UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA
E INVESTIGAÇÃO OPERACIONAL

EXTREME VALUE THEORY: AN APPLICATION TO SPORTS

Sérgio Luís Ganhão Vicente

Dissertação
Mestrado em Estatística
2012

UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA
E INVESTIGAÇÃO OPERACIONAL

EXTREME VALUE THEORY: AN APPLICATION TO SPORTS

Sérgio Luís Ganhão Vicente

Dissertação orientada por:
Prof. Doutora Maria Isabel Fraga Alves
Prof. Doutora Maria Ivette Leal de Carvalho Gomes

Dissertação
Mestrado em Estatística
2012

Acknowledgements

First of all, I would like to thank my beloved sister, for her unconditional support, understanding and great patience during my period of isolation and inaccessibility, necessary for writing this master thesis.

This thesis would not have been possible without all the help, support, patience and endeavour of my principal supervisor, Professora Doutora Maria Isabel Fraga Alves, who never stopped believing in me and who became progressively more than a teacher, acting like a mother and a good friend.

I also thank my second supervisor, Professora Doutora Maria Ivette Leal de Carvalho Gomes, whose advices and meticulous corrections were beneficial to improve the quality of my exposition.

I would like to express my very special thanks to Professora Doutora Maria Antónia Amaral Turkman for all my current knowledge and achieved level with the **R** software, a masterpiece of this thesis.

I am specially grateful to my Swiss friends and teachers from Geneva, my beloved and wonderful native city, for their importance in my life since my childhood and for all their support during this master thesis.

I would like to express a particular warm thanks to my friends Helena Avelar and Luís Ribeiro for their eternal support and interest along all my thesis.

I would also like to extend my gratitude to my special pupils of Biology of Faculdade de Ciências de Lisboa for all their interest in my thesis and for believing in my capacities as a teaching professional.

Finally, I thank my parents for giving me life, without which this thesis would not have been created.

Sérgio Vicente

Contents

Abstract

Extreme Value Theory can be applied to several areas, where the existence of extreme events is a daily reality. Characterized by constant record breaks such as minimal times or maximal speeds, Sports can obtain benefits from Extreme Value Theory, when used as indicator of records' quality. In particular, the 100 metres race, whose current lowest record of 9.58 seconds is held by Usain Bolt, requires an exceptional cardiorespiratory capacity, monitored by the Maximal Oxygen Uptake, or $\dot{V}O_2max$, which measures the maximal amount of oxygen used during intense efforts, in millilitres per bodyweight and per minute (ml/kg/min). The highest $\dot{V}O_2max$ (96 ml/kg/min) was recorded for the skiers Bjørn Dæhlie and Espen Harald Bjerke. What is the probability of exceeding the aforementioned records? Is there a finite limit for these quantities?

Extreme Value Theory is the most appropriate tool to answer these questions, offering two possible approaches: a parametric and a semi-parametric one. The former focuses on estimating the parameters of a proposed underlying model, using the Maximum Likelihood or the Probability Weighted Moments methods. In particular, the Block Maxima method proposes the Generalized Extreme Value distribution as a suitable model to be fitted to the whole dataset and the POT method proposes the Generalized Pareto distribution as a suitable one to be fitted only to observations above a fixed level. Concerning the semi-parametric approach, there is no distribution proposal. Assuming only that the underlying distribution's tail satisfies Gnedenko's Theorem, the goal is to estimate the shape parameter of the underlying distribution, known as Extreme Value Index, which determines the weight of its tail. All the inference is based on a portion of the sample above a random level to be determined.

With the obtained estimates, both approaches answer the previous questions computing exceedance probabilities and endpoint estimates.

Keywords: Extreme Value Theory, Block Maxima method, POT method, Semi-parametric approach, Sports.

Resumo

A Teoria dos Valores Extremos tem as suas origens na década de 1920, com o trabalho pioneiro de Leonard Tippett, ao qual foi solicitado que encontrasse uma forma de tornar os fios de algodão produzidos pela indústria algodoeira britânica mais resistentes. Nos seus estudos, rapidamente se apercebeu de que a resistência dos fios estava directamente relacionada com a força das fibras mais fracas. Surgiu então a necessidade de criar uma teoria probabilística que permitisse lidar com situações em que a quantificação e modelação de acontecimentos ditos extremos passasse a ser o alvo de interesse do investigador, uma vez que a Teoria Clássica era insuficiente para fornecer respostas às questões que se colocavam. Com a ajuda de Sir Ronald Fisher, Leonard Tippett lançou as bases de todo um corpo probabilístico teórico que viria a adqurir uma importância fundamental e crescente em ramos onde a existência de acontecimentos extremos acaba por ser uma condição sine qua non, podendo colocar sérios problemas e entraves se não houver uma compreensão e controlo do fenómeno que os origina. Dada a escassez de dados que caracteriza tais fenómenos, pela sua natureza extrema, e mesmo até rara, a Teoria dos Valores Extremos adquire um papel crucial no sentido de expurgar informação estatística a partir dos elementos disponíveis.

Desde então, são inúmeras as áreas que, cada vez mais, recorrem à Teoria dos Valores Extremos no sentido de obter uma maior compreensão acerca do mecanismo de produção dos fenómenos extremos que regem e justificam a existência dessas áreas. Encontramos assim a sua presença em áreas como a Hidrologia, onde a constante ameaça de cheias, ruptura de diques e elevação do nível das águas do mar pode pôr em risco inúmeras vidas humanas. O Mercado Financeiro, assolado pela flutuação constante dos indicadores financeiros, vê na Teoria dos Valores Extremos uma ferramenta preciosa para poder lidar com as graves consequências económicas que podem surgir quando tais indicadores atingem níveis extremos. O aumento da temperatura global do planeta, cujos níveis extremos podem ameaçar a sobrevivência de muitas espécies, obriga a área do Ambiente a socorrer-se e obter respostas junto da Teoria dos Valores Extremos.

A área do Desporto não foge à regra. Em particular, o Atletismo é caracterizado por um constante aperfeiçoamento dos atletas, onde a prossecução e manutenção de recordes acaba por ser um factor determinante e um objectivo comum, no sentido de alcançarem prestígio, reconhecimento e realização profissional. Variáveis como tempos mínimos, alturas máximas e comprimentos máximos caracterizam por si só as diversas modalidades que compõem as provas atléticas. Uma das mais famosas modalidades do Atletismo, pelo interesse crescente que suscita e pela natureza excpecional dos recordes alcançados, é sem dúvida a prova dos 100 metros. Nomes tais como Carl Lewis, Ben Johnson ou, mais recentemente, Usain Bolt são indissociáveis desta modalidade, onde o talento do atleta é medido pelo tempo mínimo que demora a percorrer uma distância de 100 metros. Actualmente, o recorde mundial é detido pelo jamaicano Usain Bolt, que conseguiu percorrer 100 metros em 9.58 segundos, no Campeonato Mundial de Atletismo de 2009, em Berlim. Face a este recorde, quais são as possibilidades actuais de vencer este recorde? Qual a probabilidade de manter este nível ou então de reduzi-lo para um nível inferior? Ou então, será que se chegou a um patamar abaixo do qual um atleta não consegue descer mais? Quaisquer que sejam as respostas a estas perguntas, é consenso universal que a prestação dum atleta de corrida de alta competição está directamente relacionada com a sua capacidade cardiorespiratória. É precisamente a monitorização e aperfeiçoamento dessa capacidade que conduz um atleta no caminho do sucesso, permitindo-lhe, assim, atingir níveis extremos, quer em termos de tempo, quer em termos de velocidade. Uma das variáveis usadas na medição da capacidade cardiorespiratória dum atleta é o consumo máximo de oxigénio, mais conhecido por $\dot{V}O_2max$, que representa a quantidade máxima de oxigénio que o corpo humano consegue assimilar, transportar e usar durante um exercício físico intenso, medida em mililitros por quilo de peso corporal e por minuto (ml/kg/min). O controlo permanente desta variável é de importância vital não só em atletas de corrida de velocidade, como também em ciclistas de alta competição e esquiadores de fundo. A manutenção dum nível elevado do VO2max acaba por ser um factor de preocupação constante por parte deste tipo de atletas, dada a sua ligação íntima com um alto desempenho durante as provas atléticas. O $\dot{V}O_2max$ mais elevado até à actualidade foi registado nos esquiadores noruegueses Bjørn Dæhlie e Espen Harald Bjerke, que atingiram um nível de 96 ml/kg/min. Qual a probabilidade de um atelta de alta competição ultrapassar este valor? Será que o corpo humano tem a possibilidade de exceder muito mais este limite? Qual o valor mais elevado do $\dot{V}O_2max$ que, nas circunstâncias actuais, pode ser atingido por um atleta de alta competição?

Para responder a todas as questões colocadas no parágrafo anterior, a Teoria dos Valores Extremos é sem dúvida a ferramenta mais adequada. As respostas podem ser então obtidas seguindo duas perspectivas: uma perspectiva paramétrica e uma perspectiva semi-paramétrica. A perspectiva paramétrica tem por pressuposto base a existência dum modelo paramétrico subjacente à obtenção dos dados provenientes de acontecimentos extremos, em que o objectivo central passa pela estimação dos parâmetros desse modelo por métodos de estimação pontual, tais como o método da Máxima Verosimilhança e o método dos Momentos Ponderados de Probabilidade. A partir dessas estimativas, as perguntas anteriormente colocadas encontram as suas respostas em parâmetros estimados, tais como o limite superior (ou inferior) do suporte do modelo subjacente aos dados ou ainda a probabilidade de excedência de níveis elevados (ou baixos). A adopção dum modelo paramétrico adequado é então a questão-chave da abordagem paramétrica. Uma vez que esse modelo é evidentemente desconhecido, surgem então várias propostas dentro do âmbito paramétrico. O método dos Máximos por Blocos (vulgo método dos Máximos Anuais, quando os dados são obtidos de forma anual) propõe um ajustamento da família Generalizada de Valores Extremos aos dados disponíveis, considerando que estes são réplicas independentes duma variável aleatória que selecciona apenas o máximo de cada bloco previamente definido. Por outro lado, o método POT (do inglês Peaks-Over-Threshold) propõe o ajustamento da família Generalizada Pareto às observações que excedem um determinado nível fixado a priori, considerando que essas observações representam uma amostra proveniente da cauda direita (ou esquerda) do modelo subjacente aos dados disponíveis.

A perspectiva semi-paramética não propõe nenhum modelo paramétrico para ajustar aos dados e centra a sua atenção na estimação do parâmetro de forma do modelo subjacente desconhecido, que se designa por Índice de Valores Extremos, e que está directamente relacionado com o peso da cauda direita (ou esquerda) do modelo. Para essa estimação ser então possível, a cauda do modelo subjacente deve obedecer a certas condições, uniformizadas e formalizadas por Boris Gnedenko em 1943, que, de acordo com a abordagem semi-paramétrica, são assumidas como estando satisfeitas pelo modelo desconhecido. A estimação do parâmetro de forma é então feita seleccionando as observações da amostra que se encontram acima dum determinado nível aleatório, que não está fixo à partida e que depende do tamanho da amostra em causa, uma vez que se considera que as observações de topo transportam a informação necessária acerca da cauda do modelo subjacente. A determinação do nível aleatório óptimo a considerar perante uma determinada amostra é então uma questão de importância central, sem a qual a obtenção duma estimativa para o Índice de Valores Extremos fica seriamente comprometida. Uma vez obtida a estimativa desse parâmetro, a abordagem semi-paramétrica também permite

responder às questões atrás colocadas, focando-se na estimação do limite superior (ou inferior) do suporte do modelo subjacente ou na obtenção de probabilidades de excedência de níveis elevados (ou baixos).

Palavras-chave: Teoria dos Valores Extremos, Método dos Máximos por Blocos, Abordagem POT, Abordagem semi-paramétrica, Desporto.

List of Figures

List of Tables

List of acronyms e abbreviations

ABias: Asymptotic Bias
AMSE: Asymptotic Mean Squared Error
a.s.: almost sure
AVar: Asymptotic Variance
CI: Confidence Interval
d.f.: distribution function
ERV: Extended Regular Variation
$\mathcal{ERV}_\gamma$: Class of Extended Regular Varying functions of index γ
EVI: Extreme Value Index
EVT: Extreme Value Theory
GEV: Generalized Extreme Value
GEVd: Generalized Extreme Value distribution
GP: Generalized Pareto
GPd: Generalized Pareto distribution
i.i.d.: independent e identically distributed
LAN: Locally Asymptotically Normal
LRT: Likelihood Ratio Test
ME-plot: Mean Excess plot
ML: Maximum Likelihood
p.d.f.: probability density function
PORT: Peaks Over Random Threshold
POT: Peaks Over Threshold
PWM: Probability Weighted Moments
QQ-plot: Quantile Quantile plot
r.v.: random variable
RV: Regular Variation
$\mathcal{RV}_\alpha$: Class of Regular Varying functions of index α

Chapter 1

Introduction

Extreme Value Theory has its roots in the 1920s, with the pioneering contributions of Leonard Tippett, who was requested to find a way of strengthening the cotton threads of the British cotton industry. During his studies, he soon discovered that the strength of the threads was ruled by the resistance of the weakest fibres. It was then necessary to create a probabilistic theory to be applied in situations where quantifying and modelling extreme events was the main focus of the investigator, as the Classical Statistical Theory was insufficient to find answers to the emerging questions. With the help of Sir Ronald Fisher, Leonard Tippett laid the foundations of a theoretical probabilistic framework, which would quickly become important in areas where the existence of extreme phenomena is a *sine qua non* condition, as the misunderstanding and lack of control of such phenomena can lead to severe damages and problems. Because of their extremal, or even rare, nature, the investigator frequently faces data scarcity. Therefore, Extreme Value Theory is a valuable tool that makes it possible to deal with such situations. The decisive step is taken in 1943, when Boris Gnedenko synthesized and formalized all the knowledge to date about extreme events in his famous theorem, known as the first theorem of Extreme Value Theory.

Since then, we assist to an increasing demand coming from areas where the existence of extreme phenomena is a *raison d'être*, forcing them to understand the underlying mechanisms responsible for the emergence of extreme events. We find the presence of Extreme Value Theory in areas such as Hydrology, where flood threats, dam bursts and high sea levels may put human lives at risk. The Financial market, characterized by constant fluctuations of financial indicators, can use the Extreme Value Theory as a powerful tool to prevent catastrophic economic damages caused by extreme levels of their indicators. The so called global warming, whose extreme levels can jeopardize the survival of several species, is a serious environmental problem that can find some answers in Extreme Value Theory.

Even Sports do not escape from the influence of Extreme Value Theory. In particular, Athletics is characterized by a constant improvement of athletes, where achieving and maintaining records is the basic rule of all its modalities. Variables such as minimum time, maximum speed, maximum length and maximum height define completely each modality. One of the most famous athletics' events is the 100 metres race, where the athletes are distinguished by their capacity of running 100 metres in the shortest possible time. Currently, the record is hold by the Jamaican Usain Bolt, who ran the aforementioned distance in 9.58 seconds, at the 2009 Berlin World Championships. In the present circumstances, what are the possibilities of breaking this record? Have we achieved a steady state with no possibility of reducing Bolt's record? Independently from the answers, it is universally accepted that the performance of a runner is directly related to his cardiorespiratory capacity and the constant enhancement of this capacity is a matter of eternal concern for every athlete. One of the variables that can be used to monitor this capacity is the *Maximal Oxygen Uptake* (shortly, $\dot{V}O_2max$), which represents the maximum quantity of oxygen that can be assimilated and used by the human body during an intense effort, measured in millilitres per kilogram of bodyweight and per minute (ml/kg/min). The maintenance of a high $\dot{V}O_2max$ level is then one of the keys for success, not only for runners, but even for cyclists or cross-country skiers. To date, the highest $\dot{V}O_2max$ was recorded for the Norwegian skiers Bjørn Dæhlie and Espen Harald Bjerke, who attained a high level of 96 ml/kg/min. In the present circumstances, is is then possible to surpass this value? What is the maximal $\dot{V}O_2max$ that a current top athlete can achieve?

To answer all the previous questions, the Extreme Value Theory is undoubtedly the most suitable tool. In Chapter 2, we present the Extreme Value Theory, as the result of the works of Leonard Tippett, Sir Ronald Fisher and Boris Gnedenko, enriched by the more recent contributions of Emil Gumbel and Laurens de Haan. Chapter 3 covers the two main approaches of Extreme Value Theory used to answer the aforementioned questions: the parametric approach and the semi-parametric approach, discussing the main tools that characterize each approach. In Chapter 4, we turn back to the $\dot{V}O_2max$ variable and to the 100 metres race, in order to seek answers for the aforementioned questions, by means of the methodologies presented in Chapter 3. Finally, in Chapter 5, we close this thesis extracting some conclusions from Chapter 4 and letting space for open problems and unanswered questions. Finally, Appendix A includes the software **R** scripts used for all the computations in the $\dot{V}O_2max$ analysis. Since the 100 metres analysis follows exactly the same paths as the $\dot{V}O_2max$, the scripts for the 100 metres analysis will not be presented. They are exactly the same, changing only variables names and some numbers.

Chapter 2

The Extreme Value Theory

2.1 Introduction

Extreme Value Theory (EVT) is a statistical and theoretical framework, which deals with modelling the behaviour of sample extremes, such as the sample minimum and the sample maximum. The behaviour of such order statistics may be assessed by their exact distribution function (d.f.) or by their limiting distribution function, the asymptotic distribution function, if we increase the sample size towards infinite.

Let $(X_1, X_2, ..., X_n)$ be a sample of n independent and identically distributed (i.i.d.) random variables (r.v.'s), with d.f. F. The corresponding ordered sample in non-decreasing order is denoted by $(X_{1:n}, X_{2:n}, ..., X_{n:n})$, where $X_{i:n}$, $i = 1, ..., n$, stands for the i-th order statistic. In particular, $X_{1:n}$ and $X_{n:n}$ represent the *sample minimum* and the *sample maximum*, respectively. In this thesis, we will focus only on the results about the sample maximum, since the corresponding results for the sample minimum can be obtained from those of the sample maximum. Then, consider the sequence of maxima $M_1 = X_1, M_n = X_{n:n} = \max(X_1, X_2, \ldots, X_n)$, for $n \geq 2$, obtained from the above sample. As mentioned, all the results for the sample minimum can be obtained from those of the sample maximum, since $m_n = \min(X_1, X_2, \ldots, X_n) = -\max(-X_1, -X_2, \ldots, -X_n)$.

The exact distribution of M_n can be obtained from the d.f. F. Indeed, for all $x \in \mathbb{R}$,

$$F_{M_n}(x) = P(M_n \leq x) = P(X_1 \leq x, X_2 \leq x, ..., X_n \leq x) = \prod_{i=1}^{n} P(X_i \leq x) = F^n(x).$$

But the interest of this thesis is the behaviour of the sample maximum, when the sample size increases towards infinity:

Theorem 2.1 *Let F be the underlying d.f. of a sequence of r.v.'s and x^F its right endpoint, i.e., $x^F = sup\{x : F(x) < 1\}$, which may be infinite. Then*

$$M_n \xrightarrow[n\to\infty]{p} x^F,$$

where $\xrightarrow[n\to\infty]{p}$ means convergence in probability.

Proof. We know that a sequence of r.v.'s $X_1, X_2, ..., X_n, ...$ converges in probability towards the r.v. X, if and only if, $\forall \epsilon > 0$,

$$\lim_{n\to\infty} P(\mid X_n - X \mid \geq \epsilon) = 0.$$

So,

$$\begin{aligned} P(\mid M_n - x^F \mid \geq \epsilon) &= P(M_n \geq x^F + \epsilon \quad \vee \quad M_n \leq x^F - \epsilon) \\ &= P(M_n \geq x^F + \epsilon) + P(M_n \leq x^F - \epsilon) \\ &= 0 + P(M_n \leq x^F - \epsilon) \\ &= F^n(x^F - \epsilon). \end{aligned}$$

We know that $x^F = \sup\{x : F(x) < 1\}$. Consequently,

$$\lim_{n\to\infty} P(\mid M_n - x^F \mid \geq \epsilon) = \lim_{n\to\infty} F^n(x^F - \epsilon) = 0,$$

since $F(x^F - \epsilon) < 1$.

□

On the same way, we have

$$M_n \xrightarrow[n\to\infty]{d} D,$$

where D is a r.v. with a degenerate distribution in x^F and $\xrightarrow[n\to\infty]{d}$ means convergence in distribution. Indeed,

$$F_{M_n}(x) = F^n(x) \xrightarrow[n\to\infty]{} \begin{cases} 0, & \text{if } x < x^F, \\ 1, & \text{if } x \geq x^F. \end{cases}$$

Therefore, M_n has a degenerate asymptotic distribution. So, in order to do some kind of inference, we need to have a non-degenerate behaviour for M_n. Then, as with the Central Limit Theorem, a normalization is required. This theorem is concerned with the asymptotic behaviour of the sequence of sums $X_1, X_1 + X_2, ..., \sum_{i=1}^n X_i, \ ...$, as $n \to \infty$:

Theorem 2.2 *Consider a sequence of i.i.d. r.v.'s,* $X_1, X_2, ..., X_n, ...$, *with* $E(X_i) = \mu$ *and* $Var(X_i) = \sigma^2 < +\infty$. *Therefore,*

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \quad \xrightarrow[n\to\infty]{d} \quad Z \frown \mathcal{N}(0,1).$$

In order to look for an appropriate non-degenerate limiting distribution for the sequence of sample maxima, we need a similar theorem; that is, we look for normalizing sequences $a_n > 0$ and b_n real such that

$$\frac{M_n - b_n}{a_n} \quad \xrightarrow[n\to\infty]{d} \quad W \frown G, \tag{2.1}$$

with W non-degenerate, i.e.,

$$\lim_{n\to\infty} P(M_n \leq a_n x + b_n) = \lim_{n\to\infty} F^n(a_n x + b_n) = G(x),$$

for every continuity point x of G.

2.2 The extremal limit problem

The first problem is to determine which d.f.'s G may appear on the limit in (2.1). These distributions are called *extreme value distributions*. In order to provide the answer to this question, we need to introduce two important concepts:

Definition 2.3. (max-domain of attraction) A d.f. F belongs to the max-domain of attraction of G, if there exist sequences $\{a_n > 0\}$ and $\{b_n\}$ real, such that

$$\lim_{n\to\infty} P(M_n \leq a_n x + b_n) = \lim_{n\to\infty} F^n(a_n x + b_n) = G(x), \tag{2.2}$$

for all the continuity points x of G and we write $F \in \mathcal{D}(G)$.

Definition 2.4. (distribution functions of the same type) Two d.f.'s F_1 and F_2 are said to be of the same type if there exist constants $a > 0$ and $b \in \mathbb{R}$ such that

$$F_2(ax + b) = F_1(x). \tag{2.3}$$

It means that F_1 and F_2 are the same, apart from location and scale parameters, i.e., they belong to the same location-scale family.

In addition to these concepts, we need to invoke the *Convergence to Types Theorem* from Khinchin:

Theorem 2.5 (**Convergence to Types Theorem - Khinchin**)

1. *Let W and $\widetilde{W}$ be two r.v.'s with non-degenerate d.f.'s G and $\widetilde{G}$, respectively. Suppose that $\{X_n\}_{n\in\mathbb{N}}$ is a sequence of r.v.'s with d.f. F_n and that we have real sequences $a_n, \widetilde{a}_n > 0$ and $b_n, \widetilde{b}_n \in \mathbb{R}$, such that $\frac{X_n - b_n}{a_n} \xrightarrow[n\to\infty]{d} W \frown G$ and $\frac{X_n - \widetilde{b}_n}{\widetilde{a}_n} \xrightarrow[n\to\infty]{d} \widetilde{W} \frown \widetilde{G}$. Then, there exist constants $A > 0$ and $B \in \mathbb{R}$ such that,*

$$\frac{\widetilde{a}_n}{a_n} \underset{n\to\infty}{\longrightarrow} A \qquad \text{and} \qquad \frac{\widetilde{b}_n - b_n}{a_n} \underset{n\to\infty}{\longrightarrow} B \tag{2.4}$$

and

$$\widetilde{G}(x) = G(Ax + B), \tag{2.5}$$

for every continuity point x of G and $\widetilde{G}$.

2. *Conversely,*

if $\frac{\widetilde{a}_n}{a_n} \underset{n\to\infty}{\longrightarrow} A$, $\frac{\widetilde{b}_n - b_n}{a_n} \underset{n\to\infty}{\longrightarrow} B$ *and* $\frac{X_n - b_n}{a_n} \xrightarrow[n\to\infty]{d} W \frown G$, *then*

$$\frac{X_n - \widetilde{b}_n}{\widetilde{a}_n} \xrightarrow[n\to\infty]{d} \frac{W - B}{A} \frown \widetilde{G},$$

with $\widetilde{G}(x) = G(Ax + B)$ and for every continuity point x of G and $\widetilde{G}$.

Important conclusions can be drawn from this Theorem for the sample maximum:

(i) According to (2.3) and (2.5), G and $\widetilde{G}$ are of the same type. Therefore, $\frac{M_n - b_n}{a_n}$ has an asymptotic distribution of the same type as $\frac{M_n - \widetilde{b}_n}{\widetilde{a}_n}$.

(ii) The choice of the sequences a_n and b_n is not unique: if we choose normalizing sequences a_n and b_n or $\widetilde{a}_n$ and $\widetilde{b}_n$ that are asymptotically equivalent, i.e. such that (2.4) holds, the d.f. F will belong to the max-domain of attraction of two d.f.'s of the same type. So, a d.f. F cannot be in the max-domain of attraction of two d.f.s of different types.

The problem of finding the extreme value distributions has been solved by Fisher and Tippett (1928), completed by Gnedenko (1943) and later revived and streamlined by de Haan (1970). They demonstrate that, if (2.2) holds, the limiting distribution G must be one of just three types. Formally,

Theorem 2.6 (**Asymptotic Distribution of the Sample Maximum, Fisher and Tippett, 1928, Gnedenko, 1943**) *If $F \in \mathcal{D}(G)$, the limiting d.f. G of the sample maximum, suitably normalized, is of the same type of one of the following distributions:*

(i) Type I: $G^{(I)}(x) = \Lambda(x) = \exp(-\exp(-x)), \quad x \in \mathbb{R};$

(ii) Type II: $G^{(II)}(x|\alpha) = \Phi_\alpha(x) = \begin{cases} 0, & x \leq 0, \\ \exp(-x^{-\alpha}), & x > 0, \alpha > 0; \end{cases}$

(iii) Type III: $G^{(III)}(x|\alpha) = \Psi_\alpha(x) = \begin{cases} \exp(-(-x)^\alpha), & x < 0, \alpha > 0, \\ 1, & x \geq 0; \end{cases}$

where the shape parameter α of $G^{(II)}$ and $G^{(III)}$ describes the tail's behaviour of the underlying d.f. F.

Theorem 2.6 shows us that, in contrast with the Central Limit Theorem, the limiting distribution is non-normal and depends on F only through its tail behaviour.

The three types can be generalized, with the introduction of a location (λ) and scale (δ) parameters:

$$\Lambda(x|\lambda,\delta) = \Lambda\left(\frac{x-\lambda}{\delta}\right); \quad \Phi_\alpha(x|\lambda,\delta) = \Phi_\alpha\left(\frac{x-\lambda}{\delta}\right); \quad \Psi_\alpha(x|\lambda,\delta) = \Psi_\alpha\left(\frac{x-\lambda}{\delta}\right), \tag{2.6}$$

for $\lambda \in \mathbb{R}, \delta > 0$.

Over the years, each type was labelled with a name, as a tribute to the work of their authors EVT. Therefore, $\Lambda(x|\lambda,\delta)$ is known as ***max-Gumbel-type d.f.***, $\Phi_\alpha(x|\lambda,\delta)$ as ***max-Fréchet-type d.f.*** and $\Psi_\alpha(x|\lambda,\delta)$ as ***max-Weibull-type d.f.*** The max-Weibull distribution must be distinguished from the classical and well-known Weibull distribution, used in survival analysis and in reliability theory, among other areas,

$$F(x|\lambda,\delta,\alpha) = 1 - \exp\left(-\left(\frac{x-\lambda}{\delta}\right)^\alpha\right). \tag{2.7}$$

The Weibull distribution was originally developed to address problems for minima arising in material sciences and this form of the distribution is commonly used in practice. For this reason, the max-Weibull distribution $\Psi_\alpha(x|\lambda,\delta)$, related to EVT, is commonly called the Reversed Weibull Distribution.

The three types of d.f.'s above seem unrelated. However, Jenkinson (1955) identified them as the only members of the following family:

$$G(x|\gamma) = G_\gamma(x) = \begin{cases} \exp\left(-(1+\gamma x)^{-\frac{1}{\gamma}}\right), & 1+\gamma x > 0 \quad \text{if } \gamma \neq 0, \\ \exp(-\exp(-x)), & x \in \mathbb{R} \quad \text{if } \gamma = 0, \end{cases} \tag{2.8}$$

where the shape parameter γ is known as the ***extreme value index*** (EVI).

The parametrization in (2.8) is due to von Mises (1936) and Jenkinson (1955) and is known as the ***Generalized Extreme Value distribution (GEVd)*** or the ***von Mises-Jenkinson family***, which unifies all possible non-degenerate weak limits of the maximum M_n:

(i) for $\gamma = 0$, taken as the continuity limit for $\gamma \to 0^+$ and for $\gamma \to 0^-, G_\gamma$ and $G^{(I)}$ are of the same type.
This case is proved calculating the limit of $G_\gamma(x)$ at $\gamma = 0$:

$$\lim_{\gamma\to 0^+} G_\gamma(x) = \lim_{\gamma\to 0^+} \exp\left(-(1+\gamma x)^{-\frac{1}{\gamma}}\right) = \exp\left(-\lim_{\gamma\to 0^+}(1+\gamma x)^{-\frac{1}{\gamma}}\right).$$

Now, define $\tau = \frac{1}{\gamma}$. For $\gamma \to 0^+$, we have $\tau \to +\infty$.

Using the well-known limit result $\lim_{k\to+\infty} \exp\left(1+\frac{x}{k}\right)^k = \exp(x)$, we have

$$\exp\left(-\lim_{\gamma\to 0^+}(1+\gamma x)^{-\frac{1}{\gamma}}\right) = \exp\left(-\lim_{\tau\to+\infty}\left(1+\frac{x}{\tau}\right)^{-\tau}\right) = \exp(-\exp(-x)).$$

We obtain the same result with $\lim_{\gamma\to 0^-} G_\gamma(x)$ and, therefore, $\lim_{\gamma\to 0} G_\gamma(x) = G^{(I)}(x)$. In order to grant the continuity of $G_\gamma(x)$ for $\gamma = 0$, me must then define $G_\gamma(x) = \exp(-\exp(-x))$.

(ii) for $\gamma > 0$ and taking $\gamma = \frac{1}{\alpha}, G_\gamma$ and $G^{(II)}$ are of the same type.
Indeed, following de Haan and Ferreira (2006) and (2.3), since $\gamma > 0$, we have:

$$\begin{aligned} G_\gamma\left(\frac{1}{\gamma}x - \frac{1}{\gamma}\right) &= \exp\left\{-\left(1+\gamma\left(\frac{1}{\gamma}x-\frac{1}{\gamma}\right)\right)^{-\frac{1}{\gamma}}\right\} \\ &= \exp\left(-(1+x-1)^{-\frac{1}{\gamma}}\right) \\ &= \exp\left(-x^{-\frac{1}{\gamma}}\right) \\ &= \exp(-x^{-\alpha}) \\ &= G^{(II)}(x|\alpha). \end{aligned}$$

(iii) for $\gamma < 0$ and taking $\gamma = -\frac{1}{\alpha}, G_\gamma$ and $G^{(III)}$ are of the same type.

Following the same references, but this time, with $\gamma < 0$, we have:

$$\begin{aligned} G_\gamma\left(-\frac{1}{\gamma}x - \frac{1}{\gamma}\right) &= \exp\left\{-\left(1+\gamma\left(-\frac{1}{\gamma}x-\frac{1}{\gamma}\right)\right)^{-\frac{1}{\gamma}}\right\} \\ &= \exp\left(-(1-x-1)^{-\frac{1}{\gamma}}\right) \\ &= \exp -\left(-x^{-\frac{1}{\gamma}}\right) \\ &= \exp(-(-x^{-\alpha})) \\ &= G^{(III)}(x|\alpha). \end{aligned}$$

This way, the GEVd is the unified version of the three types Λ, Φ and Ψ. So, taking up (2.1), we have

$$\frac{M_n - b_n}{a_n} \xrightarrow[n\to\infty]{d} W \frown G_\gamma \Longleftrightarrow F \in \mathcal{D}(G_\gamma).$$

As we did for the three types above, we can obtain a more general version of the GEVd, by incorporating a location parameter (λ) and a scale parameter (δ):

$$G_\gamma(x|\lambda,\delta) = G_\gamma\left(\frac{x-\lambda}{\delta}\right), \quad \lambda \in \mathbb{R}, \delta > 0. \tag{2.9}$$

The shape parameter γ (the EVI) is directly related with the right tail of the d.f. F: it determines the weight of the right tail of the underlying d.f. F. For this reason, the shape parameter is also known as ***tail index***.

Definition 2.7. Let F be a distribution function. We define the **right tail of a distribution function** as the following function:

$$\overline{F}(x) = P(X > x) = 1 - F(x).$$

which may be represented graphically in Figure 2.1

The tail index γ tells us how the tail function $\overline{F}(x)$ decays to zero as $x \to x^F$:

(i) For $\gamma = 0$, we are in the max-Gumbel-type domain of attraction, which contains exponential right-tailed distributions, with finite or infinite right endpoint x^F. In this case, all moments exist. The Gumbel domain contains distributions such as Normal, Exponential, Gamma, Lognormal and Gumbel itself;

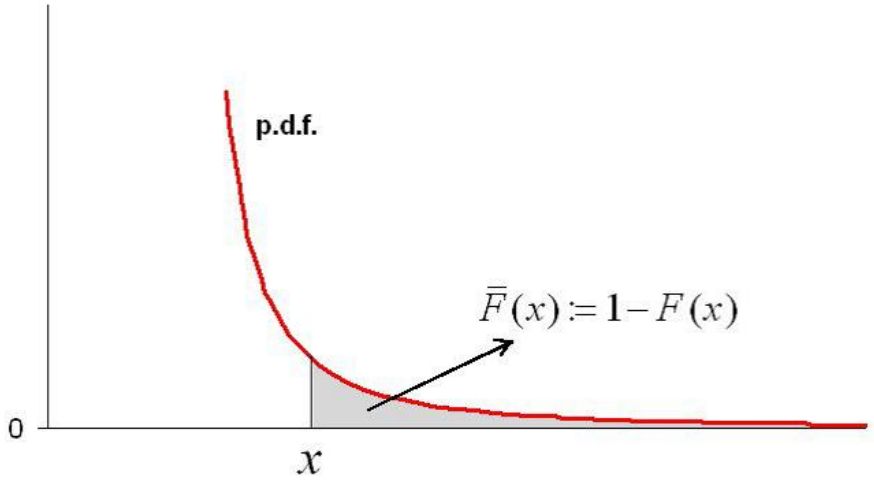

Figure 2.1: *Right tail of a distribution function*

(ii) For $\gamma > 0$, we are in the max-Fréchet-type domain of attraction, which contains heavy right-tailed distributions, with polynomial decay and infinite right endpoint x^F. Moments of order greater than or equal to $\frac{1}{\gamma}$ do not exist. The Fréchet domain contains distributions such as Pareto, Cauchy, Student's and Fréchet itself;

(iii) For $\gamma < 0$, we are in the max-Weibull-type domain of attraction, which contains light right-tailed distributions, with short decay and finite right endpoint x^F. The Weibull domain contains distributions such as Uniform, Beta and Weibull itself.

In their article, Fisher and Tippett (1928) deduced the three types of extreme value distributions by an ingenious and important argument: the maximum of a sample of size kn may be regarded as the largest element of a sample of k maxima obtained from samples of size n. So, consider k independent samples $(X_1, X_2, ..., X_n)$ of size n, extracted from a population with d.f. F, with $k \in \mathbb{N}$. For each sample $(X_1, X_2, ..., X_n)$, if the limiting distribution exists, we know that $\lim_{n\to\infty} F^n(a_n x + b_n) = G(x)$. Thus, for the k independent replicated samples, we have $\lim_{n\to\infty} F^{kn}(a_n x + b_n) = G^k(x)$. If we consider the whole sample of size kn, the same argument is valid and we have $\lim_{n\to\infty} F^{kn}(a_{kn} x + b_{kn}) = G(x)$, where new normalizing sequences $\{a_{kn} > 0\}$ and $\{b_{kn}\}$ real are chosen such that F^{kn} converges to G. But following Fisher and Tippett (1928), the two perspectives are equivalent, so the two limiting distributions must be of the same type and the normalizing constants must be asymptotically equivalent. Then, by the *Convergence to Types Theorem* (cf. Theorem 2.5), there exist $A_k > 0$ and $B_k \in \mathbb{R}$ such that,

$$\frac{a_{kn}}{a_n} \underset{n\to\infty}{\longrightarrow} A_k \quad , \quad \frac{b_{kn} - b_n}{a_n} \underset{n\to\infty}{\longrightarrow} B_k$$

and

$$G(x) = G^k(A_k x + B_k). \tag{2.10}$$

The functional equation (2.10) is known as ***stability equation*** and the solutions for $G(x)$ in this functional equation give all the possible limiting distributions for sequences of sample maxima suitably normalized. The distribution functions G satisfying (2.10) are called ***max-stable distribution functions***.

Fisher and Tippett (1928) determined all the possible solutions for this equation: the only solutions of this functional equation are precisely the types $G^{(I)}$, $G^{(II)}$ and $G^{(III)}$ from Theorem 2.6. This important result has the following meaning: the largest observation of a sample of k independent observations drawn from an extremal d.f. G must itself have G as limiting distribution, after a suitable normalization. It means that the three types $G^{(I)}$, $G^{(II)}$ and $G^{(III)}$ belong to their own max-domain of attraction. So, the limiting distribution G must be a *max-stable* distribution and the classes of extreme value and *max-stable* d.f.'s actually coincide.

The asymptotic theory for minima, $m_n = \min(X_1, X_2, \ldots, X_n)$, is a direct consequence from the EVT for maxima, as noted at the beginning of this chapter:

Theorem 2.8 (**Asymptotic Distribution of the Sample Minimum**) *Suppose there exist sequences $a_n^* > 0$ and b_n^* real, such that*

$$\frac{m_n - b_n^*}{a_n^*} \quad \xrightarrow[n\to\infty]{d} \quad W^* \frown G^*,$$

with G^* a non-degenerate d.f. Then, G^* must be one of the following types:

(i) Type I:* $G^{(I^*)}(x) = 1 - G^{(I)}(-x) = \Lambda^*(x) = 1 - \exp(-\exp(x)), \quad x \in \mathbb{R}$;

(ii) Type II:* $G^{(II^*)}(x|\alpha) = 1 - G^{(II)}(-x|\alpha) = \Phi_\alpha^*(x) = \begin{cases} 1 - \exp(x^{-\alpha}), & x < 0, \alpha > 0, \\ 1, & x \geq 0; \end{cases}$

(iii) Type III:* $G^{(III^*)}(x|\alpha) = 1 - G^{(III)}(-x|\alpha) = \Psi_\alpha^*(x) = \begin{cases} 0, & x \leq 0, \\ 1 - \exp(-x)^\alpha, & x > 0, \alpha > 0. \end{cases}$

As for the sample maximum, we can unify the three cases with the correspondent GEVd for minima:

$$G^*(x|\gamma) = G_\gamma^*(x) = 1 - G_\gamma(-x) = \begin{cases} 1 - \exp\left(-(1-\gamma x)^{-\frac{1}{\gamma}}\right), & 1 - \gamma x > 0 \quad \text{if } \gamma \neq 0, \\ 1 - \exp(-\exp(x)), & x \in \mathbb{R} \quad \text{if } \gamma = 0, \end{cases}$$

which includes the three *min-stable* types Λ^*, Φ^* and Ψ^* for $\gamma = 0$, $\gamma > 0$ and $\gamma < 0$, respectively, as for the maxima. Additionally, we can obtain more general forms of the distributions above, including a location parameter, $\lambda \in \mathbb{R}$, and a scale parameter, $\delta > 0$:

$$\Lambda^*(x|\lambda,\delta) = \Lambda^*\left(\frac{x-\lambda}{\delta}\right) \; ; \; \Phi^*_\alpha(x|\lambda,\delta) = \Phi^*_\alpha\left(\frac{x-\lambda}{\delta}\right) \; ; \; \Psi^*_\alpha(x|\lambda,\delta) = \Psi^*_\alpha\left(\frac{x-\lambda}{\delta}\right),$$

for the three types distributions and

$$G^*_\gamma(x|\lambda,\delta) = G^*_\gamma\left(\frac{x-\lambda}{\delta}\right), \quad \lambda \in \mathbb{R}, \delta > 0,$$

for the GEVd for minima. Λ^* is known as ***min-Gumbel-type d.f.***, Φ^* as ***min-Fréchet-type d.f.*** and Ψ^* as ***min-Weibull-type d.f.***

The min-Weibull distribution Ψ^* corresponds to the classical Weibull distribution mentioned in (2.7). The max-Gumbel and max-Fréchet distributions are commonly used in practice for maxima, as they were developed to deal with such problems. For this reason, the correspondent distributions for minima, Λ^* and Ψ^*, are known as *Reversed Gumbel* and *Reversed Fréchet* distributions.

2.3 The max-domain of attraction problem

Now that we have solved the first problem of determining which d.f. G may appear as a limiting distribution for a suitably normalized sequence of maxima, we have to solve another problem: assuming G as a possible limiting d.f. for the sequence $\frac{M_n - b_n}{a_n}$, what are the necessary and sufficient conditions that F must satisfy in order to belong to the max-domain of attraction of G? von Mises (1936) provided a set of conditions that ensures that F belongs to the domain of attraction of G. These conditions are known as ***von Mises' conditions***.

Theorem 2.9 (**von Mises' sufficient conditions for $F \in \mathcal{D}(G_\gamma)$, von Mises, 1936**) *Let F be an absolutely continuous d.f. Existing the probability density function (p.d.f.), $f(x) = F'(x)$, and the second derivative, $F''(x)$, let $h(x) = \frac{f(x)}{\overline{F}(x)}$ represent the **hazard function** or **hazard rate** from Reliability Theory.*

(i) Suppose $h(x) \neq 0$ and differentiable for x next to x^F (or for large x, if $x^F = \infty$). If

$$\lim_{x \to x^F} \frac{d}{dx}\left(\frac{1}{h(x)}\right) = 0,$$

then $F \in \mathcal{D}(G^{(I)})$.

(ii) Suppose $x^F = \infty$ *and* F' *exists. If, for some* $\gamma > 0$,

$$\lim_{x\to\infty} xh(x) = \frac{1}{\gamma} = \alpha,$$

then $F \in \mathcal{D}(G^{(II)})$.

(iii) Suppose $x^F < \infty$ *and* F' *exists for* $x < x^F$. *If, for some* $\gamma < 0$,

$$\lim_{x\to x^F} (x^F - x)h(x) = -\frac{1}{\gamma} = \alpha,$$

then $F \in \mathcal{D}(G^{(III)})$.

The proof of this theorem may be found in de Haan (1976).

These three conditions may be unified in a unique sufficient condition for F to belong to any of the only three max-domain of attraction, also derived in von Mises (1936).

Theorem 2.10 (**von Mises' sufficient conditions for** $F \in \mathcal{D}(G_\gamma)$) *Under the conditions of Theorem 2.9, if*

$$\lim_{x\to x^F} \left(\frac{1}{h(x)}\right)' = \gamma,$$

we have

$$F \in \mathcal{D}(G_\gamma).$$

Von Mises' conditions are very easy to check, requiring only the existence of the first or second derivative of F, but are only applicable to absolutely continuous d.f.'s F. Besides, they are only sufficient conditions, and not necessary.

We have to wait for Gnedenko (1943) for a set of necessary and sufficient conditions for maximal attraction to the three types of limit laws:

Theorem 2.11 (**Gnedenko's necessary and sufficient conditions for** $F \in \mathcal{D}(G_\gamma)$)

(i) $F \in \mathcal{D}(G^{(I)})$ *if and only if*

$$x^F \leq \infty \quad \textit{and} \quad \lim_{t\to x^F} \frac{\overline{F}\,(t + xg(t))}{\overline{F}(t)} = \exp(-x), \forall x \in \mathbb{R}, \tag{2.11}$$

where $g(t)$ *is a continuous and monotone positive function.*

(ii) $F \in \mathcal{D}(G^{(II)})$ *if and only if, for* $\gamma > 0$ *and* $\alpha = 1/\gamma$,

$$x^F = \infty \quad \textit{and} \quad \lim_{x\to\infty} \frac{\overline{F}(tx)}{\overline{F}(x)} = t^{-1/\gamma} = t^{-\alpha}, \forall t > 0. \tag{2.12}$$

(iii) $F \in \mathcal{D}(G^{(III)})$ *if and only if, for* $\gamma < 0$ *and* $\alpha = -1/\gamma$,

$$x^F < \infty \quad \textit{and} \quad \lim_{x \to 0^-} \frac{\overline{F}(x^F - tx)}{\overline{F}(x^F - x)} = t^{-1/\gamma} = t^{\alpha}, \forall t > 0.$$

According to Theorem 2.11, the max-Fréchet-type distribution Φ_α only attracts d.f.'s where $F(x) < 1, \forall x$, i.e., where $x^F = \infty$, and the max-Weibull-type distribution Ψ_α only attracts d.f.'s where $F(x^F) = 1$, for $x^F < \infty$, and $F(x) < 1, \forall x < x^F$. However, Gnedenko refers that the conditions for the Gumbel domain are neither definitive nor convenient for practical use. For this case, the von Mises' condition is better, but not necessary, as seen.

We must emphasize one important detail about Gnedenko's conditions: they do not grant the existence of a limiting distribution for the sequence of suitably normalized maxima. They only ensure that, *if this limiting distribution exists*, it must be one of the three types mentioned.

Laurens de Haan brought important improvements to the domain of attraction conditions. In order to examine such improvements, we have to define some concepts.

Definition 2.12. (Tail quantile function) Let F be a continuous function with inverse

$$F^{-1}(u) = \inf\{x : F(x) \geq u\}.$$

We define the ***tail quantile function*** as

$$U(t) = F^{-1}\left(1 - \frac{1}{t}\right), \qquad t \in [1, \infty[$$

or equivalently

$$U(t) = \left(\frac{1}{1 - F}\right)^{-1}(t), \qquad t \in [1, \infty[.$$

In particular, we have

(i) $U(t)$ is non-decreasing over the interval $[1, \infty[$;

(ii) $U(1) = \inf\{x : F(x) \geq 0\} = x_F$, where x_F stands for the left endpoint of the d.f. F;

(iii) $U(\infty) = \lim_{t \to +\infty} U(t) = \inf\{x : F(x) \geq 1\} = \sup\{x : F(x) < 1\} = x^F$.

Definition 2.13. (Regular Variation) A ultimately positive (for large x) and measurable function $f : \mathbb{R}^+ \to \mathbb{R}$ is said to be of **Regular Variation (RV)** (at infinity) with index α, if and only if, for some $\alpha \in \mathbb{R}$,

$$\lim_{t \to \infty} \frac{f(tx)}{f(t)} = x^{\alpha}, \quad x > 0. \tag{2.13}$$

We write $f \in \mathcal{RV}_\alpha$ and we call α the **index of regular variation**. A function satisfying (2.13) with $\alpha = 0$ is called slowly varying.

It may be useful to examine a more general class of functions, that represents a generalization of the $\mathcal{RV}_\alpha$ class of functions. For that, we need an additional Theorem that weakens the conditions of Definition 2.13 (see Appendix B of de Haan and Ferreira, 2006):

Theorem 2.14 *Suppose $f : \mathbb{R}^+ \to \mathbb{R}$ is measurable, eventually positive, and*

$$\lim_{t\to\infty} \frac{f(tx)}{f(t)}$$

exists, is finite and positive, for all x in a positive measurable set. Then $f \in \mathcal{RV}_\alpha$.

With this result, we can reformulate the RV property in a different way.

Definition 2.15. A measurable function $f : \mathbb{R}^+ \to \mathbb{R}$ is said to be of Regular Variation if it is possible to find a real function $a > 0$ such that the limit

$$\lim_{t\to\infty} \frac{f(tx)}{a(t)}$$

exists and is positive, for all $x > 0$.

The more general class of functions mentioned above may now be defined.

Definition 2.16. (Extended Regular Variation and Π-class) Suppose $f : \mathbb{R}^+ \to \mathbb{R}$ is measurable and there exists a real function $a > 0$ such that

$$\lim_{t\to\infty} \frac{f(tx) - f(t)}{a(t)} = \tau(x), \quad \forall x > 0, \tag{2.14}$$

where $\tau(.)$ is a non-constant function defined as

$$\tau(x) = \begin{cases} c\frac{x^\gamma - 1}{\gamma}, & \text{if } \gamma \neq 0, \\ c \log x, & \text{if } \gamma = 0, \end{cases}$$

with $c \neq 0$. Moreover, (2.14) holds if $a \in \mathcal{RV}_\gamma$.
It is also possible to incorporate the c constant into the a function. In particular, a measurable function $f : \mathbb{R}^+ \to \mathbb{R}$ is said to be of **extended regular variation** (ERV), if there exists a function $a > 0$ such that

$$\tau(x) = \begin{cases} \frac{x^\gamma - 1}{\gamma}, & \text{if } \gamma \neq 0, \\ \log x, & \text{if } \gamma = 0, \end{cases}$$

for all $x > 0$.
We write $f \in \mathcal{ERV}_\gamma$ and a is called an auxiliary function for f. For the case $\gamma = 0$, we say that f belongs to the **class** Π and write $f \in \Pi$ or $f \in \Pi(a)$. The results for functions satisfying (2.14) are similar to those satisfying (2.13).

We are now able to expose some improvements presented by Laurens de Haan. All the results and proofs may be found in de Haan and Ferreira (2006).

A first result is an alternative formulation of the limit relation (2.2):

$$\begin{aligned} \lim_{n\to\infty} F^n(a_n x + b_n) &= G(x) \Leftrightarrow \\ \Leftrightarrow \lim_{n\to\infty} n \log F(a_n x + b_n) &= \log G(x) \Leftrightarrow \\ \Leftrightarrow \lim_{n\to\infty} n(-\log F(a_n x + b_n)) &= -\log G(x). \end{aligned}$$

Considering the first order Mac Laurin's expansion of the function $f(x) = -\log(1-x)$, we get

$$-\log(1-x) \simeq x,$$

and rewriting $\log F(x) = \log\{1 - (1 - F(x))\}$, we have

$$\begin{aligned} \lim_{n\to\infty} n(-\log F(a_n x + b_n)) &= -\log G(x) \Leftrightarrow \\ \Leftrightarrow \lim_{n\to\infty} n\{-\log\{1 - (1 - F(a_n x + b_n))\}\} &= -\log G(x) \Leftrightarrow \\ \Leftrightarrow \lim_{n\to\infty} n(1 - F(a_n x + b_n)) &= -\log G(x) \Leftrightarrow \\ \Leftrightarrow \lim_{n\to\infty} \frac{1}{n(1 - F(a_n x + b_n))} &= -\frac{1}{\log G(x)}. \end{aligned}$$

Using now the tail quantile function from definition 2.12, we can write

$$\lim_{n\to\infty} \frac{1}{n} U^{-1}(a_n x + b_n) = -\frac{1}{\log G(x)}. \tag{2.15}$$

To simplify this identity, we need the following Lemma, from de Haan and Ferreira (2006):

Lemma 2.17 *Suppose f_n is a sequence of nondecreasing functions and g is a nondecreasing function. Suppose that for any x in some open interval $]a, b[$ that is a continuity point of g,*

$$\lim_{n\to\infty} f_n(x) = g(x).$$

Let f_n^{-1} and g^{-1} be the left-continuous inverses of f_n and g. Then, for each x in the interval $]g(a), g(b)[$ that is a continuity point of g^{-1}, we have

$$\lim_{n\to\infty} f_n^{-1}(x) = g^{-1}(x).$$

From this lemma, it follows that (2.15) can be written as

$$\lim_{n\to\infty} \frac{U(nx) - b_n}{a_n} = G^{-1}(\exp(-1/x)) = D(x), \quad \forall x > 0,$$

which can be written in a continuous version:

$$\lim_{t\to\infty} \frac{U(tx) - b(t)}{a(t)} = G^{-1}(\exp(-1/x)) = D(x), \quad \forall x > 0,$$

where $a(t) = a_{[t]}$ and $b(t) = b_{[t]}$ (with $[t]$ the integer part of t).

All the results about the limit relation (2.2) can now be summarized:

Theorem 2.18 (**de Haan and Ferreira, 2006, Theorem 1.1.2**) *Let $a_n > 0$ and b_n be two real sequences and consider a d.f. G, non-degenerate. The following statements are equivalent:*

(i)

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G(x), \tag{2.16}$$

for each continuity point x of G.

(ii)

$$\lim_{t\to\infty} t\{1 - F(a(t)x + b(t))\} = -\log G(x),$$

for each continuity point x of G for which $0 < G(x) < 1$, $a(t) = a_{[t]}$ and $b(t) = b_{[t]}$ (with $[t]$ the integer part of t).

(iii)

$$\lim_{t\to\infty} \frac{U(tx) - b(t)}{a(t)} = D(x), \tag{2.17}$$

for each continuity point $x > 0$ of $D(x) = G^{-1}(\exp(-1/x))$, $a(t) = a_{[t]}$ and $b(t) = b_{[t]}$.

From Section 2.2, we know the form of the function $G(x)$: if relation (2.16) holds, $G(x)$ is given by (2.8) and we have $G(x) = G_\gamma(x)$. We can obtain the inverse function of $G_\gamma^{-1}(x)$ easily:

$$G_\gamma^{-1}(x) = \begin{cases} \frac{1}{\gamma(-\log x)^\gamma} - \frac{1}{\gamma} & \text{if } \gamma \neq 0, \\ -\log(-\log x) & \text{if } \gamma = 0, \end{cases}$$

for $0 < x < 1$. Therefore, the function $D(x)$ of (2.17) may be calculated:

$$D(x) = G_\gamma^{-1}(\exp(-1/x)) = D_\gamma(x) = \begin{cases} \frac{1}{\gamma\{-\log(\exp(-1/x))\}^\gamma} - \frac{1}{\gamma} = \frac{x^\gamma - 1}{\gamma} & \text{if } \gamma \neq 0, \\ -\log\{-\log(\exp(-1/x))\} = \log x & \text{if } \gamma = 0, \end{cases} \tag{2.18}$$

for all $x > 0$.

We can now reformulate Theorem 2.18 with the knowledge of G.

Theorem 2.19 (**de Haan and Ferreira, 2006, Theorem 1.1.6**) *Let be* $\gamma \in \mathbb{R}$ *and consider the d.f.* G_γ *in* (2.8). *The following statements are equivalent:*

i. There exist two real constants $a_n > 0$ *and* b_n, *such that*

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G_\gamma(x), \tag{2.19}$$

for each continuity point x *of* G_γ. *Recall that, if this relation holds, we say that the d.f.* F *belongs to the max-domain of attraction of* G_γ;

ii. There exists a positive function a *such that, for* $x > 0$

$$\lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} = D_\gamma(x) = \frac{x^\gamma - 1}{\gamma}, \tag{2.20}$$

where, for $\gamma = 0$, *the right-hand side is interpreted as* $\log x$;

iii. There exists a positive function a *such that*

$$\lim_{t\to\infty} t\{1 - F(a(t)x + U(t))\} = -\log G(x) = (1+\gamma x)^{-1/\gamma},$$

for each continuity point x *of* G, *where* $1 + \gamma x > 0$;

iv. There exists a positive function g *such that*

$$\lim_{t\to x^F} \frac{\overline{F}(t + xg(t))}{\overline{F}(t)} = (1+\gamma x)^{-1/\gamma}, \tag{2.21}$$

for each x, *where* $1+\gamma x > 0$. *Moreover,* (2.19) *holds with* $b_n = U(n)$ *and* $a_n = a(n)$. *Also,* (2.21) *holds with* $g(t) = a\left(\frac{1}{\overline{F}(t)}\right)$.

Taking a closer look at condition (2.20), we notice we can relate it with condition (2.17). Indeed, from relation (2.17), we have

$$\lim_{t\to\infty} \frac{U(tx) - b(t)}{a(t)} = D(x),$$

for $D(x) = D_\gamma(x)$, defined in (2.18). Therefore,

$$\begin{aligned}
\lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} &= \lim_{t\to\infty} \frac{U(tx) - b(t) + b(t) - U(t)}{a(t)} \\
&= \lim_{t\to\infty} \frac{U(tx) - b(t)}{a(t)} - \frac{U(t) - b(t)}{a(t)} \\
&= D_\gamma(x) - D_\gamma(1).
\end{aligned}$$

From (2.18), we have $D_\gamma(1) = 0$. Then

$$\lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} = D_\gamma(x).$$

Another interesting point about condition (2.20) is its correspondence with our definition of $\mathcal{ERV}_\gamma$ class of functions in (2.14). So, a necessary and sufficient condition for $F \in \mathcal{D}(G_\gamma)$ is that $U \in \mathcal{ERV}_\gamma$. This condition is known as *first order extended regular variation property* (de Haan, 1984) and we will refer to this condition as *first order condition.*

Laurens de Haan also worked on the von Mises' sufficient conditions, using a formulation with the tail quantile function, $U(t)$:

Theorem 2.20 *Under the conditions of Theorem 2.9, if*

$$\lim_{x \to x^F} \frac{1}{h'(x)} = \gamma$$

or equivalently

$$\lim_{x \to x^F} \frac{1}{h(x)} \frac{f'(x)}{f(x)} = -\gamma - 1 \tag{2.22}$$

or, using the tail quantile function,

$$\lim_{t \to \infty} \frac{tU''(t)}{U'(t)} = \gamma - 1,$$

then

$$F \in \mathcal{D}(G_\gamma).$$

The sufficient conditions for the three types were also reformulated by Laurens de Haan:

Theorem 2.21 *Under the conditions of Theorem 2.9,*

(i) If

$$\lim_{x \to x^F} \frac{1}{h(x)} \frac{f'(x)}{f(x)} = -1, \tag{2.23}$$

then, $F \in \mathcal{D}(G^{(I)})$.

(ii) If

$$\lim_{t \to \infty} \frac{tU'(t)}{U(t)} = \gamma,$$

then, $F \in \mathcal{D}(G^{(II)})$.

(iii) If

$$\lim_{t \to \infty} \frac{tU'(t)}{U(\infty) - U(t)} = -\gamma,$$

then, $F \in \mathcal{D}(G^{(III)})$.

Condition (2.23) is obtained from (2.22) with $\gamma = 0$.

