - b_n)/a_n > u\right\}$$

may be written as

$$\begin{aligned}
\frac{\mathrm{P}\left\{X > b_n + a_n(x+u)\right\}}{\mathrm{P}\left(X > b_n + a_n u\right)} &= \frac{n\left[1 - F\left\{b_n + a_n(x+u)\right\}\right]}{n\left\{1 - F\left(b_n + a_n u\right)\right\}} \\
&= \frac{\Lambda_n(x+u)}{\Lambda_n(u)} \\
&\to \frac{\Lambda(x+u)}{\Lambda(u)}, \quad n \to \infty, \\
&= \frac{\{1 + \xi(x + u - \eta)/\tau\}_+^{-1/\xi}}{\{1 + \xi(u - \eta)/\tau\}_+^{-1/\xi}} \\
&= \begin{cases} (1 + \xi x/\sigma_u)_+^{-1/\xi}, & \xi \neq 0, \\ \exp(-x/\sigma_u), & \xi = 0, \end{cases}
\end{aligned}$$

provided that $\sigma_u = \tau + \xi(u - \eta) > 0$, so that $\Lambda(u) > 0$. Thus the limiting probability that $(X - b_n)/a_n < x + u$, conditional on $(X - b_n)/a_n > u$, is given by the GPD.

Largest order statistics

☐ Consider the r largest order statistics $Y_r \leq \cdots \leq Y_1$ of the rescaled variables $\{(X_j - b_n)/a_n : j = 1, \ldots, n\}$, and suppose that $n \to \infty$.

☐ As Y_1 has the limiting distribution of the rescaled maximum $(M_n - b_n)/a_n$,

$$\Pr\left(Y_1 \leq y_1\right) = \exp\left\{-\Lambda\left(y_1\right)\right\}, \quad y_1 \in \mathbb{R}$$

and $f_{Y_1}\left(y_1\right) = \left\{-\dot{\Lambda}\left(y_1\right)\right\} \exp\left\{-\Lambda\left(y_1\right)\right\}$.

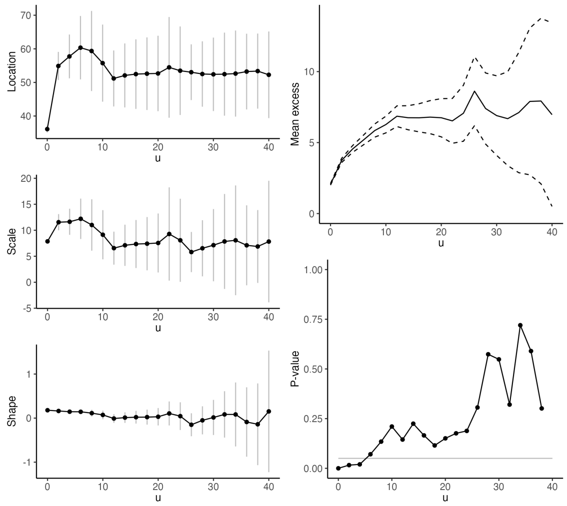
☐ The second-largest variable Y_2 is also the largest of an infinite number of these rescaled variables, so its distribution is also G, but conditioned on $Y_2 < Y_1$. Hence

$$\mathrm{P}(Y_2 \leq y_2 \mid Y_1 = y_1) = \exp\left\{\Lambda(y_1) - \Lambda(y_2)\right\}, \quad y_2 < y_1,$$

and it follows that the limiting joint density of the **r-largest order statistics** $Y_r < \cdots < Y_1$ is

$$f(y_1, \ldots, y_r) = \exp\{-\Lambda(y_r)\} \prod_{j=1}^{r} \{-\dot{\Lambda}(y_j)\}, \quad y_r < \cdots < y_1. \qquad (11)$$

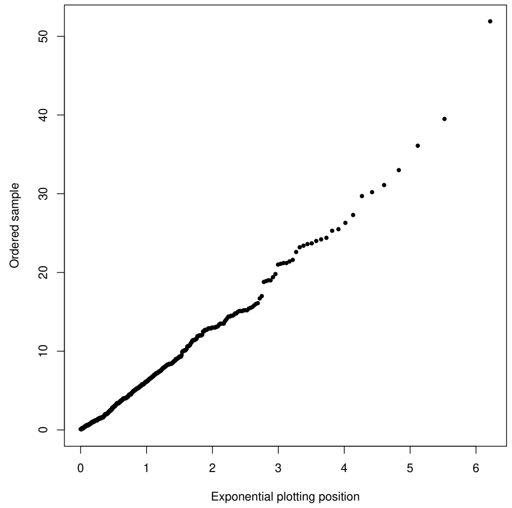
Abisko threshold analysis



☐ All panels suggest that $u_{\min} = 10$ is reasonable.

Exploratory plot

☐ The natural plot here is of ordered exceedances against exponential plotting positions:

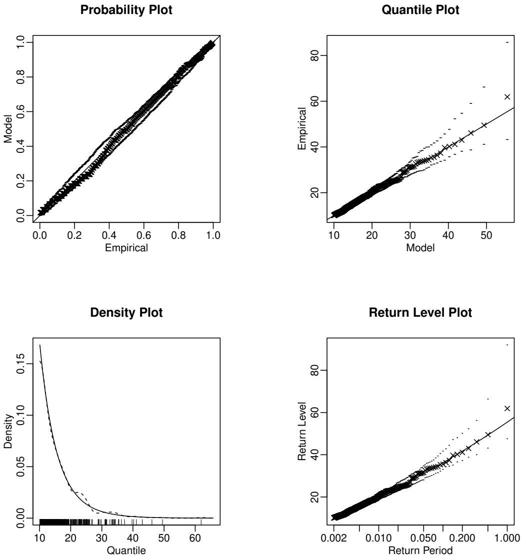


GPD fit

```
(fit.gpd <- fpot(abisko$precip,threshold=10))

Deviance: 2828.05

Threshold: 10
Number Above: 499
Proportion Above: 0.033

Estimates
  scale    shape
5.83261  0.07025

Standard Errors
  scale    shape
0.39483  0.05088

Optimization Information
  Convergence: successful
  Function Evaluations: 16
  Gradient Evaluations: 6
```

Abisko POT fit

- ☐ Let's check the fit using `plot(fit.gpd)`:



Summary

☐ The two fits agree fairly well:
  - Maxima: $\widehat{\eta} = 20.4_{0.649}$, $\widehat{\tau} = 5.84_{0.483}$, $\widehat{\xi} = 0.08_{0.072}$;
  - POT: $\widehat{p}_u = 0.033$, $\widehat{\sigma}_u = 5.83_{0.394}$, $\widehat{\xi} = 0.07_{0.051}$.

☐ The location and scale parameters are estimated quite well, but the shape much less well.

☐ The shape parameter estimate is slightly positive, but not significantly so (some hydrologists claim that rainfall has $\xi \approx 0.1 \ldots$).

☐ The fit appears to be good.

☐ In applications one would need to check that the threshold fits are robust to the choice of u (above $u_{\min}$).

☐ It is tempting to fit the model with $\xi = 0$, which will give much smaller standard errors for the other parameters. But as we do not know that $\xi = 0$, this reduction in uncertainty may be unrealistic, and it may introduce bias in extrapolation.

2.3 Targets of Inference

Return levels and return periods

☐ In basic analyses, typically aim to estimate risk measures such as

$$\mathrm{P}(X > x) = 1 - F_X(x), \quad x_p = F_X^{-1}(1-p),$$

where $X \sim F_X$ is a background observation and x and x_p are larger than any value yet observed.

☐ We often express risk in terms of blocks of m background observations, often daily measurements, with the blocks being years; then $m = 365.25$.

☐ We then call x_p a **T-year return level** with a **return period** of $1/p$ observations or T years (i.e., $N_p = Tm$ background observations),
  - e.g., the law states that nuclear installations should withstand the highest windspeed in $T = 10^7$ years(!), so if X is a daily maximum windspeed, then $N_p = 365.25 \times T$ and $p = 1/(365.25T)$.

☐ Hence a return level solves the equation

$$\mathrm{P}(X > x_p) = 1 - F_X(x_p) = p = 1/N_p \tag{12}$$

for some small p.

Return levels and return periods II

☐ If $x_p > u$, so p is less than the probability $\mathrm{P}(X > u) = p_u$ that a background observation exceeds threshold u, then solving $1 - F_X(x_p) = p$ in the POT model gives

$$x_p = \begin{cases} u + \frac{\sigma_u}{\xi}\left\{(p_u/p)^{\xi} - 1\right\}, & \xi \neq 0, \\ u + \sigma_u \log(p_u/p), & \xi = 0. \end{cases} \tag{13}$$

☐ The GEV applies to maxima of blocks of m background observations, so we approximate the upper tail of F_X by $G^{1/m}$, giving

$$1 - p = G^{1/m}(x_p), \tag{14}$$

which yields

$$x_p = \begin{cases} \eta + \frac{\tau}{\xi}\left[\{-m\log(1-p)\}^{-\xi} - 1\right], & \xi \neq 0, \\ \eta - \tau \log\left\{-m\log(1-p)\right\}, & \xi = 0. \end{cases} \tag{15}$$

☐ In both cases

- $-\log(1-p) \doteq p = 1/N_p$ for large N_p, giving simpler expressions,
- point estimates are obtained by replacing the unknown parameters by their estimates,
- uncertainty is best assessed using the profile log likelihood for x_p.

Note on computation of return levels

☐ For the POT model in which the GPD is fitted to exceedances of u, and provided $x_p > u$, we have

$$\begin{aligned} \mathrm{P}(X > x_p) &= \mathrm{P}(X > x_p \mid X > u)\mathrm{P}(X > u) \\ &= \mathrm{P}(X - u > x_p - u \mid X > u)\mathrm{P}(X > u) \\ &= \{1 + \xi(x_p - u)/\sigma_u\}_+^{-1/\xi} \times p_u, \end{aligned}$$

and we seek x_p such that

$$1 - p = \mathrm{P}(X \leq x_p) = 1 - p_u\{1 + \xi(x_p - u)/\sigma_u\}_+^{-1/\xi},$$

which leads to the stated expression for x_p.

☐ If the GEV model is fitted to the maxima of blocks of m background observations then we have

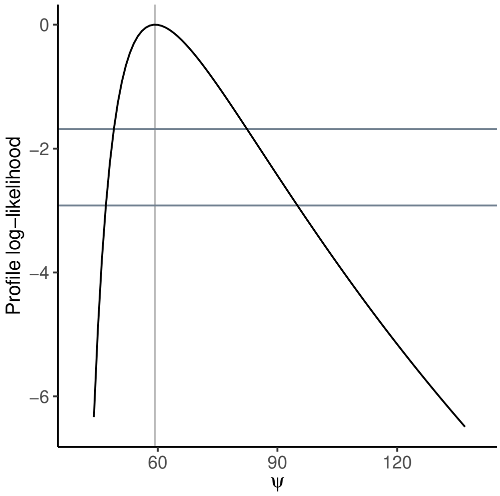
$$1 - p = G^{1/m}(x_p) = \exp\left[-\left\{1 + \xi(x_p - \eta)/\tau\right\}_+^{-1/\xi}/m\right],$$

which gives the stated expression for x_p.

Profile log-likelihood

☐ Here $\psi = x_p$ is the 100-year return level for daily precipitation at Abisko based on the GEV fit.

☐ The strong asymmetry means that symmetric confidence intervals could be very misleading.



Other measures of risk

☐ In environmental applications it may be important to estimate amounts of rain falling into an entire catchment area, or the length and impact of a heatwave, or . . .

☐ The Basel Accords regulate measures of risk to be used by financial institutions:

- the **Value at Risk** VaR_p is another name for a quantile/return level x_p;
- the **Expected Shortfall** is defined as the expected loss conditional on VaR_p being exceeded,

$$\mathrm{E}(X - \text{VaR}_p \mid X > \text{VaR}_p),$$

  where in both cases X represents a potential loss.
- More sophisticated measures such as **expectiles** are also used.

Comments

☐ The T-year return level is often called ‘the level exceeded once on average every T years’, and is easily misinterpreted:

- ‘on average’ does not mean that disasters arise at regular T-year intervals!
- selection is often discounted — if M independent time series are monitored, then we expect M/T T-year events each year;
- the assumption of stationarity is rarely true, so large events may cluster together in periods of elevated risk.

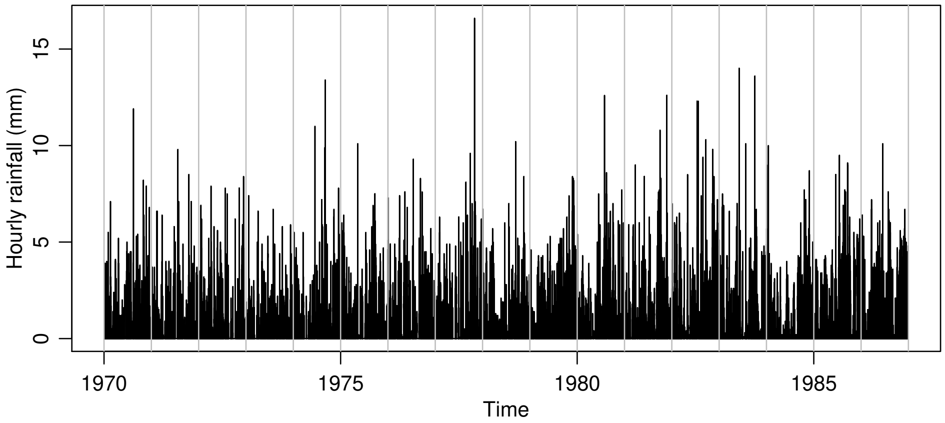
☐ Preferable to refer to quantiles — but probably impossible to change a cultural icon!

☐ Return levels and return periods are parameters of distributions, but future events are as-yet unobserved random variables, and it may be useful to consider their distributions. The distribution of the largest value X_T to be observed over T blocks of future background observations is $G^T(y)$, and it may be better to use this for risk analysis, in a Bayesian approach (later, probably).

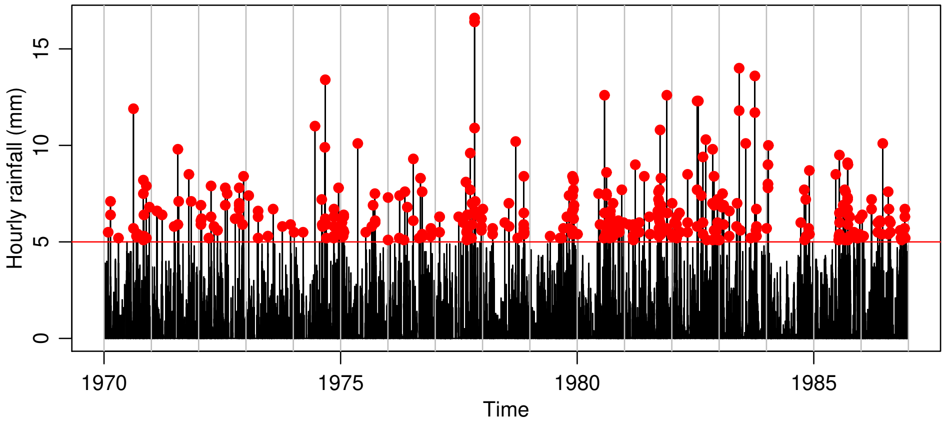
Improved inferences

- ☐ The block maximum method can be inefficient if other data are available. Alternative methods include:
  - – peaks over thresholds,
  - – r-largest order statistics.
- ☐ Both are special cases of a **point process** representation, under which we use a Poisson process to approximate the occurrence of those values that exceed a (high) threshold.
- ☐ The Poisson process representation is also very powerful in more general settings, so we discuss it before returning to data analysis.

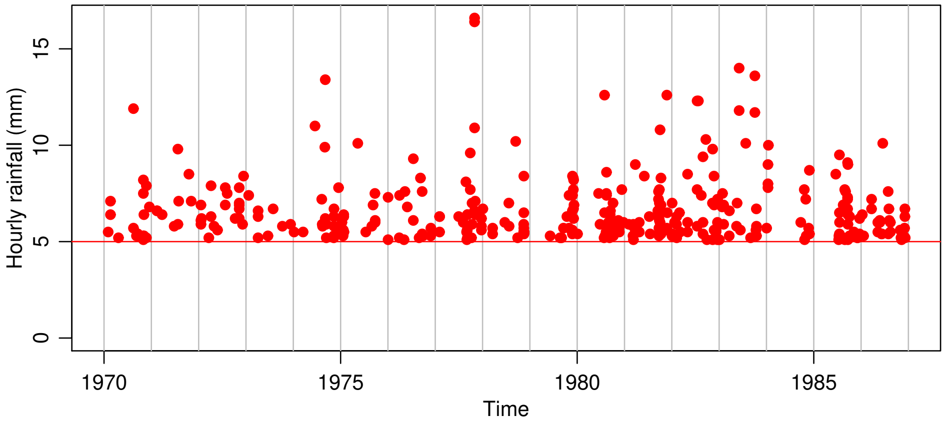
Example: Hourly rainfall at Eskdalemuir, 1970–86



Example: Eskdalemuir rainfall



Example: Eskdalemuir rainfall



Comments

☐ The fixed number of annual maxima has been replaced by a random number of exceedances over the threshold.

☐ We now retain more observations in the tail of the distribution.

☐ Dependence in the underlying series means that exceedances occur in clusters, which we may need to model.

☐ For now we suppose that the underlying series comprises independent identically distributed observations, whose maxima have a non-degenerate limit, after renormalisation.

3.1 Point Processes

Point process

☐ A **point process** is a stochastic model for a **point pattern** $\mathcal{P} = \{x_1, x_2, \ldots\}$ lying in a **state space** $\mathcal{E}$. We also call a point an **event**.

☐ We visualise $\mathcal{E} \subset \mathbb{R}^2$, but $\mathcal{E}$ might be more complex, e.g., $\mathcal{E} = \mathbb{R} \times \mathcal{C}$, where $\mathcal{C}$ is a space of functions—then a 'point' would be $x = (u, f) \in \mathcal{E}$, with $u \in \mathbb{R}$ and $f \in \mathcal{C}$.

☐ The set $\mathcal{E}$ must allow us to count how many points of $\mathcal{P}$ lie in any suitable subset $\mathcal{A} \subset \mathcal{E}$, giving

$$N(\mathcal{A}) = |\mathcal{P} \cap \mathcal{A}| = \sum_x I(x \in \mathcal{P} \cap \mathcal{A}), \quad \mathcal{A} \subset \mathcal{E},$$

where $I(\cdot)$ is an indicator function.

☐ Two points cannot exactly coincide: $\mathcal{P}$ must be **simple** (or **orderly**) — otherwise we would not know how many points there are.

☐ If you know about measures . . . the function $N(\mathcal{A})$ is

- a **counting measure** on $\mathcal{E}$, since it counts the number of elements of $\mathcal{P}$ in any (measurable) set $\mathcal{A}$,
- a **Radon measure** if $N(\mathcal{A}) < \infty$ for any $\mathcal{A}$ compact (in a suitable topology on $\mathcal{E}$),
- a **random measure** if the points $\mathcal{P}$ arise at random, since then $N(\mathcal{A})$ is a random variable computed from the (random) $\mathcal{P}$.

Laplace transform

☐ If it exists, the **Laplace transform** of a scalar random variable X is defined as

$$\mathrm{E}\{\exp(-tX)\} = M_X(-t),$$

where M_X is the **moment-generating function (MGF)**. This is useful because

- there is a bijection between distributions and MGFs, i.e., if we recognise M_X, then we know the corresponding distribution;
- the **continuity theorem** tells us that if $\{X_n\}$, X have CDFs $\{F_n\}$, F for which the MGFs $M_n(t)$, $M(t)$ exist and there exists $a > 0$ such that

$$\lim_{n\to\infty} M_n(t) = M(t), \quad 0 \le |t| < a,$$

  then $X_n \xrightarrow{D} X$, i.e., X_n converges in distribution (weakly, in law) to X.
- Hence for large enough n we can approximate the distribution of X_n by that of X.

☐ On the next slide we will extend this to point processes, but first, a simple example:

Theorem 7 (Law of small numbers) *If $X_n \sim B(n, p_n)$ and $np_n \to \lambda > 0$ when $n \to \infty$, then the limiting distribution of X_n is $\mathrm{Pois}(\lambda)$, i.e., $X_n \xrightarrow{D} X$, where $X \sim \mathrm{Pois}(\lambda)$.*

Note to Theorem 7

☐ The MGF of X is

$$M(t) = \mathrm{E}(e^{tX}) = \sum_{x=0}^{\infty} e^{tx}\lambda^x e^{-\lambda}/x! = e^{-\lambda}\sum_{x=0}^{\infty}(\lambda e^t)^x/x! = \exp\left\{\lambda(e^t-1)\right\}, \quad t \in \mathbb{R}.$$

☐ The MGF of X_n is

$$M_n(t) = \mathrm{E}(e^{tX_n}) = \sum_{x=0}^{n} e^{tx}\binom{n}{x}p_n^x(1-p_n)^{n-x} = \left(1-p_n+p_ne^t\right)^n, \quad t \in \mathbb{R}.$$

Let $p_n = \lambda_n/n$, where $\lambda_n \to \lambda$, and note that as $n \to \infty$ and for any real t,

$$\left(1-p_n+p_ne^t\right)^n = \left(1+\frac{\lambda_n(e^t-1)}{n}\right)^n \to \exp\left\{\lambda(e^t-1)\right\}.$$

☐ As $M_n(t) \to M(t)$ for all real t, the continuity theorem implies that $X_n \xrightarrow{D} X$.

Laplace functional

- ☐ We specify properties of $\mathcal{P}$ through the finite-dimensional distributions of $N(\cdot)$, i.e.,

$$\mathrm{P}\{N(\mathcal{A}_1)=n_1,\ldots,N(\mathcal{A}_k)=n_k\},\quad n_1,\ldots,n_k\in\{0,1,2,\ldots,\},$$

for all possible choices of sets $\mathcal{A}_1,\ldots,\mathcal{A}_k$, and all $k=0,1,2,\ldots$.

- ☐ An efficient way to do this is through the **Laplace functional**,

$$\mathcal{L}_{\mathcal{P}}(f)=\mathrm{E}\left\{\exp\left(-\int f\mathrm{d}\mathcal{P}\right)\right\},\quad \text{where}\quad \int f\mathrm{d}\mathcal{P}=\int f(x)\mathcal{P}(\mathrm{d}x)=\sum_{x\in\mathcal{P}}f(x),$$

for functions $f\geq 0$ that are positive only on a bounded set. If $f(x)=\sum_r t_r I(x\in\mathcal{A}_r)$, then $\mathcal{L}_{\mathcal{P}}(f)$ is the joint MGF for the $N(\mathcal{A}_r)$.

- ☐ Under mild conditions, there is
  - a bijection between point processes and Laplace functionals; and
  - the continuity theorem can be generalised.

Convergence of point processes

Definition 8 *A sequence of random variables* $\{X_n\}$ *with corresponding CDFs* $\{F_n\}$ **converges weakly (or in distribution)** *to a random variable* X *with CDF* F *if*

$$\lim_{n\to\infty}F_n(x)=F(x)\text{ at every } x \text{ where } F \text{ is continuous.}$$

Definition 9 *A sequence of point processes* $\{\mathcal{P}_n\}$ *with corresponding counts* $\{N_n(\cdot)\}$ *on* $\mathcal{E}$ **converges weakly (or in distribution)** *to a point process* $\mathcal{P}$ *with count* $N(\cdot)$*, written* $\mathcal{P}_n\xrightarrow{D}\mathcal{P}$*, if for all choices of* k *and all compact sets* $\mathcal{A}_1,\ldots,\mathcal{A}_k\subset\mathcal{E}$ *such that*

$$\mathrm{P}\{N(\partial\mathcal{A}_j)=0\}=1,\quad j=1,\ldots,k,$$

where $\partial\mathcal{A}_j$ *is the boundary of* $\mathcal{A}_j$,

$$\{N_n(\mathcal{A}_1),\ldots,N_n(\mathcal{A}_k)\}\xrightarrow{D}\{N(\mathcal{A}_1),\ldots,N(\mathcal{A}_k)\},\quad n\to\infty.$$

Theorem 10 (No proof) *The point processes* $\mathcal{P}_1,\mathcal{P}_2,\ldots$ *converge weakly to the point process* $\mathcal{P}$ *on* $\mathcal{E}$ *if and only if the corresponding Laplace functionals converge for every continuous non-negative function* f *on* $\mathcal{E}$ *with compact support, i.e., as* $n\to\infty$,

$$\mathcal{L}_{\mathcal{P}_n}(f)=\mathrm{E}\left\{\exp\left(-\int f\mathrm{d}\mathcal{P}_n\right)\right\}\to\mathcal{L}_{\mathcal{P}}(f)=\mathrm{E}\left\{\exp\left(-\int f\mathrm{d}\mathcal{P}\right)\right\}.$$

Kallenberg's theorem

☐ **Kallenberg's theorem** gives another way to establish the weak convergence of $\{\mathcal{P}_n\}$ to a simple process $\mathcal{P}$ when $\mathcal{E} \subset \mathbb{R}$.

☐ For any $\mathcal{A} \subset \mathcal{E}$, let $N_n(\mathcal{A}) = |\mathcal{P}_n \cap \mathcal{A}|$. Then if

- $\mathcal{B} \subset \mathcal{E}$ is any interval,
- $\mathcal{C}$ is any finite union of disjoint sub-intervals of $\mathcal{E}$,

and if

$$\mathrm{E}\{N_n(\mathcal{B})\} \to \mathrm{E}\{N(\mathcal{B})\}, \quad \mathrm{P}\{N_n(\mathcal{C}) = 0\} \to \mathrm{P}\{N(\mathcal{C}) = 0\}, \quad n \to \infty, \tag{16}$$

then $\mathcal{P}_n$ converges weakly to $\mathcal{P}$.

