UNIVERSITE CATHOLIQUE DE LOUVAIN

FACULTE DES SCIENCES

ECOLE DE STATISTIQUE, BIOSTATISTIQUE
ET SCIENCES ACTUARIELLES

Temporal analysis of the evolution of extreme values using climatological data

Promoteur : Johan Segers

Co-promoteur : Prénom Nom

Lecteurs : Anna Kiriliouk

Michel Crucifix

Mémoire présenté en vue de l'obtention du

Master en statistiques, orientation générale

par : **Antoine Pissoort**

Juin 2017

Contents

List of Figures

Abstract

This thesis aims to

We took advantage of a high-level language (c++) to make our analysis efficient and also made use of parallel computing to decrease computation time for time consuming...

Acknowledgements

I would first like to thank my thesis supervisor Johan Segers for all his help and his guidance during this whole year. The repeated appointments we have made have

I also would like to thank the "Institut Royal de Météorologie" (IRM) of Belgium for his help, his guidance and his provided quality datasets.

List of Abbreviations

For convenience, we place a list of all the abbreviations we will use during the text. However, they will always be defined for the first occurrence in the text.

GEV ..... Generalized Extreme Value

df ..... distribution function

GPD ..... Generalized Pareto Distribution or Generalized Pareto Distribution function

EVI ..... Extreme Value Index (ξ)

EVT ..... Extreme Value Theory

MCMC ..... Marko Chain Monte Carlo

MH ..... Metropolis-Hastings (algorithm)

Introduction : Presentation of the Problem

Unlike his counterparts (see for example credit risk analysis, financial applications,...), the extreme value analysis applied on the broad environemental area like here for the meteorological data, has strong impacts on the people lives

An important question is still whether climate changes caused by anthropogenic activities will change the intensity and frequency of extreme events **?**.

The problem we are here facing in climate change evidence is that of le lack of past data to compare with her

Also, for such an analysis, the number of parameters to take into account is considerable (and tend to infinity)

Can make a parallelism with Chaos Theory and the well-known butterfly effect which have strong applications in weather models

We highly expect the climate change to affect the extreme weather

[extremes in climate change p.347]

It has been proven that winter become warmer in context of RC. (see naveau,...)

?

"The first myth about climate extremes, which has been purported by researchers in climatology or hydrology, among them prominent names, is that “extremes are defined as rare events” or similar. This myth is debunked by a simple bimodal PDF (Fig. 6.12a). The events sitting in the tails of that distribution are not rare" (Mudelsee, 2014, pp.257)

Until now, studies on climate extremes that consider Europe have usually had a strong national signature , or have had to make use of either a dataset with daily series from a very sparse network of meteorological stations (e.g. eight stations in Moberg et al. (2000)) or standardized data analysis performed by different researchers in different countries along the lines of agreed methodologies (e.g. Brazdil et al., 1996; Heino et al., 1999) Klein Tank et al. (2002)

During this project, we will try a novel approach, that is to link directly theory and practice and hence present the concepts theoretically and then illustrate by one example retrieved from our application. Even if that could be difficult, we think this approach is advantageous for several reasons :

Extrapolation !!!! See p154 [statistical analysis of extreme book]

Voir effet de l'ilot de chaleur –> urbanisation sur les tempés !

—> artificial warming on cities stations which were not(less) urbanized 100 years ago.

[In this thesis, efforts have been made to use power of (hyper)references into the text. While this not (yet ?...) usable in printed versions, the reader may feel more comfortable in a numeric version to more easily handle the vast amount of sections, equations, references, etc... and the links that are made between them.]

Presentation of the Analysis : Temperatures from Uccle

Comparisons with freely available data

TX = Tmax, TN=Tmin.

A similar dataset is freely available on the internet. (http://lstat.kuleuven.be/Wiley/Data/ecad00045TX.txt and which was a project initially performed by the KMNI). However, we were reticent to simply analyze these data as we know that it is hard to trust internet's data, even if they come from well-known "authorities". After having made all these comparisons analysis (see start of code...), we remark effectively that there are differences in these two datasets, and hence large errors of measures can easily occur in unofficial data. It confirms the fact that is important to get reliable data if one wants to make reliable analysis. However, these differences tend to be much smaller when considering the "open shelter" version (54% of equal measurements in closed shelter VS 14.4% in the closed case). For this reason, we have confidence that this public datasets is dealing with open shelter temperatures data.

However, for **meteorological considerations**, it is always better to consider temperature's analysis in **closed shelters**. Indeed, thanks to gratefull advices from mr Tricot working for the IRM :

- It can
- It

First analysis

As expected (line 200 code local), we see that there is an upward trend for both (yearly) maxima and minima. However, we remark that this one is less pronounced for minima. That makes sense as (as mentioned above), the global warming is

Add constraint to the broken linear trend to make it continuous :

We will see that the decrease (around 1940 to 1970) is more from randomness than a real decrease (...)

TEST (statistic) for the Differences in trend !!!! !!!!!!!

From the figure 2, we can already have some remarks :

- The drop in the series around 1950 to 1975 that is made visible by the LOESS estimator (red) is probably more due to a random effect than a real decrease or freezing of the maximum temperatures at this time. (to assessfofrmally if possible)

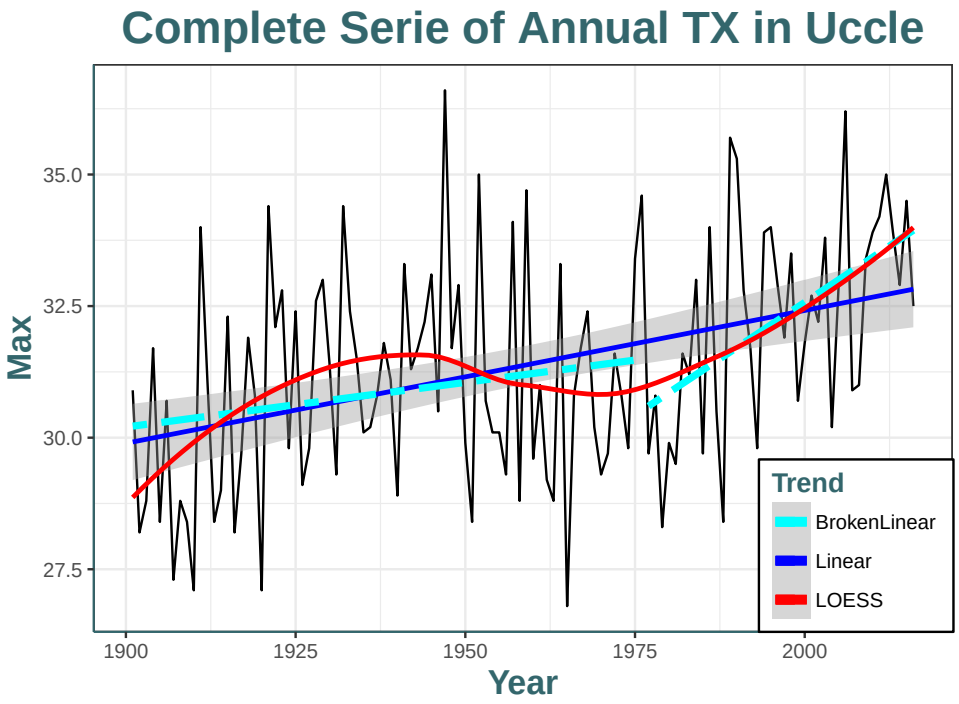

Figure 1: First plot representing the **yearly** maxima (**above**) and minima (**below**), shaded grey line representing the standard error of the linear trend (regression), red line representing the polynomial nonparametric fit by LOESS.

-

PUT the examples right in the place where it is mentioned in the theory.! "As we have seen in section 2.1.1.... and in section 2.2.2....."

We must choose a block-length which is large enough for the limiting arguments supporting the GEV approximation (see (1.5)) to be valid, either a large bias in the estimates could occur. For example, if this is too short, the maxima may be too close of each other to assume independence. But a large block-length implies less data to work with, and thus a large variance of the estimates.. So we must find a compromise between bias and variance.

TABLE with nested models (gumbel, GEV, + linear trend, etc etc)

From figure 2,

The code which provide all the tools to retrieve the presented results are left in appendix, but in the numeric version only be cause it is very heavy. this also enables you to get all insights (..)

GAM and splines

Simultaneous confidence intervals : Following Ruppert et al. (2003, section 3, 4.9, 6.5) which uses a simulation-based approach to generate a simultaneous interval **?**

From the pointwise confidence intervals we can say that (example) $f(1980)$ has 95% chance

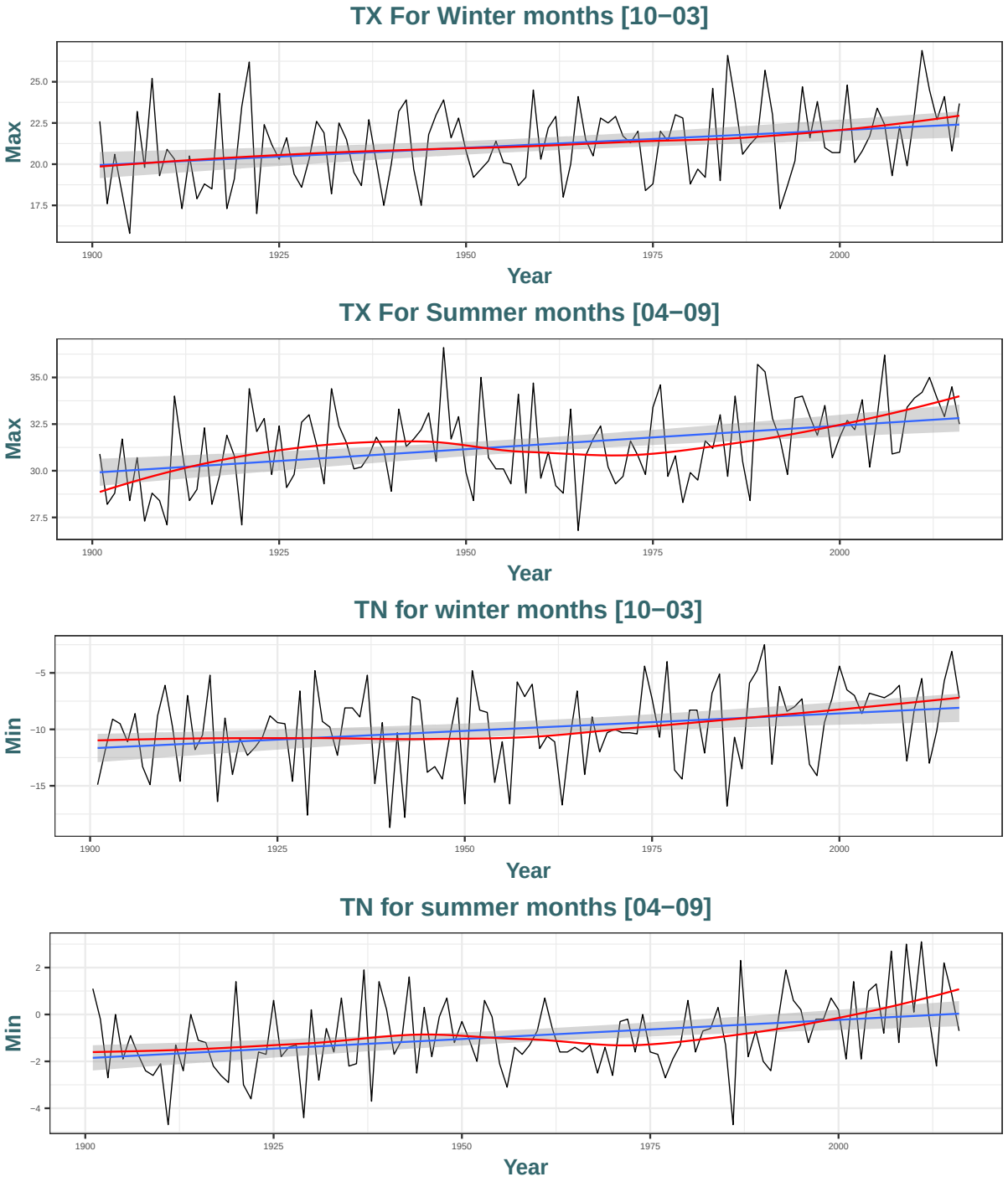

Figure 2: First plot representing the **yearly** maxima (above) and minima (below) taking only the summer months (April to September) and winter months (October to March), shaded grey line around the linear trend represents its standard error. See how the polynomial trend (red line) also changes. Obv, TX for smummer and TN for winter are the same series as for the global serie

to lie within (-1,0) (say) and $f(2000)$ has also 95% to lie within (0.2,1.2) BUT it is a fallacy to say simultaneously that both are contained in these intervals at the same time with .95 confidence. Ruppert et al. (2003, section 6.5)

We can see that only the 2 increasing periods from the start and at the end of the series are significant, while the decreasing period from... (see the red line) is more likely to be the subject of randomness. Moreover, we remark that the

After having... we will now go through with the more specific subjec of this thesis, that is the extreme value analysis.

Part I

Theoretical Framework

Chapter 1

Extreme Value Theory : Basics

Contents

There are two approaches, the block-maxima and the peaks-over-threshold approach (see section **2**) yielding to different extreme value distribution. The former aims at while the latter models the...

In this section, we will present the basics of EVT and we will consider a *block-maxima approach*. After defining some useful concepts in section **1.1**, we will

1.1 Preliminaries

Some useful definitions to start with !!

Definition 1.1 (Similar distribution functions)**.** *We say that two distributions functions G and G^* are* **similar** *or are of the* **same type** *if, for constants $a > 0$ and b*

$$G^*(az + b) = G(z), \qquad \forall z, \tag{1.1}$$

which means that the distributions differ only in location and scale. In the sequel, the concept of *similar* distributions will be useful to derive the three different families of extreme value distributions from other distributions of the *same type*. This is directly linked with max-stable process that we will define...

Definition 1.2 (Max-stability)**.** From Leadbetter et al. (1983) or Resnick (1987)*, we say that a distribution G is* **max-stable** *if, for each $n \in \mathbb{N}$*

$$G^n(a_n z + b_n) = G(z), \qquad n = 2, 3, ... \tag{1.2}$$

for some constants $a_n > 0$ and b_n.

In other words, taking powers of G results only in a change of location and scale. **??** This concept will be closely connected with the extremal limit laws in the following (). However, max-stable process are more used in a multivariate setting, see for example for an introduction.

Definition 1.3 (Min-stability)**.** *Anageously, from (Reiss and Thomas, 2007, pp.23), we say that a distribution function G is* **min-stable** *if*

$$\Pr\{X_{(1)} > d_n + c_n z\} = \bar{G}^n(d_n + c_n z) = \bar{G}(z), \tag{1.3}$$

where $c_n = a_n$, $d_n = -b_n$ and $X_{(1)}$ the minimum of the sample of size n, see (A.3).

Principles of stability Behind all the principles about Extreme Value Theory that will be covered during this thesis, will be influenced by the principle of *stability*.

As this will be .. in the following, we think useful to define precisely the concept of *non-degenerate distribution functions*.

Definition 1.4 (Non-degenerate distribution functions)**.** *We say that a distribution function is* **non-degenerate** *if*

We illustrate this by the most common theorem in statistics, the *Central Limit Theorem* (CLT) which plays typically with the empirical mean $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. We know that (..) $\bar{X}_n$ converges to the true mean μ in probability(?) and thus in distribution, that is to a non-random single point, i.e. to a *degenerate* distribution

$$\Pr\{\bar{X}_n \leq x\} = \begin{cases} 0, & x < \mu; \\ 1, & x \geq \mu. \end{cases}$$

That is not very useful, in particular for inferential purposes.

For this reason, CLT aims at finding a non-degenerate limiting distribution for $\bar{X}_n$, after allowance for normalization by sequences of constants. We will state it in his most basic form :

Theorem 0 (Central Limit Theorem)**.** *Write*[1] $\{X_i\}$ *as a sequence* (or "stochastic process"?!.. Check if written like this is fine ! for the following too) *of n iid random variables with* $E(X_i^2) < \infty$*. Then, as* $n \to \infty$*,*

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2),$$

where $\mu = E(X_i)$ *and* $\sigma^2 = V(X) > 0$*. d means convergence in distribution and the reader may refer to* appendix **A.** *for a useful short review of most important concepts of convergence for EVT.*

Then, by a proper choice of some normalizing constants, μ and $\sqrt{n}$ (as location and scale parameters respectively), we find the non-degenerate Normal distribution in the limit for the empirical mean $\bar{X}_n$.

With the same logic, we find that this is the same for the distribution of $X_{(n)}$

$$\lim_{n\to\infty} \Pr\{X_{(n)} \leq x\} = \lim_{n\to\infty} \Pr\{X_i \leq x\}^n = \begin{cases} 0, & F(x) < 1; \\ 1, & F(x) = 1. \end{cases} \tag{1.4}$$

That is, another degenerate distribution. This is exactly what Extreme Value Theory aims to achieve for (typically) the maximum order statistics $X_{(n)}$, that is finding a non-degenerate distribution in the limit by means of normalization. This will be the main subject of the next sections.

1.2 Extremal Types Theorem

Introduced by Fisher and Tippett Fisher and Tippett (1928), later revised by Gnedenko (1943) and finally streamlined by **?**, the *extremal types* theorem is very important for its applications. We remind the distribution of maxima is $\Pr\{X_{(n)} \leq x\} = F^n(x)$, from A.4 in appendix **A**. It states the following:

Theorem 1.1 (Extremal types theorem)**.** *If the distribution of partial maxima of an iid sequence of random variables with common (unknown) distribution F, say,* $X_{(n)}$*, properly normalized, converges to a non-degenerate (see ...) limiting distribution G, i.e.*

$$\lim_{n\to\infty} \Pr\Big\{a_n^{-1}(X_{(n)} - b_n) \leq z\Big\} = F^n(a_n z + b_n) = G(z), \qquad (\forall z \in \mathbb{R})), \tag{1.5}$$

and for some constants $a_n > 0$*,* $b_n \in \mathbb{R}$*.*

[extreme value and cluster ana. of euro... clustering 2011] , meaning that F is said to be in the **domain of attraction**[2] of G, denoted by $F \in D(G)$. This theorem considers an i.i.d. random

[1] We adopt this notation w.l.o.g. in the following, if there is no confusion possible. It is the same as $X_1, \ldots, X_n$

[2] ?? We will more precisely define this concept in the next section.

sample, but it holds true if the original scheme being no longer i.d. still remains independent (we will present the stationary case in section **3.1**). However, even the stationary assumption is often poor in practical applications (see for our applications in our case, the temperature....) **?** (see application...) but we will handle that in section **3.2** . with G the *Generalized Extreme Value* (GEV) distribution :

$$G(z) = \exp\left\{ -\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]_+^{-\xi^{-1}} \right\} := G_\xi(z), \tag{1.6}$$

from which we introduce the notation $y_+ = \max(y, 0)$, denoting in the above that $\{z : 1 + \xi\sigma^{-1}(z-\mu) > 0\}$ to ensure the term in the exponential is negative and the distribution function converges to 1. We will use this notation in the following so it is important to remind that this yields a vital condition(?). This will define the endpoints of the different distribution functions from the values of the shape parameter, that is $\{\xi > 0;\ \xi < 0;\ \xi = 0\}$, more details will be provided in next section . Moreover, $-\infty < \mu, \xi < \infty$ and $\sigma > 0$ with μ, σ and ξ being the three parameters of the model characterizing location, scale and shape respectively.

We think important to point out that here, the location parameter μ does not represent the mean as in the classic statistical view, but does represent the "center" of the distribution, and the scale parameter σ is not the standard deviation, but does govern the "size" of the deviations around μ. This can already be pointed out on figures in appendix where we demonstrate for little variations of the parameters.

From Coles (2001), we introduce an important theorem in Extreme Value Theory and that has many implications. This theorem simply says the following :

Theorem 1.2. *For any distribution function F,*

$$F \text{ is } \textit{max-stable} \iff F \text{ is GEV.} \tag{1.7}$$

Any distribution functions that are *max-stables* (see definition **1.2**) are also GEV (theorem **1.2**), and vice-versa. To gain interesting insights of the implications of this theorem, we think useful to give a proof but only for the "$\Leftarrow$" as the converse requires too much mathematical backgrounds.