Finally, it is important to refer that the alternative versions of Gnedenko's necessary and sufficient conditions, created once again by Laurens de Haan, brought some completion to Gnedenko's work. Theorem 2.11 can be reformulated in a more uniform way, using statement (2.21) as a definition of max-domain of attraction:

Theorem 2.22 *The d.f. F is in the max-domain of attraction of the extreme value distribution G_γ if and only if, for some positive function g,*

$$\lim_{t \to x^F} \frac{\overline{F}(t + xg(t))}{\overline{F}(t)} = (1 + \gamma x)^{-1/\gamma}, \tag{2.24}$$

for each x, where $1 + \gamma x > 0$. If (2.24) holds for some $g > 0$, then it also holds with

$$g(t) = \begin{cases} \frac{\int_t^{x^F} \overline{F}(s)ds}{\overline{F}(t)} = E(X - t | X > t), & \text{if } \gamma = 0, \\ \gamma t, & \text{if } \gamma > 0, \\ -\gamma(x^F - t), & \text{if } \gamma < 0. \end{cases} \tag{2.25}$$

As we can see, Laurens de Haan suggests a function g for the Gumbel-domain of attraction, in condition (2.11) of Theorem 2.11:

$$g(t) = \frac{\int_t^{x^F} \overline{F}(s)ds}{\overline{F}(t)} = E(X - t | X > t), \text{ for } t < x^F \text{ and } \int_t^{x^F} \overline{F}(s)ds < \infty. \tag{2.26}$$

In the reliability literature, this particular function g is known as *mean residual life* or *mean excess function*.

The necessary and sufficient conditions were also rewritten using the tail quantile function U. But before presenting the reformulated version of Gnedenko's necessary and sufficient conditions, it is pertinent to introduce a useful Lemma, using also the concepts defined in Definitions 2.13 and 2.16:

Lemma 2.23 *If $F \in \mathcal{D}(G_\gamma)$ or equivalently, if $U \in \mathcal{ERV}_\gamma$, for $\gamma \in \mathbb{R}$, we can state the following conditions on the tail of the underlying d.f. F:*

(i) For $\gamma = 0$, or identically, for $U \in \Pi$,

$$\lim_{t \to \infty} \frac{U(tx)}{U(t)} = 1 \Longleftrightarrow U \in \mathcal{RV}_0, \text{ for all } x > 0, \quad \text{and} \quad \lim_{t \to \infty} \frac{a(t)}{U(t)} = 0.$$

Moreover, if $U(\infty) < \infty$,

$$\lim_{t \to \infty} \frac{U(\infty) - U(tx)}{U(\infty) - U(t)} = 1 \Longleftrightarrow U(\infty) - U \in \mathcal{RV}_0, \text{ for all } x > 0,$$

and

$$\lim_{t\to\infty} \frac{a(t)}{U(\infty) - U(t)} = 0.$$

The function a is then a slowly varying function, i.e.,

$$\lim_{t\to\infty} \frac{a(tx)}{a(t)} = 1 \Longleftrightarrow a \in \mathcal{RV}_0, \text{ for all } x > 0.$$

(ii) For $\gamma > 0$, we have

$$U(\infty) = \infty \quad \text{and} \quad \lim_{t\to\infty} \frac{U(t)}{a(t)} = \frac{1}{\gamma}.$$

(iii) For $\gamma < 0$, we have

$$U(\infty) < \infty \quad \text{and} \quad \lim_{t\to\infty} \frac{U(\infty) - U(t)}{a(t)} = -\frac{1}{\gamma}.$$

In particular, this implies that $\lim_{t\to\infty} a(t) = 0$.

We can still state the following:

1. For $\gamma > 0$, relation (2.20) is equivalent to

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^\gamma \Longleftrightarrow U \in \mathcal{RV}_\gamma, \text{ for all } x > 0.$$

2. For $\gamma < 0$, relation (2.20) is equivalent to $U(\infty) < \infty$ and

$$\lim_{t\to\infty} \frac{U(\infty) - U(tx)}{U(\infty) - U(t)} = x^\gamma \Longleftrightarrow U(\infty) - U \in \mathcal{RV}_\gamma, \text{ for all } x > 0.$$

The proofs may be found in de Haan (1976) and de Haan and Ferreira (2006).

With this Lemma, we can now reformulate Gnedenko's conditions for $\gamma > 0$ and $\gamma < 0$:

Theorem 2.24 *$F \in \mathcal{D}(G_\gamma)$, for $\gamma \neq 0$, if and only if*

1. for $\gamma > 0$ and for $x > 0$,

$$\lim_{t\to\infty} \frac{U(tx)}{U(t)} = x^\gamma \Longleftrightarrow U \in \mathcal{RV}_\gamma;$$

2. for $\gamma < 0$ and for $x > 0$,

$$\lim_{t\to\infty} \frac{U(\infty) - U(tx)}{U(\infty) - U(t)} = x^\gamma \Longleftrightarrow U(\infty) - U \in \mathcal{RV}_\gamma.$$

The proofs may also be found in de Haan (1976) and de Haan and Ferreira (2006).

2.4 The choice of the normalizing sequences a_n and b_n

The last problem that needs to be solved is the choice of suitable normalizing sequences a_n and b_n for the basic limit relation (2.2). As we have seen in Theorem 2.5, this choice is not unique. Moreover, the choice of such sequences depends on the G function that appears on the limit. The most common choices are indicated in the following Theorem:

Theorem 2.25 (**Normalizing constants for $F \in \mathcal{D}(G^{(I)})$, $F \in \mathcal{D}(G^{(II)})$ and $F \in \mathcal{D}(G^{(III)})$, Gnedenko, 1943, and de Haan and Ferreira, 2006**) *If $F \in \mathcal{D}(G_\gamma)$, then,*

(i) for $\gamma = 0$,

$$\lim_{n\to\infty} F^n(a_n x + b_n) = \exp(-\exp(-x)) = \Lambda(x)$$

holds for all $x \in \mathbb{R}$, with

$$a_n = F^{-1}\left(1 - \frac{1}{ne}\right) - F^{-1}\left(1 - \frac{1}{n}\right) = U(ne) - U(n) \quad \text{or} \quad a_n = g(U(n))$$

and

$$b_n = F^{-1}\left(1 - \frac{1}{n}\right) = U(n),$$

with g defined as in (2.25) for $\gamma = 0$;

(ii) for $\gamma > 0$,

$$\lim_{n\to\infty} F^n(a_n x + b_n) = \exp\left(-x^{-\frac{1}{\gamma}}\right) = \Phi_{1/\gamma}(x)$$

holds for all $x > 0$, with

$$a_n = F^{-1}\left(1 - \frac{1}{n}\right) = U(n)$$

and

$$b_n = 0;$$

(iii) for $\gamma < 0$,

$$\lim_{n\to\infty} F^n(a_n x + b_n) = \exp\left(-(-x)^{-\frac{1}{\gamma}}\right) = \Psi_{-1/\gamma}(x)$$

holds for all $x < 0$, with

$$a_n = x^F - F^{-1}\left(1 - \frac{1}{n}\right) = x^F - U(n)$$

and

$$b_n = x^F.$$

Von Mises's sufficient conditions also include normalizing sequences a_n and b_n for the general case $F \in \mathcal{D}(G_\gamma)$, rewritten by Laurens de Haan, with the tail quantile function $U(t)$.

Theorem 2.26 (**Normalizing constants for** $F \in \mathcal{D}(G_\gamma)$) *Under the conditions of Theorem 2.20, we have*

$$b_n = F^{-1}\left(1 - \frac{1}{n}\right) = U(n)$$

and

$$a_n = \frac{1}{h(b_n)} = \frac{1}{nf(b_n)} = nU'(n).$$

It must be emphasized that these latter sequences are distinct from the normalizing sequences presented in Theorem 2.25. They are used to normalize the sample maximum M_n, when F belongs to the domain of attraction of the GEVd. However, the sequences presented in Theorem 2.25 are used to normalize the sample maximum when F belongs to the domain of attraction of one of the three standard types of Theorem 2.6.

Chapter 3

Estimation of parameters in Extreme Value Theory

Estimating the GEVd parameters $(\gamma, \lambda, \delta)$ constitutes an important task in EVT, since it is a starting point for statistical inference about extreme values of a population. In particular, the EVI, the shape parameter γ in (2.9), measures the right tail's weight of the underlying d.f. F, allowing us to understand and describe the behaviour of the extreme values of a population. With the estimated EVI, it is possible to estimate other parameters of extreme events like the *right endpoint* of the underlying d.f. F, *extremes quantiles*, the *return period* and the *probability of exceedance* of a high level. We can follow basically two approaches in order to obtain estimates for the GEVd parameters, a ***parametric approach*** and a ***semi-parametric approach***.

3.1 Parametric approaches

3.1.1 Introduction

Following a parametric approach, the main assumption is that we can use the limiting distribution of the sample extremes as an exact distribution that can be fitted to data, i.e., the data in hand form an i.i.d sample coming from an "exact" GEVd defined in (2.9). As already said, the focus of this thesis is on the sample maximum, $M_n = \max(X_1, \ldots, X_n)$, and in this chapter, we will continue with the same line. As we are dealing with extreme events, in particular with maxima, every kind of inference must be carefully done, because of the extremal nature of such events and the serious consequences of a misleading inference. Specifically, EVT usually deals with estimating the probability of severe shocks that are more extreme than any other that has been observed until now.

In Chapter 2, we saw that the right tail of the underlying d.f. F governs the max-domain of attraction. Then, we must focus only on the behaviour of the high-order statistics and, consequently, the remaining data are not so essential. So, in order to perform a correct inference about extreme events from the available data, it is important to consider which of them are considered extremes, under some criterion. Parametric approaches are distinguished according to how many specific observations are picked up among the available sample data. Within EVT framework, there are two primary approaches to select such observations: ***the Gumbel's approach*** or ***the Block Maxima method*** and ***the Peaks over Threshold method*** (shortly, POT). The first approach chooses the largest observation from successive periods, defining an appropriate length of the periods as the blocks; the second approach focuses on the observations that exceed a fixed (high) threshold.

3.1.2 The Gumbel's approach or Block Maxima method

Under this approach, the sample of size n is divided into m sub-samples of size k or m blocks of size k, with $n = m \times k$ and k sufficiently large. Usually, a block corresponds to a period of one year, with k observations per year. Thus, this method is also called ***the Annual Maxima method***. In each block, the largest observation is selected, so that we obtain a sample of m independent sample maxima. Formally, let Y be the r.v. that represents the maximum of a block of size k, i.e.

$$Y \equiv M_k = \max(X_1, \ldots, X_k). \tag{3.1}$$

Considering now m blocks, we obtain a collection of m sample maxima, $(Y_1, \ldots, Y_m)$. Consequently, we can fit a parametric GEVd given by (2.9) and obtain estimates of the EVI (γ), as well as the location (λ) and the scale (δ) parameters, which replace the attraction coefficients b_n and a_n in (2.1), respectively. This way, the Gumbel's method can be used whenever our dataset consists of independent samples maxima. Several estimation methods for the GEVd are available in the literature; the most popular are *the Maximum Likelihood* (ML) method and *the Probability Weighted Moments* (PWM) method. With the parameters estimates in hand, we can then obtain estimates of the other parameters mentioned at the beginning of this Chapter and make some inference with confidence intervals.

3.1.2.1 Maximum Likelihood Estimation

Since we are in a traditional parametric estimation environment, the first method we bear in mind is obviously the ML method. Let $(Y_1, \ldots, Y_m)$ be a random sample of the

r.v. Y defined in (3.1), taken from a GEVd. For $\gamma \neq 0$, the log-likelihood function of an observed random sample $(y_1, \ldots, y_m)$ is given by

$$\ell(\gamma, \lambda, \delta | y_1, \ldots, y_m) = -m \log \delta - \left(\frac{1}{\gamma} + 1\right) \sum_{i=1}^{m} \log \left(1 + \gamma \frac{y_i - \lambda}{\delta}\right) - \sum_{i=1}^{m} \left(1 + \gamma \frac{y_i - \lambda}{\delta}\right)^{-\frac{1}{\gamma}}, \tag{3.2}$$

for $1 + \gamma \frac{y_i - \lambda}{\delta} > 0, i = 1, \ldots, m$.

When $\gamma = 0$, the log-likelihood function reduces to

$$\ell(0, \lambda, \delta | y_1, \ldots, y_m) = -m \log \delta - \sum_{i=1}^{m} \exp\left(-\frac{y_i - \lambda}{\delta}\right) - \sum_{i=1}^{m} \frac{y_i - \lambda}{\delta}. \tag{3.3}$$

The ML estimator $(\hat{\gamma}, \hat{\lambda}, \hat{\delta})$ for the unknown parameters $(\gamma, \lambda, \delta)$ is obtained by maximization of (3.2) or equivalently (3.3). Differentiating these expressions, we obtain the likelihood system of equations, which has no explicit solution. Thus, such a system must be solved iteratively, by numerical methods. Details on computational methods can be found in Prescott and Walden (1980), Prescott and Walden (1983), Hosking (1985), Smith (1985) and Macleod (1989).

Despite its attractiveness, the ML method can have some problems in EVT. The asymptotic properties of ML estimators are valid only under the known regularity conditions of ML theory (Cox and Hinkley, 1974). However, as the GEVd support depends on unknown parameters, the regularity conditions are not satisfied and, although we have reliable numerical processes to find ML estimates, they lack for the usual asymptotic properties of ML estimators. Smith (1985) showed that the usual asymptotic properties of ML estimators for the GEVd depend on the value of the unknown EVI (γ). He proved that the ML estimator exists for $\gamma > -1$, but was only able to demonstrate that the classical asymptotic properties of consistency and asymptotic normality hold for $\gamma > -0.5$, letting the case $-1 < \gamma \leq -0.5$ unsolved. In particular, the author stated that, for $\gamma > -0.5$, we have

$$\sqrt{m}((\hat{\gamma}, \hat{\lambda}, \hat{\delta}) - (\gamma; \lambda; \delta)) \quad \xrightarrow[m \to \infty]{d} \quad Z \frown \mathcal{N}(0, \mathcal{I}^{-1}),$$

where $\mathcal{I}^{-1}$ is the inverse of the Fisher Information matrix.

For $\gamma \leq -1$, the ML procedure is not applicable, since the log-likelihood function has no local maximum. The density is J-shaped and the corresponding log-likelihood function always tends to $+\infty$ along some path in the log-likelihood space. Many authors, defending ML estimation, state that this disadvantage is of little practical importance, since distributions with $\gamma \leq -1$ have a very light upper tail, situation which is rarely encountered in typical cases of EVT. More recently, Zhou (2009) and Zhou (2010) solved

the open problem let by Smith (1985), proving that the ML estimators fulfill the two asymptotic properties mentioned for $\gamma > -1$.

Another problem with ML estimation is the convergence of the iterative process of maximization: sometimes, the computational process does not converge and we may not be able to find a suitable estimator.

Despite its drawbacks, ML estimation can handle with missing data, non-stationarity and temporal dependence, with only a few modifications, while this task is very tedious or even impossible with other estimation methods.

3.1.2.2 Probability Weighted Moments Estimation

The PWM method is also very popular for fitting the GEVd to the available data sample, $(y_1, \ldots, y_m)$. This method is a generalization of the usual method of moments, but with an increasing weight for tail observations, and its application to GEVd is explained in detail by Hosking et al. (1985). Generally speaking, the PWM of a r.v. X with d.f. F are given by the quantities

$$M_{p,r,s} = E\{X^p(F(X))^r(1-F(X))^s\}, \quad \text{for } p,r,s \in \mathbb{R}. \tag{3.4}$$

Consider now a sample $(Y_1, \ldots, Y_m)$ of GEVd i.i.d. r.v.'s. In this context, we can use PWM with the form

$$M_{1,r,0} = E\{Y(F(Y))^r\}, \quad \text{for } r = 0, 1, \ldots. \tag{3.5}$$

For $\gamma \neq 0$, Hosking et al. (1985) derived the quantity in (3.5), with $F(y) = G_\gamma(y|\lambda,\delta)$:

$$M_{1,r,0} = \frac{1}{r+1}\{\lambda - \frac{\delta}{\gamma}(1-(r+1)^\gamma\Gamma(1-\gamma))\}, \quad \text{for } \gamma < 1, \tag{3.6}$$

where $\Gamma(.)$ stands for the *gamma function*, $\Gamma(t) = \int_0^{+\infty} x^{t-1}e^{-x}dx,\ t \geq 0$.

As we can see, the PWM for the GEVd exist only for $\gamma < 1$, but, according to Hosking et al. (1985), this is not a problem, because in hydrology, the field they work on, the EVI lies generally between $-\frac{1}{2}$ and $\frac{1}{2}$. Thus, the PWM estimator $(\hat{\gamma}, \hat{\lambda}, \hat{\delta})$ for the unknown parameters $(\gamma, \lambda, \delta)$ is obtained solving the following system of equations resulting from (3.6), with $r = 0, 1, 2$:

$$\begin{cases} M_{1,0,0} = \lambda - \frac{\delta}{\gamma}(1-\Gamma(1-\gamma)), \\ 2M_{1,1,0} - M_{1,0,0} = \frac{\delta}{\gamma}\Gamma(1-\gamma)(2^\gamma - 1), \\ \frac{3M_{1,2,0}-M_{1,0,0}}{2M_{1,1,0}-M_{1,0,0}} = \frac{3^\gamma-1}{2^\gamma-1}, \end{cases}$$

leading to

$$\begin{cases} \lambda = M_{1,0,0} + \frac{\delta}{\gamma}(1 - \Gamma(1-\gamma)), \\ \delta = \frac{\gamma(2M_{1,1,0} - M_{1,0,0})}{\Gamma(1-\gamma)(2^{\gamma}-1)}, \\ \frac{3M_{1,2,0} - M_{1,0,0}}{2M_{1,1,0} - M_{1,0,0}} = \frac{3^{\gamma}-1}{2^{\gamma}-1}. \end{cases} \tag{3.7}$$

For the r.v. $Y \frown GEVd$, Landwehr et al. (1979) demonstrate that an unbiased estimator for $M_{1,r,0}$ is given by

$$\widehat{M}_{1,r,0} = \frac{1}{m} \sum_{i=1}^{m} \left(\prod_{j=1}^{r} \frac{(i-j)}{(m-j)} \right) Y_{i:m}. \tag{3.8}$$

Thus, replacing $M_{1,r,0}$ in (3.7) by its unbiased estimator given by (3.8), for $r = 0, 1, 2$, we obtain the PWM estimator, $(\hat{\gamma}, \hat{\lambda}, \hat{\delta})$:

$$\begin{cases} \hat{\lambda} = \widehat{M}_{1,0,0} + \frac{\hat{\delta}}{\hat{\gamma}}(1 - \Gamma(1-\hat{\gamma})), \\ \hat{\delta} = \frac{\hat{\gamma}(2\widehat{M}_{1,1,0} - \widehat{M}_{1,0,0})}{\Gamma(1-\hat{\gamma})(2^{\hat{\gamma}}-1)}, \\ \frac{3\widehat{M}_{1,2,0} - \widehat{M}_{1,0,0}}{2\widehat{M}_{1,1,0} - \widehat{M}_{1,0,0}} = \frac{3^{\hat{\gamma}}-1}{2^{\hat{\gamma}}-1}. \end{cases} \tag{3.9}$$

Note that, to obtain $\hat{\gamma}$, the last equation of (3.9) has to be solved numerically.

For the Gumbel-case, $\gamma = 0$, the PWM defined in (3.5) are given by

$$M_{1,r,0} = \frac{1}{r+1}\{\lambda + \delta(\log(1+r) - \psi(1))\}, \tag{3.10}$$

where $\psi(x) = \frac{d}{dx} \log \Gamma(x)$ stands for the *digamma function*, with $\Gamma(x)$ defined in (3.6), giving in particular $-\psi(1) \simeq 0.57721567$, the Euler-Mascheroni constant.

As for the case $\gamma \neq 0$, the PWM estimator $(\hat{\lambda}, \hat{\delta})$ for the unknown parameters (λ, δ) results from the following system of equations, obtained from (3.10) for $r = 0, 1$:

$$\begin{cases} \lambda = M_{1,0,0} + \delta\psi(1), \\ \delta = \frac{2M_{1,1,0} - M_{1,0,0}}{\log 2}. \end{cases} \tag{3.11}$$

Following the same reasoning and replacing $M_{1,r,0}$ in (3.11) by its unbiased estimator defined in (3.8) for $r = 0, 1$, the PWM estimator results in

$$\begin{cases} \hat{\lambda} = \widehat{M}_{1,0,0} + \hat{\delta}\psi(1), \\ \hat{\delta} = \frac{2\widehat{M}_{1,1,0} - \widehat{M}_{1,0,0}}{\log 2}. \end{cases}$$

As for the ML estimator, the PWM estimator is asymptotically Normal, after a suitable normalization. For details, see Hosking et al. (1985).

Despite the popularity and attractiveness of ML estimation due to its asymptotic properties, this method looses its primacy in a small-sample context, where it is outperformed by the PWM estimation. Indeed, Hosking et al. (1985) showed that, in small samples, the PWM estimator has lower variance than the ML estimator. The weak performance of this latter estimator in small samples, making it a poor competitor to the PWM estimator, remains a serious criticism, since it is not unusual to make inference about extremes with only a few data, if we are dealing with rare extreme events. Nevertheless, ML is the preferred parametric estimation method because of its flexibility of modification, to incorporate more complex problems, as mentioned above. On the opposite, PWM estimation has a serious difficulty in dealing with more complex structures.

3.1.2.3 Estimation of other parameters of extreme events

Let $\mathcal{Y}_{1-p}$ be the **extreme quantile** of order $(1-p)$ of the GEVd underlying the r.v. Y, defined in (3.1), with p sufficiently small. Estimates of extreme quantiles can be obtained inverting the GEVd. Straightforward calculations lead us to the following expression:

$$\mathcal{Y}_{1-p} = G_\gamma^{-1}(1-p|\lambda,\delta) = \begin{cases} \lambda + \frac{\delta}{\gamma}\{(-\log(1-p))^{-\gamma} - 1\}, & \text{if } \gamma \neq 0, \\ \lambda - \delta \log(-\log(1-p)), & \text{if } \gamma = 0, \end{cases} \tag{3.12}$$

where we can replace $(\gamma, \lambda, \delta)$ by its corresponding ML or PWM estimator.

We can express extreme quantiles in terms of the quantile function, given in Definition 2.12, defining $t = \frac{1}{p}$:

$$U_{G_\gamma}\left(\frac{1}{p}\right) = G_\gamma^{-1}(1-p|\lambda,\delta) = \mathcal{Y}_{1-p}. \tag{3.13}$$

In particular, dealing with Annual Maxima, we define the **T-period level**, $U(T)$, as the (high) level exceeded on average by the r.v. Y, for every period of length T. For this reason, T is called the **return period** of the level u:

$$T = \frac{1}{1 - G_\gamma(u|\lambda,\delta)} = \frac{1}{P(Y > u)}, \tag{3.14}$$

where the denominator is called **exceedance probability**.

The return period T can be seen as the mean value of a geometric r.v. Let N_u be the number of periods needed to exceed the level u for the first time. So, N_u is a geometric r.v. with mean value $\frac{1}{p_u}$, with $p_u = P(N_u > u)$. Therefore, we have $T = E(N_u)$.

If our objective is to obtain the quantiles of the original data X and make some inference about the population underlying the m blocks of observations $X_1, \ldots, X_k$, with $X \frown F$, we know that the r.v. Y defined in (3.1) is distributed as

$$F_Y \equiv F_{M_k} = F^k \simeq G_\gamma, \tag{3.15}$$

where k is the block length. As for the r.v. Y, we denote by $\mathcal{X}_{1-p}$ the quantile of order $(1-p)$ of the d.f. F underlying the r.v. X, i.e., a quantity such that $F(\mathcal{X}_{1-p}) = 1-p$. Using now relation (3.15), we have:

$$F^k(\mathcal{X}_{1-p}) = (1-p)^k \simeq G_\gamma(\mathcal{X}_{1-p}|\lambda, \delta).$$

Then,

$$\mathcal{X}_{1-p} \simeq G_\gamma^{-1}((1-p)^k|\lambda,\delta) = \begin{cases} \lambda + \frac{\delta}{\gamma}\{(-\log(1-p)^k)^{-\gamma} - 1\}, & \text{if } \gamma \neq 0, \\ \lambda - \delta \log(-\log(1-p)^k), & \text{if } \gamma = 0. \end{cases} \tag{3.16}$$

Again, estimates of (3.14) and (3.16) are obtained replacing $(\gamma, \lambda, \delta)$ by its ML or PWM estimator.

From Chapter 2, we know that, if $\gamma < 0$, the right endpoint of the GEVd is finite. Therefore, we can estimate the finite right endpoint of the respective underlying d.f. F by

$$x^F = U_{G_\gamma}(\infty) = \mathcal{Y}_1 = \mathcal{X}_1 = \lambda - \frac{\delta}{\gamma}, \tag{3.17}$$

replacing again $(\gamma, \lambda, \delta)$ by its ML or PWM estimator.

3.1.2.4 Inference: confidence intervals for the extreme value index

Since ML and PWM estimators are asymptotically normal, we can construct Confidence Intervals (CI's) and make other forms of inference on the GEVd parameters $(\gamma, \lambda, \delta)$. For instance, denoting by θ any of the three parameters of the GEVd, we can obtain a 95% CI for θ by the traditional way:

$$\hat{\theta} \pm 1.96\sqrt{\frac{\hat{v}_{\hat{\theta}}}{m}},$$

where $\hat{\theta}$ is the ML or PWM estimate of θ and $\hat{v}_{\hat{\theta}}$ denotes the respective diagonal element of the covariance-matrix of the limiting Normal distribution, replacing the unknown parameters by their estimates.

However, Beirlant et al. (2004) suggest another way to construct such CI, since using the Normal distribution as an approximation to the true sampling distribution of the ML

or PWM estimators may result in a poor inference. To improve the quality of the CI, it is preferable to use the ***profile likelihood function***.

Definition 3.1. (**profile likelihood function**) The profile likelihood function for γ, $\mathcal{L}_p(\gamma)$, is given by

$$\mathcal{L}_p(\gamma) = \max_{\lambda,\delta|\gamma} \mathcal{L}(\lambda, \delta|\gamma, y_1, \ldots, y_m),$$

i.e., for each value of γ, the profile likelihood function gives the maximized likelihood function in order to the other parameters, λ and δ.

The profile likelihood ratio statistic

$$\mathfrak{L}_p = \frac{\mathcal{L}_p(\gamma_0)}{\mathcal{L}_p(\hat{\gamma})}$$

is used, as the classical likelihood ratio statistic, for testing the hypotheses

$$H_0 : \gamma = \gamma_0 \quad vs \quad H_1 : \gamma \neq \gamma_0.$$

So, under H_0,

$$-2\log \mathfrak{L}_p \quad \underset{m\to\infty}{\xrightarrow{d}} \quad V \frown \chi^2_{(1)}.$$

For the asymptotic confidence level α, H_0 is then rejected if $-2\log \mathfrak{L}_p > \chi^2_{(1)}(1-\alpha)$ and the profile likelihood-based $100(1-\alpha)\%$ CI for γ is

$$CI_\gamma = \{\gamma : -2\log \mathfrak{L}_p \leq \chi^2_{(1)}(1-\alpha)\} = \left\{\gamma : \log \mathcal{L}_p(\gamma) \geq \log \mathcal{L}_p(\hat{\gamma}) - \frac{\chi^2_{(1)}(1-\alpha)}{2}\right\}.$$

Profile likelihood-based CI for the other GEVd parameters can be constructed following the same reasoning.

3.1.2.5 Statistical choice of extreme value models

As we have seen, the choice of the log-likelihood function depends on the EVI value. Therefore, the following hypotheses test has been widely used in this approach for testing the Gumbel hypothesis as a d.f. of $\{Y_i\}_{i=1,\ldots,k}$, defined in (3.1):

$$H_0 : \gamma = 0 \quad vs \quad H_1 : \gamma \neq 0. \tag{3.18}$$

The Gumbel type d.f. is a favorite one among statisticians because of the great simplicity of statistical inference associated to Gumbel populations. The particular case $\gamma = 0$ can be considered as a transition point: for $\gamma < 0$, data come from a d.f. with finite right

endpoint and for $\gamma > 0$, the d.f. has an infinite right endpoint. For this reason, it is a common practice to separate extreme value models, with the Gumbel-type d.f. playing a central role. The Gumbel-type d.f. is of particular interest, too, because of the great variety of distributions possessing an exponential right-tail, with finite or infinite right endpoint.

The literature about this subject is prolific and we cannot build an exhaustive list covering all the works. Among the articles concerned with this test, we can mention Van Montfort (1970), Tiago de Oliveira (1981), Tiago de Oliveira and Gomes (1984), Hosking (1984) and Marohn (2000). The problem in (3.18) may also be assessed with goodness-of-fit tests for the Gumbel model. The most popular are based on the following goodness-of-fit statistics: Kolmogorov-Smirnov, Cramér-von Mises and Anderson-Darling statistics. Details about this subject can be found in Stephens (1976), Stephens (1977) and Stephens (1986).

3.1.3 The Peaks Over Threshold (POT) method and the Generalized Pareto distribution

This approach focuses on and selects only the observations of the whole sample that exceed a given high and fixed threshold, fitting the appropriate parametric model to the excesses over that high level. Attention is then restricted to a random number of observations exceeding a deterministic level, admitting that sufficient data are available above the chosen threshold. If this can be assumed, we have to look for the convenient conditional distribution of these excesses. For fitting such a distribution to available data, we use the ML method or the PWM method, as for the Block Maxima method.

3.1.3.1 Generalized Pareto distribution

Given a random sample $(X_1, \ldots, X_n)$, with $X \frown F$, and a high and deterministic threshold u, smaller than the right endpoint of the support of F, let

$$N_u = \sharp\{i : X_i > u, i = 1, \ldots, n\}$$

be the number of X_i among $(X_1, \ldots, X_n)$ which exceed the threshold u. Notice that, N_u is a r.v. with Binomial distribution, i.e. $N_u \frown \mathcal{B}(n, 1 - F(u))$. We define ***exceedances*** over u or ***Peaks Over Threshold*** (POT) u as

$$\{W_i\}_{i=1}^{N_u} = \{X_i : X_i > u,\ i = 1, \ldots, N_u\}. \tag{3.19}$$

We say that an exceedance occurs if the observed value is larger than the threshold u. We can, hence, represent the exceedances by the r.v. $X|X > u$. The conditional d.f. of $X|X > u$ is given by

$$F_{X|X>u}(x) = P(X \leq x|X > u) = \frac{F(x) - F(u)}{\overline{F}(u)}, \text{ for } x \geq u. \tag{3.20}$$

Balkema and de Haan (1974) and Pickands III (1975) proved that, for a sufficiently high threshold u and under adequate conditions, the conditional d.f. $F_{X|X>u}(x)$ may be well approximated by the ***Generalized Pareto distribution*** (GPd). More specifically, they proved that the GPd is the limiting distribution of suitably normalized exceedances.

Definition 3.2. (Generalized Pareto distribution) The Generalized Pareto distribution is defined by

$$H_\gamma(x) = \begin{cases} 1 - (1+\gamma x)^{-1/\gamma}, 1 + \gamma x > 0, x \geq 0, & \text{if } \gamma \neq 0, \\ 1 - \exp(-x), x \geq 0, & \text{if } \gamma = 0. \end{cases} \tag{3.21}$$

We can obtain a full parametric model family adding location and scale parameters, $\lambda \in \mathbb{R}$ and $\delta > 0$, respectively:

$$H_\gamma(x|\lambda, \delta) = H_\gamma\left(\frac{x - \lambda}{\delta}\right).$$

For $\gamma < 0$, $\gamma = 0$ and $\gamma > 0$, GPd reduces to Beta, Exponential and type II Pareto d.f., respectively.

The strong connection between GPd and EVT comes from condition (2.21) of Theorem 2.19 or from condition (2.24) of Theorem 2.22. If we pay more attention to these conditions, which define a necessary and sufficient condition for the max-domain of attraction, we see that the left-hand side of the conditions can be rewritten as follows:

$$\frac{\overline{F}(t + xg(t))}{\overline{F}(t)} = P(X > t + xg(t)|X > t) = \overline{F}_{X|X>t}(t + g(t)x).$$

Therefore, the left-hand side of the aforementioned conditions can be interpreted as the conditional survival function of the exceedances over the threshold t, taken at $t+g(t)x$. So, according to (2.21) or (2.24),

$$\lim_{t \to x^F} \overline{F}_{X|X>t}(t + g(t)x) = \begin{cases} (1+\gamma x)^{-1/\gamma}, 1 + \gamma x > 0, & \text{if } \gamma \neq 0, \\ \exp(-x), & \text{if } \gamma = 0, \end{cases}$$

or equivalently,

$$\lim_{t\to x^F} \overline{F}_{X|X>t}(t+g(t)x) = \overline{H}_\gamma(x) \Longleftrightarrow \lim_{t\to x^F} F_{X|X>t}(x) = H_\gamma\left(\frac{x-t}{g(t)}\right),$$

which gives a distributional approximation for the exceedances over the (high) threshold t by the GPd, where $g(t) > 0$ is a scaling factor. We can, then, interpret t as a location parameter and $g(t)$ as a scale parameter, which will be called σ_t. Note that the limit function $D_\gamma(x)$ in (2.20) is the tail quantile function of the GPd, showing that the upper tail of F is, in some sense, close to the upper tail of the GPd.

Recall that our goal is modelling the exceedances d.f., $F_{X|X>t}(x)$, over high thresholds and, hence, it only make sense to consider thresholds that tend to the right endpoint of the underlying d.f. F. Making use from our previous notation in (3.20) , we can restate Theorem 2.22 as follows:

Theorem 3.3 *$F \in \mathcal{D}(G_\gamma)$ if and only if $F_{X|X>u}(x) \simeq H_\gamma(x|u,\sigma_u)$, where*

$$H_\gamma(x|u,\sigma_u) = \begin{cases} 1-(1+\gamma\frac{x-u}{\sigma_u})^{-1/\gamma}, & 1+\gamma\frac{x-u}{\sigma_u} > 0, x \geq u, \quad \text{if } \gamma \neq 0, \\ 1-\exp(-\frac{x-u}{\sigma_u}), & x \geq u, \quad \text{if } \gamma = 0, \end{cases} \tag{3.22}$$

or equivalently

$$H_\gamma(x|u,\sigma_u) = \begin{cases} 1-(1+\gamma\frac{x-u}{\sigma_u})^{-1/\gamma}, & x \geq u, \quad \text{if } \gamma > 0, \\ 1-\exp(-\frac{x-u}{\sigma_u}), & x \geq u, \quad \text{if } \gamma = 0, \\ 1-(1+\gamma\frac{x-u}{\sigma_u})^{-1/\gamma}, & u \leq x \leq u-\frac{\sigma_u}{\gamma}, \quad \text{if } \gamma < 0. \end{cases}$$

We can summarize the results by the ***Pickands-Balkema-de Haan Theorem*** (Balkema and de Haan, 1974 and Pickands III, 1975) , often called the second theorem of EVT, after Fisher-Tippett-Gnedenko Theorem:

Theorem 3.4 (**Pickands-Balkema-de Haan Theorem**) *$F \in \mathcal{D}(G_\gamma)$ if and only if*

$$\lim_{u\to x^F} |F_{X|X>u}(x) - H_\gamma(x|u,\sigma_u)| = 0, \tag{3.23}$$

for some GPd with shape, location and scale parameters γ, u and $\sigma_u > 0$, respectively. If (3.23) *holds, we say that F belongs to the* **POT-domain of attraction** *of the GPd, H_γ.*

This important Theorem makes the connection between GPd and EVT: a continuous d.f. F has a Generalized Pareto upper tail if and only if it belongs to the max-domain of attraction of some extreme value distribution. The GPd is, hence, an important parametric family as it gives the asymptotic behaviour of the upper tail of a d.f. F. The shape

parameter γ, as for the GEVd, is closely related to the upper tail heaviness of the d.f. F: $\gamma = 0$ refers to exponential right tails, $\gamma > 0$ to heavy right tails with no finite right endpoint and $\gamma < 0$ to light right tails with finite right endpoint. The shape parameter $\gamma \in \mathbb{R}$ is the EVI and is the same in both H_γ and G_γ approximations.

Instead of working with the exceedances, we can work with the ***excesses***. Excesses can be represented by the r.v. $Y = X - u$. We can rewrite Theorems 3.3 and 3.4 as follows:

Theorem 3.5 *$F \in \mathcal{D}(G_\gamma)$ if and only if $F_{Y|Y>0}(y) \simeq H_\gamma(y|0, \sigma_u)$, where*

$$H_\gamma(y|0,\sigma_u) = \begin{cases} 1 - (1 + \gamma\frac{y}{\sigma_u})^{-1/\gamma}, & 1 + \gamma\frac{y}{\sigma_u} > 0, y \geq 0, \quad \text{if } \gamma \neq 0, \\ 1 - \exp(-\frac{y}{\sigma_u}), & y \geq 0, \quad \text{if } \gamma = 0, \end{cases} \tag{3.24}$$

or equivalently

$$H_\gamma(y|0,\sigma_u) = \begin{cases} 1 - (1 + \gamma\frac{y}{\sigma_u})^{-1/\gamma}, & y \geq 0, \quad \text{if } \gamma > 0, \\ 1 - \exp(-\frac{y}{\sigma_u}), & y \geq 0, \quad \text{if } \gamma = 0, \\ 1 - (1 + \gamma\frac{y}{\sigma_u})^{-1/\gamma}, & 0 \leq y \leq -\frac{\sigma_u}{\gamma}, \quad \text{if } \gamma < 0. \end{cases}$$

Theorem 3.6 (**Pickands-Balkema-de Haan Theorem**) *$F \in \mathcal{D}(G_\gamma)$ if and only if*

$$\lim_{u \to x^F} |F_{Y|Y>0}(y) - H_\gamma(y|0,\sigma_u)| = 0,$$

for some GPd with shape and scale parameters γ and $\sigma_u > 0$, respectively. In this case, the GPd is the limit distribution of the scaled excesses.

3.1.3.2 Maximum Likelihood estimation

Let $(X_1, \ldots, X_n)$ be the original random sample of the r.v. X, with $X \frown F$. Given a value of the threshold u, let m be the number of exceedances of this sample. We obtain, then, a collection of m excesses, denoted by $Y_j = X_i - u | X_i > u$, for $i = 1, \ldots, n$ and $j = 1, \ldots, m$. In order to pursue a parametric approach, we assume that the actual excess d.f., $F_{Y|Y>0}$, can be replaced by some GPd, as defined in (3.24). We can also work with the exceedances $X_j = X_i | X_i > u$, fitting the parametric family defined in (3.22). In both cases, the problem is the estimation of the parameters γ and σ_u. In this Section, this estimation will be performed through the ML method.

For $\gamma \neq 0$, the log-likelihood function for a given random sample $(y_1, \ldots, y_m)$ of the r.v Y with GPd is given by

$$\ell(\gamma, \sigma_u | y_1, \ldots, y_m) = -m \log \sigma_u - \left(\frac{1}{\gamma} + 1\right) \sum_{i=1}^{m} \log\left(1 + \frac{\gamma y_i}{\sigma_u}\right), \tag{3.25}$$

where $1 + \frac{\gamma y_i}{\sigma_u} > 0, \;\; i = 1, \ldots, m.$

For $\gamma = 0$, the log-likelihood function reduces to the following expression:

$$\ell(0, \sigma_u | y_1, \ldots, y_m) = -m \log \sigma_u - \frac{1}{\sigma_u} \sum_{i=1}^{m} y_i \quad .$$

Usually, for computational purposes, we prefer a reparametrization of the log-likelihood function in (3.25). Defining $\tau = \frac{\gamma}{\sigma_u}$, the log-likelihood function may be rewritten as

$$\ell(\gamma, \tau | y_1, \ldots, y_m) = -m \log \gamma + m \log \tau - \left(\frac{1}{\gamma} + 1\right) \sum_{i=1}^{m} \log(1 + \tau y_i),$$

where $1 + \tau y_i > 0, \;\; i = 1, \ldots, m.$

The ML estimator $(\hat{\gamma}, \hat{\tau})$ of the parameters (γ, τ) follows then from

$$\frac{1}{\hat{\tau}} - \left(\frac{1}{\hat{\gamma}} + 1\right) \frac{1}{m} \sum_{i=1}^{m} \frac{y_i}{1 + \hat{\tau} y_i} = 0, \tag{3.26}$$

where $\hat{\gamma} = \frac{1}{m} \sum_{i=1}^{m} \log(1 + \hat{\tau} y_i)$.

The main objective of this reparametrization, introduced by Davison (1984), is to get $\hat{\gamma}$ explicitly as a function of $\hat{\tau}$, obtained numerically through (3.26), but after the replacement of $\hat{\gamma}$ by $\frac{1}{m} \sum_{i=1}^{m} \log(1 + \hat{\tau} y_i)$. For $\gamma = 0$, we have the classical case of the exponential distribution, yielding $\hat{\sigma}_u = \overline{Y}$.

As discussed in Section 3.1.2.1, Zhou (2009) and Zhou (2010) established the asymptotic normality and consistency of the ML estimators in the EVT framework, for $\gamma > -1$. In particular, Smith (1987) specified the following result:

$$\sqrt{m}((\hat{\gamma}, \hat{\sigma_u}) - (\gamma, \sigma_u)) \quad \xrightarrow[n \to \infty]{d} \quad Z \frown \mathcal{N}(0, V),$$

with $V = \begin{bmatrix} (1+\gamma)^2 & -\sigma_u(1+\gamma) \\ -\sigma_u(1+\gamma) & 2\sigma_u^2(1+\gamma) \end{bmatrix}$.

3.1.3.3 Probability Weighted Moments estimation

As the ML estimators can be numerically hardly tractable for the case $\gamma \neq 0$, Hosking and Wallis (1987) suggest the use of PWM estimators. Recalling the definition of PWM in (3.4), we consider, for the GPd, $M_{p,r,s}$ with $p = 1$, $r = 0$ and $s = 0, 1, \ldots$, yielding

$$M_{1,0,s} = \frac{\sigma_u}{(s+1)(s+1-\gamma)}, \text{ for } \gamma < 1. \tag{3.27}$$

As for the case of fitting a GEVd, in the Block Maxima method, we can replace $M_{1,0,s}$ by its empirical counterpart

$$\widehat{M}_{1,0,s} = \frac{1}{m}\sum_{i=1}^{m}\left(\prod_{j=1}^{s}\frac{m-i-j+1}{m-j}\right)Y_{i:m}$$

and solving (3.27), for $s = 0, 1$ with respect to γ and σ, we obtain the PWM estimators

$$\begin{cases} \hat{\gamma} = 2 - \frac{\widehat{M}_{1,0,0}}{\widehat{M}_{1,0,0}-2\widehat{M}_{1,0,1}}, \\ \hat{\sigma_u} = \frac{2\widehat{M}_{1,0,0}\widehat{M}_{1,0,1}}{\widehat{M}_{1,0,0}-2\widehat{M}_{1,0,1}}. \end{cases}$$

Note that, in the GPd case, the r-th moments only exist for $\gamma < 1/r$.

As for the ML estimation, the PWM estimators are asymptotically Normal-distributed. For details, we refer to Hosking and Wallis (1987). In particular, they showed that, for the GPd with shape parameter in the range $0 \leq \gamma \leq 0.4$ and specially for small samples, the PWM estimators perform better than the ML estimators, since they exhibit small dispersion. The difference is less pronounced as the sample size increases. They also noted that the traditional Method of Moments is preferable when $\gamma < 0$. Nevertheless, the PWM estimation has some problems: on one hand, for $\gamma \geq 1$, PWM estimators do not exist and on the other hand, we can obtain estimates that can be inconsistent with the data, in the sense that some of the observations may fall above the estimate of the right endpoint, x^F.

3.1.3.4 Estimation of other parameters of extreme events

We can now estimate the same quantities defined in Section 3.1.2.3 but under the POT method. Defining, once again, $\mathcal{Y}_{1-p}$ as the **extreme quantile** of order $(1-p)$ of the GPd underlying the excesses Y, with p sufficiently small, we can obtain estimates of extreme quantiles inverting the GPd given by (3.24), yielding

$$\mathcal{Y}_{1-p} = H_\gamma^{-1}(1-p|0,\sigma_u) = \begin{cases} \frac{\sigma_u}{\gamma}(p^{-\gamma}-1), & \text{if } \gamma \neq 0, \\ -\sigma_u \log p, & \text{if } \gamma = 0, \end{cases} \tag{3.28}$$

and replacing (γ, σ_u) by its ML or PWM estimator.

We can use, instead, the tail quantile function of Definition 2.12, with $t = \frac{1}{p}$, to express extreme quantiles of the GPd:

$$U_{H_\gamma}\left(\frac{1}{p}\right) = H_\gamma^{-1}(1-p|0,\sigma_u) = \mathcal{Y}_{1-p}.$$

If $\gamma < 0$, the right endpoint of the GPd is finite and is given by

$$U_{H_\gamma}(\infty) = \mathcal{Y}_1 = -\frac{\sigma_u}{\gamma},$$

which can be estimated replacing again (γ, σ_u) by its ML or PWM estimator.

One important point has to be noted: provided that, under this parametric approach, $F_{Y|Y>0}(y) \simeq H_\gamma(y|0,\sigma_u)$, the quantiles estimates obtained from (3.28) are the estimated quantiles of the d.f. $F_{Y|Y>0}(y)$. However, if we want to estimate the extreme quantiles of the original and unknown d.f. F, associated with the r.v. X, we can use the identity (3.20). Provided that $Y = X - u$, we have

$$\begin{aligned} F_{Y|Y>0}(y) &= P(Y \le y|Y>0) \\ &= P(X \le u+y|X>u) \\ &= F_{X|X>u}(u+y) \\ &= \frac{F(u+y)-F(u)}{\overline{F}(u)}. \end{aligned}$$

Therefore, we can establish the following identities:

$$\begin{aligned} F_{Y|Y>0}(x-u) &= \frac{F(x)-F(u)}{\overline{F}(u)} \Leftrightarrow \\ \Leftrightarrow F(x) &= F_{Y|Y>0}(x-u)(1-F(u))+F(u), \end{aligned}$$

or, equivalently, in terms of the tail quantile function,

$$\begin{aligned} 1-F(x) &= 1-F_{Y|Y>0}(x-u)(1-F(u))-F(u) \Leftrightarrow \\ \Leftrightarrow \overline{F}(x) &= \overline{F}(u)(1-F_{Y|Y>0}(x-u)). \end{aligned}$$

As $F_{Y|Y>0}(y) \simeq H_\gamma(y|0,\sigma_u)$, we have

$$\overline{F}(x) \simeq \overline{F}(u)(1-H_\gamma(x-u|0,\sigma_u)).$$

Estimating $\overline{F}(u)$ by the sample frequency of the observations that exceed u in the original sample $(X_1,\ldots,X_n)$, $\frac{m}{n}$, and replacing the parameters of H_γ by their ML or PWM estimators, we get

$$\widehat{\overline{F}}(x) = \frac{m}{n}(1-H_{\hat{\gamma}}(x-u|0,\hat{\sigma}_u)). \tag{3.29}$$

Defining $\mathcal{X}_{1-p}$ as the extreme quantile of order $(1-p)$ of the d.f. F underlying the r.v. X, i.e., a quantity such that $F(\mathcal{X}_{1-p}) = 1-p$, with p sufficiently small, we can estimate these quantiles, for instance for $\gamma \neq 0$, using (3.29):

$$\hat{\mathcal{X}}_{1-p} = \widehat{U}\left(\frac{1}{p}\right) = \widehat{F}^{-1}(1-p) = u + \frac{\hat{\sigma}_u}{\hat{\gamma}}\left[\left(\frac{np}{m}\right)^{-\hat{\gamma}} - 1\right],$$

and, in particular for $\gamma < 0$, an estimate for the **right endpoint** of the d.f. F is given by

$$\hat{x}^F = \widehat{U}(\infty) = u - \frac{\hat{\sigma}_u}{\hat{\gamma}}. \tag{3.30}$$

3.1.3.5 Inference: confidence intervals for the extreme value index

Following the same reasoning as in Section 3.1.2.4, we can construct CI's for the parameters of the GPd, on the basis of the asymptotic normality of the ML and PWM estimators.

As for the case of Block Maxima, Beirlant et al. (2004) recommend the use of the *profile likelihood function* to construct better CI's. In the case of the GPd, the profile likelihood-based $100(1-\alpha)\%$ CI for γ is

$$CI_\gamma = \left\{\gamma : \log \mathcal{L}_p(\gamma) \geq \log \mathcal{L}_p(\hat{\gamma}) - \frac{\chi^2_{(1)}(1-\alpha)}{2}\right\}. \tag{3.31}$$

Profile likelihood-based CI's for the other GPd parameters can be constructed by the same way.

3.1.3.6 Statistical choice of GPd models

In the POT method, the following hypotheses test is mainly considered, for the same reasons seen in the case of the GEVd fitting:

$$H_0 : \gamma = 0 \quad vs \quad H_1 : \gamma \neq 0.$$

This test gives priority to the Exponential d.f. for modelling the excesses above a high threshold and received some attention in the literature, specially from hydrologists. We can find the test procedures in Gomes and van Monfort (1986), Marohn (2000), Reiss and Thomas (2007) and Kozubowski et al. (2009), among the great variety of literature. The problem of goodness-of-fit for the GPd model was studied by Choulakian and Stephens (2001), among others, and Lilliefors (1969) presented the special case of the Kolmogorov-Smirnov test, applied to the Exponential distribution with unknown parameters.

3.1.3.7 The choice of the threshold

The choice of the threshold u is still an unsolved problem and in the literature of the POT method, not so much attention has been given to this issue. The choice of the threshold is not straightforward; indeed, a compromise has to be found between high values of u, where the bias of the estimators is smaller, and low values of u, where the variance is smaller.

Davison and Smith (1990) suggest the use of the ***Mean Excess function*** defined in (2.26). In the GPd case, this function is given by:

$$e(u) = E(X - u|X > u) = E(Y|Y > 0) = \frac{\sigma_u + \gamma u}{1 - \gamma}, \text{ for } \gamma < 1. \tag{3.32}$$

If the GPd assumption is valid, the plot of $e(u)$ versus u, called ***Mean Excess plot*** (or shortly ME-plot), should follow a straight line with intercept $\frac{\sigma_u}{1-\gamma}$ and slope $\frac{\gamma}{1-\gamma}$. In practice, based on a sample of size n, $(x_1, x_2, \ldots, x_n)$, $e(u)$ is estimated by its empirical counterpart, the ***sample Mean Excess function***:

$$\hat{e}_n(u) = \frac{\sum_{i=1}^{n} x_i I_{]u,+\infty[}(x_i)}{\sum_{i=1}^{n} I_{]u,+\infty[}(x_i)} - u, \text{ where } I_{]u,+\infty[}(x_i) = \begin{cases} 1, \text{ if } x_i \in]u, +\infty[, \\ 0, \text{ if } x_i \in]-\infty, u]. \end{cases} \tag{3.33}$$

To view this function, we generally construct the *sample ME-plot*

$$\{(X_{n-k:n}, \hat{e}_n(X_{n-k:n})) : 1 \leq k \leq n-1\},$$

where $X_{n-k:n}$ denotes the the $(k+1)$-th largest observation and where $\hat{e}_n(u)$ may be rewritten as

$$\hat{e}_n(x_{n-k:n}) = \frac{1}{k} \sum_{j=1}^{k} x_{n-j+1:n} - x_{n-k:n}. \tag{3.34}$$

If the data support a GPd over a high threshold, we would expect the sample ME-plot to become linear in view of (3.32). At least, this is the ideal situation. But even for data that are genuinely GP-distributed, the sample ME-plot is seldom perfectly linear, particularly toward the right-hand end, where we are averaging a small number of large excesses. In fact we often omit the final few points from consideration, as they can severely distort the plot. Consequently, the threshold u is chosen at the point to the right of which a rough linear pattern appears in the plot.

Another procedure consists in choosing one of the sample points as a threshold, i.e. $u = X_{n-k:n}, \quad k = 1, \ldots, n-1$. With such a random threshold, we work then with the $k+1$ top order statistics associated to the whole sample of size n, $X_{n:n}, X_{n-1:n}, \ldots, X_{n-k:n}$.

This procedure has given rise to the semi-parametric approaches and to the Peaks Over Random Threshold (PORT) methodology, which will be discussed in the next Section. In these methods, the choice of k is an important issue.

Finally, another approach is to try different levels u. If the model produces very different results for different choices of u, the results should be viewed with caution.

3.2 Semi-Parametric Inference

3.2.1 Introduction

The use of parametric approaches has been raising some questions and doubts about the validity of its assumptions. Since we are using an approximate d.f. like the GEVd or the GPd instead of the exact d.f. F for the data, subjecting them to an asymptotic parametric model, just like an exact model, may seem very rigid and somehow unrealistic. Moreover, in many applications of EVT, the main interest is not to describe the data at the expense of a theoretical and unrealistic model, but to describe the "behaviour" of extreme values.

During the seventies, we assist to the birth of another type of approach for statistical inference in EVT, with the pioneering works of Pickands III (1975) and Hill (1975), among others. This new approach is known as ***semi-parametric approach***. This term reflects the fact that our main interest remains in the estimation of parameters of extreme events, specially the EVI, but at the expense of only partial assumptions about the unknown d.f. F. The subsequent development of this area is due once again to Laurens de Haan and, currently, the estimation of parameters of extreme events is often developed under this framework.

Following a semi-parametric approach, there is no fit of a specific parametric model depending on a location parameter (λ), a scale parameter (δ) and a shape parameter (γ), as in a parametric approach. No assumption is then made about the global form of the underlying d.f. F. The only assumption is about its tail behaviour, where we want do make some inference. Therefore, in this approach, the only assumption is that the d.f. F belongs to the max-domain of attraction of some extreme value distribution, i.e., $F \in \mathcal{D}(G_\gamma)$.

From condition (2.21) of Theorem 2.19, we known that

$$F \in \mathcal{D}(G_\gamma) \Longleftrightarrow \lim_{t \to x^F} \frac{\overline{F}(t + xg(t))}{\overline{F}(t)} = (1 + \gamma x)^{-1/\gamma},$$

for any x, where $1 + \gamma x > 0$.

In Section 3.1.3.1, it was seen that

$$\frac{\overline{F}(t + xg(t))}{\overline{F}(t)} = \overline{F}_{X|X>t}(t + g(t)x),$$

and also that

$$\lim_{t \to x^F} \overline{F}_{X|X>t}(t + g(t)x) = \overline{H}_\gamma(x) = \begin{cases} (1+\gamma x)^{-1/\gamma}, 1 + \gamma x > 0, & \text{if } \gamma \neq 0, \\ \exp(-x), & \text{if } \gamma = 0, \end{cases}$$

or equivalently

$$\lim_{t \to x^F} F_{X|X>t}(x) = H_\gamma \left(\frac{x-t}{g(t)} \right),$$

where $H_\gamma(.)$ is the GPd and $g(t)$ a scaling factor, as seen before.

From relation (3.20) and Section 3.1.3.4, we know that we can rewrite this last condition as

$$\overline{F}(x) \simeq \overline{F}(t) \left\{ 1 - H_\gamma \left(\frac{x-t}{g(t)} \right) \right\}, \tag{3.35}$$

which means that, from some high threshold t onwards, i.e. for $X > t$, the tail of the d.f. F may be well approximated by relation (3.35). Then, inference about the tail of the d.f. F can be drawn from observations above the threshold t.

As said in Section 3.1.3.7, the threshold is chosen at one of the random sample point $X_{n-k:n}$. The inference is then based on the $k+1$ top order statistics and not only on the sample maximum M_n. Indeed, it would be unrealistic to consider that only the sample maximum (one single observation) contains valuable information about the tail of the d.f. In many extreme situations, we have only a finite sample and we cannot use only the largest observation for inference. It is thus plausible to base inference on a set of top order statistics, since it is only these that lie in the region of F we believe it has the specified form.

Intuitively, as n increases, so should k, or we can miss the benefit of an increasing sample size. The choice of k is then intimately related with the sample size n, i.e.

$$k = k_n, \quad \text{where } k = k_n \to \infty \quad \text{as} \quad n \to \infty, \tag{3.36}$$

but at a slower rate than n such that

$$\frac{k_n}{n} \to 0 \quad \text{as} \quad n \to \infty. \tag{3.37}$$

A sequence of integers k_n is said to be *intermediate* if it satisfies (3.36) and (3.37). Consequently, order statistics $X_{n-k:n}$, with k satisfying (3.36) and (3.37) are called ***intermediate order statistics***.

Determining k is then an important issue in a semi-parametric approach. However, the choice of k is not an easy task and several authors have suggested many solutions, but none of them have been universally adopted. The choice of k is a problem parallel to the determination of the threshold u in Chapter 3.1.3.7.

Chosen the threshold $t = X_{n-k:n}$, we can expect the approximation (3.35) to hold for intermediate order statistics. Let then $X_1, X_2, \ldots, X_n$ be i.i.d r.v.s with d.f. F. The simplest unbiased estimator for F is the *empirical distribution function*, F_n, given by

$$F_n(x) = n^{-1} \sum_{i=1}^{n} I_{]-\infty,x]}(x_i), \tag{3.38}$$

with $I(.)$ defined in (3.33).

Using the approximation (3.35) for the random threshold $t = X_{n-k:n}$ and choosing $k = k_n \to \infty$, $\frac{k}{n} \to 0$ and $n \to \infty$, we get

$$\overline{F}(x) \simeq \overline{F}(X_{n-k,n}) \left[1 - H_\gamma \left(\frac{x - X_{n-k:n}}{g(X_{n-k:n})} \right)\right], \quad x > X_{n-k:n}.$$

and provided that $\overline{F}(X_{n-k,n}) \simeq \overline{F}_n(X_{n-k,n}) = \frac{k}{n}$ and that $g(X_{n-k,n}) = a\left(\frac{1}{\overline{F}(X_{n-k,n})}\right)$ (see (2.21)), it follows that

$$\overline{F}(x) \simeq \frac{k}{n} \left\{ 1 - H_\gamma \left(\frac{x - X_{n-k:n}}{a\left(\frac{n}{k}\right)} \right) \right\}, \; x > X_{n-k:n}. \tag{3.39}$$

Therefore, in order to concretize this latter approximation, we need to estimate the parameter γ and the normalizing scaling constant $a\left(\frac{n}{k}\right)$. This approximation is valid for any x larger than $X_{n-k,n}$ and even for $x > X_{n:n}$, i.e., outside the range of the observations, which is the key behind EVT. Consequently, γ is the main parameter of extreme events to be estimated, using a set of top observations and an adequate methodology.

In Section 3.2.3, we present some of the most well-known semi-parametric estimators for the EVI, ranging from the first pioneering contributions to more recent progresses. Since the semi-parametric approach is a very vast field, it is impossible to cover all the works developed in this area. So, it is not our objective to present an exhaustive list of all the existent estimators. As mentioned before, the main parameter of interest in this approach is γ, as it is the basis for the estimation of other parameters of extreme events. Therefore, we must obtain estimators for γ with desirable properties in order to do proper inference. While some properties are only dependent on the behaviour of k

and on the first order extended regular variation property defined in expression (2.21) of Theorem 2.19, other properties, namely the asymptotic normality of the estimators, require that the underlying d.f. F satisfies another condition, apart from those mentioned. This additional condition is known as ***the second order extended regular variation property*** and as for the first order condition, we will refer to this new property as *the second order condition*.