☐ When $\mathcal{E} \subset \mathbb{R}^D$, the same result holds if intervals are replaced by **rectangles**,

$$(a, b] = \{x = (x_1, \ldots, x_D) : a_d < x_d \le b_d, d = 1, \ldots, D\} \subset \mathcal{E},$$

where $a_d < b_d$ for each d.

☐ Thus weak convergence of point processes to a simple limiting process in $\mathbb{R}^D$ entails establishing convergence of expected counts for rectangles and of the **void probabilities** of finite unions of rectangles.

☐ See Kingman (1993) *Poisson Processes* and Daley and Vere-Jones (2002, 2008), *An Introduction to the Theory of Point Processes*.

3.2 Poisson Processes

Poisson process

Definition 11 *A* **Poisson process** *is a random countable subset $\mathcal{P}$ of a state space $\mathcal{E}$ such that*

☐ *the random variables $N(\mathcal{A}_1), \ldots, N(\mathcal{A}_k)$ corresponding to any collection of disjoint subsets $\mathcal{A}_1, \ldots, \mathcal{A}_k$ of $\mathcal{E}$ are independent; and*

☐ *for any $\mathcal{A} \subset \mathcal{E}$, $N(\mathcal{A})$ has the Poisson distribution with mean $\mu(\mathcal{A})$, where $0 \le \mu(\mathcal{A}) \le \infty$, and $\mu(\mathcal{A}) < \infty$ for compact $\mathcal{A}$.*

Comments:

☐ if $\mathcal{A} = \bigcup_j \mathcal{A}_j$ is a countable union of disjoint sets, then $N(\mathcal{A}) = \sum_j N(\mathcal{A}_j)$, so $\mu(\mathcal{A}) = \sum_j \mu(\mathcal{A}_j)$, and μ is a measure; called the **mean measure** of $\mathcal{P}$;

☐ μ must be **diffuse**, i.e., $\mu(\{x\}) = 0$ for every $x \in \mathcal{E}$;

☐ if $\mathcal{E} \subset \mathbb{R}^D$, $\mathcal{A} = [a_1, x_1] \times \cdots \times [a_D, x_D]$, and if

$$\dot{\mu}(x_1, \ldots, x_D) = \frac{\partial^D \mu(\mathcal{A})}{\partial x_1 \cdots \partial x_D}$$

exists and is finite, then $\dot{\mu}$ is called the **intensity function** of $\mathcal{P}$;

☐ if $\dot{\mu}(x) \equiv \dot{\mu}$, then $\mathcal{P}$ is called **homogeneous**. Otherwise it is **inhomogeneous**.

☐ We simplify notation by replacing $\mu\{(a, b]\}$ by $\mu(a, b]$, etc.

Conditioning

Theorem 12 (Conditioning) *Let $\mathcal{P}$ be a Poisson process with mean measure μ, and suppose that $\mathcal{A} \subset \mathcal{E}$ is such that $0 < \mu(\mathcal{A}) < \infty$. Conditional on the event $N(\mathcal{A}) = n$, the n points of $\mathcal{P} \cap \mathcal{A}$ have the same distribution as n points generated independently at random in $\mathcal{A}$ with measure $\mu_{\mathcal{A}}(\mathcal{B}) = \mu(\mathcal{B})/\mu(\mathcal{A})$, for $\mathcal{B} \subset \mathcal{A}$.*

- ☐ If μ has intensity $\dot{\mu}(x)$, then we can generate points of $\mathcal{P}$ in $\mathcal{A}$ by
  - generating a value n of $N(\mathcal{A}) \sim \text{Poiss}\{\mu(\mathcal{A})\}$;
  - then generating $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \dot{\mu}(x)/\mu(\mathcal{A})$ for $x \in \mathcal{A}$.
- ☐ The process generated at the second step is a **binomial process**.

Lemma 13 *The* **Laplace functional** *of a Poisson process $\mathcal{P}$ on $\mathcal{E}$ with mean measure μ is*

$$\mathcal{L}_{\mathcal{P}}(f) = \exp\left[- \int_{\mathcal{E}} \left\{ 1 - e^{-f(x)} \right\} \mu(\mathrm{d}x) \right].$$

slide 101

Note to Theorem 12

If we observe a Poisson process with intensity $\dot{\mu}(x)$ on the set $\mathcal{A}$, and there are points at $\{x_1, \ldots, x_n\}$, then the corresponding probability element is

$$\exp\{-\mu(\mathcal{A})\} \times \prod_{j=1}^{n} \dot{\mu}(x_j), \quad \{x_1, \ldots, x_n\} \subset \mathcal{A}.$$

Properties of the Poisson process imply that $N(\mathcal{A})$ has a Poisson distribution with mean $\mu(\mathcal{A})$, so the conditional density of the n points in $\mathcal{A}$, given that $N(\mathcal{A}) = n$, is the ratio

$$\frac{\exp\{-\mu(\mathcal{A})\} \times \prod_{j=1}^{n} \dot{\mu}(x_j)}{\mu(\mathcal{A})^n \exp\{-\mu(\mathcal{A})\}/n!} = n! \prod_{j=1}^{n} \left\{ \frac{\dot{\mu}(x_j)}{\mu(\mathcal{A})} \right\}, \quad \{x_1, \ldots, x_n\} \subset \mathcal{A}.$$

Now consider the measure $\mu_{\mathcal{A}}(\mathcal{B}) = \mu(\mathcal{B})/\mu(\mathcal{A})$, for $\mathcal{B} \subset \mathcal{A}$, which is a probability measure on subsets of $\mathcal{A}$, because it is non-negative and $\mu_{\mathcal{A}}(\mathcal{A}) = 1$. The corresponding probability density is $\dot{\mu}(x)/\mu(\mathcal{A})$ ($x \in \mathcal{A}$), so the joint density for independent identically distributed variables $X_1, \ldots, X_n$ with distribution $\mu_{\mathcal{A}}$ is $\prod_{j=1}^{n}\{\dot{\mu}(x_j)/\mu(\mathcal{A})\}$, which is almost the conditional probability above. The additional factor $n!$ arises because the point process is unlabelled: the same density would arise for any of the $n!$ permutations of $X_1, \ldots, X_n$ that gave the outcome $\{x_1, \ldots, x_n\}$.

note 1 of slide 101

Note to Lemma 13

Let $f \geq 0$ have support only on a compact $\mathcal{A}$, so $\mu(\mathcal{A}) < \infty$. Conditional on $N(\mathcal{A}) = n$, $\int f(x)\mathcal{P}(\mathrm{d}x) = \sum_{j=1}^{n} f(X_j)$, where $\{X_1, \ldots, X_n\} \subset \mathcal{A}$ are independent with density $\dot{\mu}(x)/\mu(\mathcal{A})$. Thus

$$\begin{aligned} \mathrm{E}\left[\exp\left\{-\int f(x)\mathcal{P}(\mathrm{d}x)\right\} \,\middle|\, N(\mathcal{A}) = n\right] &= \mathrm{E}\left[\exp\left\{-\sum_{j=1}^{n} f(X_j)\right\} \,\middle|\, N(\mathcal{A}) = n\right] \\ &= \left\{\int_{\mathcal{A}} e^{-f(x)}\mu(\mathrm{d}x)/\mu(\mathcal{A})\right\}^{n}. \end{aligned}$$

Hence

$$\begin{aligned} \mathrm{E}\left[\exp\left\{-\int f(x)\mathcal{P}(\mathrm{d}x)\right\}\right] &= \sum_{n=0}^{\infty}\left\{\int_{\mathcal{A}} e^{-f(x)}\mu(\mathrm{d}x)/\mu(\mathcal{A})\right\}^{n} \frac{\mu(\mathcal{A})^n}{n!} e^{-\mu(\mathcal{A})} \\ &= \exp\left[\int_{\mathcal{A}} e^{-f(x)}\mu(\mathrm{d}x) - \mu(\mathcal{A})\right] \\ &= \exp\left[-\int_{\mathcal{A}}\left\{1 - e^{-f(x)}\right\}\mu(\mathrm{d}x)\right] \\ &= \exp\left[-\int_{\mathcal{E}}\left\{1 - e^{-f(x)}\right\}\mu(\mathrm{d}x)\right], \end{aligned}$$

as required, since $1 - \exp\{-f(x)\} \equiv 0$ outside $\mathcal{A}$.

Superposition and colouring

Theorem 14 (Superposition) *If $\mathcal{P}_1$, $\mathcal{P}_2$ are independent Poisson processes on $\mathbb{R}^D$ with mean measures μ_1, μ_2, then their union $\mathcal{P}_1 \cup \mathcal{P}_2$ is a Poisson process with mean measure $\mu_1 + \mu_2$.*

Theorem 14 extends to a countable number of Poisson processes.

Theorem 15 (Colouring) *Let $\mathcal{P}$ be a Poisson process with intensity $\dot{\mu}(x)$. Colour a point of $\mathcal{P}$ at x red with probability $\gamma(x)$; otherwise colour it green. Then the red and green sets of points $\mathcal{P}_{\mathrm{red}}$ and $\mathcal{P}_{\mathrm{green}}$ are independent Poisson processes with intensity functions*

$$\dot{\mu}_{\mathrm{red}}(x) = \dot{\mu}(x)\gamma(x), \quad \dot{\mu}_{\mathrm{green}}(x) = \dot{\mu}(x)\{1 - \gamma(x)\}.$$

The colouring theorem is in some sense the inverse of the superposition theorem, and it too applies with a countable number of colours.

Note to Theorem 14

☐ This looks easy using the Laplace functional for $\mathcal{P}_1 \cup \mathcal{P}_2$, which is

$$\mathcal{L}_{\mathcal{P}_1 \cup \mathcal{P}_2}(f) = \mathrm{E}\left[\exp\left\{-\int f(x)(\mathcal{P}_1 \cup \mathcal{P}_2)(\mathrm{d}x)\right\}\right].$$

Now

$$\int f(x)(\mathcal{P}_1 \cup \mathcal{P}_2)(\mathrm{d}x) = \int f(x)\mathcal{P}_1(\mathrm{d}x) + \int f(x)\mathcal{P}_2(\mathrm{d}x),$$

and the two processes are independent, so

$$\begin{aligned}
\mathrm{E}\left[\exp\left\{-\int f\mathrm{d}(\mathcal{P}_1 \cup \mathcal{P}_2)\right\}\right] &= \mathrm{E}\left\{\exp\left(-\int f\mathrm{d}\mathcal{P}_1\right)\right\} \times \mathrm{E}\left\{\exp\left(-\int f\mathrm{d}\mathcal{P}_2\right)\right\} \\
&= \exp\left\{-\int_{\mathcal{E}}\left(1 - e^{-f}\right)\mathrm{d}\mu_1\right\} \times \exp\left\{-\int_{\mathcal{E}}\left(1 - e^{-f}\right)\mathrm{d}\mu_2\right\} \\
&= \exp\left\{-\int_{\mathcal{E}}\left(1 - e^{-f}\right)\mathrm{d}(\mu_1 + \mu_2)\right\},
\end{aligned}$$

which is the Laplace functional of a Poisson process with mean measure $\mu_1 + \mu_2$.

☐ The catch with the argument above is the assumption that points of $\mathcal{P}_1$ and $\mathcal{P}_2$ do not coincide, so that

$$\mathrm{P}(\mathcal{P}_1 \cap \mathcal{P}_2 \cap \mathcal{A} = \emptyset) = 1$$

for any $\mathcal{A}$ for which $\mu_1(\mathcal{A})$, $\mu_2(\mathcal{A})$ are both finite. This is intuitively obvious but takes a bit of measure-theoretic work to prove.

Mapping

Theorem 16 (Mapping) *Let $\mathcal{P}$ be a Poisson process on $\mathcal{E}$ with mean measure μ, and suppose that the function $g : \mathcal{E} \to \mathcal{E}^*$ maps $\mathcal{E}$ into $\mathcal{E}^*$. Define*

$$\mu^*(\mathcal{A}^*) = \mu\{g^{-1}(\mathcal{A}^*)\}, \quad \mathcal{A}^* \subset \mathcal{E}^*.$$

If

(i) $\mu^*(\{x^*\}) = \mu^*(x^*) = 0$ for every $x^* \in \mathcal{E}^*$, and

(ii) $\mu^*(\mathcal{A}^*) < \infty$ for any compact $\mathcal{A}^*$,

then $\mathcal{P}^* = g(\mathcal{P})$ is a Poisson process on $\mathcal{E}^*$ with mean measure μ^*.

Here

☐ (i) implies that g does not create atoms in $\mathcal{E}^*$,

☐ (ii) implies that no compact set $\mathcal{A}^* \subset \mathcal{E}^*$ has infinite measure,

which are both needed for $\mathcal{P}^*$ to be Poisson.

Example 17 *If $\mathcal{P}$ is a homogeneous Poisson process of unit rate on $(0, \infty)$, and $g(x) = 1/x$, show that $g(\mathcal{P})$ is a Poisson process and find its intensity function. What if $g(x) = \lceil x \rceil$ or $g(x) = |\sin x|$?*

Note to Example 17

☐ The mean measure of $\mathcal{P}$ is given by $\mu[a,b] = (b-a)$, for $0 < a < b < \infty$.
(i) The function g maps $(0,\infty)$ to $(0,\infty)$, and $g \equiv g^{-1}$, so $g^{-1}(x^*) = 1/x^*$ satisfies $\mu[1/x^*, 1/x^*] = (1/x^* - 1/x^*) = 0$ for any $0 < x^* < \infty$.
(ii) Any compact set $\mathcal{A}$ of $(0,\infty)$ is a subset of a set $[a,b]$ for some $0 < a < b < \infty$, so

$$\mu^*(\mathcal{A}) = \mu(g^{-1}\mathcal{A}) = \int_{g^{-1}\mathcal{A}} \dot{\mu}(x)\,dx \le \int_{g^{-1}[a,b]} 1\,dx = \int_{1/b}^{1/a} dx = (1/a - 1/b) < \infty.$$

Hence $g(\mathcal{P})$ is indeed a Poisson process, and since $\mu[a,b] = (1/a - 1/b)$, its intensity function is $d\mu[a,b]/db = 1/b^2$, for $b > 0$.
A sketch shows what happens to the intensities of $\mathcal{P}$ and $g(\mathcal{P})$.

☐ With $g(x) = \lceil x \rceil$, where $\lceil x \rceil$ is the smallest integer greater than or equal to x, then condition (i) fails whenever $x^* \in \mathbb{N}$, so the resulting process is not Poisson, as points of $\mathcal{P}^*$ could be superposed on the positive integers and thus $N^*(\cdot)$ is not well-defined. Equivalently,

$$\mu^*(\{n\}) = \mu[g^{-1}(\{n\})] = \mu\{(n-1, n]\} = 1, \quad n \in \mathbb{N},$$

so μ^* has atoms on every positive integer and thus is not diffuse.

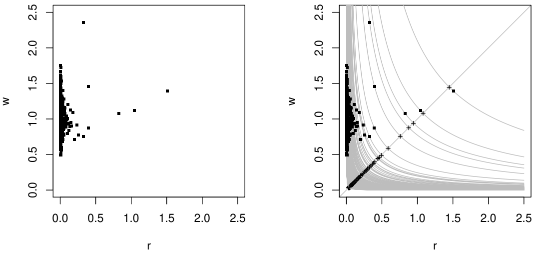
☐ With $g(x) = |\sin(x)|$ we have $\mathcal{E}^* = [0,1]$, and it is easy to check that while condition (i) is satisfied, $\mu^*([a,b]) = \infty$ for any $0 < a < b < 1$, so condition (ii) fails; $\mathcal{P}^*$ has an infinite number of points in any interval.

Example

Example 18 *Let $\mathcal{P}$ be a Poisson process with $\mathcal{E} = \mathbb{R}_+^2$ with $x = (r,w)$ generated by*

$$\mu\{(r,\infty)\times(w,\infty)\} = \frac{1}{r} \times \{1 - F(w)\}, \quad r, w > 0,$$

where F is the CDF of a positive continuous random variable W with unit expectation. Show that $q = rw$ defines a Poisson process and find its intensity.



Left panel: first 1000 points (r,w) of a Poisson process sequentially generated on $\mathbb{R}_+^2$. Right panel: mapping of the points shown in the left panel to $q = rw$, shown as + on the diagonal, with the mapping function shown by the curved grey lines.

Note to Example 18

☐ In the picture $\mathcal{P} = \{(R_i, W_i) : i = 1, 2, \ldots\}$, where $R_1 > R_2 > \cdots > 0$ are generated sequentially by setting $R_i = (E_1 + \cdots + E_i)^{-1}$, with $E_i \overset{\text{iid}}{\sim} \exp(1)$, and $W_i = \exp(\sigma\varepsilon_i - \sigma^2/2)$, where $\varepsilon_i \overset{\text{iid}}{\sim} N(0, 1)$, independent of the E_i; note that $\mathrm{E}(W_i) = 1$. The first 1000 points of a realisation of such a process are shown in the left-hand panel of the figure; the full realisation would have an infinity of points at the left-hand edge of the panel, because $\mu\{(r, \infty) \times (0, \infty)\} = 1/r \to \infty$ as $r \to 0$.

☐ In the general case the mean measure has an *intensity function* $\dot\mu$ given by its derivative at the upper right corner of a rectangle $(r', r) \times (w', w)$, i.e.,

$$\begin{aligned} \dot\mu(r, w) &= \frac{\partial^2 \mu\{(r', r) \times (w', w)\}}{\partial r \partial w} \\ &= \frac{\partial^2}{\partial r \partial w} \{\mu(r', w') - \mu(r, w') - \mu(r', w) + \mu(r, w)\} \qquad (17) \\ &= \frac{1}{r^2} \times f(w), \quad r, w > 0, \qquad (18) \end{aligned}$$

where we have written $\mu(r, w) = \mu\{(r, \infty) \times (w, \infty)\}$ and so forth, and f denotes the density function corresponding to F.

☐ Let $g(r, w) = rw$, corresponding to setting $Q_i = R_i W_i$, which amounts to collapsing the points shown in the left-hand panel onto the diagonal line shown in the right-hand panel. For any $q > 0$, $\mu^*(q) = \mu[\{(r, q/r) : r > 0\}] = 0$ because μ has a density with respect to Lebesgue measure and the set $\{(r, q/r) : r > 0\}$ has Lebesgue measure zero, so this transformation does not create atoms. We can check the second property of μ^* once it is calculated. Note that $Q = RW > q$ if and only if $R > q/W$, and that $\mathcal{A}_q = \{(r, w) : rw > q\}$ has measure

$$\begin{aligned} \mu^*(q) = \mu(\mathcal{A}_q) &= \int_0^\infty f(w) \int_{r=q/w}^\infty \frac{1}{r^2} \, \mathrm{d}r \, \mathrm{d}w \\ &= \int_0^\infty f(w) \left[-\frac{1}{r}\right]_{q/w}^\infty \mathrm{d}w \\ &= \int_0^\infty f(w) \frac{1}{q/w} \, \mathrm{d}w \\ &= \frac{1}{q}\mathrm{E}(W) = \frac{1}{q}, \quad q > 0. \qquad (19) \end{aligned}$$

Hence $Q_i = R_i W_i$ is also Poisson, with the same mean measure as the R_i. This implies that the second property is also satisfied: any compact set $\mathcal{A}^*$ is a subset of (q_1, q_2) for some $q_2 > q_1$, so

$$\mu^*(\mathcal{A}^*) \le \mu^*(q_1, \infty) = \mu^*(q_2, \infty) = q_1^{-1} - q_2^{-1} < \infty.$$

☐ The restriction of $\mathcal{P}$ to a subset $\mathcal{E}'$ of $\mathcal{E}$ clearly also follows a Poisson process, with mean measure $\mu'(\mathcal{A}) = \mu(\mathcal{E}' \cap \mathcal{A})$. For example, if we let $\mathcal{E} = (0, \infty)$, consider $R_1, R_2, \ldots$ and let $\mathcal{E}' = (z', \infty)$ for some $z' > 0$, then we retain only those points R_i exceeding z'. As $\mu(\mathcal{E}') = 1/z'$ is finite, these R_i can be generated by first simulating a Poisson variable N' with mean $1/z'$, and if $N' = n$, simulating n independent variables on the interval (z', ∞) with survivor function z'/z; these Pareto variables have probability density function z'/z^2 $(z > z')$.

Marking

Theorem 19 (Marking) *Let $\mathcal{P}$ be a Poisson process on $\mathcal{E}$ with mean measure μ. Attach a random variable y_x, called the mark, to each point x of $\mathcal{P}$; the distribution of $y_x \in \mathcal{Y}$ may depend on x but not on any other point of $\mathcal{P}$. Then the points (x, y_x) form a Poisson process $\mathcal{P}^*$ in the product space $\mathcal{E} \times \mathcal{Y}$ with mean measure*

$$\mu(\mathcal{C}) = \iint_{(x,y)\in\mathcal{C}} \nu_x(\mathrm{d}y)\mu(\mathrm{d}x), \quad \mathcal{C} \subset \mathcal{E} \times \mathcal{Y},$$

where $\nu_x(\cdot)$ is the conditional probability measure of y_x given x.

- ☐ This provides an approach to making new Poisson processes, by attaching random variables to existing processes, and (perhaps) then applying the mapping theorem.
- ☐ If y_x takes a countable number of values ($\equiv$ colours), then the colouring theorem shows that the corresponding subsets of $\mathcal{P}$ are independent Poisson processes.

Note to Theorem 19

The Laplace functional of $\mathcal{P}^*$ is

$$\mathrm{E}\left\{\exp\left(-\int f \mathrm{d}\mathcal{P}^*\right)\right\} = \mathrm{E}_{\mathcal{P}}\left[\mathrm{E}\left\{\exp\left(-\int f \mathrm{d}\mathcal{P}^*\right) \,\middle|\, \mathcal{P}\right\}\right]$$

and the inner expectation on the right-hand side is

$$\prod_{x\in\mathcal{P}} \int_{\mathcal{Y}} e^{-f(x,y)}\, \nu_x(\mathrm{d}y) = \exp\left(-\int f^*\, \mathrm{d}\mathcal{P}\right),$$

say, where

$$f^*(x) = -\log \int_{\mathcal{Y}} e^{-f(x,y)}\, \nu_x(\mathrm{d}y).$$

Thus the Laplace functional of $\mathcal{P}^*$ is that of the Poisson process $\mathcal{P}$ with f replaced by f^*. But since

$$\begin{aligned}\int_{\mathcal{E}} \left\{1 - e^{-f^*(x)}\right\} \mu(\mathrm{d}x) &= \int_{\mathcal{E}} \left\{1 - \int_{\mathcal{Y}} e^{-f(x,y)}\, \nu_x(\mathrm{d}y)\right\} \mu(\mathrm{d}x) \\ &= \int_{\mathcal{E}} \int_{\mathcal{Y}} \left\{1 - e^{-f(x,y)}\right\}\, \nu_x(\mathrm{d}y)\mu(\mathrm{d}x),\end{aligned}$$

the result is established.

3.3 Poisson Modelling of Extremes

Basic result

Theorem 20 *Let* $X_1, \ldots, X_{nt_0} \overset{\text{iid}}{\sim} F$ *form* t_0 *blocks each of* n *observations, and suppose that sequences* $\{a_n\} > 0$ *and* $\{b_n\}$ *exist such that*

$$\mathrm{P}\left[\{\max(X_1, \ldots, X_n) - b_n\}/a_n \le x\right] \to G(x), \quad n \to \infty,$$

where G *is non-degenerate. Then as* $n \to \infty$ *the point processes*

$$\mathcal{P}_n = \{(j/(n+1), (X_j - b_n)/a_n) : j = 1, \ldots, nt_0\}$$

on $\mathcal{E} = [0, t_0] \times \mathcal{E}_x$ *converge in distribution to a Poisson process* $\mathcal{P}$ *with mean measure*

$$\mu\{(t', t) \times [x, \infty)\} = (t - t')\Lambda(x), \quad 0 \le t' < t \le t_0, \quad x \in \mathcal{E}_x = \{x' \in \mathbb{R} : \Lambda(x') < \infty\}, \tag{20}$$

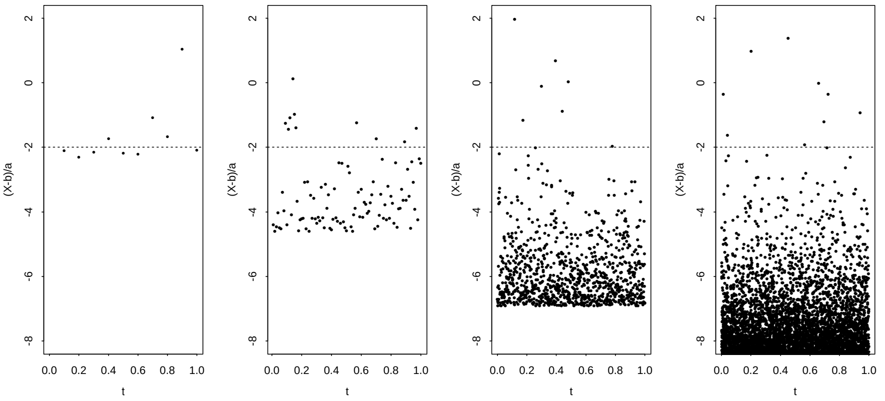
where

$$\Lambda(x) = \left(1 + \xi\frac{x - \eta}{\tau}\right)_+^{-1/\xi}$$

depends on parameters $\eta, \xi \in \mathbb{R}$ *and* $\tau > 0$ *and* $a_+ = \max(a, 0)$ *for real* a. *The corresponding intensity function is*

$$-\dot{\Lambda}(x) = \tau^{-1}\left(1 + \xi\frac{x - \eta}{\tau}\right)_+^{-1/\xi - 1} \ge 0.$$

Point process limit



☐ Here $\mathcal{E} \subset \mathbb{R}^D$, so we only need Kallenberg's theorem: for $\mathcal{A} \subset \mathcal{E}$, let $N_n(\mathcal{A}) = |\mathcal{P}_n \cap \mathcal{A}|$. Then if $\mathcal{B} \subset \mathcal{E}$ is any rectangle, and $\mathcal{C}$ is any finite union of disjoint rectangles of $\mathcal{E}$, and if

$$\mathrm{E}\{N_n(\mathcal{B})\} \to \mathrm{E}\{N(\mathcal{B})\}, \quad \mathrm{P}\{N_n(\mathcal{C}) = 0\} \to \mathrm{P}\{N(\mathcal{C}) = 0\}, \quad n \to \infty, \tag{21}$$

then $\mathcal{P}_n \xrightarrow{D} \mathcal{P}$ as $n \to \infty$.