Proof :

- If $a_n^{-1}(X_{(n)} - b_n)$ has limit distribution G for large n as in (1.5), then

$$\Pr\Big\{a_n^{-1}(X_{(n)} - b_n) \le z\Big\} \approx G(z).$$

  Hence for any integer k, since nk is large, we have

$$\Pr\Big\{a_{nk}^{-1}(X_{(n)k} - b_{nk}) \le z\Big\} \approx G(z). \tag{1.8}$$

- Since $X_{(n)k}$ is the maximum of k variables having identical distribution as $X_{(n)}$,

$$\Pr\Big\{a_{nk}^{-1}(X_{(n)k} - b_{nk}) \leq z\Big\} = \Big[\Pr\{a_{nk}^{-1}(X_{(n)} - b_{nk}) \leq z\}\Big]^k, \tag{1.9}$$

giving two expressions for the distribution of M_n, by (1.8) and (1.9) :

$$\Pr\{X_{(n)} \leq z\} \approx G\Big(a_n^{-1}(z - b_n)\Big) \quad \text{and} \quad \Pr\{X_{(n)} \leq z\} \approx G^{1/k}\Big(a_{nk}^{-1}(z - b_{nk})\Big).$$

- It follows that G and $G^{1/k}$ are identical apart from location and scale coefficients. Hence, G is *max-stable* and therefore GEV. This gives proof of the **extremal types theorem**, 1.1.

□

1.3 Characterization of the GEV distributions : 3 Types

...

The quantity $\xi \in \mathbb{R}$ in (1.6) is called the *extreme value index* (EVI) and is at the center of the analysis in extreme value theory. It determines, in some degree of accuracy, the type of the underlying distribution. Hence, from this general definition of the GEV distribution (1.6), we can directly retrieved three principal classes of EV distributions, from their *standard form*, in the α-*parametrization*, with $\alpha = \xi^{-1}$ (just show in the ξ param. ? for convenience) :

$$\boxed{\mathbf{I}} \quad G_1(z) = \exp\{-e^{-z}\}, \qquad -\infty < z < \infty. \tag{1.10}$$

$$\boxed{\mathbf{II}} \quad G_{2,\alpha}(z) = \begin{cases} 0, & z \leq 0; \\ \exp\{-(z)^{-\alpha}\}, & z > 0,\ \alpha > 0. \end{cases} \tag{1.11}$$

$$\boxed{\mathbf{III}} \quad G_{3,\alpha}(z) = \begin{cases} \exp\{-(-z)^{\alpha}\}, & z > 0,\ \alpha > 0?; \\ 1, & z \geq 0. \end{cases} \tag{1.12}$$

[Faire tableau]

(mettre les indice aux fctions G + le shape parameter est "correct" ????)

The II and III can be reformulated (in the ξ-parametrization) as.....

$$G_\xi(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]_+^{-\xi^{-1}}\right\}, \qquad \xi \neq 0, \tag{1.13}$$

where we added explicitly the location and scale parameters μ and σ in order to obtain the three *extreme value distributions*, see for example (Reiss and Thomas, 2007, pp.16) among others. The parameters are such that $\sigma > 0$ and $-\infty < \mu, \xi < \infty$. This statement will hold for the rest of this thesis

By simply coming back in the ξ-parametrization by using $\xi = \alpha^{-1}$ in the above distribution functions, all these three classes of extreme distributions can be expressed in the same functional form as special cases of this single three-parameter (Actually, there are just location and scale

parameters in the type **I** extremal model in (1.10) as $\xi \to 0$) distribution (1.6). That is, when $\xi \to 0$ we retrieve the **type I** or *Gumbel* family (1.10) while $\xi > 0$ and $\xi < 0$ leads to the **type II** or *Fréchet* family and to the **type III** or *Weibull* family, see (1.11) and (1.12) respectively. Both the Gumbel and Fréchet limiting distributions are unbounded (In fact, the Fréchet distribution has a finite left endpoint in $\mu - \sigma\xi^{-1}$, but this has no really interest here); that is, the upper endpoint tends to $+\infty$ while the Weibull distribution has a finite right endpoint in $\mu - \sigma\xi^{-1}$. In the following, we define the left and the right endpoint of a particular df F, respectively ${}_*x$ and x_*, by :

$$ {}_*x = \inf\{x : F(x) > 0\}, \qquad \text{and} \qquad x_* = \sup\{x : F(x) < 1\}. $$

Density We give a representation of the density of these functions by considering the density of the GEV distribution (1.6), that is $g_\xi(z) = \frac{dG_\xi(z)}{dz}$ (we can assume absolute continuity). This is shown in in figure (**??**) for various shape parameters. We also give in appendix. *Or: As the relation of the (density) distribution with the location and scale parameters is trivial, we do not illustrate that here. However, for the shape parameter it is a bit more subtil... and we let the figure*

For the case $\xi \neq 0$,

$$ g_\xi(z) = \sigma^{-1}\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}-1} \exp\left\{ -\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]_+^{-\xi^{-1}}\right\}, \tag{1.14} $$

[table + two case for $Xi = 0$ and $xi \neq 0$] [appendix graphs for various location/scale parameter values]

For the case $\xi = 0$, we have

$$ g_0(z) = \sigma^{-1} \exp\left\{ -\left(\frac{z-\mu}{\sigma}\right)\right\} \exp\left\{ -\exp\left[-\left(\frac{z-\mu}{\sigma}\right)\right]\right\}, \tag{1.15} $$

Note that the support varies equally as for the distribution functions wrt sign of ξ

?? dernier graphe weibull

In some ways, some people will feel this was unfortunate, because now it is common for people to model and fit the GEV without thinking very clearly about the specific form of their data and distributions [Extremes, distribution, etc..] That is the reason why we think it can be useful to explain in the following some examples of how we can construct such extreme distributions for the three classes in concrete cases (see next section), playing with the appropriate choice of sequences a_n and b_n to retrieve the pertaining distribution family.

1.4 Concrete applications : examples of convergence to GEV

This is well not easy to find the sequences in practice. http://stats.stackexchange.com/questions/105745/extreme-value-theory-show-normal-to-gumbel/105749#105749

1.4.1 Convergence to Gumbel distribution

Type I or **Gumbel** distribution $G_1(x)$ can be retrieved by considering, for example, a iid exponential distributed sequence $\{X_j\}$ of n random variables, that is $X_j \overset{iid}{\sim} Exp(\lambda)$ and consider

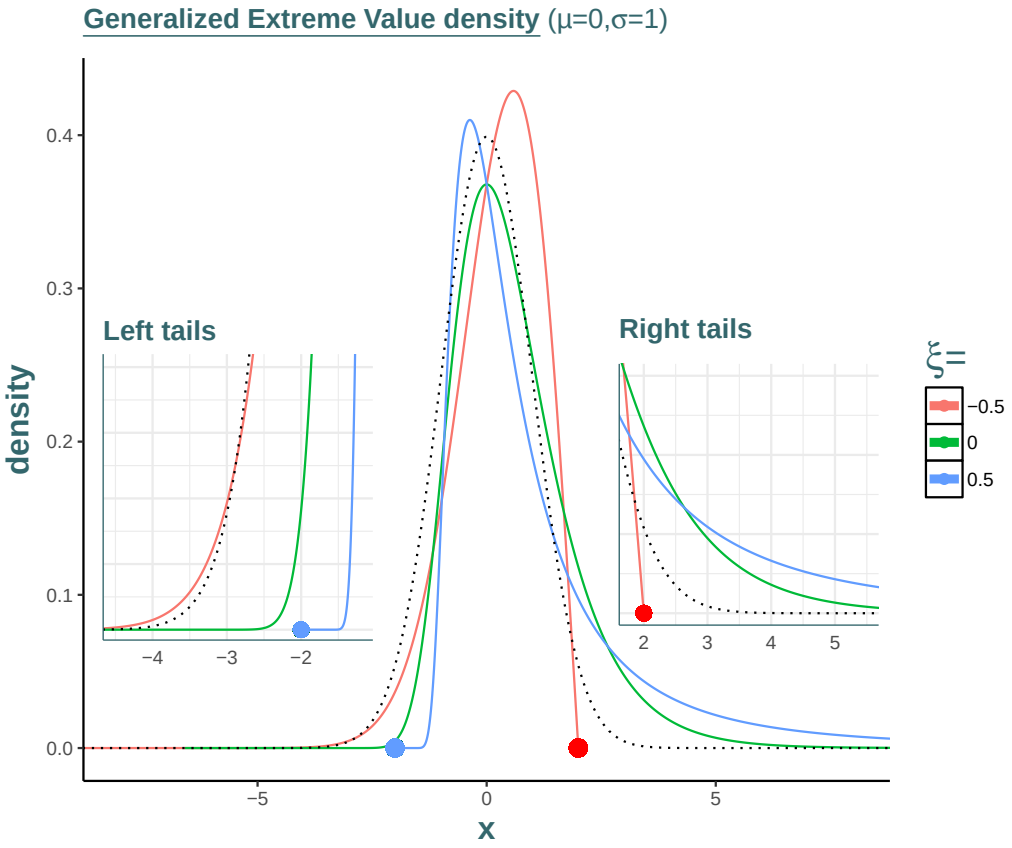

Figure 1.1: GEV distribution with the normal in dotted lines and a zoom on the part of interest, the tails ...

the largest of these values $X_{(n)}$ as defined earlier. By definition, we know $F(x) = 1 - \exp^{-x}$. Our goal is to find non-random sequences $\{b_n\}$, $\{a_n > 0\}$ such that

$$\lim_{n\to\infty} \Pr\{a_n^{-1}(X_{(n)} - b_n) \leq z\} = G_1(z). \tag{1.16}$$

We can easily find that

$$\begin{aligned}\Pr\{a_n^{-1}(X_{(n)} - b_n) \leq z\} &= \Pr\{X_{(n)} \leq b_n + a_n z\} \\ &= \Big[\Pr\{X_1 \leq b_n + a_n z\}\Big]^n \\ &= \Big[1 - \exp\{-\lambda(b_n + a_n z)\}\Big]^n,\end{aligned}$$

from the iid assumption of the random variables and their exponential distribution. Hence, by choosing the sequences $a_n = \lambda^{-1}\log n$ and $b_n = \lambda^{-1}$ and reminding that

$$\Big[1 - \exp\{-\lambda(b_n + a_n z)\}\Big]^n = \Big[1 - \frac{1}{n}e^{-z}\Big]^n \overset{n\to\infty}{\longrightarrow} \exp(-e^{-z}) := G_1(z).$$

Recall: $\boxed{\lim_{n\to\infty}\Big(1 + \frac{x}{n}\Big)^n = \exp(x)}$

intro Falk et al. (2011) somewhR

we find the the so-called standard *Gumbel* distribution in the limit.

We can show the same with iid standard normal random variables, $X_j \overset{iid}{\sim} N(0,1)$, with sequences $a_n = -\Phi^{-1}(1/n)$ and $b_n = 1/a_n$. (see appendix [extremes, distributions pdf])

Typically, unbounded distributions like the Exponential and Normal (as well as the Gamma, Lognormal, Weibull, etc.) whose tails fall off exponentially or faster will have this same Gumbel limiting distribution for the maxima, and will have medians (and other quantiles) that grow as $n \to \infty$ at the rate of (some power of) $\log n$. This is typical example of light-tailed distribution (i.e., decays exponentially as defined in section **1.1**).

1.4.2 Convergence to Fréchet distribution

Type II or **Fréchet-Pareto type** distrbution $G_2(x)$

When starting with a sequence $\{X_j\}$ of n iid random variables (block-maxima?) following a *basic*(or *generalized*, with scale parameter set to 1) Pareto distribution with shape parameter $\alpha \in (0, \infty)$, $X_j \sim Pa(\alpha)$, we have that

$$F(x) = 1 - x^{-\alpha}, \qquad x \in [1, \infty), \tag{1.17}$$

so that we can write, by choosing appropriately $b_n = 0$

$$\begin{aligned}-n\bar{F}(a_n z + b_n) &= -n(a_n z + b_n)^{-\alpha} \\ &= \Big[Q(1 - \frac{1}{n})\Big]^{\alpha}(a_n)^{-\alpha}(-z^{-\alpha}),\end{aligned}$$

where $Q(1 - \frac{1}{n})$ is the quantile function (see). Hence, it is easy to see that by setting the constant $a_n = Q(1 - 1/n)$ and keeping $b_n = 0$, we have that

$$\Pr\{a_n^{-1}X_{(n)} \leq z\} \to \exp(-z^{-\alpha}),$$

showing that for this particular values of the normalizing constants, we retrieve the Fréchet distribution in the limit from a strict Pareto distribution. The fact that b_n is set to zero can be understood intuitively since for heavy-tailed distribution (see) such as the Pareto distribution, a correction for location is not necessary to obtain non-degenerate limit distribution. (Beirlant et al., 1996, pp.51)

[see p.28 memoire other si ft autre exemple]

More generally, we can state the more general following theorem :

Theorem 1.3 (Pareto-type distributions)**.** *?? For the same choice of normalizing constants as above, that is* $a_n = Q(1 - \frac{1}{n})$ *and* $b_n = 0$ *and for any* $x \in \mathbb{R}$*, if*

$$n[1 - F(a_n x)] = \frac{1 - F(a_n x)}{1 - F(a_n)} \to x^{-\alpha}, \qquad n \to \infty \tag{1.18}$$

then we obtain the Fréchet distribution in the limit, or written formally "$\bar{F}$ is of Pareto-type*" or, more technically, "*$\bar{F}$ is regularly varying with index -α*".*

We let the concepts of **regularly varying functions**, together with **slowly varying functions** be defined in appendix A.1 with some useful theorems and properties, according to Beirlant et al. (1996, pp.51-54) and supported by Beirlant et al. (2006, pp.49, 77-82).

Beirlant et al. (2006, pp.75) !!!!

1.4.3 Convergence to Weibull distribution

Type III or **Weibull** family (?) of distributions $G_3(x)$ are, for example, in the limit of n iid uniform random variables $X_j \sim U[L, R]$ where L and $R > L$ are both in $\mathbb{R}$ and denote respectively the Left and the Right endpoint of the domain of definition. We have by definition

$$F(x) = \frac{x}{R - L}, \qquad x \in [L, R].$$

It is $= 0$ for $x < L$, $= 1$ for $x > R$. Assuming the general case ($[L, R]$ can be $\neq [0, 1]$), we have for the maximum $X_{(n)}$:

$$\begin{aligned} \Pr\{a_n^{-1}(X_{(n)} - b_n) \leq z\} &= \Pr\{X_{(n)} \leq b_n + a_n z\} \\ &= \Big[1 - \frac{R - b_n - a_n z}{R - L}\Big]^n, \quad \text{if} \quad L \leq b_n + a_n z \leq Rn, \\ &= (1 + \frac{z}{n})^n \to e^z, \qquad\quad \text{if} \quad z \leq 0 \text{ and } n > |z|. \end{aligned}$$

When choosing $a_n = R$ and $b_n = (R - L)/n$, we find the unit Reversed Weibull distribution $We(1, 1)$ in the limit as expected.

However, for inferential purpose, this is not of particular interest, because from the expression in 1.5, we know that

$$a_n^{-1}(X_{(n)} - b_n) \xrightarrow{d} G_{\xi,\mu,\sigma}(z), \qquad \text{as } n \to \infty. \tag{1.19}$$

After some algebra, this leads to

$$X_{(n)} \xrightarrow{d} G_{\xi,\mu^*,\sigma^*}(z), \qquad \text{as } n \to \infty, \tag{1.20}$$

with the sequences a_n and b_n being absorbed into the new location and scale parameters μ^* and σ^*. We can then ignore the normalizing constants in practical applications and fit directly the GEV in our set of maxima $X_{(n),k}$. The pertaining estimated parameters will implicitly take the normalization into account, i.e. it will estimate μ^* and σ^*. As also stated in.., the shape parameter is invariant.

But what about the fact that $X_{(n)}$ non-normalized is degenerate (see intro) ?

1.4.4 Some Conditions/comments (?) (Continuity condition)

As we have noticed in (?), F needs certain conditions at its right endpoint x_* for the limit to be convergent in (1.1). The *continuity condition* ensures that a discrete distribution cannot have a non-degenrate limit distribution as in (1.1). Examples are well documented by **?**[pp.118-119] for the case of poisson, Geometric and negative binomial distribution. Thus, we cannot have a limit for these distributions.

1.5 Maximum domain of attraction

The preceding results can be more easily summarized and obtained when considering *maximum domain of attraction* (MDA). The term "*maximum*" is typically used to distinguish from *sum-stable* distribution. As we study here only the maxima, there are no confusion possible in our work. We will then preferably write only *domain of attraction* in the following for convenience, and thus consider these two names as synonyms.

Definition 1.5 (Domain of attraction)**.** *We say that a distribution F is in the (***maximum***) **domain of attraction** of an extreme value family G_k (see (1.10)-(1.12)), denoted by $F \in D(G_k)$, if there exist $a_n > 0$ and $b_n \in \mathbb{R}$ such that the distribution of $a_n^{-1}(X_{(n)} - b_n)$ converges weakly (see **??**) to G_k where $X_{(n)}$ is as defined earlier with distribution F.*

The definition is well-defined in the sense that $F \in D(G_i)$ and $F \in D(G_j)$ implies $\xi_i = \xi_j$, writing by ξ_k the extreme value index pertaining to the extreme value distribution G_k.

Before going further with the characterization of the three domains of attraction of our purpose, we think important to introduce a new theorem from Gnedenko **?**

Theorem 1.4 (Convergence to Types Theorem)**.** *Let F_n be a sequence of random variables converging weakly (see appendix A.1.1) to F. Let $a_n > 0$ and $b_n \in \mathbb{R}$ such that $a_n F_n + b_n \Rightarrow F'$, where both F and F' are non-degenerate (see **??**). Then,*

$$a_n \to a \quad \text{and} \quad b_n \to b, \qquad a > 0 \quad \text{and} \quad b \in \mathbb{R}.$$

Equivalently, if G_n, G, G' are distribution functions with G, G' being non-degenerate, and there exists $a_n, a'_n > 0$ and $b_n, b'_n \in \mathbb{R}$ such that

$$G_n(a_n x + b_n) \xrightarrow{d} G(x) \quad \text{and} \quad G_n(a'_n x + b'_n) \xrightarrow{d} G'(x),$$

at all continuity points of F, respectively F', then there exists constants $A > 0$ and $B \in \mathbb{R}$ such that

$$\lim_{n\to\infty} \frac{a_n}{a'_n} \to A, \qquad \lim_{n\to\infty} \frac{(b_n - b'_n)}{a'_n} \to B,$$

and $G'(Ax + B) = G(x)\ \forall x \in \mathbb{R}$.

We have now all the necessary tools to the pertaining domains of attractions. But, before proceeding, we would like to point out that the fact that the characterization of the first domain of attraction (Gumbel class) is much more complex than the two following (Fréchet and Weibull class) and requires much more technicalities going beyond the scope of this thesis. Moreover, despite this class is important in theory, it is less relevant for our purpose of modelling extremes. It often requires other generalizations, for instance with additional parameters to surpass the issues of fitting empirical data. Pinheiro and Ferrari (2015) In the last paragraph, we will present the unified framework, the domain of attraction pertaining to the GEV distributions, which is a kind of summary for the three first domains of attraction presented.
In each of the characterization of the domains of attractions, we will present some of their most useful, necessary and sufficient conditions ... together with their *von Mises conditions*, initially from but revisited in **?**. These conditions are very important in practice and sometimes more intuitive because they make use of the *hazard function*, defined by, for sufficiently smooth distributions :

$$r(x) = \frac{f(x)}{\bar{F}(x)} = \frac{f(x)}{1 - F(x)}. \tag{1.21}$$

It involves the density function $f(x) = \frac{dF(x)}{d(x)}$ in the numerator.**?**[We can conversely define the *reciprocal hazard function* simply by $\tilde{r}(x) = 1/r(x)$]. This function will be useful to characterize insightful conditions for each domains of attraction, known as the *von Mises criterion*.

1.5.1 Domain of attraction for Gumbel distribution (G_1)

We derive here two ways of formulating necessary and sufficient condition for a distribution function F to be in the domain of attraction of G_1, namely $F \in D(G_1)$.

- From (mettre vrmt??) (Haan and Ferreira, 2006, pp.20), for finite or infinite right endpoint x_* with $\int_x^{x_*} \int_t^{x_*} \bar{F}(s)dsdt < \infty$, the function

$$h(x) := \frac{(\bar{F}(x)) \int_x^{x_*} \int_t^{x_*} (\bar{F}(s))\ ds\ dt}{\left(\int_x^{x_*} (\bar{F}(s))ds \right)^2},$$

  must satisfy $\lim_{t\uparrow x_*} h(t) = 1$. [Reminder: $\lim_{t\uparrow y}(\cdot)$ means that t is approaching y from below, i.e. from values smaller than y in a increasing manner, and vice-versa for $\lim_{t\downarrow y}(\cdot)$].

- From (Beirlant et al., 2006, pp.72), for some auxiliary function b, for every $v > 0$, the condition

$$\frac{\bar{F}(x + b(x)v)}{\bar{F}(x)} \to e^{-v}, \tag{1.22}$$

must hold as $x \to x_*$. Then,

$$\frac{b(x + vb(x))}{b(x)} \to 1.$$

voir lien avec la GPD!! ecrire en hazard rate?