3.2.2 The second order extended regular variation property

In a semi-parametric approach, apart from the first order condition, we often need a second order condition, to guarantee desirable properties for the estimators of the EVI. The only assumption in this approach is that $F \in \mathcal{D}(G_\gamma)$, which is equivalent to assume that the first order extended regular variation property is satisfied, i.e., from Theorem 2.19, we have

$$F \in \mathcal{D}(G_\gamma) \Longleftrightarrow \lim_{t\to\infty} \frac{U(tx)-U(t)}{a(t)} = D_\gamma(x) = \begin{cases} \frac{x^\gamma - 1}{\gamma}, & \gamma \neq 0, \\ \log x, & \gamma = 0, \end{cases}$$

for every $x > 0$ and some positive measurable auxiliary function a, where necessarily we have $a \in \mathcal{RV}_\gamma$, according to definition of extended regular variation (see Definition 2.16).

As mentioned before, an increase of the sample size implies an increase of k, the number of intermediate order statistics used to estimate γ. However, this rise of k may introduce some bias in the EVI-estimators, which can be controlled if we have additional information about the tail of F, in order to control the speed of convergence in the first order condition, i.e., the speed of convergence of maximum values, linearly normalized, towards the limit law G_γ. This condition is known as *the second order extended regular variation property*. Therefore, the choice of k will also be decided by this second order condition. We must then quantify the speed of convergence, imposing a precise rate. For that, we merely need to assume that there exists a function $A(t)$, not changing sign eventually, such that $\lim_{t\to\infty} A(t) = 0$, which measures not only the speed of convergence of the sequence of maximum values to a non-degenerate limit law but also the bias of the estimators. This function $A(t)$ may be either positive or negative. As $A(t)$ measures the speed of convergence of $\frac{U(tx)-U(t)}{a(t)}$ towards $D_\gamma(x)$,

$$\lim_{t\to\infty} \frac{\frac{U(tx)-U(t)}{a(t)} - D_\gamma(x)}{A(t)} \tag{3.40}$$

must exist, for all $x > 0$. Let be $H(x)$ the limit function of (3.40). We can now define the second order condition as follows:

Definition 3.7. (Second order condition) The function U (or the associated d.f. F) is said to satisfy the second order condition if, for some positive function a and for some positive or negative function A, with $\lim_{t\to\infty} A(t) = 0$,

$$\lim_{t\to\infty} \frac{\frac{U(tx)-U(t)}{a(t)} - D_\gamma(x)}{A(t)} = H(x), \quad x > 0. \tag{3.41}$$

As for the function a of the first-order condition, we call A the *second order auxiliary function.*

We need now to determine which functions $H(x)$ are eligible for the limit relation in (3.41). Following de Haan and Ferreira (2006), we can then state the following result for the function $H(x)$:

Theorem 3.8 (**de Haan and Ferreira (2006), Theorem 2.3.3 and Corollary 2.3.4**) *Suppose relation* (3.41) *holds and the function H is not a multiple of D_γ and is not identically zero. Then, there exist functions a, positive, and A, positive or negative, and a parameter $\rho \leq 0$ such that*

$$\lim_{t\to\infty} \frac{\frac{U(tx)-U(t)}{a(t)} - D_\gamma(x)}{A(t)} = H_{\gamma,\rho}(x) = \frac{1}{\rho}\left(\frac{x^{\gamma+\rho}-1}{\gamma+\rho} - \frac{x^\gamma - 1}{\gamma}\right), \tag{3.42}$$

for $x > 0$.
ρ is a second order parameter controlling the speed of convergence of the first order condition. For the cases $\gamma = 0$ and/or $\rho = 0$, $H_{\gamma,\rho}(x)$ is understood to be equal to the respective limit in (3.42)*, by continuity arguments.*
Moreover, $A(t)$ is such that

$$\lim_{t\to\infty} \frac{A(tx)}{A(t)} = x^\rho,$$

that is, $|A| \in \mathcal{RV}_\rho$.

The function A, describing the rate of convergence of the first order condition, is regularly varying with index ρ. So, if $\rho < 0$, we have an algebraic speed of convergence and if $\rho = 0$, the speed of convergence is slower (logarithmic, for example). Therefore, the rate at which the intermediate sequence k_n tends to infinity must be in accordance with the rate of convergence in the first order condition, as quantified by the rate function A. Note that the second-order condition implies the domain of attraction condition.

Dekkers and de Haan (1989) showed that the second order condition holds for most of well-known d.f.'s (Normal, Gamma, GEVd, Exponential, Uniform, Cauchy).

3.2.3 Estimation of the extreme value index

To unify notation and avoid confusion with the number of intermediate order statistics involved in the following estimators of the EVI, let $(X_1, \ldots, X_n)$ be a random sample of size n taken from a d.f. F, such that $F \in \mathcal{D}(G_\gamma)$, for some $\gamma \in \mathbb{R}$, with G_γ defined in (2.8). We denote by $X_{1:n} \leq X_{2:n} \leq ... \leq X_{n:n}$ the corresponding order statistics in non-descending order and by $X_{(i)} = X_{n-i+1:n},\ i = 1, \ldots, n$ the descending order statistics. The purpose of this Section is the estimation of the parameter γ, considering a random threshold $X_{n-k:n},\ k = 1, \ldots, n-1$, and basing the estimation on the $k+1$ top order statistics, $X_{n:n}, X_{n-1:n}, \ldots, X_{n-k:n}$ or equivalently $X_{(1)}, X_{(2)}, \ldots, X_{(k+1)}$. It is very important to stress that, in this Section, $k+1$ represents the top-portion of the sample which is selected in order to estimate the EVI and not the number of observations used on the computation of the estimate. Some estimators select a top-portion of the sample but use only a few observations from this portion. This is the case for the first EVI estimator presented below: the Pickands estimator.

3.2.3.1 The Pickands estimator

Pickands III (1975) was the first to present a semi-parametric estimator for a real EVI, $\gamma \in \mathbb{R}$. The ***Pickands estimator*** for $\gamma \in \mathbb{R}$, $\hat{\gamma}^P_{n,k}$, is given by

$$\hat{\gamma}^P_{n,k} = \frac{1}{\log 2} \log \left(\frac{X_{\left([\frac{k+1}{4}]\right)} - X_{\left(2[\frac{k+1}{4}]\right)}}{X_{\left(2[\frac{k+1}{4}]\right)} - X_{\left(4[\frac{k+1}{4}]\right)}} \right), \quad k = 1, \ldots, n, \tag{3.43}$$

where $[x]$ stands for the integer part of x. Notice that this estimator involves $k+1$ of the top observations, for $k \geq 3$.

The properties of $\hat{\gamma}^P_{n,k}$ were studied by Pickands III (1975) and extended by Dekkers and de Haan (1989). In particular, they proved that this estimator is strongly consistent (and therefore weakly consistent as well) and asymptotically Normal-distributed. We refer to those authors for the proofs.

Theorem 3.9 (**Weak consistency, Pickands III, 1975**) *Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of i.i.d. r.v.'s with d.f. F. If $F \in \mathcal{D}(G_\gamma)$ and k is an intermediate sequence of integers defined in* (3.36) *and* (3.37)*, then*

$$\hat{\gamma}^P_{n,k} \xrightarrow[n\to\infty]{p} \gamma,$$

where $\xrightarrow[n\to\infty]{p}$ means convergence in probability.

Theorem 3.10 (**Strong consistency, Dekkers and de Haan, 1989**) *Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of i.i.d. r.v.'s with d.f. F. If $F \in \mathcal{D}(G_\gamma)$ and k_n is an intermediate sequence of integers defined in* (3.36) *and* (3.37) *such that $\frac{k_n}{\log(\log n)} \to \infty$, then*

$$\hat{\gamma}^P_{n,k} \quad \xrightarrow[n\to\infty]{a.s.} \quad \gamma,$$

where $\xrightarrow[n\to\infty]{a.s.}$ means almost sure convergence.

The Pickands estimator is very appellative because of its relative simplicity, its scale/location invariance and its applicability to the general case $\gamma \in \mathbb{R}$. However, this estimator is characterized by a large asymptotic variance and a high volatility as a function of k, since it is very sensitive to the choice of intermediate order statistics that are used for the estimation. All these features motivated more recent modifications and proposals for the Pickands estimator. Among recent improvements, we can refer Pereira (1994), Fraga Alves (1995), Drees (1995) and Segers (2005).
Pereira (1994) and Fraga Alves (1995) proposed a generalization of the Pickands estimator, with the introduction of a *tuning* or *control parameter* M, defined as follows:

$$\hat{\gamma}^P_{n,k,M} = \frac{1}{\log M} \log \left(\frac{X_{\left(\left[\frac{k+1}{M^2}\right]\right)} - X_{\left(M\left[\frac{k+1}{M^2}\right]\right)}}{X_{\left(M\left[\frac{k+1}{M^2}\right]\right)} - X_{\left(M^2\left[\frac{k+1}{M^2}\right]\right)}} \right), \quad k = 1, \ldots, n \quad \text{and} \quad M \in \mathbb{N} \setminus \{1\},$$

which involves $k+1$ of the top observations, for $k \geq M^2 - 1$. The traditional Pickands estimator corresponds to $\hat{\gamma}^P_{n,k,2}$. This generalized estimator was defined since, according to Fraga Alves (1995), there seems to be no particular reason for the choice of $M = 2$ in the classical Pickands estimator. $\hat{\gamma}^P_{n,k,M}$ is proved to be consistent (weakly and strongly) and asymptotically Normal, permitting the construction of CI's for γ.

Drees (1995) also presents a refined version of $\hat{\gamma}^P_{n,k}$, consisting in a mixture of Pickands estimators, which has been generalized later by Segers (2005).

3.2.3.2 The Hill estimator

A few months after the publication of the Pickands estimator, Hill (1975) proposed another estimator for γ, restricted however, to heavy tails d.f.'s, which belong to Fréchet max-domain of attraction ($\gamma > 0$). We will denote this estimator by $\hat{\gamma}^H_{n,k}$, so that the ***Hill estimator*** for $\gamma > 0$ is given by

$$\hat{\gamma}^H_{n,k} = \frac{1}{k} \sum_{i=1}^{k} (\log X_{(i)} - \log X_{(k+1)}), \tag{3.44}$$

which involves $k+1$ of the top order statistics.

The properties of the Hill estimator have been studied by many authors and we may find many references in the literature. We can find condensed results in de Haan and Ferreira (2006) and Embrechts et al. (1997):

Theorem 3.11 (**Weak consistency**) *Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of i.i.d. r.v.'s with d.f. F. If $F \in \mathcal{D}(G_\gamma)$, with $\gamma > 0$, and k is an intermediate sequence of integers, as defined in* (3.36) *and* (3.37)*, then*

$$\hat{\gamma}^H_{n,k} \xrightarrow[n\to\infty]{p} \gamma.$$

Theorem 3.12 (**Strong consistency**) *Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of i.i.d. r.v.'s with d.f. F. If $F \in \mathcal{D}(G_\gamma)$, with $\gamma > 0$, and k_n is an intermediate sequence of integers, as defined in* (3.36) *and* (3.37)*, such that $\frac{k_n}{\log(\log n)} \to \infty$, then*

$$\hat{\gamma}^H_{n,k} \xrightarrow[n\to\infty]{a.s.} \gamma.$$

The Hill estimator obtained a high degree of popularity since it can be derived and interpreted from different points of view and because it has appealing theoretical properties. Despite these advantages, this estimator presents some problems: first, unlike the Pickands estimator, the Hill estimator is not invariant to shifts of the data, while the scale invariance remains. On the order hand, as for many estimators, the Hill estimator is sensitive to the growth speed of k with respect to n. In several instances, a severe bias can appear, depending on the rate of increase of k_n.

As for the Pickands estimator, the Hill estimator caught the attention of many authors in an attempt to look for modifications. As examples, we can mention Peng (1998) and Fraga Alves (2001).

3.2.3.3 The Moment estimator

The ***Moment estimator***, which will be denoted by $\hat{\gamma}^M_{n,k}$, was developed by Dekkers et al. (1989) as an extension of the Hill estimator for $\gamma \in \mathbb{R}$ and not only for $\gamma > 0$.

Define, for $j = 1, 2$,

$$M^{(j)}_{n,k} = \frac{1}{k}\sum_{i=1}^{k}(\log X_{(i)} - \log X_{(k+1)})^j, \tag{3.45}$$

and also

$$\hat{\gamma}^+_{n,k} = M^{(1)}_{n,k} = \hat{\gamma}^H_{n,k} \quad \text{and} \quad \hat{\gamma}^-_{n,k} = 1 - \frac{1}{2}\left[1 - \frac{\left(M^{(1)}_{n,k}\right)^2}{M^{(2)}_{n,k}}\right]^{-1}. \tag{3.46}$$

The Moment estimator for $\gamma = \gamma_+ + \gamma_-$ is defined by the following expression:

$$\hat{\gamma}^M_{n,k} = \hat{\gamma}^+_{n,k} + \hat{\gamma}^-_{n,k},$$

where $\gamma_+ = \max(0, \gamma)$ e $\gamma_- = \min(0, \gamma)$. As for the previous estimators, the Moment estimator deals with $k+1$ top observations.

We observe that the Moment estimator has two pieces: $\hat{\gamma}^+_{n,k}$, which is the Hill estimator defined in (3.44), valid for $\gamma > 0$, and $\hat{\gamma}^-_{n,k}$, which will be called the ***Negative Moment estimator***, valid for $\gamma < 0$.

As the Moment estimator is an extension of the Hill estimator for the more general case $\gamma \in \mathbb{R}$, it satisfies the properties of weak and strong consistency as well:

Theorem 3.13 (**Weak consistency**) *Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of i.i.d. r.v.'s with d.f. F. If $F \in \mathcal{D}(G_\gamma)$ and k is an intermediate sequence of integers, as defined in* (3.36) *and* (3.37)*, then*

$$\hat{\gamma}^M_{n,k} \xrightarrow[n\to\infty]{p} \gamma.$$

Theorem 3.14 (**Strong consistency**) *Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of i.i.d. r.v.'s with d.f. F. If $F \in \mathcal{D}(G_\gamma)$ and k_n is an intermediate sequence of integers, as defined in* (3.36) *and* (3.37)*, such that $\frac{k_n}{(\log n)^\eta} \to \infty$, for some $\eta > 0$, then*

$$\hat{\gamma}^M_{n,k} \xrightarrow[n\to\infty]{a.s.} \gamma.$$

As for the Hill estimator, the Moment estimator is not location invariant, conserving the scale invariance property.

3.2.3.4 The Negative Hill estimator

Apart from the Negative Moment estimator defined in (3.46), the ***Negative Hill estimator***, $\hat{\gamma}^{NH}_{n,k}$, is another estimator for $\gamma < 0$, introduced by Falk (1995), as an alternative for the POT-ML estimator discussed in Section 3.1.3.2, which was seen to have desirable properties only for $\gamma > -0.5$, until the recent works of Zhou (2009) and Zhou (2010), who extended its properties for $\gamma > -1$. Nevertheless, the Negative Hill estimator was developed at a time when the desirable properties for the POT-ML estimator were only proved for $\gamma > -0.5$. Therefore, it is intended to be used only for $\gamma < -0.5$.

As we know from Chapter 2, if $F \in \mathcal{D}(G_\gamma)$ and $\gamma < 0$, then the right endpoint of F, x^F, is finite. Consequently, according to de Haan and Ferreira (2006), the distribution function of

$$\widetilde{X} = \frac{1}{x^F - X}$$

is in the max-domain of attraction of $G_{-\gamma}$. We can thus apply the Hill estimator to $\widetilde{X}$, but, since x^F is unknown, it has to be estimated, in order to produce a statistic. As $\gamma < 0$, the right endpoint is finite and can be well approximated by the sample maximum, $X_{n:n}$. We obtain finally, the Negative Hill estimator:

$$\hat{\gamma}_{n,k}^{NH} = \frac{1}{k}\sum_{i=1}^{k-1} \log(X_{(1)} - X_{(i+1)}) - \log(X_{(1)} - X_{(k+1)}),$$

which involves again $k+1$ top order statistics. Unlike the Hill estimator, the Negative Hill estimator is shift invariant and scale invariant as well.

As for the others estimators, the Negative Hill estimator is consistent for $\gamma < -0.5$. de Haan and Ferreira (2006) enunciate the conditions for weak consistency:

Theorem 3.15 (**Weak consistency, de Haan and Ferreira (2006),Theorem 3.6.4**) *Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of i.i.d. r.v.'s with d.f. F. If $F \in \mathcal{D}(G_\gamma)$ and k is an intermediate sequence of integers, as defined in* (3.36) *and* (3.37)*, such that $\frac{k^\eta}{\log n} \to \infty$ for $\eta \to 0$, then*

$$\hat{\gamma}_{n,k}^{NH} \xrightarrow[n\to\infty]{p} \gamma.$$

3.2.3.5 The Generalized Hill estimator

The ***Generalized Hill estimator***, $\hat{\gamma}_{n,k}^{GH}$ is a scale but not location invariant estimator introduced by Beirlant et al. (1996), as another attempt to generalize the Hill estimator for the case $\gamma \in \mathbb{R}$ and is defined as follows:

$$\hat{\gamma}_{n,k}^{GH} = \hat{\gamma}_{n,k}^{H} + \frac{1}{k}\sum_{i=1}^{k}(\log \hat{\gamma}_{n,i}^{H} - \log \hat{\gamma}_{n,k}^{H}),$$

where $\hat{\gamma}_{n,k}^{H}$ stands for the Hill estimator defined in (3.44). Since this estimator is based on $\hat{\gamma}_{n,k}^{H}$, it involves $k+1$ top observations.

The Generalized Hill estimator is consistent for $\gamma \in \mathbb{R}$, provided that k_n is an intermediate sequence of integers defined in (3.36) and (3.37). For a complete study of its properties, we refer to Beirlant et al. (2005).

3.2.3.6 The Mixed Moment estimator

More recently, the ***Mixed Moment estimator***, $\hat{\gamma}_{n,k}^{MM}$, was developed by Fraga Alves et al. (2009) from a combination of Theorems 2.6.1 and 2.6.2 of de Haan (1970), again for the general case $\gamma \in \mathbb{R}$:

$$\hat{\gamma}_{n,k}^{MM} = \frac{\hat{\varphi}_n(k) - 1}{1 + 2\min(\hat{\varphi}_n(k) - 1, 0)},$$

where

$$\hat{\varphi}_n(k) = \frac{M_{n,k}^{(1)} - L_{n,k}^{(1)}}{\left(L_{n,k}^{(1)}\right)^2},$$

with

$$L_{n,k}^{(1)} = \frac{1}{k}\sum_{i=1}^{k}\left(1 - \frac{X_{(k+1)}}{X_{(i)}}\right),$$

and $M_{n,k}^{(j)}$ defined in (3.45), for $j = 1$. As its predecessors, this estimator works with $k+1$ top observations.

Since the Mixed Moment estimator is not location invariant, Fraga Alves et al. (2009) propose an alternative estimator, dependent on a *tuning parameter* q, $0 \leq q < 1$. This alternative estimator is based on transformed data obtained from the original sample $(X_1, \ldots, X_n)$. Each value of the original sample is then replaced by

$$X_i^* = X_i - X_{[nq]+1},\ 0 \leq q < 1,\ 1 \leq i \leq n,$$

where $[x]$ stands for the integer part of x.

We can then obtain the corresponding Mixed Moment estimator, $\hat{\gamma}_{n,k}^{MM(q)}$, which will be location invariant. This methodology, known as the PORT-methodology, whose terminology was introduced by Araújo Santos et al. (2006), is discussed in the next Section. As for the other estimators, the Mixed Moment estimator was proved to be consistent for any $\gamma \in \mathbb{R}$.

3.2.3.7 The Peaks Over Random Threshold (PORT) Methodology

It was seen that some of the presented semi-parametric estimators for γ are not location invariant: the Hill and the Moment estimators suffer from this drawback. In order to fill this gap, recent methodologies try to introduce some modifications in those estimators, in order to induce the location invariance property.

The most well-known modification was introduced in Araújo Santos et al. (2006) and consists in working with a sample of excesses over a random threshold $X_{n_q:n}$,

$$\underset{\sim}{X}^{(q)} = (X_{n:n} - X_{n_q:n}, X_{n-1:n} - X_{n_q:n}, \ldots, X_{n_q+1:n} - X_{n_q:n}), \tag{3.47}$$

where $n_q = [nq] + 1$, with

(i) $0 < q < 1$, for d.f.'s with finite or infinite left endpoint $x_F = \inf\{x : F(x) > 0\}$, and, consequently, the random threshold is an empirical quantile,

(ii) $q = 0$, for d.f.'s with finite left endpoint x_F and the random threshold can then be the sample minimum.

As already seen, q is called a *tuning parameter*.

Therefore, a methodology based on a sample of excesses $\underset{\sim}{X}^{(q)}$, as defined in (3.47) is called a ***PORT-methodology***, with PORT standing for Peaks Over Random Threshold. With this methodology, we can obtain consistent estimators for the EVI from any of the classical estimators, made location/scale invariant using the transformed sample in (3.47). For more details, we refer to Araújo Santos et al. (2006). These authors base the classical Hill and Moment estimators on the transformed data $\underset{\sim}{X}^{(q)}$, obtaining the ***PORT-Hill***, $\hat{\gamma}_{n,k}^{H(q)} = \hat{\gamma}_{n,k}^{H}(\underset{\sim}{X}^{(q)})$, and the ***PORT-Moment***, $\hat{\gamma}_{n,k}^{M(q)} = \hat{\gamma}_{n,k}^{M}(\underset{\sim}{X}^{(q)})$ estimators, respectively. An exhaustive study of their asymptotic properties is also performed.

The choice of a convenient value for q is crucial because a misleading value can conduct to an inconsistent PORT-estimator. The choice $q = 0$ is tempting, but it must be used with some care, as it is the case for the PORT-Hill estimator, where the use of the minimum as a threshold produces an inconsistent estimator for γ, whenever the model underlying the data has an infinite left endpoint (see Gomes et al., 2007).

In Fraga Alves et al. (2009), the same methodology is applied to the Mixed Moment estimator, producing the ***PORT-Mixed Moment*** estimator, $\hat{\gamma}_{n,k}^{MM(q)} = \hat{\gamma}_{n,k}^{MM}(\underset{\sim}{X}^{(q)})$.

3.2.4 Semi-parametric estimation of other extreme events

As in Sections 3.1.2.3 and 3.1.3.4, related to parametric approaches, we can obtain semi-parametric estimates of other parameters of interest, based on the EVI-estimators. **Extreme quantiles** are one of those parameters and they were properly estimated following parametric approaches, in Section 3.1.

Defining again $\mathcal{X}_{1-p} = U(\frac{1}{p})$ as the extreme quantile of order $(1-p)$ of the underlying d.f. F of the r.v. X, with p small, and following now a semi-parametric approach, the only assumption is that $F \in \mathcal{D}(G_\gamma)$. Therefore, from relation (2.17) of Theorem 2.18, we have

$$F \in \mathcal{D}(G_\gamma) \Longleftrightarrow \lim_{t\to\infty} \frac{U(tx) - b(t)}{a(t)} = D_\gamma(x),$$

with $D_\gamma(x)$ defined in (2.18). Then, for a high t, we have

$$\begin{aligned} \frac{U(tx) - b(t)}{a(t)} \simeq D_\gamma(x) \Leftrightarrow \\ \Leftrightarrow U(tx) \simeq b(t) + a(t) D_\gamma(x) \Leftrightarrow \\ \Leftrightarrow U(x) \simeq b(t) + a(t) D_\gamma\left(\frac{x}{t}\right). \end{aligned}$$

We can use this approximation to estimate extreme quantiles $\mathcal{X}_{1-p}$, but, in order to make this approximation applicable, we need to estimate γ and the normalizing constants $a(t)$ and $b(t)$, under a semi-parametric framework.

Regarding the location attraction coefficient, $b(t)$, Theorem 2.19 states that (3.48) holds for $b(t) = U(t)$, with $U(.)$ defined in Definition 2.12. The same result can be obtained from Theorem 2.26 for the choice of the normalizing constants. As inference is made with $k+1$ intermediate order statistics, we can choose $t = \frac{n}{k}$, with $k = k_n \to \infty, \frac{k}{n} \to 0, n \to \infty$, yielding

$$U\left(\frac{1}{p}\right) \simeq U\left(\frac{n}{k}\right) + a\left(\frac{n}{k}\right) D_\gamma\left(\frac{k}{np}\right). \tag{3.48}$$

The quantity $U(\frac{n}{k})$ can be estimated by an intermediate order statistic, as it represents the extreme quantile of order $\left(1 - \frac{k}{n}\right)$ of the d.f. F. Since the empirical d.f. F_n defined in (3.38) is the natural estimator of the d.f. F, we get, for a large $t = \frac{n}{k}$,

$$\hat{b}\left(\frac{n}{k}\right) = \hat{U}\left(\frac{n}{k}\right) = \hat{F}^{-1}\left(\frac{n-k}{n}\right) = F_n^{-1}\left(\frac{n-k}{n}\right) = X_{n-k:n}. \tag{3.49}$$

Concerning the scale coefficient $a(\frac{n}{k})$, Ferreira et al. (2003) and de Haan and Ferreira (2006) present a suitable estimator for $a(\frac{n}{k})$, based on the Moment estimator and for the general case $\gamma \in \mathbb{R}$:

$$\hat{a}\left(\frac{n}{k}\right) = X_{n-k:n} M_{n,k}^{(1)} (1 - \hat{\gamma}_{n,k}^{-}), \tag{3.50}$$

with $M_{n,k}^{(j)}$ given in (3.45) for $j = 1$ and $\hat{\gamma}_{n,k}^{-}$ defined in (3.46).

Consequently, taking back (3.48), we can define the following estimators for the extreme quantiles $U\left(\frac{1}{p}\right)$:

1. for $\gamma \neq 0$,

$$\hat{U}\left(\frac{1}{p}\right) = X_{n-k:n} + \hat{a}\left(\frac{n}{k}\right) \frac{\left(\frac{k}{np}\right)^{\hat{\gamma}} - 1}{\hat{\gamma}}; \tag{3.51}$$

2. for $\gamma = 0$,

$$\hat{U}\left(\frac{1}{p}\right) = X_{n-k:n} + \hat{a}\left(\frac{n}{k}\right) \log\left(\frac{k}{np}\right),$$

with $\hat{a}\left(\frac{n}{k}\right)$ given by (3.50) and $\hat{\gamma}$ standing for any of the EVI-estimators of Section 3.2.3.

In particular, for the simpler case of heavy tails ($\gamma > 0$), a straightforward estimator of the extreme quantile $\mathcal{X}_{1-p} = U(\frac{1}{p})$ results from (3.48). From condition (2.21) of Theorem 2.19, it was seen that $a(t)$ is linked to $g(t)$. More specifically,

$$g(t) = a\left(\frac{1}{\overline{F}(t)}\right).$$

Therefore, for $t = X_{n-k:n}$, we have

$$g(X_{n-k:n}) = a\left(\frac{1}{\overline{F}(X_{n-k:n})}\right) \simeq a\left(\frac{n}{k}\right),$$

since we know that $\overline{F}(X_{n-k:n}) \simeq \overline{F}_n(X_{n-k:n}) = \frac{k}{n}$.

On the other hand, from (2.25), we have $g(t) = \gamma t$ for $\gamma > 0$. Therefore, after replacing γ by any of the EVI-estimators valid for $\gamma > 0$, we get

$$\hat{a}\left(\frac{n}{k}\right) = \hat{\gamma} X_{n-k:n}$$

and taking back (3.48), we get the estimator of $U(\frac{1}{p})$, for $\gamma > 0$:

$$\hat{U}\left(\frac{1}{p}\right) = X_{n-k:n}\left(\frac{k}{np}\right)^{\hat{\gamma}},$$

which is called the ***Weissman estimator*** (Weissman, 1978).

From (3.51), we can also estimate the **right endpoint** of the d.f. F. If $F \in \mathcal{D}(G_\gamma)$, for $\gamma < 0$, the right endpoint x^F is finite. Replacing p by zero in (3.51), we get

$$\hat{U}(\infty) = \hat{x}^F = \hat{U}\left(\frac{n}{k}\right) + \hat{a}\left(\frac{n}{k}\right) D_\gamma(\infty) = X_{n-k:n} - \frac{X_{n-k:n} M^{(1)}_{n,k}(1-\hat{\gamma}^-_{n,k})}{\hat{\gamma}}, \tag{3.52}$$

where the normalizing constants were replaced by their respective estimators, given in (3.49) and (3.50), and γ was replaced by any of the previous EVI-estimators, valid for $\gamma < 0$, or by $\hat{\gamma}^-_{n,k}$. But, as we have the obvious restriction $X_{n:n} \leq x^F$, incorporating this restriction in (3.52), we can define a more accurate estimator for the right endpoint:

$$\hat{x}^F = \max\left(X_{n:n}, X_{n-k:n}\left(1 - \frac{M^{(1)}_{n,k}(1-\hat{\gamma}^-_{n,k})}{\hat{\gamma}}\right)\right). \tag{3.53}$$

Note that, with the normalizing constants estimates, we can obtain a **tail probability** estimator using condition (3.39):

$$\widehat{\overline{F}}(x) = \frac{k}{n}\left(1 - H_{\hat{\gamma}}\left(\frac{x - X_{n-k:n}}{\hat{a}\left(\frac{n}{k}\right)}\right)\right), \; x > X_{n-k,n}. \tag{3.54}$$

This latter estimator may be used as an indicator of "excellence" of any value above $X_{n-k:n}$, since it measures the **exceedance probability** mentioned in (3.14). In particular, it can be used to measure the "quality" of the sample maximum $X_{n:n}$, through the probability of finding another sample maximum "better" than the current one.

3.2.5 The asymptotic normality of the estimators of the extreme value index and corresponding confidence intervals

All previous EVI-estimators enjoy the property of consistency. This property depends only on the behaviour of the intermediate sequence $k = k_n$ as $n \to \infty$. But in order to do some proper inference such as finding CI's for γ, we need to assume some distributional characterization for these estimators. However, the existence of an asymptotic non-degenerate distribution requires that the underlying d.f. F satisfies the second order condition discussed in Section 3.2.2.

Let $\hat{\gamma}^E_{n,k}$ be an arbitrary estimator of γ, like those presented in Section 3.2.3. From de Haan and Ferreira (2006), we obtain the following Theorem:

Theorem 3.16 *Suppose that the d.f. F satisfies the second-order condition in Theorem 3.8. Consequently, for an intermediate sequence $k = k_n$ such that* (3.36) *and* (3.37) *hold and also such that*

$$\lim_{n\to\infty} \sqrt{k} A\left(\frac{n}{k}\right) = \lambda, \tag{3.55}$$

with λ finite, $\exists \upsilon_E \in \mathbb{R}$ and $\sigma_E > 0$, such that

$$\sqrt{k}(\hat{\gamma}^E_{n,k} - \gamma) \quad \xrightarrow[n\to\infty]{d} \quad Z \frown \mathcal{N}(\lambda \upsilon_E, \sigma^2_E). \tag{3.56}$$

Some specific extra mild conditions may be needed for some estimators $\hat{\gamma}^E_{n,k}$.

The values υ_E and σ^2_E are components of the *asymptotic bias* and of the *asymptotic variance* of $\hat{\gamma}^E_{n,k}$, respectively. Estimating these quantities allows us to establish CI's for γ, based on any estimator $\hat{\gamma}^E_{n,k}$. We can find expressions for υ_E and σ^2_E in Chapter 3 of de Haan and Ferreira (2006) or in Chapters 4 and 5 of Beirlant et al. (2004), for the general case of the auxiliary function $A(t)$ such that $|A(t)| \in \mathcal{RV}_\rho$. For instance, in the case of the Hill estimator, we have:

$$\upsilon_H = \frac{1}{1-\rho} \quad \text{and} \quad \sigma^2_H = \gamma^2$$

and, consequently,

$$\sqrt{k}(\hat{\gamma}^H_{n,k} - \gamma) \quad \xrightarrow[n\to\infty]{d} \quad Z \frown \mathcal{N}\left(\frac{\lambda}{1-\rho}, \gamma^2\right).$$

The construction of approximate CI's requires then the choice of k and the knowledge of ρ, from the auxiliary function $A(t)$, and of γ itself! In many applications of EVT, it is often assumed the special case $A(t) = \beta t^\rho$ for the auxiliary function, since $A(t) \in \mathcal{RV}_\rho$, valid only for $\rho < 0$. But this assumption introduces one more parameter to be known if

we want to construct proper CI's. The presence of unknown parameters in the asymptotic Normal distribution is a not problem only for the Hill estimator, since it is a common characteristic in every estimator of γ.

The estimation of ρ and β is not an easy task. Among the articles concerned with the estimation of ρ, we can refer Gomes et al. (2002) and Fraga Alves et al. (2003). Suitable estimators for β can be found in Gomes and Martins (2002) and Caeiro and Gomes (2006). As mentioned, the estimation of the second-order parameters is complex and some questions are still open in this estimation. For the calculus of CI's for γ, de Haan and Ferreira (2006) recommend the assumption $\lambda = 0$ in (3.55), i.e. $\lim_{n\to\infty} \sqrt{k} A\left(\frac{n}{k}\right) = 0$, so that the limiting distribution in (3.56) has zero mean. This avoids then the bias estimation, υ_E, which usually is a function of ρ and β (and possibly γ). The $100(1-\alpha)\%$ approximate CI for γ is then given by

$$\hat{\gamma}^E_{n,k} - z_{1-\alpha/2}\sqrt{\frac{\sigma^2_E}{k}} < \gamma < \hat{\gamma}^E_{n,k} + z_{1-\alpha/2}\sqrt{\frac{\sigma^2_E}{k}}, \tag{3.57}$$

where z_ε is the ε-quantile of the Normal distribution.

As σ^2_E is a function of γ, it is replaced by its estimator $\hat{\sigma}^2_E$, which is obtained replacing γ by its estimate in the expression of σ^2_E.

Under the conditions of the Theorem 3.16, we can also obtain a CI for the right endpoint of the underlying d.f. F (c.f. de Haan and Ferreira, 2006):

$$X_{n:n} < x^F < \hat{x}^F_E + z_{1-\alpha}\frac{\hat{a}\left(\frac{n}{k}\right)}{(\hat{\gamma}^E_{n,k})^2}\sqrt{\frac{\hat{\sigma}^2_E}{k}}, \tag{3.58}$$

with $\hat{x}^F_E$ given by (3.53) for the respective estimator $\hat{\gamma}^E_{n,k}$ and $\hat{a}\left(\frac{n}{k}\right)$ given by (3.50). Remember that the right endpoint x^F cannot lie under the sample maximum and, as a matter of fact, the left-hand side of the CI must be limited by $X_{n:n}$. Therefore, the confidence level of the interval is undetermined, but lower than $1-\alpha$.

We can define the following expressions of σ^2_E for each estimator discussed in Section 3.2.3:

$$
\begin{aligned}
\sigma_P^2 &= \begin{cases} \dfrac{\gamma^2\left(2^{2\gamma+1}+1\right)}{(\log 2)^2\left(2^\gamma-1\right)^2}, & \text{if } \gamma \neq 0, \\ \frac{3}{(\log 2)^4} & \text{if } \gamma = 0; \end{cases} \\
\sigma_H^2 &= \gamma^2, \ \text{ if } \gamma > 0; \\
\sigma_M^2 &= \begin{cases} \gamma^2+1 & \text{if } \gamma \geq 0, \\ \dfrac{(1-\gamma)^2(1-2\gamma)(1-\gamma+6\gamma^2)}{(1-3\gamma)(1-4\gamma)}, & \text{if } \gamma < 0; \end{cases} \\
\sigma_{NH}^2 &= \gamma^2, \ \text{ if } -1 < \gamma < -0.5; \\
\sigma_{GH}^2 &= \begin{cases} \gamma^2+1 & \text{if } \gamma \geq 0, \\ \dfrac{(1-\gamma)(1+\gamma+2\gamma^2)}{1-2\gamma} & \text{if } \gamma < 0; \end{cases} \\
\sigma_{MM}^2 &= \begin{cases} (1+\gamma)^2 & \text{if } \gamma \geq 0, \\ (1-2\gamma)^4 \dfrac{(1-\gamma)^2(6\gamma^2-\gamma+1)}{(1-2\gamma)^3(1-3\gamma)(1-4\gamma)} & \text{if } \gamma < 0; \end{cases} \\
\sigma_{H(q)}^2 &= \sigma_H^2; \\
\sigma_{M(q)}^2 &= \sigma_M^2; \\
\sigma_{MM(q)}^2 &= \sigma_{MM}^2.
\end{aligned}
\tag{3.59}
$$

All these expressions can be found in de Haan and Ferreira (2006), for Pickands, Hill, Moment and Negative Hill estimators. For the Generalized Hill estimator, we refer to Beirlant et al. (2005) and for the Mixed Moment estimator, our reference is Fraga Alves et al. (2009). Finally, the identities for the PORT versions of the estimators are found in Araújo Santos et al. (2006) and Fraga Alves et al. (2009).

It must be remembered that these variances are only valid if Theorem 3.56 is satisfied and if extra specific conditions for each estimator are fulfilled, too. All these conditions will be assumed in this thesis.

The choice of k still remains and will be discussed in Section 3.2.7.

3.2.6 Testing the extreme value index sign

All the results for the semi-parametric approach rely on the extreme value condition, i.e., assume that the underlying d.f. F belongs to the max-domain of attraction of the GEVd, or equivalently, that the tail of the underlying d.f. F is close to the tail of the GPd, in some sense. As we have seen, some of the semi-parametric estimators require the knowledge about the tail's type of the underlying d.f., described by the sign of the

EVI. Therefore, if we can do some *a priori* assumptions about the most appropriate type of decay of the d.f.'s tail, then, every kind of statistical inference will be greatly improved. For $\gamma < 0$, we know that the underlying d.f. has a light right tail, with a finite right endpoint x^F, and for $\gamma > 0$, it has a heavy right tail with an infinite right endpoint. A previous conjecture about the sign of γ help us to select a specific procedure of estimation, more reliable than procedures valid for a general $\gamma \in \mathbb{R}$. Note, for instance, that the estimator of the right endpoint is only applicable if $\gamma < 0$.

The case $\gamma = 0$ is of particular interest, as it can be seen as a transitional value of the EVI, separating light right tails with finite right endpoint from heavy right tails with no finite right endpoint. As mentioned, this case is also attractive because the Gumbel max-domain attracts a great variety of d.f.'s having an exponential type of right tail's decay and because of the great simplicity of inference within the Gumbel max-domain. Therefore, as well as in a parametric context, separating statistical inference procedures according to the most appropriate max-domain of attraction has gained much interest in a semi-parametric context. With the rise of the PORT methodology, tests for Gumbel versus non-Gumbel max-domain have received recent attention from statisticians.

This subject was covered by several articles since the eighties, specially within a parametric framework, and some of the tests presented were treated in Section 3.1. The tests developed under a semi-parametric approach are based on the k excesses over the random threshold $X_{n-k:n}$, with k satisfying (3.36) and (3.37):

$$Z_i = X_{n-i+1:n} - X_{n-k:n}, \quad i = 1, \cdots, k. \tag{3.60}$$

In this Section, we present the articles of Neves et al. (2006), Neves and Fraga Alves (2007) and Neves and Fraga Alves (2008), which present three statistics for the following test:

$$H_0 : F \in \mathcal{D}(G_0) \quad vs. \quad H_1 : F \in \mathcal{D}(G_\gamma)_{\gamma \neq 0} \tag{3.61}$$

or against one-sided alternatives:

$$H_0 : F \in \mathcal{D}(G_0) \quad vs. \quad H_1 : F \in \mathcal{D}(G_\gamma)_{\gamma > 0} \tag{3.62}$$

or

$$H_0 : F \in \mathcal{D}(G_0) \quad vs. \quad H_1 : F \in \mathcal{D}(G_\gamma)_{\gamma < 0}. \tag{3.63}$$

The designed three statistics for testing (3.61) (or the one-sided alternative versions) are based on the k excesses defined in (3.60) and presented below:

(i) the Greenwood test statistic, $R_n(k)$:

$$R_n(k) = \frac{k^{-1}\sum_{i=1}^{k} Z_i^2}{\left(k^{-1}\sum_{i=1}^{k} Z_i\right)^2}, \tag{3.64}$$

(ii) the Hasofer-Wang test statistic, $W_n(k)$:

$$W_n(k) = k^{-1}\frac{\left(k^{-1}\sum_{i=1}^{k} Z_i\right)^2}{k^{-1}\sum_{i=1}^{k} Z_i^2 - \left(k^{-1}\sum_{i=1}^{k} Z_i\right)^2} = \frac{1}{k}\frac{1}{R_n(k)-1}, \tag{3.65}$$

(iii) the Ratio test statistic, $T_n(k)$:

$$T_n(k) = \frac{Z_1}{k^{-1}\sum_{i=1}^{k} Z_i}. \tag{3.66}$$

The three test statistics are location/scale invariant, since they are based on ratios and spacings of high order statistics.

The Greenwood test statistic was introduced in 1946, turning out to be useful for detecting the presence of heavy-tailed distributions, while the Hasofer-Wang test statistic dates back to 1992 and was built on the Shapiro-Wilk goodness-of-fit statistic, being the most powerful of the three, for tests concerning alternatives in the Weibull max-domain of attraction.

Neves and Fraga Alves (2007) reformulated the asymptotic properties of these two statistics, when $k = k_n$ behaves as an intermediate sequence, rather than remaining fixed when the sample size n increases (as it was the case originally). The Ratio test statistic was introduced by Neves et al. (2006), as a complementary test statistic, motivated by the different contributions of the sample maximum to the sum of the k excesses above the random threshold. Since the test based on this statistic tends to be conservative but with a reasonable power, it is seen as a good complement to the remainder statistics. Its asymptotic properties are properly exposed in the aforementioned article.

The normalized versions of (3.64), (3.65) and (3.66) are, respectively:

$$R_n^*(k) = \sqrt{\frac{k}{4}}\left(R_n(k) - 2\right), \tag{3.67}$$

$$W_n^*(k) = \sqrt{\frac{k}{4}}\left(kW_n(k) - 1\right), \tag{3.68}$$

$$T_n^*(k) = T_n(k) - \ln k. \tag{3.69}$$

Under the null hypothesis that $F \in \mathcal{D}(G_0)$ and under extra second order conditions on the right tail of F and on the growth of convergence of the intermediate sequence k_n to infinity, the normalized statistics converge in distribution according to:

$$R_n^* \xrightarrow[n\to\infty]{d} Z_1 \frown \mathcal{N}(0,1),$$

$$W_n^* \xrightarrow[n\to\infty]{d} Z_2 \frown \mathcal{N}(0,1),$$

$$T_n^* \xrightarrow[n\to\infty]{d} Z_3 \frown \Lambda.$$

The critical region of asymptotic size α, for the two-sided test (3.61), is given by

$$V_n^* < \upsilon_{\alpha/2} \text{ or } V_n^* > \upsilon_{1-\alpha/2}, \tag{3.70}$$

where V has to be conveniently replaced by T, R or W and υ_ε denotes the ε-quantile of the corresponding limiting distribution.

Following a one-sided alternative approach, the critical regions of asymptotic size α, for the one-sided tests (3.62) and (3.63), are:

1. for the Fréchet max-domain of attraction ($H_1 : \gamma > 0$)

$$R_n^*(k) > z_{1-\alpha}, \qquad W_n^*(k) < z_\alpha, \qquad T_n^*(k) > \mathcal{G}_{1-\alpha}; \tag{3.71}$$

2. for the Weibull max-domain of attraction ($H_1 : \gamma < 0$)

$$R_n^*(k) < z_\alpha, \qquad W_n^*(k) > z_{1-\alpha}, \qquad T_n^*(k) < \mathcal{G}_\alpha, \tag{3.72}$$

where z_ε and $\mathcal{G}_\varepsilon$ are the ε-quantiles of the standard Normal and Gumbel distributions, respectively.

3.2.7 The adaptive selection of the tail sample fraction

As mentioned in Section 3.2.1, an important issue in semi-parametric approaches is the consideration of k, which defines the tail sample fraction. Its "correct choice" is crucial for the semi-parametric estimators to have desirable properties in order to do proper inference. It was seen that the choice of k must be indexed to the sample size n, so that when the sample size increases, so does the value of k. At least, a correct choice must satisfy the following properties:

$$k = k_n \to \infty, \frac{k}{n} \to 0 \text{ when } n \to \infty.$$

Therefore, k must be large enough, but not too large: it must increase moderately as the sample size increases. If k is sufficiently low, we stay close to the sample maximum and few order statistics will be used, resulting in estimators with large variances. If k is sufficiently high, the number of order statistics used increases, allowing a decrease of estimators' variances but resulting in a larger bias, since we are using observations which do not actually converge to the hypothesized limiting d.f. Indeed, from de Haan and Ferreira (2006), we have the following result:

Theorem 3.17 (**de Haan and Ferreira, 2006, Theorem 2.2.1**) *Suppose von Mises' conditions for the max-domain of attraction of an extreme value distribution G_γ are fulfilled (cf. Theorem 2.10). Then, if $k = k_n \to \infty, \frac{k}{n} \to 0$ when $n \to \infty$,*

$$\sqrt{k}\frac{X_{n-k:n} - U(\frac{n}{k})}{\frac{n}{k}U'(\frac{n}{k})} \quad \xrightarrow[n\to\infty]{d} \quad Z \frown \mathcal{N}(0,1).$$

The optimal value is then a result of balancing bias and variance, which is not so trivial. Several methods have been proposed, most of them by the authors of the articles that introduce estimators for extreme value parameters. For example, in his article, Pickands III (1975) suggests a specific criterion for choosing k, together with his estimator, but that method was never widely adopted, contrary to the estimator itself. Embrechts et al. (1997) proposed a more practical way for the choice of k: the *Pickands-plot*. For each value of $k = 1, 2, \ldots, n$, we calculate the Pickands estimator and plot it against k. The range of k's that corresponds to a *plateau*, i.e. those values of k that correspond to a reasonable horizontal plot, is considered for an elective value of the estimator. The same strategy is proposed for the Hill estimator, but Embrechts et al. (1997) warn us that the results of Hill-plots can be very misleading. They refer to these plots as "Hill-horror plots": due to the high volatility in Hill-plots, they are far from being constant and it may be very difficult to choose the range where a stable plot is evident. Drees et al. (2000) proved that Hill-plots are most effective only when the underlying d.f. is Pareto or very close to it.

One of the most serious objections raised against semi-parametric methods is their sensitivity towards the choice of k. Consequently, in the literature, many efforts have been made to find the optimal k that achieves the best compromise between bias and variance. An important criterion, very popular among statisticians and used in most of the articles, is choosing k in order to minimize the ***Asymptotic Mean Squared Error (AMSE)***. However, as we will see, this criterion is mainly dependent on the second order assumptions about the underlying d.f. F discussed in Section 3.2.2. The optimal k can be determined when the underlying d.f. F is known, provided that F has a second order expansion involving extra parameters. But in practice, we do not know the

exact analytical form of the underlying d.f. and the extra parameters are very difficult to estimate. Still, Dekkers and de Haan (1989) proved that the second order conditions hold for most of the well-known d.f.'s.

For an arbitrary estimator of γ, $\hat{\gamma}^E_{n,k}$, like those presented in Section 3.2.3, we define the AMSE as follows:

$$AMSE(\hat{\gamma}^E_{n,k}) = AVar(\hat{\gamma}^E_{n,k}) + \left(ABias(\hat{\gamma}^E_{n,k})\right)^2,$$

where AVar and ABias stand for *Asymptotic Variance* and *Asymptotic Bias*, respectively.

In Section 3.2.5, we saw that any arbitrary estimator $\hat{\gamma}^E_{n,k}$ has a limiting Normal distribution and that σ^2_E and υ_E are components of AVar and ABias, respectively. So, from (3.56), we get

$$AVar(\hat{\gamma}^E_{n,k}) = E_\infty\left((\hat{\gamma}^E_{n,k} - \gamma)^2\right) = \frac{\sigma^2_E}{k}$$

and

$$ABias(\hat{\gamma}^E_{n,k}) = E_\infty(\hat{\gamma}^E_{n,k} - \gamma) = \frac{\lambda \upsilon_E}{\sqrt{k}},$$

where E_∞ stands for the asymptotic mean value. Then, as an optimality criterion for k, we search for the value which minimizes the AMSE plot $\{(k, AMSE(\gamma^E_{n,k}))\}$.

Unfortunately, as it was also seen in Section 3.2.5, σ^2_E is a function of γ and υ_E may eventually depend on γ. We can replace γ by its estimator $\hat{\gamma}^E_{n,k}$, but this latter presumes that the value of k has already been chosen, which is precisely our problem now! The choice of k can then be made recursively, but it induces a high volatility in its estimate and an important loss of efficiency. In order to avoid such issue, other methods of determining k have been recently used, such as bootstrapping methods (cf. Gomes and Oliveira, 2001) and regression methods (cf. Beirlant et al., 2004) for an optimal adaptive choice of k.

In this thesis, we will follow the heuristic methodology proposed by Henriques-Rodrigues et al. (2011). Let $\hat{\gamma}^{(i)}_{n,k}$, $i \in \mathcal{E} = \{1, 2, 3, 4, 5, 6, 7\}$, be the set of the seven EVI-estimators of Section 3.2.3. Then consider

$$k^{opt} = \arg\min_k \sum_{(i,j)\in\mathcal{E}: i\neq j} \left(\hat{\gamma}^{(i)}_{n,k} - \hat{\gamma}^{(j)}_{n,k}\right)^2 \tag{3.73}$$

the optimal value chosen for k, as we expect that there will be a region where all the estimators are concordant, i.e., a region where all the estimates have close values and the smallest value possible for (3.73). According to Henriques-Rodrigues et al. (2011), (3.56) still holds if we replace k with k^{opt}.

The same procedure can be used for the estimation of the right endpoint x^F or for the exceedance probability.

Chapter 4

Case Studies

4.1 The Maximal Oxygen Uptake or $\dot{V}O_2max$

The maximal oxygen uptake, or $\dot{V}O_2max$, represents the maximum amount of oxygen an individual can absorb and use during an intense physical exercise, per unit time. It can be expressed either as an absolute value, in litres of oxygen per minute (l/min), or as a relative value, in millilitres of oxygen per kilogram of bodyweight per minute (ml/kg/min). This latter unit is the most commonly used in sports for athletes' comparisons. $\dot{V}O_2max$ is a useful quantitative measure for the capacity of sports' practitioners, since it is directly related to the ability of realize a high intensity exercise at a high effort's level, for extra four or five minutes. When the $\dot{V}O_2max$ is attained during an intense physical effort, any additional increment of the exercise's intensity has no effect (or only a residual one), on the consumption of oxygen.

The knowledge of this quantity is crucial in sports where "pain tolerance", "obstinacy" or "fight capacity" are essential to surpass discomforts of exhaustive exercises. A very high level of $\dot{V}O_2max$ results in a better cardiorespiratory capacity and subsequent ability to produce energy at high levels of efforts. It is, then, more useful as an indicator of personal aerobic potential, rather than a predictor of success in high efforts exercises. Other factors are identically important to obtain a better performance, such as efficiency, recovery, confidence, genetics, among others. Thus, taken on its own, $\dot{V}O_2max$ is a rough guide of an athlete's potential. However, it remains one of the indicators most studied and of most interest for athletes.

In an average sedentary and healthy adult male, the $\dot{V}O_2max$ runs around 45 ml/kg/min, but in a high level athlete, it can reach values between 80 and 90 ml/kg/min. For women, the same measure rounds 35 ml/kg/min in sedentary and healthy adult fe-

males, attaining scores around 75 ml/kg/min in high level female athletes. Athletes who attain the highest levels of $\dot{V}O_2max$ are the cross-country skiers, the cross-country runners and the road racing cyclists. $\dot{V}O_2max$ is then highly related to sex and to sports modality. Typically, women achieve $\dot{V}O_2max$ scores 15 to 30% below those of men of the same category. The sport's category is also an important variable, which is highly correlated with the level of $\dot{V}O_2max$, depending on the amount of oxygen solicited per unit time, dictated by the sport's category involved. A common soccer player, for instance, reach levels between 42 and 60 ml/kg/min, while a typical cross-country skier can rise his $\dot{V}O_2max$ beyond 90 ml/kg/min. In sports where endurance is an important component, athletes usually have higher levels of $\dot{V}O_2max$. Currently, the world's record belongs to the Norwegian cross-country skiers Bjørn Dæhlie and Espen Harald Bjerke, with a $\dot{V}O_2max$ of 96 ml/kg/min. However, according to the physiologist Erlend Hem, Dæhlie's score was obtained out of season and he believes that the skier can exceed 100 ml/kg/min at the peak of his physical form.

In this thesis, we have no intention to conduct a regression analysis on $\dot{V}O_2max$, in order to study the effect of the most important factors affecting its levels. Instead, the scope of this Section is to apply EVT to a random sample of $\dot{V}O_2max$, obtained from a population of world athletes. To reduce as much as possible the impact of "confounding variables" and respect the i.i.d rule, we focus only on masculine athletes picked up in three categories: cross-country skiing, cross-country running and road racing cycling. As seen, these modalities produce the highest levels of $\dot{V}O_2max$, which are of great interest, since we are dealing with extremes, in particular, maxima. Our intention is to work with a population as homogeneous as possible and obtain a collection of i.i.d random variables.

The dataset consists of a list of 74 observations collected from several fonts, such as Saltin and Åstrand (1967), Bangsbo and Larsen (2001), Noakes (2003) and McArdle et al. (2009). Apart from the aforementioned works, the data were also gathered from a set of websites listed below:

1. `http://www.letsrun.com`, with the data available in the following pages:
   - `http://www.letsrun.com/forum/flat_read.php?thread=4691852`
   - `http://www.letsrun.com/forum/flat_read.php?thread=1477013`
   - `http://www.letsrun.com/forum/flat_read.php?thread=1477013&page=2`
   - `http://www.letsrun.com/forum/flat_read.php?thread=4858418`
2. `http://www.brianmac.co.uk/vo2max.htm`
3. `http://www.topendsports.com/testing/records/vo2max.htm`

4. http://www.cyclisme-dopage.com/portraits/armstrong.htm

5. http://www.sportvital.cz/sport/trenink/vo2-max-meritko-nasi-kondice/

Each observation represents the highest $\dot{V}O_2max$ obtained by an athlete, picked up from one of the three categories mentioned above. We assume then homogeneity between these categories. In sports measures, it is very frequent to obtain repeated values for several measurements. This occurs because of the lack of precision of the measurement's instruments. So, in order to avoid estimation problems due to values' discretization, we use the same smoothing technique than Einmahl and Magnus (2008) and Einmahl and Smeets (2011). For instance, when r athletes have a $\dot{V}O_2max$ of 84 ml/kg/min, these r results are smoothed equally over the interval $]83.95, 84.05[$ as follows:

$$\dot{V}O_2max_j = 83.95 + 0.1\frac{2j-1}{2r}, \quad j = 1, \ldots, r.$$

4.1.1 Parametric data analysis

The first parametric approach we will follow on this first case study is the Gumbel's *Block Maxima method*, presented in Section 3.1.2. Under this methodology, we can consider each athlete as an individual block. For each athlete, we pick up his highest $\dot{V}O_2max$ and, therefore, we obtain a collection of several maxima, as much as the number of blocks considered, in this case, 74 blocks, provided that we have 74 athletes. However, it is important to clarify one essential point. The observed sample is a sample of *maxima*, since we keep only the highest value of each athlete. Recalling the notation of Section 3.1.2, we represent the characteristic under study by the r.v. X, which represents the maximum amount of oxygen consumed per unit time (our $\dot{V}O_2max$), with unknown d.f. F. As this quantity can be observed repeatedly on the same athlete, for every individual, we can observe a collection of k $\dot{V}O_2max$, possibly with a high level of correlation. Considering m athletes, we have then m blocks of k observations. But, provided that we keep only the highest result of each athlete, represented by the r.v. Y defined in (3.1), each block consists only of one observation. Therefore, our observed random sample $(y_1, \ldots, y_m)$ is formed by $m = 74$ blocks of one individual observation. It should be noted that we *do not have access* to the k observations of each athlete and, therefore, we cannot accede to the original r.v. X. Every kind of inference is then solely made in terms of the r.v. Y, which represent a maximum itself: the maximal $\dot{V}O_2max$ of an athlete of our population under study.

The second parametric approach presented in Section 3.1.3, the *POT methodology*, may seem unapplicable in this case study, provided that our sample has a modest size

of 74 observations and that the consideration of a threshold will exclude observations, reducing the size of the available sample. Conscious of this handicap and of the modest size of our sample, we still decided to conduct a POT approach, as a matter of comparison with the Gumbel's approach and to see what this methodology has to tell us in a modest sample context. We will use exactly the same sample as for Gumbel's approach, since the POT methodology is only valid if we are working with a random sample. Since each athlete has only one observation, his current highest $\dot{V}O_2max$, we can then assume that the i.i.d rule is satisfied on the construction of our sample. But before proceeding further with the POT methodology, it is important to emphasize a noticeable difference between the two parametric approaches, namely concerning the definition of the r.v. Y. The Gumbel's approach considers the available sample as a realization of the r.v. Y, defined in (3.1). Therefore, the random sample at hand is considered as a sample of m *maxima*, replicas of the r.v. Y, which is assumed to follow a GEVd. The POT methodology considers that the sample at hand is a realization of the r.v. X under study, which follows a d.f. F. Consequently, this methodology assumes that the available data are replicas of X taken from the d.f. F, in particular from the right tail of F, since we are working with top results. Therefore, focusing on the observations above some high threshold u, we assume we can fit a GPd to the excesses over that threshold, represented by the r.v. Y, discussed in Section 3.1.3.1. Since the choice of the fixed threshold u is the central point of the POT approach, we can use the sample ME-plot described in Section 3.1.3.7 to select the most appropriate threshold. This plot can be seen in Figure 4.1, obtained with the **R** software (see Appendix A.1).

Remember from (3.32), that for the GPd, defined in (3.24), the mean excess function is given by

$$e(u) = E(X - u|X > u) = E(Y|Y > 0) = \frac{\sigma_u + \gamma u}{1 - \gamma},\ \gamma < 1.$$

Therefore, if we assume correctly the GPd as the underlying parametric model for the excesses defined by the r.v. Y, the ME-plot should follow a straight line with intercept $\frac{\sigma_u}{1-\gamma}$ and slope $\frac{\gamma}{1-\gamma}$. Since $e(u)$ is estimated by its empirical counterpart, $\hat{e}_n(u)$, defined in (3.34), we expect the sample ME-plot to conserve the linear property, exhibiting a linear pattern of its path. Consequently, according to Davison and Smith (1990), if we are able to identify a point one the plot, above which the sample path is reasonably linear, then we have found an appropriate threshold, meaning that the excesses Y over this threshold follow a GPd.