Forms of $\Lambda(x)$

- ☐ $\Lambda(x)$ is decreasing, but has three distinct forms:
  - – when $\xi > 0$,

$$\Lambda(x) = \begin{cases} +\infty, & x \le \eta - \tau/\xi, \\ \left(1 + \xi\frac{x-\eta}{\tau}\right)_+^{-1/\xi}, & x > \eta - \tau/\xi, \end{cases}$$

    which is finite only for $x > \eta - \tau/\xi$, so $\Lambda(\mathcal{A}) = +\infty$, giving infinite counts, for any set $\mathcal{A}$ that goes below $\eta - \tau/\xi$;
  - – for $\xi = 0$ we take the limit when $\xi \to 0$, giving

$$\Lambda(x) = \exp\{-(x-\eta)/\tau\}, \quad x \in \mathbb{R},$$

    which is finite for all x;
  - – when $\xi < 0$,

$$\Lambda(x) = \begin{cases} \left(1 + \xi\frac{x-\eta}{\tau}\right)_+^{-1/\xi}, & x < \eta - \tau/\xi, \\ 0, & x \ge \eta - \tau/\xi, \end{cases}$$

    which is finite for all x.
- ☐ When $\xi \le 0$ the limiting mass at $-\infty$ is infinite, so any compact set $\mathcal{A}$ considered must have a finite lower bound.

Implications: Maxima

- ☐ A rescaled block maximum $Y_n = \{\max(X_1, \ldots, X_n) - b_n\}/a_n$ satisfies

$$\begin{aligned} \mathrm{P}(Y_n \le y) &= \mathrm{P}[N_n\{(0,1) \times [y,\infty)\} = 0] \\ &\to \mathrm{P}[N\{(0,1) \times [y,\infty)\} = 0] \quad n \to \infty, \\ &= \exp[-\mu\{(0,1) \times [y,\infty)\}], \\ &= \exp\{-\Lambda(y)\}, \quad y \in \mathbb{R}, \end{aligned}$$

  so a block maximum has a limiting **generalized extreme-value (GEV)** distribution,

$$G(y) = \exp\left\{-\left(1 + \xi\frac{y-\eta}{\tau}\right)_+^{-1/\xi}\right\}, \quad y \in \mathbb{R}.$$

Implications: Threshold exceedances I

☐ Consider the ‘forgetting’ mapping $g : \mathbb{R}^2 \to \mathbb{R}$ with $g(t,x) = x$, giving the process of event sizes $\mathcal{P}^* = g(P)$ without their times. The mapping theorem (Theorem 16) implies that $\mathcal{P}^*$ is Poisson with mean measure

$$\mu^*\{[x,\infty)\} = \mu[g^{-1}\{[x,\infty)\}] = \mu\{[0,t_0] \times [x,\infty)\} = t_0\{1+\xi(x-\eta)/\tau\}_+^{-1/\xi}.$$

☐ The conditional property (Theorem 12) implies that conditional on $N^*(\mathcal{A}_u = [u,\infty)) = n$, these n threshold exceedances have the same distribution as n points generated independently on $\mathcal{A}_u$ with measure

$$\frac{\mu^*(\mathcal{A}_{u+x})}{\mu^*(\mathcal{A}_u)} = \frac{t_0\{1+\xi(x-u-\eta)/\tau\}_+^{-1/\xi}}{t_0\{1+\xi(u-\eta)/\tau\}_+^{-1/\xi}} = \left(1+\xi\frac{x}{\sigma_u}\right)_+^{-1/\xi}, \quad x > 0,$$

where $\sigma_u = \tau + \xi(u-\eta)$. This corresponds to the **generalized Pareto distribution (GPD)**.

☐ The mapping $g_1(t,x) = t$ giving the rescaled times of those events that exceed u clearly satisfies the conditions of the mapping theorem and has measure

$$\mu_1^*\{(s,t]\} = \mu[g_1^{-1}\{(s,t]\}] = \mu\{(s,t] \times (u,\infty]\} = (t-s)\Lambda(u),$$

and intensity function $\Lambda(u)$. Thus the rescaled times of those events whose sizes exceed u form a homogeneous Poisson process on $(0,t_0)$ with rate $\Lambda(u)$.

Implications: Threshold exceedances II

This yields two fitting approaches:

☐ estimate η, τ, and ξ directly by fitting the Poisson process likelihood

$$\exp\{-\mu(\mathcal{A}_u)\} \times \prod_{j=1}^{n} \dot{\mu}(t_j, x_j),$$

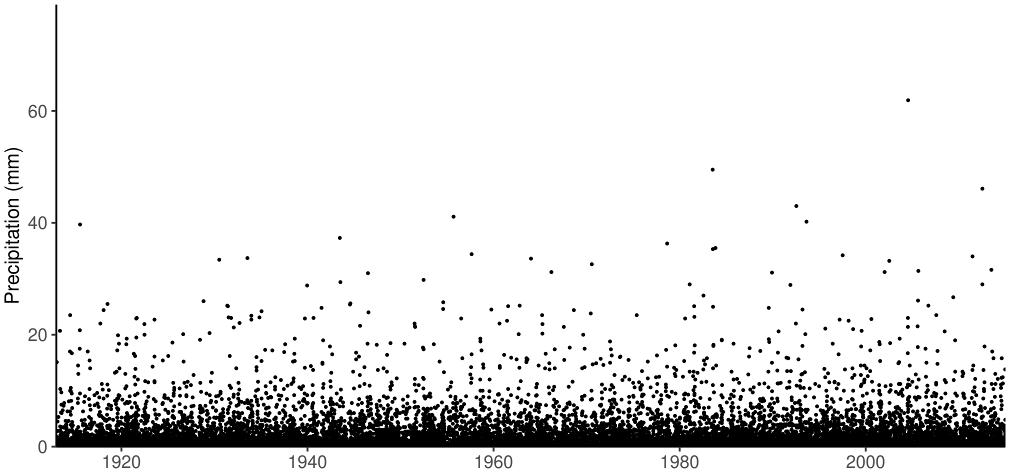
for the region $\mathcal{A}_u = [0,t_0] \times [u,\infty)$ containing $\{x_1,\ldots,x_n\}$ and where the intensity function for threshold exceedances is

$$\begin{aligned}\dot{\mu}(t,x) &= \frac{\partial^2 \mu\{[s,t]\times[u,x]\}}{\partial t \partial x} \\ &= \frac{\partial^2}{\partial t \partial x}\left((t-s)\left[\{1+\xi(u-\eta)/\tau\}_+^{-1/\xi} - \{1+\xi(x-\eta)/\tau\}_+^{-1/\xi}\right]\right) \\ &= \frac{1}{\tau}\{1+\xi(x-\eta)/\tau\}_+^{-1/\xi-1}.\end{aligned}$$

☐ estimate σ_u and ξ from the exceedances and p_u from the number of exceedances, n_u.

Abisko daily rainfall data

☐ Daily precipitation in Abisko, in northern Sweden, 1913–2014. The largest value is 61.9 mm, but many values are zero and most of the positive values are quite small.

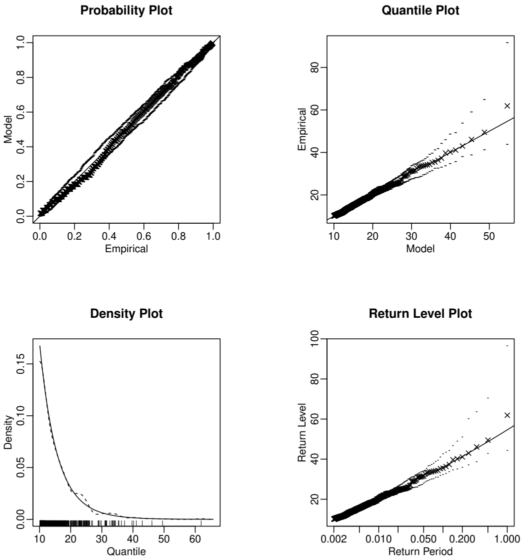


Abisko Poisson process fit

```
(fit.pp <- fpot(abisko$precip, threshold=10, model="pp", npp=365.25,
   start=list(loc=20,scale=6.5,shape=0.01)))
# needs initial values and number of points/block

Deviance: 2241.606

Threshold: 10
Number Above: 499
Proportion Above: 0.0134

Estimates
     loc     scale     shape
19.79658   6.52110   0.07026

Standard Errors
    loc    scale    shape
0.55597  0.37895  0.05088

Optimization Information
  Convergence: successful
  Function Evaluations: 20  ...  Gradient Evaluations: 8
```

Abisko Poisson process fit

☐ Let's check the fit using `plot(fit.pp)`:



Note 1 to Theorem 20

☐ If a limiting distribution G for rescaled maxima exists, then

$$\begin{aligned}\mathrm{P}[\{\max(X_1,\ldots,X_n)-b_n\}/a_n\le y] &= \mathrm{P}\{\max(X_1,\ldots,X_n)\le b_n+a_ny\}\\ &= F^n(b_n+a_ny)\\ &= \left[1-\frac{n\{1-F(b_n+a_ny)\}}{n}\right]^n.\end{aligned}$$

Hence a limiting function $\Lambda(y)$ must exist such that

$$\Lambda_n(y)=n\{1-F(b_n+a_ny)\}\to\Lambda(y),\quad n\to\infty.$$

☐ Let $H(x)=-\log\{1-F(x)\}$ denote the *cumulative hazard function* corresponding to F, and choose $b_n=b_n^*$ such that $H(b_n^*)=-\log n$, so that

$$\log\Lambda_n(y)=H(b_n+a_ny)-H(b_n).$$

☐ We suppose that F is continuous, places probability in an interval $[x_*,x^*]$, where either or both limits might be infinite, F is not defective (so there is no mass at x^*), that H is twice continuously differentiable with reciprocal hazard function $r(x)=1/H'(x)$, and that $\lim_{x\to x^*}r'(x)=\xi$ is real and finite. These are sometimes called the *von Mises conditions*.

☐ Then

$$H(b_n+a_ny)-H(b_n)=a_n\int_0^y\frac{1}{r(b_n+a_nx)}\,\mathrm{d}x=a_n\int_0^y\frac{1}{r(b_n)+a_nxr'\{b_n+s_n(x)\}}\,\mathrm{d}x,$$

where $s_n(x)$ lies between zero and x. If we now choose $a_n=a_n^*=r(b_n^*)$, which is positive because $r(x)=\{1-F(x)\}/f(x)$, we have

$$H(b_n^*+a_n^*y)-H(b_n^*)=\int_0^y\frac{1}{1+xr'\{b_n^*+s_n(x)\}}\,\mathrm{d}x=\int_0^y\frac{1}{1+\xi_nx}g_n(x)\,\mathrm{d}x,$$

where $\xi_n=r'(b_n^*)$ and $g_n(x)=(1+\xi_nx)/\{1+xr'\{b_n^*+s_n(x)\}$.

☐ The implicit function theorem implies that $s_n(x)$ is continuous in x and so is r', so $g_n(x)$ is continuous in x, and one can check that $g_n(x)\to1$ as $n\to\infty$. Hence in the interval where $1+\xi_nx$ does not change sign, we can use a mean value theorem for integrals and choose y^* such that

$$H(b_n^*+a_n^*y)-H(b_n^*)=g_n(y^*)\int_0^y\frac{1}{1+\xi_nx}\,\mathrm{d}x=g_n(y^*)\times\xi_n^{-1}\log(1+\xi_nx)_+,$$

where we add the $(\cdot)_+$ to remind us that the term in brackets must be positive. Now $\xi_n=r'(b_n^*)\to\xi$ and $0<y^*<y$ as $n\to\infty$, so

$$\lim_{n\to\infty}H(b_n^*+a_n^*y)-H(b_n^*)=\xi^{-1}\log(1+\xi y)_+=\log\Lambda(y),$$

as required. This establishes sufficient conditions under which a maximum has limiting distribution $\exp\{-\Lambda(y)\}$, with $\eta=0$ and $\tau=1$. We need the more general case to allow for the fact that b_n and a_n are unknown in applications (because F is unknown).

note 1 of slide 115

Note 2 to Theorem 20

☐ To establish the Poisson convergence, define the binomial processes

$$\mathcal{P}_n = \{(j/(n+1), (X_j - b_n)/a_n) : j = 1, \ldots, nt_0\}, \quad n = 1, 2, \ldots,$$

and the corresponding count process $N_n(\cdot)$ on $\mathcal{E} = [0, t_0] \times \mathcal{E}_x$.

☐ Let $0 < t_1 < t_2 \leq t_0$ and $x_1 < x_2$ determine the rectangle $\mathcal{A} = (t_1, t_2] \times (x_1, x_2]$, let

$$\mu(\mathcal{A}) = (t_2 - t_1)\{\Lambda(x_1) - \Lambda(x_2)\}, \quad \mathcal{A} \subset \mathcal{E},$$

and let $\mathcal{P}$ denote a Poisson process on $\mathcal{E}$ with mean measure μ.

☐ We now check Kallenberg's conditions. If $\lfloor x \rfloor$ is the integer part of x, then

$$\begin{aligned} \mathrm{E}\{N_n(\mathcal{A})\} &= \lfloor (n+1)t_2 - (n+1)t_1 \rfloor \times \mathrm{P}\{x_1 < (X_j - b_n)/a_n \leq x_2\} \\ &= \frac{\lfloor (n+1)(t_2 - t_1) \rfloor}{n} \times \Lambda_n(x_1, x_2] \\ &\to (t_2 - t_1)\Lambda(x_1, x_2] = \mu(\mathcal{A}), \quad n \to \infty, \end{aligned}$$

which verifies the first condition.

☐ For the second condition, let $\mathcal{C}$ be a union of a finite number of disjoint rectangles of $\mathcal{E}$, and note that we can write $\mathcal{C} = \bigcup_{i=1}^{k} \mathcal{T}_i \times \bigcup_{l=1}^{L_i} \mathcal{X}_{i,l}$, where the $\mathcal{T}_i \subset [0, t_0]$ are disjoint intervals, and the intervals $\mathcal{X}_{i,l} \subset \mathbb{R}$ are disjoint for each i. Let $\mathcal{T}_1 = (t_1, t_2]$, let $\mathcal{X}_i = \bigcup_{l=1}^{L_i} \mathcal{X}_{i,l}$ and $\mathcal{B}_i = \mathcal{T}_i \times \mathcal{X}_i$, and note that independence and identical distribution of the X_j gives

$$\begin{aligned} \mathrm{P}\{N_n(\mathcal{B}_1) = 0\} &= \mathrm{P}\{(X_1 - b_n)/a_n \notin \mathcal{X}_1\}^{\lfloor (n+1)(t_2 - t_1) \rfloor} \\ &= \left[\left\{1 - \frac{\Lambda_n(\mathcal{X}_1)}{n}\right\}^n\right]^{\lfloor (n+1)(t_2 - t_1) \rfloor / n} \\ &\to \exp\{-|\mathcal{T}_1|\Lambda(\mathcal{X}_1)\}, \quad n \to \infty, \\ &= \exp\{-\mu(\mathcal{T}_1 \times \mathcal{X}_1)\}. \end{aligned}$$

This applies for each $\mathcal{T}_i$, and the corresponding variables X_j are independent, so

$$\begin{aligned} \mathrm{P}\{N_n(\mathcal{C}) = 0\} &= \prod_{i=1}^{k} \mathrm{P}\{N_n(\mathcal{B}_i) = 0\} \\ &\to \prod_{i=1}^{k} \exp\{-\mu(\mathcal{T}_i \times \mathcal{X}_i)\}, \quad n \to \infty, \\ &= \exp\left\{-\sum_{i=1}^{k} \mu(\mathcal{B}_i)\right\} \\ &= \exp\{-\mu(\mathcal{C})\}, \end{aligned}$$

which establishes the second condition. Thus $\mathcal{P}_n \xrightarrow{D} \mathcal{P}$.

4 Complications

slide 116

4.1 Introduction

slide 117

Basic ideas

☐ Chapters 2 and 3 described 'vanilla' statistical analyses for rare events using the GEV, GPD, and point process methods.

☐ The basic derivations of these models assume that

$$X_1, \ldots, X_m \overset{\text{iid}}{\sim} F, \quad m \to \infty.$$

☐ In applications these assumptions are generally false:

- m is finite;
- the background data may show trend, seasonality or other forms of **non-stationarity**, so $X_j \sim F_j$;
- time series are typically **dependent**, as cold weather, heatwaves, . . . occur over several days;
- some (maybe subtle) **selection** mechanism may apply, e.g., when an analysis is performed immediately after a rare event.

☐ This chapter will describe methods for detecting and dealing with these problems.

slide 118

4.2 Nonstationarity

slide 119

Vanilla analysis of maxima

☐ Our previous analyses supposed that

- block maxima satisfy $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{GEV}(\eta, \tau, \xi)$,
- exceedances of a threshold u satisfy $X_1 - u, \ldots, X_n - u \overset{\text{iid}}{\sim} \text{GPD}(\sigma, \xi)$,

but often we observe additional variation, either due to

- **systematic** changes in the background data (e.g., due to trend or seasonality), or to
- **haphazard** variation (e.g., due to weather conditions) that we have not accounted for.

☐ We'll pass most time looking at systematic changes.

☐ For an example of haphazard variation, consider annual maximum daily rainfall $M = \max(X_1, \ldots, X_{365})$, where X_j is total rainfall on day j. On many days $X_j = 0$, so

$$M = \max(X_1, \ldots, X_N),$$

where $N \ll 365$ is the (random) number of rainy days. If N varies a lot from year to year, then M might be much smaller in some years than in others, so the GEV is a poor model (remember we derive it assuming that $X_j \sim F$, where F is continuous . . .).

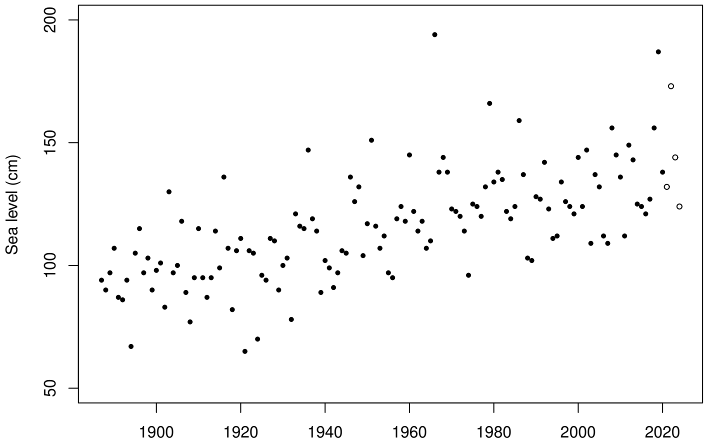
slide 120

A damp day in Venice

Punta della Dogana and Santa Maria della Salute

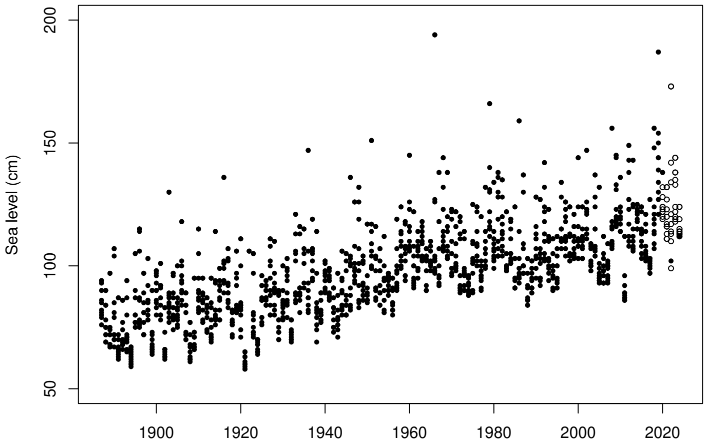
Annual maximum sea levels, 1887–2024

In October 2020, the MOdulo Sperimentale Elettromeccanico (MOSE) system was inaugurated: rows of mobile gates are raised when particularly high tides are predicted, in order to limit how much water from the Adriatic Sea can enter the Venetian lagoon. Data with MOSE operational are shown by circles. The record: 196 cm in 1996.



Ten largest annual sea levels, 1887–2024

In 1935, only the six largest values are available, and in 1922 only the largest value is available. The data sources for 1887–1981 and 1982 onwards are different.



Non-stationarity

- ☐ Obvious approach is to suppose that the GEV parameters can depend on external factors, i.e., $Y_t \sim \text{GEV}(\eta_t, \tau_t, \xi_t)$, where the dependence might be specified as

$$\begin{aligned}
\eta_t(\beta) &= \beta_0 + \beta_1 t, \\
\eta_t(\beta) &= \beta_0 + \sum_{k=1}^{K} \{\beta_{2k-1}\cos(2\pi kt/365) + \beta_{2k}\sin(2\pi kt/365)\}, \\
\eta_t(\beta) &= \beta_0 + \beta_1 x(t), \\
\tau_t(\beta) &= \exp(\beta_0 + \beta_1 t), \\
\xi_t(\beta) &= \begin{cases} \beta_1, & t \le t_0, \\ \beta_2, & t > t_0, \end{cases}
\end{aligned}$$

  where $x(t)$ is some physical quantity that varies over time (e.g., ENSO, NAO, or global average temperature).
- ☐ In applications we typically find that
  - the location parameter η varies,
  - the scale parameter τ might or might not vary,
  - the shape parameter ξ is constant (it is difficult to estimate, and anyway often is regarded as an intrinsic aspect of the background process).

Parametric inference

- ☐ Example model specification: $y_t \overset{\text{ind}}{\sim} \text{GEV}(\eta_t, \tau_t, \xi_t)$, where η_t, τ_t, ξ_t depend on parameters β.
- ☐ If $y_1, \ldots, y_n$ are assumed to be independent, then the log likelihood for β is

$$\ell(\beta) = \sum_{t=1}^{n} \log g\{y_t; \eta_t(\beta), \tau_t(\beta), \xi_t(\beta)\},$$

  where g is the GEV density.
- ☐ Maximization of $\ell(\beta)$ yields maximum likelihood estimates and the observed information matrix, from which we compute standard errors, confidence intervals, etc.
- ☐ We say that model $\mathcal{M}_0$ is nested within a model $\mathcal{M}_1$ if $\mathcal{M}_1$ reduces to $\mathcal{M}_0$ by fixing (say) d parameters. Then the corresponding maximised log likelihoods satisfy $\widehat{\ell}_1 \ge \widehat{\ell}_0$, and the likelihood ratio statistic (or equivalently difference in deviances) is

$$W = 2(\widehat{\ell}_1 - \widehat{\ell}_0).$$

- ☐ If $\mathcal{M}_0$ is adequate, then asymptotic likelihood theory implies that $W \dot{\sim} \chi^2_d$, so values of W larger than the $1-\alpha$ quantile of the χ^2_d distribution would lead to a rejection of $\mathcal{M}_0$ in favour of $\mathcal{M}_1$, at significance level α.

Model diagnostics

☐ If $Y_t \sim \text{GEV}(\eta_t, \tau_t, \xi_t)$ for $t = 1, \dots, n$, then

$$Z_t = \frac{1}{\xi_t} \log\left(1 + \xi_t \frac{Y_t - \eta_t}{\tau_t}\right) \quad \overset{\text{iid}}{\sim} \quad \text{standard Gumbel},$$

i.e.,

$$\text{P}(Z_t \le z) = \exp\{-\exp(-z)\}, \quad z \in \mathbb{R}, \quad t = 1, \dots, n.$$

☐ If we replace the parameters by their estimates $\widehat{\eta}_t = \eta_t(\widehat{\beta})$, etc., these results should still hold (approximately) for the **Gumbel residuals**

$$\widehat{z}_t = \frac{1}{\widehat{\xi}_t} \log\left(1 + \widehat{\xi}_t \frac{y_t - \widehat{\eta}_t}{\widehat{\tau}_t}\right), \quad t = 1, \dots, n.$$

☐ We use the $\widehat{z}_t$ in diagnostic plots, e.g.,

– the **probability plot**, showing $\{j/(n+1), \exp\{-\exp(-\widehat{z}_{(j)})\};\ j = 1, \dots, n\}$, or

– the **quantile plot**, showing $\left\{\left(-\log\left[-\log\{j/(n+1)\}\right], \widehat{z}_{(j)}\right);\ j = 1, \dots, n\right\}$, or

plots of the $\widehat{z}_j$ against appropriate variables, to see if any patterns remain after fitting the model.

Example: Venice sea levels, 1887–2019

☐ We ignore the data from 2020 onwards, when MOSE is operational.