A lot of more precise characterizations and conditions together with proofs can be found, for example in (Haan and Ferreira, 2006, pp.20-33). We can also mention a condition that is based on the von Mises function.

However, we present his ***von Mises criterion*** as in (Beirlant et al., 2006, pp.73):

If the *hazard function* $r(x)$ (1.21) is ultimately positive in the neighbourhood of x_*, is differentiable there and satisfies

$$\lim_{x \uparrow x_*} \frac{dr(x)}{dx} = 0, \tag{1.23}$$

then $F \in D(G_1)$. (compare hazar convergence rates of the three types !!!!!)

Examples of distributions in $D(G_1)$ Intuitively, we can remark that all distributions which are exponentially decaying will have this propensity to be in the Gumbel domain of attraction. For instance, the *Exponential*, the *Gamma*, the *Weibull*, the *logistic*, etc. To see that, by a Taylor expansion, we have that

$$\bar{G}_1(x) = 1 - \exp(-e^{-x}) \sim e^{-x}, \qquad x \to \infty.$$

Hence, the Gumbel domain of attractions G_1 decays exponentially (as tend their pertaining distributions).

1.5.2 Domain of attraction for Fréchet distribution ($G_{2,\alpha}$)

Let $\alpha := \xi^{-1} > 0$ be the *index* of the Fréchet distribution $G_{2,\alpha}$ (see (1.11)). Then, $F \in G_{2,\alpha}$ if and only if

PAS F !!!!!!!!

$$\bar{F}(x) = x^{-\alpha} L(x), \tag{1.24}$$

for some slowly varying function L. See for example theorem **2.4** (?). In this case and with $b_n = 0$,

$$F^n(a_n x) \to G_2(x), \qquad x \in \mathbb{R},$$

with

$$a_n := F^{\leftarrow}(1 - \frac{1}{n}) = \Big(\frac{1}{1 - F}\Big)^{\leftarrow}(n),$$

where we define the quantity $F^{\leftarrow}(t) = \inf\{x \in \mathbb{R} : F(x) \geq t\}$ for $t < 0 < 1$ as the *generalized inverse of* F with which we can retrieve $x_t = F^{\leftarrow}(t)$, the t-quantile of F. Even if we deal in this text only with continuous and strictly increasing distribution functions (?), we think it

is more reliable(?) to consider generalized inverse instead of the ordinary inverse, for sake of generalization.

This previous theorem informs us that all distribution functions $F \in D(G_{2,\alpha})$ have necessarily an infinite right endpoint, that is $x_* = \sup\{x : F(x) < 1\} = \infty$. These distributions are all with regularly varying right-tail with index $-\alpha$. In short,

$$F \in D(G_{2,\alpha}) \Longleftrightarrow \bar{F} \in R_{-\alpha}.$$

Finally, we must also present the (revisited) ***Von Mises condition*** for this domain of attraction which state the following in Falk and Marohn (1993) : if F is absolutely continuous with density f and right endpoint $x_* = \infty$, such that

$$\lim_{x \uparrow \infty} x \; r(x) = \alpha > 0,$$

where $r(x)$ is the *hazard function* defined in (1.21), then $F \in D(G_{2,\alpha})$. In words, it means that... We illustrate this with the standard Pareto distribution case (as previsouly in), that is

$$F(x) = \left(1 - (\frac{x_m}{x})^{\alpha}\right) 1_{x \geq x_m}, \qquad \alpha > 0 \;\text{ and }\; x_m > 0.$$

Clearly, we can see that by setting $K = x_m^{\alpha}$, we have

$$\bar{F}(x) = K x^{-\alpha}.$$

Therefore, we have that $a_n = (Kn)^{\alpha^{-1}}$ and $b_n = 0$.

Examples of distributions in $D(G_{2,\alpha})$ These distributions are typically very-fat tailed (and hence, heavy-tailed, see) distributions, such that $E(X_+)^{\delta} = \infty$ for $\delta > \alpha$. This class of distributions is appropriate for phenomena with extremely large maxima (like...). **?** Common distributions include Pareto, Cauchy, Burr, stable distributions with $\alpha < 2$, etc. An example to see that, is again by Taylor expansion at the tail of $G_{2,\alpha}$ with $\alpha > 0$

$$\bar{G}_{2,\alpha}(x) = 1 - \exp(-x^{-\alpha}) \sim x^{-\alpha}, \qquad x \to \infty, \tag{1.25}$$

showing that $G_{2,\alpha}$ tends to decrease as a *power law*. See for example eq.

1.5.3 Domain of attraction for Weibull distribution ($G_{3,\alpha}$)

We say that $F \in G_{3,\alpha}$ (1.12) with index $\alpha > 0$ if and only if there exists finite right endpoint $x_* \in \mathbb{R}$ such that

$$\bar{F}(x_* - x^{-1}) = x^{-\alpha} L(x), \tag{1.26}$$

where $L(\,\cdot\,)$ is a slowly varying function (see).

For $F \in D(G_{3,\alpha})$, we have also

$$a_n = x_* - F^{\leftarrow}(1 - n^{-1}), \qquad b_n = x_*.$$

Hence

$$a_n^{-1}(X_{(n)} - b_n) \xrightarrow{d} G_{3,\alpha}.$$

[see references [domain of attraction course]]

Finally, we still present the ***Von Mises condition*** from Falk and Marohn (1993) related to the $G_{3,\alpha}$) domain of attraction. It states that for F having positive derivative on some $[x_0, x_*)$, with finite right endpoint $x_* < \infty$, then $F \in D(G_{3,\alpha})$ if

$$\lim_{x \uparrow x_*} (x_* - x) r(x) = \alpha > 0, \qquad \int_{-\infty}^{x_*} \bar{F}(u) du < \infty, \tag{1.27}$$

where $r(x)$ is again the *hazard function* defined in (1.21).

Examples of distributions in $D(G_{3,\alpha})$ Weibull's domain of attraction thus includes all the distribution functions that are bounded to the right ($x_* < \infty$). As most phenomena are typically bounded, we will think as the Weibull for the most attractive and flexible class for modelling extremes. But, in practice, the Fréchet one is often more preferable in an extreme analysis context because allowing for arbitrarily large values.

[put general case pp.73-75 beirlant] ?

1.5.4 Closeness under tail equivalence property

An interesting property of all the three types of domain of attraction $D(G_{k,\alpha})_{k=1,2,3}$ we have derived, is that those are *closed under tail-equivalence* **??**. This is useful for characterizing tail's types of the distributions falling in the domains of attraction. the It this sense,

1. For the **Gumbel** domain of attraction, let $F \in D(G_{1,\alpha})$. If H is another distribution function such that, for some $b > 0$,

$$\lim_{x \uparrow x_*} \frac{\bar{F}(x)}{\bar{H}(x)} = e^b, \tag{1.28}$$

   then $H \in D(G_{1,\alpha})$. This emphasizes the exponential type of the tails for the distributions H falling in the Gumbel domain of attraction.

2. For the **Fréchet** domain of attraction, let $F \in D(G_{2,\alpha})$. If H is another distribution function such that, for some $c > 0$,

$$\lim_{x \to \infty} \frac{\bar{F}(x)}{\bar{H}(x)} = c^\alpha, \tag{1.29}$$

   then $H \in D(G_{2,\alpha})$.

3. For the **Weibull** domain of attraction, let $F \in D(G_{3,\alpha})$. If H is another distribution function such that, for some $c > 0$,

$$\lim_{x\uparrow x_*} \frac{\bar{F}(x)}{\bar{H}(x)} = c^{-\alpha}, \tag{1.30}$$

then $H \in D(G_{3,\alpha})$.

This emphasizes the polynomial types for the tails of the distributions falling in the Fréchet or in the Weibull domain of attraction.

For more informations about the characterizations of the, one can refer to ....

1.5.5 Domain of attraction of the GEV

[Coles slides 30] (and see stat extremes beirlant!!) The conditions that have been stated the three preceding domains of attraction can be restated under this "unified" framework for the GEV distribution defined in (1.6) For a given distribution function F, by letting the sequences b_n, a_n, and the shape parameter such that[3]

$$b_n = F^{\leftarrow}(1-1/n), \qquad a_n = r(b_n) \qquad \text{and} \qquad \xi = \lim_{n\to\infty} \tilde{r}(x),$$

with $r(\cdot)$ the *hazard* function defined (in 1.21). Hence, $a_n^{-1}(X_{(n)} - b_n)$ has limiting distribution

$$\begin{cases} \exp\Big\{ -\big[1+\xi\sigma^{-1}(x-\mu)\big]_+^{-\xi^{-1}}\Big\}, & \xi \neq 0; \\ \exp\big\{-e^{-x}\big\}, & \xi = 0, \end{cases}$$

which is the GEV (see eq1.10-1.13) [see if put location an scale parameters]

Among a lot of characterizations available, we present the most []:

Theorem 1.5. *If there exist a positive, measurable function $u(\cdot)$, then for $-\infty < \xi < \infty$, $F \in D(GEV)$ if and only if :*

$$\lim_{v\uparrow x_*} Pr\left\{ \frac{X-v}{u(v)} > x \mid X > v \right\} := \lim_{v\uparrow x_*} \frac{\bar{F}(v+xu(v))}{\bar{F}(v)} = \begin{cases} \big(1+\xi x\big)_+^{-\xi^{-1}}, & \xi \neq 0; \\ e^{-x}, & \xi = 0. \end{cases} \tag{1.31}$$

1.6 The Skewed Generalized Extreme Value Distribution

?

1.7 The Concepts of Return Levels and Return Periods

Return levels play a major role in environmental analysis. For such tasks, it is usually more convenient to interpret EV models in terms of insightful return levels rather than individual parameter estimates, following Coles (2001, pp.49,pp.81).

[3] We think important to recall again the reader the difference of parametrization $\xi = \alpha^{-1}$

Assuming for this introductory example our time unit reference is in year -as usually assumed in meteorological analysis-, let us consider the *m-year return level* r_m and define it, (at first sight), as the high quantile for which the probability that the annual maximum exceeds this quantile is $1/\lambda m$, where λ is the mean number event will be obviously equal to 1 here for yearly blocks. m is called the *return period* and is the expected time between the occurrence of two so-defined high-quantiles. Let $\{X_{(n),y}\}$ denote the iid sequence of n random variables representing the annual maximum for yeay y. From (1.6), we then have (check the implication)

$$\begin{aligned} F(r_m) = \Pr\{X_{(n),y} \leq r_m\} &= 1 - 1/m \\ \Leftrightarrow \quad \left[1 + \xi\left(\frac{r_m - \mu}{\sigma}\right)\right]^{-\xi^{-1}} &= \frac{1}{m}. \end{aligned}$$

Hence, by inverting this relation, and letting $y_m = -\log(1 - m^{-1})$, we can get the quantile of the GEV, namely the *return level*

$$r_m = \begin{cases} \mu + \sigma\xi^{-1}(y_m^{\xi} - 1), & \xi \neq 0; \\ \mu + \sigma \log(y_m), & \xi = 0. \end{cases} \tag{1.32}$$

(See for eq. index ())

Hence, we can directly retrieve it from our estimation of the three GEV parameters.

However, we recall that the definition of return period is easily misinterpreted and the given above is thus not universally accepted. To evaporate (vanish) this issue, it is important to distinguish stationary from non-stationary sequences.

Explore why the return leels go beyond the right endpoint of the distribution (when $\xi < 0$ as here), for which return period, etC...

1.7.1 Return Level Plot

Standard errors of the estimates As usual, the standard errors of these estimates are important to compute, for example to construct confidence intervals (but they can be quite misleading!), and hence the return level plot. We naturally expect these to increase with the return period. As r_m is a function of the GEV parameters, we can use the *delta method* (see..)to approximate the variance of $\hat{r}_m$. Specifically,

$$\mathrm{Var}(\hat{r}_m) \approx \nabla r_m^{'} V \nabla r_m,$$

with V the variance-covariance matrix of the estimated parameters $(\hat{\mu}, \hat{\sigma}, \hat{\xi})'$ and

$$\begin{aligned} \nabla r_m^{'} &= \left[\frac{\partial r_m}{\partial \mu}, \frac{\partial r_m}{\partial \sigma}, \frac{\partial r_m}{\partial \xi}\right] \\ &= \left[1,\ \xi^{-1}(y_m^{-\xi} - 1),\ \sigma\xi^{-2}(1 - y_m^{-\xi}) - \sigma\xi^{-1} y_m^{-\xi} \log y_m\right], \end{aligned} \tag{1.33}$$

with $y_m = -\log(1 - m^{-1})$, with the gradient evaluated at the estimates $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$. But a problem arise for the so-computed standard errors when considering long-range return levels. (GRAPH??)

They can increase so drastically with the return period that the confidence intervals of the *return level plot* can become difficult to work with. To try to get rid of this issue with in section **3..** by constructing intervals on the basis of the *profile* log-likelihood.

Interpretation When plotted against the return period on a logarithmic scale, the return levels has different shapes depending on the value of the shape parameter ξ, namely :

- If $\boldsymbol{\xi = 0}$, then return level plot will be **linear**.
- If $\boldsymbol{\xi < 0}$, then return level plot will be **convex**.
- If $\boldsymbol{\xi > 0}$, then return level plot will be **concave**.

1.8 INFERENCE

put here? and piut in section 4 only the "advanced" (improved) methods

Chapter 2

Peaks-Over-Threshold Methods

Contents

Seuils meteo : 0C (gel permanent), 25C et 30C pour les Tx 0C (gel) et 20C pour les Tn

—> Use this for thresholds ?

In this chapter we will focus on the other kind of EV models, modelling excess over a threshold. This... In section **2.1**, we briefly introduce the concepts to be able to more precisely characterize the distributions in section **2.2**. In section 2.3, we introduce the Poisson models,

2.1 Preliminaries: Intuitions

The threshold models relying on the *Peaks-Over-Threshold* (POT) method are useful to propose a better (?) alternative than the blocking method in **2.1**. With this new method, we consider a more natural way of determining whether an observation is extreme or not, by focusing only on all observations that are greater than a pre-specified *threshold*. As we saw, estimates of the GEV parameters are sensitive to the size of block chosen to identify extremes (see) while we will investigate that the estimates of the *Generalized Pareto Distribution* (GPD)[1] parameters are more stable in this sense. Henceforth POT avoids the problem that can arise by considering the maximum of blocks only (), but this method also brings its own problems (). Be aware that this method brings lots of problems with the independence condition... And, especially for temperature data, where for example during heat or cold waves...

Let's consider a sequence $\{X_j\}$ of n iid random variables having marginal distribution function F. We are then regarding for observations that exceed a well-chosen (see) threshold u, which must obviously be smaller than the right endpoint $x_* = \sup\{x : F(x) < 1\}$ of F. The aim here is to find a "child" probability distribution function (fig.? -video youtube), say H, from the underlying (parent) distribution F, that will allow us to model the exceedance $Y = X - u$, and with H then expressed as $\boxed{H(y) = \Pr\{X - u \leq y \mid X > u\}}$. Typically, threshold models can therefore be regarded as the conditional survival function of the exceedances Y, knowing that the threshold u is exceeded (Beirlant et al., 2006, pp.147) :

$$\Pr\{Y > y \mid Y > 0\} = \Pr\{X - u > y \mid X > u\} = \frac{\bar{F}(u+y)}{\bar{F}(u)}. \tag{2.1}$$

or in terms of the exceedance distribution function $F^{[u]}(x) = \Pr\{X \leq u + x | X > u\}$ (Reiss and Thomas, 2007, pp.12), **?** and Rosso (2015) :

$$F^{[u]}(x) = \frac{\Pr\{X - u \leq x, X > u\}}{\Pr\{X > u\}} = \frac{F(x+u) - F(u)}{\bar{F}(u)}, \tag{2.2}$$

making use of the well-known conditional probability law. One can remark that (2.1) is actually the survivor of the exceedance distribution function, that is $\bar{F}^{[u]}$.

These intuitive characterizations we have given above about the modelling of the threshold exceedances in term of probability distribution function can be useful to understand the following.

However, if the parent distribution F were known, we would be able to compute the distribution of the threshold exceedances in (2.1). (Coles, 2001, pp.74) But as for the GEV in the method of block-maxima (section **2.1**), the distribution F is not known in practice, as we will see also in (..). Hence, and as usual in statistics[2], we must again rely on approximations. This time, we will try to approximate (2.2)

[1] Notice that, as an abuse of language and for smoother readability, we will use the abbreviations "GPD" to denote both the Distribution and the Distribution *function*

[2] ?

2.2 Characterization of the Generalized Pareto Distribution

Anageously to the *Fisher-Tippett* theorem in section **2.1** which applies for the block maxima, we have now to define a new theorem which applies for values above a predefined threshold. From this result 1.6(?), these two theorems form together the basis of Extreme Value Theory.

Theorem 2.1 (POT-stability)**.** Reiss and Thomas (2007, pp.25) *The max-stability theorem in* **??** *can be applied and are formulated here by the fact that the GP distribution functions H are the only continuous one such that, for certain choice of constants a_u and b_u,*

$$F^{[u]}(a_u x + b_u) = F(x).$$

This will be useful for modelling the exceedances in the following theorem (?). And for the examples (see ex. p.25)

Theorem 2.2 (Pickands–Balkema–de Haan)**.** *discovered by* Balkema and de Haan (1974) *and*Iii (1975) *which showed that the distribution of a threshold u of normalized excesses $F^{[u]}(x)(b_u + a_u x)$, as the threshold approaches the right endpoint x_* of F, is the Generalized Pareto Distribution (**GPD**) $H_{\xi,\sigma_u}(y)$. That is, if X is a random variable for which (1.5) holds, and for the approximating GP distribution function possessing the same left endpoint u as the exceedance distribution function $F^{[u]}$, we have* Reiss and Thomas (2007, pp.27)*:*

$$|F^{[u]}(x) - H_{\xi,\sigma_u}(x)| \longrightarrow 0, \qquad u \to x_*.$$

Or, in an other, maybe more intuitive formulation (the same : delete) Coles (2001) :

$$\Pr\{X(-u) \leq y \mid X > u\} \longrightarrow H_{\xi,\sigma_u}(y), \qquad u \to x_*, \tag{2.3}$$

where the **GPD** is defined as :

$$H_{\xi,\sigma_u}(y) = \begin{cases} 1 - \left(1 + \frac{\xi y}{\sigma_u}\right)_+^{-\xi^{-1}}, & \xi \neq 0; \\ 1 - \exp\left\{-\frac{y}{\sigma_u}\right\}, & \xi = 0. \end{cases} \tag{2.4}$$

We recall again that $y = x - u > 0$, and where the scale parameter is denoted σ_u to emphasize its dependency with the chosen threshold u :

$$\sigma_u = \sigma + \xi(u - \mu), \tag{2.5}$$

where one can also remark that the location parameter μ does not appear anymore in (**??**) as it does appear in 2.9.

2.2.1 Outline proof of the GPD and justification from GEV

As we did for block-maxima approach in section **2.1.1** (1.2-1.2), we think it is interesting to have a formal and comprehensive, and still not too technical, intuitive view of where are the GPD from. We remind that we aim here at retrieving the GPD $H_{\xi,\sigma_u}(y)$ (2.3-**??**) from probability distributions as expressed in (2.1-2.2).

Proof :

- We start with X having distribution function F. From the GEV theorem in section **2.1.** (see 1.5-1.6), we have for the largest order statistic, for large enough n,

$$F_{X_{(n)}}(z) = F^n(z) \approx \exp\left\{ -\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\xi^{-1}}\right\}, \tag{2.6}$$

  with $\mu, \sigma > 0$ and ξ the GEV parameters. hence, by simply taking logarithm on both sides, we have

$$n \ln F(z) \approx -\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\xi^{-1}}. \tag{2.7}$$

- We also have that, from Taylor expansion, $\ln F(z) \approx -[1 - F(z)]$ as both sides go to zero when $z \to \infty$. Therefore, substituting into (2.7), we get the following for large u :

$$1 - F(u + \boldsymbol{y}) \approx n^{-1}\left[1 + \xi\left(\frac{u + \boldsymbol{y} - \mu}{\sigma}\right)\right]^{-\xi^{-1}}.$$

  where we specially added the therm $\boldsymbol{y} > 0$ for our purpose of retrieving something in the form of (2.1)-(2.2).