Now, Figure 4.1 reveals a clear decreasing linear pattern of the whole sample path. If seems then difficult to find a point above which the sample path is roughly linear. However, taking a closer look at the sample path of Figure 4.1, we discover genuinely two

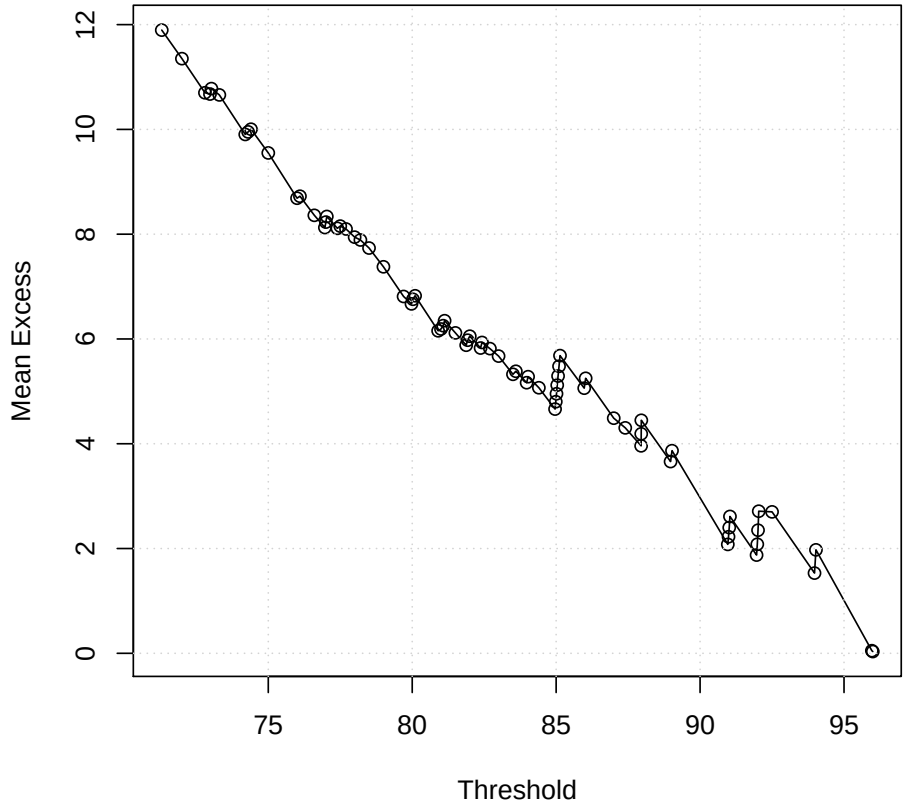

Figure 4.1: *Sample ME-plot for the $\dot{V}O_2max$ data.*

linear trends and not only one, as it seemed at first look. Fitting then a straight line to the first 25 observations and another straight line to the remaining observations of the ordered sample, we obtain Figure 4.2, by means of the **R** software (see Appendix A.2).

As the fitted lines show, we denote a change in the linear trend slope around 80 ml/kg/min. Surprisingly, this value is not innocent. Indeed, in 1954, Per-Olof Åstrand, one of the founding fathers of modern exercise physiology and pioneer in $\dot{V}O_2max$ studies, created with Irma Ryhming the famous *Åstrand-Ryhming nomogram*, a graphical calculator which allows to estimate a personal $\dot{V}O_2max$, based on the heart frequency and the exercise intensity. For more details, we refer to Åstrand and Ryhming (1954). With the help of the nomogram, the authors defined a classification chart with several levels, according to the personal $\dot{V}O_2max$. On the top of the chart, we find the value 80 ml/kg/min, above which an athlete is considered an "exceptional athlete". Now, this top value of the chart is precisely the value detected on the sample ME-plot. For this reason, we will choose $u = 80$ as the fixed threshold for our data and fit a GPd to the excesses above this threshold. Since we have 49 observations above the threshold $u = 80$, we end

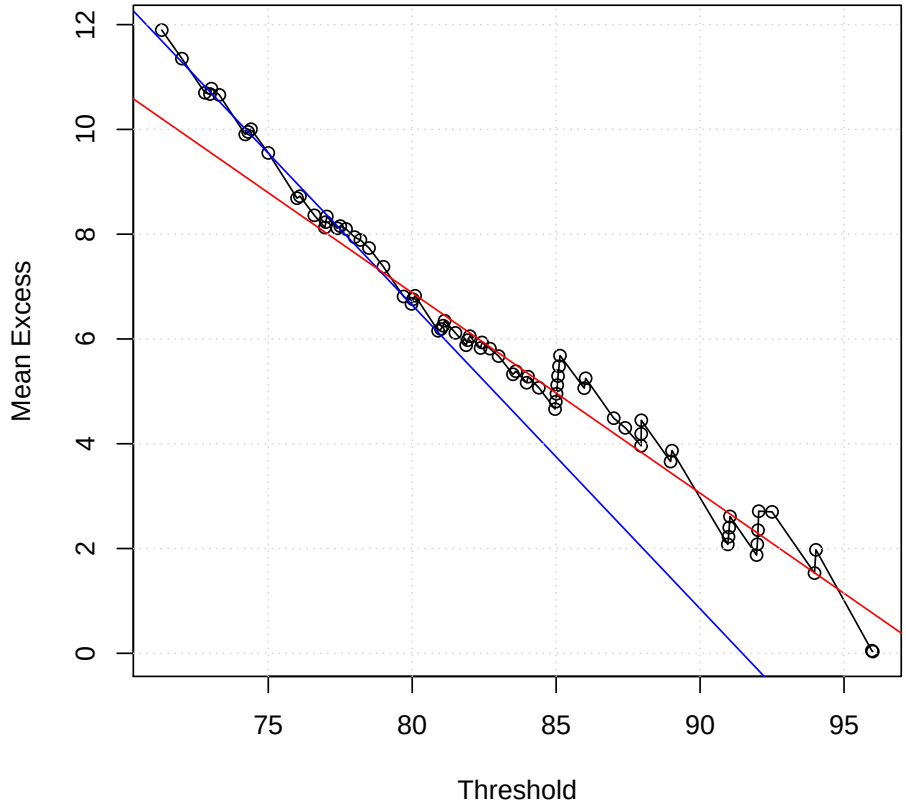

Figure 4.2: *Sample ME-plot for the $\dot{V}O_2max$ data, with fitted straight lines.*

up with a sample of $m = 49$ excesses. We alert once more that the consideration of this threshold will reduce the sample size, but we cannot remain indifferent to the coincidence between the threshold detected by the ME-plot and the top level of the Åstrand-Ryhming nomogram.

1) The Block Maxima method

a) *Preliminary statistical analysis*

As a starting point before Gumbel's approach, it is very useful to have an idea about the right tail of the underlying d.f. F. A preliminary graphical analysis may help us to suspect which type of right tail is probably at play in our d.f. F, linking us to one of the three types of extreme value distributions defined in Theorem 2.6. We remember that Weibull max-domain contains light right-tailed distributions with a finite right endpoint, while Fréchet max-domain contains heavy right-tailed distributions with no finite right

endpoint. The Gumbel max-domain acts as a boundary between the previous two types, comprising exponential right-tailed distributions with finite or infinite right endpoint. Then, it seems reasonable to start with Gumbel max-domain and check if the balance tilts significantly towards one of the sides. Moreover, from the relation between the GEVd and the GPd stated by the Pickands-Balkema-de Haan Theorem (cf. Theorem 3.4), we know that a d.f. which belongs to Gumbel max-domain of attraction has a right tail that can be modeled by an Exponential d.f. Since available data are considered extremal in some sense, we can try to fit an Exponential model to the data and ascertain whether such a distribution can be suitably fitted to the right tail of the unknown d.f. F, so that $\overline{F}\left(\frac{x-\lambda}{\delta}\right) \simeq \exp\left(-\frac{x-\lambda}{\delta}\right)$ for large x, with $x > \lambda$. The goodness-of-fit of the Exponential distribution to the right tail of F can be easily checked by eye with a classical graphical tool used for this purpose: the *Quantile-Quantile plot* (shortly, QQ-plot). The QQ-plot relies on an important property, which characterizes important classes of distributions: the theoretical quantiles of a parametric family are linearly related to the corresponding quantiles of the standard member of this family.

Turning back to our case study under the Block Maxima approach, as the available data are m replicas of the r.v. Y, they are considered large, since the sample is a sample of m maxima. Therefore, if the underlying d.f. is exponential right-tailed, we expect that $\overline{F}\left(\frac{y-\lambda}{\delta}\right) \simeq \exp\left(-\frac{y-\lambda}{\delta}\right)$. As stated before, the linearity property behind the QQ-plot is verified for important classes of distributions. This is the case namely for location and scale parameter families, within which we find the Exponential distribution. Let then $Q_{\lambda,\delta}(p) = F^{-1}(p)$ be the theoretical quantiles of order p of any location-scale parametric family, with $p \in]0,1[$. In particular, for the Exponential distribution, we have

$$Q_{\lambda,\delta}(p) = \lambda - \delta \log(1-p), \quad 0 < p < 1.$$

The quantiles of the corresponding standard member of the class can then be defined by $Q_{0,1}$, yielding, for the Exponential distribution,

$$Q_{0,1}(p) = -\log(1-p), \quad 0 < p < 1.$$

The aforementioned linear relationship between quantiles is thus visible for the Exponential distribution, since we have

$$Q_{\lambda,\delta}(p) = \lambda + \delta Q_{0,1}(p), \quad 0 < p < 1.$$

Hence, for the Exponential case, the plot of $Q_{\lambda,\delta}(p)$ versus $Q_{0,1}(p)$ produces a straight line, with slope and intercept given by δ and λ, respectively. However, since the location and scale parameters are unknown, we cannot obtain the theoretical quantiles, $Q_{\lambda,\delta}(p)$,

needed for the plot. Thus, information about these quantiles can only be found in our available data, here represented by $(y_1, \ldots, y_m)$. It is well known that the natural unbiased estimator of any d.f. F is the empirical d.f. F_n, defined in (3.38). Then, with our available data $(y_1, \ldots, y_m)$, the theoretical quantiles of order p, $Q_{\lambda,\delta}(p)$, can be estimated by the empirical quantiles of order p, yielding

$$\hat{Q}_{\lambda,\delta}(p) = F_m^{-1}(p) = \inf\{y : F_m(y) \geq p\}, \quad 0 < p < 1.$$

If $Y \frown Exp(\lambda, \delta)$, the linear relationship between $Q_{\lambda,\delta}(p)$ and $Q_{0,1}(p) = -\log(1-p)$ still holds approximately, if we replace $Q_{\lambda,\delta}(p)$ by $\hat{Q}_{\lambda,\delta}(p)$. We have just to define the values for p in order to trace the plot. Since we have m data points, the coordinates $(-\log(1-p), \hat{Q}_{\lambda,\delta}(p))$ can be plotted for m values of $p \in]0,1[$. More precisely, we can plot the points $(-\log(1-p_i), \hat{Q}_{\lambda,\delta}(p_i)), i = 1, \ldots, m$, where p_i are called *plotting positions*. The most common choice for p_i is $p_i = \frac{i}{m+1}, i = 1, \ldots, m$, yielding $\hat{Q}_{\lambda,\delta}(p_i) = y_{i:m}$ as estimates of the theoretical quantiles. The plotted points $(-\log(1-p_i), y_{i:m})$ produce then the QQ-plot presented in Figure 4.3, with the help of the **R** software (see Appendix A.3).

As we can see, the plot exhibits a concave pattern specially for large values of $y_{i:m}$, which means that replacing $Q_{\lambda,\delta}(p_i)$ by $\hat{Q}_{\lambda,\delta}(p_i) = y_{i:m}$ ends up with the linear relationship between quantiles. Therefore, the sample $(y_1, \ldots, y_m)$ may not have been generated by an Exponential model, which cannot be fitted to the right tail of the $\dot{V}O_2max$.

According to Beirlant et al. (2004), a concave pattern of the QQ-plot suggests that the underlying d.f. F has a lighter right tail than expected from an Exponential distribution. Thus, a suitable parametric model for Y must have a lighter right tail than the Exponential distribution, as it is the case for some d.f.'s of the Gumbel max-domain, such as the Normal or Gumbel distributions, and for all the d.f.'s of Weibull max-domain.

In a parametric approach, all the inference relies on the parameters' estimates of a suitable parametric model, obtained from point estimation methods. Therefore, to pursue a parametric analysis, we must select a suitable parametric model for our data. As our random sample $(Y_1, \ldots, Y_m)$ is a sample of maxima, the first immediate candidate that comes to our mind is the GEVd, defined in (2.9). Thereupon, in order to pursue a parametric approach, we can fit this limiting distribution to our data, just as an "exact" parametric model. Once again, we can use the QQ-plot to have a rough and quick confirmation of the plausible fit of the GEVd to our data. Since the case $\gamma = 0$, representing the Gumbel d.f., is seen as the separating line between the cases $\gamma < 0$, encompassing light right-tailed d.f's, and $\gamma > 0$, encompassing heavy right-tailed d.f.'s, we can ascertain the goodness-of-fit of the Gumbel distribution to our data through the QQ-plot. Therefore,

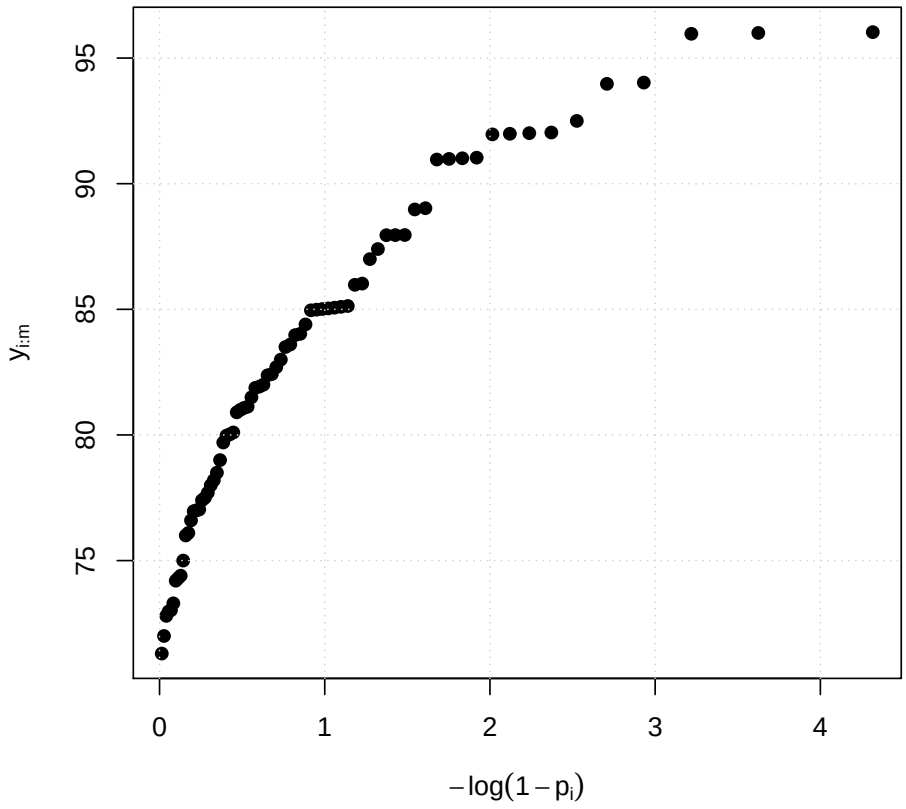

Figure 4.3: *Exponential QQ-plot for the $\dot{V}O_2max$ data*

for the Gumbel d.f., $F\left(\frac{y-\lambda}{\delta}\right) = \exp\left[-\exp\left(-\frac{y-\lambda}{\delta}\right)\right],\ y \in \mathbb{R}$, we have

$$\hat{Q}_{\lambda,\delta}(p_i) = y_{i:m},\ i = 1, \ldots, m$$

and

$$Q_{0,1}(p_i) = -\log(-\log p_i),\ i = 1, \ldots, m,$$

for $p_i = \frac{i}{m+1}$. The resulting plot can be seen in Figure 4.4,obtained with the help of the **R** software (see Appendix A.4).

This time, the plot exhibits a roughly linear pattern for the plotted points, despite of a visible concave pattern located at upper values of $\dot{V}O_2max$. The Gumbel distribution seems then to be a suitable candidate that fits our data. Remember that Figure 4.3 revealed a non-linear pattern when the Exponential model was proposed as a parametric model to be fitted to the data, pointing to an underlying d.f. F with a lighter right tail than the Exponential distribution. As the Gumbel distribution has a lighter right tail than the Exponential distribution, it may appear as a suitable candidate for Y. Since the

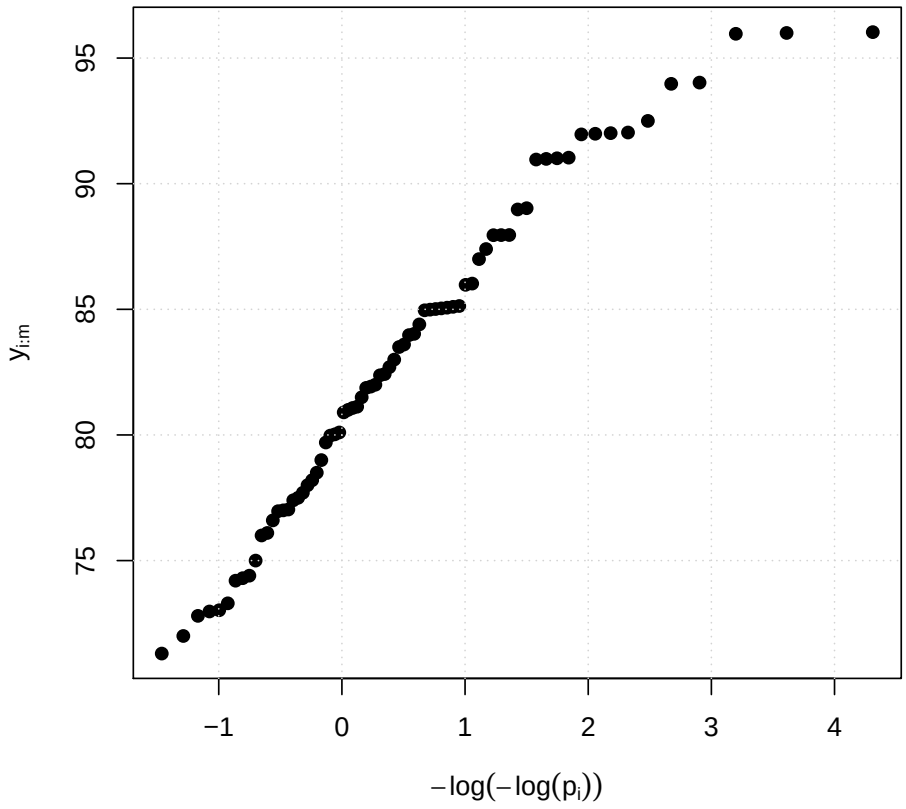

Figure 4.4: *Gumbel QQ-plot for the $\dot{V}O_2max$ data.*

QQ-plot in Figure 4.4 presents a more linear pattern than Figure 4.3 does, the Gumbel distribution is a more appropriate parametric family to be fitted to the r.v. Y.

As mentioned above, the QQ-plot reflects the linear relationship between the theoretical quantiles of a family of distributions and the corresponding theoretical quantiles of the standard member of the involved family. We can take advantage of the linear relationship to obtain preliminary estimates of the parameters of the theoretical family. Let then $Q_{\lambda,\delta}(p) = \Lambda^{-1}(p|\lambda,\delta)$ be the quantiles of order p of the Gumbel family, for $p \in]0,1[$. We have, in particular, $Q_{0,1}(p) = \Lambda^{-1}(p)$ as the quantiles of order p of the standard Gumbel distribution. From Theorem 2.6 and condition (2.6), we get, for $0 < p < 1$,

$$Q_{0,1}(p) = -\log(-\log p)$$

and

$$Q_{\lambda,\delta}(p) = \lambda - \delta \log(-\log p).$$

It follows that

$$Q_{\lambda,\delta}(p) = \lambda + \delta Q_{0,1}(p),$$

which makes clear the linear relationship between $Q_{\lambda,\delta}(p)$ and $Q_{0,1}(p)$. Thus, fitting a line to the points in Figure 4.4, we can use the intercept as a preliminary estimate for λ and the slope as a preliminary estimate for δ. Using the Ordinary Least Squares method of the **R** software (see Appendix A.5), we obtain

```
Call:
lm(formula = vo2max ~ Qg)

Coefficients:
(Intercept)           Qg
    80.0841       5.3099
```

which yields

$$\hat{\lambda} = 80.0841 \quad \text{and} \quad \hat{\delta} = 5.3099, \tag{4.1}$$

as preliminary estimates of the parameters of Gumbel distribution.

Using the **R** software (see Appendix A.5), we can add the fitted line to the QQ-plot of Figure 4.4, presented in Figure 4.5.

As stated before, all the d.f.'s belonging to Weibull max-domain have lighter right tails than the Exponential distribution. Hence, it may be reasonable to fit a GEVd with $\gamma < 0$ to the data to ascertain whether we obtain a better fit than that obtained with the Gumbel distribution. Once again, the goodness-of-fit can be assessed using the QQ-plot. Then, for the GEVd, we have, for $0 < p < 1$,

$$Q_{\gamma,\lambda,\delta}(p) = \lambda + \delta \, \frac{(-\log p)^{-\gamma} - 1}{\gamma} \tag{4.2}$$

and

$$Q_{\gamma,0,1}(p) = \frac{(-\log p)^{-\gamma} - 1}{\gamma}, \tag{4.3}$$

resulting in the linear relationship

$$Q_{\gamma,\lambda,\delta}(p) = \lambda + \delta \, Q_{\gamma,0,1}(p).$$

However, contrary to the Exponential and Gumbel cases, the quantiles of the GEV standard model depends on the shape parameter γ. Therefore, the QQ-plot for the GEVd can only be obtained after specifying a value for γ. Following Beirlant et al. (2004), we look for the value of γ in the neighbourhood of 0, which maximizes the coefficient of correlation between $\hat{Q}_{\gamma,\lambda,\delta}(p)$ and $Q_{\gamma,0,1}(p)$ on the QQ-plot.

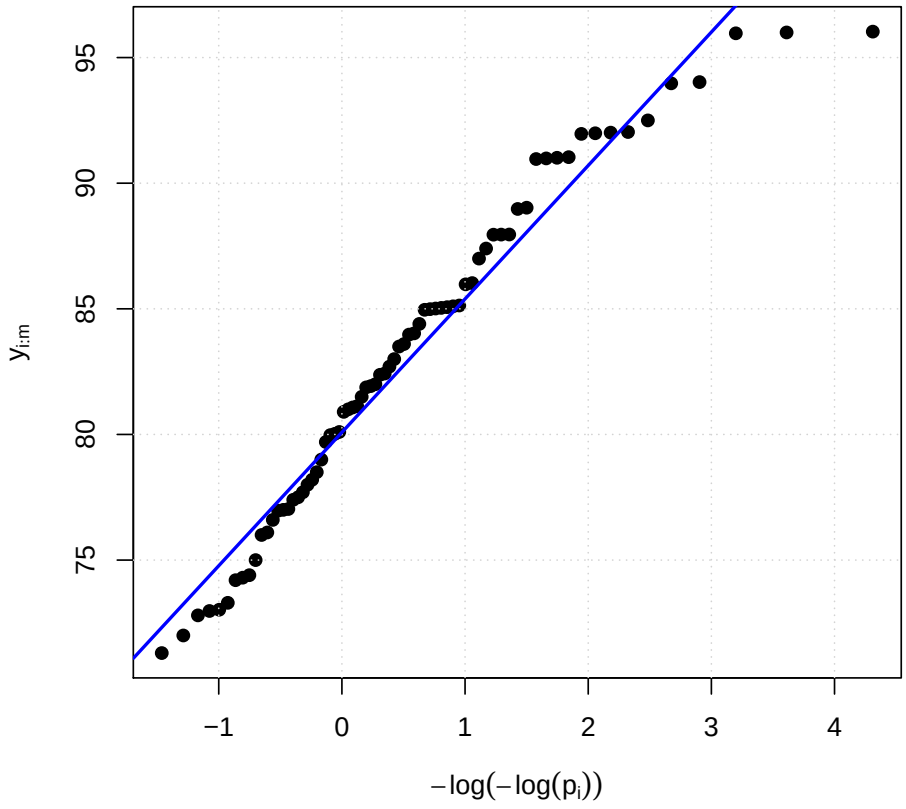

Figure 4.5: *Gumbel QQ-plot for the $\dot{V}O_2max$ data, with fitted straight line.*

This value can be obtained using the **R** software (see Appendix A.6):

```
$maximum
[1] -0.225207

$objective
[1] 0.9949305
```

which can be visualized graphically in Figure 4.6.

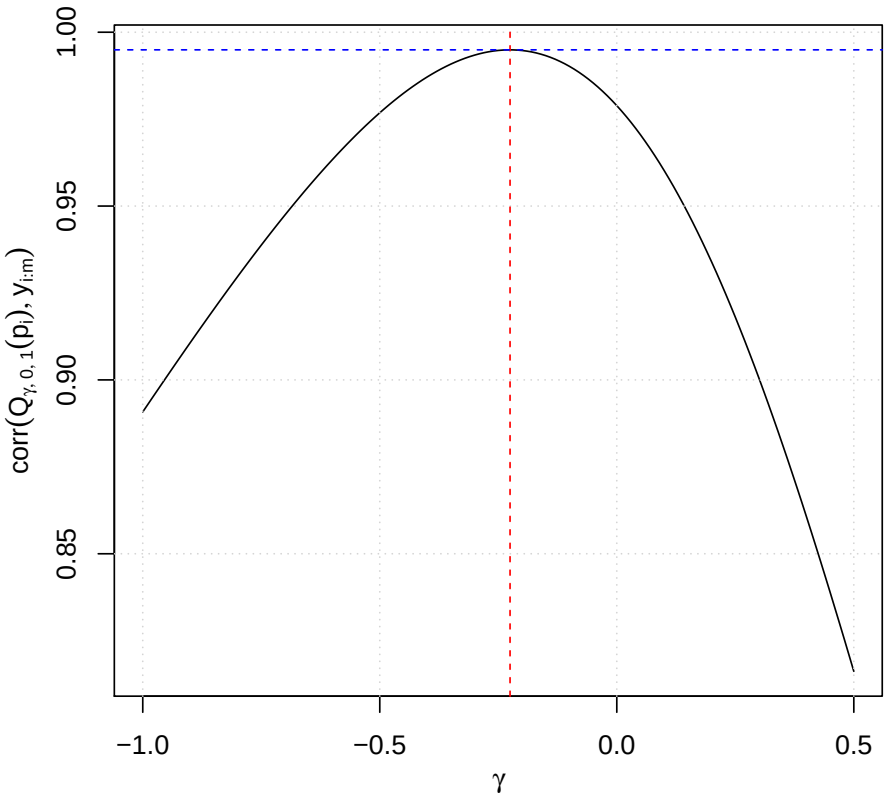

Figure 4.6: *Correlation plot between quantiles of the standard GEVd and of the location-scale GEV family for the $\dot{V}O_2max$ data.*

Using the **R** software (see Appendix A.7) with $\hat{\gamma} = -0.225207$, the corresponding QQ-plot is shown in Figure 4.7. Comparing the GEVd QQ-plot with the Gumbel QQ-plot of Figure 4.4, the former seems to reveal a better fit of the GEVd, which puts this model in a apparently better position to be chosen for the r.v. Y. As we did for the Gumbel QQ-plot, we can fit a straight line to the points of the plot, obtaining preliminary estimates for the

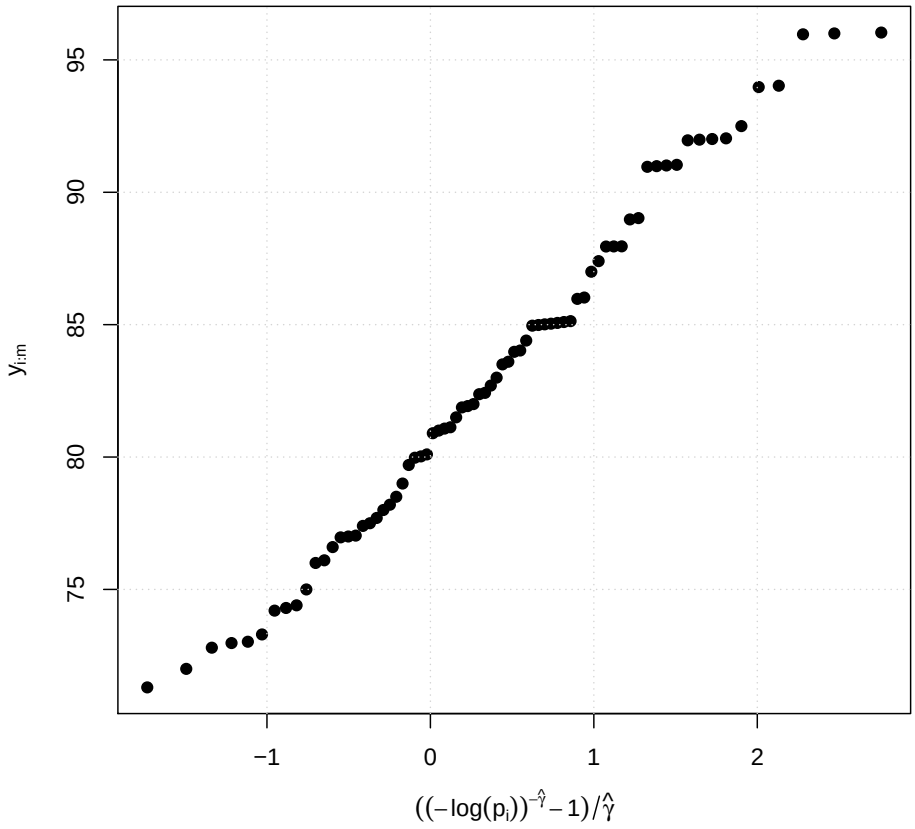

Figure 4.7: *GEVd QQ-plot for the $\dot{V}O_2max$ data.*

location and scale parameters of the GEVd. The results are obtained with the **R** software (see Appendix A.8) and the resulting fitted line can be seen in Figure 4.8.

```
Call:
lm(formula = vo2max ~ Qgev)

Coefficients:
(Intercept)         Qgev
     80.508        6.491
```

The preliminary estimates for the GEVd parameters are then

$$(\hat{\gamma}, \hat{\lambda}, \hat{\delta}) = (-0.225207, 80.508, 6.491). \tag{4.4}$$

We can notice a negative estimate for the scale parameter γ. The inference about γ will confirm whether we can assume $\gamma < 0$, pointing then to a d.f. for Y with a lighter right tail than the Exponential distribution.

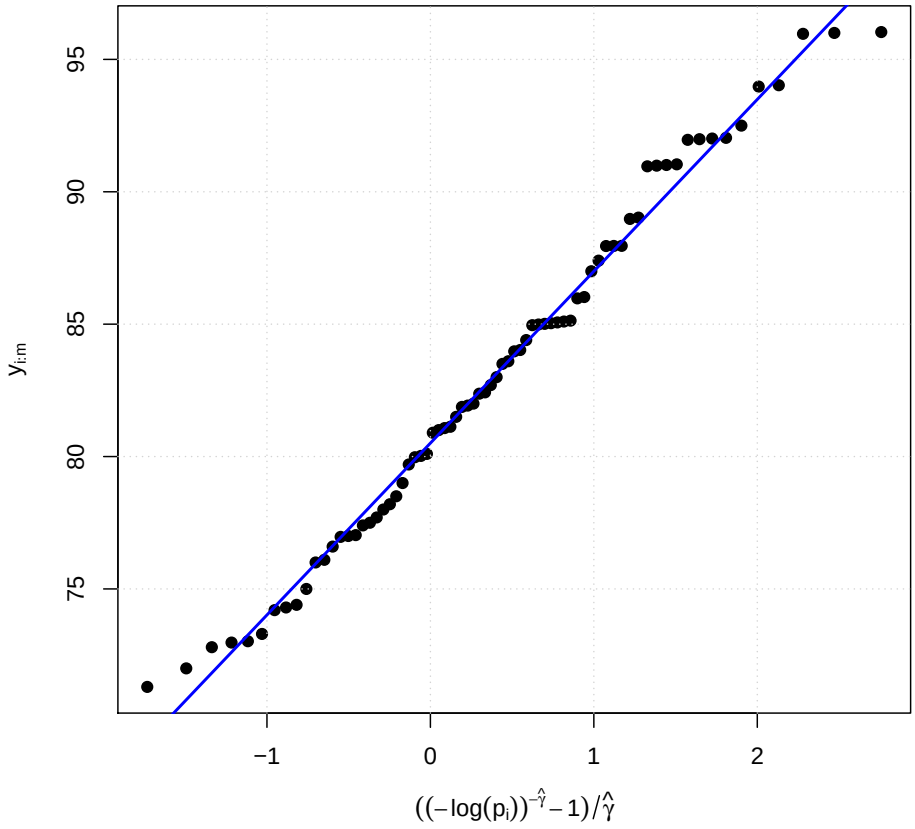

Figure 4.8: *GEVd QQ-plot for the $\dot{V}O_2max$ data, with fitted line.*

b) *Statistical choice of extreme value models*

From the preliminary analysis of last paragraph, we kept two parametric families as plausible candidates to be fitted the $\dot{V}O_2max$ data. To explore this question more deeply, we need more objectives tests. So, we can use some of the statistical tests mentioned in Section 3.1.2.5. As we are in a parametric approach, we are particularly interested in selecting a parametric model that best fits our $\dot{V}O_2max$ data. As seen, the model is chosen among the GEVd family, with a particular interest in Gumbel model, the transitional case between GEVd with $\gamma < 0$ and GEVd with $\gamma > 0$. Therefore, we can perform the following test, in order to check if the Gumbel model is suitable for our data or not:

$$H_0 : \gamma = 0 \quad vs \quad H_1 : \gamma \neq 0. \tag{4.5}$$

But if we want to have an alternative distribution for our data, in case of rejection of Gumbel model, we can perform an unilateral test,

$$H_0 : \gamma = 0 \quad vs \quad H_1 : \gamma < 0, \tag{4.6}$$

and if H_0 is rejected, we can choose a GEVd with $\gamma < 0$ as a parametric model for our data. Here, we are particularly interested in the case $\gamma < 0$, motivated by the conclusion drawn in the graphical preliminary analysis.

The first test presented can be found in Tiago de Oliveira and Gomes (1984) and allows us to test (4.6). The statistic to be used is given by

$$GS_m = \frac{Y_{m:m} - Y_{[m/2]+1:m}}{Y_{[m/2]+1:m} - Y_{1:m}},$$

which is called the *Gumbel statistic*.

Under the validity of H_0, we have the following result:

$$GS_m^* = \frac{GS_m - \beta_m}{\alpha_m} \quad \xrightarrow[n\to\infty]{d} \quad Z \frown \Lambda, \tag{4.7}$$

where the attraction coefficients β_m and α_m are given by

$$\beta_m = \frac{\log m + \log(\log 2)}{\log(\log m) - \log(\log 2)}$$

and

$$\alpha_m = \frac{1}{\log(\log m)}.$$

Therefore, at the asymptotic size α, for the test in (4.6), we reject H_0 if

$$GS_m^* \leq \mathcal{G}_\alpha,$$

where $\mathcal{G}_\epsilon$ stands for the standard Gumbel ϵ-quantile. For further details, we refer to Tiago de Oliveira and Gomes (1984).

We can also obtain the corresponding p-value as follows:

$$p(GS_m^*) = \Lambda(GS_m^*).$$

With the **R** software (see Appendix A.9), we obtain the following results:

```
[1] gs_m= 1.169591   gs*_m= -1.440154   p-value= 0.01467885
```

Considering the results, H_0 is then rejected at the asymptotic size $\alpha = 0.05$, leading us to a GEVd family with $\gamma < 0$ as a parametric model to be fitted to our data.

Another test that can be performed comes from Hosking (1984), where the author presents, among others, the Likelihood Ratio Test (LRT) and the Rao's score test, both for testing (4.5). Let $(Y_1, \ldots, Y_m)$ be a random sample of maxima, where $Y_i \frown G_\gamma$, for $i = 1, \ldots, m$, with G_γ defined in (2.9). Let $\ell(\gamma, \lambda, \delta | y_1, \ldots, y_m)$ be the respective unrestricted log-likelihood function, where $\ell(0, \lambda, \delta | y_1, \ldots, y_m)$ denotes the restricted log-likelihood function, which corresponds to the Gumbel case. The LRT statistic is given by

$$\mathbf{L} = -2\left(\ell(0, \hat{\lambda}_{G_0}, \hat{\delta}_{G_0} | Y_1, \ldots, Y_m) - \ell(\hat{\gamma}_{G_\gamma}, \hat{\lambda}_{G_\gamma}, \hat{\delta}_{G_\gamma} | Y_1, \ldots, Y_m)\right), \tag{4.8}$$

with $(\hat{\lambda}_{G_0}, \hat{\delta}_{G_0})$ and $(\hat{\gamma}_{G_\gamma}, \hat{\lambda}_{G_\gamma}, \hat{\delta}_{G_\gamma})$ denoting the ML estimators for G_0 and G_γ models, respectively.

Under H_0, we have

$$\mathbf{L} \xrightarrow[m\to\infty]{d} Z \frown \chi^2_{(1)}.$$

To achieve a higher accuracy in the χ^2-approximation, Hosking (1984) recommends the Bartlett correction, yielding the statistic

$$\mathbf{L}^* = \frac{\mathbf{L}}{1 + 2.8/m} \xrightarrow[m\to\infty]{d} Z \frown \chi^2_{(1)}. \tag{4.9}$$

For the test in (4.5), at the asymptotic size α, H_0 is rejected if

$$\mathbf{L}^* \geq \chi^2_{1,1-\alpha},$$

where $\chi^2_{1,\varepsilon}$ stands for the $\chi^2_{(1)}$ ε-quantile.

The corresponding p-value can be calculated as follows:

$$p(\mathbf{L}^*) = 1 - \chi^2_{(1)}(\mathbf{L}^*).$$

The **R** software (see Appendix A.10) allows us to obtain the parameters ML estimates required by (4.8):

```
[1] Gumbel ML estimates
 lambda= 79.87891   delta= 5.737552
[2] GEV ML estimates
 gamma= -0.2431824  lambda= 80.64481   delta= 6.166826
```

Consequently, we have

$$(\hat{\lambda}_{G_0}, \hat{\delta}_{G_0}) = (79.87891, 5.737552) \tag{4.10}$$

and

$$(\hat{\gamma}_{G_\gamma}, \hat{\lambda}_{G_\gamma}, \hat{\delta}_{G_\gamma}) = (-0.2431824, 80.64481, 6.166826). \tag{4.11}$$

Note the closeness of the final ML estimates for the Gumbel model in (4.10) to the initial estimation performed with the preliminary statistical analysis presented in (4.1). The same observation applies to the final ML estimates for the GEVd in (4.11), which preliminary estimates were obtained in (4.4).

Now, with the final ML estimates in hand, we can proceed to the test with the statistic given by (4.9). The **R** software produces the following results (see Appendix A.11):

```
[1] l= 4.272373   l*= 4.116609   p-value= 0.04246411
```

Once again, at the same asymptotic size $\alpha = 0.05$, we reject the null hypothesis of Gumbel model, but we note that the p-value is very close to the asymptotic size α.

Considering now Rao's score test, still from Hosking (1984), but previously presented by Tiago de Oliveira (1981), we continue with the same test in (4.5), as for the LRT. The score test checks whether the derivative of the log-likelihood function with respect to γ, at point $\gamma = 0$, is significantly different from zero. Significant nonzero values suggest $\gamma \neq 0$ and imply the rejection of H_0. Let then $g_\gamma(y|\lambda, \delta)$ denote the p.d.f. corresponding to GEVd, $G_\gamma(y|\lambda, \delta)$, defined in (2.9). The log-likelihood function for an observed random sample $(y_1, \ldots, y_m)$ is given by

$$\ell(\gamma, \lambda, \delta | y_1, \ldots, y_m) = \sum_{i=1}^{m} \log g_\gamma(y_i|\lambda, \delta),$$

from which we can obtain the ***score function*** with respect to γ:

$$V(\gamma|\lambda,\delta,y_1,\ldots,y_m)=\frac{\partial\ell(\gamma,\lambda,\delta|y_1,\ldots,y_m)}{\partial\gamma}=\sum_{i=1}^{m}\frac{\partial\log g_\gamma(y_i|\lambda,\delta)}{\partial\gamma}.$$

Let $(\hat{\lambda}_{G_0},\hat{\delta}_{G_0})$ be the parameters ML estimators for the G_0 model. In order to test (4.5), we define the score statistic as

$$V_m=\sum_{i=1}^{m}\lim_{\gamma\to 0}\frac{\partial\log g_\gamma(Y_i|\hat{\lambda}_{G_0},\hat{\delta}_{G_0})}{\partial\gamma}=\sum_{i=1}^{m}\left(\frac{1}{2}Z_i^2-Z_i-\frac{1}{2}Z_i^2\exp(-Z_i)\right),$$

where $Z_i=\frac{Y_i-\hat{\lambda}_{G_0}}{\hat{\delta}_{G_0}}$, for $=1,\ldots,m$.

Again under H_0, we have

$$V_m^*=\frac{V_m}{\sqrt{2.09797m}}\quad\xrightarrow[m\to\infty]{d}\quad Z_1\frown\mathcal{N}(0,1) \tag{4.12}$$

or, equivalently,

$$V_m^{*2}=\frac{V_m^2}{2.09797m}\quad\xrightarrow[m\to\infty]{d}\quad Z_2\frown\chi^2_{(1)}. \tag{4.13}$$

Details about the normal version (4.12) are presented in Tiago de Oliveira (1981) and the chi-square version (4.13) is presented by Hosking (1984). The normal version of the statistic allows us to perform the test in (4.6). Therefore, for the tests in (4.6) and (4.5), H_0 is rejected at the asymptotic size α if $V_m^*\leq z_\alpha$ or if $V_m^{*2}\geq\chi^2_{1,1-\alpha}$, respectively, where z_ε and $\chi^2_{1,\varepsilon}$ are the Normal and the $\chi^2_{(1)}$ ε-quantiles. The corresponding p-values are:

$$p(V_m^*)=\Phi(V_m^*)$$

and

$$p(V_m^{*2})=1-\chi^2_{(1)}(V_m^{*2}),$$

where $\Phi(.)$ represents the standard Normal d.f.

Using (4.10) for $(\hat{\lambda}_{G_0},\hat{\delta}_{G_0})$ and the **R** software (see Appendix A.12), we get the following results for both statistics:

```
[1] Normal Test:  v_m= -15.88231   v_m*= -1.274671   p-value= 0.1012128
```

```
[2] Chi-square Test:  v^2_m= 252.2478   v^2_m*= 1.624787
    p-value= 0.2024256
```

This time, the results of Rao's score test differ from those of the Gumbel statistic and of the LRT: at the asymptotic size of $\alpha = 0.05$, the null hypothesis of Gumbel model is not rejected, driving us away from the GEVd with $\gamma < 0$.

We can use a last test mentioned in Section 3.1.2.5 to get one more witness for the Gumbel model or for the GEVd with $\gamma < 0$. Details about this last test can be found in Marohn (2000), which is known as *Locally Asymptotically Normal* (LAN) *test*. As for the score test, we can use the LAN test for (4.5) or for (4.6).

Let $(\hat{\lambda}_{G_0}, \hat{\delta}_{G_0})$ be the parameters ML estimators for the G_0 model. The LAN statistic for testing the Gumbel hypothesis has the following expression:

$$T_m = \frac{1}{3.451}\left(\frac{1.6449}{\sqrt{m}}S_{1,m} - \hat{\delta}_{G_0}\frac{0.5066}{\sqrt{m}}S_{2,m} - \hat{\delta}_{G_0}\frac{0.8916}{\sqrt{m}}S_{3,m}\right),$$

where

$$\begin{aligned}
S_{1,m} &= \sum_{i=1}^{m}\left\{\frac{1}{2}\left(\frac{Y_i - \hat{\lambda}_{G_0}}{\hat{\delta}_{G_0}}\right)^2 - \frac{Y_i - \hat{\lambda}_{G_0}}{\hat{\delta}_{G_0}} - \frac{1}{2}\left(\frac{Y_i - \hat{\lambda}_{G_0}}{\hat{\delta}_{G_0}}\right)^2 \exp\left(-\frac{Y_i - \hat{\lambda}_{G_0}}{\hat{\delta}_{G_0}}\right)\right\},\\
S_{2,m} &= \sum_{i=1}^{m}\left\{-\frac{1}{\hat{\delta}_{G_0}} + \left(\frac{Y_i - \hat{\lambda}_{G_0}}{\hat{\delta}^2_{G_0}}\right)\left(1 - \exp\left(-\frac{Y_i - \hat{\lambda}_{G_0}}{\hat{\delta}_{G_0}}\right)\right)\right\},\\
S_{3,m} &= \sum_{i=1}^{m}\left\{\frac{1}{\hat{\delta}_{G_0}} - \frac{1}{\hat{\delta}_{G_0}}\exp\left(-\frac{Y_i - \hat{\lambda}_{G_0}}{\hat{\delta}_{G_0}}\right)\right\}.
\end{aligned}$$

Note that $S_{1,m}, S_{2,m}$ and $S_{3,m}$ are the components of the *score function*, i.e., the first derivatives of the log-likelihood function from the GEVd defined in (2.9), with respect to each parameter, γ, λ and δ, under H_0 or, equivalently, at point $\gamma = 0$:

$$\begin{aligned}
S_{1,m} &= \sum_{i=1}^{m}\lim_{\gamma\to 0}\frac{\partial \log g_\gamma(Y_i|\lambda,\delta)}{\partial\gamma},\\
S_{2,m} &= \sum_{i=1}^{m}\frac{\partial \log g_0(Y_i|\lambda,\delta)}{\partial\delta},\\
S_{3,m} &= \sum_{i=1}^{m}\frac{\partial \log g_0(Y_i|\lambda,\delta)}{\partial\lambda},
\end{aligned}$$

replacing λ and δ with their ML estimators, $\hat{\lambda}_{G_0}$ and $\hat{\delta}_{G_0}$, respectively.

According to Marohn (2000), under the validity of H_0, we have

$$T^*_m = \frac{T_m}{0.6904} \quad \xrightarrow[m\to\infty]{d} \quad Z \frown \mathcal{N}(0,1) \tag{4.14}$$

and, at the asymptotic size of α, H_0 is rejected if $|T^*_m| \geq z_{1-\alpha/2}$ or if $T^*_m \leq z_\alpha$, for the tests in (4.5) or in (4.6), respectively. The corresponding p-values are obtained with

$$p(T^*_m) = 2 - 2\Phi(|T^*_m|)$$

or

$$p(T_m^*) = \Phi(T_m^*).$$

For testing (4.6), the following results are produced using the **R** software (see Appendix A.13) and taking (4.10) for $\hat{\lambda}_{G_0}$ and $\hat{\delta}_{G_0}$:

```
[1] t_m= -0.8798613   t_m*= -1.274423   p-value= 0.1012569
```

As for Rao's score test, we do not reject the null hypothesis of Gumbel model, at the asymptotic level $\alpha = 0.05$.

To complement the previous tests, we can turn to the goodness-of-fit tests for the Gumbel model, invoked in Section 3.1.2.5, equivalent to the hypotheses test defined in (4.5).

Let $(Y_1, \ldots, Y_m)$ be a random sample of maxima, where $Y_i \frown G_0$, for $i = 1, \ldots, m$, with G_0 defined in (2.9). The test in (4.5) can be checked equivalently by a goodness-of-fit test for the Gumbel model G_0, with the following statistics:

1. Kolmogorov-Smirnov

$$D_m = \max_{1 \leq i \leq m} \left\{ \left| G_0(Y_{i:m}|\hat{\lambda}_{G_0}, \hat{\delta}_{G_0}) - \frac{i}{m} \right|, \left| G_0(Y_{i:m}|\hat{\lambda}_{G_0}, \hat{\delta}_{G_0}) - \frac{i-1}{m} \right| \right\}, \quad (4.15)$$

2. Cramér-von Mises

$$W_m^2 = \sum_{i=1}^{m} \left(G_0(Y_{i:m}|\hat{\lambda}_{G_0}, \hat{\delta}_{G_0}) - \frac{2i-1}{2m} \right)^2 + \frac{1}{12m}, \quad (4.16)$$

3. Anderson-Darling

$$A_m^2 = -m - \frac{1}{m} \sum_{i=1}^{m} \Big\{ (2i-1) \log(G_0(Y_{i:m}|\hat{\lambda}_{G_0}, \hat{\delta}_{G_0})) + \\ + (2m+1-2i) \log(1 - G_0(Y_{i:m}|\hat{\lambda}_{G_0}, \hat{\delta}_{G_0})) \Big\}, \quad (4.17)$$

where $\hat{\lambda}_{G_0}, \hat{\delta}_{G_0}$ represent the ML estimators for the Gumbel model, G_0.

With $\hat{\lambda}_{G_0}, \hat{\delta}_{G_0}$ given in (4.10) and turning to **R** software (see Appendix A.14), we get

```
Kolmogorov-Smirnov statistic:  0.06815551
Cramer-von Mises statistic:  0.07324065
Anderson-Darling statistic:  0.540129
```

To complete the test, the appropriate test statistic is compared with simulated upper quantiles of the statistic's sampling distribution, called *upper tail percentage points*. Tables of simulated quantiles can be found in Chandra et al. (1981) for Kolmogorov-Smirnov statistic and in Stephens (1977) for Cramér-von Mises and Anderson-Darling statistics. H_0 is then rejected if the observed statistic value exceeds the appropriate quantile at the asymptotic level α.

For Kolmogorov-Smirnov statistic, part of the table found in Chandra et al. (1981) is transcribed in Table 4.1.

Table 4.1: *Upper tail percentage points for Kolmogorov-Smirnov statistic, modified for the Gumbel distribution.*

		Upper tail significance level α			
Statistic	m	.10	.05	.025	.01
$\sqrt{m}\ D_m$	10	.760	.819	.880	.944
	20	.779	.843	.907	.973
	50	.790	.856	.922	.988
	∞	.803	.874	.939	1.007

Chandra et al. (1981)

The observed statistic is then

$$\sqrt{m}\ d_m = \sqrt{74} \times 0.06815551 \simeq 0.586$$

and for $m = 74$ and $\alpha = 0.05$, the upper quantile is not exceeded and, consequently, we do not reject H_0.

For Cramér-von Mises and Anderson-Darling statistics, we transcribe part of the table from Stephens (1977) in Table 4.2.

Table 4.2: *Upper tail percentage points for Cramér-von Mises and Anderson-Darling statistics, modified for the Gumbel distribution.*

		Upper tail percentage points, α				
Statistic	Modification	.75	.90	.95	.975	.99
W_m^2	$W_m^2(1 + 0.2/\sqrt{m})$	.073	.102	.124	.146	.175
A_m^2	$A_m^2(1 + 0.2/\sqrt{m})$	.474	.637	.757	.877	1.038

Stephens (1977)

The observed modified statistics are then

$$w_m^2(1 + 0.2/\sqrt{m}) = 0.07324065 \times (1 + 0.2/\sqrt{74}) \simeq 0.075$$

and

$$a_m^2(1+0.2/\sqrt{m}) = 0.540129 \times (1+0.2/\sqrt{74}) \simeq 0.553.$$

We notice that neither of the modified statistics exceed the upper quantile at size $\alpha = 0.05$ and, hereupon, we do not reject H_0.

The three goodness-of fit tests lead us to the non-rejection of H_0, favouring once again the Gumbel model. However, these tests must be done with some care, because they tend to be conservative. To summarize all the results from the previous tests, we can present the Table 4.3 with the decision of each test:

Table 4.3: *Results from the statistical choice of extreme value models, for the $\dot{V}O_2max$ data.*

Test	Hypotheses	Observed statistic	p-value	Decision ($\alpha = 0.05$)
Gumbel statistic	$H_0 : \gamma = 0$ vs $H_1 : \gamma < 0$	$gs_m^* = -1.440154$	0.01467885	reject H_0
LRT	$H_0 : \gamma = 0$ vs $H_1 : \gamma \neq 0$	$\mathbf{l}^* = 4.116609$	0.04246411	reject H_0
Rao's score test	$H_0 : \gamma = 0$ vs $H_1 : \gamma \neq 0$	$v_m^{*2} = 1.624787$	0.2024256	not reject H_0
LAN test	$H_0 : \gamma = 0$ vs $H_1 : \gamma < 0$	$t_m^* = -1.274423$	0.1012569	not reject H_0

Concerning the three goodness-of-fit tests in (4.15), (4.16) and (4.17), we concluded that the Gumbel model G_0 was not rejected.

Most of the tests elect the Gumbel model as a parametric model to be fitted to the $\dot{V}O_2max$ data. Even the LRT has a suspicious p-value near the boundary $\alpha = 0.05$. For $\alpha = 0.01$, the statistic is just not significant, leading us to non-rejection of Gumbel family. Therefore, the underlying d.f. F of the $\dot{V}O_2max$ can have a finite or an infinite right endpoint. However, $\dot{V}O_2max$ is pre-eminently a physiological variable and, consequently, is naturally limited by many physiological factors. In this matter, there is no consensus among physiologists, since each of them claim the relative importance of different factors. But all of them agree about one point: $\dot{V}O_2max$ does not have the capacity to grow infinitely, provided that it relies on limited physiological factors. One of the most important factors is undoubtedly the heart rate. According to Cerretelli and Di Prampero (1987), 70-85% of the limitation in $\dot{V}O_2max$ can be attributed to heart rate. Other appointed limiting factors are the pulmonary diffusion, i.e. the exchange of oxygen and carbon dioxide between the lungs and the blood, and the oxygen carrying capacity of the blood (blood volume and flow). Details on how these factors can limit $\dot{V}O_2max$ are very abundant in physiological literature. We can cite Bassett and Howley (2000), Kravitz and Dalleck (2002), Warpeha (2003) and McArdle et al. (2009). Despite of these recent contributions, the fact that $\dot{V}O_2max$ has an upper bound limit has already been established in the 1920s with the works of Hill and Lupton (1923) and Hill et al. (1924)

and today, it is universally accepted that there is a physiologically upper limit to the body's ability to consume oxygen.

c) *Parametric estimation of extreme events*

From last paragraph, we conclude that, even if we do not reject the Gumbel model, Physiology tells us that it does make any sense to establish an infinite right endpoint for the underlying distribution F of $\dot{V}O_2max$. The Gumbel model has precisely this type of flexibility, since its max-domain of attraction embraces d.f.'s with finite or infinite right endpoints. Thereupon, we proceed with the parameters estimation of the Gumbel model, the location parameter (λ) and the scale parameter (δ), turning to the estimation methods presented in Sections 3.1.2.1 and 3.1.2.2: the ML method and the PWM method. The ML method was already applied since it was needed to perform the hypotheses tests in (4.5) with (4.9). The results were presented in (4.10).

The adequacy of the Gumbel fit via ML method can be assessed by graphical diagnosis tools of the **R** software (see Appendix A.15). They are depicted in Figure 4.9.

The Gumbel fit is globally satisfactory, despite the asymmetry present at the beginning and at the end of the sample, which is very common in the latter case.

We can now apply the PWM method to obtain the estimates for the same parameters using the **R** software (see Appendix A.16):

```
[1] Gumbel PWM estimates
 lambda= 79.91804   delta= 5.39941
```

We can gather the results of both methods in Table 4.4:

Table 4.4: *ML and PWM estimates of the location and scale parameters of the Gumbel model, for the $\dot{V}O_2max$ data.*

Estimation method	$\hat{\lambda}$ (location)	$\hat{\delta}$ (scale)
ML	79.87891	5.737552
PWM	79.91804	5.39941

As we can see, the estimates are very similar, using both ML and PWM methods. The values can be taken as estimates for the attraction coefficients of Section 2.4, with $\hat{b}_n = \hat{\lambda}$ and $\hat{a}_n = \hat{\delta}$.

With the estimates in hand, we can turn to the construction of CI's for the location and scale parameters of the Gumbel model. As mentioned in Section 3.1.2.4, we can obtain

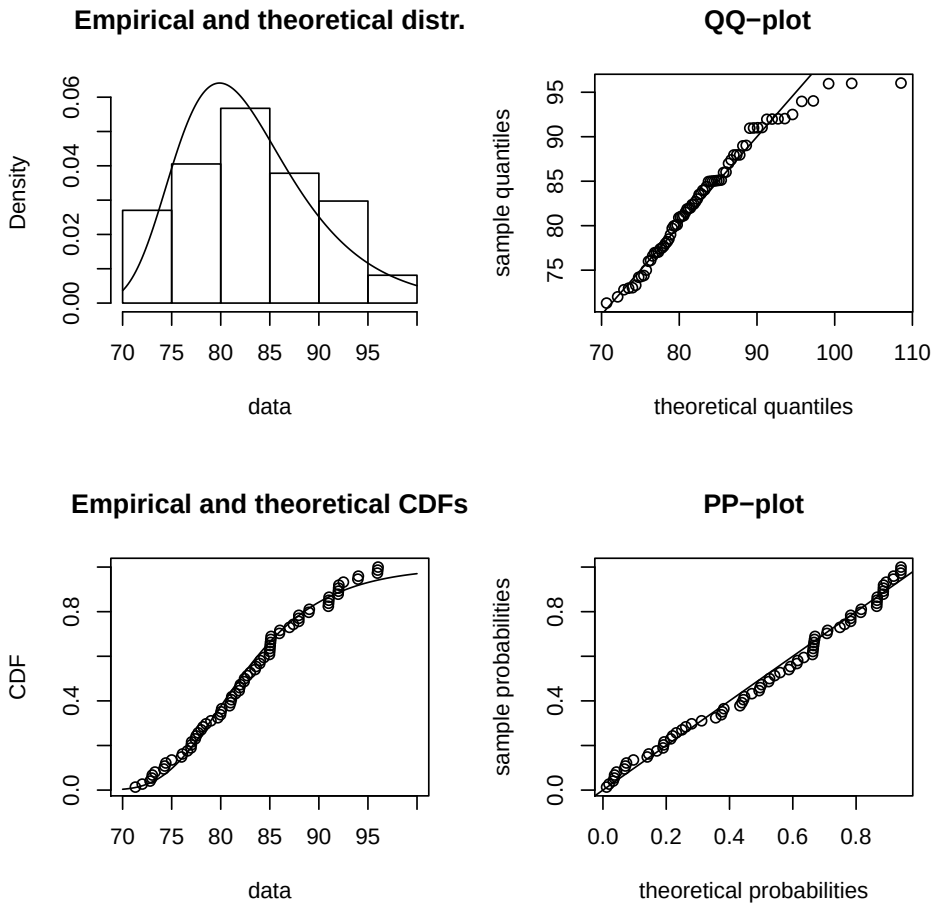

Figure 4.9: *Graphical diagnosis of the Gumbel fit for the $\dot{V}O_2max$ data.*

more accurate intervals if they are based on the profile likelihood function. We obtain the following results with our usual **R** software (see Appendix A.17), at the asymptotic 95% confidence level:

```
[1] "profiling loc"
[1] "profiling scale"
          lower     upper
loc   78.504117 81.304361
scale  4.851143  6.890512
```

The profile likelihood-based CI's can also be plotted, yielding Figure 4.10 (see Appendix A.17 for **R** details).