☐ Analysis of maxima uses straight-line regression model,

$$\eta_t = \beta_0 + \beta_1 x_t, \quad t = 1, \dots, n = 133,$$

with $(x_1, \dots, x_{133}) = (1887 - 1900, \dots, 2019 - 1900)/100$ chosen so that

– β_0 equals the location parameter in the year 1900,

– β_1 denotes the change in maximum sea level over 100 years,

☐ We fit two nested models, both with constant scale and shape parameters, i.e.,

$$\mathcal{M}_0: \quad \eta_t = \beta_0, \quad \tau_t \equiv \tau, \quad \xi_t \equiv \xi,$$
$$\mathcal{M}_1: \quad \eta_t = \beta_0 + \beta_1 x_t, \quad \tau_t \equiv \tau, \quad \xi_t \equiv \xi.$$

☐ The code prints a ‘deviance’ $D = -2\widehat{\ell}$ (or `nllh`$= -\widehat{\ell}$) for the fitted model, which allows model comparison using the likelihood ratio statistic:

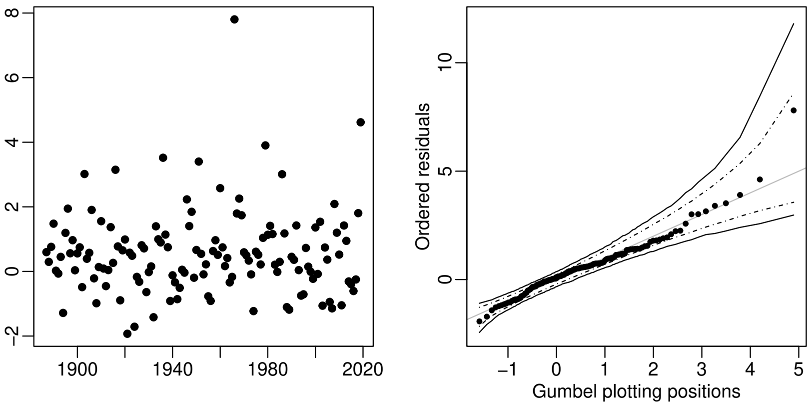
$$w = 2(\widehat{\ell}_1 - \widehat{\ell}_0) = D_0 - D_1.$$

Example: Fitting models

```
y <- venice$y[venice$year<2020,1]
x <- (venice$year[venice$year<2020]-1900)/100
(fit0 <- evd::fgev(y))

Call: evd::fgev(x = y)
Deviance: 1193.487

Estimates
    loc    scale    shape
106.517   20.050   -0.139

Standard Errors
    loc    scale    shape
1.89487  1.29297  0.04412

Optimization Information
  Convergence: successful ...

(fit1 <- evd::fgev(y,nsloc=x))  # nsloc specifies the x variable for the non-stationary locat

Call: evd::fgev(x = y, nsloc = x)
Deviance: 1122.072

Estimates
     loc  loctrend     scale     shape
 89.8087   35.0291   15.0816   -0.1023

Standard Errors
     loc  loctrend     scale     shape
 2.34431   3.51218   0.96584   0.04071

Optimization Information
  Convergence: successful ...
```

Example: Venice sea levels, 1887–2019

Model-checking for fit to Venice maximum sea-level data. Left panel: Gumbel-scale residuals, $\widehat{z}_t$. Right: ordered $\widehat{z}_t$ plotted against Gumbel plotting positions, with pointwise (dot-dash) and overall (solid) 95% confidence bands obtained by simulating 10,000 Gumbel samples.



r-largest analysis

☐ The limiting joint density of the largest r order statistics $Y_1 > \cdots > Y_r$ in a large sample is

$$\exp\{-\Lambda(y_r)\} \times \prod_{j=1}^{r} \left\{-\dot{\Lambda}(y_j)\right\}, \quad y_1 > \cdots > y_r,$$

where $\Lambda(y;\eta,\tau,\xi) = \{1+\xi(y-\eta)/\tau\}_+^{-1/\xi}$ and $\dot{\Lambda}(y) = \mathrm{d}\Lambda(y)/\mathrm{d}y$.

☐ In the Venice data there are r_t (usually 10) largest values in each year, say

$$y_{t,1} > \cdots > y_{t,r_t}, \quad t = 1,\ldots,n,$$

so if the data for different years are independent, and if we again use parameters $\eta_t(\beta)$, $\tau_t(\beta)$, $\xi_t(\beta)$ in year t, the likelihood is

$$L(\beta) = \prod_{t=1}^{n} \exp\left[-\Lambda\{y_{t,r_t};\eta_t(\beta),\tau_t(\beta),\xi_t(\beta)\}\right] \times \prod_{j=1}^{r_t} \left[-\dot{\Lambda}\{y_{t,j};\eta_t(\beta),\tau_t(\beta),\xi_t(\beta)\}\right].$$

☐ To fit the model we just maximise the corresponding log likelihood, compute the observed information matrix, and proceed as before ...

☐ Of course this assumes that this is a reasonable model, which might be questioned: are the background data really independent?

r-largest: model-checking

☐ The transformed values $\Lambda(Y_j)$ form a Poisson process of unit rate on the positive half-line:

$$0 < \Lambda(Y_1) < \Lambda(Y_2) < \Lambda(Y_3) < \cdots,$$

so

$$\Lambda(Y_1), \quad \Lambda(Y_2) - \Lambda(Y_1) \quad \Lambda(Y_3) - \Lambda(Y_2), \ldots \overset{\text{iid}}{\sim} \exp(1).$$

☐ Recall that if $E \sim \exp(1)$, then $-\log E$ has a standard Gumbel distribution.

☐ Hence if the model is adequate and we replace the parameters by their estimates, the

$$-\log\left\{\widehat{\Lambda}(Y_j) - \widehat{\Lambda}(Y_{j-1})\right\}, \quad j = 2, \ldots,$$

should be approximately independent Gumbel variables.

☐ The theory above is OK in principle, but if the observations are heavily rounded, the values of $\widehat{\Lambda}(Y_j)$ might be very similar, so that the approach above fails. Alternatively we might note that

$$\Lambda(Y_j) = \Lambda(Y_1) + \sum_{i=1}^{j-1}\{\Lambda(Y_{i+1}) - \Lambda(Y_i)\} \sim \text{Gamma}(j, 1), \quad j = 1, \ldots, r,$$

and then compare the ordered values of the $\Lambda(Y_j)$ for different years with quantiles of the gamma distribution.

Venice data

☐ We shall use the r largest observations to fit several models:

$$\mathcal{M}_0: \quad \eta_t = \beta_0, \quad \tau_t \equiv \tau, \quad \xi_t \equiv \xi,$$
$$\mathcal{M}_1: \quad \eta_t = \beta_0 + \beta_1 x_t, \quad \tau_t \equiv \tau, \quad \xi_t \equiv \xi,$$
$$\mathcal{M}_2: \quad \eta_t = \beta_0 + \beta_1 x_t + \beta_2 I(t > 1981), \quad \tau_t \equiv \tau, \quad \xi_t \equiv \xi,$$
$$\mathcal{M}_3: \quad \eta_t = \beta_0 + \beta_1 x_t + \beta_3 \cos(2\pi t/18.6) + \beta_4 \sin(2\pi t/18.6), \quad \tau_t \equiv \tau, \quad \xi_t \equiv \xi.$$

☐ Reasoning:

- we expect the baseline IID model $\mathcal{M}_0$ to be terrible (there is an obvious trend);
- we expect $\mathcal{M}_1$ to be much better than $\mathcal{M}_0$, as it allows for the trend;
- if $\mathcal{M}_2$ improves significantly on $\mathcal{M}_1$ then the data sources pre- and post-1982 disagree;
- $\mathcal{M}_3$ is suggested by the discussion in Pirazzoli (1982, *Acqua Aria*) who suggested that an 18.6-year astronomical cycle may influence the maxima.

☐ We could add trend in the scale and shape parameters, but will avoid this here.

☐ We take $r = 2$ for now.

Example: Venice sea levels

Code to fit the models using the function `rlarg.fit` of the `ismev` package:

```
y <- venice$y[venice$year<2020,]
year <- venice$year[venice$year<2020]
X <- cbind(x,(year>=1982),cos(2*pi*year/18.6),sin(2*pi*year/18.6) )

> head(y)  # 10 largest values for each year, starting in 1887
   y1  y2 y3 y4 y5 y6 y7 y8 y9 y10
1  94  93 90 86 85 82 81 80 79  76
2  90  84 84 78 75 75 72 72 69  69
3  97  75 74 72 72 68 68 68 67  67
4 107 104 85 81 79 72 72 70 70  67
5  87  72 70 67 66 66 65 64 63  62
6  86  77 74 70 70 69 69 68 68  66

> head(X)  # matrix of covariates for different fits
         x
[1,] -0.13 0 -0.9541393  0.29936312
[2,] -0.12 0 -0.9994295 -0.03377414
[3,] -0.11 0 -0.9317526 -0.36309386
[4,] -0.10 0 -0.7587581 -0.65137248
[5,] -0.09 0 -0.5000000 -0.86602540
[6,] -0.08 0 -0.1847261 -0.98279005

fit0 <- ismev::rlarg.fit(xdat=y, r=2)

fit1 <- ismev::rlarg.fit(xdat=y, r=2, ydat=X, mul=c(1)) # mul says which columns of X to use

fit2 <- ismev::rlarg.fit(xdat=y, r=2, ydat=X, mul=c(1,2))

fit3 <- ismev::rlarg.fit(xdat=y, r=2, ydat=X, mul=c(1,3,4))
```

Fits of $\mathcal{M}_0$ and $\mathcal{M}_1$

```
> fit0 <- ismev::rlarg.fit(xdat=y, r=2)
$conv
[1] 0
$nllh
[1] 1035.521
$mle
[1] 112.1400606  18.3991733  -0.1485854
$se
[1] 1.48482972 0.80384786 0.03128211

> fit1 <- ismev::rlarg.fit(xdat=y, r=2, ydat=X, mul=c(1))
$model
$model[[1]]
[1] 1
$model[[2]]
NULL
$model[[3]]
NULL

$link
[1] "c(identity, identity, identity)"
$conv
[1] 0
$nllh
[1] 973.297
$mle
[1] 93.946101 31.725728 14.160470 -0.103406
$se
[1] 1.68660910 2.42460867 0.65001011 0.03125164
```

Fits of $\mathcal{M}_2$ and $\mathcal{M}_3$

```
fit2 <- ismev::rlarg.fit(xdat=y, r=2, ydat=X, mul=c(1,2))
$model
$model[[1]]
[1] 1 2
...
$conv
[1] 0
$nllh
[1] 969.336
$mle
[1] 91.7010233 41.0735757 -9.7734237 13.8874345 -0.1025608
$se
[1] 1.84707056 4.05490122 3.40265002 0.64056193 0.03303069

> fit3 <- ismev::rlarg.fit(xdat=y, r=2, ydat=X, mul=c(1,3,4))
$model
$model[[1]]
[1] 1 3 4
...
$conv
[1] 0
$nllh
[1] 973.0675
$mle
[1] 93.8443140 31.9199292 -0.4317245 -0.7943716 14.1372635 -0.1050128
$se
[1] 1.69570726 2.44300955 1.36448615 1.34041204 0.64842570 0.03166116
```

Example: Venice sea levels, 1887–2019

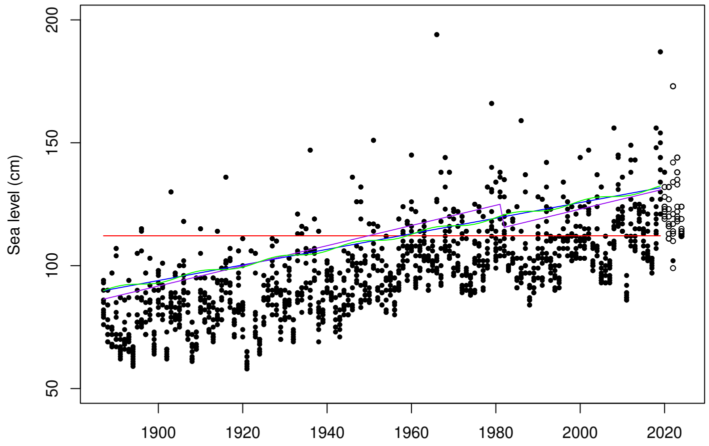
Summaries of fitted models for Venice sea level data analysis, with estimates$_{\mathrm{SEs}}$:

Model	r	$-2\widehat{\ell}$	ξ	τ (cm)	β_0 (cm)	β_1 (cm/century)	β_2 (cm)	β_3 (cm)	β_4 (cm)
$\mathcal{M}_1$	1	1122.07	$-0.102_{0.041}$	$15.1_{0.97}$	$89.8_{2.3}$	$35.0_{3.5}$			
$\mathcal{M}_1$	2	1946.59	$-0.103_{0.031}$	$14.2_{0.65}$	$94.0_{1.7}$	$31.7_{2.4}$			
$\mathcal{M}_1$	3	2605.49	$-0.106_{0.025}$	$13.3_{0.52}$	$95.9_{1.4}$	$31.5_{1.9}$			
$\mathcal{M}_1$	4	3185.07	$-0.104_{0.022}$	$12.8_{0.46}$	$96.8_{1.3}$	$31.3_{1.7}$			
$\mathcal{M}_1$	9	5263.20	$-0.090_{0.016}$	$11.5_{0.36}$	$98.2_{1.0}$	$30.3_{1.1}$			
$\mathcal{M}_0$	2	2071.04	$-0.149_{0.031}$	$18.4_{0.80}$	$112.1_{1.5}$				
$\mathcal{M}_1$	2	1946.59	$-0.103_{0.031}$	$14.2_{0.65}$	$94.0_{1.7}$	$31.7_{2.4}$			
$\mathcal{M}_2$	2	1938.67	$-0.103_{0.033}$	$13.9_{0.64}$	$91.7_{1.9}$	$41.1_{4.0}$	$-9.8_{3.4}$		
$\mathcal{M}_3$	2	1946.14	$-0.105_{0.032}$	$14.1_{0.65}$	$93.8_{1.7}$	$31.9_{2.4}$		$-0.43_{1.46}$	$-0.79_{1.34}$
$\mathcal{M}_4$	2	1924.37	$0.057_{0.063}$	$10.6_{0.87}$	$93.2_{2.0}$	$40.0_{4.3}$	$-8.4_{3.6}$		

Note that:

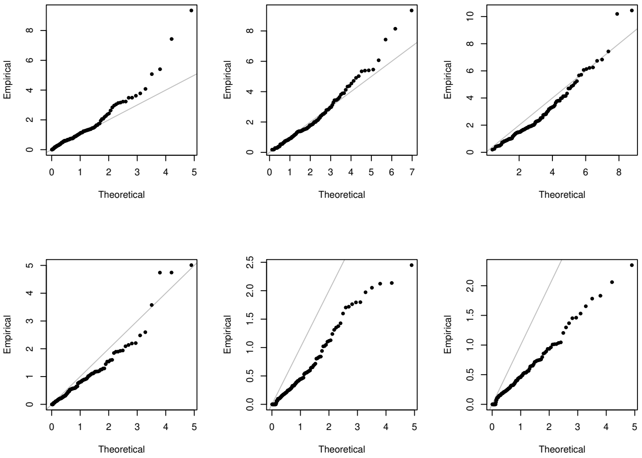
- ☐ the log likelihoods are only comparable for the same values of r, because different values of r use different subsets of the data;
- ☐ if the data were independent, we'd expect the SEs for $r = 1$ to reduce by factors of roughly 2 and 3 for $r = 4$ and $r = 9$;
- ☐ there is strong evidence for trend (surprise!), a change due to the data sources, and $\xi < 0$;
- ☐ there is no evidence of the astronomical cycle.

Example: Venice sea levels, 1887–2019



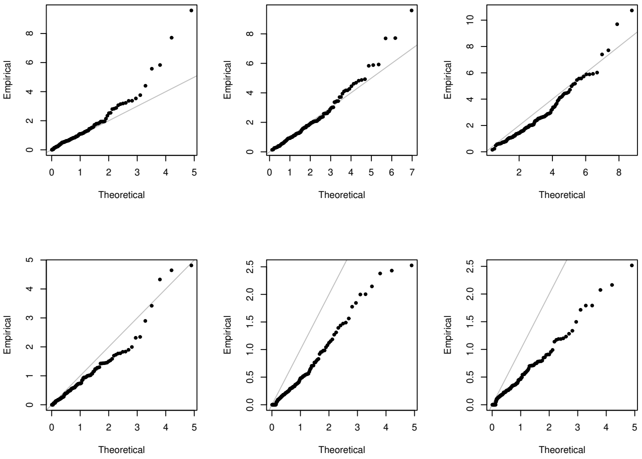
Largest ten annual sea levels at Venice, with fits from models $\mathcal{M}_0$ (red), $\mathcal{M}_1$ (blue), $\mathcal{M}_2$ (purple), and $\mathcal{M}_3$ (green), when $r = 2$.

Example: Venice sea levels, 1887–2019



Residual plots for fit of $\mathcal{M}_1$ with $r = 2$: Top row: comparison of $\widehat{\Lambda}(y_j)$ with corresponding gamma distributions for $j = 1, 2, 3$. Bottom row: comparison of $\widehat{\Lambda}(y_{j+1}) - \widehat{\Lambda}(y_j)$ for $j = 1, 2, 3$. The model does not seem to fit very well!

Example: Venice sea levels, 1887–2019



Residual plots for fit of $\mathcal{M}_2$ with $r = 2$: Top row: comparison of $\widehat{\Lambda}(y_j)$ with corresponding gamma distributions for $j = 1, 2, 3$. Bottom row: comparison of $\widehat{\Lambda}(y_{j+1}) - \widehat{\Lambda}(y_j)$ for $j = 1, 2, 3$. The model does not seem to fit very well!

Haphazard variation

☐ Above we modelled **systematic variation** by allowing the parameters to depend on known quantities.

☐ There may be **haphazard variation** that can be modelled by adding extra randomness.

☐ Suppose that conditional on ε, the data have rate $\varepsilon\Lambda(y)$, where $\varepsilon \sim$ Gamma$(\nu, 1/\nu)$, i.e.,

$$f(\varepsilon) = \frac{\nu^\nu}{\Gamma(\nu)}\varepsilon^{\nu-1}e^{-\nu\varepsilon}, \quad \varepsilon > 0, \quad \nu > 0,$$

which is the usual gamma density with parameters $\alpha = \nu$, $\lambda = \nu$, so $\mathrm{E}(\varepsilon) = \alpha/\lambda = 1$, $\mathrm{var}(\varepsilon) = \alpha/\lambda^2 = 1/\nu \to 0$ as $\nu \to \infty$. Hence the baseline model corresponds to $\nu = \infty$.

☐ The marginal density for $Y_1 > \cdots > Y_r$ is then

$$f(y_1, \ldots, y_r) = \int_0^\infty f(y_1, \ldots, y_r \mid \varepsilon) f(\varepsilon)\,\mathrm{d}\varepsilon = \cdots = \prod_{j=1}^r \{-\dot\Lambda(y_j)\} \frac{\Gamma(\nu + r)}{\Gamma(\nu)\nu^r} \frac{1}{\{1 + \Lambda(y_r)/\nu\}^{\nu+r}},$$

so in particular the maximum has density

$$f(y_1) = \{-\dot\Lambda(y_1)\} \frac{1}{\{1 + \Lambda(y_1)/\nu\}^{\nu+1}},$$

i.e., a model with parameters (η, τ, ξ, ν), where $\nu \to \infty$ gives the basic model.

A better model?

☐ Take $r = 2$ and fit

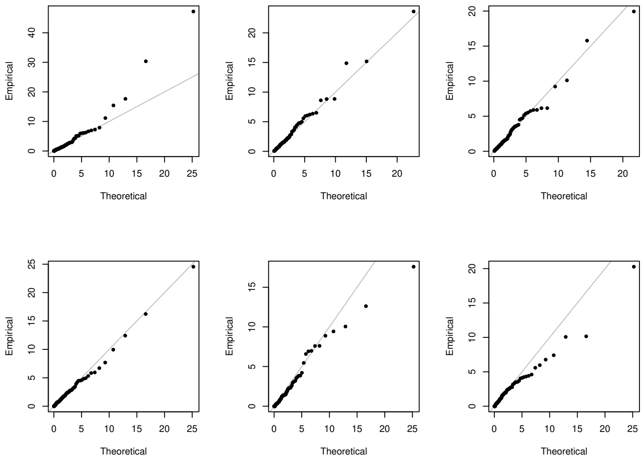
$$\mathcal{M}_4: \quad \eta_t = \beta_0 + \beta_1 x_t + \beta_2 I(t > 1981), \quad \tau_t \equiv \tau, \quad \xi_t \equiv \xi, \quad \nu,$$

gives the results shown on slide 23:

- there is a big reduction in $-2\widehat{\ell}$, so the model is a clear improvement on the others: it is worthwhile to include ν;
- $\widehat{\nu} = 1.72_{0.63}$, giving strong evidence of overdispersion relative to the baseline model, with ε having standard deviation $\widehat{\nu}^{-1/2} = 0.76$;
- the estimates of the βs are similar, but $\widehat{\tau}$ is smaller and now $\widehat{\xi} \approx 0$, because including ν accounts for some of the variation not accounted for in the other models;
- most of the standard errors are larger, because of the additional variation that ν accommodates.

☐ Residual plots on the next slide are (somewhat) better, though the largest values are still too big relative to the model.

Example: Venice sea levels, 1887–2019



Residual plots for Venice data fit of $\mathcal{M}_4$ with $r = 2$: Top row: comparison of $\widehat{\Lambda}(y_j)$ with corresponding F distributions for $j = 1, 2, 3$. Bottom row: comparison of $\widehat{\Lambda}(y_{j+1}) - \widehat{\Lambda}(y_j)$ for $j = 1, 2, 3$. The model fits better than before, but still the largest values are not well modelled.

Measures of risk

☐ Under non-stationarity the quantiles and thus the return levels vary with time, so the interpretation as 'the level exceeded once on average every T years' needs thought ...

☐ Suppose there are m IID background observations in each block (year, say), but that their distributions F_t differ for the different blocks, and let M_T be the maximum for T blocks.

☐ If the maximum in year t has GEV distribution G_t, then $F_t \approx G_t^{1/m}$, and we solve

$$1 - p = \left\{ \prod_{t=1}^{T} F_t(x_p) \right\}^{1/T} = \left\{ \prod_{t=1}^{T} G_t^{1/m}(x_p) \right\}^{1/T} = \mathrm{P}(M_T \le x_p)^{1/(mT)}.$$

☐ Likewise, in the POT setup, we suppose that independent observations X_j have thresholds u_j, exceedance probabilities p_{u_j} and GP distributions $H_j(x) = 1 - \overline{H}_j(x)$, and then solve

$$1 - p = \left[\prod_{j=1}^{mT} \left\{ 1 - p_{u_j} \overline{H}_j(x_p - u_j) \right\} \right]^{1/mT} = \mathrm{P}(M_T \le x_p)^{1/(mT)}.$$

☐ Note that

- $\mathrm{P}(X > x_p)$ will vary over time, so x_p may not be a very useful summary of risk,
- both formulae reduce to the previous ones when the data are stationary,
- there are no explicit formulae for x_p, which must be found numerically.

Comments

☐ Similar techniques are applicable for the threshold exceedance and point process models, but the threshold may need to be time-varying and thus needs care.

☐ Using the r-largest model may be preferable, as the threshold is replaced with a choice of r.

☐ Under the GPD, changing the threshold $u \mapsto u'$ changes the scale parameter:

$$\sigma_u \mapsto \sigma_{u'} = \sigma_u + \xi(u' - u),$$

so, for example, the formulation

$$(\sigma_u, \xi) = (g(x_1), h(x_2))$$

at threshold u will become

$$(\sigma_{u'}, \xi) = \left(g(x_1) + h(x_2)(u - u'), h(x_2)\right),$$

at threshold u', so interpretation depends on threshold—undesirable.

☐ GEV, r-largest and Poisson process fits use the parameters (η, τ, ξ), invariant to type of model, which is preferable.

☐ In some investigations it is preferable to use the GP model: do increasing rainfall maxima come from increases in the number of days with heavy rainfall, but no changes in the amounts, or increased amounts when it rains?

4.3 Dependence

Modelling issues

☐ Environmental time series data typically show:
  - long–term trends (e.g., gradual climatic change);
  - seasonality (e.g., annual cycles in meteorology);
  - other forms of non–stationarity (e.g., the effect of ENSO or NAO); and
  - short term dependence (due to volatility, storms, . . .).

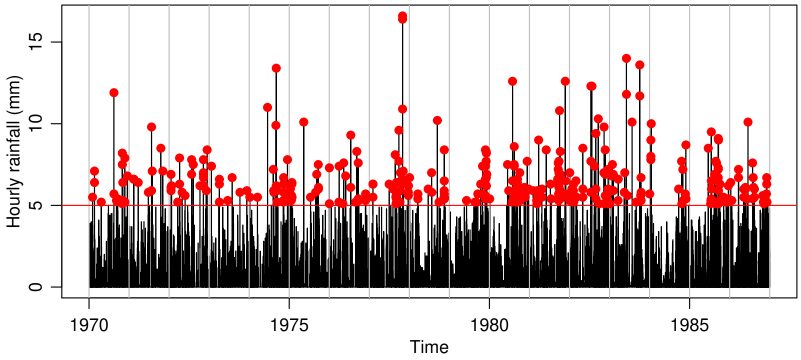
☐ We have discussed non-stationarity. Now we discuss dependence. In brief:
  - the previous limiting theory for maxima also applies, with small changes, provided long-range dependence of extremes is sufficiently weak; but
  - clustering of extremes due to short range dependence arises and must be dealt with.