- Finally, we get for (2.1), with some mathematical manipulations, as $u \to x_*$:

$$\begin{aligned}\Pr\{X > u + y \mid X > u\} = \frac{\bar{F}(u+y)}{\bar{F}(u)} &\approx \frac{n^{-1}\left[1 + \xi\sigma^{-1}(u+y-\mu)\right]^{-\xi^{-1}}}{n^{-1}\left[1 + \xi\sigma^{-1}(u-\mu)\right]^{-\xi^{-1}}} \\ &= \left[1 + \frac{\xi\sigma^{-1}(u+y-\mu)}{1 + \xi\sigma^{-1}(u-\mu)}\right]^{-\xi^{-1}} \\ &= \left[1 + \frac{\xi y}{\sigma_u}\right]^{-\xi^{-1}},\end{aligned}$$

  where σ_u is still linear in the threshold u (you will see in (2.9), that is $\sigma_u = \sigma + \xi(u - \mu)$. By simply reverting the probability as in (2.2), we have then

$$\begin{aligned}\Pr\{X - u \le y \mid X > u\} &= 1 - \Pr\{X > u + y \mid X > u\} \\ &= 1 - \left(1 + \frac{\xi y}{\sigma_u}\right)^{-\xi^{-1}},\end{aligned} \tag{2.8}$$

  which is $GPD(\xi, \sigma_u)$ as required and σ_u is as defined in (2.9).

□

More comprehension can come from (Reiss and Thomas, 2007, pp.27-28) or if one wants to analyse rates of convergence.

2.2.2 Dependence of the scale parameter σ

We chose to express the scale parameter as σ_u to emphasize its dependency with the threshold u. If we increase the threshold, say to $u' > u$, then the scale parameter will be adjusted following :

$$\sigma_{u'} = \sigma_u + \xi(u' - u), \tag{2.9}$$

and in particular, this adjusted parameter $\sigma_{u'}$ will increase if $\xi > 0$ and decrease if $\xi < 0$. If $\xi = 0$, there would be no change in the scale parameter[3]. We think important to point out the fact that, similarly as mentioned for the GEV models in (1.6), the scale parameter σ_u for GPD models is not the usual standard deviation, but does govern the "size" of the excesses. (AghaKouchak et al., 2013, pp.20)

We will later discuss the threshold choice in section **3.**

2.2.3 Three different types of GPD and duality with GEV

One will remark the similarity with the GEV distributions as the parameters of the GPD of the threshold excesses are uniquely determined by the corresponding GEV distribution parameters of block-maxima (see outline proof in the above to convince yourself). Hence, the shape parameter ξ of the GPD is equal to that of the corresponding GEV and, most of all, it is invariant[4] while the computation of σ_u will not be affected by changes of the corresponding μ or σ in the GEV, from the self-compensation arising in (2.9). (Coles, 2001, pp.76)

Hence, as for the block-maxima approach, there are also three possible families of the GPD depending on the value of the shape parameter ξ which determines the qualitative behaviour of the GPD. Hosking and Wallis (1987), Singh and Guo (1995)

- The **first** type, call it $H_{0,\sigma_u}(y)$, comes by letting the shape parameter $\xi \to 0$ in **??**, giving :

$$H_{0,\sigma}(y) = 1 - \exp\Big(-\frac{y}{\sigma}\Big), \qquad y > 0. \tag{2.10}$$

  One can easily notice that it corresponds to an **exponential** distribution function, and hence light-tailed, with parameter $1/\sigma_u$, namely $Y \sim \exp(\sigma_u^{-1})$.

- The **second** and the **third** types, that is when $\xi < 0$ and $\xi > 0$ (resp.), differ only by their support :

$$H_{\xi,\sigma_u}(y) = 1 - \left(1 + \frac{\xi y}{\sigma_u}\right)^{-\xi^{-1}}, \qquad \text{for} \quad \begin{cases} y > 0, & \xi > 0; \\ 0 < y < \sigma_u \cdot |\xi|^{-1}, & \xi < 0. \end{cases} \tag{2.11}$$

  Therefore, if $\xi > 0$ the corresponding GPD is of **Pareto**-type, hence is heavy-tailed, and has no upper limit while if $\xi < 0$, the associated GPD has an upper bound $y_* = u + \sigma_u/|\xi|$

[3] This is consistent with the *memoryless property* of the exponential distribution H_{0,σ_u} (**??**), for which we give more details in

[4] For instance, choosing different block size in the GEV modelling would shift its (estimated) parameters while GPD (estimated) parameters are *stable*.

and is then **Beta**-type distribution. A special case arise when $\xi = -1$ where the pertaining distribution becomes Uniform$(0, \sigma)$. (**?**, pp.186)

Some plots ?

After looking at the behaviour of the density of these functions, we will procure a more comprehensive view by defining some examples of how to retrieve these different types of Generalized Pareto Distributions.

Density functions of the GPD

$$h_{\xi,\sigma_u}(y) = \frac{\xi}{\sigma_u}\left(1 + \xi\frac{y}{\sigma_u}\right)^{-\xi^{-1}-1} \tag{2.12}$$

2.2.4 Examples of the GPD as limiting distribution for exceedances

We have seen in the previous paragraph that if we can have an approximate distribution G for block-maxima, then threshold excess will have a corresponding distribution given by a member of the Generalized Pareto family. Whence the shape parameter ξ, as for GEV distributions, is determinant for controlling the behaviour of the GPD, and thus leads to the three different types in (2.10)-(2.11).

1. The first type

The choice of a threshold will be discussed in section **3.5.1**.

From (Beirlant et al., 2006, p.147-),

2.3 Return Levels

In a similar way as for method of block-maxima (see section **1.6**). From (2.4), we obtain the quantiles of the GPD simply by setting this equation equal to $1 - 1/m$ and inverting.

However, differently as for *block*-maxima, the quantiles of the GPD cannot be as readily interpreted as return levels because the observations no longer derive from predetermined *blocks* of equal length. Instead, it is now required to estimate the *probability of exceeding the threshold*, ζ_u.

We can now retrieve the return level r_m, i.e. the **value that is exceeded on average once every m observations**. This value is given by

$$r_m = \begin{cases} u + \sigma_u\ \xi^{-1}\Big[(m\zeta_u)^{\xi} - 1\Big], & \xi \neq 0; \\ u + \sigma_u \log(m\zeta_u), & \xi = 0. \end{cases} \tag{2.13}$$

provided m is sufficiently large.

Interpretation

Whereas the interpretation of the plot in function shape parameter value is the same as for the block-maxima method (see the end of section **1.8**, it is more convenient to replace the value of m by $N \cdot n_y$ in (2.13), where n_y is the number of observations per year, to give return levels on an annual scale. This method allows us to obtain the *N-year return level* which is now commonly defined as the level expected to be exceeded once every N years.

2.4 Point Process Approach

Following mainly Coles (2001), with some further concepts taken from **?** ,

As for the two preceding methods, the point process approach aims at modelling some sequences which are initially assumed to be independent (??)

[coles, pp.124] Here, Point Process could be seen as a kind of summary of the two previous methods (respectively in chapter **1** and in rest of this chapter), leading to nothing new. However, this approach if often preferred :

1. Its interpretation **unifies** the **models** considered so far.
2. Its likelihood enables a more natural formulation of non-stationarity in excess models from the Generalized Pareto model, see section **2.2**.

Furthermore, we will see that the parametrization of the point process model is invariant to threshold choice so that this variation would only affect the well-known (already mentioned) bias-variance trade-off in the inference. Interesting if seasonal modelling.

Hence, we recall that if Y is Poisson distributed with parameter λ, then

$$\Pr[Y = k] = e^{-\lambda}\frac{\lambda^k}{k!}, \qquad k \in \mathbb{N}. \tag{2.14}$$

2.4.1 Non-homogeneous Poisson Process

$$N(A) \sim \text{Poi}(\Lambda(A)), \qquad \Lambda(A) = \int_A \lambda(x)dx. \tag{2.15}$$

"If a process is stationary and satisfies an asymptotic lack of “clustering” condition for values that exceed a high threshold, then its limiting form is non-homogeneous Poisson with intensity measure" Λ, on a set of the form $A = (t_1, t_2) \times (x, \infty)$, given by

$$\Lambda(A) = (t_2 - t_1) \cdot \left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-\xi^{-1}} \tag{2.16}$$

2.5 INFERENCE

put here?

2.6 Standard Threshold Selection (Methods)

2.6.1 Threshold choice for the excess models

Single threshold selection involves a **bias-variance trade-off**. That is, (raccourcir)

- **Lower threshold** will induce **higher bias** due to model misspecification. In other words, the threshold must be sufficiently high to ensure that the asymptotics underlying the GPD approximation are reliable.
- **Higher threshold** will induce higher estimation uncertainty, i.e. **higher variance** of the parameter estimate as the sample size is reduced for high threshold.

(Following Leadbetter et al. (1983), this is practically equivalent to estimation of the k^{th} upper order statistic $X_{(n-k+1)}$ called the "tail fraction" below. To ensure tail convergence, as $n \to \infty$, $k \to \infty$ but at a reduced rate such that $k/n \to 0$, i.e. the quantile level of the threshold increases at a faster rate as the sample size n grows.

)

Based on Mean Residual Life

function or *mean excess function* , following again (Beirlant et al., 2006, pp.14-19), (Coles, 2001, pp.78-80),

$$mrl(u_0) := E(X - u_0 \mid X > u_0) \; = \frac{\int_{u_0}^{x_*} \bar{F}(u)du}{\bar{F}(u_0)}, \tag{2.17}$$

for X having survival function $\bar{F}(u_0)$ computed at u_0, with $x_* = \sup\{x : F(x) < 1\}$ denoting the right endpoint of the support of F. It denotes, in an actuarial context, the expected remaining quantity or amount to be paid out when a level u_0 has been chosen. However, even if it is mainly applied in an actuarial context or in survival analysis in the literature (see **?** for a well-known example), there are also interesting and reliable applications in our more environmental purposes as we will see in the following. Moreover, this function has interesting properties about the tail of the underlying distribution of X (Beirlant et al., 2006, pp.16). In fact, we expect the following :

- If $mrl(u_0)$ is constant, then X has exponential distribution.
- If $mrl(u_0)$ ultimately increases, then X has a heavier tail than the exponential distribution.
- If $mrl(u_0)$ ultimately decreases, then X has a lighter tail than the exponential one.

(and vice-ersa, goes it in the two sens??)

This can be particularly interesting for our purpose when considering threshold models. For this case, we can suppose the excesses of a threshold generated by the sequence $\{X_i\}$ follow a generalized Pareto distribution (see 2.2). Knowing the theoretical mean of this distribution, we retrieve, provided the shape parameter $\xi < 1$ and denoting σ_u the scale parameter corresponding to excess of a threshold $u > u_0$,

$$\begin{aligned} mrl(u) := E(X - u \mid X > u) &= \frac{\sigma_u}{1-\xi} \\ &= \frac{\sigma_{u_0} + \xi u}{1-\xi}, \end{aligned} \tag{2.18}$$

from the threshold u dependence with the scale parameter σ (see 2.9). Hence, we remark that $mrl(u)$ is linearly increasing in u, with gradient $\xi(1-\xi)^{-1}$ and intercept $\sigma_{u_0}(1-\xi)^{-1}$. Furthermore, we can estimate empirically this function intuitively by

$$\widehat{mrl}(u) = \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{[i]} - u), \tag{2.19}$$

where we let the $x_{[i]}$ denoting the (i-th out of the) n_u observations that exceed u.

Mean residual life plot This leads to an interesting tool for our purpose, the *mean residual life plot*. It comes from combining the linearity detected between $mrl(u)$ and u in (2.18) with (2.19). Therefore, a reliable information can be retrieved from the point of the points

$$\left\{ \left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{[i]} - u) \right) : u < x_{max} \right\}. \tag{2.20}$$

Even if its interpretation is not easy, this graphical procedure will give insights for the choice of a suitable threshold u_0 to model extremes via general Pareto distribution, that is the threshold u_0 above which we can detect linearity in the plot. Relying on this well-chosen threshold u_0, the generalized Pareto distribution should be a good approximation. Remind however that its interpretation is often subjective. Furthermore, information in the far right-hand-side of this plot is unreliable. Variability is high due to the limited amount of data (exceedances) above very high thresholds. This can be seen for example on larger confidence intervals.

From (Coles, 2001, pp.83-84)

"Substantial subjectivity in interpreting these diagnostic plots, and the resulting uncertainty. Similar challenges are seen with the River Nidd data, shown in Tancredi et al. (2006), and many other exam- ples in the literature. These examples suggests that a more 'objective' thresh- old estimation approach is needed and that uncertainty must be accounted for."

see mixture pdf]

Based on the stability of the parameter's estimates

see section 4.3.4 of coles.

montrer pr varying threshold si les estimateurs changent bcp ? Stability plots avec IC grisé qui sajuste complement ds cette region.

The aim is to plot MLE's of the parameters gainst the threshold. These MLE's are supposed to be independent of the threshold choice.

From its simplicity, it forms one of the main tools for the practitioners (as said by e.g.).

But this method is also highly criticized, especially for its lack of interpretability, and the pointwise confidence intervals which are strongly dependent across the range of thresholds (here e.g. we took only....).

Other techniques have thus been proposed, see e.g. Wadsworth (2016) which propose complementary plots with greater interpretability, with a "simple" likelihood-based procedure allowing for automated (more formal ?) threshold selection.

In two words, this method .... "To identify a treshold that provides the best fit to the likelihood (8)", we maximize the profile likelihood $L_p(j) = L(\hat{\beta}_j, \hat{\gamma}_j, j)$, with $(\hat{\beta}_j, \hat{\gamma}_j)$ the MLE's for a fixed j. After computing $j^* = \text{argmax}_j\, L_p(j)$, the question is whether $L(\hat{\beta}_{j^*}, \hat{\gamma}_j, j^*$ provides a significantly better fit to $\xi*$ than $L(0,1,0) = \prod_{i=1}^{k-1} \phi(\xi_i^*; 0, 1)$. This can be answered by a likelihood ratio test, with test statistic

$$T = \frac{L(\hat{\beta}_{j^*}, \hat{\gamma}_{j^*}, j^*)}{L(0,1,0)}. \tag{2.21}$$

If this is significant at level α, there is evidence against a hypothesis of white noise and then we select the threshold $u^* = u_{j^*+1}$ which provides the best fit.

"The lowest threshold that one entertains, u1, may also have an impact upon the selected threshold, and might thus be regarded as a tuning parameter. "

"how many thresholds k one should choose. There should be some link to the sample size of the data: if k is too large compared to the sample size n, then the asymptotic theory will not provide a good approximation to the distribution."

Based on the Dispersion Index Plot

As we have seen, the methods considered above lead to a huge amount of subjectivity. Following Ribatet (2006), this method is particularly useful for time series. In section **2.4** we have proven that occurrences of the excesses are represented by a Poisson process, see 2.14. Hence, $\mathbb{E}[X] = \text{Var}[X]$ and the *Dispersion Index* statistic introduced by **?** is defined by $DI = s^2 \cdot \lambda - 1$, where s^2 is the intensity of the Poisson process and λ can be interpreted as the mean number of events in a block.

A confidence interval can also be computed :

Based on *L-Moments* plot

They are linear combinations of the ordered data values. From the GPD, we have

$$\tau_4 = \tau_3 \cdot \frac{1 + 5\tau_3}{5 + \tau_3}, \tag{2.22}$$

where τ_4 is the *L-Kurtosis* and τ_3 is the *L-Skewness*. See e.g. Hosking and Wallis (1997) for more details on L-moments or Peel et al. (2001) for a known application of this method in hydrology.

We can then construct the *L-Moment plot* :

$$\left\{(\hat{\tau}_{3,u}, \hat{\tau}_{4,u}) : u \leq x_{\max}\right\} \tag{2.23}$$

where $\hat{\tau}_{3,u}$ and $\hat{\tau}_{4,u}$ are estimations of L-kurtosisand L-skewness based on u and $x_{\max}$ is the maximum observation.

2.7 "Varying" Threshold : Mixture Models

Dey and Yan (2016)

see application with pdf small thesis !!! –> unconclusive.

The threshold is either implicitly or explicitly defined as a parameter to be automatically estimated, and in most cases the un- certainty associated with the threshold choice can be accounted for naturally in the inferences.

The so-called "*fixed threshold approach*" (named in **?**, among others) which include thus the diagnostics discussed in section

There is a wide literature on the subject. The model can be presented in a general way :

$$f(x) = (1 - \phi_u) \cdot b_t(x) + \phi_u \cdot g(x), \tag{2.24}$$

with $b_t(x)$ the density of the bulk model, and where we ignored the parameter dependence for clarity.

"A guiding principle in developing, or choosing, extreme value mixture models is to combine a suitable bulk model, or at least a flexible bulk model, with the tail model. If this is successfully achieved then these models and inference schemes can provide an automated and objective approach to threshold and tail estimation, including uncer- tainty quantification." book risk pp.62

Problem is the discontinuity which (can) occur in the pdf (not the case for the cdf). this can present bias and uncertainty when the quantity of interest considered is close to the threshold. "Nonstationary extensions of such models can be particularly problematic with the extent of discontinuity varying along the threshold function."

Possible alternaties are possible to force continuity on the pdf.

"If the bulk model is correctly specified, then the parametric mixture models are easy to understand and quick to fit so are preferred. However, in more usual situation of unknown population distribution, the nonparametric mixture models perform consistently well for low and high quantiles." evmix package thesis.

Nonstationary extremes

see thesis2012 p.155

Cross-validation **?** Besides all these methods that are very subjective,...

or see gelman bayesian book pp.169

Chapter 3

Relaxing Independence Assumption

Contents

In environmental applications, the independence assumption is questionable. It is rarely fulfilled , and never completely. From hydrological process (see Milly et al. (2008) for stationarity) to temperature analysis (see ref) or even the broader area of meteorological applications (see ref), theoretical assumptions that have been made for the models are not sustainable. This sounds also obvious in extreme values analysis of temperature data, since we expect the temperature .... Whereas it was not really problematic for block-maxima, it was much more painful for POT as we have seen both in section 2. However, here we can see in our case that it does not really happens...(verif.. and why??)

See (Beirlant et al., 2006, pp.375)

....

3.1 Stationary Extremes

From now, we considered $X_{(n)} = \max_{1\leq i\leq n} X_i$ where we assumed $X_1, \ldots, X_n$ are independent random variables. For sake of simplicity, we abandon this notation. In the sequel, this will be denoted by $\tilde{X}_{(n)} = \max_{1\leq i\leq n} \tilde{X}_i$ where $\tilde{X}_1, \ldots, \tilde{X}_n$ will typically denote a sequence of independent random variables, so that the maximum $\tilde{X}_{(n)}$ is composed of independent random variables only.

We are now interested by modelling $X_{(n)} = \max_{1\leq i\leq n} X_i$ where $\{X_i\}$ will now denote a *stationary* sequence of n random variables sharing the same marginal distribution as the sequence $\{\tilde{X}_i\}$ of independent random variables, F.

Definition 3.1 (Stationarity)**.** *We say that the sequence $\{X_i\}$ of n random variables is* **stationary** *if*

More generally, for $h \geq 0$ and $n \geq 1$, the distribution of the lagged random vector $(X_{1+h}, \ldots, X_{n+h})$ does not depend on h when the sequence is said to be (strongly) stationary.

Note that we will only focus on weak(?) stationarity.

For now, we denote $F_{i_1,\ldots,i_p}(u_1, \ldots, u_p) := \Pr\{X_{i_1} \leq u_1, \ldots, X_{i_p} \leq u_p\}$ as the joint distribution function of $(X_{i_1}, \ldots, X_{i_p})$ for any arbitrary positie integers $(i_1, \ldots, i_p)$.

Definition 3.2 ($D(u_n)$ dependence condition)**.** *From ? and following (Beirlant et al., 2006; Coles, 2001, pp.373-374, pp.93) Let $\{u_n\}$ be a sequence of real numbers. The* $\boldsymbol{D(u_n)}$ **condition** *holds if for any set of integers $i_1 < \cdots < i_p$ and $j_1 < \cdots < j_q$ such that $j_1 - i_p > \ell$, we have that*

$$|F_{i_1,\ldots,i_p,j_1,\ldots,j_q}(u_n, \ldots, u_n; u_n, \ldots u_n) - F_{i_1,\ldots,i_p}(u_n, \ldots, u_n)F_{j_1,\ldots,j_q}(u_n, \ldots u_n)| \leq \beta_{n,\ell} \ , \qquad (3.1)$$

where $\beta_{n,\ell}$ is nondecreasing and $\lim_{n\to\infty} \beta_{n,\ell_n} = 0$, for some sequence $\ell_n = o(n)$, as $n \to \infty$.

This condition ensures that, when the sets of variables are separated by a relatively short distance, typically $s_n = o(n)$, the long-range dependence between such events is limited, in a sense that is sufficiently close to zero to have no effect on the limit extremal laws.

From this result, we can retrieve the *extreme-value theorem*

Result is remarkable in the sense that, provided a series has limited long-range dependence at extreme levels ($D(u_n)$) condition makes precise), maxima of stationary series follow the same distributional limit laws as those of independent series. [S.Coles 2001 p.94]

For specific sequence of theresholds u_n that increase with n.