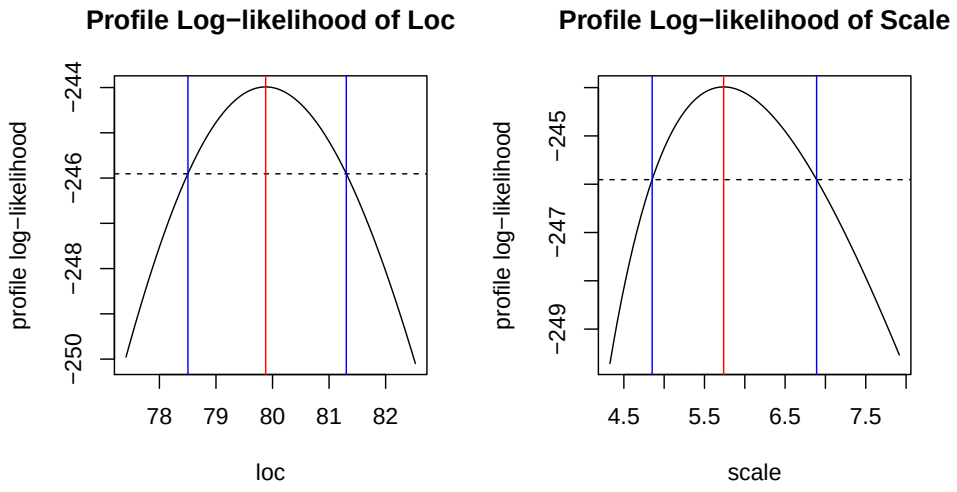

Figure 4.10: *Profile likelihood-based 95% confidence intervals for the Gumbel model parameters, for the $\dot{V}O_2max$ data.*

As we are exploring the Gumbel model, the last estimate we can introduce is the estimate of the exceedance probability, defined in (3.14). As we are dealing with non-temporal data, it does make any sense to estimate the return period in (3.14). Concerning extreme quantiles, obtained in expressions (3.12) and (3.13), we have no specific choice for estimating a particular extreme quantile. Finally, as we are in a Gumbel model, we cannot use expression (3.17) for estimating the right endpoint of the underlying d.f. F, since it is only valid for $\gamma < 0$.

Turning back to the exceedance probability, in our context of $\dot{V}O_2max$, we can be interested by one relevant estimate: the possibility of any athlete of our population surpassing the current record of 96 ml/kg/min, i.e. $P(Y > 96)$. We can consider this

probability as an excellence measure of the current record. Defining X as the r.v. that represents the $\dot{V}O_2max$ of an athlete of our homogeneous population of athletes, delineated at the beginning of this Chapter, and Y as the r.v. in (3.1), which represents the *maximal* $\dot{V}O_2max2$ of such an athlete, we can now stipulate that $Y \frown \Lambda(\hat{\lambda}, \hat{\delta})$, where $(\hat{\lambda}, \hat{\delta})$ are the ML or PWM estimators of (λ, δ). We obtain the following results with the **R** software (see Appendix A.18):

```
[1] Maximum Likelihood: P(Y>96)= 0.05844263
[2] Probability Weighted Moments: P(Y>96)= 0.04959853
```

In the present conditions, any top athlete has then a 5-6% probability of surpassing the current $\dot{V}O_2max$ record.

Before leaving the Block Maxima method, we will also consider the parametric GEVd, with $\gamma < 0$. Indeed, when LRT was performed, we needed to estimate the parameters of the GEVd model and, as we noted in (4.11), we have $\hat{\gamma} < 0$. Additional reasons can be appointed for this decision. First of all, as mentioned above, it is universally accepted that $\dot{V}O_2max$ is a physiological factor with a finite upper bound. Consequently, an estimate of the right endpoint x^F is appropriate. But this estimate is only calculable when we have $\gamma < 0$. Second, the concave pattern of the QQ-plot in Figure 4.3 suggests us a light right tail of the underlying d.f. F and the same conclusion can be drawn from the sample ME-plot in Figure 4.1. Indeed, according to Beirlant et al. (2004), a decreasing trajectory for the sample ME-plot corresponds to d.f.'s with a light right tail. Finally, some of the tests performed rejected the null hypothesis $H_0 : \gamma = 0$. Even if the LRT has a p-value close to the usual size $\alpha = 0.05$, the null hypothesis remains rejected. Parametric tests for this matter are abundant in the literature and we only tried some of them. We cannot discard the possibility of other tests that haven't been used here to lead us to the rejection of H_0. As we are in a modest sample, a comparison of the performed tests in terms of their power would be desirable. But this is out of the scope of this thesis.

The results for the GEVd are presented below, applying the **R** software (see Appendix A.19) to extract the same type of estimates obtained in the Gumbel model. For the parameters estimates, we get:

```
[1] GEV ML estimates
  gamma= -0.2431824  lambda= 80.64481   delta= 6.166826

[2] GEV PWM estimates
  gamma= -0.2023684  lambda= 80.46483   delta= 6.307003
```

We observe that the respective estimates for γ are relatively similar and since we have $\hat{\gamma} > -1$, we can cross our fingers and expect that we also have $\gamma > -1$, to grant the consistency and asymptotic Normality of the estimators. The adequacy of the GEVd fit via ML method can be assessed by graphical diagnosis tools of the **R** software (see Appendix A.20). They are depicted in Figure 4.11.

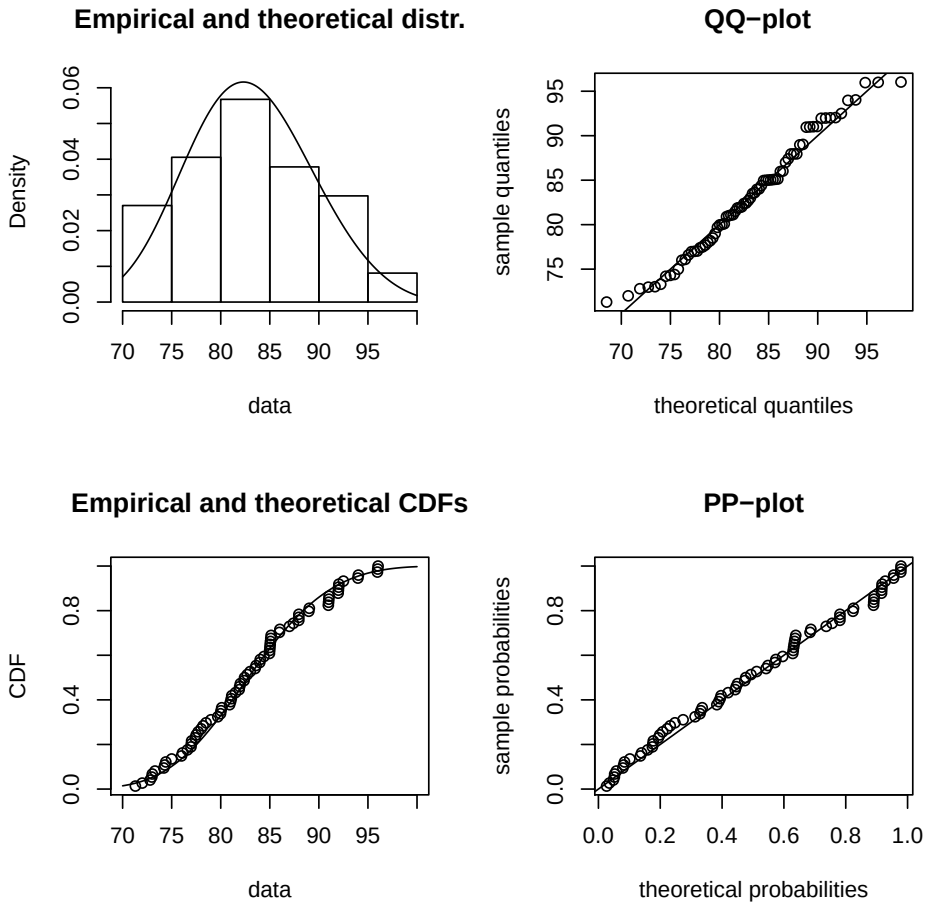

Figure 4.11: *Graphical diagnosis of the GEVd fit for the $\dot{V}O_2max$ data.*

The GEVd fit is globally satisfactory, despite the asymmetry present at the beginning and at the end of the sample, which is very common in the latter case. Comparing the diagnosis for the GEVd with the same diagnosis for the Gumbel model in Figure 4.9, we can argue a better fit for the GEVd than for the Gumbel model and choose the first as a suitable parametric model.

We can obtain CI's using the profile likelihood function, as we did for the Gumbel family, plotting the results in Figure 4.12 (see Appendix A.21 for **R** details).

```
[1] "profiling loc"
[1] "profiling scale"
[1] "profiling shape"
           lower       upper
loc   79.0378578 82.29713706
scale  5.1364086  7.63658965
shape -0.4388217 -0.01412509
```

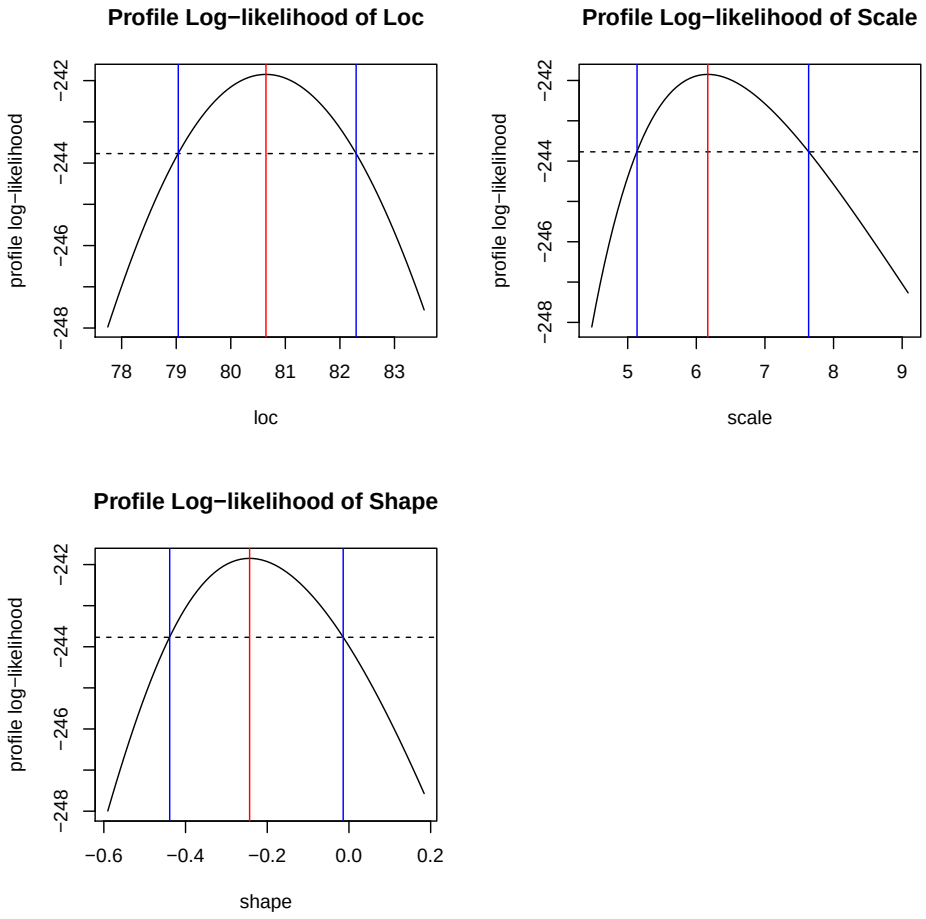

Figure 4.12: *Profile likelihood-based 95% confidence intervals for the GEVd parameters, for the $\dot{V}O_2max$ data.*

The CI for the EVI alerts us to the proximity of $\gamma = 0$, since the upper limit of the interval is very close to zero. It can be one of the reasons for the discordances of the statistical tests performed above.

Finally, we can estimate the exceedance probability for the current record of 96 ml/kg/min, as we did for the Gumbel family (see Appendix A.22 for **R** details):

```
[1] Maximum Likelihood: P(Y>96)= 0.02158176
[2] Probability Weighted Moments: P(Y>96)= 0.03250008
```

Compared with the case $\gamma = 0$, we see that the estimated probabilities are lower, with a reduction from 5-6% to 2-3%.

And now, as we are in the world of $\gamma < 0$, we can obtain an estimate for the right endpoint of the underlying d.f. of the $\dot{V}O_2max$. Turning to expression (3.17), the estimate for the right endpoint x^F can be calculated either by ML method or by PWM method (see Appendix A.23 for **R** details):

```
[1] Maximum Likelihood: x^F= 106.0037
[2] Probability Weighted Moments: x^F= 111.6308
```

The ML method establishes the level 106.0037 ml/kg/min as an estimate for the $\dot{V}O_2max$ upper bound, while the PWM method pushes this limit further, presenting the estimate 111.6308 ml/kg/ml as finite right endpoint. These values are plausible ones, in the light of what was exposed above, in physiological terms. Notice that the current record holder is at a distance of 10 ml/kg/min of the highest possible value, in terms of ML estimation. All the results for Gumbel and GEV models are summarized in Table 4.5:

Table 4.5: *Estimation results for Gumbel and GEVd models, for the $\dot{V}O_2max$ data.*

		Gumbel model $\gamma = 0$	GEV model $\gamma < 0$
$\hat{\gamma}$	ML	-	-0.2431824
	PWM	-	-0.2023684
	profile CI (95%)	-	(-0.4388217,-0.01412509)
$\hat{\lambda}$	ML	79.87891	80.64481
	PWM	79.91804	80.46483
	profile CI (95%)	(78.504117,81.304361)	(79.0378578,82.29713706)
$\hat{\delta}$	ML	5.737552	6.166826
	PWM	5.39941	6.307003
	profile CI (95%)	(4.851143,6.890512)	(5.1364086,7.63658965)
$P(Y > 96)$	ML	0.05844263	0.02158176
	PWM	0.04959853	0.03250008
x^F	ML	-	106.0037
	PWM	-	111.6308

2) The POT method

a) *Preliminary statistical analysis*

As a preview of the POT approach, we can perform a quick preliminary analysis, ascertaining the goodness-of-fit of the Exponential model to the excesses above the chosen threshold. The Exponential distribution is particularly important in this approach, since we know from (3.21) that the GPd reduces to the Exponential d.f. when $\gamma = 0$. Therefore, trying to fit an Exponential model to the excesses Y above the chosen threshold reduces to ascertain the goodness-of-fit of the GPd to these excesses, for $\gamma = 0$. Recall that we chose the threshold $u = 80$ based on the sample ME-plot of Figure 4.1. We can then use the QQ-plot of Figure 4.13 to check the adequacy of the Exponential model to the excesses above $u = 80$ (see Appendix A.24 for **R** details).

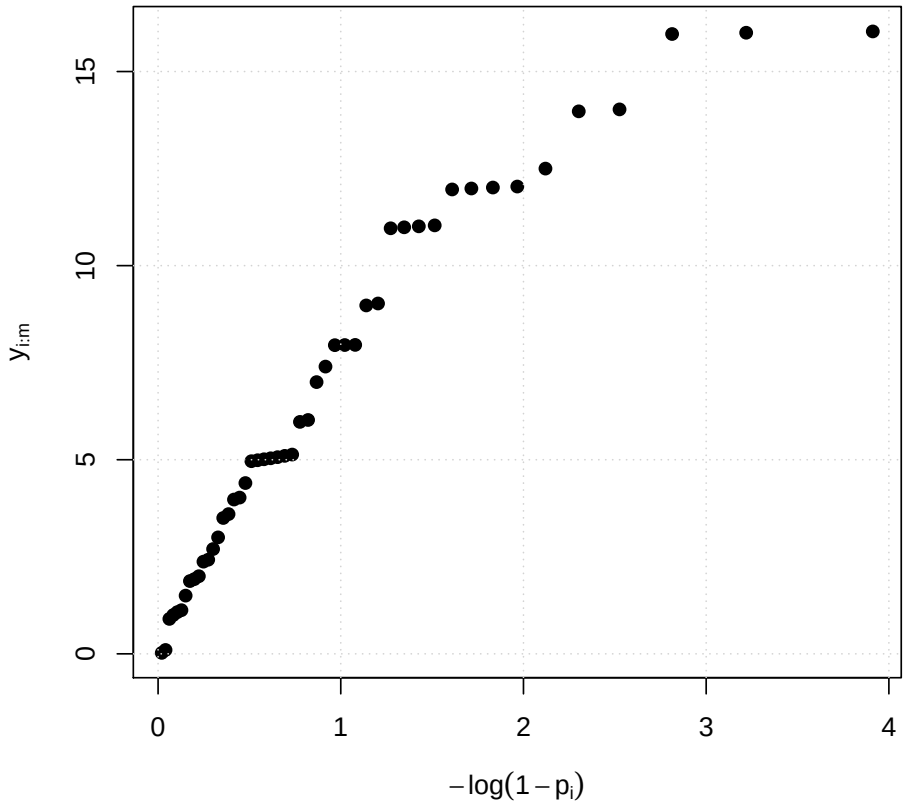

Figure 4.13: *Exponential QQ-plot for the $m = 49$ excesses of the $\dot{V}O_2max$ data*

As in the Block Maxima method, the sample path of the Exponential QQ-plot is characterized by a concave pattern and does not follow a linear trend. We saw the same

pattern in Figure 4.3, for high values of $\dot{V}O_2max$. Thus, we can conclude that the Exponential d.f. is not a suitable parametric model to be fitted to the excesses Y, since the underlying d.f. has a lighter right tail, i.e. a Beta-type right tail. The GPd with $\gamma < 0$, defined in (3.24), embodies such d.f.'s and its goodness-of-fit can be assessed again with the help of the QQ-plot. Based on (3.24), we can obtain the theoretical quantiles of order p for the GP family, for $\gamma \neq 0$:

$$Q_{\gamma,\sigma_u}(p) = H_\gamma^{-1}(p|0,\sigma_u) = \sigma_u \frac{(1-p)^{-\gamma}-1}{\gamma}, \quad 0 < p < 1. \tag{4.18}$$

The quantiles of the standard GPd are given then by

$$Q_{\gamma,1}(p) = \frac{(1-p)^{-\gamma}-1}{\gamma}, \quad 0 < p < 1, \tag{4.19}$$

yielding the linear relationship

$$Q_{\gamma,\sigma_u}(p) = \sigma_u\, Q_{\gamma,1}(p).$$

Plotting then $Q_{\gamma,\sigma_u}(p)$ against $Q_{\gamma,1}(p)$, we obtain a straight line with no intercept and slope σ_u. As for the GEVd in the Block Maxima approach, we have to specify a value for γ in order to construct the QQ-plot. Following again the technique described in Beirlant et al. (2004), we choose the value of γ that maximizes the correlation coefficient between $\hat{Q}_{\gamma,\sigma_u}(p)$ and $Q_{\gamma,1}(p)$ in the QQ-plot. With the help of **R** software (see Appendix A.25), we obtain

```
$maximum
[1] -0.5532608

$objective
[1] 0.994463
```

which can be visualized graphically in Figure 4.14 (see Appendix A.25 for **R** details). With $\hat{\gamma} = -0.5532608$, the corresponding QQ-plot is shown in Figure 4.15 (see Appendix A.26 por **R** details).

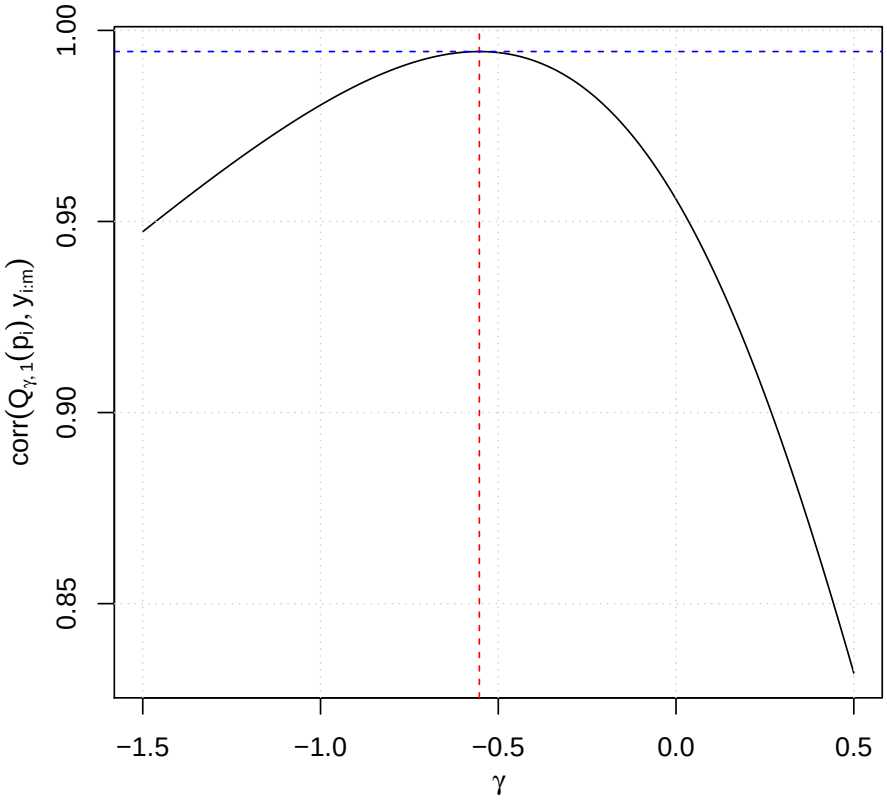

Figure 4.14: *Correlation plot between quantiles of the standard GPd and GP model for the* $m = 49$ *excesses of the* $\dot{V}O_2max$ *data.*

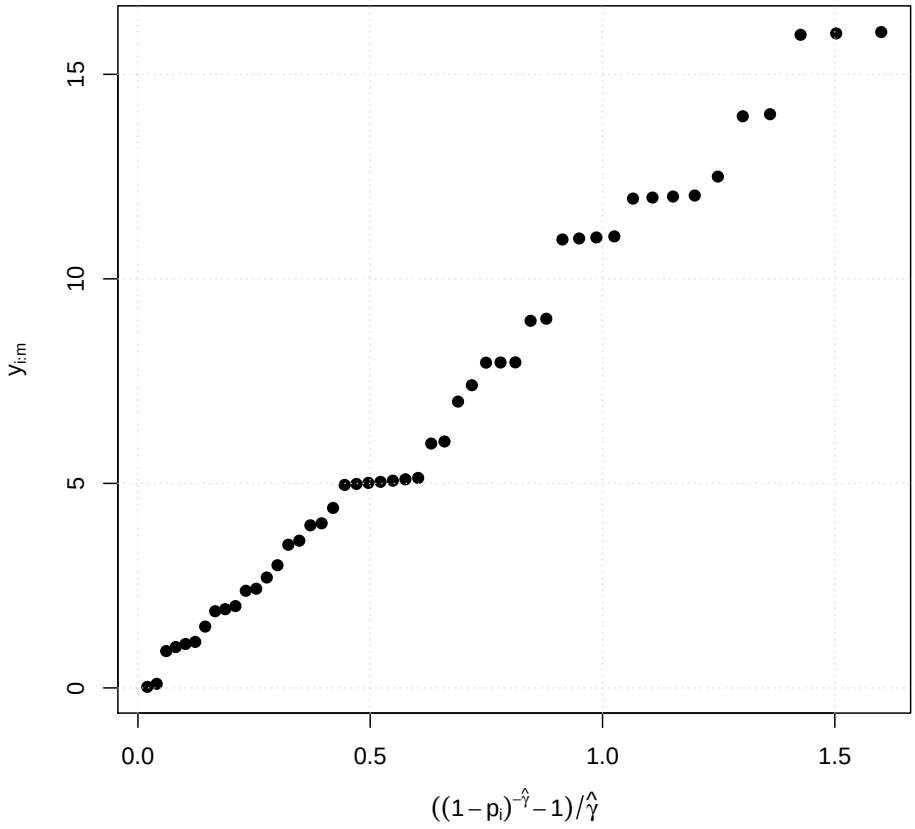

Figure 4.15: *GPd QQ-plot for the* $m = 49$ *excesses of the* $\dot{V}O_2max$ *data.*

The QQ-plot for the GPd exhibits a more satisfactory linearity than the QQ-plot for the Exponential case of Figure 4.13 does, in spite of some irregularity. The GPd with $\gamma \neq 0$ seems then to provide a better fit for the r.v. Y. Now, based on the QQ-plot of Figure 4.15, we can fit a least squares straight line to the plotted points, which slope gives us a preliminary estimate of the scale parameter σ_u. As stated before, the straight line relating the GPd quantiles has no intercept. Therefore, least squares straight line must be fitted without intercept. The QQ-plot with the fitted line in presented in Figure 4.16, with the help of the **R** software (see Appendix A.27).

```
Call:
lm(formula = excess ~ Qgpd - 1)

Coefficients:
 Qgpd
10.48
```

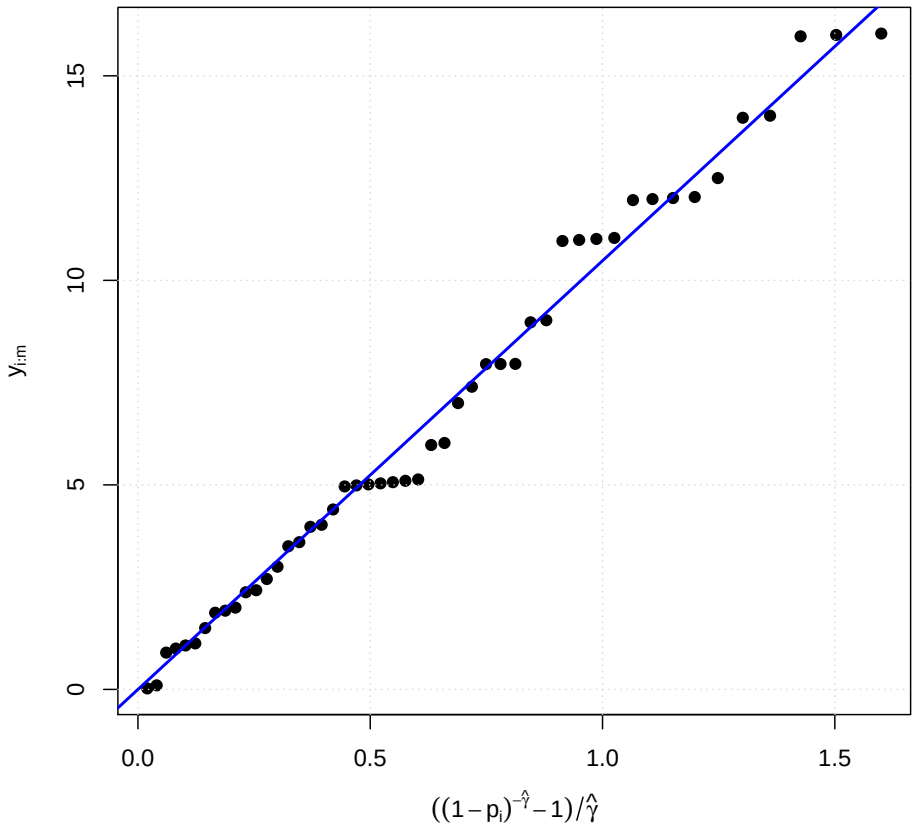

Figure 4.16: *GPd QQ-plot for the $m = 49$ excesses of the $\dot{V}O_2max$ data,with fitted line.*

The preliminary estimates for the GPd parameters are then

$$(\hat{\gamma}, \hat{\sigma_u}) = (-0.5532608, 10.48). \tag{4.20}$$

As for the Gumbel's approach, note the negative estimate obtained for γ, which points us to a GPd with $\gamma < 0$, when confirmed with some inference about γ.

b) *Statistical choice of GPd models*

The preliminary statistical analysis of the POT approach selected the GPd with $\gamma < 0$ as a suitable candidate for the r.v. Y, which represents here the excesses above the chosen threshold, to the detriment of the GPd with $\gamma = 0$, i.e. the Exponential distribution. However, since the preliminary analysis is not exempted from some subjectivity, we need to rely on objective statistical tests, in order to take a decision. Such tests were mentioned in Section 3.1.3.6 and will be applied to the $\dot{V}O_2max$ data. Let then X be the r.v. that represents the $\dot{V}O_2max$ of an athlete of our defined population. Given our fixed threshold $u = 80$, we denote by $Y = X - u$ the r.v. that represents the excess above that fixed threshold u, as defined in Section 3.1.3.1. In this parametric approach, it is assumed that Y is GP-distributed, i.e. $Y \frown H_\gamma$, with H_γ defined in (3.24). We are particulary interested in the following tests:

$$H_0 : \gamma = 0 \quad vs \quad H_1 : \gamma \neq 0, \tag{4.21}$$

from a two-sided point of view, or

$$H_0 : \gamma = 0 \quad vs \quad H_1 : \gamma < 0, \tag{4.22}$$

for its one-sided version.

Since we are in a POT context, the statistical tests will only be performed with the exceedances obtained from the available data. Let then $(W_1, \ldots, W_m)$ denote the m exceedances over a non-random threshold u, as defined in (3.19), extracted from the available random sample $(X_1, \ldots, X_n)$. The first test discussed was proposed by Gomes and van Monfort (1986) and will be used to test (4.22), using the following test statistic:

$$G_m = \frac{W_{m:m}}{W_{[m/2]+1:m}}$$

Under the validity of H_0, we have

$$G_m^* = \log 2 \, G_m - \log m \quad \xrightarrow[m\to\infty]{d} \quad Z \frown \Lambda \tag{4.23}$$

and, at the asymptotic level α, H_0 is rejected if $G_m^* \leq \mathcal{G}_\alpha$, where $\mathcal{G}_\varepsilon$ represents the standard Gumbel ε-quantile. This decision rule allows us to calculate the corresponding p-value:

$$p(G_m^*) = \Lambda(G_m^*).$$

Hence, applying the test statistic to our observed sample of exceedances $(w_1, \ldots, w_{49})$, the **R** software yields (see Appendix A.28):

```
[1] g_m= 1.128476   g_m*= -3.10962   p-value= 1.846555e-10
```

At the asymptotic level $\alpha = 0.05$, the observed value is highly significant, leading us to the rejection of the null hypothesis. The Exponential model is then not selected by this test procedure as a suitable parametric model for the excesses Y. Before proceeding to the next statistical test, one point must be stressed. We used the asymptotic distribution of the statistic G_m^* to take a decision about H_0. However, Gomes and van Monfort (1986) propose simulated critical points for small and moderate sample sizes, as $n = 20$, $n = 100$ and $n = 250$, at levels $\alpha = 0.05$ and $\alpha = 0.1$. These critical points can be found in Table 4.6, where $x\downarrow$ denotes values smaller than x.

Table 4.6: *Simulated critical points for the test statistic G_m^* for statistical choice of GPd models.*

		Critical region for $H_1 : \gamma < 0$	
Statistic	m	.10	.05
G_m^*	20	$-.89\downarrow$	$-1.19\downarrow$
	100	$-.94\downarrow$	$-1.21\downarrow$
	250	$-.86\downarrow$	$-1.13\downarrow$
	∞	$-.83\downarrow$	$-1.09\downarrow$

Gomes and van Monfort (1986)

The table does not consider our sample size $m = 49$, where $m = 20$ is the nearest size available. However, the observed value $g_m^* = -3.10962$ is significantly lower than any of the tabled critical points, even for the asymptotic ones. We maintain then the rejection decision of H_0. We are now able to go further with other statistical tests.

After presenting a LAN test for checking the null hypothesis of a Gumbel distribution against a GEVd, in the context of a Block Maxima approach, Marohn (2000) includes a GPd-test procedure in his article, based on the sample coefficient of variation, for testing (4.21) and (4.22). The necessary test statistic to perform the mentioned tests was already discussed in Gomes and van Monfort (1986). The test statistic to be used is given by

$$T_m = \frac{1}{2}\left(\frac{S_W^2}{(\overline{W} - u)^2} - 1\right),$$

where $S_W^2 = \frac{1}{m}\sum_{i=1}^{m}(W_i - \overline{W})^2$ is the sample variance.

Under H_0, we have

$$T_m^* = \sqrt{m}\,T_m \quad \xrightarrow[m\to\infty]{d} \quad Z \frown \mathcal{N}(0,1) \tag{4.24}$$

and, at the asymptotic size of α, H_0 is rejected if $|T_m^*| \geq z_{1-\alpha/2}$ or if $T_m^* \leq z_\alpha$, for the tests in (4.21) or in (4.22), respectively. The corresponding p-values are obtained with

$$p(T_m^*) = 2 - 2\Phi(|T_m^*|)$$

or

$$p(T_m^*) = \Phi(T_m^*).$$

Using simulation, Marohn (2000) showed that, for the two-sided version of the test, the test statistic is biased, with very poor power for small and moderate sample sizes, leading to reasonable results only for large sample sizes ($m \geq 500$). As we are working with a sample of $m = 49$ exceedances, we will only perform the left-sided version of the test given by (4.22). The **R** software yields the following results for $u = 80$ (see Apendix A.29):

```
[1] One-sided Test
 t_m= -0.2595148   t_m*= -1.816604   p-value= 0.03463891
```

At the asymptotic size of $\alpha = 0.05$, the null hypothesis is rejected, favouring then the GPd with $\gamma < 0$ as a suitable model to be fitted to the excesses Y, as already pointed in the preliminary analysis.

As for the Block Maxima approach, we can apply a LRT to the sample of exceedances in order to test (4.21). Let $(W_1, \ldots, W_m)$ be a random sample of exceedances, where $W_i \frown H_\gamma$, for $i = 1, \ldots, m$, with H_γ defined in (3.22). Let $\ell(\gamma, u, \sigma_u | w_1, \ldots, w_m)$ be the respective unrestricted log-likelihood function, where $\ell(0, u, \sigma_u | w_1, \ldots, w_m)$ denotes the restricted log-likelihood function, which corresponds to the Exponential case. The LRT statistic is given by

$$\mathbf{L} = -2\left(\ell(0, u, \hat{\sigma}_{u,H_0} | w_1, \ldots, w_m) - \ell(\hat{\gamma}_{H_\gamma}, u, \hat{\sigma}_{u,H_\gamma} | w_1, \ldots, w_m)\right), \tag{4.25}$$

with $\hat{\sigma}_{u,H_0}$ and $(\hat{\gamma}_{H_\gamma}, \hat{\sigma}_{u,H_\gamma})$ denoting the ML estimators for H_0 and H_γ models, respectively.

Under the null hypothesis, we have

$$\mathbf{L} \quad \xrightarrow[m\to\infty]{d} \quad Z \frown \chi^2_{(1)}.$$

To achieve a higher accuracy in the χ^2-approximation, Reiss and Thomas (2007) recommend the Bartlett correction, yielding the statistic

$$\mathbf{L}^* = \frac{\mathbf{L}}{1 + 4/m} \quad \xrightarrow[m\to\infty]{d} \quad Z \frown \chi^2_{(1)}. \tag{4.26}$$

For the test in (4.21), at the asymptotic size α, the null hypothesis is rejected if

$$\mathbf{L}^* \geq \chi^2_{1,1-\alpha},$$

where $\chi^2_{1,\varepsilon}$ stands for the $\chi^2_{(1)}$ ε-quantile.

The corresponding p-value can be calculated as follows:

$$p(\mathbf{L}^*) = 1 - \chi^2_{(1)}(\mathbf{L}^*).$$

The **R** software allows us to obtain the parameters ML estimates required by (4.25) (see Appendix A.30):

```
[1] Exponential ML estimates
 sigma_u= 6.644596
[2] GPd ML estimates
 gamma= -0.7268204   sigma_u= 12.04526
```

The final ML estimates are then

$$\hat{\sigma}_u = 6.644596, \tag{4.27}$$

for the Exponential distribution, and

$$(\hat{\gamma}, \hat{\sigma}_u) = (-0.7268204, 12.04526), \tag{4.28}$$

for the GPd. These latter estimates can be compared with the preliminary estimates given by (4.20).

Now, with the final ML estimates in hand, we can proceed to the LRT by means of the statistic given by (4.26). The **R** software produces the following results (see Appendix A.31):

```
[1] l= 12.93565   l*= 11.95937   p-value= 0.0005437322
```

At the asymptotic size $\alpha = 0.05$, the Exponential null hypothesis is rejected, leading to the same decision as for the tests presented by Gomes and van Monfort (1986) and Marohn (2000).

Until now, all the performed tests rejected the Exponential distribution as a suitable parametric model to be fitted to the r.v. Y. Notice that, the preliminary analysis already pointed to the GPd as a suitable model for our data. We can now complete this Section performing the same goodness-of-fit tests discussed in the Block Maxima approach: the Kolmogorov-Smirnov, Cramér-von Mises and Anderson-Darling tests. As we are not in a heavy-tail context, the test proposed by Kozubowski et al. (2009) will not be applied here.

Lilliefors (1969) studied the Kolmogorov-Smirnov test in the context of the Exponential distribution with unknown parameters. Since the Exponential distribution is embodied in the GPd distribution when $\gamma = 0$, we can use the procedures described in his work to check the goodness-of-fit of the Exponential distribution to the r.v. Y. Based on (4.15), the Kolmogorov-Smirnov statistic for the null hypothesis of Exponential model is given by

$$D_m = \max_{1\leq i\leq m}\left(\left|1-\exp\left(-\frac{Y_{i:m}}{\hat{\sigma}_u}\right)-\frac{i}{m}\right|,\left|1-\exp\left(-\frac{Y_{i:m}}{\hat{\sigma}_u}\right)-\frac{i-1}{m}\right|\right), \tag{4.29}$$

where $\hat{\sigma}_u$ stands for the ML estimator of σ_u for the Exponential model.

With $\hat{\sigma}_u$ given by (4.27), we can easily compute the Kolmogorov-Smirnov statistic with the **R** software (see Appendix A.32):

```
Kolmogorov-Smirnov statistic:   0.1383868
```

The observed value must be compared with the critical values given by Lilliefors (1969), rejecting then the null hypothesis of Exponential distribution if the observed value exceeds the respective critical point. We transcribe some critical values in Table 4.7.

Table 4.7: *Simulated critical values of the Kolmogorov-Smirnov statistic adapted to the Exponential distribution with unknown parameters.*

		Level of significance for D_m		
Statistic	m	.10	.05	.01
D_m	5	.406	.442	.504
	10	.295	.325	.380
	15	.244	.269	.315
	20	.212	.234	.278
	30	.174	.192	.226
	>30	$.96/\sqrt{m}$	$1.06/\sqrt{m}$	$1.25/\sqrt{m}$

Lilliefors (1969)

Provided that $m > 30$, the critical values are $\frac{.96}{\sqrt{49}} = 0.137$ and $\frac{1.06}{\sqrt{49}} = 0.151$, for $\alpha = 0.1$ and $\alpha = 0.05$, respectively. The decision seems jeopardized , since the

observed value lies between the two critical values. In any case, we know that the goodness-of-fit tests are reputed to be conservative. So, because of the proximity of the critical points, we cannot maintain the null hypothesis with conviction. The Exponential model is then rejected.

For the Cramér-von Mises and Anderson-Darling goodness-of-fit tests, we refer to Choulakian and Stephens (2001). This time, the distribution postulated by the null hypothesis is the GPd, with unknown parameters. According to (4.16) and (4.17), the two statistics are given by the following expressions, for the GPd:

1. Cramér-von Mises

$$W_m^2 = \sum_{i=1}^{m} \left(H_{\hat{\gamma}}(Y_{i:m}|\hat{\sigma}_{u,H_\gamma}) - \frac{2i-1}{2m} \right)^2 + \frac{1}{12m}, \tag{4.30}$$

2. Anderson-Darling

$$A_m^2 = -m - \frac{1}{m} \sum_{i=1}^{m} \big\{ (2i-1) \log(H_{\hat{\gamma}}(Y_{i:m}|\hat{\sigma}_{u,H_\gamma})) + \\ + (2m+1-2i) \log(1 - H_{\hat{\gamma}}(Y_{i:m}|\hat{\sigma}_{u,H_\gamma})) \big\}, \tag{4.31}$$

where $\hat{\gamma}, \hat{\sigma}_{u,H_\gamma}$ represent the ML estimators for the GPd, H_γ.

With the estimates $(\hat{\gamma}, \hat{\sigma}_{u,H_\gamma})$ given by (4.28), the observed values for the two statistics are given below, using the **R** software (see Appendix A.33):

```
Cramer-von Mises statistic:    0.09454922
Anderson-Darling statistic:    0.5972052
```

The observed values are now compared with the critical values given by Choulakian and Stephens (2001). The null hypothesis of GPd is rejected is the observed values exceed the critical tabled values. Part of the tabled critical points are presented in Table 4.8.

The asymptotic critical points were obtained by simulation, for γ between -0.5 and 0.9, and, according to Choulakian and Stephens (2001), they can be used with good accuracy for $m \geq 25$. Since γ was estimated, the table should be entered at $\hat{\gamma}$ and if $\hat{\gamma} < -0.5$, the table should be entered at $\gamma = -0.5$. As we have $\hat{\gamma} = -0.7268204$, we follow the suggestion of the authors and enter the table at $\gamma = -0.5$. At the asymptotic level $\alpha = 0.05$, the null hypothesis of GPd is not rejected, for both statistics, since the observed values do not exceed the tabled critical points. Considering other significance levels do not change our decision. Although the tests based on Cramér-von Mises and

Table 4.8: *Simulated critical values of the Cramér-von Mises (normal style) and Anderson-Darling (bold) statistics adapted to the GPd with unknown parameters.*

	Upper-Tail Asymptotic Percentage Points		
γ	.10	.05	.01
.9	.094	.115	.165
.9	**.641**	**.771**	**1.086**
.5	.101	.124	.179
.5	**.685**	**.830**	**1.180**
.1	.116	.144	.210
.1	**.766**	**.935**	**1.348**
0	.124	.153	.224
0	**.796**	**.974**	**1.409**
-.1	.129	.160	.236
-.1	**.831**	**1.020**	**1.481**
-.5	.174	.222	.338
-.5	**1.061**	**1.321**	**1.958**

Choulakian and Stephens (2001)

Anderson-Darling statistics are conservative, the observed values are not in a doubtful region and, in particular, the Anderson-Darling observed value of the statistic is somehow distant from the critical point.

All the tests performed in this Section are concordant: the GPd with $\gamma < 0$ must be selected, to the detriment of the Exponential distribution, as a suitable parametric model to be fitted to the r.v. Y. Provided that we chose a parametric model, we can now proceed to parametric estimation and respective inference.

c) *Parametric estimation of extreme events*

From last Section, we selected the GPd with $\gamma < 0$ as an appropriate parametric family to be fitted to the excesses Y above the fixed threshold $u = 80$. Notice that, contrary to the Block Maxima approach, the preliminary analysis and the statistical tests of the POT method pointed to the same direction: the underlying d.f. F, associated to the r.v. X, the $\dot{V}O_2max$ of a top-athlete, is not exponential right-tailed, but light right-tailed. In particular, the ML estimate of the EVI, given by (4.28), is somewhat small, indicating a very light right tail.

As stressed in Section 3.1.3.3, the PWM estimators perform better than the ML estimators in a small sample context. Since our sample consists of $m = 49$ excesses, PWM estimators may be particularly useful. The **R** software gives the following estimates for the GPd (see Appendix A.34) and the estimation results of both methods are presented

in Table 4.9.

```
[1] GPd PWM estimates
gamma= -0.5086472   sigma_u= 10.0245
```

Table 4.9: *ML and PWM estimates for the parameters of the GPd for the $\dot{V}O_2max$ data.*

Estimation method	$\hat{\gamma}$ (shape)	$\hat{\sigma}_u$ (scale)
ML	-0.7268204	12.04526
PWM	-0.5086472	10.0245

The difference between the two methods is more evident for the EVI, since the PWM method also elects a light right tail for the d.f. F, but heavier than the ML method. The PWM scale estimate is lower than the ML one, pointing to a smaller dispersion of the $\dot{V}O_2max$ excesses. As we did for the Block Maxima approach, we can check the fit quality of the GPd with diagnosis tools of the **R** software. Here, we apply the tools to the ML and PWM fits, which can be seen in Figures 4.17 and 4.18 (see Appendix A.35 for **R** details).

Based on the first two plots, the two estimation methods provide a satisfactory fit of the GPd. The main difference appears on the Density plot, where the PWM method shows a lower discrepancy between the kernel density and the fitted density. Note the very light right tail of the ML fitted density. Recall that the ML estimation provided a low estimate for the EVI, available in Table 4.9.

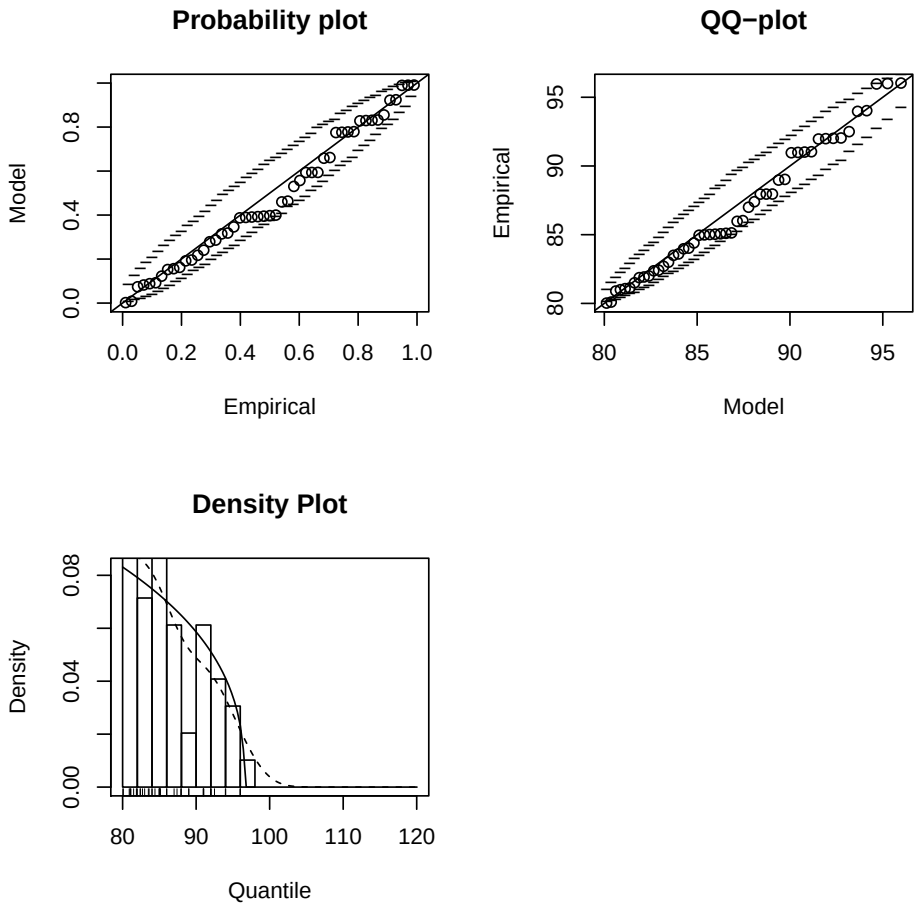

Figure 4.17: *Diagnosis plots for the ML fit of the GPd.*

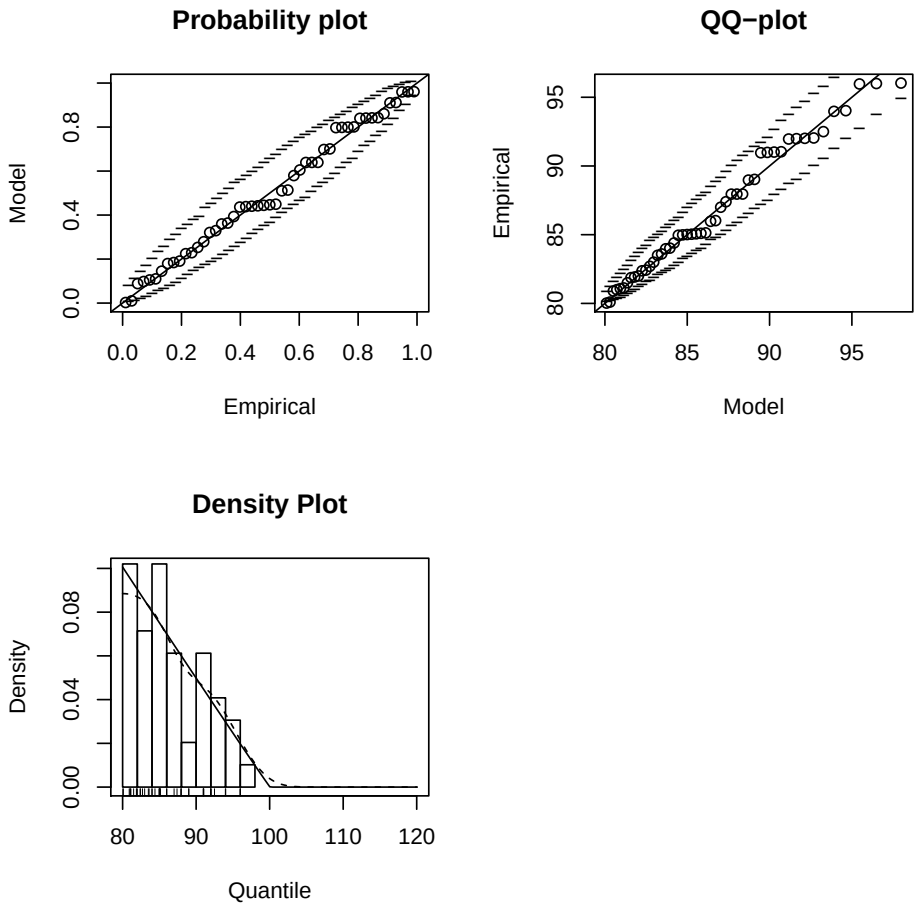

Figure 4.18: *Diagnosis plots for the PWM fit of the GPd.*

Now we have got rid of the point estimates of the GPd parameters, we can obtain CI's for the same parameters using the profile likelihood method of Section 3.1.3.5. However, notice that the estimate of γ is dangerously near -1. Since the profile likelihood method is a likelihood-based method, we know from Section 3.1.2.1 that the ML procedure is not applicable for $\gamma \leq -1$, compromising then the construction of CI's. The profile log-likelihood function for γ is depicted in Figure 4.19 (see Appendix A.36 for **R** details).

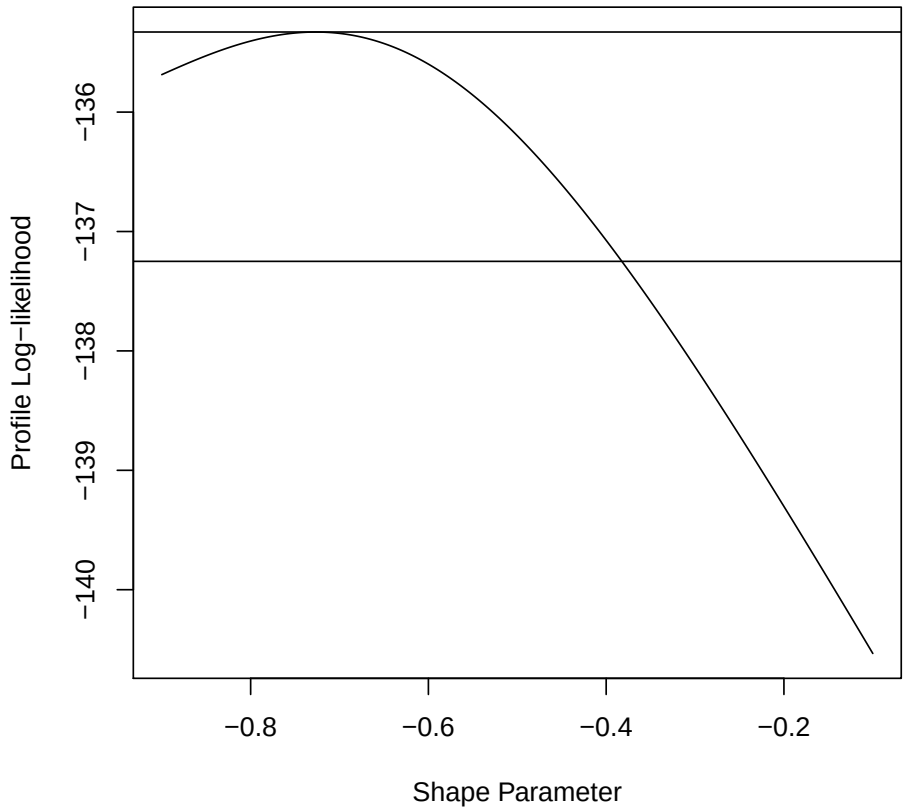

Figure 4.19: *Profile log-likelihood function for the POT approach of* $\dot{V}O_2max$.

The plot confirms the non-existence of the profile log-likelihood function for $\gamma \leq -1$. Thus, the obtention of a 95% profile CI for γ with the lower horizontal line on the plot is compromised.

We can opt for the traditional method of constructing CI's described in Section 3.1.2.4, but, as stressed by Beirlant et al. (2004), these intervals may be very misleading, since the Normal approximation may result in a poor inference, penalized by a small size sample. We could seek alternative methods of construction CI's, but this is out of the scope of

this thesis. Consequently, under the POT approach in this case study, we do not make any inference based on the estimates of the GPd parameters.

Before leaving the POT approach of the $\dot{V}O_2max$, we can still obtain estimates of exceedances probabilities for the r.v. X under study and of the right endpoint of F, provided that $\gamma < 0$, using (3.29) and (3.30), respectively. As for the Block Maxima approach, we can estimate the probability of exceeding the actual sample maxima of 96 ml/kg/min. Using then (3.29) for the threshold $u = 80$ and getting the estimates $(\hat{\gamma}, \hat{\sigma}_u)$ from Table 4.9, we can compute $P(X > 96)$ using the **R** software:

```
[1] Maximum Likelihood: P(X>96)= 0.006457043
[2] Probability Weighted Moments: P(X>96)= 0.02481165

[1] Maximum Likelihood: x^F= 96.57254
[2] Probability Weighted Moments: x^F= 99.70816
```

We notice that the ML estimate of the right endpoint is very close to the sample maximum, leaving almost no more space for an improvement of the current record. Consequently, this estimation method points to a steady state of the current $\dot{V}O_2max2max$ record, with a very low probability of exceeding it. This result is not surprising, because of the low estimate of the EVI obtained by the ML method, as stressed before. With a higher estimate of the EVI, available on Table 4.9, the PWM method gives some space for improving the current $\dot{V}O_2max$ record, with a 2.5% probability of surpassing it. Notice the heavier right tail of the fitted PWM density on Figure 4.18, which results in a higher estimate of the right endpoint of F.

To close this Section, it would be interesting to compare the POT results with the Block Maxima results. The comparison between the GEVd and the GPd estimates makes more sense, by analogy issues. All the results are then gathered in Table 4.10.

As we can see, the two approaches lead to substantial differences between estimates. By analogy issues, the chosen threshold $u = 80$ and the estimate of the GEVd location parameter, $\hat{\lambda}$, were juxtaposed, since u can be seen as a location parameter of the GPd. Notice how the ME-plot of Figure 4.2 selected the threshold $u = 80$, which is very close to the GEVd location parameter estimate. Remember that we are not working with a large sample, which is sufficient to create differences between alternative approaches. However, we have a battery of estimates at our disposal, which provides us some guidance and some conclusions about the $\dot{V}O_2max$. The POT approach attributes a better quality to the current record of 96 ml/kg/min, with a lower right endpoint estimate and a lower probability of surpassing the current maximum level than the Block Maxima does.

Table 4.10: *Comparison of the results for the $\dot{V}O_2max$ between the Block Maxima and POT approaches.*

		POT GPd ($\gamma < 0$)	Block Maxima GEVd ($\gamma < 0$)
$\hat{\gamma}$	ML	-0.7268204	-0.2431824
	PWM	-0.5086472	-0.2023684
	profile CI (95%)	-	(-0.4388217,-0.01412509)
$u/\hat{\lambda}$	ML	80	80.64481
	PWM	80	80.46483
	profile CI (95%)	-	(79.0378578,82.29713706)
$\hat{\delta}$	ML	12.04526	6.166826
	PWM	10.0245	6.307003
	profile CI (95%)	-	(5.1364086,7.63658965)
$P(Y > 96)$	ML	0.006457043	0.02158176
	PWM	0.02481165	0.03250008
x^F	ML	96.57254	106.0037
	PWM	99.70816	111.6308

4.1.2 Semi-Parametric data analysis

a) *Testing the extreme value index sign*

Following now a semi-parametric approach as described in Section 3.2, we can start our analysis as in parametric approach, testing the EVI sign with the methodology discussed in Section 3.2.6. Our interest here is to make an *a priori* selection of the most suitable max-domain of attraction for our $\dot{V}O_2max$ data. Testing EVI sign may also be useful to select the appropriate semi-parametric estimators, since some of them rely on the particular sign of γ. This way, our concern is to perform the tests in (3.61) or in (3.63) by means to the test statistics defined in (3.67), (3.68) and (3.69).

As a preliminary general interest, we can perform the test

$$H_0 : F \in \mathcal{D}(G_0) \quad vs. \quad H_1 : F \in \mathcal{D}(G_\gamma)_{\gamma \neq 0}, \tag{4.32}$$

simply to ascertain the sign of γ.

As the statistics are a function of the random threshold k, we can represent graphically the sample paths of each statistic, plotting $R_n^*(k)$, $W_n^*(k)$ and $T_n^*(k)$ versus k. The resulting curves can be visualized in Figure 4.20 (see Appendix A.38 for **R** details).

Based on the rejection rule in (3.70) and considering the behaviour of the three statistics, we observe a clear trend of Greenwood and Ratio statistics towards the respective critical values, even getting into the rejection zone. Concerning Hasofer-Wang statistic, the respective sample path is almost always outside of the acceptance region, except for

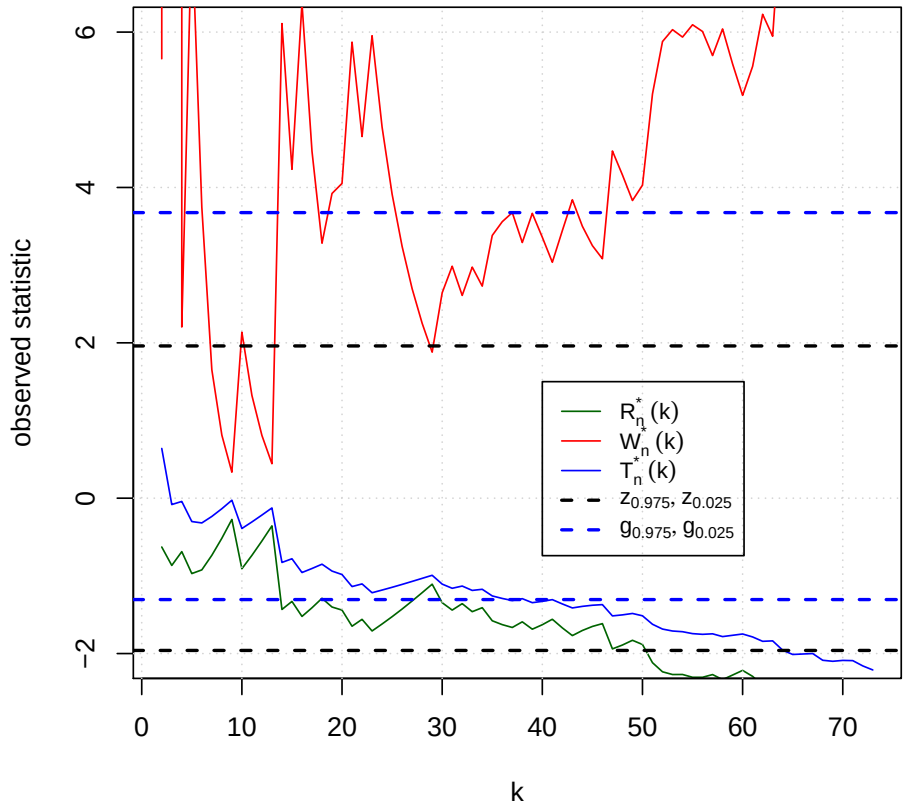

Figure 4.20: *Sample paths of Greenwood, Hasofer-Wang and Ratio statistics in a two-sided test context.*

a brief interval of k, and the curve exhibits the same trend mentioned for the two other statistics. Therefore, at the asymptotic level of $\alpha = 0.05$, we find evidence to reject the null hypothesis in (4.32), pulling away the Gumbel max-domain of attraction.