☐ If the background data were independent, then the indicators $I(X_t > u)$ would be IID Bernoulli variables with probability p_u, say, and thus
  - for any $h = 1, 2, \ldots$ we would see

$$\mathrm{P}(X_{t+h} > u \mid X_t > u) = \mathrm{P}(X_{t+h} > u) = p_u;$$

  - intervals between exceedances would be IID geometric variables with mean $1/p_u$ (approximately exponential for small p_u).

Example: Eskdalemuir rainfall



These data show:

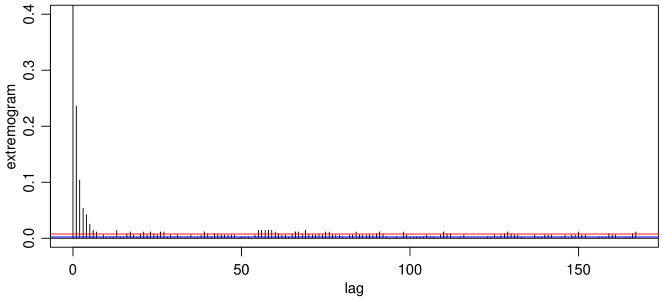
☐ apparent stationarity (with small seasonal changes?);

☐ long-range independence (rain in 1975 is independent of rain in 1980 . . .);

☐ short-range dependence, owing to clustering of hours with heavy rain?

It seems safe to assume weak dependence of extremes at long ranges, but we need to allow for local dependence.

Extremogram



The **extremogram** for a stationary time series $\{X_t\}$ estimates

$$\pi_h(u) = \mathrm{P}(X_{t+h} > u \mid X_t > u), \quad h = 1, 2, \ldots .$$

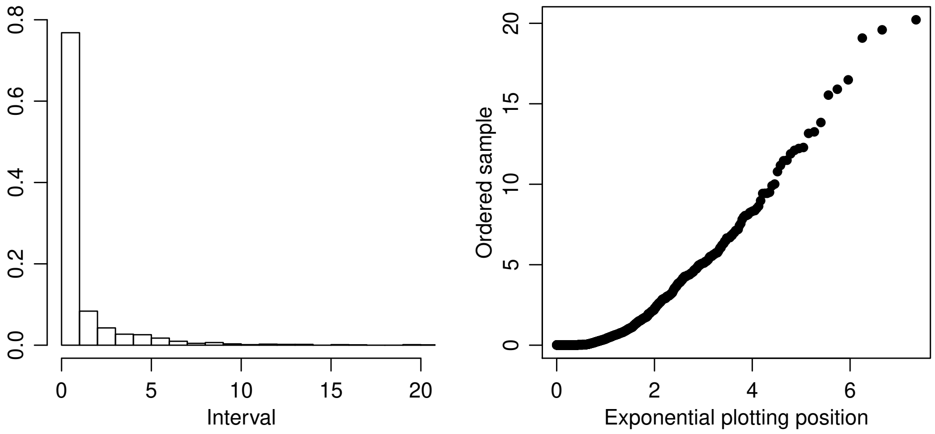
Independent data would have $\pi_h(u) \equiv \mathrm{P}(X_t > u)$ for all h (blue line in picture, upper 95% point is red line).

☐ This is the analogue of the ACF in conventional time series analysis,

☐ estimated using frequencies in place of probabilities —

☐ beware poor sampling properties of $\widehat{\pi}_h(u)$ (so don't worry about values for high lags).

Intervals between exceedances



The intervals between exceedances should be approximately exponentially distributed, but we see too many small intervals, due to clustering.

Definitions

Definition 21 *A time series* $\{X_j\}$ *is said to be* **(strictly) stationary** *if, for any finite subset* $\mathcal{A}$ *of* $\mathbb{Z}$, *the sets of variables* $X_{\mathcal{A}}$ *and* $X_{h+\mathcal{A}}$ *have the same distribution for all* $h \in \mathbb{Z}$. *In particular this means that the marginal distribution of* X_j *is invariant to location shifts, i.e.,*

$$\mathrm{P}(X_j \le x) = F(x), \quad j \in \mathbb{Z}, x \in \mathbb{R}.$$

Definition 22 *The* **matching series** *for a stationary time series* $\{X_j\}$ *with* $X_j \sim F$ *is the independent series* $\{X_j^*\}$ *for which* $X_j^* \overset{\text{iid}}{\sim} F$.

Definition 23 *If* F *is a continuous CDF then* $\{u_m\}$ *is a* **threshold sequence (for** F**)** *if there exists* $\Lambda \in (0, \infty)$ *such that* $\lim_{m\to\infty} m\{1 - F(u_m)\} = \Lambda$.

☐ If $M = \max(X_1, \ldots, X_m)$ where $X_j \overset{\text{iid}}{\sim} F$, and if the extremal types theorem (ETT) applies for sequences $\{a_m\} > 0$ and $\{b_m\}$, then taking $u_m = b_m + a_m x$ gives

$$\Lambda_m(x) = m\{1 - F(u_m)\} = m\{1 - F(b_m + a_m x)\} \to \left(1 + \xi \frac{x - \eta}{\tau}\right)_+^{-1/\xi} = \Lambda(x),$$

say, so $\{u_m\}$ is then a threshold sequence if $\Lambda(x) > 0$.

☐ If there is no Λ for which a threshold sequence exists, then the ETT does not apply.

$D(u_n)$

The usual condition used to impose near-independence of distant extremes is $D(u_n)$:

Definition 24 *Let* $\mathcal{A}, \mathcal{B}$ *be subsets of* $\{1, \ldots, n\}$ *such that* $\max \mathcal{A} < \min \mathcal{B} - l$ *for some positive integer* l*, and let* $M_{\mathcal{A}} \le u$ *denote the event* $\max_{i \in \mathcal{A}} X_i \le u$*, etc. Then* $D(u_n)$ *is satisfied if*

$$|\mathrm{P}(M_{\mathcal{A}} \le u_n, M_{\mathcal{B}} \le u_n) - \mathrm{P}(M_{\mathcal{A}} \le u_n)\mathrm{P}(M_{\mathcal{B}} \le u_n)| \le \alpha(n, l),$$

where $\alpha(n, l_n) \to 0$ *for some sequence* $l_n = o(n)$ *as* $n \to \infty$*.*

Under $D(u_n)$, maxima of subsets that are sufficiently separated are almost independent, where 'sufficiently separated' means that as $n \to \infty$, the gap l_n between $\mathcal{A}$ and $\mathcal{B}$ satisfies $l_n/n \to 0$.

Theorem 25 *Let* $X_1, \ldots, X_n$ *be a sequence from a stationary series with marginal distribution* F *that satisfies* $D(u_n)$ *for a threshold sequence* $u_n = b_n + a_n x$*. Then if*

$$\mathrm{P}\left\{\max(X_1, \ldots, X_n) \le u_n\right\} \to G(x), \quad n \to \infty,$$

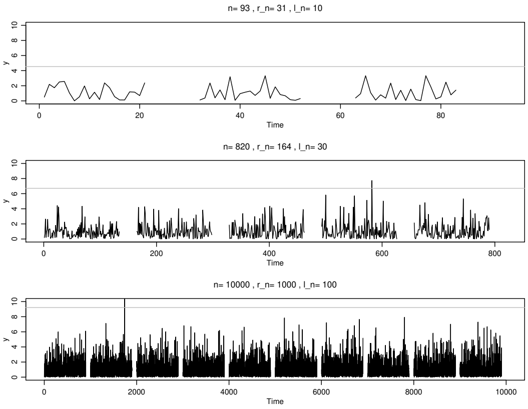
where G *is non-degenerate,* G *is a GEV distribution.*

Idea of Theorem 5

☐ We split $X_1, \ldots, X_n$ into k_n blocks of lengths r_n, where $k_n, r_n \to \infty$ as $n \to \infty$;

☐ we ensure that the block maxima are at least l_n observations apart, where $l_n \to \infty$, so if $D(u_n)$ applies these maxima become independent for large n;

☐ then we apply the ETT to the k_n (nearly independent) block maxima, and show that if these have a limiting distribution, it must be GEV.



Implications of Theorem 5

☐ The assumptions of Theorem 5 are weak, so it should hold in many applications.

☐ Hence we aim to understand the effect of local dependence by studying the properties of the maximum of a 'short' block $X_1, \ldots, X_{r_n}$ of neighbouring observations, which we compare with the maximum of an independent series $\{X_j^*\}$ with the same marginal distribution as $\{X_j\}$.

☐ Let $X_j^* \overset{\text{iid}}{\sim} F$, where $X_j \sim F$, i.e., F is the marginal distribution of $\{X_j\}$, and let

$$M_n^* = \max(X_1^*, \ldots, X_n^*), \quad M_n = \max(X_1, \ldots, X_n).$$

☐ We first consider an example.

Example 26 (Moving maximum process) *Let* $Z_j \overset{\text{iid}}{\sim} F(z) = \exp(-1/z)$ *for* $z > 0$, *and for* $a \geq 0$ *define*

$$X_0 = Z_0, \quad X_j = (a+1)^{-1} \max(aZ_{j-1}, Z_j), \quad j = 1, \ldots, n.$$

Show that

$$\mathrm{P}(M_n/n \leq x) \to \mathrm{P}(M_n^*/n \leq x)^\theta, \quad n \to \infty,$$

where $\theta = \max(1, a)/(a+1)$ *lies in the interval* $[1/2, 1]$.

Note to Example 6

☐ The marginal distribution of X_j is unit Fréchet:

$$\begin{aligned}\mathrm{P}(X_j \le x) &= \mathrm{P}\{aZ_{j-1} \le (a+1)x, Z_j \le (a+1)x\} \\ &= \exp\left\{-\frac{a}{(a+1)x}\right\}\exp\left\{-\frac{1}{(a+1)x}\right\} = \exp(-1/x), \quad x > 0,\end{aligned}$$

☐ If $X_1^*, X_2^*, \ldots$ are independent unit Fréchet variables and $M_n^* = \max(X_1^*, \ldots, X_n^*)$, then

$$\mathrm{P}(M_n^*/n \le x) = [\exp\{-1/(nx)\}]^n = \exp(-1/x),$$

whereas $M_n = \max(X_1, \ldots, X_n)$ satisfies

$$\begin{aligned}\mathrm{P}(M_n/n \le x) &= \mathrm{P}(X_1 \le nx, \ldots, X_n \le nx) \\ &= \mathrm{P}\{aZ_0 \le (a+1)nx, Z_1 \le (a+1)nx, aZ_1 \le (a+1)nx, \ldots, Z_n \le (a+1)nx\} \\ &= \mathrm{P}\{aZ_0 \le (a+1)nx\}\left[\prod_{j=1}^{n-1}\mathrm{P}\{Z_j \le (a+1)nx/\max(1,a)\}\right]\mathrm{P}\{Z_n \le (a+1)nx\}\end{aligned}$$

because the Z_j are independent. This implies that

$$\begin{aligned}\mathrm{P}(M_n/n \le x) &= \exp\left\{-\frac{a}{(a+1)nx}\right\}\left[\exp\left\{-\frac{\max(1,a)}{(a+1)nx}\right\}\right]^{n-1}\exp\left\{-\frac{1}{(a+1)nx}\right\} \\ &= \exp\left(-\theta_n/x\right),\end{aligned}$$

where

$$\theta_n = \frac{a+1+(n-1)\max(1,a)}{n(a+1)} \to \theta = \frac{\max(1,a)}{a+1} \in [1/2, 1], \quad n \to \infty.$$

☐ Hence

$$\mathrm{P}(M_n/n \le x) \to \exp(-\theta/x) = \mathrm{P}(M_n^*/n \le x)^\theta, \quad n \to \infty,$$

so although $X_j \overset{\mathrm{D}}{=} X_j^*$,

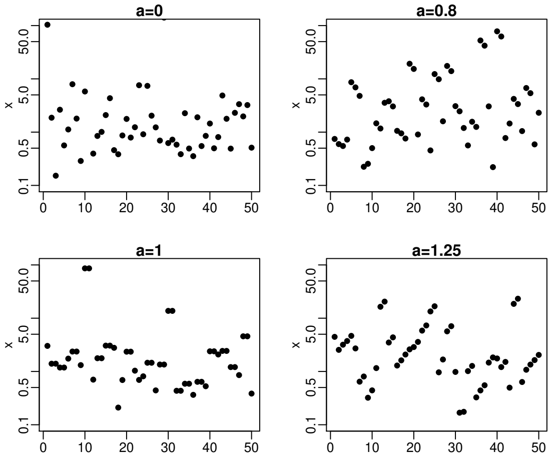
$$\mathrm{P}(M_n^* \le z) \doteq \mathrm{P}(M_n \le z)^{1/\theta} \le \mathrm{P}(M_n \le z),$$

i.e., M_n is stochastically smaller than M_n^*.

note 1 of slide 154

Moving maxima

Realisations of the moving maximum process of Example 6 with $a = 0, 0.8, 1, 1.25$. In each case the marginal distribution is unit Fréchet. The maxima show increasing clustering as $a \to 1$.



Effect of local dependence

Theorem 27 *Let* $\{X_j\}$ *be a stationary process such that* $X_j \sim F$ *and let* $X_j^* \overset{\text{iid}}{\sim} F$. *Set* $M_n = \max(X_1, \ldots, X_n)$, $M_n^* = \max(X_1^*, \ldots, X_n^*)$ *and let* $\{a_n\} > 0$ *and* $\{b_n\}$ *be sequences of real numbers. Then there exists a non-degenerate distribution function* G *such that*

$$\mathrm{P}\left\{(M_n - b_n)/a_n \le y\right\} \to G(y) = \exp\{-\Lambda(y)\}, \quad n \to \infty,$$

if and only if

$$\mathrm{P}\left\{(M_n^* - b_n)/a_n \le y\right\} \to G^*(y) = \exp\{-\Lambda^*(y)\}, \quad n \to \infty.$$

If so, $G(y) = \{G^*(y)\}^\theta$ *or equivalently* $\Lambda(y) = \theta\Lambda^*(y)$. *We call* $\theta \in (0, 1]$ *the* ***extremal index***.

☐ As G^* must be GEV(η^*, τ^*, ξ^*), say, G is also GEV, with parameters

$$\xi = \xi^*, \quad \tau = \tau^*\theta^\xi, \quad \eta = \eta^* + \tau^*(\theta^\xi - 1)/\xi \le \eta^* :$$

- the shape parameter is unchanged by the dependence but $\eta < \eta^*$, and
- M_n is stochastically smaller than M_n^*, i.e., dependence tends to reduce the sizes of the extremes for a series of given length, because

$$\lim_{n\to\infty} \mathrm{P}(M_n \le b_n + a_n y) = G(y) = \{G^*(y)\}^\theta \ge G^*(y) = \lim_{n\to\infty} \mathrm{P}(M_n^* \le b_n + a_n y).$$

Implications

- ☐ As $m \to \infty$ for independent data, the rescaled intervals T_m/m between exceedances are independent and Poisson process properties imply that

$$\mathrm{P}(T_m/m \le s) \to \mathrm{P}(S \le s) = 1 - e^{-\lambda s}, \quad s \ge 0.$$

- ☐ In the corresponding dependent case it can be shown that

$$\mathrm{P}(T_m/m \le s) \to \mathrm{P}(S \le s) = 1 - \theta e^{-\lambda \theta s}, \quad s \ge 0,$$

  i.e., in the limit,
  - exceedances arise in clusters of mean size $1/\theta \in [1, \infty)$,
  - θ is the probability that a randomly-chosen observation is the last of a cluster;
  - the expected interval $\mathrm{E}(S)$ between exceedances is unchanged,
  - but $\mathrm{E}(S \mid S > 0) \to 1/(\theta\lambda)$, so the mean interval between clusters increases by $1/\theta$,
  - the maximum of m dependent data has the same limiting distribution as the maximum of $m\theta \le m$ independent data.
- ☐ In fact a cluster maximum has the same limiting distribution as a randomly-chosen exceedance, so there is no asymptotic bias in fitting the GPD only to cluster maxima, **if we can identify the clusters . . .**

Note to Implications

☐ In the independent case, consider exceedances of a threshold sequence $u_m = b_m + a_m u$. As the X_j are independent and the process is stationary, the interval T_m between two successive exceedances satisfies

$$\begin{aligned} \mathrm{P}(T_m > k) &= \mathrm{P}(X_1 \le u_m, \ldots, X_k \le u_m \mid X_0 > u_m) \\ &= \mathrm{P}(X_1 \le u_m, \ldots, X_k \le u_m) \\ &= F(u_m)^k \\ &= \{1 - \Lambda_m(u)/m\}^k, \quad k \in \mathbb{N}, \end{aligned}$$

where $\Lambda_m(u) = m\{1 - F(u_m)\} \to \Lambda(u) \equiv \lambda$, say, as $m \to \infty$. For any $s > 0$, $\lfloor ms \rfloor / m \to s$ as $m \to \infty$, so

$$\begin{aligned} \mathrm{P}(T_m/m > s) &= \mathrm{P}(T_m > ms) \\ &= \mathrm{P}(T_m > \lfloor ms \rfloor) \\ &= \{1 - \Lambda_m(u)/m\}^{\lfloor ms \rfloor} \\ &\to \exp(-\lambda s), \quad s > 0. \end{aligned}$$

Hence $T_m/m \xrightarrow{D} S \sim \exp(\lambda)$.

☐ In the dependent case, we argue heuristically as follows. Let $\mathcal{C}$ denote the event that the exceedance at $j = 0$ is the last exceedance in a cluster. Then

$$\begin{aligned} \mathrm{P}(T_m > k) &= \mathrm{P}(X_1 \le u_m, \ldots, X_k \le u_m \mid X_0 > u_m) \\ &= \mathrm{P}(X_1 \le u_m, \ldots, X_k \le u_m \mid \mathcal{C}, X_0 > u_m)\mathrm{P}(\mathcal{C} \mid X_0 > u_m) \\ &\quad + \mathrm{P}(X_1 \le u_m, \ldots, X_k \le u_m \mid \mathcal{C}^c, X_0 > u_m)\mathrm{P}(\mathcal{C}^c \mid X_0 > u_m). \end{aligned}$$

As the data are dependent, $\mathrm{P}(X_1 \le u_m, \ldots, X_k \le u_m \mid \mathcal{C}, X_0 > u_m)$ is approximately the probability that $\max(X_1, \ldots, X_k) \le u_m$, conditional on $\mathcal{C} \cap \{X_0 > u_m\}$, and for large k we therefore have

$$\mathrm{P}(X_1 \le u_m, \ldots, X_k \le u_m \mid \mathcal{C}, X_0 > u_m) \doteq F(u_m)^{k\theta} = \{1 - \Lambda_m(u)/m\}^{k\theta},$$

whereas for large k,

$$\mathrm{P}(X_1 \le u_m, \ldots, X_k \le u_m \mid \mathcal{C}^c, X_0 > u_m) \approx 0,$$

because an observation that is not the last of cluster is highly likely to be followed by another exceedance. Thus if we let $a = \lim_{m\to\infty} \mathrm{P}(\mathcal{C} \mid X_0 > u_m)$, we have

$$\mathrm{P}(T_m/m > s) = \mathrm{P}(T_m > \lfloor ms \rfloor) \doteq \{1 - \Lambda_m(u)/m\}^{\lfloor ms \rfloor \theta}\,\mathrm{P}(\mathcal{C} \mid X_0 > u_m) \to a\exp(-\theta\lambda s), \quad s > 0$$

This distribution puts a mass of $1 - a$ at $s = 0$ and therefore has mean

$$(1 - a)0 + a/(\theta\lambda) = a/(\theta\lambda).$$

☐ But as the expected time between exceedances is $1/\lambda$, it must be the case that $a = \theta$, which gives the stated distribution.

☐ A ‘corrected’ argument is much messier, but is essentially the same as that above.

Statistical consequences of clustering

☐ Clustering affects the return levels and their interpretation:
  - if $\theta = 1$, then annual maxima are independent but the 'T-year-event' has probability

$$(1 - 1/T)^T \doteq e^{-1} \doteq 0.368$$

    of not appearing in any period of T years;
  - if $\theta < 1$, then the T-year event has probability

$$(1 - 1/T)^{T\theta} \doteq e^{-\theta}$$

    of not appearing in a period of T years, giving (for example) $e^{-0.1} \doteq 0.905$. The same number of events will occur, on average, but they will occur together when $\theta < 1$.

☐ Various estimators of θ exist. A simple procedure is
  - identify clusters, e.g., by declaring that clusters are separated by runs of more than r non-exceedances of u,
  - let $\widehat{\theta}_u = n_c/n_u$, i.e., the number of clusters divided by the number of exceedances of u.

POT fit to the Eskdalemuir data

```
> fpot(esk.rain, threshold=5)

Call: fpot(x = esk.rain, threshold = 5)
Deviance: 1058.954

Threshold: 5
Number Above: 356
Proportion Above: 0.0024

Estimates
  scale    shape
1.52239  0.06702

Standard Errors
  scale    shape
0.11488  0.05383

Optimization Information
  Convergence: successful
  Function Evaluations: 18
  Gradient Evaluations: 6
```

The fit above does not allow for any clustering.

POT fit to the Eskdalemuir data, allowing for clustering

```
> fpot(esk.rain, threshold=5, r=1, cmax=TRUE)  # cmax=TRUE means fit the GP only to the clust

Call: fpot(x = esk.rain, threshold = 5, cmax = TRUE, r = 1)
Deviance: 835.234

Threshold: 5
Number Above: 356
Proportion Above: 0.0024

Clustering Interval: 1
Number of Clusters: 272
Extremal Index: 0.764

Estimates
  scale    shape
1.63808  0.04183

Standard Errors
  scale    shape
0.14343  0.06322

Optimization Information
  Convergence: successful
  Function Evaluations: 18
  Gradient Evaluations: 5
```

The fit above uses a simple (simplistic!) approach to identifying clusters, which are supposed to end when there are r consecutive background observations below the threshold. Note that $\widehat{\theta} = n_c/n_u = 272/356$. The GP parameters are estimated from n_c cluster maxima, so their standard errors are appreciably increased relative to the previous fit, which ignores any clustering.

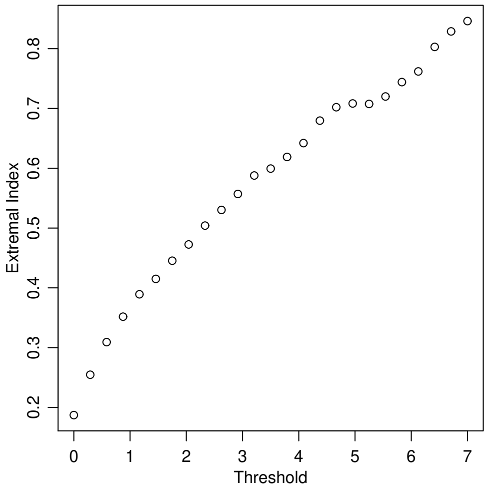
POT fit to the Eskdalemuir data, allowing for clustering

```
> fpot(esk.rain, threshold=5, r=4, cmax=TRUE)

Call: fpot(x = esk.rain, threshold = 5, cmax = TRUE, r = 4)
Deviance: 777.1452

Threshold: 5
Number Above: 356
Proportion Above: 0.0024

Clustering Interval: 4
Number of Clusters: 243
Extremal Index: 0.6826

Estimates
  scale    shape
1.79613  0.01343

Standard Errors
  scale    shape
0.16476  0.06557

Optimization Information
  Convergence: successful
  Function Evaluations: 18
  Gradient Evaluations: 4
```

The fit above uses $r = 4$. Note that $\widehat{\theta} = 0.68$ is smaller than when $r = 1$, and that the standard errors for the GP parameters are still larger than before.

Dependence of $\widehat{\theta}$ on u

Unfortunately $\widehat{\theta}$ (here estimated with $r = 1$) depends on u. The lack of a limit might throw doubt on the theory . . .



Clustering and return levels

☐ The consequences for estimation of return levels are that m dependent background observations correspond to $m\theta$ matching observations, so in the previous formulae on slide 83 we replace m by $m\theta$ (for maxima) and probability $p = 1/N_p$ for dependent observations by $p/\theta = 1/(N_p\theta)$ matching observations (for exceedances), solving $1 - F_X(x_p) = p/\theta$ in the threshold case and

$$1 - F_X(x_p) = 1 - G^{1/(m\theta)}(x_p)$$

when fitting maxima.

☐ To estimate the return level

$$x_p \doteq u + \frac{\sigma}{\xi}\left[(p_u\theta/p)^{\xi} - 1\right],$$

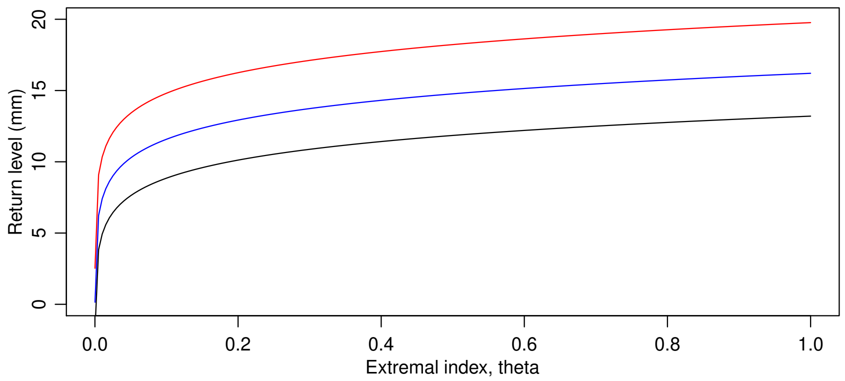
we estimate σ and ξ by fitting the GPD to threshold exceedances, and use

$$\widehat{p}_u = \frac{n_u}{n}, \quad \widehat{\theta} = \frac{n_c}{n_u},$$

where n_c is number of clusters and n_u is number of exceedances; thus $\widehat{p}_u\widehat{\theta} = n_c/n$.