Theorem 3.1 (Limit distribution of maxima under $D(u_n)$)**.** *From ?. Let $\{X_i\}$ be a stationary sequence of n iid random variables with $X_{(n)} = \max(X_1, \ldots, X_n)$. If there exists sequences $\{a_n > 0\}$ and $\{b_n\}$ such that $D(u_n)$ condition holds, then*

$$Pr\{X_{(n)} \leq u_n\} \longrightarrow H(x), \qquad n \to \infty, \qquad (3.2)$$

where H is non-degenerate as defined... and $D(u_n)$ is satisfied with $u_n = a_n x + b_n$ for every real x.

[bootstrap and other....]

Theorem 3.2 (Leadbetter 1983)**.** *From (Coles, 2001, pp.) Let $\{X_i^*\}$ be a stationary sequence and let $\{X_i\}$ be a sequence of iid random variables. By defining $X_{(n)}^* = \max\{X_n^*\}$ and $X_{(n)} = \max\{X_n\}$, we have under regularity conditions,*

$$\Pr\{a_n^{-1}(X_{(n)} - b_n) \leq x\} \longrightarrow G(x), \qquad n \to \infty$$

for normalizing sequences $\{a_n > 0\}$ *and* $\{b_n\}$*, where* G *is non-degenrate, if and only if*

$$\Pr\{a_n^{-1}(X^*_{(n)} - b_n \leq x\} \longrightarrow G^*(x), \qquad n \to \infty.$$

G^* *is the limit distribution coming from a stationary process, defined by*

$$G^*(x) = G^\theta(x), \tag{3.3}$$

for some constant $\theta \in (0, 1]$ *which is called the* **extremal index**.

3.1.1 The extremal index

The *extremal index* is an important indicator quantifying the extent of extremal dependence, or equivalently the degree at which the assumption of independence is violated. From eq.(3.3), it is clear that if $\theta = 1$, then the process is independent, but the converse does not hold while the case $\theta = 0$ will not be considered as it is too "far" from independence (check with data?) and brings problems, see for example Beirlant et al. (2006, pp.379-380). Moreover, the results of Theorem 4.2. would not hold true.

.. However, the maximum has a tendency to decrease as .... (Coles, 2001, pp.96)

Formally, it can be defined as

$$\theta = \lim_{n\to\infty} \Pr\{\max(X_2, \ldots, X_{p_n}) \leq u_n \mid X_1 \geq u_n\}, \tag{3.4}$$

where $p_n = o(n)$ and the sequence u_n is such that $\Pr\{X_{(n)} \leq u_n\}$ converges. Coles (2001)[slides]

Hence, θ can be thought as the probability that an exceedance over a high threshold is the final element in a *cluster* of exceedances.

Cluster of exceedance

From eq.(3.4), we can now state that extremes have the tendency to occur in cluster, whose *mean cluster size* is θ^{-1} at the limit. Equivalently(?), θ^{-1} is the factor with which the mean distance between cluster is increased.

Identifying clusters and declustering as the distribution of a cluster maximum is the same as the marginal distribution of an exceedance. + slide 82-83(?)

However, [pp.178 Coles], information in discarded when one considers *declustering*. And this information could be substantially important in meteorological applications, for instance to determine heat or cold waves.

New parameters

When $\theta > 0$, we have from Theorem 4.2 that G^* is an EV distribution but with different scale and location parameters than G. If we note by (μ^*, σ^*, ξ^*) the parameters pertaining to G^* and those from G kept in the usual way, we have the following relationships when $\xi \neq 0$

$$\mu^* = \mu - \sigma\xi^{-1}(1 - \theta^\xi), \qquad \sigma^* = \sigma\theta^\xi. \tag{3.5}$$

In the Gumbel case ($\xi = 0$), we have $\sigma^* = \sigma$ and $\mu^* = \mu + \log\theta$. The fact that $\xi^* = \xi$ is

Return levels

From that (see clusters), one can see that the proba ility of an exceedance is variable (see coles, pp.103 or slide 82) (...)

$$r_m = u + \sigma\xi^{-1}\left[(m\zeta_u\theta)^{\xi} - 1\right] \tag{3.6}$$

It is important to take that into account as ignorance of this "dependence" can lead to overestimation of the return level.

3.1.2 Tail dependence

From (Reiss and Thomas, 2007, section 2.6), (Coles, 2001, section 8.4) or (Beirlant et al., 1996, section 9.4.1,10.3.4) + see tail dependence function `atdf()` in `R`.

Problem with traditional tools used in standard time series analysis such as (partial-) auto-correlation functions is that heavy-tailed distributions do not have moments, whereas correlation focus on dependence in the center of the distribution and not the tails. Wada et al. (2016, pp.134) Whence it is important to focus on a *tail dependence measure*.

The auto-tail dependence function using $\chi(u)$ and/or $\bar{\chi}(u)$ employs X against itself at different lags.

a possible estimator (this used by `atdf()`) can come from the sample version

$$\rho_n(u,h) = \frac{1}{n(1-u)} \sum_{i \leq n} \mathbf{1}(\min(x_i, x_{i+h}) > x_{[nu]:n}) \tag{3.7}$$

(compare with beirlant notations!!!!)$x_{[nu]:n}$

3.1.3 Modelling : Threshold Models

Block-Maxima The modelling with the techniques provided by the GEV distributions (see chapter **1**) can be used in the similar way as we have seen from (3.3) or in section 3.1 that the shape parameter remains invariant. The difference is that the effective number of maxima $n = n\theta$ will be reduced and hence the convergence will slower.

A still unsolved problem Coles (2001)[pp.98] is related to the approximations in the limit. Indeed, as the effective number of observations is reduced from n to $n\theta$, the approximation is expected to be poorer, and this "problem" will be exacerbated with increased levels of dependence in the series.

Thresholds models Practically speaking, one might expect a threshold based analysis to result in estimates of return levels with much reduced standard errors as "all" the extremes are included in the analysis, i.e. those who excess a threshold u. The example in section .... illustrated this

However, the fact that this method deals with "all" the extremes brings also some problems, and especially the issue of *temporal dependence* (see plot of acf or pacf wrt u) which is illustrated by the fact that the extremes have a tendency to *cluster*. Inferences based on the likelihood found in eq.(4.3) which relies on the independence assumption are now invalid.

Several methods can be used :

- **Filtering out** an (approximate) independent sequence of threshold exceedances.
- **Declustering**. We compute the maximum value in each cluster and then we model these clusters maximums as independent GP random variable. In this approach, we remove **temporal dependence** but we do not estimate it.

However, Fawcett and Walshaw (2012) emphasized the fact that use of the information from *all* extremes rather than just from cluster maxima (?)can be pressed into use to estimate return levels, regardless of the how strong the extremal dependence is. Hence, declustering has no interest. (see book risk p.135) This method accounts for dependence in standard error estimates of the parameters.

3.1.4 Applications

3.2 Non-Stationary Extremes

Whereas we have considered and relaxed during the previous section the first "i" of the "iid" assumption made during the whole chapter **2**, we will now tackle the last part "**id**", i.e. the strong (?) assumption that the observations are **i**dentically **d**istributed.

The stationarity assumption is very poor to hold for climatologic data **?**. It is also the case for temperature data.

OUR AIM HERE IS TO MODEL THE ***EVOLUTION*** OF As we are dealing with time varying sequences, we can

- Positive trend
- Seasonality

The aim of our modelling will more focus on a different parametrization for the mean, thus in allowing the location parameter to vary through time/seasons.

- Variation in time through t accounting for the season : $\mu(t) = \beta_0 + \mathbb{1}_i(t)$ where i=1,2,3,4 represent the seasons.

3.2.1 Block-Maxima

As we continue to consider modelling as yearly blocks, we do only face nonstationary concerns for the trend which is (probably) imputed to the Global Warming. The evidence of seasonality arising when we decrease the length of the blocks is not an issue for yearly modelling. However, we loose information, or comparatively, we do not use all the information as at least one half

3.2.2 Diagnostics

Gumbel plot (slide 94) coles

3.3 Model Comparisons

3.3.1 Statistical Tools

In order to compare our models, that is for example to check whether a trend (or seasonality) is statistically significant, or if the nonstationnary models provide an improvement over the simpler (stationary) model, we will make use of the **deviance statistic** defined as

$$D = 2\{\ell_1(\mathcal{M}_1) - \ell_0(\mathcal{M}_0)\}, \tag{3.8}$$

for two nested models $\mathcal{M}_0 \subset \mathcal{M}_1$, where $\ell_1(\mathcal{M}_1)$ and $\ell_0(\mathcal{M}_0)$ are the maximized log-likelihoods under models $\mathcal{M}_1$ and $\mathcal{M}_0$ respectively as defined in .

Asymptotically, the distribution of D is χ_k with k (df) representing the difference of parameters between model $\mathcal{M}_1$ and $\mathcal{M}_0$. Thus, comparisons of D with the critical values from χ_k will guide our decision.

3.4 Return Levels

Stationarity Under an assumption of a stationary sequence, the return level is the same for all years, and this gives rise to the notion of the return period (or m-year event). Hence, the return period of a particular event is the inverse of the probability that the event will be exceeded in any given year. The m-year return level is associated with a return period of m years. However, there are two main interpretations in this context for return periods.

(**?**, pp.100)

Denoting $X_{(n),y}$ the annual maximum for year y. Omitting the notational dependence on block size n, we assume $\{X_{(n),y}\} \overset{iid}{\sim} F$.

1. The first interpretation of the m-year event is **the expected waiting time until an exceedance occurs**. To see that, letting T be the year of the first exceedance, we we can write

$$\begin{aligned}
\Pr\{T = t\} &= \Pr\{X_{(n),1} \leq r_m, \ldots, X_{(n),t-1} \leq r_m, X_{(n),t} > r_m\} \\
&= \Pr\{X_{(n),1} \leq r_m\} \ldots \Pr\{X_{(n),t-1} \leq r_m\} \Pr\{X_{(n),t} > r_m\} && \text{[iid assumption]} \\
&= \Pr\{X_{(n),1} \leq r_m\}^{t-1} \Pr\{X_{(n),1} > r_m\} && [\text{ stationarity}] \\
&= F^{t-1}(r_m)(1 - F(r_m)) \\
&= (1 - 1/m)^{t-1}(1/m).
\end{aligned} \tag{3.9}$$

   We easily recognize that T has geometric density with parameter $1/m$. From simple properties of geometric distributions, we found its expected values is $1/(1/m)$, that is the expected waiting time for an m-year event is m years.

2. The second interpretation of the m-year event is that **the expected number of events in a period of m years is exactly 1**. To see that, we define

$$N = \sum_{y=1}^{m} I(X_{(n),y} > r_m),$$

as the random variable representing the number of exceedances in m years (where I is indicator function). We can view each year as a "trial", and from the fact that we have assumed $\{X_{(n),y}\}$ are iid, we can compute the probability that the number of exceedances in m-years is k

$$\Pr\{N = k\} = \binom{m}{k}(1/m)^k(1 - 1/m)^{m-k},$$

from which we recognize a well-know distribution, that is $N \sim$ Bin$(m, 1/m)$. Again from properties of this distribution, we easily find that N has an expected value of 1.

Non-stationarity From the definition on non-stationary process, the modelling of return period will change over time. Hence, we introduce the notation of the distribution function F_y of a particular $X_{(n),y}$. We must study

$p(y) = \Pr(X_{(n),y} > r) = 1 - F_y(r)$. If we estimate F_y, we can retrieve easily p(y). $F_y(r_p(y)) = 1 - p$ with the exceedance level $r_p(y)$ changing with year. It shows the changing nature of "risk".

Return period as expected waiting time

Return period as expected number of events

Part II

Inferential Methods

Chapter 4

Methods of Inference

Contents

We decide to present in this section the two mains methods of inference for GEV distributions. First, the likelihood-based methods for their wide applicability, and easy interpretability. Then, we will broadly present the bayesian methods for their increasing supports in this domain, and easy adjustability. Finally, we will present some other well-known methods that are also widely used to estimate GEV parameters like the Hill or the moment estimator.

As we already discussed in section **2.1.1.** (see (1.19)-(1.20)), a great advantage for the modelling is that we do not have to find the normalizing sequences

4.1 Likelihood-based Methods

The most usual method to first consider and which generally do a good job is the Maximum Likelihood (ML) inference.

A potential difficulty with the use of likelihood methods for the GEV concerns the regularity conditions that are required for the usual asymptotic properties associated with the maximum likelihood estimator to be valid. Such conditions are not satisfied by the GEV model because the end points of the GEV distribution are functions of the parameter values, $\mu - \sigma/\xi$ is an upper end-point of the distribution when $\xi < 0$, and a lower end-point (?) when $\xi > 0$. **?**

, from decreasing order of Coles (2001, pp.55)

1. $\boldsymbol{\xi < -1}$: MLE's are unlikely to be obtainable. This is due to
2. $\boldsymbol{\xi \in (-1, -0.5)}$: MLE's are generally obtainable but their standard asymptotic properties do not hold.
3. $\boldsymbol{\xi > -0.5}$: MLE's are regular, in the sense of having the usual asymptotic properties.

But fortunately, in practice, the problematic cases in the two first situations ($\xi \leq 0.5$) are rarely encountered for environmental problems. This situation corresponds to distributions (in the Weibull or in the Beta? family) with very short bounded upper tail, see for example figure **??**. And if it is the case, Bayesian inference, which do not depend on these regularity conditions, may be preferable.

Other forms of likelihood-based methods have also emerged to remedy this problem of unstability for low values of ξ. Close to a bayesian formulation, **penalized ML** method has been proposed by Coles and Dixon (1999) which adds a penalty term to the likelihood function to "force" the shape parameter to be < 1, values close to -1 being much larger penalized. (..) We talk about other methods who try to circumvent issues of the usual likelihood computation in section or in section **4.2**.

Problems of this simple method arise when the approximate normality of the MLE cannot hold. Hence, the underlying inferences are not sustainable. This is the reason why another method is usually more preferable, the *profile likelihood*.

GEV distribution

(return level) [extremesin climate change p.106] "Approximate confidence intervals for the return level can be obtained by the delta method (Casella and Berger, 2002, Sect. 5.5.4) which relies on the asymptotic normality of maximum-likelihood estimators and produces a symmetric confidence interval. Alternatively, profile likelihood methods (Coles 2001, Sect. 2.6.6) provide asymmetric confidence intervals, which better capture the skewness generally associated with return level estimates." stationary case

We are now considering a sequence $\{Z_i\}_{i=1}^n$ of independent random variables sharing each the same GEV distribution. Let denote $\boldsymbol{z} = (z_1, \ldots, z_n)$ the vector of observations. From the densities of the GEV distribution $g_\xi(z)$ defined in (1.14)-(1.15), we derive the log-likelihood $\log\left[L(\mu, \sigma, \xi; \boldsymbol{z})\right]$, for the two different cases $\xi \neq 0$ or $\xi = 0$ respectively:

1.

$$\ell(\mu, \sigma, \xi \neq 0\,; \mathbf{z}) = -m\log\sigma - (1+\xi^{-1})\sum_{i=1}^{n}\log\left[1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right]_+ - \sum_{i=1}^{n}\left[1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right]_+^{-\xi^{-1}}. \tag{4.1}$$

2.
$$\ell(\mu, \sigma, \xi = 0 \,; \mathbf{z}) = -m \log \sigma - \sum_{i=1}^{n} \left(\frac{z_i - \mu}{\sigma} \right) - \sum_{i=1}^{n} \exp \left\{ - \left(\frac{z_i - \mu}{\sigma} \right) \right\}, \tag{4.2}$$
using the Gumbel limit $\xi \to 0$ of the GEV, see (1.15).

Generalized Pareto Distribution

As we have seen, excess-over-threshold models rely on .... From (2.12), we can write the *log-likelihood* of the GPD :

$$\ell(\mathbf{z}; \xi, \sigma_u) = -n \ln \sigma_u - (1 + \xi^{-1}) \sum_{i}^{n} \ln(1 + \xi \sigma_u^{-1} z_i), \qquad (1 + \xi \sigma_u^{-1} z_i) > 0. \tag{4.3}$$

4.1.1 Profile Likelihood

Usual likelihood methods are not the most accurate for inference. In section **2.5**, the problem was that confidence intervals computed in the usual method, with standard errors computed by the Delta method in (1.33), was not reliable for inference on return levels. This is due to rejection of the normal approximation (see...) because of the severe asymmetries that are often observed in the likelihood surface for return levels, especially for large quantiles. **?**

This is why it is useful to consider an other approach linked with the usual likelihood method, the *profile likelihood* which is often more convenient when a single parameter is of interest. Let's denote it θ_j. Now let's consider the parameter vector $\boldsymbol{\theta} = (\theta_j, \boldsymbol{\theta_{-j}}) = (\mu, \sigma, \xi)$ typically for parameter inferences in EVT in a stationary context, where $\boldsymbol{\theta_{-j}}$ corresponds to all components of $\boldsymbol{\theta}$ except θ_j. $\boldsymbol{\theta_{-j}}$ can be seen as a vector of nuisance parameters. The profile log-likelihood for θ_j is defined by

$$\ell_p(\theta_j) = \arg\max_{\boldsymbol{\theta_{-j}}} \ell(\theta_j, \boldsymbol{\theta_{-j}}). \tag{4.4}$$

Henceforth for each value of θ_j, the profile log-likelihood is the maximised log-likelihood with respect to $\boldsymbol{\theta_{-j}}$, i.e. with respect to all other components of $\boldsymbol{\theta}$ but not θ_j. Generalization where $\boldsymbol{\theta_j}$ is of dimension higher than one is possible.

Another interpretation concern the χ^2 distribution

Return levels Here, we are now specifically interested in computing the profile log-likelihood for the estimation of the return level $\theta_j = r_m$. To do that, we present a method which consists of three main steps :

1. To include r_m as a parameter of the model, by **??** we can rewrite μ as a function of ξ, σ and r_m :
$$\mu = r_m - \sigma \xi^{-1} \Big[\Big(- \log\{1 - m^{-1}\} \Big)^{-\xi} - 1 \Big].$$
By plugging it in the log-likelihood in (4.1)-(4.2), we obtain the new GEV log-likelihood $\ell(\xi, \sigma, r_m)$ as a function of r_m.

2. We maximise this new likelihood $\ell(\xi, \sigma, r_m = r_m^-)$ at some fixed low value of $r_m = r_m^- \leq r_m^+$ with respect to the "nuisance" parameters (ξ, σ) to obtain the profile log-likelihood

$$\ell_p(r_m = r_m^-) = \arg\max_{(\xi,\sigma)} \ell\Big(r_m = r_m^- \ , \ (\xi, \sigma)\Big).$$

We choose arbitrarily large value of the upper range r_m^+ and conversely for starting point of r_m^-.

3. Repeat previous step for a range of values of r_m such that $r_m^- \leq r_m \leq r_m^+$ and then choose r_m which attain the maximum value of $\ell_p(r_m)$.

From this, we easily obtain the *profile log-likelihood plot*

4.2 Other Methods

Distinct inference for EVI ξ and global inference (see "others"

Beirlant et al. (2006, pp.140)

As we have seen, the two approaches we have encountered, that is *block-maxima* and POT, share commonly the same parameter ξ. Hence, it is not necessary to differentiate between these methods for the sole estimate the shape parameter.

4.2.1 Estimators Based on Extreme Order Statistics (put with POT)?

Pickands estimator

Firstly introduced by **?**, this method can be applied $\forall \xi \in \mathbb{R}$

$$\hat{\xi}_k^P = \frac{1}{\ln 2} \ln \left(\frac{X_{n-\lceil k/4 \rceil + 1, n} - X_{n-\lceil k/2 \rceil + 1, n}}{X_{n-\lceil k/2 \rceil + 1, n} - X_{n-k+1, n}} \right), \tag{4.5}$$

where we used the definition of Beirlant et al. (2006) We recall that $\lceil x \rceil$ denotes the integer (ceil) part of x.

A condition for the consistency of this estimator is that k must be chosen such that $k/n \to 0$ as $n \to \infty$. This condition will hold for the rest rest of the estimators based on (...) in the following

A problem with this intuitive estimator is that its asymptotic variance is very large (see e.g. Dekkers and Haan (1989)) and depends highly on the value of k. To improve this, we can quote the estimator of Segers (2001) which is globally more efficient, depending on the value of an extra-"parameter"(?) and a function to choose.

Methods for heavy-tailed distributions ($\xi > 0$)

Typically, EV analysis of temperature data do not show heavy-tailedness §see). For this reason, some tools commonly used for inference on **Pareto-type** distributions are not relevant. Because of their wide use and application, we will name them for completeness

The Hill estimator ($\xi > 0$)

This is probably the most simple EVI estimator thanks to the intuition behind its construction. There exists plenty of interpretations to construct it (see e.g. Beirlant et al. (2006, pp.101-104)). Unfortunately, it only holds for heavy-tailed distributions ($\xi > 0$).