If we want to be more specific and check if we are in a Weibull max-domain, we can perform the one-sided version of the test, at the asymptotic level of $\alpha = 0.05$:

$$H_0 : F \in \mathcal{D}(G_0) \quad vs. \quad H_1 : F \in \mathcal{D}(G_\gamma)_{\gamma<0},$$

and modify the respective critical values according to the rule in (3.72). The resulting plot can be seen in Figure 4.21 (see Appendix A.39 for **R** details).

Observing the plot, we conclude exactly the same as for the two-sided test: at the asymptotic level $\alpha = 0.05$, we find evidence to reject H_0. One important point should be emphasized: as noted in Section 3.2.6, the Hasofer-Wang statistic W_n^* is specially addressed to detect a Weibull max-domain, since it it the most powerful of the three in that case. Paying attention to Hasofer-Wang statistic's sample path, we note that it is

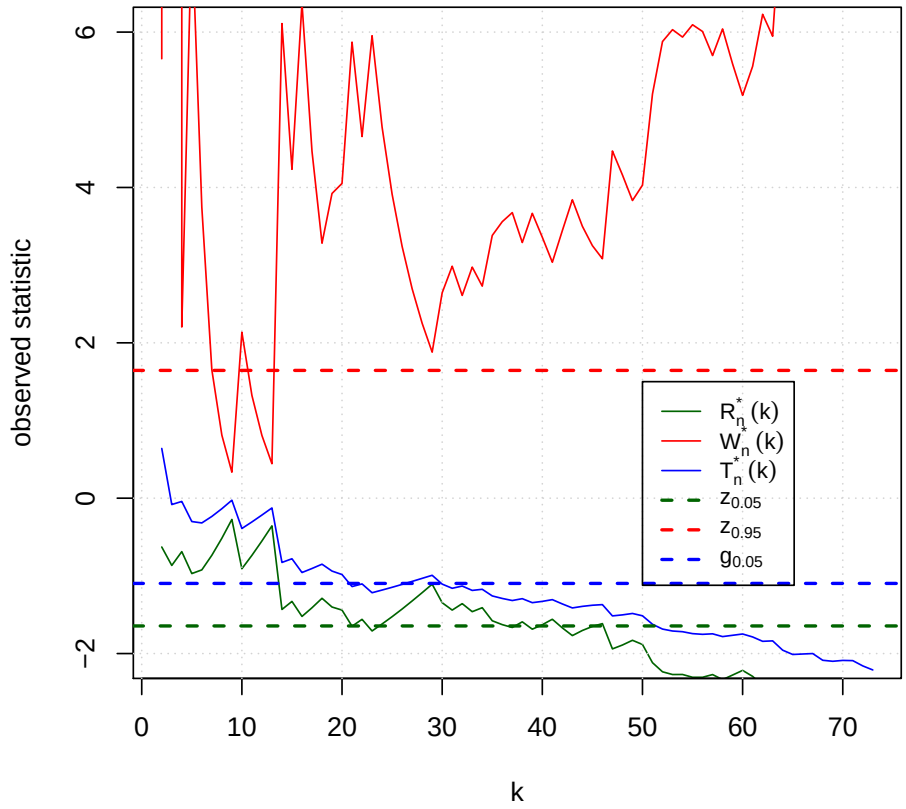

Figure 4.21: *Sample paths of Greenwood, Hasofer-Wang and Ratio statistics in a one-sided context.*

almost always in the rejection zone, except for a small interval of k, with a tendency to stay in that zone. As our intention here is ascertain the possibility of a Weibull max-domain, the Hasofer-Wang statistic is very useful to give us some guidance.

As a conclusion for these tests, we elect then a Weibull max-domain for the d.f. of the $\dot{V}O_2max$. Contrary to the parametric approach where the choice of the statistical model was doubtful, the semi-parametric approach points clearly in Weibull max-domain's direction.

b) *Heuristic choice of the random threshold*

The next step is to estimate then the EVI with some semi-parametric estimators presented in Section 3.2.3. However, as for the the test statistics discussed above, the semi-parametric estimators for γ are a function of the random threshold k. Consequently, a choice for k would be appropriate in order to obtain an estimate for the EVI. For this

choice, we follow the heuristic approach of Section 3.2.7. According to this approach, k is chosen where the estimates of each estimator are very similar. To have an idea about the similitude of the estimates, we can plot the sample path of each estimator on the same plot and choose a region where the sample paths are very close.

The first estimator seen was the Pickands estimator. This estimator may cause some troubles for the heuristic choice of k. Indeed, this estimator is known for its large asymptotic variance, given by (3.59). Since we are using a heuristic process for the choice of k based on a minimization of distances between γ estimates, the Pickands estimator may cause some distortion on the measurements. For now, we present the isolated sample path of Pickands' estimator in Figure 4.22 (see Appendix A.40 for **R** details). Taking a closer look at the scale of the plot, we see clearly a high volatility of the EVI-estimates. Consequently, the use of Pickands' estimator can lead to misleading results.

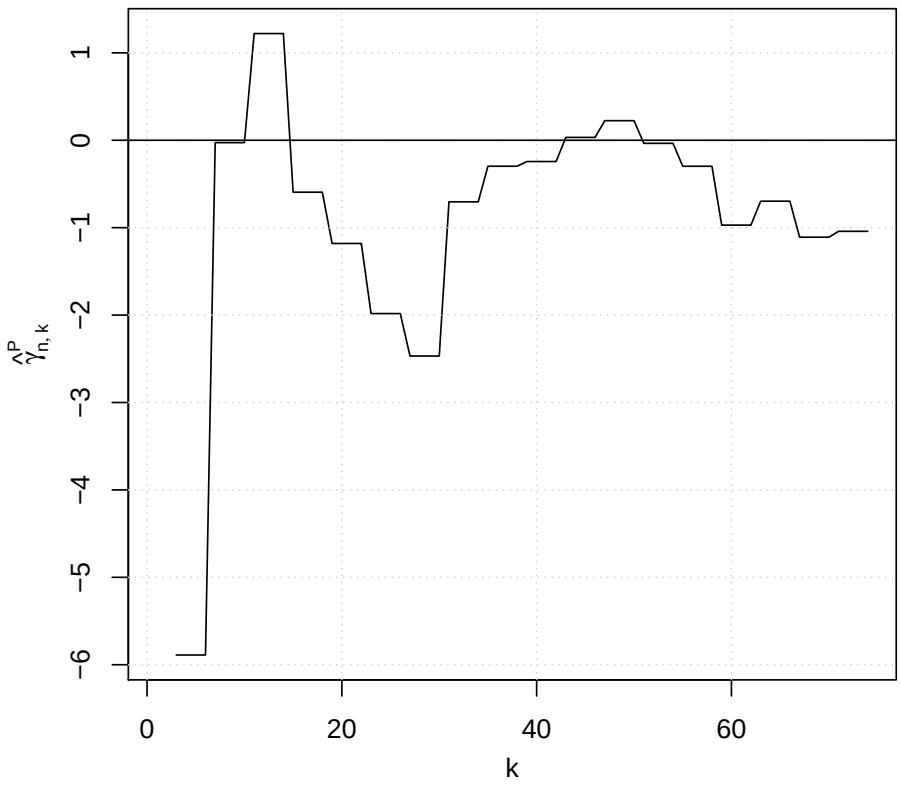

Figure 4.22: *Pickands-plot for the $\dot{V}O_2max$ data.*

Obviously, the Hill estimator will not be used, since it is only valid for $\gamma > 0$ and we are in a Weibull max-domain context. The Moment estimator, the Generalized Hill estimator and the Mixed Moment estimator can be chosen without danger, since they are

valid for $\gamma \in \mathbb{R}$. The Negative Hill estimator was developed for $\gamma < -0.5$ and as the EVI is unknown, we cannot guarantee that this condition is fulfilled. We can eventually consider the sample path of the Negative Hill estimator in Figure 4.23 (see Appendix A.41 for **R** details).

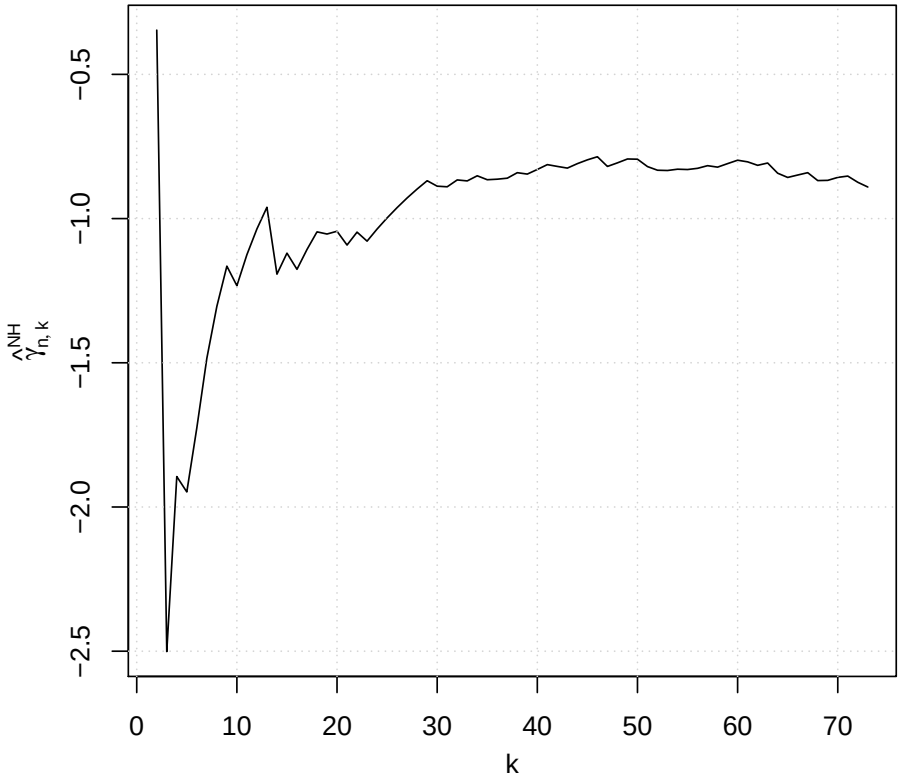

Figure 4.23: *Negative Hill estimator sample path for the $\dot{V}O_2max$ data*

The sample path exhibits a stable pattern sensibly after $k = 30$ and stays always under -0.5, except for the beginning part of the plot. To ascertain whether $\gamma < -0.5$, we can build CI's for γ from each possible value of the threshold k, using expression (3.57). However, from (3.59), we note that the asymptotic variance of $\hat{\gamma}^{NH}_{n,k}$ is only valid for $-1 < \gamma < -0.5$. We recall that the expressions of the asymptotic variances in (3.59) hold only if Theorem 3.56 and extra specific conditions are satisfied. Moreover, expression (3.57) is a simplified version of CI's for γ, intended to avoid the estimation of the second-order parameters ρ and β mentioned in Section 3.2.5. Finally, replacing γ by the Negative Hill estimate, we introduce some sampling variability, which has repercussions on the interval accuracy. All these reasons make it difficult to confirm if $\gamma < -0.5$ and, consequently, the Negative Hill estimator will be excluded from our analysis. Concerning the PORT estimators,

PORT-Hill will be excluded for the same reasons as the Hill estimator. To use the PORT-Moment and PORT-Mixed Moment estimators, we have to take into account the *tuning parameter* q, with $0 \leq q < 1$. To have some guidance, we can plot the sample path of each PORT estimator for different values of q. The sample paths are shown in Figure 4.24 (see Appendix A.42 for **R** details).

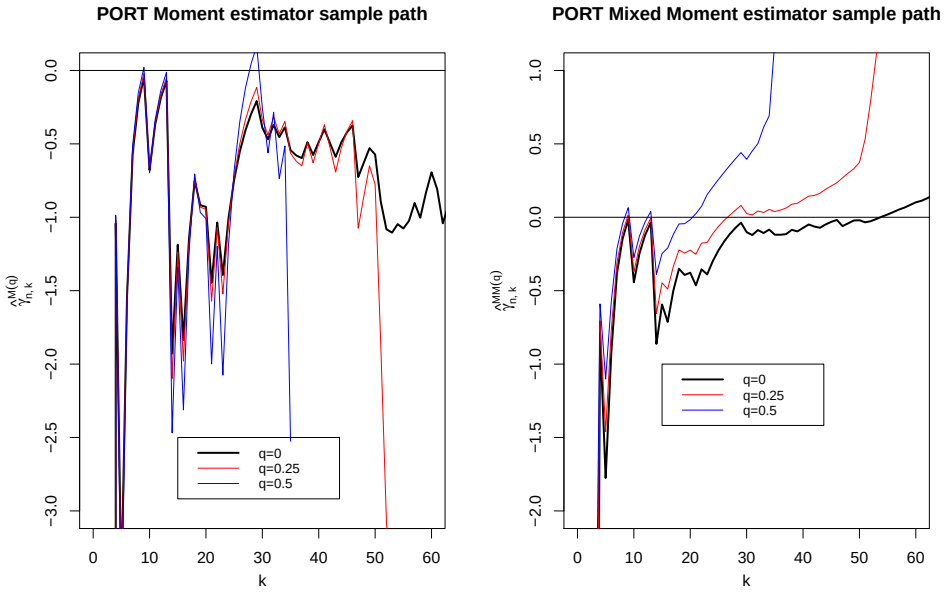

Figure 4.24: *PORT-Moment and PORT-Mixed Moment sample paths for the* $\dot{V}O_2max$ *data*

As the two plots exhibit, lower values for the *tuning parameter* q return more stable sample paths for the *PORT* estimators. Therefore, lower values of q would be desirable. Fraga Alves et al. (2009) suggest the use of $q = 0.1$ or even $q = 0.01$. We will choose the latter one.

Now that the semi-parametric estimators have been chosen for the heuristic choice of k, a plot with the sample paths of all the estimators may help us to ascertain eventual regions where a clear proximity of all the estimates is visible. This plot can be found in Figure 4.25 (see Appendix A.43 for **R** details).

Several observations can be made from Figure 4.25:

1. All the estimators are highly volatile for low values of k and exhibit lower variability sensibly after $k = 30$;

2. Despite a lower volatility after $k = 30$, none of the semi-parametric estimators are stable, provided that they all reveal a downward trend, except for the PORT-Mixed Moment estimator, which shows an upward trend;

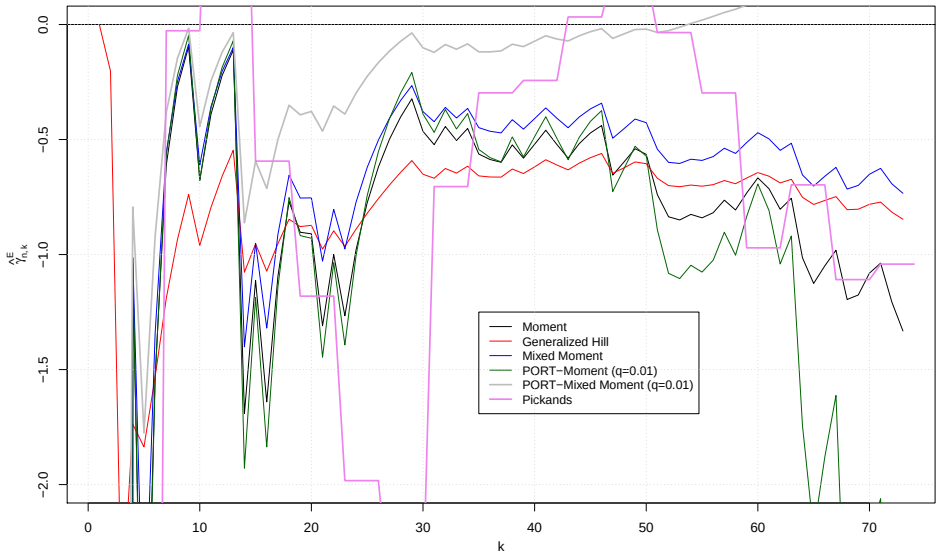

Figure 4.25: *Sample paths of Moment, Generalized Hill, Mixed Moment, PORT-Moment, PORT-Mixed Moment and Pickands estimators for the $\dot{V}O_2max$ data*

3. All the estimators are relatively close, excluding the PORT-Mixed Moment estimator, which has a distinct behaviour, remaining at the margin of the group formed by the other estimators, and the Pickands estimator;

4. Since it was stated above, the Pickands estimator exhibits an erratic behaviour, characterized by a high volatlity.

The choice ok k seems to have some obstacles and hence, to keep some homogeneity, we decided to apply the heuristic methodology of Section 3.2.7 without the PORT-Mixed Moment and Pickands estimators. Using then identity (3.73) by means of the **R** software (see Appendix A.44), the optimal value for k is given as follows:

```
[1] k opt= 19
```

The heuristic procedure selects $k = 19$ for the random threshold, choosing then $k+1 = 20$ top order statistics for semi-parametric inference. This optimal selection can be seen in Figure 4.26, where the function $\sum_{(i,j)\in\mathcal{E}:i\neq j}[\hat{\gamma}_{n,k}^{(i)} - \hat{\gamma}_{n,k}^{(j)}]^2$ from (3.73) attains its minimum at $k = 19$ (see Appendix A.45 for **R** details).

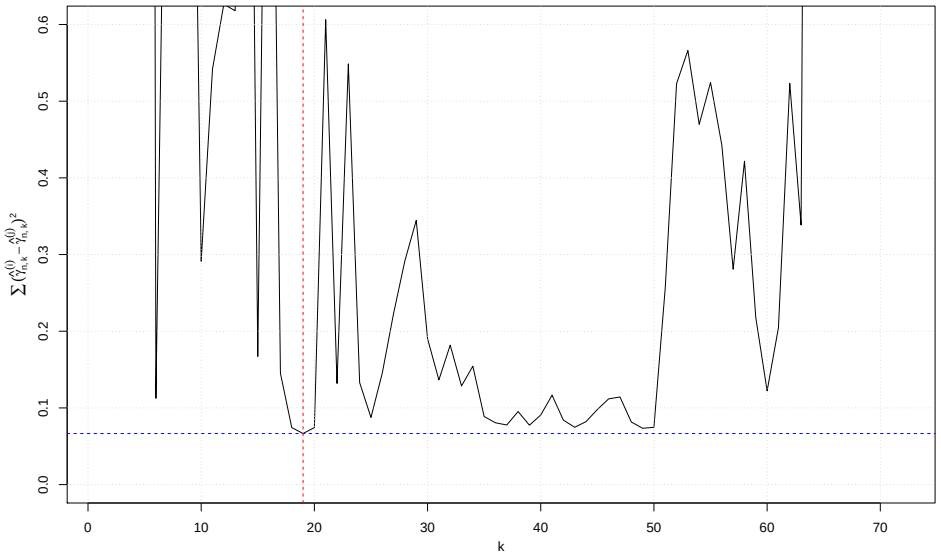

Figure 4.26: *Sample path of the function* $\sum_{(i,j)\in\mathcal{E}:i\neq j}[\hat{\gamma}_{n,k}^{(i)} - \hat{\gamma}_{n,k}^{(j)}]^2$

As we can observe in the plot of Figure 4.26, the minimum of the function is attained in a region of high variability, also visible in Figure 4.25, since the number of top order statistics is relatively low. This result is not surprising, since we know from Section 3.2.7 that a relatively low choice for k induce estimators with higher variances. We can observe another region of the plot where the values of the function are very close to the minimum attained: sensibly between $k = 35$ and $k = 50$. After $k = 50$, we have again a high volatility in the function due principally to the erratic behaviour of the PORT-Moment estimator, visible in Figure 4.25. So, an eligible value for k can also be chosen between $k = 35$ and $k = 50$. This region is more stable, with lower variance, but as values of k are higher, this induce a bias in the estimators. Since the balance between asymptotic variance and asymptotic bias, dealt with the AMSE introduced in Section 3.2.7, is out of the scope of this thesis, we will also elect a value for k between $k = 30$ and $k = 50$ by the same process of minimization. And to avoid the choice near $k = 50$, after which a high variability is visible again, we will minimize the function $\sum_{(i,j)\in\mathcal{E}:i\neq j}[\hat{\gamma}_{n,k}^{(i)} - \hat{\gamma}_{n,k}^{(j)}]^2$ for $k \in [35, 45]$ using **R** software (see Appendix A.46), yielding

```
[1] k opt= 43
```

This second choice for k involves then $k + 1 = 44$ of the top order statistics for semi-parametric inference. In Figure 4.25, the clear stability and proximity of all the estimates are also evident. For $k = 19$, we have equally a high proximity of all the estimates, but in a more volatile region.

c) *Estimation of the Extreme Value Index*

With two proposals for the threshold k, we can now proceed to the semi-parametric estimation of the EVI, turning to each estimator used during the heuristic procedure. With **R** software (see Appendix A.47) , we obtain the following results:

```
[1] k opt=19
 Moment: gamma= -0.903127    Generalized Hill: gamma= -0.8785233
 Mixed Moment: gamma= -0.7543541  PORT-Moment (q=0.01): gamma= -0.9177094

[2] k opt=43
 Moment: gamma= -0.5810163    Generalized Hill: gamma= -0.6317842
 Mixed Moment: gamma= -0.4489174
 PORT-Moment (q=0.01): gamma= -0.5879511
```

From the obtained results, we can immediately see that, for $k = 19$, the semi-parametric estimates for the EVI are lower than those of the parametric approach. For $k = 43$, the semi-parametric estimates are lower than those obtained under a Block Maxima approach, but lie between the ML and PWM estimates of the POT approach, staying closer to the latter ones. Recall that the POT approach worked with $m = 49$ exceedances and that the semi-parametric approach selected $k + 1 = 44$ top observations, almost the same portion of the sample. The semi-parametric approach seems to validate the PWM estimation, which performs better in small samples.

And now, with these estimates in hand, we are able to construct approximate CI's for the EVI, by means of expression (3.57). But first, we need to obtain estimates for the asymptotic variances of (3.59), replacing γ by the estimate of the respective estimator. For each of the four estimators used in the heuristic procedure, we can easily obtain such estimates combining (3.59) with **R** software (see Appendix A.48), yielding the following results:

```
[1] k opt=19
 Moment: s2_M= 4.037741    Generalized Hill: s2_GH= 1.13451
 Mixed Moment: s2_MM= 3.044318  PORT-Moment (q=0.01): s2_M(q)= 4.037741

 [2] k opt=43
  Moment: s2_M= 2.137555    Generalized Hill: s2_GH= 0.8409316
  Mixed Moment: s2_MM= 1.614211   PORT-Moment (q=0.01): s2_M(q)= 2.137555
```

The theoretical results are then confirmed: the estimated asymptotic variances are higher for the lowest of the two values for k. An analysis of the trade-off asymptotic variance-bias would be suitable. In this Section, we merely use these asymptotic variances estimates to construct approximate CI's for the EVI. Then, with the asymptotic variances in hand, the construction of approximate CI's for γ at the 95% asymptotic confidence level is now possible and the results are summarized in Table 4.11.

Table 4.11: *Semi-parametric approximate 95% confidence intervals for γ for the $\dot{V}O_2max$ data.*

Semi-parametric estimator of γ	$k = 19$	$k = 43$
Moment	-0.903127 **(-1.806653,0.0003987481)**	-0.5810163 (-1.018007,-0.1440253)
Generalized Hill	-0.8785233 (-1.357457,-0.3995895)	-0.6317842 (-0.9058749,-0.3576936)
Mixed Moment	-0.7543541 **(-1.538896,0.03018804)**	-0.4489174 (-0.8286641,-0.06917064)
PORT-Moment (q=0.01)	-0.9177094 (-1.821235,-0.01418371)	-0.5879511 (-1.024942,-0.1509601)

As expected, the CI's based on the lower value $k = 19$ have a larger amplitude than those based on the higher value $k = 43$, reflecting then the high variability of the estimators at lower levels of the random threshold k. Moreover, the intervals associated to the Moment and to the Mixed-Moment estimators contain zero, but at the very final of the interval. In general, these approximate CI's confirm us the Weibull max-domain of attraction for the $\dot{V}O_2max$, leading us to a finite right endpoint of the underlying d.f. F. We recall that the obtained CI's are simplified versions and more accurate intervals can be obtained with more precise techniques, involving the estimation of the second order parameters ρ and β, for instance. But, as already mentioned, this work is out of the scope of this thesis. We solely look for a simple and rough confirmation of the EVI's sign via CI's, confirming us the Weibull max-domain of attraction.

d) *Semi-parametric estimation of other extreme events*

To end this semi-parametric analysis and for some parallelism with the parametric analysis, the last step consists in estimating some other important parameters, discussed in Section 3.2.4: the normalizing constants $b\left(\frac{n}{k}\right)$ and $a\left(\frac{n}{k}\right)$, the exceedance probability of the current record of 96 ml/kg/min and finally, the right endpoint x^F of the underlying d.f. F.

For the attraction coefficients $b\left(\frac{n}{k}\right)$ and $a\left(\frac{n}{k}\right)$, we can use the identities (3.49) and (3.50) discussed in Section 3.2.4, for both values of k and each estimator used in the

heuristic process, when necessary. The identity for the location coefficient $b\left(\frac{n}{k}\right)$ is very simple, since it does not depend on γ, but only on k. Therefore, the **R** software gives us quickly the resulted needed, for both values of the random threshold (see Appendix A.49):

```
[1] k opt=19
b(n/k)= 87.4
[2] k opt=43
 b(n/k)= 81.125
```

A comparison with the parametric approach would be suitable. Since we are in a Weibull max-domain, consulting the results in Table 4.5 concerning the GEVd for $\gamma < 0$, we immediately notice a clear similitude of the ML and PWM estimates for the location parameter λ, with the semi-parametric estimate for $k = 43$. The semi-parametric approach seems to give some support to the second choice for the threshold, $k = 43$.

Focusing now on the scale coefficient $a\left(\frac{n}{k}\right)$, (3.50) shows us identically a direct dependence on k and an indirect one, via $\hat{\gamma}^{-}_{n,k}$. Therefore, as for $b\left(\frac{n}{k}\right)$, a different estimate can be calculated for each k, with our **R** software (see Appendix A.49):

```
[1] k opt=19
a(n/k)= 8.128112
[2] k opt=43
 a(n/k)= 9.956318
```

The gathered results are presented in Table 4.12.

Table 4.12: *Semi-parametric estimates of the location and scale coefficients for the $\dot{V}O_2max$ data.*

Estimator	$k = 19$	$k = 43$
$\hat{a}\left(\frac{n}{k}\right)$	8.1	9.956
$\hat{b}\left(\frac{n}{k}\right)$	87.4	81.125

As for the location coefficient, for the choice $k = 43$, the semi-parametric estimate of the scale coefficient is closer to the ones obtained for the scale parameter δ under a parametric approach (see Table 4.10). This is particularly true for the PWM method, which performs better on small and modest samples than the ML estimator. The choice $k = 43$ approximates once again the parametric (specially PWM) and semi-parametric approaches.

Now, remember the differences of the EVI estimates between the semi-parametric approach and the Block Maxima method. Comparing Table 4.10 with Table 4.11, we recall that the semi-parametric estimates of the EVI are lower in a semi-parametric context, inducing then a lighter right tail. As we have a lighter right tail of the underlying d.f. in a semi-parametric context, we surely have a lower right end-point, putting our current record closer to x^F. In order to confirm this exposition, an estimate of the right endpoint x^F is needed. For this, we can use expression (3.53). This time, the estimate of the right endpoint depends on k and γ, so that a table that condenses all the estimates would be practical. Note that the estimate of x^F depends precisely on the estimate of the scale coefficient $\hat{a}\left(\frac{n}{k}\right)$, since we can rewrite (3.53) as

$$\hat{x}^F = \max\left(X_{n:n}, X_{n-k:n} - \frac{\hat{a}\left(\frac{n}{k}\right)}{\hat{\gamma}}\right). \tag{4.33}$$

Therefore, before taking any decision about the right endpoint estimate, we have to take into account the scale effect and, as we can see in the aforementioned tables, the scale estimate is lower under a Block Maxima approach. It is time to see what our **R** software has to tell us (see Appendix A.50) .

```
[1] k opt=19
 Moment: xF= 96.39997    Generalized Hill: xF= 96.65202
 Mixed Moment: xF= 98.17493  PORT-Moment (q=0.01): xF= 96.25696

 [2] k opt=43
 Moment: xF= 98.26104    Generalized Hill: xF= 96.88405
 Mixed Moment: xF= 103.3035  PORT-Moment (q=0.01): xF= 98.05892
```

All the estimates can be gathered on Table 4.13.

Table 4.13: *Semi-parametric estimates for the right endpoint x^F of the underlying d.f. F for the $\dot{V}O_2max$ data.*

Semi-parametric estimator of γ	$k = 19$	$k = 43$
Moment	96.39997	98.26104
Generalized Hill	96.65202	96.88405
Mixed Moment	98.17493	103.3035
PORT-Moment (q=0.01)	96.25696	98.05892

Our suspicions were accurate. With a lighter right tail of the underlying d.f F in a semi-parametric context, the right endpoint x^F decreased substantially when compared

to the ML and PWM estimates of the Block Maxima method. Note the similitude with the POT estimates. Then, following a semi-parametric analysis, we conclude that the current record of 96 ml/kg/min has a slight capacity to increase much, since it is very close to the estimated upper bound. Even the highest semi-parametric estimate given by the Mixed Moment estimator for $k = 43$ does not surpass much the level 100 ml/kg/min. At this point, it would be interesting to consider the sample path of the right endpoint x^F, for each estimator covered by the heuristic process. The sample paths are visible in Figure 4.27 (see Appendix A.51).

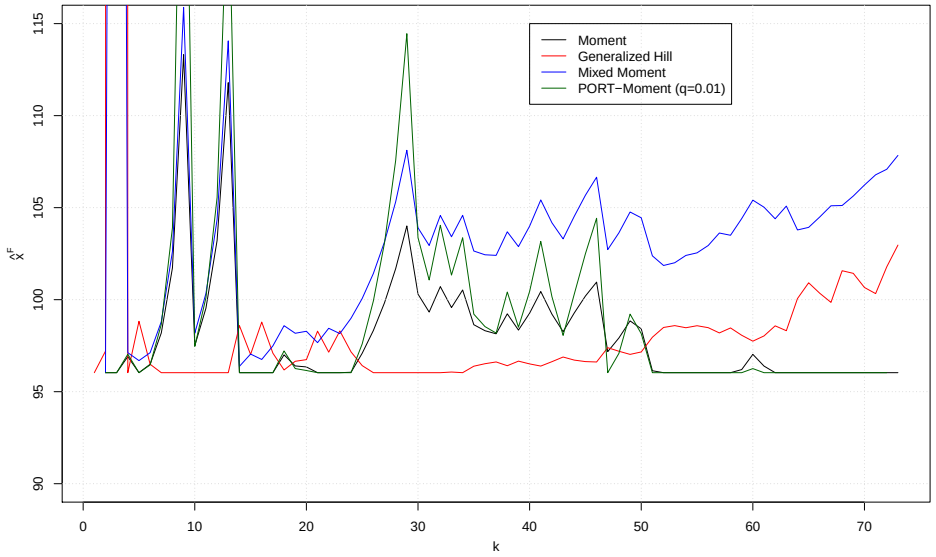

Figure 4.27: *Sample paths of Moment, Generalized Hill, Mixed Moment and PORT-Moment estimators for the right endpoint of the underlying distribution function F for the $\dot{V}O_2max$ data.*

The Moment and PORT-Moment estimators exhibit a clear volatility for low levels of the threshold k, describing a decreasing pattern sensibly after $k = 30$. This decreasing tendency causes some troubles, since a downward trend conducts the sample path of the estimator towards the sample maximum and, how it was emphasized in Section 3.2.4, the estimate of x^F cannot lie under the sample maximum, $X_{n:n}$. Remember that this restriction was imposed on the estimator of x^F, resulting in expression (3.53). Figure 4.27 makes that clear, since, sensibly after $k = 50$, the Moment and PORT-Moment estimators stay at $\hat{x}^F = 96$, since they cannot decrease anymore. The Mixed Moment estimator stays at a high level after $k = 30$ and switch to an upward trend after $k = 50$, contrasting with the two previous estimators. Concerning the Generalized Hill estimator, its behaviour is also distinct, provided that it has a stable path all the time at a very low level, taking an upward path after $k = 50$.

Looking at Figure 4.27, we can be tempted to apply the same heuristic process used for the choice of k, in order to obtain the "best" estimate of x^F. It may be interesting to determine which value of k minimize the difference between all the estimates of Figure 4.27 and compare it with the optimal k values obtained from the heuristic process applied to the EVI-estimators. Using then (3.73) and replacing $\hat{\gamma}_{n,k}$ by $\hat{x}^F$, we can obtain the solution of the minimization problem with the **R** software (see Appendix A.52):

```
[1] k opt= 6
```

The optimal selected value for k which provides the "best estimate" for the right endpoint x^F is somewhat different from the optimal values obtained from the heuristic process applied to the EVI-estimators. To gain some clarity for such a low choice, we can look at the plot of the function defined in (3.73), adapted for the right endpoint estimators, which can be seen in Figure 4.28 (see Appendix A.52 for **R** details).

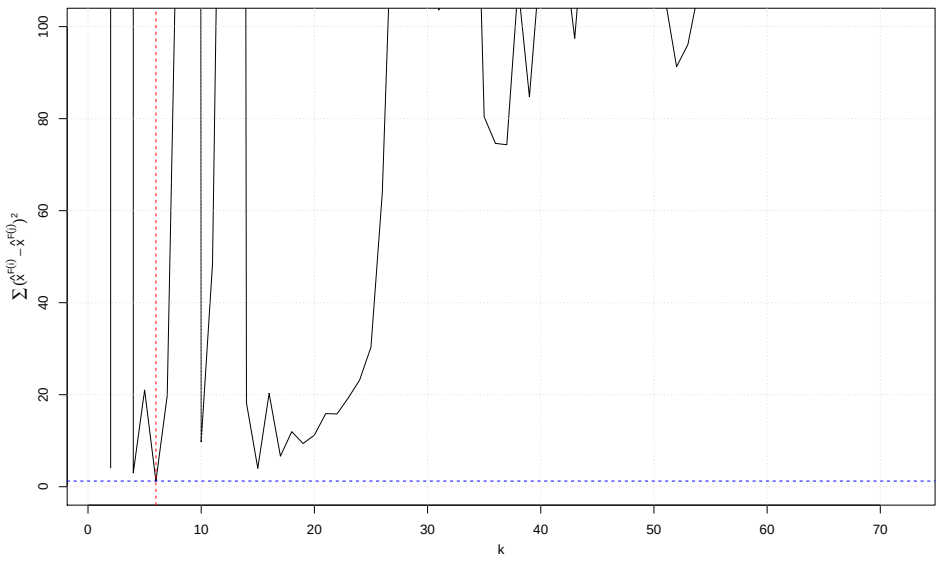

Figure 4.28: *Heuristic choice of the threshold k for the right endpoint estimation for the $\dot{V}O_2max$ data.*

The plot demonstrates a high variability of the function, which attains its lower values until $k = 20$, After that, the function is characterized by an increasing behaviour. Note that the optimal choice of $k = 6$ lies in a very unstable region of the path. Another choice of k is possible at $k = 15$, since the function attains a local minimum at this value for $10 < k < 20$, but, as for the optimal choice $k = 6$, this value lies in a very unstable region of the function. Recall that the first choice for k was $k = 19$, which was seen to be in a very unstable region of the plot depicted in Figure 4.26. The choice $k = 43$ provides more

stability of the estimates. As a curiosity, we present, in Table 4.14, the estimates of the right endpoint for the newly calculated k values.

Table 4.14: *Heuristic semi-parametric estimates for the right endpoint x^F of the underlying d.f. F for the $\dot{V}O_2max$ data.*

Semi-parametric estimator of γ	$k = 6$	$k = 15$
Moment	96.49078	96.03333
Generalized Hill	96.49083	97.0235
Mixed Moment	97.11885	97.0445
PORT-Moment (q=0.01)	96.45587	96.03333

Comparing the results with Table 4.13, we see some similarity with the estimates obtained with $k = 19$, which was selected in a k-region with the same characteristics as for the newly k values.

Taking back the firstly calculated k-values, we can use (3.58) to obtain CI's for the right endpoint, according to the different EVI-estimators used in the heuristic process. The estimated variances were already calculated, when obtaining the CI's for the EVI in Table 4.11. Taking then $\alpha = 0.05$, the CI's for x^F are presented in Table 4.15 (see Appendix A.53).

Table 4.15: *Semi-parametric approximate confidence intervals for x^F for a $\alpha = 5\%$ choice.*

Semi-parametric estimator of γ	$k = 19$	$k = 43$
Moment	96.39997 (96,100.2081)	98.26104 (96,103.712)
Generalized Hill	96.65202 (96,98.78524)	96.88405 (96,99.77561)
Mixed Moment	98.17493 (96,102.9144)	103.3035 (96,111.2383)
PORT-Moment (q=0.01)	96.25696 (96,99.945)	98.05892 (96,103.382)

Finally, to close this semi-parametric analysis, we can calculate the same exceedance probability as obtained in the parametric context, measuring the quality of the sample maximum, the current record of 96 ml/kg/min. This probability is obtained with expression (3.54), where $H_{\hat{\gamma}}$ stands for the GPd defined in (3.21). As this exceedance probability depends on k and γ, we can construct a table with the results for each threshold k, replacing γ by the estimates of the four estimators used in the heuristic process. They can be found in Table 4.16 (see Appendix A.54 for **R** details).

Table 4.16: *Semi-parametric estimates for the exceedance probability of the actual record for the $\dot{V}O_2max$ data.*

Semi-parametric estimator of γ	$k = 19$	$k = 43$
Moment	0.008170709	0.01779646
Generalized Hill	0.01253879	0.006081975
Mixed Moment	0.0307783	0.0489384
PORT-Moment (q=0.01)	0.005422953	0.01613558

The semi-parametric estimates for the exceedance probability of 96 ml/kg/min provide several results, ranging between 0.5% and 1.78%. The exception rule is the estimates of the Mixed Moment estimator, which situate this probability around 3% or 5%. We must recall that this estimator gave the highest estimates for the right endpoint, as we can see in Table 4.13 and in Figure 4.27.

Before leaving this Section, it would be interesting to confront the semi-parametric results with the parametric ones obtained in Section 4.1.1. Comparing then Tables 4.12, 4.13 and 4.16 with Table 4.10, we can conclude that the semi-parametric approach leads to more concordant results with those obtained under the POT method, except for the estimates of the scale coefficient, which are lower in a semi-parametric context. The Block Maxima method offers larger differences, when compared to the POT method or the semi-parametric framework. We see some proximity between the ML right endpoint estimate of the Block Maxima method and the Mixed Moment estimate for $k = 43$, but generally speaking, the POT and semi-parametric approaches give more quality to the current sample maximum than the Block Maxima does, with a lower probability of surpassing the current record of 96 ml/kg/min.

4.2 The 100 metres in athletics revisited

The 100 metres race is one of the most popular and prestigious outdoor events in the sports of athletics. Who cannot remember the famous sprinter Carl Lewis during the eighties? Nowadays, Jamaica is in office, with Usain Bolt as the holder of the world record of 9.58 seconds in 2009. Progressively, the 100 metres has made its own mark, even supplanting the marathon as the gold event of the running universe. Because of its popularity, this event catches the attention of many fields, including Statistics, and since it consists in running a distance of 100 metres in the shortest time possible, it is of particular interest to EVT. Among articles from EVT about this subject, we can cite Einmahl and Magnus (2008), Einmahl and Smeets (2011) and Henriques-Rodrigues et al. (2011). In

this thesis, the 100 metres race is revisited in the light of the techniques presented in Sections 3.1 and 3.2. Therefore, we will follow the same analysis as for the $\dot{V}O_2max$ and the results obtained can be seen as an update and complement of those presented in the articles cited above, since we base our analysis on data until 2012.

As for the $\dot{V}O_2max$, we try to consider a population of athletes as homogeneous as possible, controlling the presence of potential "confounding variables". Therefore, in this Section, we will also consider only masculine runners. Concretely, we use a sample of 1184 masculine world athletes, considering their 100 metres running times, measured in seconds with two decimal places and recorded from 1991 to 2012. Despite data availability before 1991, we consider this year as the starting point for our collection. Indeed, according to Einmahl and Smeets (2011), the *International Association of Athletics Federations* (IAAF) started with modern doping control procedures in 1990 and, in order to avoid doping related times as much as possible, data before January 1, 1991 were excluded from the sample. Moreover, this organization recognizes officially a recorded time if the associated wind speed is less or equal to 2 *metres per second* (m/s). Consequently, running times with a wind speed of more than 2 m/s are not taken into account. The same rule applies to doubtful and questionable timings, which are not officially recognized. Finally, we only consider electronically recorded times, to keep homogeneity of the recording procedure. Thus, times clocked by hand or by other unofficial timing systems are also excluded from the sample. All the necessary data were obtained from three websites: `http://hem.bredband.net/athletics/atb-m01.htm`, `http://www.alltime-athletics.com/m_100ok.htm` and `http://www.iaaf.org/statistics/toplists/index.html`. For each of the 1184 athletes, we picked up the recorded running times of each year from 1991 to 2012, when available. If an athlete has several records at the same year, then we keep only the *best* time for this particular year, i.e., the *lowest* time from this year. Therefore, each athlete has only one record per year when data are available at this year. This avoids correlated data within each year for each athlete. Our sample consists then of one observation per year, from 1991 to 2012, for each of the 1184 athletes.

Provided that our data are organized by year, it may be of particular interest examining the eventual presence of a downward trend of the observed running times, revealing then a progressive improvement of the athletes' performance. In order to check the eventual time-dependence of the running times, we consider Figure 4.29.

Observing then the whole set of the annual Box-Plots, we see no evident decreasing (or eventually increasing) trend. Therefore, we can assume stationarity and proceed without time-series considerations.

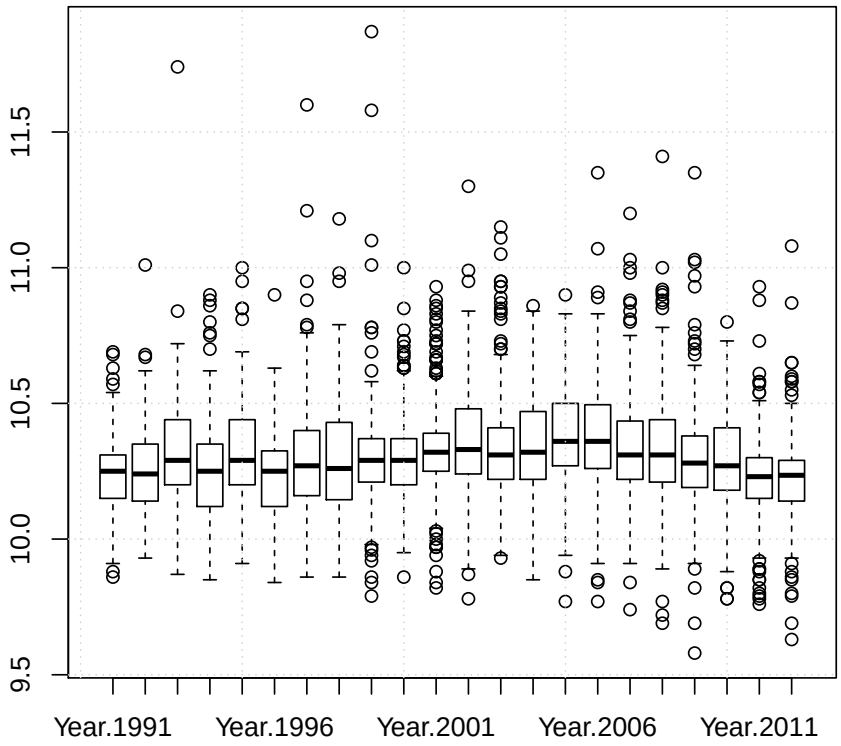

Figure 4.29: *Box-Plots of the 100 metres running times per year*

The subsequent analysis or the 100 metres running times will follow the same steps and methodologies exposed in the $\dot{V}O_2max$ case study, whereupon we refer to Section 4.1 as the guideline for all theoretical considerations, which will not be developed again in this second case study. The same rule applies to **R** scripts, which can be found in Appendix A, since they are exactly the same, changing only variables' names and some values. Using the $\dot{V}O_2max$ case as a guideline, we can easily adapt the $\dot{V}O_2max$ scripts to the 100 metres study.

4.2.1 Parametric data analysis

As for the $\dot{V}O_2max$, we will first follow the Gumbel's approach of Section 3.1.2, considering again, each of the 1184 athletes as a block. Then, for each athlete, we select the best time from all its annuals records from 1991 to 2012, independently of the year. We obtain thus 1184 blocks with one observation per block, i.e. 1184 observations. We recall that, as data are running times, the best time of each athlete represents the *lowest* of the observed values. In the $\dot{V}O_2max$ analysis, we mentioned that, in sports measures, it is very frequent to obtain repeated values for several athletes, due to the lack of precision of the measurements and discretization of data. For this reason, the obtained 1184 running times are smoothed using the same technique as for the $\dot{V}O_2max$. Hence, when for instance r athletes have the same recorded time of 10.15 seconds, these r results are smoothed equally over the interval $]10.145, 10.155[$ as follows:

$$time_j = 10.145 + 0.01\frac{2j-1}{2r}, \quad j = 1, \ldots, r.$$

However, in order to proceed, we must perform a last step. Indeed, as stated from the beginning, this thesis deals with inference based on sample maxima. Moreover, the Gumbel's approach requires a sample of m maxima. Since running times implies a sample minima analysis, we have to transpose our 1184 smoothed running times to a maxima context. This is done converting all the 1184 smoothed times in running speeds, so that a lower running time corresponds to a higher speed. This way, selecting the lowest time for an athlete is equivalent to selecting the *highest* speed. We choose to express the running speeds in metres per second (m/s), since the basic reference are 100 metres. The conversion is then done as follows:

$$speed_i = \frac{100}{smoothed\ time_i}, \quad i = 1, \ldots, 1184.$$

Defining then Y as the r.v. that represents the *maximum* running speed of an athlete who performs the 100 metres race, according to (3.1), we have a random sample

$(Y_1, \ldots, Y_m)$ of $m = 1184$ maxima at our disposal. As it was discussed in the first case study, the r.v. Y must be distinguished from the original r.v. X, which represents the running speed of a 100 metres athlete, with d.f. F. In a Gumbel's approach, we consider our data collection as a realization of the r.v. Y, which is assumed to follow a GEVd.

Concerning the POT approach, since we have a large sample, it may be reasonable to choose a fixed threshold u, above which we can fit a GPd. We will use the same sample as for Gumbel's approach, since the POT methodology is only valid if we are working with a random sample. Provided that each athlete have several measurements, surely inter-correlated, we keep the best result of each one, in order to respect the i.i.d. rule, obtaining then a random sample of 1184 observations. The philosophy behind this sample is somewhat different from Gumbel's approach. Under the POT methodology, the sample in hand is the realization of the original r.v. X with d.f. F. In particular, it is assumed that the data were taken from the right tail of F. Therefore, the random sample $(X_1, \ldots, X_n)$ is collection of top observations. Taking then a fixed threshold u, the r.v. Y represents the excesses above u, as discussed in Section 3.1.3.1, which is assumed to follow a GPd. Contrary to Gumbel's approach, which fits a GEVd to the whole sample, the POT methodology fits a GPd only to the portion of the sample above u. Since we do not have a reference value for an eventual threshold, the sample ME-plot presented in Section 3.1.3.7 can help us to fix a suitable threshold. Defining then u by $X_{n-k:n}, k = 1, \ldots, n-1$ and the empirical counterpart of the mean excess function, $\hat{e}_n$, as in (3.34), the sample ME-plot can be seen in Figure 4.30(cf. Appendix A.1).

Following Davison and Smith (1990) suggestion, already applied in the $\dot{V}O_2max$ case study, we notice two linear trends in the plot, apart from the high volatility characterizing the top sample. The two linear patterns are separated by a kink, visible at 9.7 m/s. As the objective of the POT methodology is to induce a cut-off in the sample above which the sample ME-plot follows a linear trend, we will take $u = 9.7$ as the required threshold for the POT analysis.

1) The Block Maxima approach

a) *Preliminary statistical analysis*

For a preliminary statistical analysis before proceeding to Gumbel's approach, we can fit again the Exponential model to our data in order to take a position about the weight of the right tail of the underlying d.f. F. This question can be assessed by the Exponential QQ-plot presented in Figure 4.31 (cf. Appendix A.3).

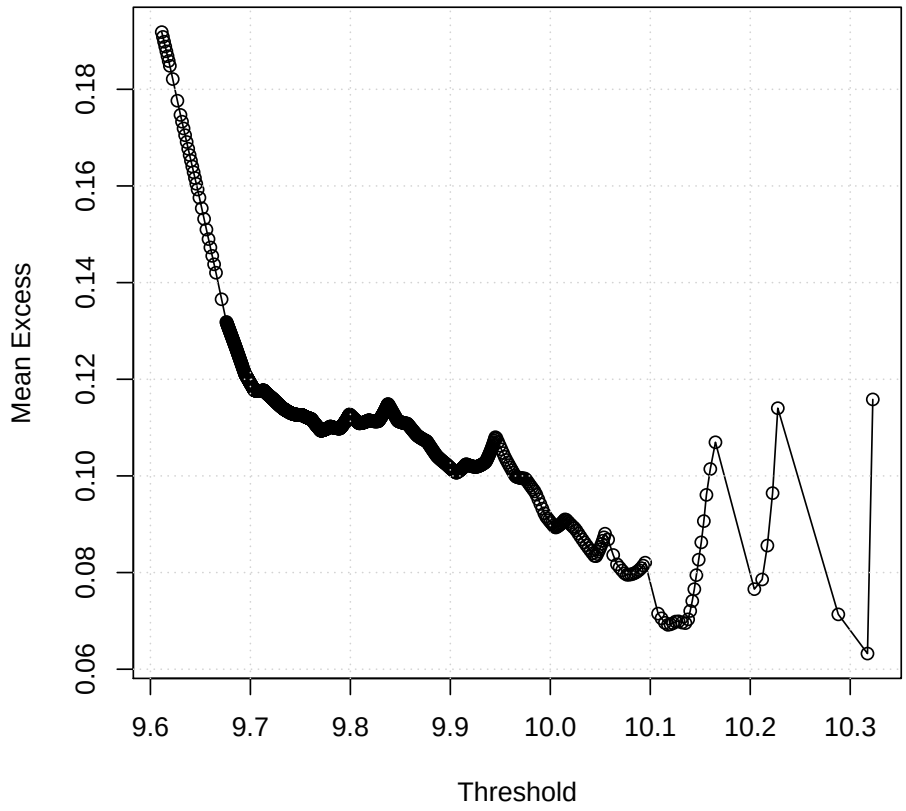

Figure 4.30: *Sample ME-plot for the 100 metres data*

This time, the plot exhibits a predominantly linear pattern, in contrast to the $\dot{V}O_2max$ case study, where a clear concave pattern was visible. However, we can argue that Figure 4.31 reveals a slight concave pattern for the 100 metres data and we cannot deny this position. Independently of the point of view chosen, we can draw the following conclusion: defining X as a r.v. that represents the running speed of an athlete who performs a 100 metres race (in contrast to Y, which represents the *maximum* running speed of a similar one), its d.f. F is exponentially right-tailed or is characterized by an almost exponential decay of its right tail. This suggests us that the d.f. of the running speed may belong to the Gumbel max-domain of attraction. Thus, the underlying d.f. of the $\dot{V}O_2max$ r.v. seems to have a lighter right tail than the d.f. of the running speed.

Provided that the Gumbel d.f. appears as a strongly potential limiting distribution for the running speed defined by the r.v. X, we can assess the goodness-of-fit of this distribution to our r.v. Y, using the classical Gumbel QQ-plot, as shown in Figure 4.32 (cf. Appendix A.4).

The Gumbel QQ-plot reveals a fairly reasonable linear pattern, with some asymmetry

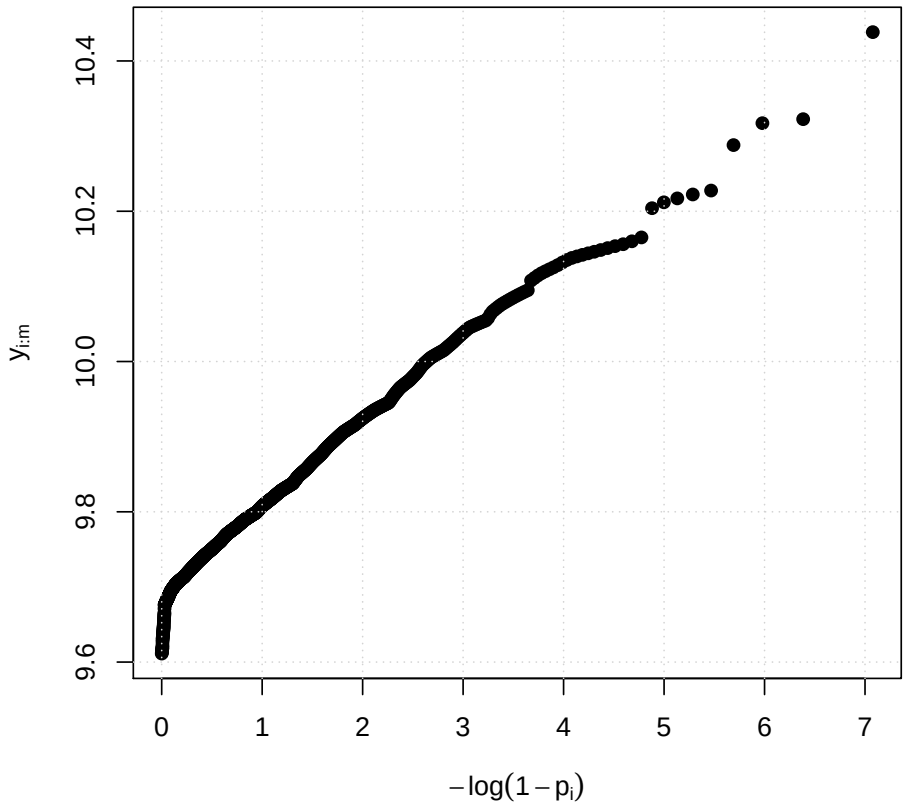

Figure 4.31: *Exponential QQ-plot for the 100 metres data*

located at lower and upper quantiles, pointing to the Gumbel model as a suitable candidate to fit our data. Once more, we can take advantage of the visible linear relation to obtain preliminary estimates of the parameters of the Gumbel distribution, λ and δ, fitting a least squares straight line to the points of Figure 4.32. The results are obtained with the **R** software (cf. Appendix A.5):

```
Call:
lm(formula = speeds ~ Qgt)

Coefficients:
(Intercept)          Qgt
   9.752466     0.087935
```

which yields

$$\hat{\lambda} = 9.752466 \quad \text{and} \quad \hat{\delta} = 0.087935, \tag{4.34}$$

as preliminary estimates of the parameters of Gumbel's distribution.

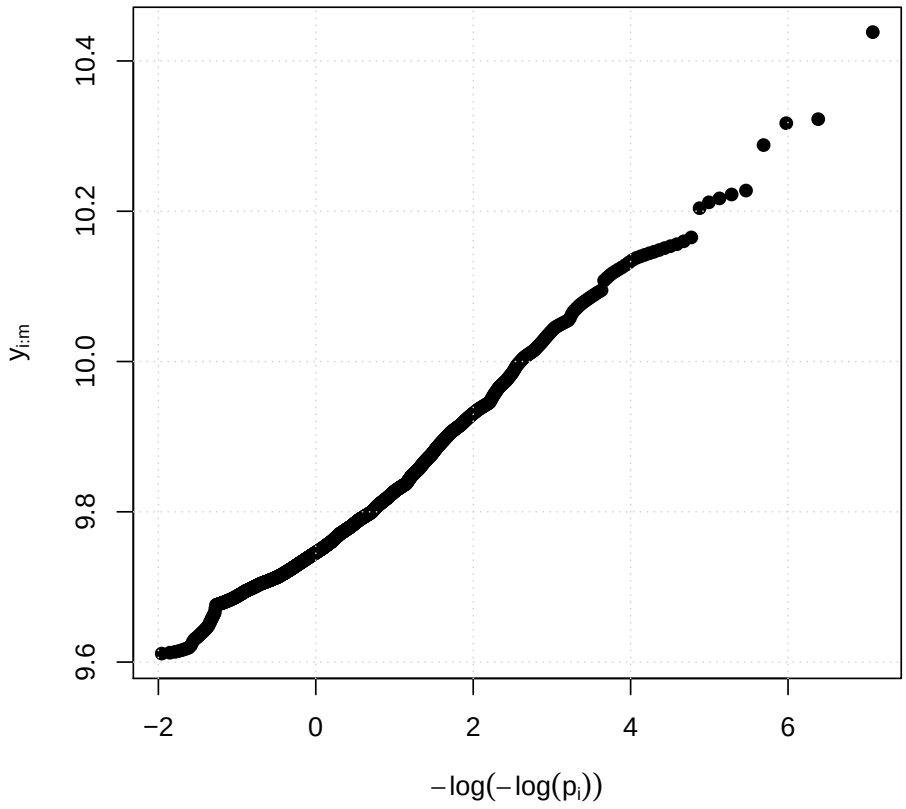

Figure 4.32: *Gumbel QQ-plot for the 100 metres data*

The least squares straight line can be added to Gumbel QQ-plot, leading us to Figure 4.33 (cf. Appendix A.5).