Return level estimation

☐ The figure below shows how θ affects estimates of the 5- (black), 20- (blue) and 100- (red) year return levels for the Eskdalemuir data with threshold $u = 5$ mm.



☐ See note on Moodle

Summary

- ☐ Under a weak (and often plausible) condition $D(u_n)$ on the dependence of distant extremes, the GEV is the limiting distribution for the maximum of a stationary dependent process.
- ☐ We compare a stationary dependent series $\{X_j\}$ such that $X_j \sim F$ with a **matching series** $\{X_j^*\} \overset{\text{iid}}{\sim} F$.
- ☐ The effect of local dependence is that extremes arise in clusters whose properties depend on the **extremal index** θ, and
  - the mean cluster size is $1/\theta \geq 1$,
  - the probability that a randomly chosen large event is the last in a cluster is θ,
  - the mean interval between clusters is $1/\theta$ times larger than for the matching series,
  - the GPD marginal distribution of a threshold exceedance is the same as that of a cluster maximum,
  - the maximum of m observations in the dependent series is approximately that of $m\theta$ observations in the matching series,
  - estimates of return levels must be modified to allow for θ.
- ☐ Various empirical estimates of θ can be computed.

5 Multivariate Extremes

5.1 Introduction

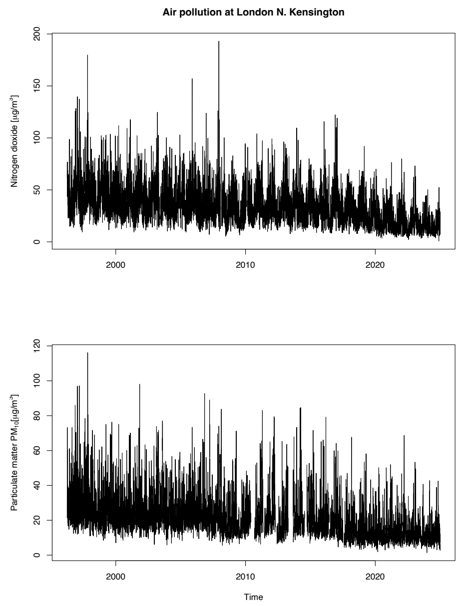
Multivariate extremes

☐ Many extremal problems are essentially multivariate in nature:
  - overwhelming of sea defences by a combination of high tides and high winds;
  - lots of traffic at different servers on a communication network;
  - flooding at many locations of a river system;
  - several successive very hot days (heat waves);
  - (near-)simultaneous downturns in several stock markets.

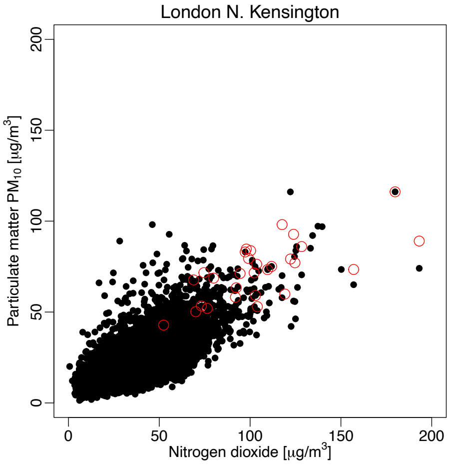
☐ Also, the (often) large variability of extreme value estimates may be reduced by incorporating information via multivariate models.

☐ In one dimension it's obvious what is 'extreme', but in addition to previous questions about suitable asymptotic models, inference, and complications, we must now consider:
  - what is 'extreme' in two or more dimensions?
  - how can we summarise extremal dependence of different variables?

Air pollution in London



Air pollution in London



Air pollution in London: Empirical tail dependence

☐ Each pollutant has its own marginal distribution: F_1 for NO_2 and F_2 for PM_{10}, say.

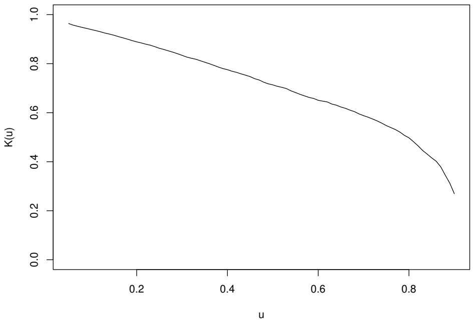
☐ To see empirically how their large values behave we first make the margins comparable.

☐ If X is a continuous random variable with CDF F_X, the random variable $U = F_X(X)$, the **probability integral transform** of X, is uniformly distributed on the unit interval.

☐ We assess the **joint behaviour** of NO_2 and PM_{10} by estimating

$$K(u) = \mathrm{P}(F_1(NO_2) > u \mid F_2(PM_{10}) > u), \quad 0 < u < 1.$$

☐ Large values for one variable lead to large values of the other. What happens when $u \to 1$?

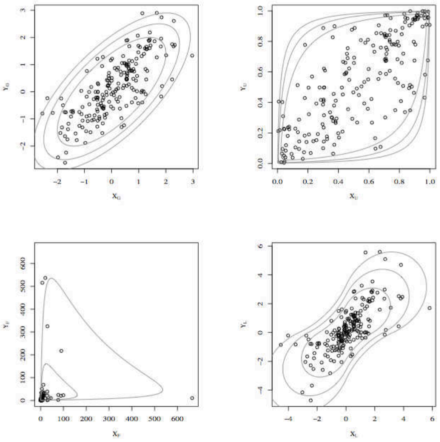


Structure variables

- ☐ Multivariate analysis is difficult, so perhaps we could simplify?
- ☐ Could consider a scalar **structure variable** $S = s(X_1, \ldots, X_D) \in \mathbb{R}$, e.g.,
  - insurance loss
    $$S = \sum_d a_d(X_d),$$
    where increasing (possibly non-linear) functions $a_d(\cdot)$ express damages to properties d due to risks X_d.
- ☐ Then, we have scalar losses $S_1, \ldots, S_n$ to which previous ideas apply, using block maxima or threshold exceedances.
- ☐ **Advantages**: simple analysis, ignores dependence between $X_1, \ldots, X_D$.
- ☐ **Disadvantages**:
  - analysis changes with S, so if new structure variable is introduced, new analysis is needed—which may disagree with original;
  - missing values of X_d not allowed;
  - don't learn which combinations of $X_1, \ldots, X_D$ yield extreme events.
- ☐ So we should study the joint distributions. First we have to look at dependence in general ...

Choice of margins

- ☐ Different aspects of dependence between multivariate data are visually highlighted by taking different marginal distributions.
- ☐ Bivariate normal data on (clockwise from top left) Gaussian, uniform, Fréchet and Laplace scales, with density contours (grey).



Standardising margins

- ☐ When studying dependence it helps to remove the effect of marginal transformations.
- ☐ Here we consider only continuous random variables, and apply the **probability integral transformation** to obtain variables with uniform margins.
- ☐ Suppose that $X \sim F$ is continuous, and takes values everywhere in an interval of $\mathbb{R}$, so the inverse

$$F^{-1}(p) = \inf\{x : F(x) \geq p\}, \quad 0 < p < 1,$$

satisfies $F\{F^{-1}(p)\} = p$ and $F^{-1}\{F(x)\} = x$.

- ☐ Then

$$\mathrm{P}\{F(X) \leq u\} = \mathrm{P}\{X \leq F^{-1}(u)\} = F\{F^{-1}(u)\} = u, \quad 0 < u < 1:$$

i.e., $F(X) \sim U(0,1)$.

- ☐ Equivalently, if $U \sim U(0,1)$, then $X = F^{-1}(U) \sim F$.
- ☐ If $X = (X_1, \ldots, X_D) \sim F$ has strictly monotone increasing marginal distributions $F_1, \ldots, F_D$, we therefore have $U_d = F_d(X_d) \sim U(0,1)$ for each $d \in \{1, \ldots, D\}$, corresponding to the top right panel on slide 59.

Copulas

- ☐ If $X \sim F$ is continuous with margins F_d then $U_d = F_d(X_d) \sim U(0,1)$ for every $d \in \{1, \ldots, D\}$, and there exists $C : [0,1]^D \to [0,1]$ such that

$$\mathrm{P}(U_1 \leq u_1, \ldots, U_D \leq u_D) = F\{F_1^{-1}(u_1), \ldots, F_D^{-1}(u_D)\} = C(u_1, \ldots, u_D),$$

for $0 \leq u_1, \ldots, u_D \leq 1$, where

$$C(0, u_2, \ldots, u_D) = 0, \quad C(u, 1, \ldots, 1) = u$$

for any permutation of the indices.

- ☐ Similarly

$$F(x) = C\{F_1(x_1), \ldots, F_D(x_D)\}.$$

- ☐ The **copula** C
  - determines the dependence structure of $U = (U_1, \ldots, U_D) = (F_1(X_1), \ldots, F_d(X_d))$,
  - is a cumulative distribution function with uniform margins,
  - is unique (Sklar's theorem) if F is continuous (as we assume), and
  - its derivatives yield joint density functions and thus constrain C.

Examples

Example 28 (Independence copula) *If $(U_1, \dots, U_D)$ are independent, then*

$$C(u_1, \dots, u_D) = \prod_{d=1}^{D} u_d, \quad , \quad 0 < u_1, \dots, u_D < 1.$$

Example 29 (Co-monotone copula) *If $(U_1, \dots, U_D)$ are totally dependent, then $U_1 = \cdots = U_D$ with probability one, and*

$$C(u_1, \dots, u_D) = \min(u_1, \dots, u_D), \quad , \quad 0 < u_1, \dots, u_D < 1.$$

Example 30 (Gaussian copula) *In terms of the D-dimensional Gaussian CDF Φ_D and the univariate Gaussian CDF Φ, the Gaussian copula is*

$$C_\Omega(u_1, \dots, u_D) = \Phi_D\left\{\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_D); \Omega\right\}, \quad 0 < u_1, \dots, u_D < 1,$$

which corresponds to the top-right panel on slide 59.

Non-extremal dependence

☐ Measures such as the usual (Pearson) correlation coefficient depend on the margins, and we seek to avoid this.

☐ Let (U_1, U_2) and (V_1, V_2) be independent pairs of variables with copula C.

☐ A standard measure of dependence is **Kendall's tau**,

$$\tau = \mathrm{corr}\{I(U_1 > V_1), I(U_2 > V_2)\} = 4\mathrm{E}\{C(U_1, U_2)\} - 1,$$

which measures the extent to which the event $(U_1 - V_1)(U_2 - V_2) > 0$ is more probable than the event $(U_1 - V_1)(U_2 - V_2) < 0$.

☐ τ measures the dependence of the entire distribution, but we wish to focus on the tails.

Example 31 *Compute Kendall's tau for the bivariate independence and co-monotone copulas.*

Note to Example 11

☐ The bivariate independence copula is $C(u_1, u_2) = u_1 u_2$, and

$$\mathrm{E}\{C(U_1, U_2)\} = \mathrm{E}(U_1 U_2) = \int_0^1 \int_0^1 u_1 u_2 \, \mathrm{d}u_1 \, \mathrm{d}u_2 = (1/2)^2 = 1/4,$$

so $\tau = 0$.

☐ The bivariate co-monotone copula is $C(u_1, u_2) = \min(u_1, u_2)$ and under this model $U_1 = U_2$ with probability one, so we integrate over just one of the variables:

$$\mathrm{E}\{C(U_1, U_2)\} = \mathrm{E}\{\min(U_1, U_2)\} = \int_0^1 u_1 \, \mathrm{d}u_1 = 1/2,$$

so $\tau = 1$.

χ

☐ The probability that equally rare values of two variables occur simultaneously is a key extremal property. When they have copula C we define the **extremal correlation** to be

$$\chi = \lim_{u\to 1} \mathrm{P}(U_2 > u \mid U_1 > u) = \lim_{u\to 1} \frac{1-2u+C(u,u)}{1-u}, \tag{22}$$

if it exists, or equivalently on general margins,

$$\chi = \lim_{u\to 1} \mathrm{P}\left\{X_2 > F_2^{-1}(u) \mid X_1 > F_1^{-1}(u)\right\}.$$

☐ X_1 and X_2 are **asymptotically dependent (AD)** if $\chi > 0$ and **asymptotically independent (AI)** if $\chi = 0$.

☐ For statistical purposes, as $u \to 1$ we replace $1-u$ and $1-C(u,u)$ in (1) by the approximations $-\log u$ and $-\log C(u,u)$ and obtain

$$\chi(u) = 2 - \frac{\log C(u,u)}{\log u}, \quad 0 < u < 1,$$

with the focus on $\chi(u)$ for $u \approx 1$.

Example 32 (Logistic copula) *The logistic copula with parameter $\alpha \in (0,1]$ is of the form*

$$C(u_1,\ldots,u_D) = \exp\left[-\left\{\sum_{d=1}^{D}(-\log u_d)^{1/\alpha}\right\}^{\alpha}\right], \quad 0 < u_1,\ldots,u_D < 1.$$

If $\alpha = 1$, this reduces to the **independence copula** *$u_1 \cdots u_D$, whereas as $\alpha \to 0$ it reduces to the completely dependent* **comonotone copula**. *Compute $\chi(u)$ for the logistic copula.*

Note to Example 12

☐ The bivariate logistic copula

$$C(u_1, u_2) = \exp\left[-\left\{(-\log u_1)^{1/\alpha} + (-\log u_2)^{1/\alpha}\right\}^{\alpha}\right], \quad \alpha \in (0, 1],$$

evaluated for $u_1 = u_2 = u$ gives

$$C(u, u) = \exp\left[-\left\{2\,(-\log u)^{1/\alpha}\right\}^{\alpha}\right] = \exp\left\{-2^{\alpha}\,(-\log u)\right\} = u^{2^{\alpha}},$$

which results in the extremal correlation

$$\chi(u) = 2 - \tfrac{\log(u^{2^{\alpha}})}{\log u} = 2 - \tfrac{2^{\alpha}\log u}{\log u} = 2 - 2^{\alpha}.$$

This is constant in u due to the max-stability of the logistic copula (more later).

☐ We have the following limiting cases:

$$\lim_{\alpha\to 1} \chi(u) = 2 - 2^1 = 0,$$

that corresponds to independence and

$$\lim_{\alpha\to 0} \chi(u) = 2 - 2^0 = 2 - 1 = 1,$$

that corresponds to perfect dependence (the co-monotone copula).

$\overline{\chi}$

☐ χ can distinguish the strength of dependence for AD distributions, but not the different rates at which $\chi(u) \to 0$ for AI distributions as $u \to 1$.

☐ We define

$$\overline{\chi}(u) = 2\frac{\log \mathrm{P}(U_1 > u)}{\log \mathrm{P}(U_2 > u, U_1 > u)} - 1 = 2\frac{\log(1-u)}{\log\{1 - 2u + C(u, u)\}} - 1, \quad 0 < u < 1, \qquad (23)$$

and use $\overline{\chi} = \lim_{u\to 1} \overline{\chi}(u)$ to measure the degree of AI.

☐ The scaling is chosen so that if

- X and Y are independent, $\overline{C}(u, u) = \mathrm{P}\{F_X(X) > u, F_Y(Y) > u\} = (1-u)^2$, and then $\overline{\chi}(u) \equiv 0$;
- if X and Y are perfectly dependent, $\overline{C}(u, u) = 1 - u$, and then $\overline{\chi}(u) \equiv 1$;
- if X and Y are asymptotically dependent, $\overline{\chi}(u) \to 1$ as $u \to 1$;
- $-1 < \overline{\chi}(u) \leq 1$, and $\overline{\chi}(u)$ increases with increasing dependence.

Example 33 (Bivariate normal) *$\overline{\chi} = \rho$ for the bivariate normal distribution.*

Note to Example 13

☐ As both probabilities tend to zero as $u \to 1$, we first use l'Hôpital's rule to obtain

$$\begin{aligned}\lim_{u\to 1} \mathrm{P}(U_2 > u \mid U_1 > u) &= \lim_{u\to 1} \frac{\mathrm{P}(U_2 > u, U_1 > u)}{\mathrm{P}(U_1 > 0)} \\ &= \lim_{u\to 1} \frac{\mathrm{P}(U_2 > u, U_1 = u) + \mathrm{P}(U_2 = u, U_1 > u)}{\mathrm{P}(U_1 = u)} \\ &= \lim_{u\to 1} 2\mathrm{P}(U_2 > u \mid U_1 = u)\end{aligned}$$

if the distribution is symmetric (as here).

☐ Note that Mill's ratio

$$\mathrm{P}(X_1 > x) = \overline{\Phi}(x) \sim \phi(x)/x, \quad x \to \infty,$$

implies that $\log \mathrm{P}(X_1 > x) \sim -x^2/2 - \log x$ as $x \to \infty$.

☐ In

$$\frac{\log \mathrm{P}(U_1 > u)}{\log \mathrm{P}(U_1 > u) + \log \mathrm{P}(U_2 > u \mid U_1 > u)}, \tag{24}$$

U_1 and U_2 can be replaced by standard normal variables X_1 and X_2, and as the distribution is symmetric, for large x we have

$$\begin{aligned}\mathrm{P}(X_2 > x \mid X_1 > x) \sim 2\mathrm{P}(X_2 > x \mid X_1 = x) &= 2\overline{\Phi}\left\{\frac{x(1-\rho)}{(1-\rho^2)^{1/2}}\right\} \\ &\sim \phi\left\{\frac{x(1-\rho)}{(1-\rho^2)^{1/2}}\right\} \div \frac{x(1-\rho)}{(1-\rho^2)^{1/2}}\end{aligned}$$

Hence

$$\begin{aligned}\frac{\log \mathrm{P}(X_1 > x)}{\log \mathrm{P}(X_1 > x) + \log \mathrm{P}(X_2 > x \mid X_1 > x)} &\sim \frac{-x^2/2 - \log x}{-x^2/2 - \log x - \frac{x^2(1-\rho)^2}{2(1-\rho^2)} - \log\left\{\frac{x(1-\rho)}{(1-\rho^2)^{1/2}}\right\}} \\ &\to \frac{1}{1 + (1-\rho)/(1+\rho)}, \quad x \to \infty, \\ &= (1+\rho)/2.\end{aligned}$$

Hence $\overline{\chi} = \rho$, thus stressing the interpretation of positive, negative and zero values of $\overline{\chi}$.

note 1 of slide 180

Comments

☐ Copulas allow the dependence between several variables to be studied without reference to their marginal distributions. They are very widely used in finance, insurance and other areas.

☐ A key distinction for rare events is between asymptotic dependence (AD) and asymptotic independence (AI), which correspond to $\chi > 0$ and $\chi = 0$ for

$$\chi = \lim_{u \to 1} \chi(u) = \lim_{u \to 1} \mathrm{P}\{F_2(X_2) > u \mid F_1(X_1) > u\}, \quad 0 < u < 1,$$

in practice replaced by the asymptotically (as $u \to 1$) equivalent

$$\chi(u) = 2 - \frac{\log C(u,u)}{\log u}, \quad 0 < u < 1,$$

where C is the copula for (X_1, X_2).

☐ A similar quantity $\overline{\chi}(u)$ is used to distinguish different strengths of AI.

☐ Plots of $\chi(u)$ and $\overline{\chi}(u)$ are essential graphical tools for looking at joint extremes, and are produced by the `R` command `evd::chiplot`.

5.3 Multivariate Models for Extremes

Extremes for $D = 2$

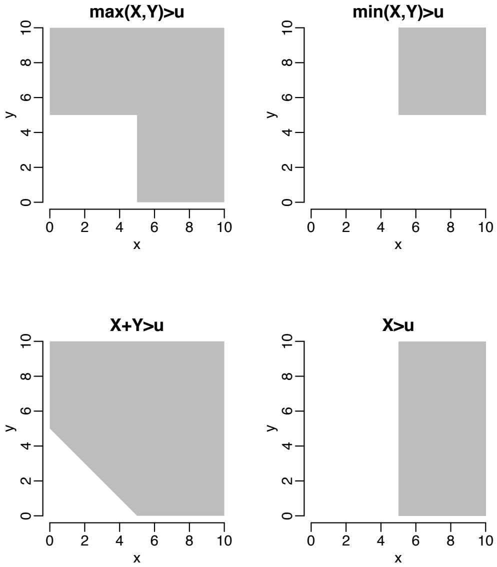
☐ Given variables (X, Y) with the same marginal distributions, and a high threshold u, we might consider any of the following scenarios:

- at least one of X and Y exceeds u, i.e., $\max(X, Y) > u$;
- both X and Y exceed u, i.e., $\min(X, Y) > u$;
- a function $s(X, Y)$ exceeds u, e.g., $X + Y > u$, though $s(\cdot)$ could also measure distance from some multivariate centre for the data; or
- given that $X > u$, we consider the distribution of Y, where Y is called a **concomitant** of X; the extremal set is $X > u$.

☐ There are other possibilities, but these already make life complicated enough.

☐ The grey regions on the next slide are considered to be extreme under these four scenarios.

Extremes for $D = 2$



Componentwise maxima

- ☐ If $(X_1, Y_1), (X_2, Y_2), \ldots \overset{\text{iid}}{\sim} F(x, y)$, define the **componentwise maxima**,

$$M_{X,n} = \max_{j=1,\ldots,n} \{X_j\}, \qquad M_{Y,n} = \max_{j=1,\ldots,n} \{Y_j\};$$

note that $M_n = (M_{X,n}, M_{Y,n})$ may not correspond to an actual observation (e.g., NO_2, PM_{10} in pollution example).

- ☐ Limiting distributions must exist for maxima of X and Y individually, because otherwise any limiting joint distribution will be degenerate, so we ask

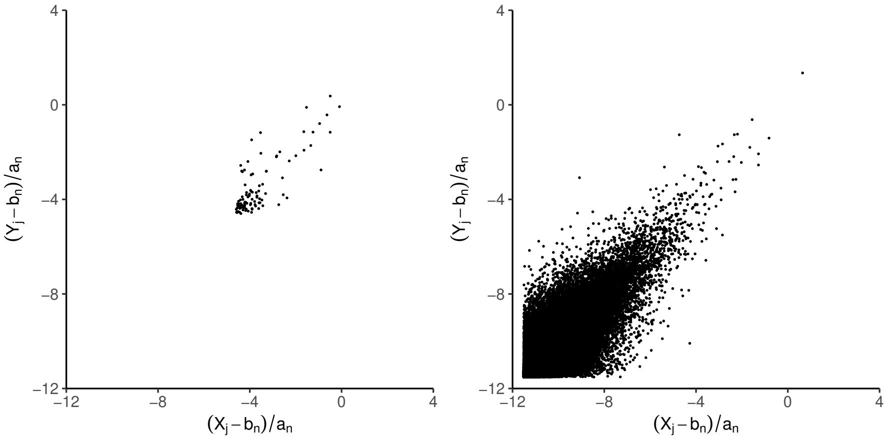
If non-degenerate limiting distributions exist for maxima of rescaled pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ as $n \to \infty$, what forms can they have?

- ☐ Considered separately, $\{X_j\}$ and $\{Y_j\}$ are sequences of independent, univariate random variables, to which our previous theory applies: if a limiting distribution exists for each margin, then we can consider the sequence of point patterns

$$\mathcal{P}_n = \{((X_j - b_{X,n})/a_{X,n}, (Y_j - b_{Y,n})/a_{Y,n}) : j = 1, \ldots, n\}, \quad n = 1, 2, \ldots.$$

- ☐ As $n \to \infty$ we will have convergence to a Poisson process, $\mathcal{P}_n \xrightarrow{D} \mathcal{P}$, with state space $\mathcal{E} = \mathbb{R}^2$ and measure μ, say, which we can use for inference on extreme values.

Point patterns for $D = 2$



Rescaled bivariate exponential datasets of sizes $n = 100$ (left) and $n = 10^4$ (right).
In this case $\Lambda(x) = e^{-x}$ on each margin, so the transformation $1/\Lambda(x)$ to unit Fréchet margins for maxima would exponentiate both axes and give many observations close to the origin.

Marginal transformation

☐ On the X margin as $n \to \infty$ we can choose the sequences $\{b_n\}$ and $\{a_n\}$ so that

$$\max\{(X_1 - b_n)/a_n, \ldots, (X_n - b_n)/a_n\} \xrightarrow{D} S \sim \exp\{-\Lambda_X(s)\},$$

where $\Lambda_X(s) = (1 + \xi_X s)_+^{-1/\xi_X}$ is monotone decreasing, so $Z = 1/\Lambda_X(S)$ has the unit Fréchet distribution

$$\begin{aligned} \mathrm{P}(Z \le z) &= \mathrm{P}\{1/\Lambda_X(S) \le z\} \\ &= \mathrm{P}\{\Lambda_X(S) \ge 1/z\} \\ &= \mathrm{P}\{S \le \Lambda_X^{-1}(1/z)\} \\ &= \exp[-\Lambda_X\{\Lambda_X^{-1}(1/z)\}] \\ &= \exp(-1/z), \quad z > 0. \end{aligned}$$

☐ The same argument applies on the Y margin, so we can apply the transformation

$$g(x, y) = (1/\Lambda_X(x), 1/\Lambda_Y(y)), \quad (x, y) \in \mathcal{E} = \mathbb{R}^2$$

to $\mathcal{P}_n$, giving sequences of point patterns

$$\mathcal{P}_n^* = g(\mathcal{P}_n) \subset \mathcal{E}^* \subset \mathbb{R}_+^2, \quad n = 1, \ldots,$$

such that the limiting maxima on each margin have unit Fréchet distributions.