It is defined as

$$\xi_k^H = k^{-1} \sum_{i=1}^{k} \ln X_{n-i+1,n} - \ln X_{n-k,n}, \qquad k \in \{1, \dots, n-1\}. \tag{4.6}$$

Following **?**, this estimator is consistent. Besides that, this estimator has several problems :

- instability with respect to the choise of k.
- Severe bias due to the heavy-tails of the distribution and thus the slowly varying component which influences negatively.
- Inadequacy with shifted data

Problem : see [pp.105]

The Moment estimator

Introduced by Dekkers et al. (1989), this estimator is a direct generalization of the Hill estimator presented in the previous section.

$$\hat{\xi}_k^M = \hat{\xi}_k^H + 1 - \frac{1}{2}\left(1 - \frac{(\hat{\xi}_k^H)^2}{\hat{\xi}_k^{H^{(2)}}}\right)^{-1}, \tag{4.7}$$

where we define

$$\hat{\xi}_k^{H^{(2)}} = k^{-1} \sum_{i=1}^{k} \left(\ln X_{n-i+1,n} - \ln X_{n-k,n}\right)^2.$$

This estimator is also consistent but

Estimator based on generalized quantile plot

To overcome the lack of graphical interpretation of the usual moment estimator,

4.2.2 The Probability-Weighted-Moment Estimator

Probability-Weighted-Moment (PWM) ...

different formulations for POT or block maxima. Look also to "other" directory **?**

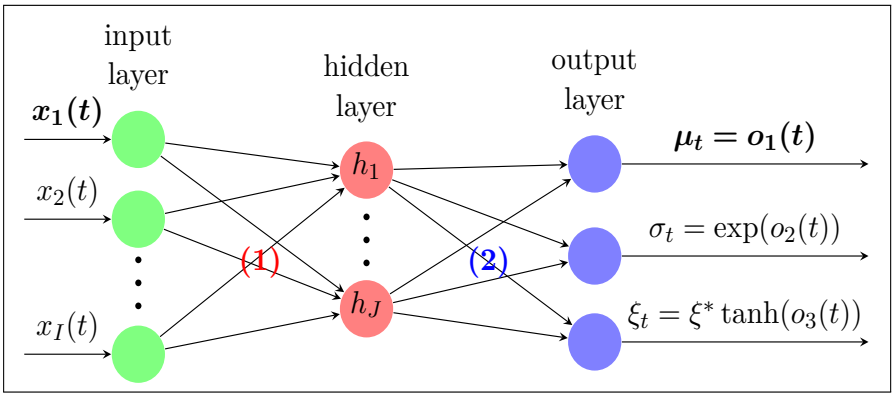

Figure 4.1: TikzFig.: *Neural Network applied to GEV. helped by Cannon (2010)*

The L-Moment Estimator

Wang (1997)

Hosking and Wallis (1997) emphasized the fact that L-moment method came historically as a modification of the PWM method.

4.2.3 Estimators based on Generalized Quantile

4.3 Improvements For Modelling Non-stationary Sequences

4.3.1 Generalized Likelihood Methods

introduced by **?** As we have seen in section **4.1**, " The ML method may diverge when sample size is small. To resolve the problems of divergence occurring in the numerical techniques used for ML, Martins and Stedinger [2000] suggest the use of a prior distribution for the shape parameter of the GEV model such that the most probable values of the parameter are included"

When dealing with nonstationnary processes, it is interesting to consider Generalized Maximum Likelihood (GML) estimators. In this case, **?** have proven that GML is likely to outperform the usual ML inference.

The GML estimator corresponds to the mode of the empirical posterior distribution...

Properties of the GML estimator see pp740. **?**

4.3.2 Neural-Network Based Inference

Neural Network (NN) ..

have the power to manage several outputs in a

"Model parameters are estimated via the GML approach using the quasi-Newton BFGS optimization algorithm, and the appropriate GEV-CDN model architecture for each location is selected by fitting increasingly complicated models and choosing the one that minimizes appropriate cost-complexity model selection criteria. For each location examined, different formu-

lations are tested with combinational cases of stationary and nonstationary parameters of the GEV distribution, linear and nonlinear architecture of the CDN and combinations of the input covariates "

? enlightens the following : Provided enough data, hidden units and an appropriate optimization, the NN can capture any smooth dependencies (relationships) of the parameters on the input, i.e., given the input, it can theoretically capture any conditional continuous density, be it asymmetric, **multimodal**, or heavy-tailed.

(see Cannon (2010) just before conclusion) One could for example expect to have particular relationships between the covariate (time or) and the parameters of interest. Only considering a linear or quadratic trends in the location parameter μ (ore more ? see section. see for other parameters) could thus be seen as a weak modelling procedure, especially when we assumed no reliable prior knowledge on the subject (see bayesian section - hyperref it). NN models have this facility of being capable of modelling any relationships without explicitly specify it *a priori*. To model correctly, one should be able to explicitly discover particular patterns (e.g., which nonlinear or linear relationship between time and TN). This is avoided here because this is done automatically through the NN process.

Physical process such as temperature or even other meteorological data (rainfall as demonstrated by Cannon (2010),...) have this tendance of demonstrating nonlinearities (see ref?) and so are NN's interesting.

As we mentioned, the NN is meant to approximate any functions with good accuracy. It comprise thus all the models considered so far, such as linear tren din μ, quadratic, etc...

Cannon (2010) recommended to use between 1 and 3 (4) hidden layers due to the relatively small sample of annual extremes (here 117).

From this, we must pay attention to the high danger of *overfitting* (see) which occurs for this sort of models. The other pitfall is its lack of interpretation of the retrieved relationships.

"It bears noting that sensitivity analysis methods, for example, the one used by Cannon and McKendry (2002), are applicable to CDN models and could be used to identify the form of nonlinear relationships between covariates and GEV distribution parameters or quantiles."

4.3.3 Bagging

Nowadays, bagging is used in many state-of-the-art algorithms such as Random Forests (see .. for comparisons of such techniques)

In a "pure" climatological point of view, *ensemble models* are of major utility, especially to make weather forecasting, see for example Suh et al. (2012) or, ... among others

For our purpose, we present another kind of ensemble modelling which is *bagging*.

" Bagging (stands for Bootstrap Aggregation) is the way decrease the variance of your prediction by generating additional data for training from your original dataset using combinations with repetitions to produce multisets of the same cardinality/size as your original data. By increasing the size of your training set you can't improve the model predictive force, but just decrease the variance, narrowly tuning the prediction to expected outcome "

"model averaging, which involves taking a weighted average of multiple models, has been recommended as a means of improving estimation performance Burnham and Anderson, 2004. This approach has been applied successfully in the context of CDN models by Carney et al.

(2005) and is worth exploring for GEV-CDN models."

?, pp.256-267 (deep learning html book) The individual classifiers' predictions (having equal weightage) are then combined by taking majority voting. This typically reduces the variance and then the (possible) overfitting

4.4 Bootstrap Methods

" like the estimated parameters themselves, the SE may not be reliable for small samples. One way to tackle this problem and improve the accuracy of SE is through the bootstrap technique (Efron, 1979). The scheme for using this technique for EV distribution function is described in detail by Katz et al. (2002), including for nonstationary cases, in which the bootstrap samples are manufactured through Monte Carlo resampling of residuals (Equation (8)) to attend to the underlying assumption that original sam- ple consist of iid data. Following this procedure, the bootstrap procedure was designed for generating 1000 samples from each original sample, considering whole year as a bloc"

In Cannon (2010), "he parametric bootstrap outperformed the residual bootstrap" Moreover, " It is possible that alternative bootstrap approaches, for example, the bias-adjusted percentile estimators evaluated by Kysely (2008), might yield better calibrated confidence intervals, although improvements were modest for stationary GEV models. "

For confidence intervals : see Cannon (2010, pp.681) following these steps

1. Fit a nonstationary model to the data
2. Transform the residuals from the fitted model so that they are identically distributed :

$$\varepsilon_t = \left[1 + \xi_t \sigma_t^{-1}(y_t - \mu_t)\right]^{-\xi^{-1}} \tag{4.8}$$

3. etc..

Monte-Carlo based methods, same as Bayesian.

Study and comparisons on the performance (coverage,..) of the methods used for the CI (boot, bayesian, likelihood, asymptotics,...)

4.4.1 Moving Block Bootstrap

[Bootstrap and other resampling in .... pp.13]

4.5 Markov models

book risk pp.136, Shaby et al. (2016) + code

4.6 Model Diagnostics : Goodness-of-Fit

After having fitted a statistical model to data, it is important to assess its accuracy in order to infer reliable conclusions from this model.

Ideally, we aim to check that our model fits well the whole population, that is the whole distribution of maxima. all the past and future temperature maxima that will arise " As this cannot be achieved in practice, it is common to assess a model with the data that were used to estimate this model. We will talk a bit about the problem that could arise from this methodology in the next section,and the problem of overfitting(?).

4.6.1 Diagnostic Plots : Quantile and Probability Plots

From (Beirlant et al., 1996, pp.18-36), together with the nice view of (Coles, 2001, pp.36-37), we present two major diagnostic tools which aims at assessing the fitting of a particular model (or distribution) against the real distribution coming from the data used to construct the model. These are called the *quantile-quantile* plot (or *qq*-plot) and the *probability* plot.

These diagnostics are popular by their easy interpretation and by the fact that they can both have graphical (i.e. subjective, qualitative, quick) view but also a more precise (i.e. objective, quantitative, rigorous) analysis can be derived, for example from the theory of linear regression. For these two diagnostic tools, we use the order statistics introduced in eq.(A.1) but now we rather consider an **ordered sample** of independent observations :

$$x_{(1)} \leq x_{(2)} \leq \ldots, \leq x_{(n)} \tag{4.9}$$

coming from a population from which we fit the estimated model (distribution) $\hat{F}$ and where $x_{(1)}$ (resp. $x_{(n)}$) is thus the minimum (resp. maximum) observation in the sample. These tools will thus help us to know if the fitted model model $\hat{F}$ is reasonable for the data.

Quantile plot

Given a ordered sample as in (4.9), a *qq-plot* consists of the locus of points

$$\left\{ \left(\hat{F}^{\leftarrow}\left(\frac{i}{n+1}\right) \, , \, x_{(i)} \right) \, : \, i = 1, \ldots, n \right\}. \tag{4.10}$$

This graphic compares the ordered quantiles $\hat{F}^{\leftarrow}\left(\frac{i}{n+1}\right)$ of the fitted model $\hat{F}$ against the ordered observed quantiles, i.e. the ordered sample from (4.9).

We used the continuity correction $\frac{i}{n+1}$.....

Probability plot

Given the same sample in (4.9), a *probability plot* consists of the points

$$\left\{ \left(\hat{F}(x_{(i)}) \, , \, \frac{i}{n+1} \right) \, : \, i = 1, \ldots, n \right\}. \tag{4.11}$$

This graph compares the estimated probability of the ordered values $x_{(i)}$ from the fitted model $\hat{F}$ against the probability...
From these two graphical diagnostic tools, the interpretation is the same and we will consider that $\hat{F}$ fits well the data if the plot looks linear, i.e. the points of the plots lie close to the unit diagonal.
Besides the fact that the probability and the quantile plots contain the same information, they are expressed in a different scale. That is, after changing the scale to probabilities or quantiles (with probability or quantile transforms), one can gain a better perception and both visualizations can sometimes lead contradictory conclusions, especially in the graphical inspection. Using both is thus preferable to make our model's diagnostic more robust.

The disadvantage of Q–Q plots is that the shape of the selected parametric distribution is no longer visible Beirlant et al. (2006)[pp.62?]

Return Level Plot

See sections ....

Overfitting problem

?? A problem of these diagnostics could arise when we focus on prediction accuracy and as we mentioned, the fact that the model is fitted from the data. This well-known problem is called *overfitting*. It can be roughly defined by the process of fitting to noise from the dataset rather than the underlying signal (put ref here). Here, it can be easily explained by the following :

- We are looking for a model which fits the data at best, i.e. for points which are the nearest possible of the diagonal line.
- But, the so-constructed model from which we put the diagnostic is fitted from these original data against which we make the comparison.
- Hence, there could be a incentive to fit a model which fits the most perfectly the available data, that is which points on the diagnostic plots is the nearest possible of the diagonal line. The model is then the best to fit the data at hand
- But, this is a catastrophy when we are seeking at making good predictions from the fitted model, that is making a guess on new, unseen, unavailable data. The model has then lost flexibility, it is not regularized and cannot generalize. (unless the feature space, hear the initial data space, has been completely explored (—>infinite data ?))

See the link with the trade-off bias-variance for threshold selection.

Gumbel plots

Z and W statistic plots

Chapter 5

Bayesian Methods

Contents

see evdbayes pdf package r

Attention : π ou f ??????

We let useful relevant tools regarding bayesian inference in appendix **A.3**

Definition 5.1 (Posterior distribution)**.** *Let* $\boldsymbol{x} = (x_1, \ldots, x_m)$ *denote the observed data of a random variable* X *distributed according to a distribution with density function*

$$\pi(\theta|\boldsymbol{x}) = \frac{\pi(\theta) \cdot L(\theta|\boldsymbol{x})}{\int_{\Theta} \pi(\theta) \cdot L(\theta|\boldsymbol{x}) \cdot d\theta} \propto \pi(\theta) \cdot L(\theta|\boldsymbol{x}) \tag{5.1}$$

where $L(\cdot)$ denotes the likelihood function, as in **??** *but there it is the log-likelhiood !!! and θ usually denotes the multidimensional set of parameters in EVT, $\theta = (\mu, \sigma, \xi)$, at least in a univariate stationary context.*

1. Whenever it is possible, it allows to introduce to introduce other source of knowledge coming from the domain at-hand, by the elicitation of a prior. The counter-argument of this advantage is that it also introduces (improper ?) subjectiveness.

2. "account- ing for parameter and threshold uncertainty is perhaps handled most easily in the Bayesian paradigm" (Dey and Yan, 2016, pp.106)

   As such, It permits an elegant way of making future predictions which is one of the most(?) important issue in EVT.

3. Bayesian framework can overcome the regularity conditions of the likelihood inference (see section **4.1**). Thus it usually provides a viable alternative in cases when MLE (for example) breaks down. And actually, we are not so fo far from the problematic situations depicted in section **3.1**. Moreover, the Highest Posterior Probability (HPD) region is constructed so that... and there is no more need to fall to asymptotic theory as in conventional methods.

4. For an asymmetric distribution, the HPD interval can be a more reasonable summary than the central probability interval (see illustration ...). For symmetric densities, HPD and central interals are the same while HPD is shorter for asymmetric densities. See Liu et al. (2015)....

As the dependence becomes stronger, the run length n must be larger in order to achieve the same precision. Dependence exists both within the output for a single parameter (autocorrelations) and across parameters (cross-correlations), we discuss this issue in section text.

5.1 Prior Elicitation

Sometimes viewed as advantage from the amount of information that can be retrieved, and sometimes viewed as an drawback due to the (rather unquantifiable) subjectivity that introduced, the construction of the prior is a key step in Bayesian analysis.

Priors are necessary in the Bayesian paradigm to be able to compute the posterior in (5.1).But, priors require the legitimate statement of domain's expert, to make this viewed the less subjective as possible

Prior may not be of great importance if the size (m) of the dataset is large. It can be seen from (5.1) where the amount of information contained in the data through $L(\theta|\boldsymbol{x})$ will be prominent compared to this contained in the prior through $\pi(\theta)$. Prior will have limited influence.

One is aware that this is not often the case in EVT cases. By design, we are dealing with rather small so-constructed datasets. And mostly for this reason, it could be important to incorporate additional information in this limited dataset through the prior distribution.

5.1.1 Non-informative Priors

Receive a correct and accepted advice from an expert is often difficult So, in many cases, we cannot inject information through the prior. We must then construct a prior which represent this lack of knowledge so that they do not influence posterior inferences.

There exists a vast amount of uninformative priors in the literature (see e.g. Yang and Berger (1996), Ni and Sun (2003)) This family of priors can be *improper*, i.e. priors for which the integral of $\pi(\theta)$ over the parameter space is not finite. It is valid to use improper priors only if the posterior target is proper.

Adjustments of these priors must always be thought in practical applications

Jeffrey's prior

is specified as

$$\pi(\theta) \propto \sqrt{\det I(\theta)}, \qquad \text{where} \qquad I_{ij}(\theta) = \mathbb{E}_\theta\left[-\frac{\partial^2 \log f(\boldsymbol{x}|\theta)}{\partial\theta_i \partial\theta_j}\right], \quad i,j = 1,\dots,d. \tag{5.2}$$

where $f(\boldsymbol{x}|\theta)$ is of course the density function of X.

This prior is invariant to reparametrization, but has complex form for EV models, and it exists only when $\xi > -0.5$ in GEV models, where it is function of ξ and σ only.

MDI prior

Maximal Data Information priors

However, it has been showed by Northrop and Attalides (2016) that both Jeffrey and MDI priors give improper posterior when there are no truncation of the shape parameter, i.e. we must restrict the fact that $\pi(\theta) \to \infty$ as $\xi \to (-)\infty$ for Jeffreys (MDI), in order to obtain a proper posterior.

Vague priors

The last and often preferred alternative to construct uninformative priors is to use proper priors which are near flat, e.g. which are uniform or with exhibits large variance for the normal distribution.

In GEV we will take independent normal-distributed priors each with a large (tuned) variance. When these variances increase, we get at the limit

$$\pi(\theta) = \pi(\mu,\nu,\xi) \overset{(\perp\!\!\!\perp)}{=} \pi(\mu)\cdot\pi(\nu)\cdot\pi(\xi) \ \propto 1, \tag{5.3}$$

where $\nu = \log\sigma$.

Taking multivariate normal distribution as prior has also been proposed (see) is often difficult as it involves 9 (hyper)parameters in total and this can be difficult

5.1.2 Informative Priors

STAN : "It can also be a huge help with computation to have less diffuse priors, even if they're not informative enough to have a noticeable impact on the posterior. "

Gamma Distributions for Quantile Differences

Beta Distributions for Probability Ratios

The Bayes Factor

5.2 Bayesian Computation : Markov Chains

Methods have been developed for sampling from arbitrary posterior distributions $\pi(\theta|\boldsymbol{x})$. Simulations of N values $\theta_1, \theta_2, \ldots, \theta_N$ that are iid from $\pi(\theta|\boldsymbol{x})$ can be used to estimate features of interest.

But simulating from $\pi(\theta|\boldsymbol{x})$ is usually not achievable and this is why we need **Markov Chain Monte Carlo** (MCMC) techniques. We use it to simulate a markov chain $\theta_1, \theta_2, \ldots, \theta_N$ that conerge to the target distribution $\pi(\theta|\boldsymbol{x})$. This means that, after some *burn-in period* B, $\theta_{B+1}, \ldots, \theta_N$ can be treated as random sample from $\pi(\theta|\boldsymbol{x})$.

Let's now (a bit weakly) define one of the most important results in Markov Chain theory.

Definition 5.2 (*First-order discrete-time* Markov Property)**.** *Let $k_0, k_1, \ldots$ be the states associated to a sequence of time-homogeneous random variables, say $\{\theta_t : t \in \mathbb{N}\}$. The Markov property states that the distribution of the future state θ_{t+1} depends only on the distribution of the current state θ_t. In other words, given θ_t, we have that θ_{t+1} is independent of all the states prior to t. We can write this as*

$$\Pr\{\theta_{t+1} = k_{t+1} \mid \theta_t = k_t,\ \theta_{t-1} = k_{t-1}, \ldots\} = \Pr\{\theta_{t+1} = k_{t+1} \mid \theta_t = k_t\}. \tag{5.4}$$

or see Angelino et al. (2016, section 2.2.3) for more in-depth results.

The samples are not independent, and the dependence influences the accuracy of the posterior estimates. As dependence becomes stronger, we must increase the run-length N to achieve the same accuracy.

5.2.1 Algorithms

We are looking for a so-generated chain that has a stationary distribution $\pi(\theta|\boldsymbol{x})$. This is the case if the chain is

1. *aperiodic*
2. *irreducible* or *ergodic*, that is if any state for θ can be reached with probability > 0 in a finite number of steps from any other state for θ.

"The Markov chains Stan and other MCMC samplers generate are *ergodic* in the sense required by the Markov chain central limit theorem, meaning roughly that there is a reasonable chance of reaching one value of theta from another." **?**

With MH or Gibbs sampler, we need to tune individually the proposal standard deviations to reach a correct acceptance, and this is often done with trial-and-error methodology.

The performance of the standard Markov chain Monte Carlo estimators depends on how effectively the Markov transition guides the Markov chain along the neighborhoods of high probability. If the exploration is slow then the estimators will become computationally inefficient, and if the exploration is incomplete then the estimators will become biased Betancourt (2016). It is then necessary to consider other form of sampling...