As a complement, we can check the eventual goodness-of-fit of the GEVd and compare it to the Exponential fit, since it is a natural candidate for extreme data. We use once again the QQ-plot, basic tool in this preliminary analysis. Inspired by the first case study, the expressions for the GEV model quantiles and for the standard GEV quantiles are given by (4.2) and (4.3):

$$Q_{\gamma,\lambda,\delta}(p) = \lambda + \delta\,\frac{(-\log p)^{-\gamma} - 1}{\gamma}, \quad 0 < p < 1$$

and

$$Q_{\gamma,0,1}(p) = \frac{(-\log p)^{-\gamma} - 1}{\gamma}, \quad 0 < p < 1,$$

resulting in the linear relationship

$$Q_{\gamma,\lambda,\delta}(p) = \lambda + \delta\, Q_{\gamma,0,1}(p).$$

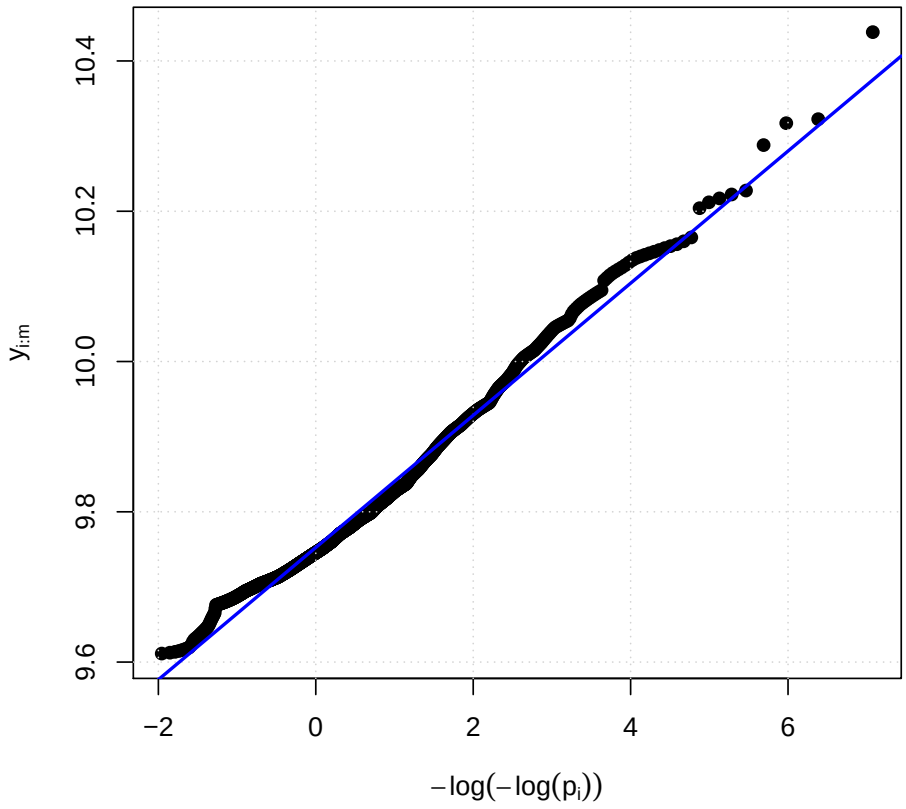

Figure 4.33: *Gumbel QQ-plot for the 100 metres data, with fitted straight line*

The obtention of the standard quantiles $Q_{\gamma,0,1}$ requires an estimate of the shape parameter. As usual, we choose γ so that the correlation between the quantiles $\hat{Q}_{\gamma,\lambda,\delta}(p) = Y_{i:m}$ and $Q_{\gamma,0,1}$ is maximized. The **R** software helps us to solve this optimization problem (cf. Appendix A.6):

```
$maximum
[1] 0.07126956

$objective
[1] 0.9963723
```

A first evidence stands out: we obtain a positive estimate for the EVI under the Block Maxima approach, $\hat{\gamma} = 0.07126956$. But since we are still in a preliminary analysis, we cannot draw any definitive conclusion about this unusual estimate. The obtention of the estimate can be confirmed graphically on Figure 4.34 (cf. Appendix A.6). We can now proceed to the GEVd QQ-plot, which can be seen on Figure 4.35 (cf. Appendix A.7).

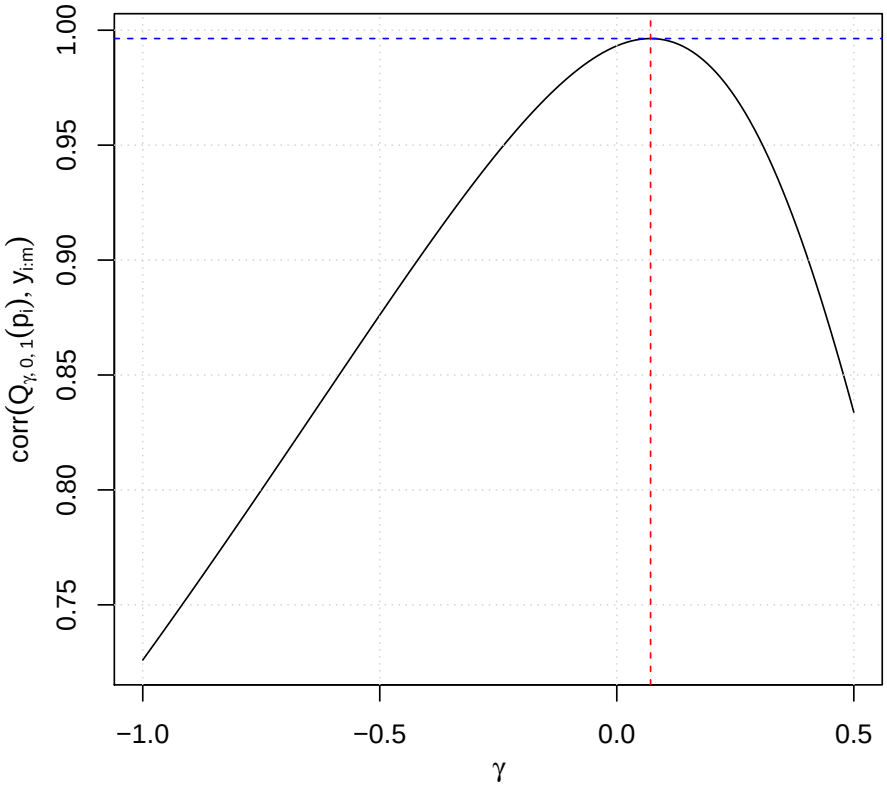

Figure 4.34: *Correlation plot between quantiles of the standard GEVd and the location-scale GEV family for the 100 metres data.*

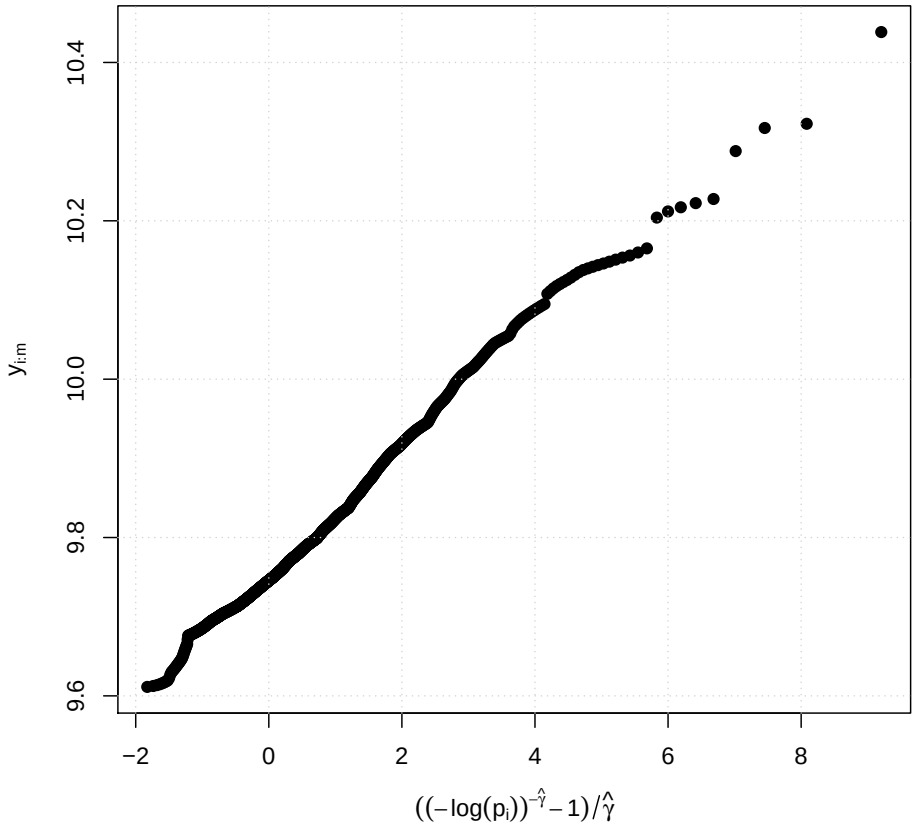

Figure 4.35: *GEVd QQ-plot for the 100 metres data.*

Compared to the Gumbel QQ-plot of Figure 4.32, we do not see major differences between the two fits. To increase the accuracy of the plot, we can add a least squares straight lines to the GEVd QQ-plot, which gives us, at the same time, preliminary estimates for λ and δ. For the line fitting, we use the **R** software which yields (cf. Appendix A.8)

```
Call:
lm(formula = speeds ~ Qgevt)

Coefficients:
(Intercept)        Qgevt
    9.75109      0.08003
```

We have then the following preliminary estimates for the GEVd parameters:

$$(\hat{\gamma}, \hat{\lambda}, \hat{\delta}) = (0.07126956, 9.75109, 0.08003) \tag{4.35}$$

which allows us to add the fitted line to the GEVd QQ-plot, available on Figure 4.36 (cf. Appendix A.8).

The choice between the Gumbel and the GEVd QQ-plots is not evident. We need more reliable and objective techniques to choose one of the parametric models. The statistical tests presented below have this mission and we hope they are able to discriminate what our eyes cannot see.

b) *Statistical choice of extreme value models*

The preliminary statistical analysis suggests the Gumbel distribution as a suitable parametric model to be fitted to our 100 metres running speeds. On the same way, the analysis eventually suggests that a GEVd may be appropriate as well, as a parametric model for the r.v. Y. Again, we have the same battle between the Gumbel model and the GEVd and the winner must be chosen with more consistent statistical tools. As for the previous case study, we will perform the same statistical tests with the same statistics. Therefore, the hypotheses at play are

$$H_0 : \gamma = 0 \quad vs \quad H_1 : \gamma \neq 0, \tag{4.36}$$

for the two-sided version of the test and

$$H_0 : \gamma = 0 \quad vs \quad H_1 : \gamma < 0, \tag{4.37}$$

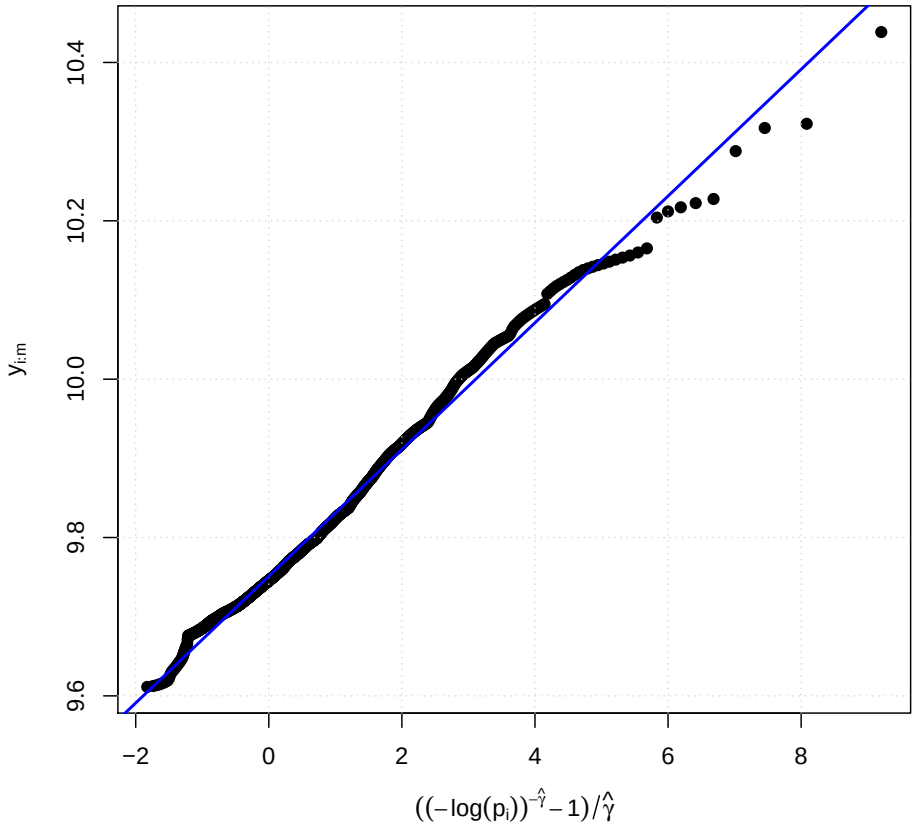

Figure 4.36: *GEVd QQ-plot for the 100 metres data,with fitted line.*

for the left one-sided one. Notice that we prefer a left one-sided version of the test since it is usual in athletics to have $\gamma < 0$. But we keep in mind the suspicious preliminary positive estimate for the EVI in (4.35). Putting it in our pocket, we proceed to the statistical tests.

The firstly discussed test was based on the standardized Gumbel statistic defined by (4.7), used for testing (4.37). Applying the **R** software to the 100 metres data, we obtain the following results (cf. Appendix A.9):

```
[1] gs_m= 4.056264   gs*_m= 2.285715   p-value= 0.9032993
```

At the asymptotic size $\alpha = 0.05$, the null hypothesis in (4.37) is not rejected and the Gumbel model is strongly recommended by this test as a suitable parametric model for the r.v. Y.

As a second statistical test, we saw the adjusted LRT defined by the statistic in (4.9), developed for the test (4.36). But as seen in (4.8), in order to perform this test, we need the ML estimates of the parameters from the Gumbel and GEV models. The **R** software allows us to obtain easily the ML estimates, necessary for the computation of the statistic (cf. Appendix A.10):

```
[1] Gumbel ML estimates
 lambda= 9.754742   delta= 0.07891165
[2] GEV ML estimates
 gamma= 0.1064495  lambda= 9.750348   delta= 0.07570596
```

The final ML estimates for the Gumbel and GEV models are then

$$(\hat{\lambda}_{G_0}, \hat{\delta}_{G_0}) = (9.754742, 0.07891165) \tag{4.38}$$

and

$$(\hat{\gamma}_{G_\gamma}, \hat{\lambda}_{G_\gamma}, \hat{\delta}_{G_\gamma}) = (0.1064495, 9.750348, 0.07570596), \tag{4.39}$$

which can now be used for the adjusted LRT using the **R** software (cf. Appendix A.11):

```
[1] l= 29.11658   l*= 29.04788   p-value= 7.061122e-08
```

With these results in hand, at the asymptotic size $\alpha = 0.05$, we find a very strong evidence for the rejection of H_0 and, hence, the Gumbel model is strongly rejected, favouring then a GEVd. Once more, note that the preliminary estimates for the Gumbel model obtained in (4.34) are very close to the final ML estimates for the same model given by

(4.38). However, we obtain a surprising result for the EVI with the ML estimates of the GEV model. Indeed, as we can see in (4.39), the estimate for γ is close to zero, but positive. Before extracting some conclusion, we must complement the ML estimation with other methods, like the PWM method, and construct some CI's for the GEVd parameters. But before this, we can perform the remaining statistical tests performed in the $\dot{V}O_2max$ analysis.

Rao's score test, defined by the statistic in (4.12) or (4.13), can now be applied, for testing again (4.37) or (4.36), respectively. Using the final ML estimates for the Gumbel model, $(\hat{\lambda}_{G_0}, \hat{\delta}_{G_0}) = (9.754742, 0.07891165)$, needed for the statistic to be computed, the **R** software provides us the following results (cf. Appendix A.12):

```
[1] Normal Test:  v_m= 290.7215   v_m*= 5.833129   p-value= 1

[2] Chi-square Test:  v^2_m= 84518.96   v^2_m*= 34.0254
    p-value= 5.439744e-09
```

Once more, we obtain surprising results. Considering the Normal version of Rao's score test, at the asymptotic level $\alpha = 0.05$, the null hypothesis of Gumbel model is not rejected, but, looking at the Chi-square version of the test, at the same asymptotic level, the Gumbel model is strongly rejected in favor of the GEVd. Recalling the results of the previous statistical tests, a surprising conclusion begins to appear: testing a Gumbel model versus a GEVd with $\gamma < 0$ leads always to a non-rejection of the Gumbel hypothesis. But, when the Gumbel model is tested against a generic GEVd with $\gamma \neq 0$, it is strongly rejected in favor of the GEVd. Therefore, the previous statistical tests suggest that a GEVd with $\gamma > 0$ is appropriate to the maxima running speeds data. Considering the results of the ML estimation for the GEVd in (4.39), we testify a positive estimate for γ. Note that the preliminary statistical analysis did not point to such a distribution as a suitable one for the r.v. Y.

Proceeding with the subsequent tests of the first case study, we can now consider the LAN test of Marohn (2000), used again to test (4.36) or (4.37), with the statistic in (4.14) (cf. Appendix A.13):

```
[1] Unilateral test:  t_m= 4.027118   t_m*= 5.833021   p-value= 1

[2] Bilateral test:  |t_m|= 4.027118   |t_m*|= 5.833021
    p-value= 5.443276e-09
```

The LAN test leads us to the same conclusion as the previous tests: the GEVd with $\gamma > 0$ is the most suitable model for the r.v. Y. As the shape parameter γ governs the heaviness of the underlying d.f. for the r.v. X, which represents the running speeds of an athlete performing a 100 metres race, a positive estimated value indicates possibly a heavy right tail.

Finally, we can close the statistical tests with the three goodness-of fit tests presented in (4.15), (4.16) and (4.17), used to check the suitability of the Gumbel model. Using again the ML estimates of the Gumbel model, $(\hat{\lambda}_{G_0}, \hat{\delta}_{G_0}) = (9.754742, 0.07891165)$ and the **R** software, we get (cf. Appendix A.14):

```
Kolmogorov-Smirnov statistic:  0.04724083
Cramer-von Mises statistic:  1.048995
Anderson-Darling statistic:  6.943542
```

The decision is ruled by the simulated critical points of the sampling distribution of each test statistic, presented in Table 4.1, for the Kolmogorov-Smirnov statistic, and in Table 4.2, for the Cramér-von Mises and Anderson-Darling statistics. The use of the Tables requires the computation of the modified statistics:

1. Modified Kolmogorov-Smirnov statistic: $\sqrt{m}\, d_m = \sqrt{1184} \times 0.04744457 \simeq 1.63$
2. Modified Cramér-von Mises statistic: $w_m^2(1+0.2/\sqrt{m}) = 1.056051 \times (1+0.2/\sqrt{1184})$ $\simeq 1.0622$
3. Modified Anderson-Darling statistic: $a_m^2(1+0.2/\sqrt{m}) = 6.963871 \times (1+0.2/\sqrt{1184})$ $\simeq 7.004$

Consulting the respective Tables at the asymptotic size of $\alpha = 0.05$, the null hypothesis of Gumbel model as a suitable model for the data is always rejected, since the modified statistics always exceed the tabled critical points. Therefore, even the conservative goodness-of-fit tests reject the Gumbel model as a suitable parametric model. We can summarize the results of all previous tests in Table 4.17.

As stated before, all the tests do not reject the Gumbel model when confronted with a GEVd with $\gamma < 0$, but reject it strongly, when tested against a generic GEVd with $\gamma \neq 0$. These results are corroborated by the three goodness-of-fit tests, which rejected the Gumbel model. We may then conclude that the tests select a GEVd with $\gamma > 0$ as an appropriate parametric model for Y. Note that the conclusions drawn from the statistical tests do not support the results extracted from the preliminary statistical, since the latter

Table 4.17: *Results for the statistical choice of extreme value models for the 100 metres data.*

Test	Hypotheses	Observed statistic	p-value	Decision ($\alpha = 0.05$)
Gumbel statistic	$H_0 : \gamma = 0$ vs $H_1 : \gamma < 0$	$gs_m^* = 2.285715$	0.9032993	not reject H_0
LRT	$H_0 : \gamma = 0$ vs $H_1 : \gamma \neq 0$	$\mathbf{l}^* = 29.04788$	7.061122e-08	reject H_0
Rao's score test	$H_0 : \gamma = 0$ vs $H_1 : \gamma \neq 0$	$v_m^{*2} = 34.0254$	5.439744e-09	reject H_0
LAN test	$H_0 : \gamma = 0$ vs $H_1 : \gamma < 0$	$t_m^* = 5.833021$	1	not reject H_0
LAN test	$H_0 : \gamma = 0$ vs $H_1 : \gamma \neq 0$	$\lvert t_m^* \rvert = 5.833021$	5.443276e-09	reject H_0

pointed to an exponential right-tailed underlying d.f. or eventually a slightly light-tailed one.

The Gumbel statistic presented in (4.7) can also be used to test a Gumbel model against a GEVd with $\gamma > 0$. Fore more details, we refer to Tiago de Oliveira and Gomes (1984). Therefore, in order to test

$$H_0 : \gamma = 0 \quad vs \quad H_1 : \gamma > 0,$$

at the asymptotic level of $\alpha = 0.05$, we reject H_0 if

$$GS_m^* = \frac{GS_m - \beta_m}{\alpha_m} \geq \mathcal{G}_{1-\alpha},$$

where $\mathcal{G}_\epsilon$ stands for the standard Gumbel ϵ-quantile.

We can also obtain the corresponding p-value as follows:

$$p(GS_m^*) = 1 - \Lambda(GS_m^*).$$

The **R** software produces the following results:

```
[1] gs_m= 4.056264   gs*_m= 2.285715   p-value= 0.09670072
```

At the asymptotic size $\alpha = 0.05$, the null hypothesis is not rejected and, once again, the Gumbel statistic selects the Gumbel model as suitable for the r.v. Y. Therefore, this test does not select a GEVd with $\gamma > 0$ as a suitable model, contrary to the other performed tests.

c) *Parametric estimation of extreme events*

The statistical tests from last section elect a GEVd with $\gamma > 0$ as a parametric model for the r.v. Y, the maximum running speed of a 100 metres athlete. We are then in the context of heavy right-tailed distributions, underlying our population of athletes.

With a parametric model in hand, we can now estimate the respective parameters by the estimation methods of Section 3.1.2. The ML estimates for the GEVd parameters, $(\hat{\gamma}_{G_\gamma}, \hat{\lambda}_{G_\gamma}, \hat{\delta}_{G_\gamma})$, were already obtained when the LRT was performed in last paragraph. The results are visible in (4.39) and we verified a positive estimate for the EVI. We can use the **R** software which offers some graphical diagnosis tools to check the adequacy of the ML fit, presented in Figure 4.37 (cf. Appendix A.20). The plots evince a satisfactory

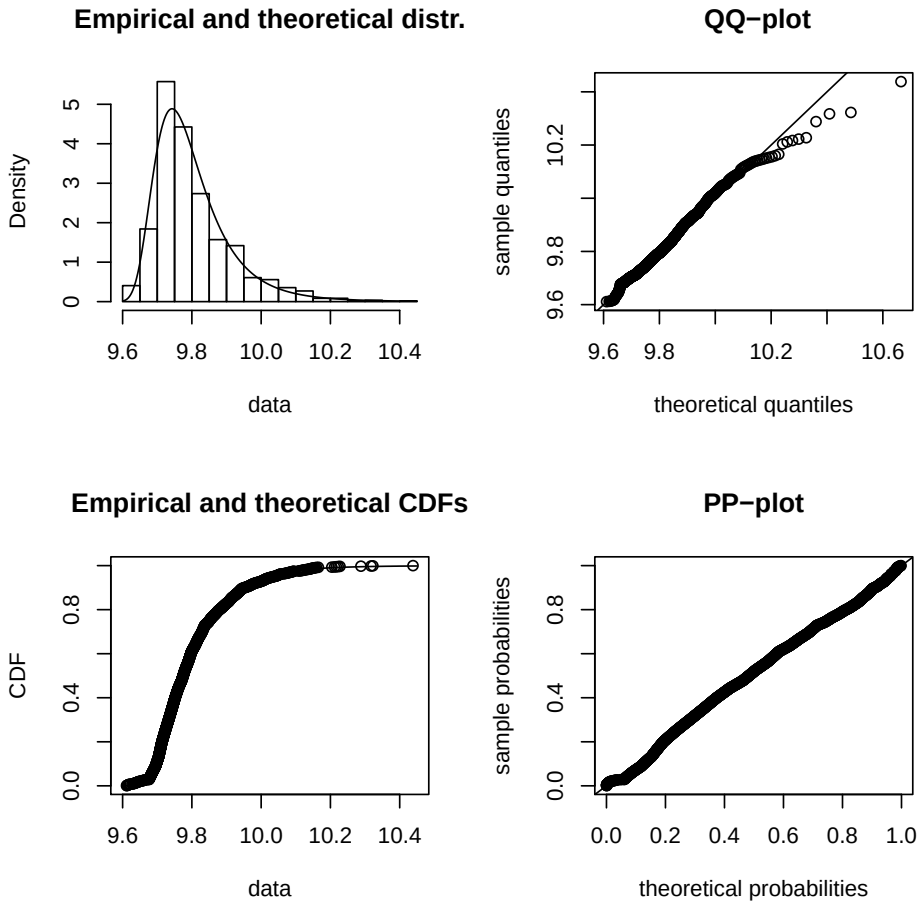

Figure 4.37: *Graphical diagnosis of the GEVd fit for the 100 metres data.*

fit for the GEVd with $\gamma > 0$ except, perhaps, for the QQ-plot, where some asymmetry is present on the top of the sample, because of the large variability characterizing this area.

Pursuing now a PWM estimation for the GEVd parameters, the **R** software provides us the following estimation results (cf. Appendix A.19):

```
[1] GEV PWM estimates
  gamma= 0.1398049  lambda= 9.748983   delta= 0.07337964
```

The estimation results are very similar to the ML estimates obtained in (4.39) and

we denote again a positive estimate for the EVI. The estimates of the two methods are summarized in Table 4.18.

Table 4.18: *ML and PWM estimates for the parameters of the GEVd, for the 100 metres data.*

Estimation method	$\hat{\gamma}$ (shape)	$\hat{\lambda}$ (location)	$\hat{\delta}$ (scale)
ML	0.1064495	9.750348	0.07570596
PWM	0.1398049	9.748983	0.07337964

Remember that these estimates can be taken as estimates for the attraction coefficients of Section 2.4, with $\hat{b}_n = \hat{\lambda}$ and $\hat{a}_n = \hat{\delta}$.

To complement the point estimates of the parameters, we can construct CI's for the GEVd parameters based on the profile likelihood method described in Section 3.1.2.4. Once more, the **R** software can be helpful (cf. Appendix A.21).

```
[1] "profiling loc"
[1] "profiling scale"
[1] "profiling shape"
           lower     upper
loc   9.74557247 9.7551860
scale 0.07221670 0.0794508
shape 0.06647845 0.1480332
```

Taking a look at the CI for the shape parameter γ, we notice that the zero value is excluded from the interval, thus pulling away the Gumbel model as a suitable parametric model, as already seen. The results can be visualized in Figure 4.38).

The last estimation that can be performed in a GEVd context with $\gamma > 0$ is the exceedance probability of an appropriate extreme quantile. Inspired by the previous case study, we can check the quality of the current record of the Jamaican Usain Bolt, computing the probability of exceeding the current sample maximum. The lowest running time of our sample is 9.58 seconds, which corresponds to the maximum running speed of approximately 10.438 m/s. We can then calculate $P(Y > 10.438)$. Provided that $Y \frown G_\gamma(\lambda, \delta)$, we can obtain the desired probability using the GEVd, replacing the parameters by its ML or PWM estimates (cf. Appendix A.22):

```
[1] Maximum Likelihood: P(Y>10.438)= 0.001736962
[2] Probability Weighted Moments: P(Y>10.438)= 0.002482753
```

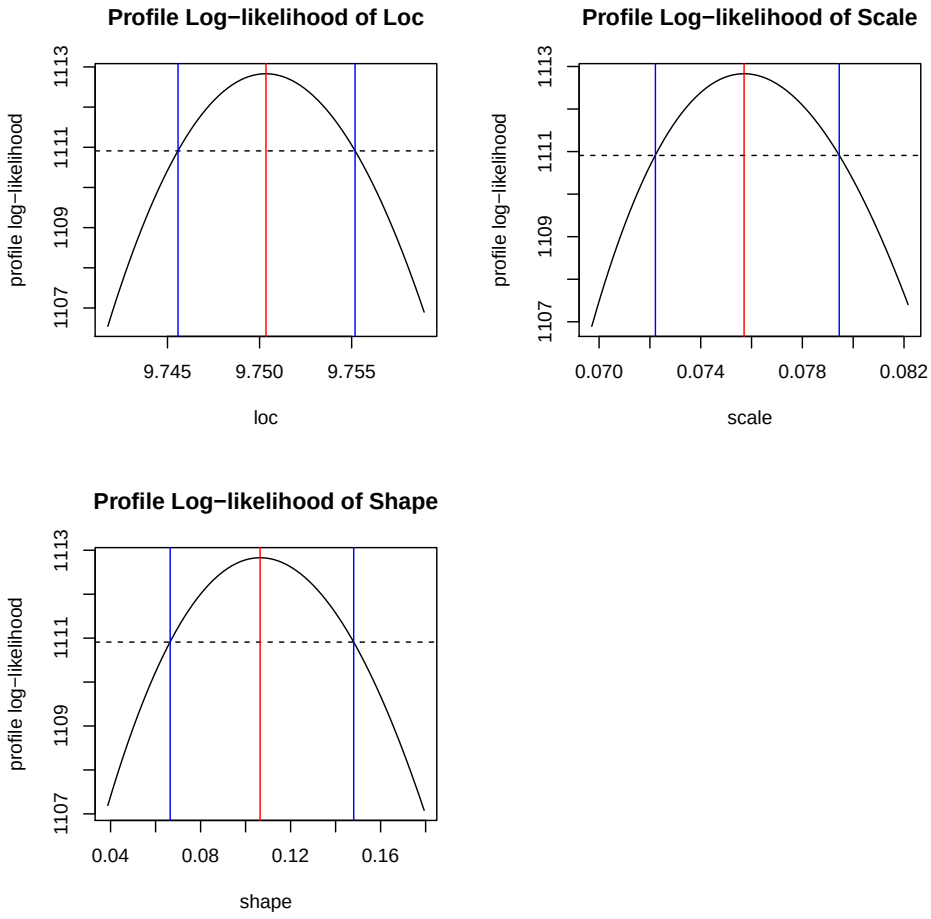

Figure 4.38: *Profile likelihood-based 95% confidence intervals for GEVd parameters for the 100m data.*

Then, following a Gumbel's parametric approach, the current record of the runner Usain Bolt can be surpassed with approximately 0.2% of probability.

Since we are in a context of $\gamma > 0$, the underlying d.f. F is unbounded on the right, with an infinite right endpoint, giving to an athlete the possibility of increase infinitely his running speed. Unless we are in a science-fiction world, this conclusion is unrealistic.

For a matter of comparison, we can examine the results provided by the Gumbel model, in spite of the decisions taken from the statistical tests. The ML estimates for (λ, δ) were already computed since they were necessary for the LRT. We can find them in (4.38). The PWM estimates can also be easily computed, providing the following results (see Appendix A.16):

```
[1] Gumbel PWM estimates
 lambda= 9.76066   delta= 0.08498496
```

The respective CI's can be also obtained using the profile likelihood approach, yielding the following results, together with Figure 4.39 (cf. Appendix A.17).

```
[1] "profiling loc"
[1] "profiling scale"
           lower      upper
loc   9.75004910 9.75946783
scale 0.07542365 0.08266502
```

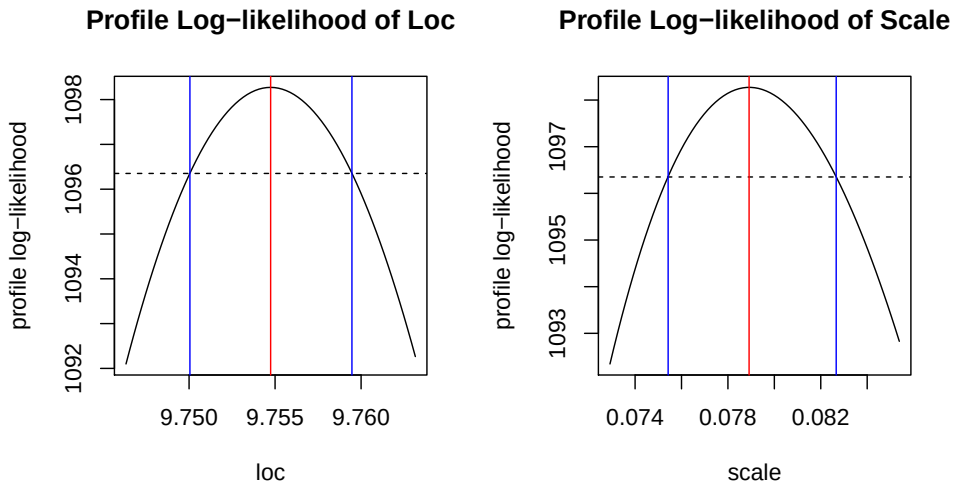

Figure 4.39: *Profile likelihood-based 95% confidence intervals for Gumbel parameters for the 100m data.*

Finally, the exceedance probability of the current world record of Usain Bolt can be estimated using the Gumbel model $G_0(\lambda, \delta)$ and replacing λ and δ by their respective ML or PWM estimates. With the help of our habitual **R** software (cf. Appendix A.18), we obtain:

```
[1] Maximum Likelihood: P(Y>10.438)= 0.0001736257
[2] Probability Weighted Moments: P(Y>10.438)= 0.0003455796
```

To ease the comparison of the results between the GEV and Gumbel models, a table like Table 4.19 would be appropriate.

Table 4.19: *Estimation results for Gumbel and GEV distributions for the 100 metres data.*

		Gumbel model $\gamma = 0$	GEV family $\gamma > 0$
$\hat{\gamma}$	ML	-	0.1064495
	PWM	-	0.1398049
	CI (95%)	-	(0.06647845,0.1480332)
$\hat{\lambda}$	ML	9.754742	9.750348
	PWM	9.76066	9.748983
	CI (95%)	(9.7500491,9.75946783)	(9.74557247,9.755186)
$\hat{\delta}$	ML	0.07891165	0.07570596
	PWM	0.08498496	0.07337964
	CI (95%)	(0.07542365,0.08266502)	(0.0722167,0.0794508)
$P(Y > 10.438)$	ML	0.0001736257	0.001736962
	PWM	0.00034558	0.00248275

As we can see, the estimates between Gumbel and GEVd models are very similar, except perhaps for the exceedance probabilities, where some critics may argue that a probability of 0.17% is significantly different from a probability of 0.017%, speaking of ML estimates. Anyway, since the conclusion of an infinite right endpoint for the underlying d.f. F may seem unsatisfactory in a human sports universe, we can choose the Gumbel model as a suitable parametric model for the 100 metres data, due to its flexibility and possibility of a finite right endpoint for the underlying d.f. Recall that the Gumbel statistic always selects the Gumbel model, whatever one-sided test is performed. We can take a look to the **R** diagnosis of Figure 4.40, applied to the Gumbel fit, in order to check the adequacy of such a distribution (cf. Appendix A.15).

Unfortunately, comparing this diagnosis with Figure 4.37, we notice that the Gumbel fit seems worse than the GEVd fit, a fact accused specially by the QQ-plot, where the empirical and theoretical quantiles correspondences seem better in the GEVd case. Thus,

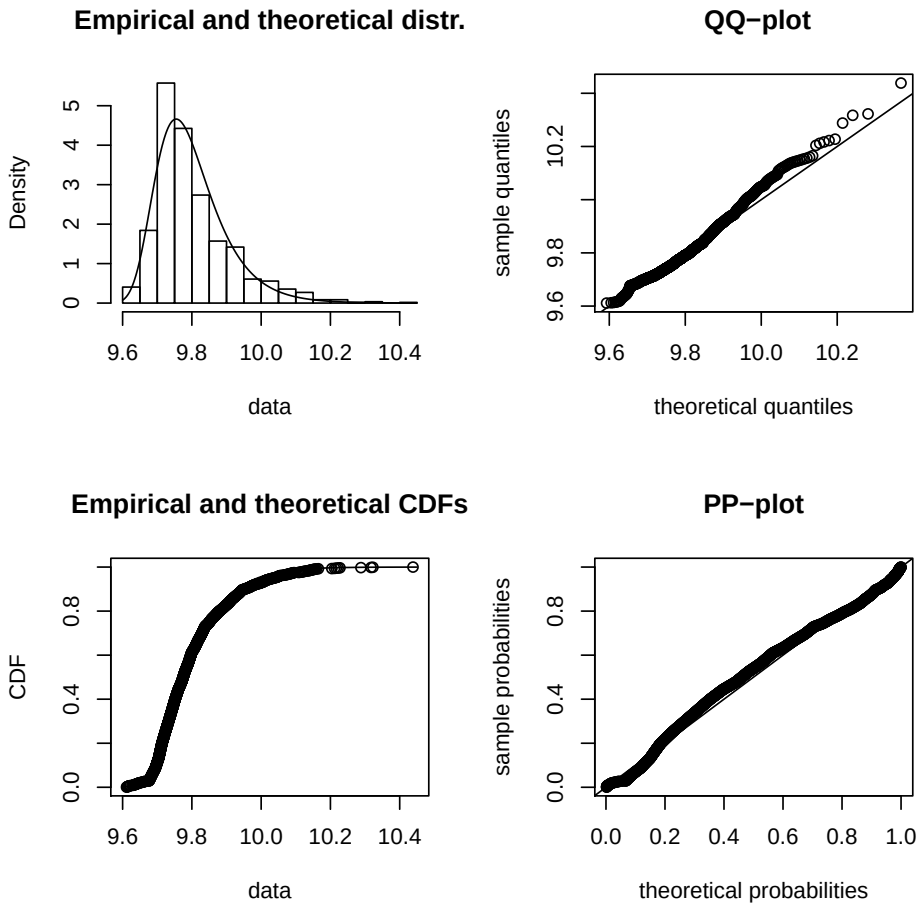

Figure 4.40: *Graphical diagnosis of the Gumbel fit for the 100 metres data*

adopting the Gumbel model as a parametric model for Y must be done with some care, keeping in mind that better results are obtained with a GEVd family with $\gamma > 0$. The Gumbel's Block Maxima approach leads us then to the suitability of a GEVd with $\gamma > 0$, or eventually, to a Gumbel parametric model, although a poorer fit is achieved in this latter case. As it makes more sense to have a finite right endpoint for the underlying d.f. F, we can look what other approaches have to tell us, namely the POT approach and the semi-parametric inference. Hence, before concluding for the superhuman capacities of our 100 metres runners, we have to wait for the results of the aforementioned approaches.

2) The POT method

a) *Preliminary statistical analysis*

Based on the ME-plot of Figure 4.30, we chose $u = 9.7$ as a suitable threshold above which we can fit a GPd. As for the $\dot{V}O_2max$ study, we begin this Section ascertaining the goodness-of fit of the particular case of the GPd, i.e. the Exponential distribution. This way, we can state an initial conjecture about an eventual exponential decay of the right tail of the underlying d.f. F by means of the QQ-plot. The plot is visible on Figure 4.41, where the Exponential distribution is fitted to the $m = 1051$ excesses above $u = 9.7$, represented by the r.v. Y (cf. Appendix A.24).

The Exponential QQ-plot exhibits a roughly linear pattern, apart from the top-sample, characterized by the usual volatility . This preliminary study contrasts with the POT approach of the $\dot{V}O_2max$, where Figure 4.13 revealed a concave pattern, questioning the suitability of the Exponential model. Using then (3.24), the theoretical quantiles of the Exponential model are given by

$$Q_{\sigma_u} = -\sigma_u \log(1-p), \quad 0 < p < 1$$

and

$$Q_1 = -\log(1-p), \quad 0 < p < 1,$$

yielding the linear relationship

$$Q_{\sigma_u} = \sigma_u Q_1.$$

Recall that, as we are working with excesses, the straight line which characterizes the linear relationship between the quantiles has no intercept.

We can then fit a least squares straight line with no intercept to the points of Figure 4.41, in order to obtain a rough estimate of the scale parameter σ_u of the Exponential distribution. The **R** software yields the following results:

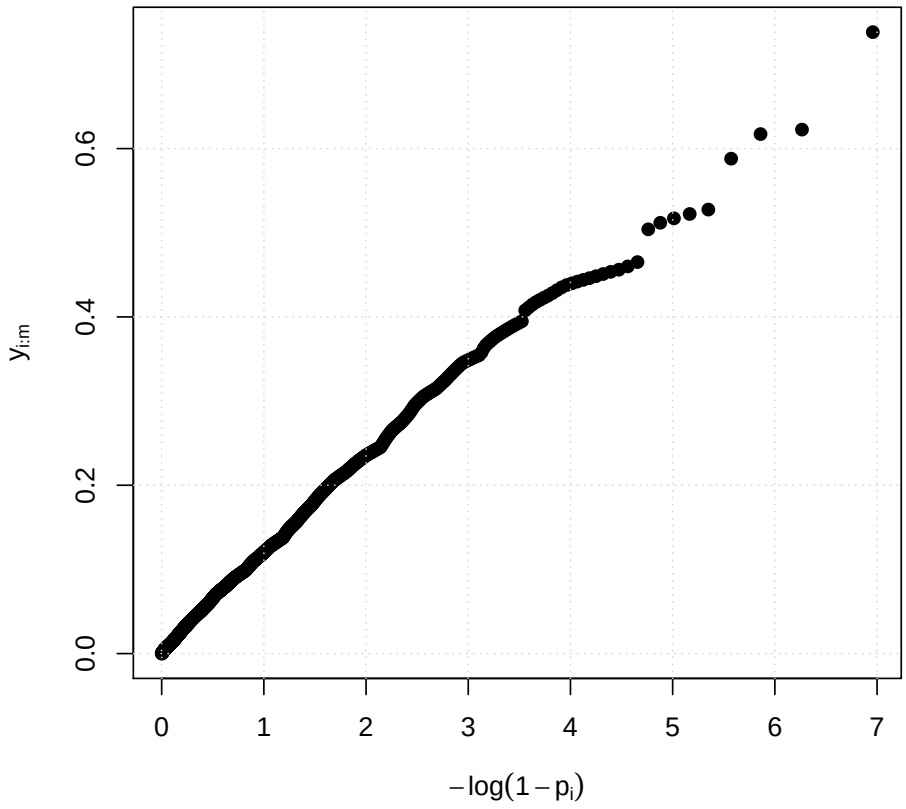

Figure 4.41: *Exponential QQ-plot for the* $m = 1051$ *excesses of the 100 metres data*

```
Call:
lm(formula = excesst ~ Qet - 1)

Coefficients:
   Qet
0.1149
```

We have then a preliminary estimate of the scale parameter of the Exponential model:

$$\hat{\sigma}_u = 0.1149 \tag{4.40}$$

and the fitted straight line can be added to the Exponential QQ-plot, resulting in Figure 4.42.

Note that the empirical quantiles grow at the same time as the theoretical Exponential quantiles, except for the top-portion of the sample, where the linear relationship between the two quantiles ceases to exist. This can be a clue about some right-asymmetry of

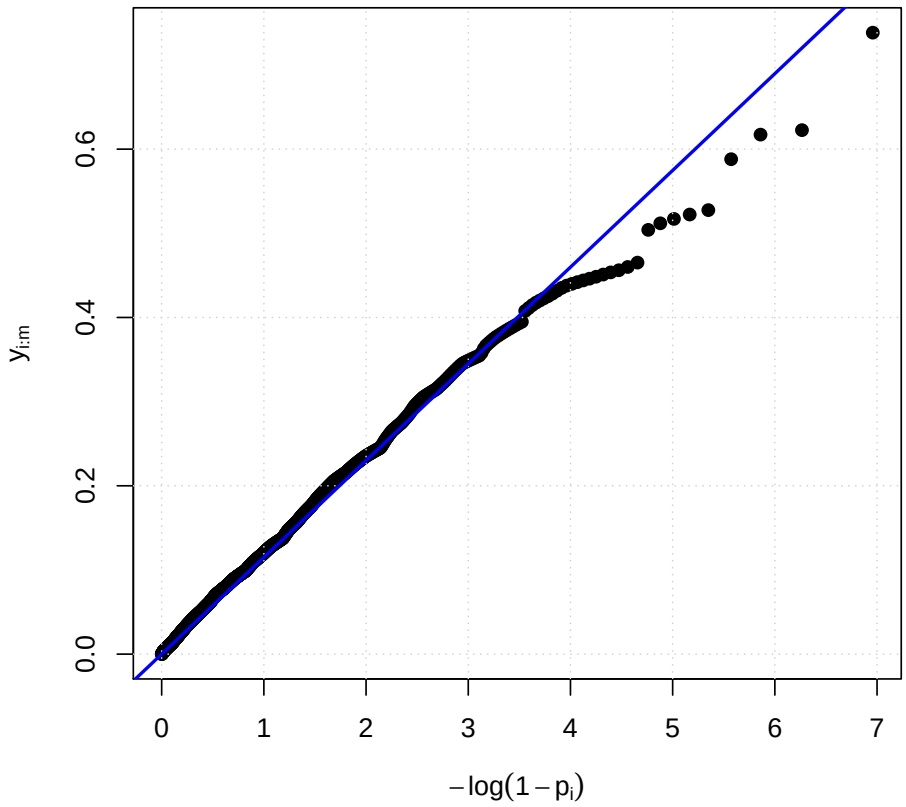

Figure 4.42: *Exponential QQ-plot with fitted line for the* $m = 1051$ *excesses in the 100 metres data*

F, when compared with the Exponential distribution. Anyway, we must not forget that the last top portion of the sample is characterized by a high volatility, as stressed in the ME-plot of Figure 4.30.

To complete this preliminary analysis, we can also fit a GPd with $\gamma \neq 0$ to the sample excesses, by means of a QQ-plot, since, until now, we do not have any reason to reject the GPd. The needed quantiles of the GPd are available in (4.18) and (4.19):

$$Q_{\gamma,\sigma_u}(p) = H_\gamma^{-1}(p|0,\sigma_u) = \sigma_u \frac{(1-p)^{-\gamma}-1}{\gamma}, \quad 0 < p < 1$$

and

$$Q_{\gamma,1}(p) = \frac{(1-p)^{-\gamma}-1}{\gamma}, \quad 0 < p < 1,$$

yielding the linear relationship

$$Q_{\gamma,\sigma_u}(p) = \sigma_u \, Q_{\gamma,1}(p).$$

As noted in the preliminary study of the $\dot{V}O_2max$, we have to specify a value for γ before the construction of the QQ-plot, since the standard quantiles of the GPd depend on this parameter. Following once again the procedure of Beirlant et al. (2004), we seek the value of γ which maximizes que correlation between $\hat{Q}_{\gamma,\sigma_u}(p) = Y_{i:m}$ and $Q_{\gamma,1}(p)$. The **R** software solves this optimization problem easily (cf. Appendix A.25):

```
$maximum
[1] -0.07146666

$objective
[1] 0.999302
```

which can be visualized graphically in Figure 4.43. With $\hat{\gamma} = -0.07146666$, the corresponding QQ-plot is shown in Figure 4.44 (cf. Appendix A.26).

We have here the first important difference between the two parametric approaches: contrary to the Block Maxima approach, we obtain a negative estimate for the EVI, which makes much more sense in the sports context. The estimate is negative and very close to zero, questioning the eventual suitability of the Exponential distribution, where $\gamma = 0$.

As for the Exponential fit, the GPd QQ-plot exhibits a linear pattern, which seems more regular than the Exponential QQ-plot of Figure 4.42. To clean this doubt, we can fit a least squares straight line to the points of Figure 4.44. Recall once more that the fitted line must have no intercept. The **R** software produces the following results (cf. Appendix A.27):

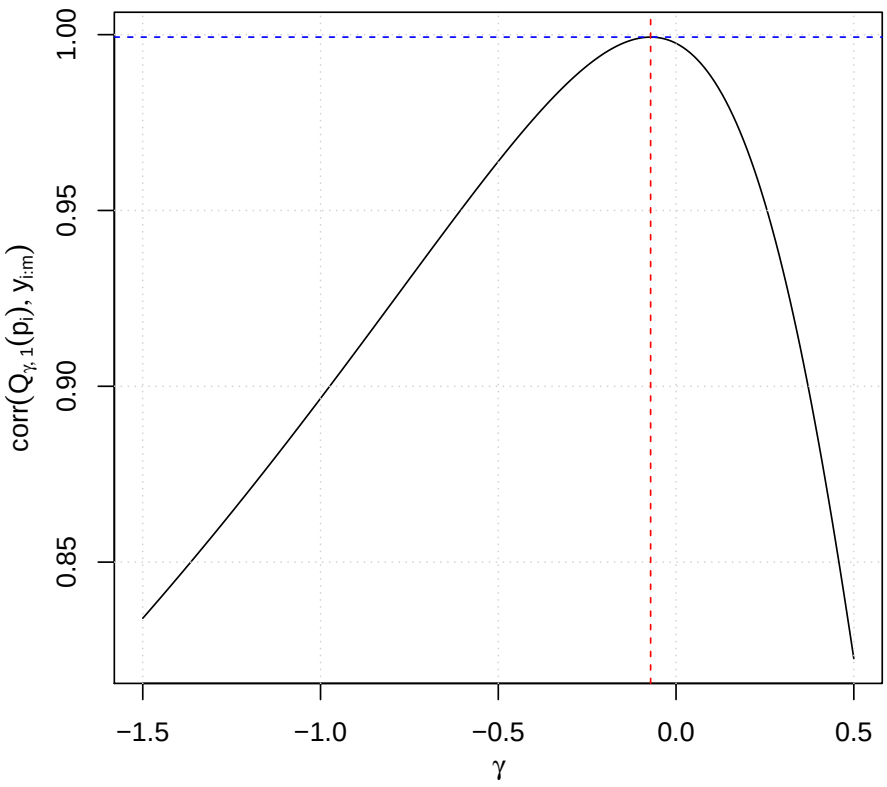

Figure 4.43: *Correlation plot between quantiles of the standard GPd and of the GP family for the 100 metres data.*

```
Call:
lm(formula = excesst ~ Qgpdt - 1)

Coefficients:
Qgpdt
0.127
```

Adding the fitted line to the GPd QQ-plot, we obtain the plot on Figure 4.45.

Looking at the GPd QQ-plot with fitted line, we confirm our guess: the GPd provides a more linear pattern than the Exponential distribution does, with less irregularity at the top of the sample. Therefore, the GPd with $\gamma < 0$ seems to be a strong candidate as a parametric model to be fitted to the sample excesses. As a preliminary analysis is always subjective, we check the suitability of the GPd model in the next Section, using objective

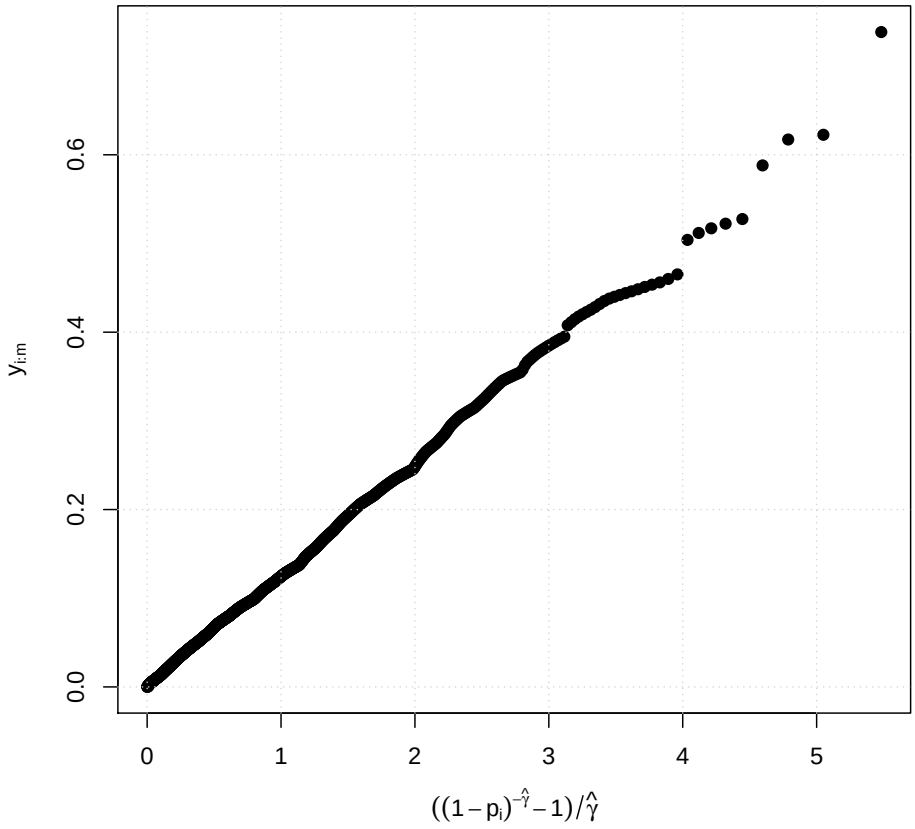

Figure 4.44: *GPd QQ-plot for the 100 metres data.*

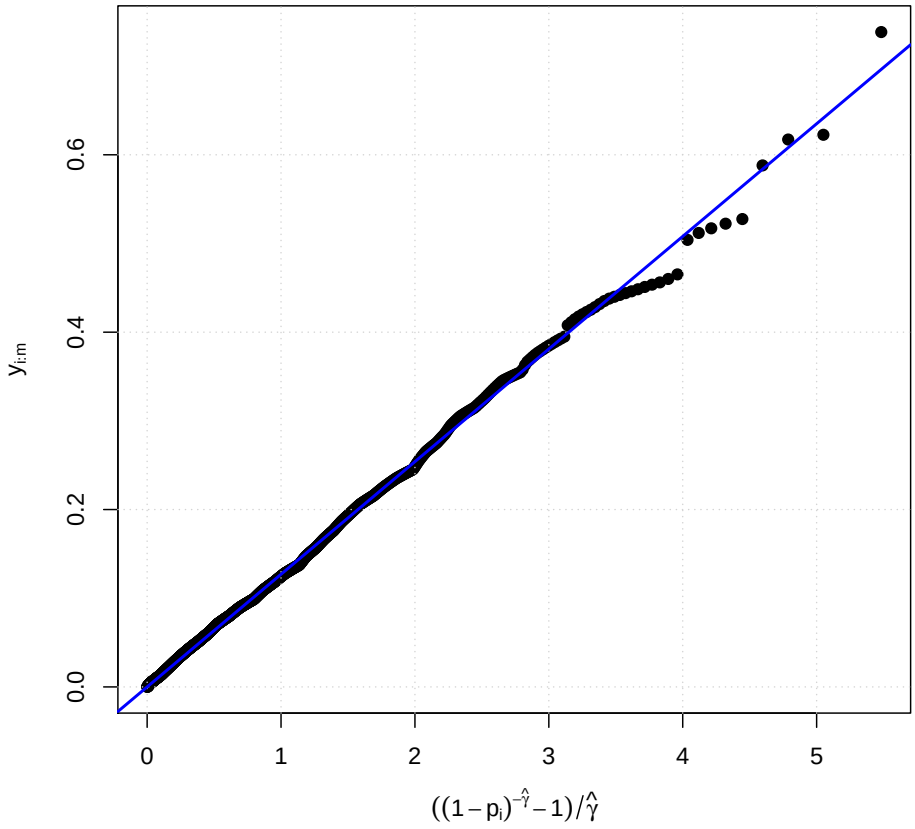

Figure 4.45: *GPd QQ-plot for the 100 metres data,with fitted line.*

statistical procedures. This plot also offers preliminary estimates of the GPd parametres:

$$(\hat{\gamma}, \hat{\sigma}_u) = (-0.07146666, 0.127). \tag{4.41}$$

b) *Statistical choice of GPd models*

From the preliminary analysis, we selected two potential candidates as a parametric model to be fitted to the excesses above the chosen threshold $u = 9.7$. The GPd with $\gamma < 0$ appeared as the first more suitable model, followed by its particular case for $\gamma = 0$, the Exponential distribution. We use then the statistical tests discussed in Section 3.1.3.6 in order to achieve (or not) some discrimination between the two candidates. We will follow the same order and methodology discussed in the $\dot{V}O_2max$ analysis, so, for theoretical questions behind the tests, we refer to the POT analysis of the previous Case Study. As for the $\dot{V}O_2max$ study, the hypotheses at play are the following:

$$H_0 : \gamma = 0 \quad vs \quad H_1 : \gamma \neq 0, \tag{4.42}$$

from a two-sided point of view, or

$$H_0 : \gamma = 0 \quad vs \quad H_1 : \gamma < 0, \tag{4.43}$$

for its one-sided version.

Recall that most of the statistical tests presented in this Section work with the exceedances above the chosen threshold. Following (3.19) and the first Case Study, we defined the exceedance above u by the r.v. W. The first discussed test was developed by Gomes and van Monfort (1986), which is used to test (4.43) with the statistic defined by (4.23). With the help of the **R** software, we get the following results (cf. Appendix A.28):

```
[1] g_m= 1.128476   g_m*= -6.218269   p-value= 1.138534e-218
```

At the usual asymptotic levels α, the statistic is highly significant and, therefore, the null hypothesis of Exponential distribution is rejected. Then, the first covered test selects the GPd with $\gamma < 0$ as a suitable parametric model to be fitted to the data. Recall from the preliminary analysis that the GPd appeared as a better candidate than the Exponential distribution.

The next test comes from the work of Marohn (2000) and can be used for (4.42) or (4.43) using the statistic in (4.24). It was seen in the $\dot{V}O_2max$ analysis that the statistic is biased when we perform the two-sided version of the test, leading to reasonable results only for sample sizes $m > 500$. As we are working with $m = 1051$ exceedances, we can

perform the two versions of the test, when we only performed the one-sided version in the $\dot{V}O_2max$ analysis. The **R** software leads us to the following results (cf. Appendix A.29):

```
[1] Two-sided Test
t_m= -0.08076792   t_m*= -2.618426   p-value= 0.008833656
[2] One-sided Test
t_m= -0.08076792   t_m*= -2.618426   p-value= 0.004416828
```

Once again, at the usual asymptotic levels α, the two-versions of the test reject the null hypothesis of $\gamma = 0$, favouring again the GPd to the detriment of the Exponential distribution.

The famous LRT is now applied to confirm (or not) our choice of the GPd as a suitable parametric model to be fitted to our data. The tested hypotheses are defined by (4.42) and checked with the statistic defined by (4.26). To compute the observed value of the statistic, we need the final ML estimates of the Exponential and GP distributions, since we only have preliminary estimates given by (4.40) and (4.41), respectively. Our usual **R** software yields the following results (cf. Appendix A.30):

```
[1] Exponential ML estimates
 sigma_u= 0.1190547
[2] GPd ML estimates
 gamma= -0.0951046   sigma_u= 0.1303735
```

The final ML estimates are then

$$\hat{\sigma}_u = 0.1190547,$$

for the Exponential distribution, and

$$(\hat{\gamma}, \hat{\sigma}_u) = (-0.0951046, 0.1303735), \tag{4.44}$$

for the GPd. Notice the quality of the preliminary estimates, since the they are very close to the final ML estimates. In particular, the final estimate of the EVI is lower, indicating a lighter right tail than expected for the underlying d.f. F.

Now, with the final ML estimates in hand, we can proceed to the LRT by means of the statistic given by (4.26). The **R** software produces the following results (cf. Appendix A.31):

```
[1] l= 8.986452   l*= 8.95238   p-value= 0.002771083
```

The LRT does not modify our decision: at the usual asymptotic levels α, the null hypothesis of Exponential distribution is one more time rejected. The LRT supports then the decisions of the previous tests.