Limiting Poisson process

☐ The function g is invertible, so Poisson convergence for $\mathcal{P}_n$ and for $\mathcal{P}_n^* = g(\mathcal{P}_n)$ is equivalent.

☐ As $n \to \infty$, the point pattern $\mathcal{P}_n^*$ converges to a Poisson process $\mathcal{P}^*$ on $\mathcal{E}^*$ whose mean measure μ^* has margins

$$\mu^*\{\mathbb{R}_+ \times (z, \infty)\} = \mu^*\{(z, \infty) \times \mathbb{R}_+\} = 1/z, \quad z > 0,$$

so we cannot compute the measure of any set containing the origin. To avoid this we define $\mathcal{P}^*$ on the 'punctured set' $\mathcal{E}^* = [0, \infty)^2 - \{(0, 0)\}$.

☐ For any $z \equiv (z_1, z_2) \in \mathcal{E}^*$ it is convenient to define

$$\mathcal{A}_z^* = \{(x, y) \in \mathcal{E}^* : x > z_1 \text{ or } y > z_2\},$$

so that the joint maxima satisfy

$$G^*(z) = \mathrm{P}(Z_1 \le z_1, Z_2 \le z_2) = \mathrm{P}\{N^*(\mathcal{A}_z^*) = 0\} = \exp\{-\mu^*(\mathcal{A}_z^*)\},$$

where $N^*(\mathcal{A}^*)$ is the number of points of $\mathcal{P}^*$ in $\mathcal{A}^* \subset \mathcal{E}^*$.

☐ In terms of the original process $\mathcal{P}$ and $\mathcal{A}^* = g(\mathcal{A})$ we have

$$\mu(\mathcal{A}) = \mu^*(\mathcal{A}^*) = \mu^*\{g(\mathcal{A})\},$$

which will be useful later.

Exponent function

☐ For ease of notation define the **exponent function**

$$V(z) = V(z_1, z_2) = \mu^*(\mathcal{A}_z^*), \quad z \in \mathcal{E}^*.$$

☐ In the scalar case we saw that the GEV is max-stable, i.e., for any $t > 0$ there exist $a_t > 0$ and b_t such that

$$G^t(b_t + a_t x) = G(x), \quad x \in \mathbb{R},$$

and when G is unit Fréchet we have $b_t = 0$ and $a_t = t$.

☐ In the multivariate case the same argument applies, giving

$$\{G^*(tz)\}^t = G^*(z) \quad \Longrightarrow \quad tV(tz) = V(z), \quad z \in \mathcal{E}^*, t > 0,$$

i.e., the function V is homogeneous of order -1.

☐ The marginal unit Fréchet distributions yield

$$V(z', \infty) = V(\infty, z') = 1/z', \quad z' > 0.$$

☐ The homogeneity of V suggests a change of variables to **angular coordinates**

$$R = Z_1 + Z_2, \quad W = (Z_1, Z_2)/R \quad \Longleftrightarrow \quad Z = (Z_1, Z_2) = RW,$$

which allow us to state the joint distribution of the maxima.

Limit distribution of componentwise maxima

Theorem 34 *If $X_1, X_2, \ldots,$ are independent copies of a D-dimensional random variable whose componentwise maxima can be linearly renormalised to converge as $n \to \infty$ to a random variable $Z = (Z_1, \ldots, Z_D)$ that has a non-degenerate distribution with unit Fréchet margins, then*

$$\mathrm{P}(Z_1 \leq z_1, \ldots, Z_D \leq z_D) = \exp\left[-D\mathrm{E}\left\{\max_{d=1}^{D}(W_d/z_d)\right\}\right], \quad z_1, \ldots, z_D > 0, \tag{25}$$

where the **angular variable** $W = (W_1, \ldots, W_D)$ *lies in the* $(D-1)$*-dimensional simplex, i.e.,*

$$W \in \mathbb{S}_{D-1} = \left\{(w_1, \ldots, w_D) : w_d \geq 0, \sum_{d=1}^{D} w_d = 1\right\}$$

and satisfies the marginal mean constraints

$$\mathrm{E}(W_d) = 1/D, \quad d = 1, \ldots, D.$$

The **angular probability distribution** ν *of* W *is otherwise arbitrary — it may or may not have an* **angular density function** $\dot{\nu}$*.*

Note: Proof of Theorem 14

☐ The proof is not more complicated in D dimensions.

☐ The preceding argument implies that after the marginal transformations the D-dimensional point processes $\mathcal{P}_n^* = g(\mathcal{P}_n)$ converge to a Poisson process on $\mathcal{E}^* = [0,\infty)^D - \{(0,\dots,0)\}$ with measure μ^* which is homogeneous of order -1, so we just have to establish the form of the distribution for maxima.

☐ Let $(q_1,\dots,q_D)$ be a point in $\mathcal{E}^*$. To see the effect of the change of variables, consider the transformation $T : \mathcal{E}^* \to (0,\infty) \times \mathbb{S}_{D-1}$ to angular variables defined by

$$T(q_1,\dots,q_D) = (r,w), \quad r = q_1 + \cdots + q_D, \quad w_d = q_d/r, \quad d = 1,\dots,D,$$

and with inverse $T^{-1}(r,w) = rw$. As T is invertible, $T(\mathcal{P}^*)$ is also a Poisson process.

☐ To compute the measure $\mu^* \circ T^{-1}$ that T induces for the angular variables (R,W), note that as $rw \in \mathcal{A}_z^*$ if and only if at least one of the rw_d exceeds z_d, or equivalently $\max_d rw_d/z_d > 1$,

$$\begin{aligned} T(\mathcal{A}_z^*) &= \{(r,w) : \max(rw_1/z_1,\dots,rw_D/z_D) > 1\} \\ &= \{(r,w) : ra_z(w) > 1\}, \end{aligned} \tag{26}$$

where $a_z(w) = \max(w/z_1,\dots,w_D/z_D)$. Moreover,

$$\mu^*(\mathcal{A}_z^*) = V(z) = V(rw) = \frac{D}{r} \times D^{-1}V(w) = \frac{D}{r} \times \nu(w), \quad r > 0, w \in \mathbb{S}_{D-1},$$

say, implying that

$$\mu^* \circ T^{-1}\{(\mathrm{d}r,\mathrm{d}w)\} = \frac{D}{r^2}\mathrm{d}r \times \nu(\mathrm{d}w). \tag{27}$$

☐ The appearance of D in (**??**) ensures that ν has unit measure, as we shall see below.

☐ Expression (**??**) is a product, so R and W are independent, and (**??**) yields

$$\begin{aligned} V(z) = \mu^*(\mathcal{A}_z^*) = \mu^* \circ T^{-1} \circ T(\mathcal{A}_z^*) &= \mu^* \circ T^{-1}\left[\{(r,w) : r > 1/a_z(w)\}\right] \\ &= D\iint_{\{(r,w):r>1/a_z(w)\}} r^{-2}\mathrm{d}r\,\nu(\mathrm{d}w) \\ &= D\int \left[-r^{-1}\right]_{1/a_z(w)}^{\infty} \nu(\mathrm{d}w) \\ &= D\int_{\mathbb{S}_{D-1}} \max_d\left(\frac{w_d}{z_d}\right)\nu(\mathrm{d}w) \\ &= D\mathrm{E}\left\{\max_d(W_d/z_d)\right\}. \end{aligned}$$

☐ As the margins of G^* are unit Fréchet, when all but one of the z_d are set to infinity we have

$$D\mathrm{E}\left\{\max_d(W_d/z_d)\right\} = D\mathrm{E}(W_d)/z_d = 1/z_d, \quad d = 1,\dots,D.$$

Hence $\mathrm{E}(W_d) = 1/D$, concluding the proof.

☐ To check that ν is a probability measure, note that

$$\int_{\mathbb{S}_{D-1}} \nu(\mathrm{d}w) = \int_{\mathbb{S}_{D-1}} (w_1 + \cdots + w_D)\,\nu(\mathrm{d}w) = \sum_{d=1}^{D} \mathrm{E}(W_d) = D \times 1/D = 1.$$

note 1 of slide 190

Bivariate maxima

- ☐ If $D = 2$, then $W_1 = 1 - W_2 = W$, say,

$$V(z_1, z_2) = 2\mathrm{E}\left\{\max\left(\frac{W}{z_1}, \frac{1-W}{z_2}\right)\right\} = 2\int_0^1 \max\left(\frac{w}{z_1}, \frac{1-w}{z_2}\right)\nu(\mathrm{d}w),$$

  and $W \sim \nu$, an **angular (or spectral) distribution function** on $[0, 1]$, such that

$$\mathrm{E}(W) = \int_0^1 w\,\nu(\mathrm{d}w) = 1/2.$$

- ☐ If ν has an angular density function $\dot{\nu}$, then

$$V(z_1, z_2) = 2\int_0^1 \max\left(\frac{w}{z_1}, \frac{1-w}{z_2}\right)\dot{\nu}(w)\,\mathrm{d}w.$$

Example 35 *Find the limiting distributions for maxima when (a)* $W \in \{0, 1\}$ *with equal probabilities, (b)* $W = 1/2$ *with probability one, (c)* $W \sim U(0, 1)$.

These cases are not useful statistically, but they illustrate why a general treatment must allow ν to have a mixture of point masses and density.

Note to Example 15

- ☐ Let $D = 2$ and let $\mathrm{P}(W = 0) = \mathrm{P}(W = 1) = 1/2$. Then

$$V(z_1, z_2) = 2\mathrm{E}\left[\max\{W/z_1, (1-W)/z_2\}\right] = 1/z_1 + 1/z_2,$$

  yielding $G^*(z_1, z_2) = \exp(-1/z_1)\exp(-1/z_2)$, corresponding to independence of Z_1 and Z_2. In this case (Z_1, Z_2) have a joint density function.
  In the corresponding D-dimensional case, W falls at the D corners of $\mathbb{S}_{D-1}$ with equal probabilities $1/D$.
- ☐ Let $D = 2$ and suppose that $\mathrm{P}(W = 1/2) = 1$. Then

$$V(z_1, z_2) = 2\mathrm{E}\left[\max\{W/z_1, (1-W)/z_2\}\right] = \max(1/z_1, 1/z_2),$$

  and hence $G^*(z_1, z_2) = \exp\{-1/\min(z_1, z_2)\}$, corresponding to total dependence of Z_1 and Z_2. Here $Z_1 = Z_2$ with probability one, so they have no joint density function; all the mass of their joint distribution lies on the line $z_1 = z_2$.
  In the corresponding D-dimensional case, W equals the barycentre $D^{-1}1_D$ of $\mathbb{S}_{D-1}$ with probability one.
- ☐ Let $D = 2$ and let W have the uniform distribution on $[0, 1]$. Then $w/z_1 \geq (1-w)/z_2$ when $w \geq z_1/(z_1 + z_2)$, and it is easy to check that

$$V(z_1, z_2) = 2\mathrm{E}\left[\max\{W/z_1, (1-W)/z_2\}\right] = 1/z_1 + 1/z_2 - 1/(z_1 + z_2).$$

  In this case both W and (Z_1, Z_2) have density functions, with no atoms.

Comments

☐ The strategy above was:
  - use g to transform the original data to standard (unit Fréchet) margins;
  - show that on these standard margins the limiting distribution for the transformed data has a specific nonparametric form, subject only to restrictions on the marginal means.

☐ We saw that $\mu(\mathcal{A})$ (for the original data) equals $\mu^*\{g(\mathcal{A})\}$ (for the transformed data), so
  - since g is monotone on each axis, $\mathcal{A}_z^* = g(\mathcal{A}_z)$ has the same shape as $\mathcal{A}_z$, and
  - if we set $z = (z_1, z_2) = ((x - b_{X,n})/a_{X,n}, (y - b_{Y,n})/a_{Y,n})$, then

$$\begin{aligned} \mathrm{P}(M_{X,n} \le x, M_{Y,n} \le y) &= \mathrm{P}\{(M_{X,n} - b_{X,n})/a_{X,n} \le z_1, (M_{Y,n} - b_{Y,n})/a_{Y,n} \le z_2\} \\ &\approx \exp\{-\mu(\mathcal{A}_z)\} \\ &= \exp[-\mu^*\{g(\mathcal{A}_z)\}] \\ &= \exp[-V\{g(z)\}], \end{aligned}$$

  gives an approximate distribution function for $(M_{X,n}, M_{Y,n})$, which we can use to find their joint density.

☐ For statistical purposes we need models for V, or equivalently ν or $\dot{\nu}$, . . .

5.4 Statistical Modelling

Introduction

☐ The previous section gave a general framework for multivariate extremes, in which a general nonparametric model appears.

☐ Two approaches to modelling are:
  - estimating the full nonparametric model — difficult because of limited data;
  - fitting a parametric sub-family of models — may be restrictive, but often is good enough.

☐ To do this we need
  - some (reasonably flexible) parametric families of models,
  - methods to fit them (likelihood),
  - methods to assess their fit and to compare them.

Parametric models

It is tricky to formulate parametric models that satisfy the mean constraints in $D \geq 3$, but numerous models exist for $D = 2$.

Example 36 *The* **logistic model** *for general D has*

$$V(z_1, \ldots, z_D) = \left(\sum_{d=1}^{D} z_d^{-1/\alpha} \right)^{\alpha}, \quad z_1, \ldots, z_D > 0, 0 < \alpha \leq 1,$$

and for $D = 2$,

$$\dot{\nu}(w) = \tfrac{1}{2}(\alpha^{-1} - 1)\{w(1-w)\}^{-1-1/\alpha}\{w^{-1/\alpha} + (1-w)^{-1/\alpha}\}^{\alpha-2}, \quad 0 < w < 1.$$

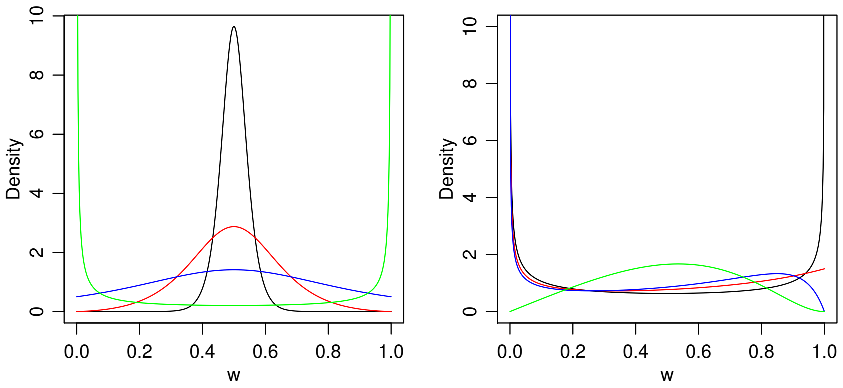
Independence and perfect dependence arise as limits as $\alpha \uparrow 1$ and $\alpha \downarrow 0$ respectively.
This model is limited by having only one parameter, which makes it symmetric and too inflexible for most purposes. The same applies to the bivariate **negative logistic model**, *which has*

$$V(z_1, z_2) = 1/z_1 + 1/z_2 - (z_1^{\alpha} + z_2^{\alpha})^{-1/\alpha}, \quad \alpha > 0,$$

for which independence and perfect dependence arise when $\alpha \to 0$ and $\alpha \to \infty$.

Logistic and Dirichlet densities

Left: logistic densities $\dot{\nu}$ with $\alpha = 0.1$ (black), 0.3 (red), 0.5 (blue), 0.9 (green).
Right: Dirichlet densities $\dot{\nu}$ with parameters $(\alpha, \beta) = (0.5, 0.5)$ (black), (0.5, 1) (red), (0.5, 2) (blue) and (2,3) (green).



Hüsler–Reiss distribution

Example 37 (Hüsler–Reiss distribution) *This is a natural analogue of the normal distribution in extremal contexts. The bivariate version has a scalar parameter $\lambda > 0$ and exponent function*

$$V(z_1, z_2) = \frac{1}{z_1}\Phi\left\{\frac{\lambda}{2} + \frac{1}{\lambda}\log\left(\frac{z_2}{z_1}\right)\right\} + \frac{1}{z_2}\Phi\left\{\frac{\lambda}{2} + \frac{1}{\lambda}\log\left(\frac{z_1}{z_2}\right)\right\}, \quad z_1, z_2 > 0, \tag{28}$$

where Φ denotes the standard normal cumulative distribution function.
For this model the angular variable W has density

$$\dot{\nu}(w) = \frac{e^{-\lambda^2/8}}{2\lambda\{w(1-w)\}^{3/2}}\phi\left\{\frac{1}{\lambda}\log\left(\frac{w}{1-w}\right)\right\}, \quad 0 < w < 1. \tag{29}$$

☐ Recall the angular coordinates $r = z_1 + \cdots + z_D$ and $w = (w_1, \ldots, w_D) = z/r \in \mathbb{S}_{D-1}$ defined in terms of $z = (z_1, \ldots, z_D) \in \mathcal{E}^*$.

☐ If it exists, we obtain the angular density $\dot{\nu}$ from V using the formula

$$\dot{\nu}(w) = -\frac{r^{D+1}}{D}\left.\frac{\partial^D V(z_1, \ldots, z_D)}{\partial z_1 \cdots \partial z_D}\right|_{z_1 = rw_1, \ldots, z_D = rw_D}, \quad w \in \mathbb{S}_{D-1}, \quad r > 0.$$

Asymmetric models

Asymmetric models include:

☐ the **bilogistic** model

$$\dot{\nu}(w) = \tfrac{1}{2}(1-\alpha)(1-w)^{-1}w^{-2}(1-u)u^{1-\alpha}\{\alpha(1-u) + \beta u\}^{-1}, \quad 0 < w < 1,$$

where $0 < \alpha, \beta < 1$, and $u = u(w, \alpha, \beta)$ satisfies

$$(1-\alpha)(1-w)(1-u)^{\beta} - (1-\beta)wu^{\alpha} = 0;$$

☐ and the **Dirichlet** model

$$\dot{\nu}(w) = \frac{\alpha\beta\Gamma(\alpha+\beta+1)(\alpha w)^{\alpha-1}\{\beta(1-w)\}^{\beta-1}}{2\Gamma(\alpha)\Gamma(\beta)\{\alpha w + \beta(1-w)\}^{\alpha+\beta+1}}, \quad 0 < w < 1,$$

for parameters $\alpha, \beta > 0$.

☐ The `R` function `evd::fbvevd` (see also `evd::dbvevd`) fits several bivariate models, including all those above.

Pickands' dependence function

☐ **Pickands' dependence function** A, determined by

$$V(z_1, z_2) = \left(\frac{1}{z_1} + \frac{1}{z_2}\right) A\left(\frac{z_1}{z_1 + z_2}\right),$$

gives a useful summary of dependence in bivariate problems. We have

(a) $\max(t, 1-t) \le A(t) \le 1$ for $t \in [0, 1]$;

(b) $A(t) = 1$ for independent data, and $A(t) = \max(t, 1-t)$ for perfectly dependent data;

(c) $A(t)$ is convex in t; and

(d) we can write

$$A(t) = 1 - t + 2\int_0^t \nu([0, w])\, \mathrm{d}w, \quad 0 \le t \le 1.$$

☐ This last formula enables the computation of ν from A, since

$$\nu([0, w]) = \begin{cases} \{1 + A'(w)\}/2, & 0 \le w < 1, \\ 1, & w = 1, \end{cases}$$

where A' is the right-hand derivative of A. Further differentiation gives $\dot{\nu}$, if it exists.

Note: Pickands' dependence function

(a) First, note that with $t = z_1/(z_1+z_2)$, we have

$$A(t) = z_1 z_2 V(z_1,z_2)/(z_1+z_2) = V\{(z_1+z_2)/z_2, (z_1+z_2)/z_1\} = V\{1/(1-t), 1/t\}, \quad 0 \le t \le 1.$$

Now since $V > 0$, comparison of the sets $\mathcal{A}_{(z_1,z_2)}$, $\mathcal{A}_{(z_1,\infty)}$ and $\mathcal{A}_{(\infty,z_2)}$ shows that

$$V(z_1,z_2) \le V(z_1,\infty) + V(\infty,z_2) = 1/z_1 + 1/z_2,$$

and hence $A(t) \le 1$, and likewise

$$V(z_1,z_2) \ge \max\{V(z_1,\infty), V(\infty,z_2)\} = \max(1/z_1, 1/z_2),$$

giving $A(t) \ge \max(t, 1-t)$.

(b) To check the values of A for dependent and independent data, note that

$$A(t) = 2\int_0^1 \max\{w(1-t), (1-w)t\}\,\nu(\mathrm{d}w),$$

and insert the appropriate ν. For example, if $\nu(\{0\}) = \nu(\{1\}) = 1/2$, then $A(t) = 2\{t/2 + (1-t)/2\} = 1$, and if $\nu(\{1/2\}) = 1$, then $A(t) = 2\max\{(1-t)/2, t/2\} = \max(t, 1-t)$, corresponding to the independent and fully dependent models respectively.

(c) For the convexity, note that the function $\max(ax, by)$ is convex for $a, b \ge 0$ and $x, y > 0$, and that linear combinations (with positive coefficients) of convex functions are convex. Thus the (possibly infinite) linear combination of such functions, $A(t)$, is convex in t.

(d) The final part is a bit more delicate. We can write

$$A(t) = 2\left\{(1-t)\int_{(t,1]} w\,\nu(\mathrm{d}w) + t\int_{[0,t]} (1-w)\,\nu(\mathrm{d}w)\right\},$$

and the first integral may be expressed as

$$\begin{aligned}
\int_{(t,1]} w\,\nu(\mathrm{d}w) &= \int_{(t,1]} \{1 + (w-1)\}\,\nu(\mathrm{d}w) \\
&= 1 - \nu([0,t]) - \left\{\tfrac{1}{2} - \int_{[0,t]} (1-w)\,\nu(\mathrm{d}w)\right\} \\
&= \tfrac{1}{2} - \nu([0,t]) + \int_{[0,t]} (1-w)\,\nu(\mathrm{d}w).
\end{aligned}$$

We can write the remaining integral as

$$\begin{aligned}
\int_{[0,t]} (1-w)\,\nu(\mathrm{d}w) &= \int_{[0,t]}\int_w^1 \mathrm{d}u\,\nu(\mathrm{d}w) = \int_0^1 \int_{[0,\min(u,t)]} \nu(\mathrm{d}w)\mathrm{d}u \\
&= \int_0^t \nu([0,u])\,\mathrm{d}u + (1-t)\nu([0,t]).
\end{aligned}$$

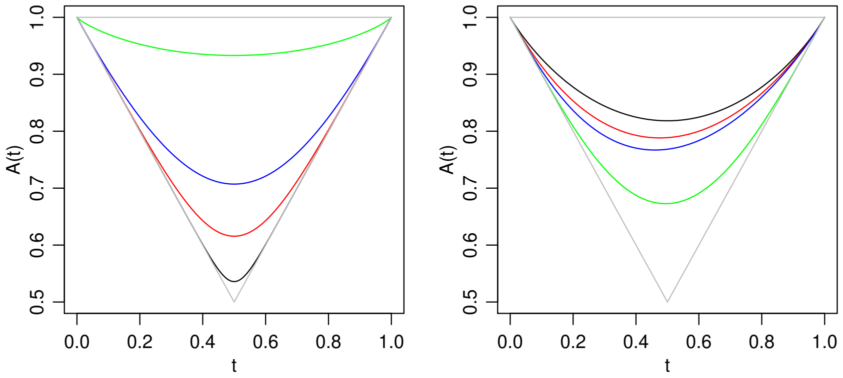
Putting the bits together we get

$$A(t) = 1 - t + 2\int_0^t \nu([0,u])\,\mathrm{d}u, \quad 0 \le t \le 1.$$

note 1 of slide 199

Pickands' dependence functions

Left: for logistic density with $\alpha = 0.1$ (black), 0.3 (red), 0.5 (blue), 0.9 (green).
Right: for Dirichlet density with parameters $(\alpha, \beta) = (0.5, 0.5)$ (black), (0.5, 1) (red), (0.5, 2) (blue) and (2,3) (green).



Extremal coefficient

- ☐ A common scalar summary of dependence between $(Z_1, \ldots, Z_D)$ with CDF

$$\mathrm{P}(Z_1 \le z_1, \ldots, Z_D \le z_D) = \exp\left\{-V(z_1, \ldots, z_D)\right\}, \quad z_1, \ldots, z_D > 0,$$

is the so-called **extremal coefficient** $\theta = V(1, \ldots, 1)$, which

- satisfies $1 \le \theta \le D$, and
- can be interpreted as the 'number of independent maxima' underlying $Z_1, \ldots, Z_D$, because the homogeneity of V gives

$$\begin{aligned} \mathrm{P}\{\max(Z_1, \ldots, Z_D) \le z\} &= \mathrm{P}(Z \le z, \ldots, Z_D \le z) \\ &= \exp\left\{-V(z, \ldots, z)\right\} \\ &= \exp\left\{-V(1, \ldots, 1)/z\right\} \\ &= (e^{-1/z})^{\theta}, \end{aligned}$$

so smaller θ corresponds to stronger dependence.