5.2.2 Hamiltonian Monte Carlo

Package Rstan

Neal and others (2011) and Betancourt and Girolami (2015) are really

HMC permit to better exploit the properties of the target distribution to make informed jumps through neighborhoods of high probability while avoiding neighborhoods of low probability entirely.

5.2.3 Computational efficiency comparison

In modern statistical area, computing methods have been widely ... And this need for compiutations will rise in the future.

We will then compare our 3 methods too see if effectively

5.3 Convergence Diagnostics

When applying MCMC algorithms to estimate posterior distributions, it is vital to assess convergence of the algorithm to try to ensure that we reached the stationary target distribution. Let's now enumerate some of the key steps we must keep in mind when thinking about convergence, an hence reliable results.

1. A sufficient *burn-in period* $B < N$ must be chosen to ensure that the convergence to the posterior distribution $\pi(\theta|\boldsymbol{x})$ has occurred.

2. For the same reason, a sufficient number of simulations N to eliminate the influence of initial conditions and ensure accuracy in the estimations ((and then make sure than we are sampling from the target stationary (posterior) distribution)).

3. Several dispersed starting values must have been simulated to ensure we explored all the regions of high probability. This is particularly important when the target distribution is complex.

4. The chains must have good mixing properties, in the sense that the whole parameter space (...) A common technique that we will apply is to run different chains several times and then combine a proportion of each chain (typically 50%) to get the final chain. This

procedure wants to ensure a proper mixing behaviour. The potential scale reduction factor (Gelman diagnostic) is also a popular tool, see .

We must keep in mind that no convergence diagnostics can prove that convergence really happened and validate the "model". However, a combined use of several relevant diagnostics will be required to increase our confidence that convergence actually happened.

5.3.1 Proposal Distribution

The main ideas are :

- If the variance of the proposal distribution is too large, most proposals will be rejected :

  ie the jumps through the chain are too large,

- If the variance of the proposal distribution is too low, then most proposals will be accepted

Both are harmful for the objective of an efficient "visit" of the whole parameter space.

Widely speaking, we consider 2 different types of algorithms in which it is preferable to target a certain acceptance rate. It is distinguished by the updating manner of the components of θ through the algorithm, i.e. the 3 univariate parameters of interest.

- When all components of θ are updated simultaneously, it is recommended to target an acceptance rate of around 0.20. Roberts et al. (1997) have shown that, under quite general conditions, the asymptotically optimal acceptance rate is 0.234. (for target density that has a symmetric product form) This quantity has been verified by Sherlock et al. (2009). It holds for the *Metropolis-Hastings* algorithm.

- When the components are updated one at a time, an acceptance rate of around 0.40 is recommended. It holds for the *Gibbs sampler* algorithm.

Let's (see **?** for example for the first case)

Gelman-Rubin diagnostic : the $\hat{R}$ statistic

As discussed in item **4** above

Geweke diagnostic

Thinning

iteration k is stored only if $k \mod \text{thin}$ is zero (and if k greater than or equal to the burn-in B).

This typically reduces the precision of posterior estimates, but it may represent a necessary computational saving.

5.3.2 The problem of auto and cross-correlations in the chains

There exists exists 2 problems of correlations in the output delivered by a MC.

- **Autocorrelation** is the
- **Cross-correlation**

5.4 Posterior Predictive

notation for posterior ? pi or f

As discussed in item **2** above, prediction is of important interest in EVT, and this is "facilitated" in the Bayesian paradigm. This also permits a more straightforward quantification of the inferential uncertainty associated.

Definition 5.3 (Posterior Predictive density)**.** *Let X_{m+1} denotes a (one-step-ahead) future observation with density $f(x_{m+1}|\boldsymbol{\theta})$. Then we define the Posterior Predictive density of a future observation X_{m+1} given $\boldsymbol{x}$ as*

$$\begin{aligned} f(x_{m+1}|\boldsymbol{x}) &= \int_\Theta f(x_{m+1},\theta|\boldsymbol{x})\cdot d\theta = \int_\Theta f(x_{m+1}|\theta)\cdot\pi(\theta|\boldsymbol{x})\cdot d\theta \\ &:= \mathbb{E}_{\theta|\boldsymbol{x}}\big[f(x_{m+1}|\theta)\big] \end{aligned} \tag{5.5}$$

where the last line emphasizes that we can evaluate $f(x_{m+1}|\boldsymbol{x})$ by averaging over the different possible parameter values.

The uncertainty in the model is reflected here through $\pi(\theta|\boldsymbol{x})$ while the uncertainty due to variability in future observations is also reflected through $f(x_{m+1}|\theta)$.

Definition 5.4 (Posterior Predictive probability)**.** *The posterior predictive probability of X_{m+1} exceeding some threshold x is accordingly given by*

$$\begin{aligned} \Pr\{X_{m+1} > x \mid \boldsymbol{x}\} &= \int_\Theta \Pr\{X_{m+1} > x \mid \theta\}\cdot\pi(\theta|\boldsymbol{x})\cdot d\theta \\ &= \mathbb{E}_{\theta|\boldsymbol{x}}\big[\Pr(X_{m+1} > x \mid \theta)\big] \end{aligned} \tag{5.6}$$

This quantity is often of interest in EVT as we are rather concerned with the probability of future unknown observable exceeding some threshold.

However, this quantity is difficult to obtain analytically. Hence, we will more rely on simulated approximations. Given a sample $\theta_1,\dots,\theta_r$ from the posterior $\pi(\theta|\boldsymbol{x})$, we use

$$\Pr\{X_{m+1} > x \mid \boldsymbol{x}\} \approx r^{-1}\sum_{i=1}^{r}\Pr\{X_{m+1} > x \mid \theta_i\}, \tag{5.7}$$

where $\Pr\{X_{m+1} > x \mid \theta_i\}$ follows directly from $f(x|\theta)$.

We will now analyse more in-depth the numerical computations in the Bayesian paradigm or how we can get numerically a sample of the posterior distribution.

5.5 Bayesian Predictive Accuracy for Model Validation

5.5.1 Cross-validation for predictive accuracy

When having large amount of data, we can use a well-known and widely used technique coming from Machine Learning. That is, dividing the dataset between a training (typically 75% of the whole set) and a test set containing the remaining observations. For example, having N draws $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)}$ coming from the posterior $\pi(\theta|x_{train})$, we can score each value using (?)

$$\log \left[N^{-1} \sum_{t(i)?=1}^{N} f(x^*|\theta^{(t)}) \right]. \tag{5.8}$$

However, we often do not have large amounts of data. Henceforth, we can use the *cross-validation* technique which is more relevant in smaller dataset, but which is computationally more demanding. There exists several variants of them.

Leave-one-out cross-validation

The *Leave-One-Out* (LOO) cross-validation is the

K-fold cross-validation

Vehtari et al. (2016)

Or we can use other criteria which avoid the computations. The basic approach is to use $\log f(x|\bar{\theta}) - p^*$ for N draws $\theta^{(1)}, \ldots, \theta^{(N)}$ from $\pi(\theta|x)$. where p^* represents the effective number of parameters and $\bar{\theta}$ the posterior mean. Several methods exists using this idea. We will see the two most important.

Deviance Information Criterion

The *Deviance Information Criterion* (DIC) was first used by Spiegelhalter et al. (2002) and use the following estimate for the effective number of parameters

$$p^* = 2 \cdot \left(\log f(x|\bar{\theta}) - N^{-1} \sum_{t=1}^{N} \log f(x|\theta^{(t)}) \right) \tag{5.9}$$

It is defined on the deviance scale and smaller DIC values indicate better models.

$$\text{DIC} = 2 \log f(x|\bar{\theta}) - \frac{4}{N} \sum_{t=1}^{N} \log f(x|\theta(t)) \tag{5.10}$$

Widely Applicable Information Criterion

The *Widely Applicable Information Criterion* (WAIC) is a more recent approach proposed by Watanabe (2010) and is given by

$$\text{WAIC} = 2\sum_{i=1}^{n}\left[\log(\mathbb{E}_{\theta|x} f(x_i|\theta)\right] - \mathbb{E}_{\theta|x}\log f(x_i|\theta) \tag{5.11}$$

or

$$\text{WAIC} = \sum_{i=1}^{n}\left[2\log\left(N^{-1}\sum_{t=1}^{N} f(x_i|\theta^{(t)})\right) - \frac{4}{N}\sum_{t=1}^{N}\log f(x_i|\theta^{(t)})\right] \tag{5.12}$$

There exists for sure other several methods, as proposed by Gelman et al. (2014).

'LOO and WAIC have various advantages over simpler estimates of predictive error such asAIC and DIC but are less used in practice because they involve additional computational steps'

For each generated chains with dispersed starting values, we evaluate separately the information criteria. The discrepancies between the chains are small (?), which is a good sign.

5.6 Bayesien Inference ?

5.6.1 Bayesian Credible Intervals

The Bayesian *credible intervals* are inherently different from the frequentist's confidence intervals. In the Bayesian intervals, the bounds are treated as fixed and the estimated parameter as a random variable, while in the frequentist's setting, bound are random variables and the parameter is a fixed value.

There exist mainly two kinds of credible interval in the Bayesian sphere :

- The *Highest Probability Interval* (HPD) which is defined as the shortest interval containing $x\%$ of the posterior probability, e.g. if we want a 95% HPD interval (ξ_0, ξ_1) for ξ :

$$\int_{\xi_0}^{\xi_1} \pi(\xi|\boldsymbol{x})d\xi = 0.95 \qquad \text{with} \qquad \pi(\xi_0|\boldsymbol{x}) = \pi(\xi_1|\boldsymbol{x}). \tag{5.13}$$

  It is often the preferred interval as it gives the parameter's values having the highest posterior probability.

- The Quantile-based credible intervals or *equal-tailed interval* picks an interval which ensures a probability of being below this interval as likely as of being above it. For some posterior distribution which are not symmetric, this could be misleading, thus it is not the most recommended interval. (see ..) However, these are often easily obtained when we have a random smaple of the posterior...(?)

5.6.2 Distribution of Quantiles : Return Levels

The Markov chains generated can be transformed to estimate quantities of interest such as quantiles and hence return levels.

The values can be retrieved in the same manner as we have done in the GEV frequentist setting in (1.32). If the df F associated is GEV then $y_m = -\log(1 - m^{-1})$, and if F is GPD then $y_m = m^{-1}$.

r_m is the quantile corresponding to the upper tail probability $p = m^{-1}$.

We can use the values of the samples generated by the posterior to estimate features of this distribution. (... see edbayes)

5.7 Bayesian Model Averaging

5.8 Applications

"It is often the case that more than one model provides an adequate fit to the data. Sensitivity analysis determines by what extent posterior inferences change when alternative models are used" book risk analysis other section pp.2.

"The basic method of sensitivity analysis is to fit several models to the same problem. Posterior inferences from each model can then be compared."

5.8.1 Own

5.8.2 `evdbayes` R package : MH algorithm

5.9 Comparisons

Hartmann and Ehlers (2016) We can calculate the effective sample size (ESS) using the posterior samples for each parameter :

$$\text{ESS} = N \cdot (1 + 2 \sum_{k} \gamma(k))^{-1} \tag{5.14}$$

where N is still the number of posterior samples and $\gamma(k)$ are the monotone lag k sample autocorrelations. **?**. We can thus interpret this as the number of effectively independent samples.

Part III

Experimental Framework (Simulation Study) : Global....

In this part, we will focus on the application of the methods seen during the theoretical part.

5.10 R package

Be careful with some functions we have created. We have put a "." for readability to separate components of the function but it does not mean inheritance relationships such as in the S3 object oriented R system. For example, `yearly.extrm()` does **not** mean the `yearly()` method for `extrm` objects.

Chapter 6

Simulation study: Performance evaluation of different methods

For this thesis, we decided to take as simulated data a mean of the parameters estimation of all the (relevant) methods considered. Then we simulate the data according to these parameters

–> look for RMSE's on MC loops

Chapter 7

Conclusion

During this thesis, we have statistically assessed the presence of a trend in the extreme temperatures in Uccle. We first detected that the trend is significative by the method of linear regression. We also discovered that the best fitted GEV model is the one with a linear trend in the location parameter.

"A key issue in applications is that inferences may be required well beyond the observed tail of the data, and so an assumption of stability is required:" **?**

"Another approach would be to use something other than time as the covariate in the model. For instance, one could imagine linking temperature data directly to CO2 level rather than time. However, linking to a climatological covariate makes extrapolation into the future more difficult, as one would need to extrapolate the covariate as well. No obvious climatological covariate comes to mind for the Red River application. "

Timescale-uncertainty effects on extreme value analyses seem not to have been studied yet. For stationary models (Sect. 6.2), we anticipate sizable effects on block extremes–GEV estimates only when the uncertainties distort strongly the blocking procedure. For nonstationary models (Sect. 6.3), one may augment confidence band construction by inserting a timescale simulation step (after Step 4 in Algorithm 6.1) Mudelsee (2014, pp.262)

!!!!! not put too much references in the text !!!!!

Managing references :

(!!!!! delete unuseful equation numbering ?!!!!)

finalize hyperref in the text. BE CAREFULL of the numbering written and the real ("hyperrefed" numbering section) +finalize also for def, thm, ... (?)

Replace "distribution functions" by "cdf",.... + all other abbrevations (et, gev,...) –> put it then in the list of abbr.

Careful with notation with X or Z in the (c)pdf, likelihood, RV,...

voir notation vectors (en gras ou avec bar en bas?)

attention aux notation homogenes (ex : partie stationnary,...) –> pour dénoter les maximums,....

notations for sequences !! attention aux "iid", "n random variables",...

Delete page counter for part pages,...

be prudent that all "methods" are listed (numbered) in (each) ToC

add square brackets [] for cite ?

Careful for boldsymbols, especiallyin bayesian

Appendix A

Statistical tools for Extreme Value Theory

A.1 Preliminaries

[MOTS DEXPLICATIONS SUR TTES LES FORMULES (domain atraction condition, etc...]

Order statistics

First of all, we write the i-th order statistics $X_{(i)}$ which are the statistics ordered by increasing value

$$X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)} \tag{A.1}$$

We adopt this simpler notation by assuming that the number of our observations will be denoted by n.

One order statistics is of particular interest for our purpose, the maximum

$$X_{(n)} := \max_{1 \leq i \leq n} X_i \tag{A.2}$$

for the minimum, $X_{(1)} = \min_{1 \leq i \leq n} X_i$ that we will still use by converse (see section .)

$$X_{(1)} := \min_{1 \leq i \leq n} X_i = -\max_{1 \leq i \leq n} (-X_i) \tag{A.3}$$

It is very important to keep in mind that all the analysis made in the following for maxima can be extended to minima by this relation (A.3). Our analysis in chapter ? will also use minima and thus make implicit use of this relations.

We can retrieve the distribution of our statistic of interest $X_{(n)}$

$$\Pr\{X_{(n)} \leq x\} = \Pr\{X_1, \ldots, X_n \leq x\} \overset{(\perp\!\!\!\perp)}{=} \Pr\{X_1 \leq x\} \ldots \Pr\{X_n \leq x\} = F^n(x), \tag{A.4}$$

(write vertically)

where independence ($\perp\!\!\!\perp$) follows from the iid assumption of the sequence $\{X_i\}$.

Miscellaneous

We also define the *survival* function $\bar{F} = 1 - F$ which is widely useful for this kind of (biostatistical) applications.

Finally, we decide to include in Appendix some concepts of convergence (A.), regularly and slowly varying functions (A.), as they will often appear in the text.

A.2 Tails of the distributions

[Heavy-tails] The distribution of a random variable X with distribution function F is said to have a **heavy** right **tail** if

$$\lim_{n\to\infty} e^{\lambda x} \Pr[X > x] = \lim_{n\to\infty} e^{\lambda x} \bar{F}(x) = \infty, \quad \forall \lambda > 0 \tag{A.5}$$

More generally, we can say that a random variable X has heavy tails if $\Pr(|X| > x) \to 0$ at a polynomial rate. In this case, some of the moments will be undefined. see stats ana. book 2007 p.30

[long right tail] The distribution of a random variable X with distribution function F is said to have a long right tail if $\forall t > 0$,

$$\lim_{n\to\infty} \Pr[X > x + t | X > x] = 1 \;\Leftrightarrow\; \bar{F}(x + t) \sim \bar{F}(x) \text{ as } x \to \infty \tag{A.6}$$

The term **risk** can be defined as tail probability p. But, because many application fields of risk analysis exist, such as actuarial science, econometrics or of course what is of interest for this thesis, climatology, many risk definitions are in usage ; Thywissen (2006) lists 22, although not completely mutually exclusive, definitions currently employed. The definition via the probability has the advantage that this is a fundamental, real number, from which the other parameters of interest, for example, the expected economic loss, can be derived.

A.3 Convergence concepts

Convergence in probability

Convergence in distribution

Weakly convergence

We say that a sequence of random variables X_n *converges weakly* to

Well other forms of convergence do exist, but these ones are the most important in regard to EVT. However, the reader may refer e.g. to **?** for more in-depth results.

A.4 Varying functions

A.5 Bayesian Inference

A.5.1 Algorithms

Metropolis–Hastings Algorithm

The *Metropolis–Hastings* algorithm is one of the first and of the pioneering algorithm discovered by Hastings (1970) to compute MCMC for Bayesian analysis.

Algorithm 1: The Metropolis–Hastings Algorithm

1. Pick a starting point θ_0 and fix some number N of simulations.
2. **For** $t = 1, \ldots, N$ **do**
   (a) Sample proposal θ_* from a proposal density $p_t(\theta_*|\theta_{t-1})$,
   (b) Compute the ratio

$$r = \frac{\pi(\theta_*|\boldsymbol{x}) \cdot p_t(\theta_{t-1}|\theta_*)}{\pi(\theta_{t-1}|\boldsymbol{x}) \cdot p_t(\theta_*|\theta_{t-1})} = \frac{\pi(\theta_*) \cdot \pi(\boldsymbol{x}|\theta_*) \cdot p_t(\theta_{t-1}|\theta_*)}{\pi(\theta_{t-1}) \cdot \pi(\boldsymbol{x}|\theta_{t-1}) \cdot p_t(\theta_*|\theta_{t-1})}.$$

   (c) Set

$$\theta_t = \begin{cases} \theta_* & \text{with probability } \alpha = \min(r, 1) \\ \theta_{t-1} & \text{otherwise.} \end{cases}$$

This algorithm remains valid when π is only proportional to a target density function and thus it can be used to approximate 5.1 .

Note that the proposal density is often chosen to be symmetric so that we will just sample under a "simple Metropolis" algorithm where r is thus simplified to be only the ratio of the posterior densities, $r = \frac{\pi(\theta_*|\boldsymbol{x}) \cdot}{\pi(\theta_{t-1}|\boldsymbol{x})}$.

We can shortly summarize the *pros* and *cons* this algorithm :

- *PROS* : Very easy to program and works even for relatively complex densities.
- *CONS* : Can be very inefficient, in the sense that it will require lots of iterations before the stationary target distribution will be reached. This requires some tuning to the algorithm through

Gibbs Sampler

The *Gibbs Sampler* can be seen as a special case of the Metropolis-Hastings algorithm. Suppose our parameter vector θ can be divided into d subvectors $(\theta_1, \ldots, \theta_d)$, and let's say in our case that each of these "subvectors" represent a single parameter, thus typically one of the three (μ, σ, ξ), for the simplest case so that $d = 3$ in this model. At each $t = 1, \ldots, N$, the Gibbs sampler samples the subvectors $\theta_t^{(j)}$ conditional on both the data $\boldsymbol{x}$ and the remaining subvectors $\theta_{t-1}^{(-j)}$

at their current values. Therefore, we have $\theta_{t-1}^{(-j)} = \left(\theta_t^{(1)}, \ldots, \theta_t^{(j-1)}, \theta_{t-1}^{(j+1)}, \ldots, \theta_{t-1}^{(d)}\right)$ and each $\theta_t^{(j)}$ is sampled from $\pi\left(\theta^{(j)}|\theta_{t-1}^{(-j)}, \boldsymbol{x}\right)$.