- ☐ For asymptotically dependent models and $D = 2$,

$$\chi = \lim_{z \to \infty} \mathrm{P}(Z_2 > z \mid Z_1 > z) = 2 - \theta = 2\{1 - A(1/2)\}.$$

Marginal transformation

☐ The models for multivariate maxima have unit Fréchet margins, but data do not, so for each margin we have

$$Z_d = \left(1 + \xi_d \frac{Y_d - \eta_d}{\tau_d}\right)_+^{1/\xi_d}, \quad d = 1, \ldots, D,$$

in terms of the original component-wise maxima $Y = (Y_1, \ldots, Y_D)^{\mathrm{T}}$, or in vector form,

$$Z_{D\times 1} = \left(1 + \xi \frac{Y - \eta}{\tau}\right)_+^{1/\xi},$$

where η, τ, ξ are vectors and addition, etc. are component-wise.

☐ The distribution of the maxima Y is therefore assumed to be

$$\mathrm{P}(Y \le y) = G^* \left\{ \left(1 + \xi \frac{y - \eta}{\tau}\right)_+^{1/\xi} \right\}, \quad x \in \mathbb{R}^D,$$

where $G^*(z) = \exp\{-V(z)\}$ is a simple extreme-value distribution.

☐ There are at least $3D + 1$ parameters (3 for each margin, and at least 1 for V).

Inference for multivariate maxima

☐ Inference involves:
  - fitting of marginal GEV distributions and transformation to standard Fréchet;
  - choice of dependence model V (or equivalently ν);
  - estimation of ν (by maximum likelihood in parametric cases);
  - model checking;
  - computation of probabilities for events of interest.

☐ Ideally all the estimation is performed at once, by fitting marginal and dependence models together (not always feasible in complex cases).

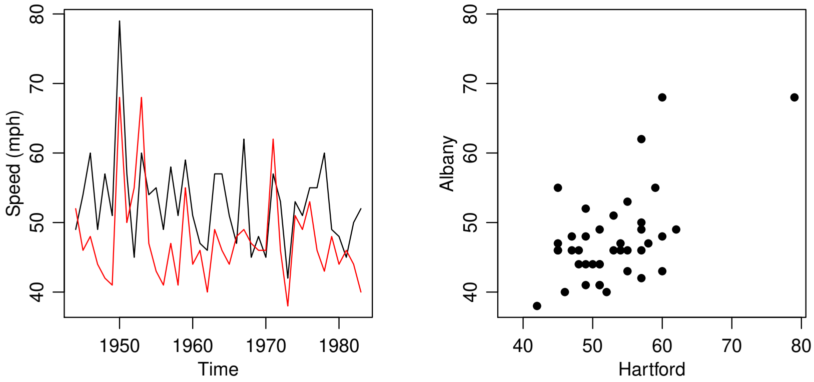
☐ In the bivariate case, the joint density for the maxima (Y_1, Y_2) can be written as

$$f(y_1, y_2) = \frac{\partial z_1}{\partial y_1}\frac{\partial z_2}{\partial y_2} \times \left\{ \frac{\partial V(z_1, z_2)}{\partial z_1}\frac{\partial V(z_1, z_2)}{\partial z_2} - \frac{\partial^2 V(z_1, z_2)}{\partial z_1 \partial z_2} \right\} \times \exp\left\{-V(z_1, z_2)\right\},$$

where the first two (Jacobian) terms depend on the marginal parameters and the remainder depend both on those parameters and those of V (or equivalently ν).

☐ For larger D the number of terms with derivatives of V increases very rapidly, but the structure of the density is the same.

Example: Wind data



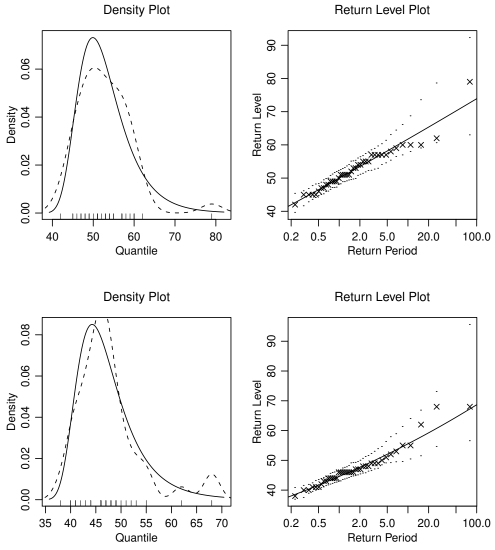
Annual maximum wind speeds at Albany, New York and Hartford, Connecticut respectively, from 1944–1983.

Example: Wind data

```
(fit1<-fbvevd(wind,model="log"))  # Fit logistic dependence function

Call: fbvevd(x = wind, model = "log")
Deviance: 492.1304
AIC: 506.1304
Dependence: 0.3658468

Estimates
    loc1     scale1     shape1       loc2     scale2     shape2        dep
49.96955    5.03097    0.01413   44.58484    4.33938    0.07879    0.70854

Standard Errors
   loc1    scale1    shape1      loc2    scale2    shape2       dep
0.87434   0.63662   0.08826   0.76813   0.56747   0.11101   0.09742

plot(fit1,mar=1,which=c(3,4))
plot(fit1,mar=2,which=c(3,4))
plot(fit1,which=c(3:6))
```

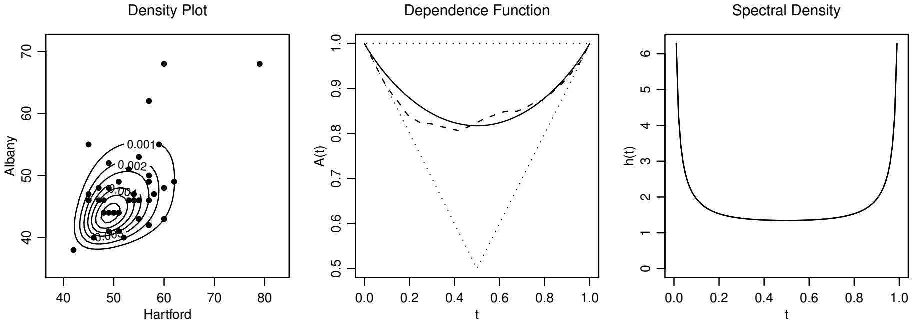
Example: Wind data

Diagnostic plots for the two marginal fits:



Example: Wind data

Diagnostic plots for the fitted logistic model:



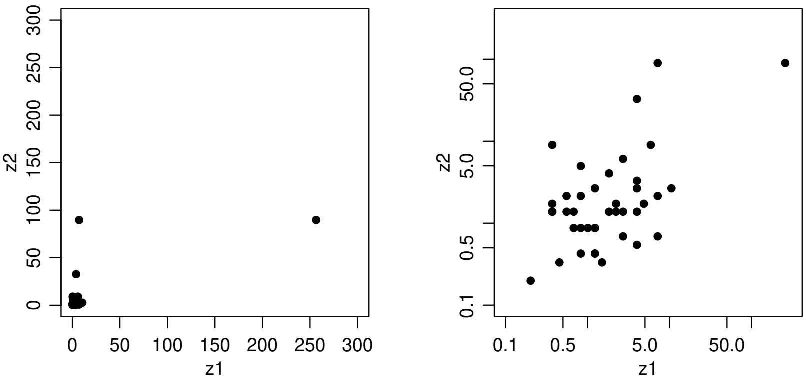
What is wrong with (a) the empirical Pickands function? (b) the spectral density?

Example: Wind data

Residuals

$$\widehat{z}_1 = \left\{1+\widehat{\xi}_1(y_1-\widehat{\eta}_1)/\widehat{\tau}_1\right\}_+^{1/\widehat{\xi}_1}, \quad \widehat{z}_2 = \left\{1+\widehat{\xi}_2(y_2-\widehat{\eta}_2)/\widehat{\tau}_2\right\}_+^{1/\widehat{\xi}_2}$$

from the fitted model, on the unit Fréchet (left) and Gumbel (right) scales. Note the size of the event in 1950.



Example: Wind data

```
# fit without 1950

(fit2<-fbvevd(wind[-7,],model="log"))

Call: fbvevd(x = wind[-7, ], model = "log")
Deviance: 466.0202
AIC: 480.0202
Dependence: 0.2858570

Estimates
    loc1     scale1     shape1       loc2     scale2     shape2        dep
50.45888    4.98736   -0.31263   44.41348    4.16650    0.08284    0.77749

Standard Errors
  loc1   scale1   shape1     loc2   scale2   shape2      dep
0.9011   0.6727   0.1355   0.7471   0.5383   0.1072   0.1004
```

The dependence parameter has increased, but still is significantly less than $\alpha = 1$. The shape parameter estimate $\widehat{\xi}_1$ is now significantly negative.

Comments

We see that:

- ☐ the marginal fits appear to be adequate (density plots not very useful);
- ☐ the joint fit appears to be reasonable (though there is not much data);
- ☐ the empirical Pickands function is not convex (!), but matches the fitted one fairly well;
- ☐ the angular density suggests that the windspeeds are not very dependent (the fitted angular density shows spikes at $w = 0, 1$), though the standard error of around 0.1 for $\widehat{\alpha} = 0.71$ shows that the data are clearly not completely independent;
- ☐ the 1950 event is quite influential—is it special in any way?

Example: Wind data

Comparison with other models: recall that the deviance equals $-2\widehat{\ell}$ and that $\mathrm{AIC} = -2\widehat{\ell} + 2p$, for a fitted model with p parameters and maximised log likelihood value $\widehat{\ell}$. Small AIC is better.

Dependence function	Paras	Deviance	AIC	$2\{1 - A(1/2)\}$
Logistic	7	492.13	506.13	0.37
Hüsler–Reiss	7	491.20	505.20	0.37
Negative logistic	7	491.57	505.57	0.36
Asymmetric negative logistic	9	491.75	509.75	0.38
Bilogistic	9	489.79	505.79	0.29
Coles–Tawn	8	489.97	505.97	0.38
Asymmetric logistic	—	—	—	—
Asymmetric mixed	—	—	—	—

More comments

- ☐ The Hüsler–Reiss model seems to be best of those that could be fitted, but there is very little difference among them.
- ☐ There is no evidence of a need for an asymmetric dependence function.
- ☐ The dependence is not very strong; this measure is $2 - \theta$, so $\widehat{\theta} \approx 1.63$, corresponding to

$$\mathrm{P}(Z_2 > z \mid Z_1 > z) \approx 2 - \theta \approx 0.37;$$

  this is appreciable but not strong dependence.
- ☐ There is probably a big loss of information due to using only the annual maxima, but at least these can (probably) be treated as independent (need to check the dates to be sure).

Poisson process approach

☐ If $X_1, \dots, X_n \overset{\text{iid}}{\sim} F$ lie in $\mathbb{R}^D$, then under the conditions for convergence of the maxima, we can define component-wise transformations

$$g(X) = \{1 + \xi(X - b_n)/a_n\}_+^{1/\xi}, \quad X, \xi, b_n \in \mathbb{R}^D, \quad a_n \in \mathbb{R}_+^D,$$

such that the sequence of point processes

$$\mathcal{P}_n^* = \{g(X_1), \dots, g(X_n)\}, \quad n = 1, 2, \dots,$$

converges to a Poisson process $\mathcal{P}^*$ on $\mathcal{E}^* = \mathbb{R}_+^D - \{0\}$ with exponent function V and measure μ^*, where $V(z) = \mu^*(\mathcal{A}_z^*)$, and $\mathcal{A}_z^* = \mathcal{E}^* - [0, z_1] \times \cdots \times [0, z_D]$

☐ The corresponding density based on the points $z_1, \dots, z_{n_{\mathcal{A}^*}}$ in an 'extreme' set $\mathcal{A}^* \subset \mathcal{E}^*$ is

$$\exp\{-\mu^*(\mathcal{A}^*)\} \times \prod_{j=1}^{n_{\mathcal{A}^*}} \mu^*(\mathrm{d}z_j),$$

with $\mu^*(\mathrm{d}z_j)$ replaced by $\dot{\mu}^*(z_j)$ when μ^* has a density function.

☐ If the limit process is assumed to be exact for events in $\mathcal{A}^*$ (see below), we can base inference for μ^* on this expression.

Marginal transformation

☐ If we are interested in an extreme region $\mathcal{B}$ in $\mathcal{E}$, we choose a region $\mathcal{A}^* \subset \mathcal{E}^*$ in the transformed data space such that $g(\mathcal{B}) \subset \mathcal{A}^*$ and use the Poisson process model.

☐ For the **marginal transformation** g, we take thresholds $u_1, \dots, u_D$, usually corresponding to the same quantile (e.g., 0.95) of each margin; n_{u_d} observations exceed these thresholds, and $\widehat{p}_d = n_{u_d}/n$ is the estimated exceedance probability for dimension d.

☐ We fit GPDs above these thresholds, giving fitted marginal distributions

$$\widehat{F}_d(x) = \begin{cases} \#\{j : x_{j,d} \le x\}/n, & x \le u_d, \\ 1 - \widehat{p}_d \left\{1 + \widehat{\xi}_d(x - u_d)/\widehat{\sigma}_d\right\}_+^{-1/\widehat{\xi}_d}, & x > u_d, \end{cases} \quad d = 1, \dots, D,$$

based on the D marginal GP parameter estimates $(\widehat{\sigma}_d, \widehat{\xi}_d)$.

☐ We then apply this estimated probability integral transformation component-wise to $x_j = (x_{j,1}, \dots, x_{j,D}) \in \mathcal{E}$ to get

$$z_j = -1/\log \widehat{F}(x_j), \quad j = 1, \dots, n,$$

which lie in $\mathcal{E}^*$ and have approximate unit Fréchet margins.

☐ There are corresponding angular variables

$$r_j = \|z_j\|_1 > 0, \quad w_j = z_j/r_j \in \mathbb{S}_{D-1}, \quad j = 1, \dots, n.$$

Extremal region $\mathcal{A}^*$

☐ We have to choose an 'extreme' region $\mathcal{A}^*$ on which to base the likelihood

$$\exp\{-\mu^*(\mathcal{A}^*)\} \times \prod_{j=1}^{n_{\mathcal{A}^*}} \mu^*(\mathrm{d}z_j), \quad \mathcal{A}^* \subset \mathcal{E}^*.$$

☐ In most cases μ^* has a tractable density $\dot{\mu}^*$, so the bottleneck is computation of $\mu^*(\mathcal{A}^*)$.

☐ In terms of the angular coordinates, $\mu^*(\mathrm{d}z_j) = r_j^{-2}\mathrm{d}r \times D\,\nu(\mathrm{d}w_j) \propto \dot{\nu}_\theta^*(w_j)$, if there is a parametric angular density.

☐ If $\mathcal{A}^* = \{(x, y) : x + y > r_0\}$ for some large r_0, then

$$\mu^*(\mathcal{A}^*) = 2\int_{\mathcal{A}^*} \frac{\mathrm{d}r}{r^2}\nu(\mathrm{d}w) = 2\int_{r=r_0}^{\infty} \frac{\mathrm{d}r}{r^2}\int_{w=0}^{1} \nu(\mathrm{d}w) = 2/r_0,$$

does not depend on parameters of ν^*. In this case the likelihood is

$$L(\theta) = \exp\{-\mu^*(\mathcal{A}^*)\}\prod_{j=1}^{n_{\mathcal{A}^*}} \dot{\mu}^*(z_j) \propto \prod_{j=1}^{n_{\mathcal{A}^*}} \dot{\nu}_\theta^*(w_j),$$

since none of the other terms depend on $\dot{\nu}_\theta^*$ (or θ).

☐ This likelihood is simple but not much used, because non-extreme data can corrupt the fit.

Censored likelihood

☐ For simplicity suppose $D = 2$ and $z = (z_1, z_2) \in \mathcal{E}^*$.

☐ Even if r_0 is large, sets such as

$$\mathcal{A}^* = \{(z_1, z_2) : z_1 + z_2 > r_0 > 0\}$$

contain values for which one of z_1 and z_2 is small, so asymptotic models may not apply.

☐ To fix this we split $\mathcal{E}^*$ into subsets

$$\begin{aligned}
\mathcal{E}_{00}^* &= \{(z_1, z_2) : z_1 \le u_1, z_2 \le u_2\}, & \mathcal{E}_{10}^* &= \{(z_1, z_2) : z_1 > u_1, z_2 \le u_2\},\\
\mathcal{E}_{01}^* &= \{(z_1, z_2) : z_1 \le u_1, z_2 > u_2\}, & \mathcal{E}_{11}^* &= \{(z_1, z_2) : z_1 > u_1, z_2 > u_2\},
\end{aligned}$$

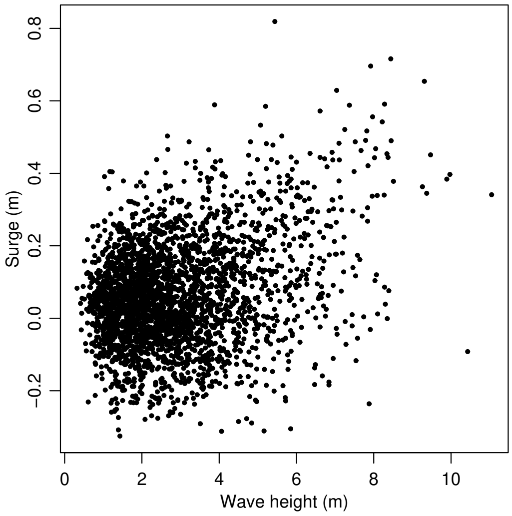
with respective likelihood contributions based on

$$\begin{array}{cc}
\mathrm{P}(Z_1 \le u_1, Z_2 \le u_2), & \dfrac{\partial \mathrm{P}(Z_1 \le z_1, Z_2 \le u_2)}{\partial z_1},\\
\dfrac{\partial \mathrm{P}(Z_1 \le u_1, Z_2 \le z_2)}{\partial z_2}, & \dfrac{\partial^2 \mathrm{P}(Z_1 \le z_1, Z_2 \le z_2)}{\partial z_1 \partial z_2}.
\end{array}$$

This uses the full information about the values of (z_1, z_2) only in $\mathcal{E}_{11}^*$, and otherwise just uses the information that $z_{1,j}$ or $z_{2,j}$ falls below the appropriate threshold.

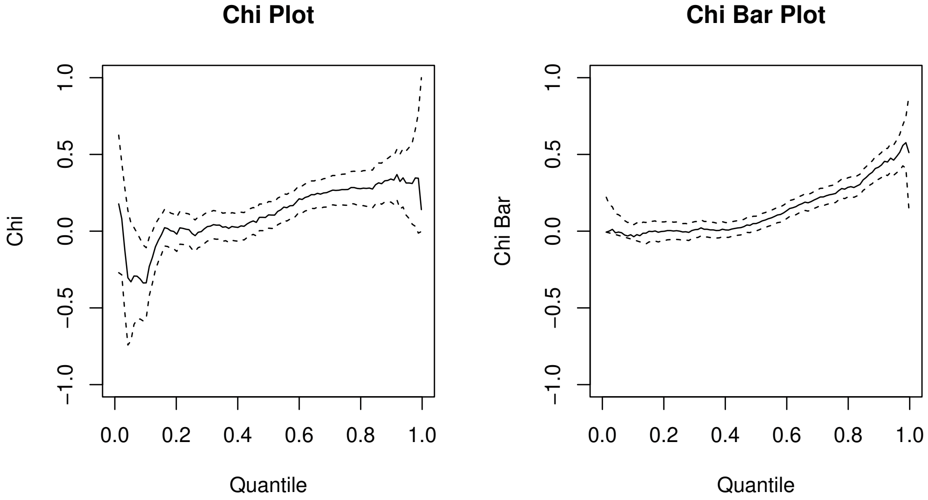
☐ This **censored likelihood** is the default for fitting Poisson process models to multivariate data.

Example: Oceanographic data



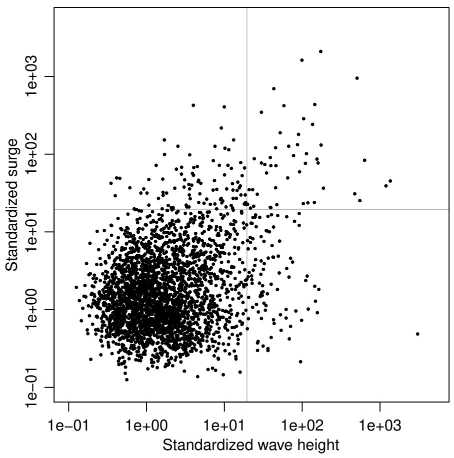
Simultaneous values of wave and surge height.

Example: Oceanographic data



Estimates of $\chi(u)$ and $\overline{\chi}(u)$. The wide confidence intervals as $u \to 1$ are typical (and indeed inevitable) and complicate the interpretation of such plots.

Example: Oceanographic data



Simultaneous values of wave and surge height, transformed to unit Fréchet scale, with regions $\mathcal{E}^*_{00}, \ldots, \mathcal{E}^*_{11}$ determined by the grey lines marking the marginal 0.95 quantiles.

Example: Oceanographic data

```
> (fit <- fbvpot(wavesurge,apply(wavesurge,2,quantile,0.95),model="log"))

Call: fbvpot(x = wavesurge, threshold = apply(wavesurge, 2, quantile,0.95), model = "log")
Deviance: 2036.076
AIC: 2046.076
Dependence: 0.3072850

Threshold: 6.08 0.322
Marginal Number Above: 144 144
Marginal Proportion Above: 0.0498 0.0498
Number Above: 49
Proportion Above: 0.0169

Estimates
   scale1     shape1     scale2     shape2        dep
 1.261341  -0.134651   0.091877   0.008904   0.759339

Standard Errors
 scale1   shape1   scale2   shape2      dep
0.13162  0.06908  0.01067  0.08568  0.02945
```

Example: Oceanographic data

Results for fits: recall that the deviance equals $-2\widehat{\ell}$ and that $\mathrm{AIC} = -2\widehat{\ell} + 2p$, for a fitted model with p parameters and maximised log likelihood value $\widehat{\ell}$. Small AIC is better.

Dependence function	Paras	Deviance	AIC	$2\{1 - A(1/2)\}$
Logistic	5	2036.08	2046.08	0.31
Hüsler–Reiss	5	2035.38	2045.38	0.30
Negative logistic	5	2034.91	2044.91	0.30
Bilogistic	6	2035.80	2047.80	0.31
Coles–Tawn	6	2035.35	2047.35	0.31
Negative bilogistic	6	2034.85	2046.85	0.31
Asymmetric mixed	6	2044.15	2056.15	0.33
Asymmetric logistic	7	2036.60	2050.60	0.31
Asymmetric negative logistic	7	2035.78	2049.78	0.31

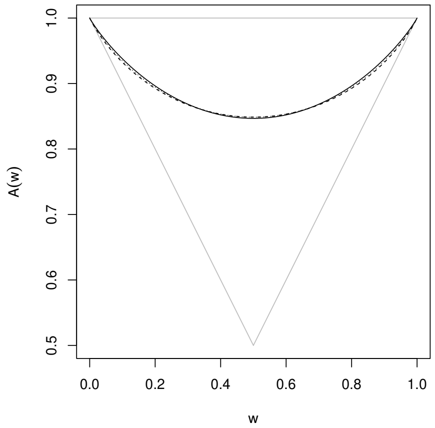
The first three fits seem best, based on the AIC values, but the differences are not large overall; it seems that the data are not complex enough to warrant a very complex (or asymmetric) model.

Example: Oceanographic data

Model	Deviance	χ	σ_1	ξ_1	σ_2	ξ_2	Dep
Logistic	2036.08	0.31	$1.26_{0.13}$	$-0.13_{0.07}$	$0.09_{0.01}$	$0.01_{0.09}$	$0.76_{0.03}$
Hüsler–Reiss	2035.38	0.30	$1.25_{0.13}$	$-0.11_{0.07}$	$0.09_{0.01}$	$0.03_{0.08}$	$0.97_{0.07}$
Negative logistic	2034.91	0.30	$1.25_{0.13}$	$-0.12_{0.07}$	$0.09_{0.01}$	$0.01_{0.08}$	$0.58_{0.06}$

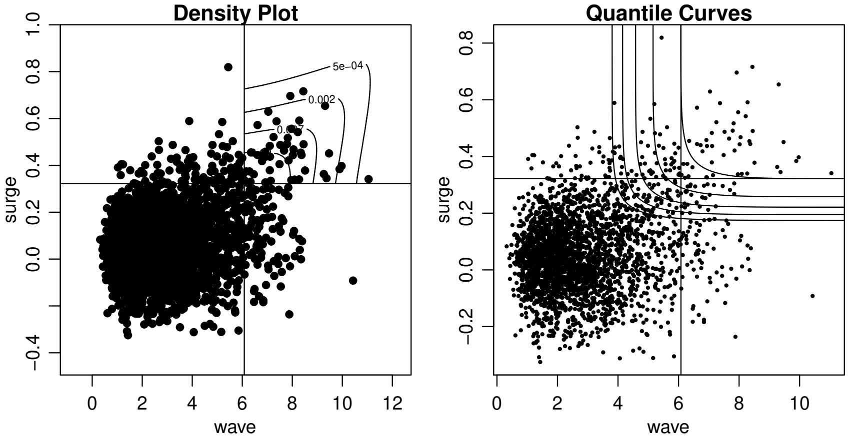
- ☐ The fits are very similar, showing data whose extremes are clearly dependent, even if the probability of high extremes on one variable given them on the other is not very high.
- ☐ The dependence parameters are not comparable (because the models are not the same), but they all give similar values of χ.
- ☐ The same censored fitting approach can be used when $D \geq 3$, but the coding becomes quite complex, with 2^D cases.

Example: Oceanographic data



Fitted Pickands functions for the logistic (solid), Hüsler–Reiss (dots) and negative logistic (dashed) models. They are essentially identical.

Example: Oceanographic data



Diagnostic plots (from `plot(fit)`) for threshold fit to oceanographic data. (A bug prevents the points on the left from being smaller.) According to the help, the quantiles are at 0.75, 0.8, . . . , 0.95 by default.