Algorithm 2: PSEUDOCODE of the Gibbs Sampler

1. Pick a starting point θ_0 and fix some number N of simulations.
2. **For** $t = 1, \ldots, N$ **do**
   For $j = 1, \ldots, d$ **do**
   (a) Sample proposal θ_* from a proposal density $p_{t,j}(\theta_*^{(j)}|\theta_{t-1}^{(j)})$,
   (b) Compute the ratio

$$\begin{aligned} r =& \frac{\pi\left(\theta_*^{(j)}|\theta_{t-1}^{(-j)}, \boldsymbol{x}\right) \cdot p_{t,j}\left(\theta_{t-1}^{(j)}|\theta_*^{(j)}\right)}{\pi\left(\theta_{t-1}^{(j)}|\theta_{t-1}^{(-j)}, \boldsymbol{x}\right) \cdot p_{t,j}\left(\theta_*^{(j)}|\theta_{t-1}\right)} \\ =& \frac{\pi\left(\theta_t^{(1)}, \ldots, \theta_t^{(j-1)}, \theta_*^{(j)}, \theta_{t-1}^{(j+1)}, \ldots, \theta_{t-1}^{(d)} \mid \boldsymbol{x}\right) \cdot p_{t,j}\left(\theta_{t-1}^{(j)}|\theta_*^{(j)}\right)}{\pi\left(\theta_t^{(1)}, \ldots, \theta_t^{(j-1)}, \theta_{t-1}^{(j)}, \theta_{t-1}^{(j+1)}, \ldots, \theta_{t-1}^{(d)} \mid \boldsymbol{x}\right) \cdot p_{t,j}\left(\theta_*^{(j)}|\theta_{t-1}\right)}, \end{aligned}$$

   (c) Set

$$\theta_t^{(j)} = \begin{cases} \theta_*^{(j)} & \text{with probability } \alpha = \min(r, 1) \\ \theta_{t-1}^{(j)} & \text{otherwise.} \end{cases}$$

(better signs for conditional bar "|" !!!!!!!

This algorithm depends on being able to simulate from $\pi\left(\theta^{(j)}|\theta_{t-1}^{(-j)}, \boldsymbol{x}\right)$ which is often impossible. However, one can use Metropolis-Hastings to $\pi\left(\theta^{(j)}|\theta_{t-1}^{(-j)}, \boldsymbol{x}\right)$, giving the above.

A special case arise if one can simulate directly so that $r = 1$, i.e. we take $p_{t,j}(\theta_*^{(j)}|\theta_{t-1}^{(j)}) = \pi(\theta_*^{(j)}|\theta_{t-1}^{(-j)}, \boldsymbol{x})$. (The proposal $p_{t,j}(\,\cdot\,)$ is also often symmetric, i.e. $p_{t,j}(\theta_*^{(j)}|\theta_{t-1}^{(j)}) = p_{t,j}(\theta_{t-1}^{(j)}|\theta_*^{(j)})$. But it cannot be simplified in the equation.)

It is important for our tasks to tune the average probability of acceptance to be roughly between 0.4 and 0.5 (see e.g. Gelman et al. (2013, chapter 11)) so that the so-generated markov-chain has desirable properties. This is doneby setting the standard deviation $\sigma(j)$ of the univariate normal distribution taken for $p_{t,j}(\theta_*^{(j)}|\theta_{t-1}^{(j)})$. Whereas our $\theta(j)$ are (often?) univariate, it is difficult to set each $\sigma^{(j)}$ to achieve average acceptance probabilities for all parameters. We will then use a trial-and-error approach.

Note also the increase of complexity with this sampler compared to the Metropolis-Hastings, where the nested loop implies that there are d iterations with each simulation.

pros and *cons* :

- *PROS* : Easy to program and, for some problems, it can also be very efficient. It is a pleasant way to split multidimensional problems into simpler (typically univariate) densities.

- *CONS* : Sometimes hard to compute analytically the conditional distributions. Not all densities can be split into pleasant conditionals equations.

Metropolis-within-Gibbs?

Hamiltonian Monte Carlo

http://deeplearning.net/tutorial/hmc.html#hmc

A difficulty we have faced and that we would like to point out for GEV models is (see p.316-317 STAN manual) t

"Most of the computation [in Stan] is done using Hamiltonian Monte Carlo. HMC requires some tuning, so Matt Hoffman up and wrote a new algorithm, Nuts (the "No-U-Turn Sampler") which optimizes HMC adaptively. In many settings, Nuts is actually more computationally efficient than the optimal static HMC! "

The *Hamiltonian Monte Carlo* (HMC)

"The Hamiltonian Monte Carlo algorithm starts at a specified initial set of parameters theta; in Stan, this value is either user-specified or generated randomly. Then, for a given number of iterations, a new momentum vector is sampled and the current value of the parameter θ is updated using the leapfrog integrator with discretization time and number of steps L according to the Hamiltonian dynamics. Then a Metropolis acceptance step is applied, and a decision is made whether to update to the new state (theta*; rho*) or keep the existing state" **?**

? have demonstrated in similar application that HMC (and Riemann manifold HMC) are much more computationally efficient than traditional MCMC algorithms such as MH.

Definition A.1 (Total energy of a closed system : Hamiltonian function)**.** *For a certain particle; Let $\pi(\theta)$ be the posterior distribution and let $\boldsymbol{p} \in \mathbb{R}^d$ denote a vector of auxiliary parameters independent of θ and distributed as $\boldsymbol{p} \sim N(\mathbf{0}, \boldsymbol{M})$. We can interpet θ as the position of the particle and $-\log \pi(\theta|\boldsymbol{x})$ describes its potential energy while $\boldsymbol{p}$ is the momentum with kinetic energy $\boldsymbol{p}'\boldsymbol{M}^{-\mathbf{1}}\boldsymbol{p} \cdot 2^{-1}$. Then the total energy of a closed system is the Hamiltonian function*

$$\mathcal{H}(\theta, \boldsymbol{p}) = -\mathcal{L}(\theta) + \boldsymbol{p}'\boldsymbol{M}^{-\mathbf{1}}\boldsymbol{p} \cdot 2^{-1}, \qquad \text{where} \qquad \mathcal{L}(\theta) = \log \pi(\theta). \tag{A.7}$$

We define $\mathcal{X} = (\theta, \boldsymbol{p})$ as *the combined state* of the particle.

The unormalized joint density of $(\theta, \boldsymbol{p})$ is

$$f(\theta, \boldsymbol{p}) \propto \pi(\theta) \cdot \exp\{-\boldsymbol{p}'\boldsymbol{M}^{-\mathbf{1}}\boldsymbol{p} \cdot 2^{-1}\} \propto \exp\{-\mathcal{H}(\theta, \boldsymbol{p})\}. \tag{A.8}$$

Following Hartmann and Ehlers (2016), the idea is to use the Hamiltonian dynamics equations (not shown here for..) to model the evolution of a particle that keep the total energy constant. Introducing the auxiliary variables $\boldsymbol{p}$ and using the gradients (..) will lead to a more efficient exploration of the parameter space

These differential equations cannot be solved so numerical integrators are required, for instance the "Störmer-Verlet" from **?** which will introduce discretization. A MH step is then required to correct the error and ensure convergence. The new proposal $\mathcal{X}_* = (\theta_*, \boldsymbol{p}_*)$ will be accepted with probability

$$\alpha(\mathcal{X}, \mathcal{X}_*) = \min\left[\frac{f(\theta_*, \boldsymbol{p}_*)}{f(\theta, \boldsymbol{p})}, 1\right] = \min\Big[\exp\left\{\mathcal{H}(\theta, \boldsymbol{p}) - \mathcal{H}(\theta_*, \boldsymbol{p}_*)\right\}\ ,\ 1\Big]. \tag{A.9}$$

As $\boldsymbol{M}$ is symmetric positie definite, $\boldsymbol{M} = m\boldsymbol{I}_d$. Then we can summarize the HMC algorithm in the following, in its 'simplest' form :

marie est la best : $Moyenne = \frac{1}{n}\sum_{i=1}^{n} X_i$

Algorithm 3: The Hamiltonian Monte Carlo algorithm

1. Pick a starting point θ_0 and set $i = 1$.
2. **Until** convergence has been reached **do**
   (a) Sample $\boldsymbol{p}_* \sim N_d(\boldsymbol{0}, \boldsymbol{I}_d)$ and $u \sim U(0, 1)$,
   (b) Set $(\theta_I, \boldsymbol{p}_I) = (\theta_{i-1}, \boldsymbol{p}_*)$ and $\mathcal{H}_0 = \mathcal{H}(\theta_I, \boldsymbol{p}_I)$,
   (c) **repeat** L times
   ▷ $\boldsymbol{p}_* = \boldsymbol{p}_* + \frac{\epsilon}{2}\nabla_\theta \mathcal{L}(\theta_{i-1}$
   ▷ $\theta_{i-1} = \theta_{i-1} + \epsilon \cdot \boldsymbol{p}_*$
   ▷ $\boldsymbol{p}_* = \boldsymbol{p}_* + \frac{\epsilon}{2}\nabla_\theta \mathcal{L}(\theta_{i-1})$,
   (d) Set $(\theta_L, \boldsymbol{p}_L = (\theta_{i-1}, \boldsymbol{p}_*$ and $\mathcal{H}^{(1)} = \mathcal{H}(\theta_L, \boldsymbol{p}_L)$,
   (e) Compute $\alpha\Big[(\theta_I, \boldsymbol{p}_I, (\theta_L, \boldsymbol{p}_L)\Big] = \min\Big[\exp\{H^{(0)} - H^{(1)},\ 1\Big]$,
   (f) **If** $\alpha\Big[(\theta_I, \boldsymbol{p}_I), (\theta_L, \boldsymbol{p}_L)\Big] > u$ **then** set $\theta_i = \theta_L$
   else set $\theta_i = \theta_I$,
   (g) Increment $i = i + 1$ and return to step (a).

As you can see, it is not trivial. The basic idea to keep in mind is that jumping rules are much more efficient than for traditional algorithms because they learn from the gradient of the log posterior density, so they know better where to jump to. As a result, it can be MUCH more efficient.

Chains are expected to reach stationarity faster as it proposes moves to regions of higher probabilities.

pros and *cons* :

- *PROS* : Easy to program as we just have to write down the model. Very efficient in general, and works for all types of problems.
- *CONS* : Need to learn how to use STAN, less control over the sampler bu maybe it is for the best?

do not forget **?**

Beirlant: 2004 !!

Bibliography

Amir AghaKouchak, David Easterling, Kuolin Hsu, Siegfried Schubert, and Soroosh Sorooshian, editors. *Extremes in a Changing Climate*, volume 65 of *Water Science and Technology Library*. Springer Netherlands, Dordrecht, 2013. ISBN 978-94-007-4478-3 978-94-007-4479-0.

Elaine Angelino, Matthew James Johnson, and Ryan P. Adams. Patterns of Scalable Bayesian Inference. *Foundations and Trends® in Machine Learning*, 9(2-3):119–247, 2016. ISSN 1935-8237, 1935-8245.

A. A. Balkema and L. de Haan. Residual Life Time at Great Age. *The Annals of Probability*, 2 (5):792–804, 1974. ISSN 0091-1798.

Jan Beirlant, Jozef L. Teugels, and Petra Vynckier. *Practical Analysis of Extreme Values*. Leuven University Press, 1996. ISBN 978-90-6186-768-5. Google-Books-ID: ylR3QgAACAAJ.

Jan Beirlant, Yuri Goegebeur, Johan Segers, and Jozef Teugels. *Statistics of Extremes: Theory and Applications*. John Wiley & Sons, March 2006. ISBN 978-0-470-01237-6. Google-Books-ID: jqmRwfG6aloC.

Michael Betancourt. Diagnosing Suboptimal Cotangent Disintegrations in Hamiltonian Monte Carlo. *arXiv:1604.00695 [stat]*, April 2016. arXiv: 1604.00695.

Michael Betancourt and Mark Girolami. Hamiltonian Monte Carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, 79:30, 2015.

Alex J. Cannon. A flexible nonlinear modelling framework for nonstationary generalized extreme value analysis in hydroclimatology. *Hydrological Processes*, 24(6):673–685, March 2010. ISSN 08856087.

Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. Springer London, London, 2001. ISBN 978-1-84996-874-4 978-1-4471-3675-0.

Stuart G. Coles and Mark J. Dixon. Likelihood-Based Inference for Extreme Value Models. *Extremes*, 2(1):5–23, March 1999. ISSN 1386-1999, 1572-915X.

A. L. M. Dekkers, J. H. J. Einmahl, and L. De Haan. A Moment Estimator for the Index of an Extreme-Value Distribution. *The Annals of Statistics*, 17(4):1833–1855, December 1989. ISSN 0090-5364, 2168-8966.

Arnold L. M. Dekkers and Laurens De Haan. On the Estimation of the Extreme-Value Index and Large Quantile Estimation. *The Annals of Statistics*, 17(4):1795–1832, December 1989. ISSN 0090-5364, 2168-8966.

Dipak K. Dey and Jun Yan. *Extreme Value Modeling and Risk Analysis: Methods and Applications*. CRC Press, January 2016. ISBN 978-1-4987-0131-0. Google-Books-ID: PYhUCwAAQBAJ.

Michael Falk and Frank Marohn. Von Mises Conditions Revisited. *The Annals of Probability*, 21(3):1310–1328, July 1993. ISSN 0091-1798, 2168-894X.

Michael Falk, Jürg Hüsler, and Rolf-Dieter Reiss. *Laws of Small Numbers: Extremes and Rare Events*. Springer Basel, Basel, 2011. ISBN 978-3-0348-0008-2 978-3-0348-0009-9.

Lee Fawcett and David Walshaw. Estimating return levels from serially dependent extremes: ESTIMATING RETURN LEVELS FROM SERIALLY DEPENDENT EXTREMES. *Environmetrics*, 23(3):272–283, May 2012. ISSN 11804009.

R. A. Fisher and Leonard H. C. Tippett. Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample. *ResearchGate*, 24(02):180–190, January 1928. ISSN 1469-8064.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis, Third Edition*. CRC Press, November 2013. ISBN 978-1-4398-4095-5. Google-Books-ID: ZXL6AQAAQBAJ.

Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding Predictive Information Criteria for Bayesian Models. *Statistics and Computing*, 24(6):997–1016, November 2014. ISSN 0960-3174.

B. Gnedenko. Sur La Distribution Limite Du Terme Maximum D'Une Serie Aleatoire. *Annals of Mathematics*, 44(3):423–453, 1943. ISSN 0003-486X.

L. de Haan and Ana Ferreira. *Extreme value theory: an introduction*. Springer series in operations research. Springer, New York ; London, 2006. ISBN 978-0-387-23946-0. OCLC: ocm70173287.

Marcelo Hartmann and Ricardo Ehlers. Bayesian Inference for Generalized Extreme Value Distributions via Hamiltonian Monte Carlo. *Communications in Statistics - Simulation and Computation*, pages 0–0, March 2016. ISSN 0361-0918, 1532-4141. arXiv: 1410.4534.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, April 1970. ISSN 0006-3444.

J. R. M. Hosking and J. R. Wallis. Parameter and Quantile Estimation for the Generalized Pareto Distribution. *Technometrics*, 29(3):339, August 1987. ISSN 00401706.

J. R. M. (Jonathan Richard Morley) Hosking and James R Wallis. *Regional frequency analysis : an approach based on L-moments*. Cambridge ; New York : Cambridge University Press, 1997. ISBN 0521430453 (hardbound).

James Pickands Iii. Statistical Inference Using Extreme Order Statistics. *The Annals of Statistics*, 3(1):119–131, January 1975. ISSN 0090-5364, 2168-8966.

A. M. G. Klein Tank, J. B. Wijngaard, G. P. Können, R. Böhm, G. Demarée, A. Gocheva, M. Mileta, S. Pashiardis, L. Hejkrlik, C. Kern-Hansen, R. Heino, P. Bessemoulin, G. Müller-Westermeier, M. Tzanakou, S. Szalai, T. Pálsdóttir, D. Fitzgerald, S. Rubin, M. Capaldo, M. Maugeri, A. Leitass, A. Bukantis, R. Aberfeld, A. F. V. van Engelen, E. Forland, M. Mietus, F. Coelho, C. Mares, V. Razuvaev, E. Nieplova, T. Cegnar, J. Antonio López, B. Dahlström, A. Moberg, W. Kirchhofer, A. Ceylan, O. Pachaliuk, L. V. Alexander, and P. Petrovic. Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. *International Journal of Climatology*, 22(12):1441–1453, October 2002. ISSN 1097-0088.

M. R. Leadbetter, Georg Lindgren, and Holger Rootzén. *Extremes and Related Properties of Random Sequences and Processes*. Springer Series in Statistics. Springer New York, New York, NY, 1983. ISBN 978-1-4612-5451-5 978-1-4612-5449-2.

Ying Liu, Andrew Gelman, and Tian Zheng. Simulation-efficient shortest probability intervals. *Statistics and Computing*, 25(4):809–819, July 2015. ISSN 0960-3174, 1573-1375.

P. C. D. Milly, Julio Betancourt, Malin Falkenmark, Robert M. Hirsch, Zbigniew W. Kundzewicz, Dennis P. Lettenmaier, and Ronald J. Stouffer. Stationarity Is Dead: Whither Water Management? *Science*, 319(5863):573–574, February 2008. ISSN 0036-8075, 1095-9203.

Manfred Mudelsee. *Climate Time Series Analysis*, volume 51 of *Atmospheric and Oceanographic Sciences Library*. Springer International Publishing, Cham, 2014. ISBN 978-3-319-04449-1 978-3-319-04450-7.

Radford M. Neal and others. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2:113–162, 2011.

Shawn Ni and Dongchu Sun. Noninformative priors and frequentist risks of bayesian estimators of vector-autoregressive models. *Journal of Econometrics*, 115(1):159–197, July 2003. ISSN 0304-4076.

P. J. Northrop and N. Attalides. Posterior propriety in Bayesian extreme value analyses using reference priors. *Statistica Sinica*, 26(2), April 2016. ISSN 1017-0405.

Murray C. Peel, Q. J. Wang, Richard M. Vogel, and THOMAS A. McMAHON. The utility of L-moment ratio diagrams for selecting a regional probability distribution. *Hydrological Sciences Journal*, 46(1):147–155, 2001.

E. C. Pinheiro and S. L. P. Ferrari. A comparative review of generalizations of the Gumbel extreme value distribution with an application to wind speed data. *arXiv:1502.02708 [stat]*, February 2015. arXiv: 1502.02708.

Rolf-Dieter Reiss and Michael Thomas. *Statistical analysis of extreme values: with applications to insurance, finance, hydrology and other fields ; [includes CD-ROM]*. Birkhäuser, Basel, 3. ed edition, 2007. ISBN 978-3-7643-7230-9 978-3-7643-7399-3. OCLC: 180885018.

Sidney I. Resnick. *Extreme Values, Regular Variation and Point Processes*. Springer Series in Operations Research and Financial Engineering. Springer New York, New York, NY, 1987. ISBN 978-0-387-75952-4 978-0-387-75953-1.

Mathieu Ribatet. A User's Guide to the POT Package (Version 1.4). *month*, 2006.

G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120, February 1997. ISSN 1050-5164, 2168-8737.

Gianluca Rosso. Extreme Value Theory for Time Series using Peak-Over-Threshold method. *arXiv preprint arXiv:1509.01051*, 2015.

David Ruppert, M. P. Wand, and R. J. Carroll. Semiparametric Regression. Cambridge Books, Cambridge University Press, 2003.

Johan Segers. Generalized Pickands Estimators for the Extreme Value Index: Minimal Asymptotic Variance and Bias Reduction. 2001.

Benjamin A. Shaby, Brian J. Reich, Daniel Cooley, and Cari G. Kaufman. A Markov-switching model for heat waves. *The Annals of Applied Statistics*, 10(1):74–93, March 2016. ISSN 1932-6157.

Chris Sherlock, Gareth Roberts, and others. Optimal scaling of the random walk Metropolis on elliptically symmetric unimodal targets. *Bernoulli*, 15(3):774–798, 2009.

V. P. Singh and H. Guo. Parameter estimation for 3-parameter generalized pareto distribution by the principle of maximum entropy (POME). *Hydrological Sciences Journal*, 40(2):165–181, April 1995. ISSN 0262-6667, 2150-3435.

David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

M.-S. Suh, S.-G. Oh, D.-K. Lee, D.-H. Cha, S.-J. Choi, C.-S. Jin, and S.-Y. Hong. Development of New Ensemble Methods Based on the Performance Skills of Regional Climate Models over South Korea. *Journal of Climate*, 25(20):7067–7082, May 2012. ISSN 0894-8755.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, August 2016. ISSN 0960-3174, 1573-1375. arXiv: 1507.04544.

R. Wada, T. Waseda, and P. Jonathan. Extreme value estimation using the likelihood-weighted method. *Ocean Engineering*, 124:241–251, 2016. ISSN 0029-8018.

J. L. Wadsworth. Exploiting Structure of Maximum Likelihood Estimators for Extreme Value Threshold Selection. *Technometrics*, 58(1):116–126, January 2016. ISSN 0040-1706.

Q. J. Wang. LH moments for statistical analysis of extreme events. *Water Resources Research*, 33(12):2841–2848, December 1997. ISSN 1944-7973.

Sumio Watanabe. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec): 3571–3594, 2010.

Ruoyong Yang and James O. Berger. *A catalog of noninformative priors*. Institute of Statistics and Decision Sciences, Duke University, 1996.