Finally, to close the battery of statistical tests for the model choice, we end this Section with the three usual goodness-of-fit tests. The Kolmogorov-Smirnov statistic is used to test the null hypothesis of an Exponential fit and the Cramér-von Mises and Anderson-Darling statistics are used to test the null hypothesis of a GP fit. Considering the Kolmogorov-Smirnov statistic given by (4.29) and presented by Lilliefors (1969), we obtain the following results with the **R** software (cf. Appendix A.32):

```
Kolmogorov-Smirnov statistic:   0.03480792
```

The observed statistic is then compared with the simulated critical values of Table 4.7. As we have $m = 1051$, the critical values for $\alpha = 0.05$ and $\alpha = 0.01$ are given by $\frac{1.06}{\sqrt{1051}} = 0.03269674$ and $\frac{1.25}{\sqrt{1051}} = 0.03855748$, respectively. The null hypothesis of Exponential fit is then rejected at the asymptotic level $\alpha = 0.05$, but not at the asymptotic level $\alpha = 0.01$. However, we know that the goodness-of-fit tests are conservative. On the other hand, the observed value of the statistic lies between the two critical values and is not far from the higher one. Thus, it is not recommended to maintain the null hypothesis, because of the presented arguments, and the Exponential fit is then rejected. Respecting the Cramér-von Mises and Anderson-Darling statistics defined in (4.30) and in (4.31) and studied by Choulakian and Stephens (2001), we can look at what the **R** software has to tell us (cf. Appendix A.33):

```
Cramer-von Mises statistic:   0.089607
Anderson-Darling statistic:   0.5375702
```

The critical points can be found in Table 4.8. As we have $\hat{\gamma} = -0.0951046$ the table is entered at -0.1. At the usual asymptotic levels α, the two statistics do not exceed their respective critical points and, consequently, the null hypothesis of GP is not rejected. Despite their conservativeness, we decided to maintain the null hypothesis, since the observed values are not so close to the critical values, contrasting with the Kolmogorov-Smirnov statistic where the observed value was involved with the critical points.

We close then this Section electing the GPd with $\gamma < 0$ as a suitable parametric model to be fitted to the 100 metres speeds above the threshold $u = 9.7$. Such a model states that the underlying d.f. F of the 100 metres running speed, represented by the r.v. X, is light right-tailed with a finite right endpoint, contrasting strongly with the Block Maxima approach.

c) ***Parametric estimation of extreme events***

With a parametric model in hand, we can now proceed to the estimation of the GPd parameters. The ML estimates were already computed, when the LRT was performed, and are available in (4.44). Concerning the PWM method, the results are given by the **R** software (cf. Appendix A.34):

```
[1] GPd PWM estimates
gamma= -0.1085283   sigma_u= 0.1319759
```

The estimation results of both methods are presented in Table 4.20.

Table 4.20: *ML and PWM estimates of the shape and scale parameters of the GPd for the 100 metres data.*

Estimation method	$\hat{\gamma}$ (shape)	$\hat{\sigma}_u$ (scale)
ML	-0.0951046	0.1303735
PWM	-0.1085283	0.1319759

As we can see, the PWM estimates are very similar to the ML estimates. Provided that we are in a large sample context, the two methods provide satisfactory results. Following then a parametric approach, we obtain very similar results for the EVI-estimate when compared to the estimate obtained by Einmahl and Magnus (2008), but regarding Einmahl and Smeets (2011), we obtained higher estimates for the EVI, inducing a heavier right tail of the underlying d.f. F.

We can use the **R** software to extract some diagnosis tools for the ML and PWM fits. The diagnosis plots can be found on Figures 4.46 and 4.47 (cf. Appendix A.35). The diagnosis plots show similar and satisfactory results for both estimation methods.

As a reinforcement for the point estimates, we can calculate CI's for the GPd parameters. In contrast to the $\dot{V}O_2max$ POT approach, we can obtain such intervals based on the profile likelihood function, since our EVI estimate is far from -1, and so does the true value of γ (at least, we are expecting it), under which the ML procedure is not applicable. Extracting the profile likelihood method from the suitable package of the **R** software, we obtain the following results, at an asymptotic confidence level of 95% (cf. Appendix A.36):

```
[1] profiling shape

   conf.inf    conf.sup
-0.14519520 -0.03528529
```

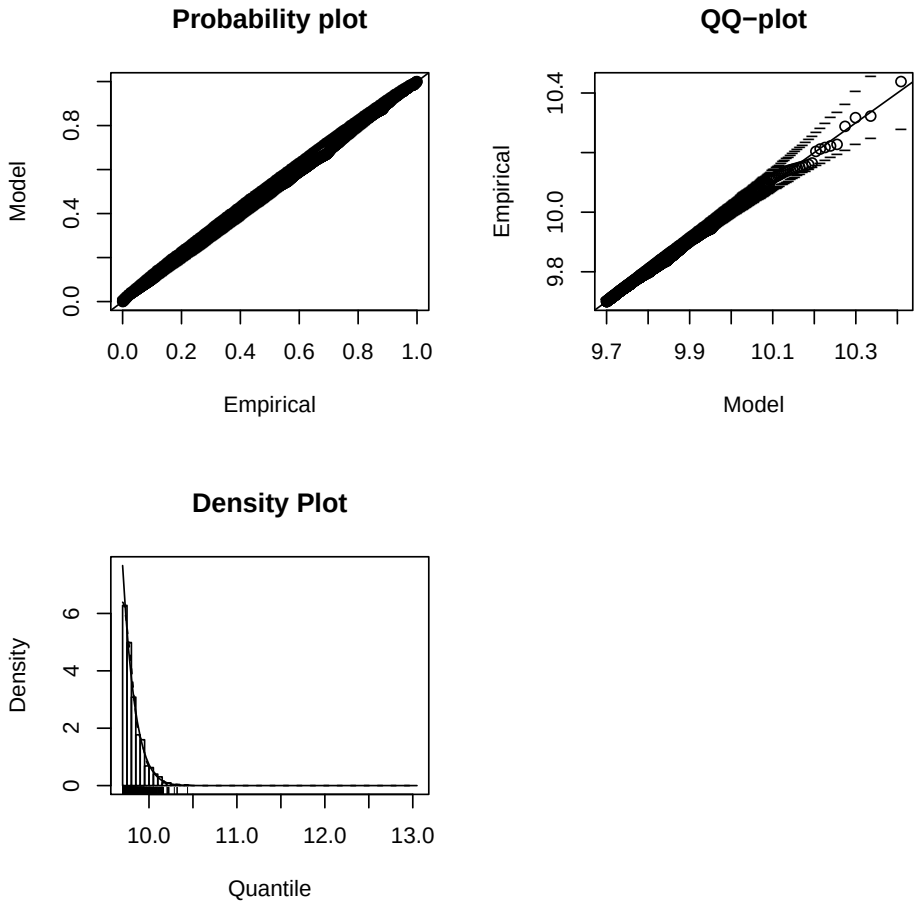

Figure 4.46: *Diagnosis plots for the ML fit of the GPd for the 100 metres data.*

```
[2] profiling scale

 conf.inf  conf.sup
0.1200450 0.1412162
```

The obtention of the 95% profile CI's can be visualized graphically. For instance, concerning the EVI, Figure 4.48 follows the procedure described in (3.31).

Looking at the CI for γ, we notice that the value 0 is excluded from the interval, pulling away the Exponential distribution as a suitable parametric model, at least for a 95% asymptotic confidence level.

To end the POT approach, we can estimate the usual probability of surpassing the current world record of Usain Bolt and present an estimate of the right endpoint of the underlying d.f. F, based on the ML and PWM estimates. Recall that the current world record of Usain Bolt is 9.58 seconds, which gives the maximal speed of 10.438 m/s. Using (3.29) and (3.30) and turning to our **R** software (cf. Appendix A.37), we get

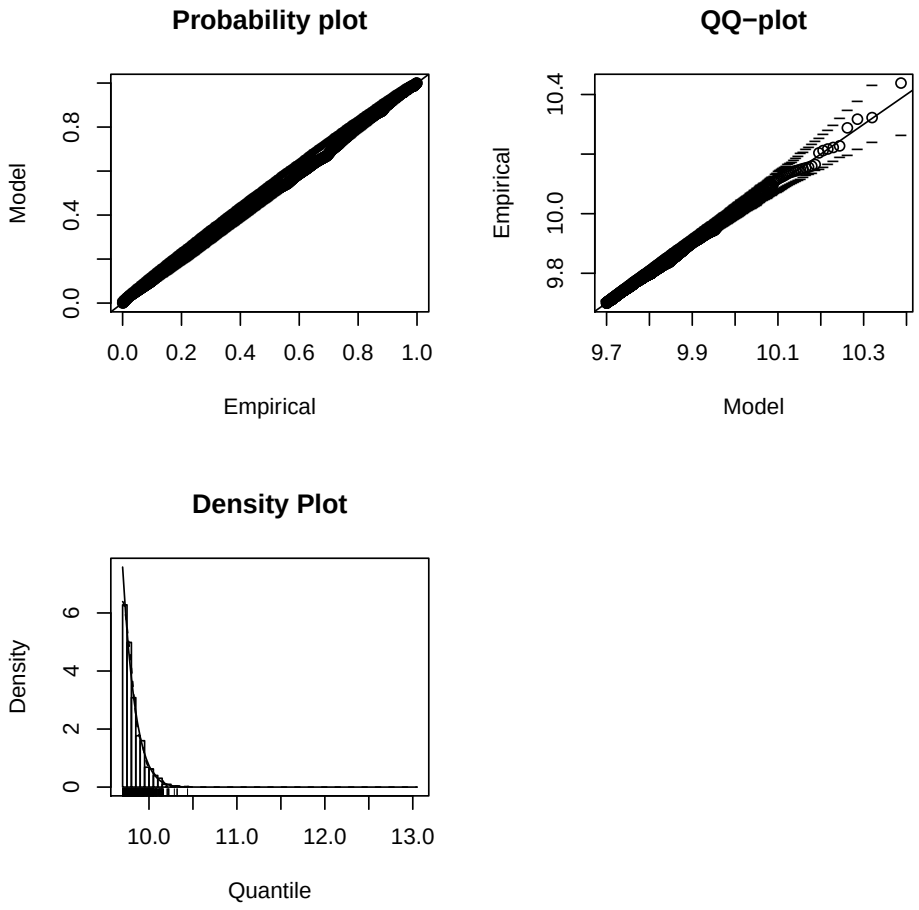

Figure 4.47: *Diagnosis plots for the PWM fit of the GPd for the 100 metres data.*

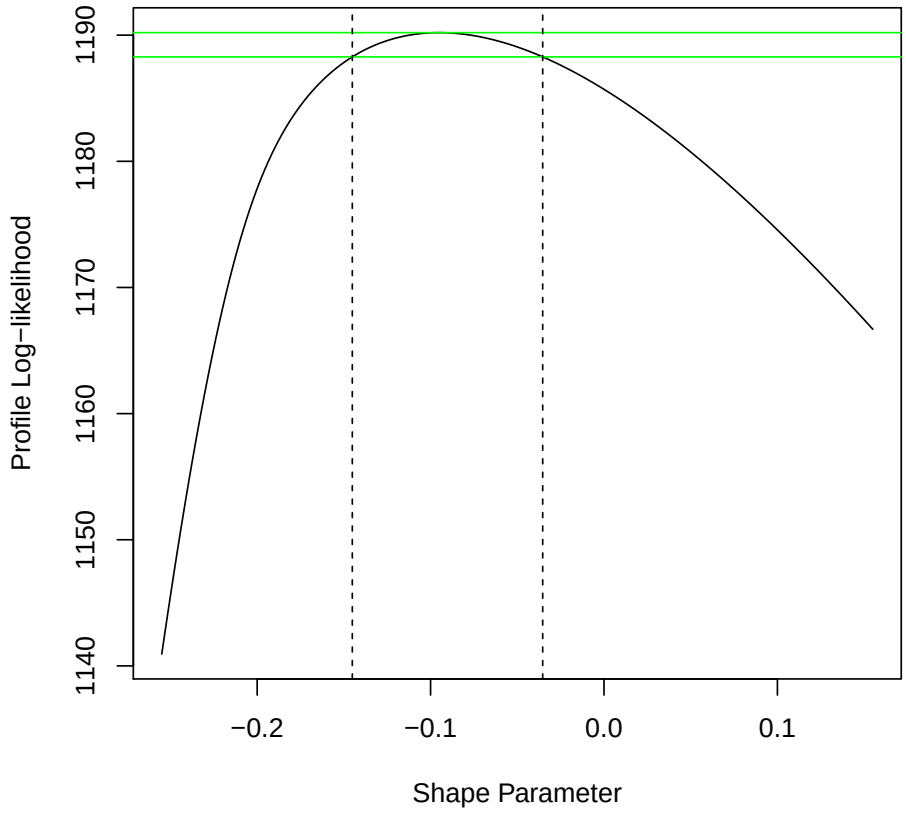

Figure 4.48: *Profile log-likelihood function for γ under the POT approach of the 100 metres data.*

```
[1] Maximum Likelihood: P(X>10.438)= 0.000260351
[2] Probability Weighted Moments: P(X>10.438)= 0.0001616754


[1] Maximum Likelihood: x^F= 11.07084
[2] Probability Weighted Moments: x^F= 10.91605
```

The right endpoint estimates are very similar, either by the ML method or by the PWM method. In particular, the ML provides the highest estimate for x^F, which corresponds to a running time of 9.03 m/s, while the PWM method obtains a boundary time of 9.16 m/s. Therefore, in the present circumstances, a 100 metres athlete has still some space for improvement. Compared to the current world record of Usain Bolt, a top-athlete can still reduce the 100 metres running time by approximately 0.42 seconds, in terms of PWM, or by approximately 0.55 seconds, in terms of ML. The estimated probabilities of getting such improvement are given by 0.0167% and 0.026%, following the PWM and ML methods, respectively.

The POT approach brings more interesting results when compared to the Block Maxima approach. Recall that this latter approach determined an infinite right endpoint for the running speed of the athletes, which is plausible in a science fiction world, but absurd in our reality. The POT approach states a finite right endpoint for the light right-tailed d.f. F. Notice finally that the estimates of the right endpoint x^F presented in this thesis under a parametric framework are a little higher than those obtained by Einmahl and Magnus (2008) and Einmahl and Smeets (2011), in terms of running speed, or a little lower, in terms of running time. Considering then only data above the threshold $u = 9.7$ is sufficient to obtain evidence for $\gamma < 0$ and to remove the absurd conclusion of $\gamma > 0$ obtained under the Block Maxima approach. The data under this threshold seem to have enough power to modify the decision about the EVI sign. If we take a closer look at the Gumbel and GEVd QQ-plots of Figure 4.33 and 4.36, we note a convex pattern of the plotted points precisely until 9.7, the chosen threshold based on the kink of the ME-plot in Figure 4.30. This convex pattern, symptom of heavy tails, is sufficient to create differences if we consider a threshold, under a POT approach, or not, under a Block Maxima approach. Note equally that the convex pattern corresponds to the worst results. Hence, are the data under 9.7 m/s really needed for the analysis? Are those data originating from another population of athletes with other d.f., where 9.7 may establish a boundary between the two populations? Are we facing a mixture of distributions? All these questions are pertinent and interesting, but, unfortunately, they are beyond of the scope of this thesis.

4.2.2 Semi-Parametric data analysis

a) *Testing the extreme value index sign*

We may begin this Section with the semi-parametric tests developed in Section 3.2.6, to achieve some discrimination of the EVI sign. The hypotheses at play are then

$$H_0 : F \in \mathcal{D}(G_0) \quad vs. \quad H_1 : F \in \mathcal{D}(G_\gamma)_{\gamma \neq 0}, \tag{4.45}$$

tested with the Greenwood ($R_n^*(k)$), Hasofer-Wang ($W_n^*(k)$) and Ratio ($T_n^*(k)$) standardized statistics.

As the statistics depend on the random threshold k, we can visualize their respective sample paths in Figure 4.49, where $R_n^*(k)$, $W_n^*(k)$ and $T_n^*(k)$ are plotted against k (cf. Appendix A.38).

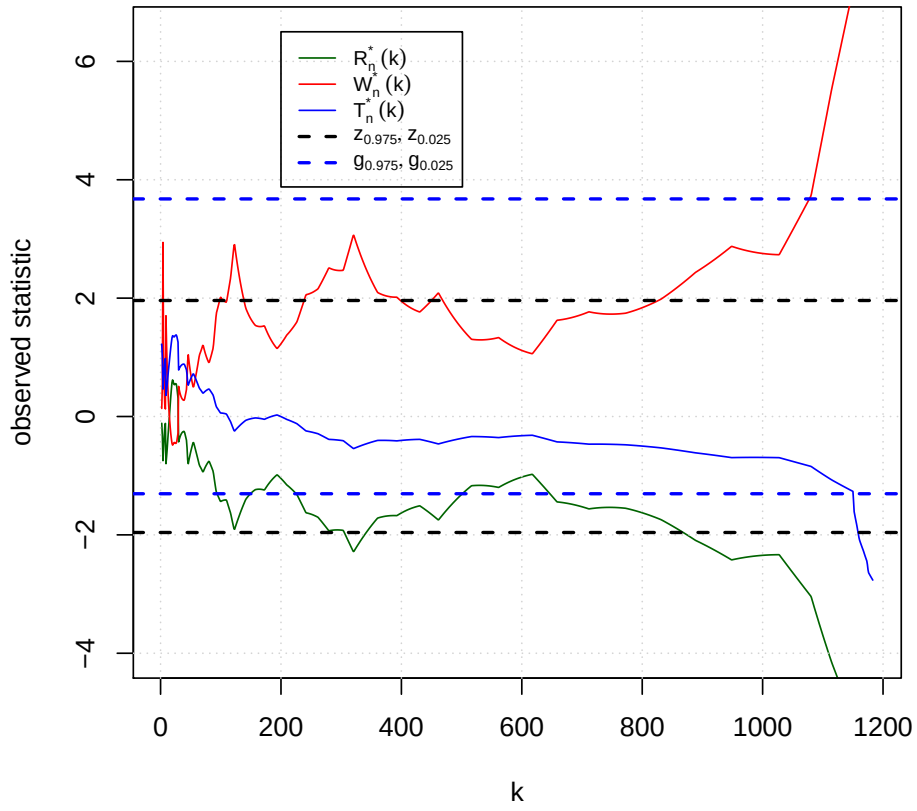

Figure 4.49: *Sample paths of Greenwood (G_n^*), Hasofer-Wang (W_n^*) and Ratio (T_n^*) statistics for testing $H_0 : F \in \mathcal{D}(G_0)$ vs. $H_1 : F \in \mathcal{D}(G_\gamma)_{\gamma \neq 0}$ in the 100 metres data.*

The Greenwood ($R_n^*(k)$) and the Hasofer-Wang ($W_n^*(k)$) statistics are useful to detect right heavy-tailed and light-tailed d.f.'s, respectively. Looking at their sample paths and

considering the decision rule in (3.70), we can notice a oscillatory behaviour of the paths around the Normal quantiles, ending up by surpassing the respective Normal quantiles, for large values of k. If we change the asymptotic size α, the oscillatory behaviour can be more or less pronounced. Anyway, it can be stated that, the sample paths head for the rejection region, since they oscillate dangerously around the critical values given by the Normal quantiles. The Ratio statistic $(T_n^*(k))$, used as a complement, is always between the Gumbel quantiles, indicating a non-rejection of H_0. However, as stated in Section 3.2.6, the Ratio test tends to be conservative, specially if the true value of γ is close to zero. This may explain the difficult decision that can be taken based on the sample paths. At the asymptotic size of $\alpha =$, we reject H_0 with some doubts, provided that the EVI has probably a value close to zero.

If we want to be more specific about the EVI sign, we can perform a one-sided test. Remember that the parametric Block Maxima approach selected a GEVd with $\gamma > 0$ as a suitable d.f. to be fitted to the 100 metres data. We can now check if the semi-parametric approach selects a Fréchet max-domain of attraction for the 100 metres data, i.e. we want to test

$$H_0 : F \in \mathcal{D}(G_0) \quad vs. \quad H_1 : F \in \mathcal{D}(G_\gamma)_{\gamma>0},$$

using the same statistics as for (4.45).

But we can check what happens if we are interested in a Weibull max-domain of attraction, performing then the test

$$H_0 : F \in \mathcal{D}(G_0) \quad vs. \quad H_1 : F \in \mathcal{D}(G_\gamma)_{\gamma<0}.$$

The sample paths of the three aforementioned statistics for both one-sided tests can be visualized in Figure 4.50 (cf. Appendix A.39).

Considering the decision rule in (3.71) and the first plot of Figure 4.50, we observe that the sample paths of the three statistics never surpass their respective critical values. Even the Greenwood statistic, a powerful tool to detect a Fréchet max-domain of attraction, never leads us to the rejection of the null hypothesis. Therefore, contrasting with the parametric Gumbel's approach, the semi-parametric inference does not consider a right heavy-tailed d.f. as an underlying model for our 100 metres athletes' population. As we are interested in a Weibull max-domain of attraction, we can focus on the Hasofer-Wang statistic. Considering the decision rule in (3.72) and the second plot of Figure 4.50, we verify that the sample path of this statistic is mainly in the rejection zone and, consequently, the null hypothesis of Gumbel max-domain of attraction is rejected for the major part of the k values, at the asymptotic size of $\alpha = 0.05$. The Ratio statistic is almost always within the non-rejection region, as it was already seen. Since the test based in

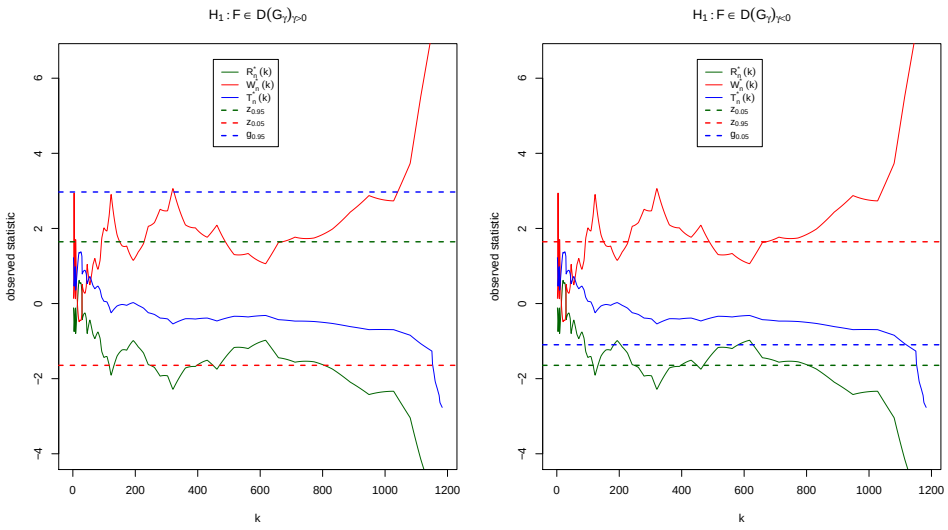

Figure 4.50: *Sample paths of Greenwood, Hasofer-Wang and Ratio statistics of the one-sided tests for the 100 metres data.*

this latter statistic is conservative, the non-rejection of the Gumbel max-domain may be explained by the negative, but close to zero, value of the EVI. The Weibull max-domain is then the most appropriate domain for the 100 metres running speeds.

b) *Heuristic choice of the random threshold*

Now that the max-domain of attraction has been defined through the sign of the EVI, we can proceed to the appropriate choice of the random threshold k in order to estimate the EVI with the semi-parametric estimators of Section 3.2.3. As we did for the $\dot{V}O_2max$, we will follow the heuristic procedure of Section 3.2.7. According to (3.73), k is chosen so that the difference between the EVI estimates obtained with the semi-parametric estimators of Section 3.2.3 is the lowest. Therefore, we have to specify which semi-parametric estimators are chosen for this optimization problem.

The first candidate to be included is the Pickands estimator, defined in (3.43). However, as seen in the $\dot{V}O_2max$ case study, this estimator is reputed for its large asymptotic variance, which can compromise the optimization problem required by the heuristic process. As a curiosity, we present the sample path of the Pickands estimator in Figure 4.51 (cf. Appendix A.40).

As for the previous case study, the Hill estimator will not be used for the heuristic process, since it is valid only for $\gamma > 0$ and the Fréchet max-domain of attraction was rejected by the semi-parametric tests. On the contrary, the Moment and Generalized Hill estimators can be used without any problem, since they are valid for $\gamma \in \mathbb{R}$. Concerning

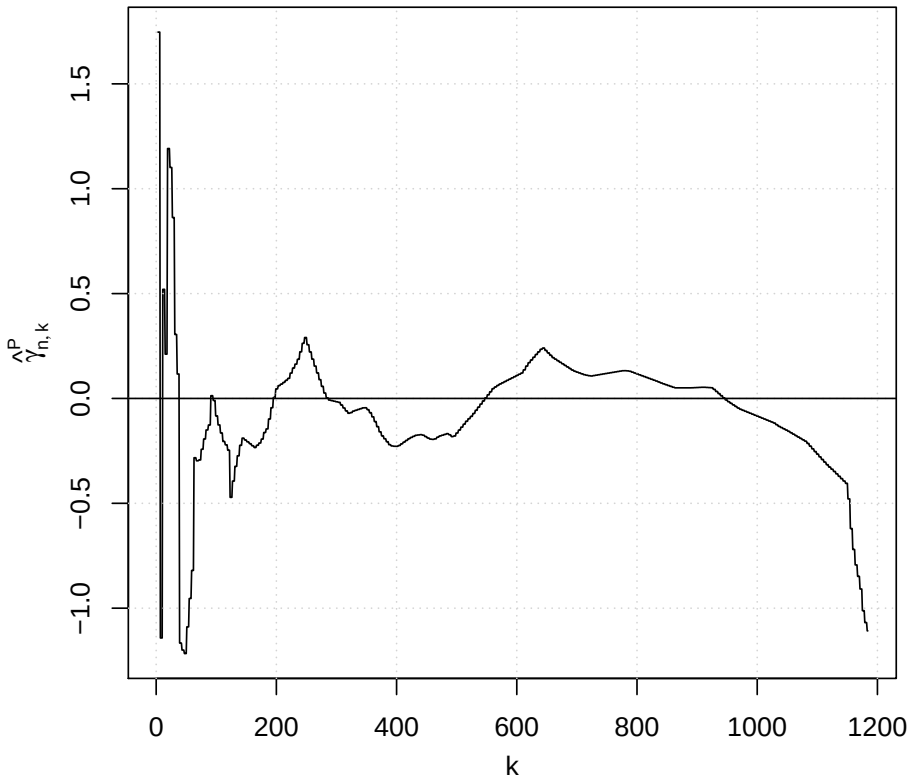

Figure 4.51: *Pickands-plot for the 100m data.*

the Negative Hill estimator, we know it is valid only for $\gamma < -0.5$. As for the $\dot{V}O_2max$ case study, we can check the sample path of this estimator, to ascertain the eventual validity of $\gamma < -0.5$ (cf. Appendix A.41). From Figure 4.52, we note that the sample path of the estimator is always above -0.5 and seems to stabilize between -0.2 and -0.25 for most of the k values. Consequently, the validity of $\gamma < -0.5$ is very questionable and difficult to prove. Thus, as for the previous case, the Negative Hill estimator will not be included as a participant of the heuristic process.

Finally, the PORT estimators depend on the tuning parameter q. Following the same rationale as the previous case, we can plot the sample path of the PORT-Moment and the PORT-Mixed Moment estimators for different values of q to have some clues about the "best" choice of q. Consider then Figure 4.53 (cf. Appendix A.42).

We extract the same conclusion from the two plots: a choice of a lower value for q provides a more stable path of the corresponding EVI-estimator. Recall that the same conclusion was drawn from the $\dot{V}O_2max$ analysis. Following then the previous case study in the light of Fraga Alves et al. (2009), we will choose again $q = 0.01$.

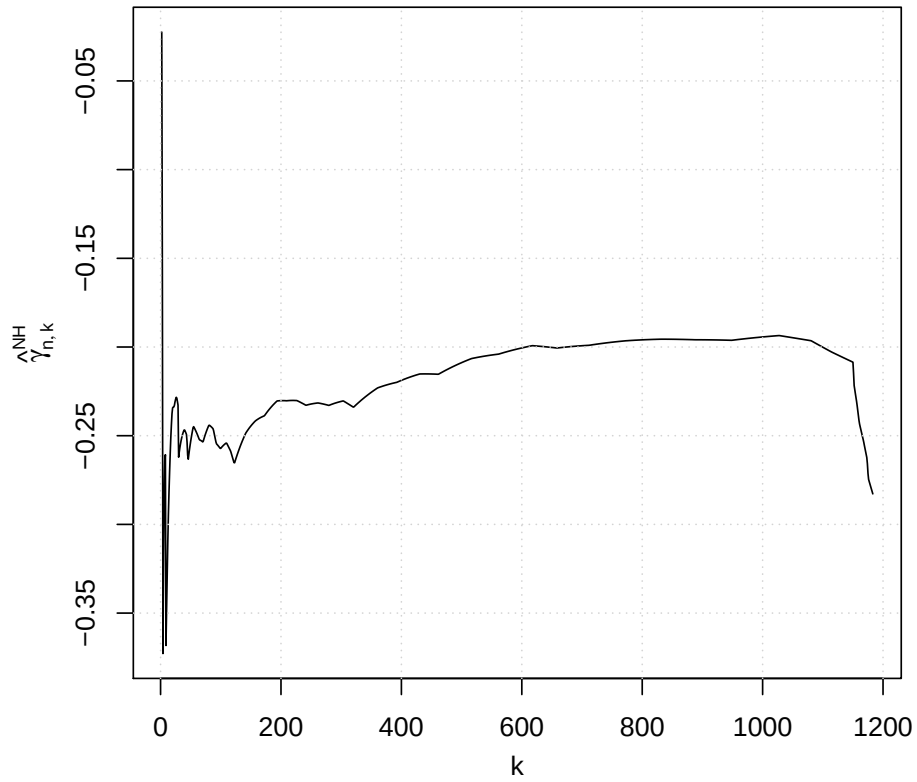

Figure 4.52: *Negative Hill estimator sample path for the 100m data*

All the chosen estimators can now be plotted together in order to seek a region where all the candidates provide concordant values for the EVI. We can see the corresponding plot in Figure 4.54 (cf. Appendix A.43).

The plot helps us to make a suitable choice of the estimators to be used in the heuristic procedure:

1. The Moment, Generalized Hill and Mixed Moment estimators are concordant along the whole plot, with concordant estimates;

2. The PORT estimators show an erratic behaviour: the PORT-Mixed Moment estimator has an upward trend above zero, as for the $\dot{V}O_2max$ Case Study, and the PORT-Moment estimator keeps some distance from the group formed by the Moment, Generalized Hill and Mixed Moment estimators, specially after $k = 350$. Consequently, they will be excluded from the analysis;

3. The Pickands estimator shows its reputed high variability and its use may distort

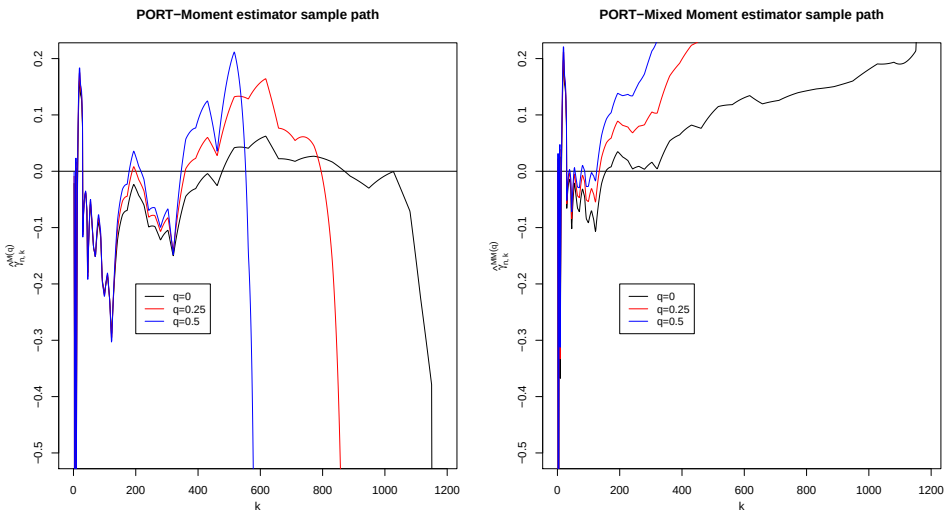

Figure 4.53: *PORT-Moment and PORT-Mixed Moment estimators sample paths for the 100m data*

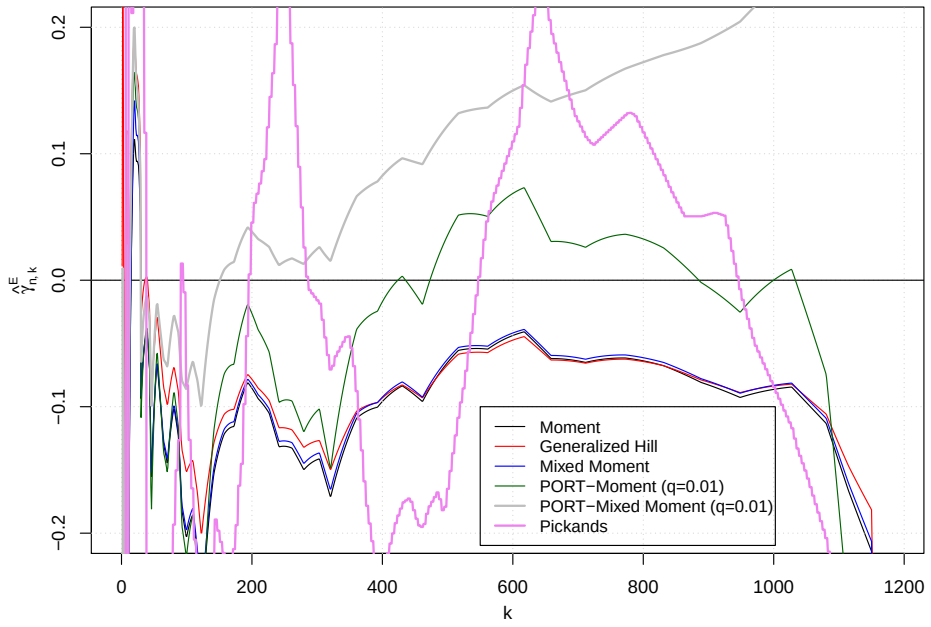

Figure 4.54: *Sample paths of Moment, Generalized Hill, Mixed Moment, PORT-Moment, PORT-Mixed Moment and Pickands estimators for the 100m data*

the optimization problem of the heuristic procedure. Therefore, it will be excluded from the set of candidates.

We decided to keep the Moment, Generalized Hill and Mixed Moment for the heuristic procedure since they exhibit concordant sample paths. We can now use (3.73) to obtain the most suitable choice of k with the help of the **R** software (cf. Appendix A.44):

```
[1] k opt= 860
```

The heuristic procedure selects then $k + 1 = 861$ top observations on which all the inference about the extreme value events will be based. But before that, it would be pertinent to understand all the calculus procedure behind the optimal value $k = 860$. Remember from Figure 4.54 that the paths of the chosen semi-parametric estimators are very close along all the plot, suggesting eventually many possibilities for k. Consider then Figure 4.55 (cf. Appendix A.45).

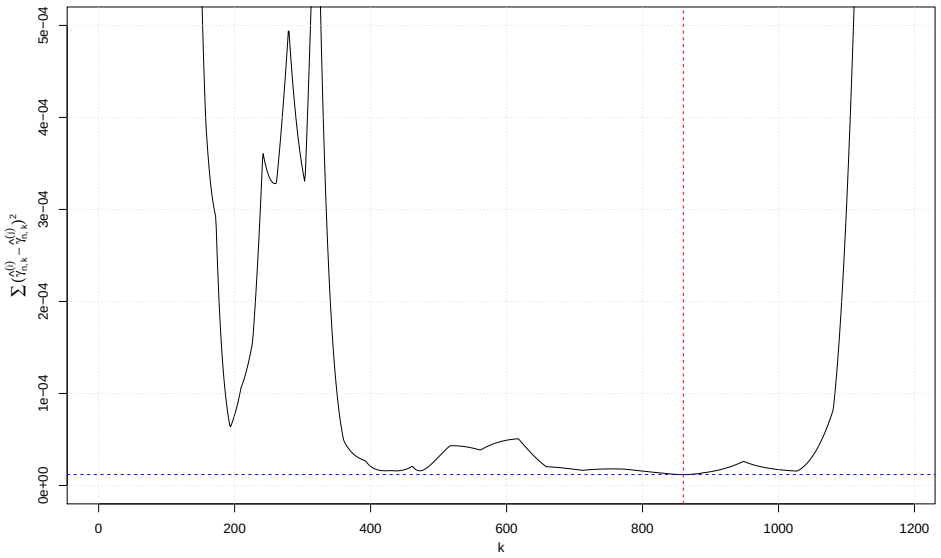

Figure 4.55: *Heuristic choice of the threshold k for the EVI estimation of the 100m data.*

As expected and considering the scale, we notice that the distances between the EVI-estimates are very small considering the three chosen estimators. The best choice of $k = 860$ is visible in the plot. Generally speaking, the region $400 < k < 1000$ seems a good one in order to select the most suitable value for k. We could keep, for instance, a value of k between 400 and 500 as well. But looking at Figure 4.54, we see that this region seems slightly more unstable than the region where we chose $k = 860$. This kind of choice and consideration is always arguable. Since we need a value for k in order to proceed,

we decided to keep the value $k = 860$ that results from the optimization problem. Recall that under a POT approach, we worked with $m = 1051$ observations above the selected threshold 9.7. Under a semi-parametric approach, fewer observations are needed, since the heuristic process elected $k + 1 = 861$ observations.

c) *Estimation of the Extreme Value Index*

Now we have elected an optimal k-value, we can proceed to the estimation of the EVI, considering the chosen estimators. We obtain the following results with the **R** software (cf. Appendix A.47):

```
[1] k opt=860
 Moment: gamma= -0.07384262    Generalized Hill: gamma= -0.07292183
 Mixed Moment: gamma= -0.07108349
```

The selected estimators provide very similar results for the EVI estimate, around -0.07. At this point, it would be interesting to compare the semi-parametric estimate of γ with those obtained under parametric approaches. The comparison is only made with the POT approach, since the Block Maxima approach led to unrealistic results, in a context of heavy right tails. Consulting then Table 4.20, we find very similar results, mainly with the ML method. Comparing the semi-parametric results for the EVI estimation with Einmahl and Magnus (2008) and Einmahl and Smeets (2011), we obtain higher estimates for γ, as it was stressed in POT analysis.

We can obtain richer estimates for γ through CI's computed using (3.57). For that, estimates for the asymptotic variances σ_E^2 are needed. They just can be found in (3.59) and computed with our **R** software (cf. Appendix A.48):

```
[1] k opt=860
 Moment: s2_M= 0.9255103    Generalized Hill: s2_GH= 0.878037
 Mixed Moment: s2_MM= 0.9261752
```

Using then the estimated asymptotic variances, we can now use expression 3.57 and construct 95% CI's for the EVI. Results are concentrated in Table 4.21.

Looking at the approximate CI's for γ, we notice that none of them include zero, confirming then the Weibull max-domain of attraction and labeling the underlying d.f. F as a light right-tailed one, with a finite right endpoint. But recall that the obtained CI's are approximated intervals, which can be improved, but with techniques out of this thesis.

Table 4.21: *Semi-parametric approximate 95% confidence intervals for γ for the 100 metres data.*

Semi-parametric estimator of γ	$k = 860$
Moment	-0.07384262 (-0.1381406,-0.009544618)
Generalized Hill	-0.07292183 (-0.1355491,-0.0102946)
Mixed Moment	-0.07108349 (-0.1354046,-0.006762394)

d) ***Semi-parametric estimation of other extreme events***

The last step consists in obtaining estimates for the location coefficient, $b\left(\frac{n}{k}\right)$, the scale coefficient, $a\left(\frac{n}{k}\right)$, the right endpoint, x^F, and for a suitable exceedance probability. The attraction coefficients can be estimated easily with conditions (3.49) and (3.50), with the participation of our usual **R** software (cf. Appendix A.49):

```
[1] k opt=860
b(n/k)= 9.727626

[2] k opt=860
a(n/k)= 0.1233697
```

Once more, note the extreme proximity between the semi-parametric approach and the POT method. Concerning the estimate of the location coefficient, $\hat{b}\left(\frac{n}{k}\right) = 9.727626$, we find an extraordinary proximity with the threshold u obtained with the ME-plot of the POT methodology, depicted in Figure 4.30. The same applies to the estimate of the scale coefficient, $\hat{a}\left(\frac{n}{k}\right) = 0.1233697$, almost like the ones obtained in Table 4.20, with a higher proximity obtained by the ML method.

With the estimates of the attraction coefficients, we can now proceed to the right endpoint estimate with (3.53) or (4.33). The **R** software reveals the following results (cf. Appendix A.50):

```
[1] k opt=860
 Moment: xF= 11.39834    Generalized Hill: xF= 11.41943
 Mixed Moment: xF= 11.46319
```

We find similar results in the POT methodology, specially for the ML estimation, although the semi-parametric methodology produces slightly superior results for the right

endpoint. Converting the speeds to running times, we obtain a left endpoint between 8.72 and 8.77 seconds, since the running times belong to a minimum context. Therefore, the semi-parametric approach consider that an athlete can lower the current time record of Usain Bolt (9.58 seconds) under 9 seconds, while the POT approach stayed a little above 9 seconds. There is then some space for improvement and it would not be surprising if Usain Bolt itself breaks his own record sooner or later. The point estimates of the right endpoint can be completed with CI's computed from (3.58), where the asymptotic variance estimates, $\hat{\sigma}_E^2$ were already calculated, when we constructed CI's for γ presented in Table 4.21. We present then the results for the 100 metres data in Table 4.22, in terms of speed (normal display) or in terms of time (bold). Recall that the expression for the CI's was left-truncated and, therefore, we cannot state the confidence level of the resulting intervals. All that we can say is that this level is lower than $1-\alpha$ (see Appendix A.53).

Table 4.22: *Semi-parametric approximate confidence intervals for x^F ($\alpha = 5\%$) for the 100 metres data, in terms of speed (m/s) or* ***time (seconds)***.

Semi-parametric estimator of γ	$k = 860$
Moment	11.39834 (10.438,12.0136) **8.773207** **(8.3239,9.58)**
Generalized Hill	11.41943 (10.438,12.03394) **8.757005** **(8.30983,9.58)**
Mixed Moment	11.46319 (10.438,12.12738) **8.723575** **(8.245804,9.58)**

The same strategy can be applied to select the optimal k-value in terms of the right endpoint estimate. For a first impression, we can look at Figure 4.56, which depicts the sample paths of the right endpoint estimator defined in (4.33), considering the three EVI-estimators used in the heuristic procedure (cf. Appendix A.51).

Just as Figure 4.54, we notice that the sample paths of the three estimators are very close, along all the plot, giving similar estimates for almost every k value. After $k = 600$, we observe a decreasing trend of the three estimators towards the sample maximum. The optimal value is choosing again by means of expression (3.73), replacing $\hat{\gamma}_{n,k}$ by $\hat{x}^F$ and using the **R** software (cf. Appendix A.52) to obtain the optimal solution of the minimization problem:

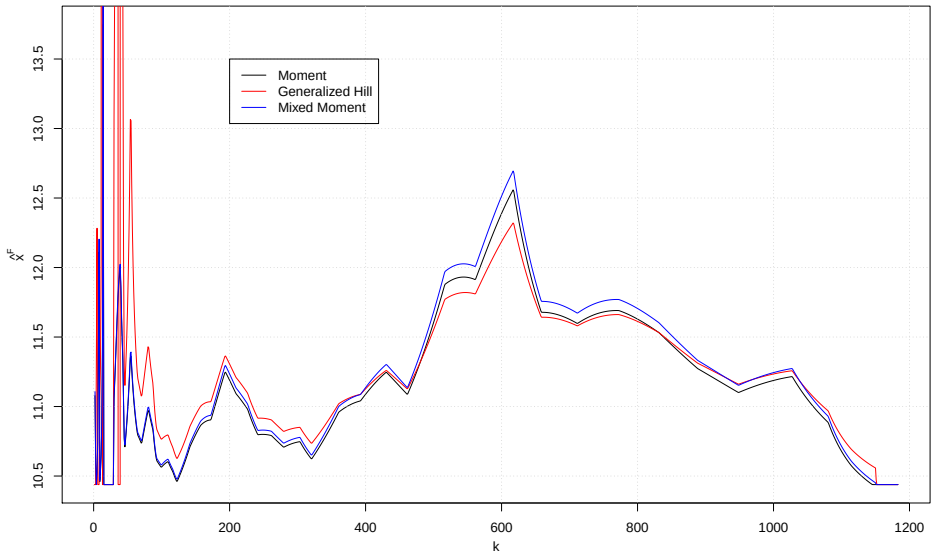

Figure 4.56: *Sample paths of Moment, Generalized Hill and Mixed Moment estimators for the right endpoint of the underlying d.f. for the 100 metres data.*

```
[1] k opt= 15
```

Mathematically, the optimal threshold found by the heuristic procedure applied to the right endpoint estimators is unfeasible, since it lies in an extremely unstable region of the plot, characterized by a high volatility of the estimators. A more adequate choice of k is then required. To seek another regions of local minima, a plot of the function defined by (3.73) and adapted for $\hat{x}^F$ would be appropriate. Consider then Figure 4.57 (cf. Appendix A.52).

As the plot demonstrates, the minimum of the distance function defined by (3.73) is effectively attained at $k = 15$. However, as it was previously observed, this k value lies in a very unstable region of the plot and cannot be maintained. Two other regions seem more adequate: the region around $k = 400$ and the region $800 < k < 1100$. Notice that the original $k = 860$ lies within the second region. In terms of stability, the region $800 < k < 1100$ seems preferable and, applying the same minimization procedure by means of the **R** software, we obtain:

```
[1] k opt= 1027
```

The heuristic procedure applied to the right endpoint estimators selects then $k + 1 = 1028$ top observations, inducing a further approximation between the semi-parametric

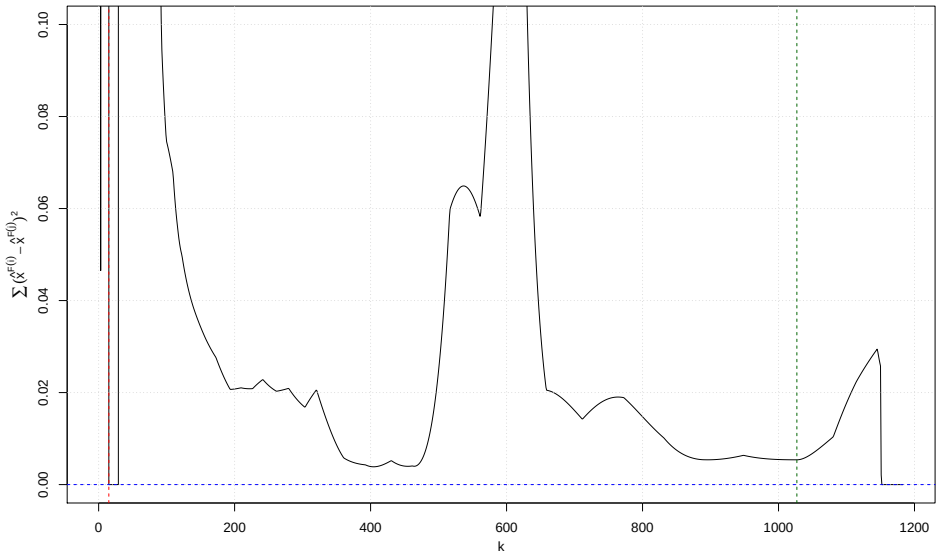

Figure 4.57: *Heuristic choice of the threshold k for the right endpoint estimation for the 100 metres data.*

approach and the parametric POT method, which selected $m = 1051$ top observations. The newly computed k value is visible in Figure 4.57.

With this new k value, the **R** software gives the following right endpoint estimates:

```
[1] k opt=1027
 Moment: xF= 11.21529    Generalized Hill: xF= 11.25726
 Mixed Moment: xF= 11.27333
```

Once more time, these estimates are closer to the estimates obtained under the POT approach, specially to ML estimates, as it was already discussed.

In the present circumstances, what is then the probability of surpassing Usain Bolt's record? This last question can be answered with our usual exceedance probability. As the current record in terms of time is 9.58 seconds, the corresponding speed is given by 10.438 m/s. We can use then condition (3.54), with $H_{\hat{\gamma}}$ given by (3.21), to obtain the required estimated probability, $P(X > 10.43841)$, recalling that the r.v. X represents the running speed of a 100 metres athlete from population defined since the beginning of this case study. Giving way to our **R** software, we obtain the following estimated probabilities (cf. Appendix A.54):

```
[1] k opt=860
 Moment: P(X>10.438)= 0.0004022769
```

```
Generalized Hill: P(X>10.438)= 0.0004150361
Mixed Moment: P(X>10.438)= 0.0004412586
```

As for all the semi-parametric inference, the exceedance probability shows us similar results when confronted with the POT analysis, where the victory is once again awarded to the ML method.

The comparison between the POT analysis and the semi-parametric approach leads us to an interesting conclusion: with fewer top observations, the semi-parametric approach presents very similar results to those obtained with the POT methodology.

Chapter 5

Conclusions and open questions

EVT offers interesting conclusions when applied to the world of Sports, even in a simplified context of analysis, where some assumptions have been made. The most important lesson is undoubtedly the clear affinity between semi-parametric approaches and the parametric POT method. In both of the Case Studies, the obtained results and estimates are very similar, for the aforementioned approaches. On the contrary, the Block Maxima method stays at the margin, yielding particular results, different from its two contenders. The 100 meters Case Study is the most flagrant situation, where this latter methodology produces a peculiar and unusual positive estimate of the EVI.

For the $\dot{V}O_2max$ analysis, the POT methodology yields similar results when compared to a semi-parametric framework, specially if we choose the PWM method as a point estimation method. Recall that this method performs better in small and modest samples than the ML does. The results seem to support this theoretical result. Generally speaking, the analysis demonstrates that the current record of Bjørn Dæhlie and Espen Harald Bjerke is arriving at a steady-state, with little space for improvement. Indeed, the estimates for the right endpoint of the underlying d.f. are very close to the sample maximum. For the 100 metres analysis, a large sample example, the ML method, applied with a POT approach, provides more concordant results with the semi-parametric analysis. As mentioned above, the Block Maxima method do not demonstrate similar affinity with the two other methods. The analysis shows that, in the present circumstances, a 100 metres athlete can improve his running time, with the possibility to reduce the current record of Usain Bolt under 9 seconds.

As stated above, all this thesis was made within simplified premises and we consider the performed work as a starting point and guideline for future and deeper investigation, refining the tools used along all the analysis. Indeed, many questions have not been treated and their consideration can improve the achieved results. The estimation of the

second order parameters of Section 3.2.2 can improve the accuracy of the semi-parametric CI's, if they have been taken into account. The knowledge of the second order parameters estimates also permits the use of the AMSE criterion for a threshold choice, since it is very popular among statisticians. Another analysis that can be done is the power comparison between the performed tests, in order to select the more suitable decision. Again, we have space for going beyond this thesis. A time-series study would also be appropriate as an additional tool that can be used for comparison.

To sum up, this thesis is neither complete, nor restrictive. It provides us with a constructive framework, which can be used as a starting point for further investigation. It was our intention to gather the most well known techniques from the EVT in order to construct an organized methodology, with a logical framework, which permits to obtain interesting conclusions, when applied to Sports. This thesis whets then our appetite for other sports' modalities that can be analyzed with the methodology presented along this work. Tennis, skiing, swimming, weightlifting, all these sports abound in data that are waiting for treatment, with the tools of this thesis. We are then just at the beginning of the story, with a maximum of curiosity and a minimum of wasted time...

Appendix A

R scripts for the $\dot{V}O_2max$ Case Study

A.1 Sample ME-plot

```
Data<-read.table("VO2max.txt",header=T)
vo2max<-Data$VO2max
vo2max<-sort(vo2max)
library(evir)
meplotvo2max<-meplot(vo2max,type="o",omit=1)
grid()
```

A.2 Sample ME-plot, with fitted straight lines

```
linreginit1<-lm(meplotvo2max$y[1:25]~meplotvo2max$x[1:25])
linreginit2<-lm(meplotvo2max$y[26:74]~meplotvo2max$x[26:74])

meplot(vo2max,type="o",omit=1)
abline(linreginit1,col="blue")
abline(linreginit2,col="red")
grid()
```

A.3 Exponential QQ-plot

```
m<-length(vo2max)
i<-c(1:m)
fun<-function(x) -log(1-x)
Q<-fun(i/(m+1))
plot(Q,vo2max,pch=19,xlab=expression(-log(1-p[i])),ylab=expression(y[i:m]))
grid()
```

A.4 Gumbel QQ-plot

```
fung<-function(x) -log(-log(x))
Qg<-fung(i/(m+1))
plot(Qg,vo2max,pch=19,xlab=expression(-log(-log(p[i]))),ylab=expression(y[i:m]))
grid()
```

A.5 Linear fit for the Gumbel QQ-plot

```
linreg<-lm(vo2max~Qg)
print(linreg,digits=5)
abline(linreg,col="blue",lwd=2)
```

A.6 Correlation for the GEVd QQ-plot

```
correlgev<-function(g){
x<-((-log(i/(m+1)))^(-g)-1)/g
cor(vo2max,x)
}
(gopt<-optimize(correlgev,interval=c(-.5,.5),maximum=T))

# Correlation plot
seqg<-seq(-1,.5,.01)
correl<-sapply(seqg,correlgev)
correl[101]<-cor(Qg,vo2max)
plot(seqg,correl,type="l",xlab=expression(gamma),
  ylab=expression(corr(Q[list(gamma,0,1)](p[i]),y[i:m])),mgp=c(2.2,1,0))
grid()
abline(v=gopt$maximum,lty=2,col="red"); abline(h=gopt$objective,lty=2,col="blue")
```

A.7 GEVd QQ-plot

```
Qgev<-((-log(i/(m+1)))^(-gopt$maximum)-1)/(gopt$maximum)
plot(Qgev,vo2max,xlab=expression(((-log(p[i]))^-hat(gamma)-1)/hat(gamma)),
  ylab=expression(y[i:m]),pch=19)
grid()
```

A.8 Linear fit for the GEVd QQ-plot

```
(linereggev<-lm(vo2max~Qgev))
abline(linereggev,col="blue",lwd=2)
```

A.9 Gumbel test statistic

```
GS<-(max(vo2max)-vo2max[floor(m/2)+1])/(vo2max[floor(m/2)+1]-min(vo2max))
bn0<-(log(m)+log(log(2)))/(log(log(m))-log(log(2)))
an0<-1/(log(log(m)))
library(evd)
pvaluewn<-pgumbel((GS-bn0)/an0)
cat("[1] gs_m=",GS,"  gs*_m=",(GS-bn0)/an0,"  p-value=",pvaluewn,"\n")
```

A.10 Gumbel and GEVd ML estimation

```
dGumbel <- function(x,a,b) 1/a*exp((b-x)/a)*exp(-exp((b-x)/a))
pGumbel <- function(q,a,b) exp(-exp((b-q)/a))
qGumbel <- function(p,a,b) b-a*log(-log(p))
a_Gumbel=as.vector(linreg$coefficients[2])
b_Gumbel=as.vector(linreg$coefficients[1])
library(fitdistrplus)
fGumb<-fitdist(vo2max,"Gumbel",start=list(a=a_Gumbel,b=b_Gumbel))
b<-as.vector(fGumb$estimate[2])
a<-as.vector(fGumb$estimate[1])

fGev<-fgev(vo2max)
shapgev<-as.vector(fGev$param[3])
locgev<-as.vector(fGev$param[1])
scalgev<-as.vector(fGev$param[2])
cat(" [1] Gumbel ML estimates","\n"," lambda=",b,"  delta=",a,"\n",
  "[2] GEV ML estimates","\n"," gamma=",shapgev," lambda=",locgev,
  "  delta=",scalgev,"\n")
```

A.11 LRT-Block Maxima

```
library(evd)
loglikgumb<-fGumb$loglik
loglikgev<-as.vector(logLik(fgev(vo2max)))
LR<--2*(loglikgumb-loglikgev)
LRstar<-LR/(1+2.8/m)
pvalueLR<-pchisq(LRstar,1,lower.tail=F)
cat("[1] l=",LR,"  l*=",LRstar,"  p-value=",pvalueLR,"\n")
```

A.12 Rao's score test

```
zi<-(vo2max-b)/a
Vm<-sum(.5*zi^2-zi-.5*zi^2*exp(-zi))
Vm2<-Vm^2
```

```
Raonorm<-Vm/sqrt(2.09797*m)
Raochi<-Vm2/(2.09797*m)

pvalueraonorm<-pnorm(Raonorm)
pvalueraochi<-pchisq(Raochi,1,lower.tail=F)
cat("[1] Normal Test:  v_m=",Vm,"  v_m*=",Raonorm,"  p-value=",pvalueraonorm,"\n")
cat("[2] Chi-square Test:  v^2_m=",Vm2,"  v^2_m*=",Raochi,
  "  p-value=",pvalueraochi,"\n")
```

A.13 LAN test

```
s1m<-sum(-(vo2max-b)/a+.5*((vo2max-b)/a)^2*(1-exp(-(vo2max-b)/a)))
s2m<-sum(-1/a+(vo2max-b)/(a^2)*(1-exp(-(vo2max-b)/a)))
s3m<-sum(1/a-1/a*exp(-(vo2max-b)/a))

Tm<-1/3.451*(1.6449/sqrt(m)*s1m-a*.5066/sqrt(m)*s2m-a*.8916/sqrt(m)*s3m)
pvaluelan<-pnorm(Tm/.6904)
cat("[1] t_m=",Tm,"  t_m*=",Tm/.6904,"  p-value=",pvaluelan,"\n")
```

A.14 Goodness-of-fit tests-Block Maxima

```
gofstat(fGumb,print.test=F)
```

A.15 Gumbel fit diagnosis

```
plot(fGumb)
```

A.16 Gumbel PWM estimation

```
y <-vo2max
y<-sort(y); yy<-c();yyy<-c()
for(i in 1:m) {
yy[i]=(i-1)/(m-1)*y[i]
yyy[i]=(i-1)*(i-2)/((m-1)*(m-2))*y[i]
}
M100=mean(y); M110=mean(yy); M120=mean(yyy)
dgpwm<-(2*M110-M100)/log(2)
lgpwm<-M100+digamma(1)*dgpwm

cat(" [1] Gumbel PWM estimates","\n"," lambda=",lgpwm,"  delta=",dgpwm,"\n")
```

A.17 Profile likelihood CI's for Gumbel model

```
confint(profile(fgev(vo2max,shape=0)))

# profile likelihood plot
par(mfrow=c(1,2))
lCIgumb<-confint(profile(fgev(vo2max,shape=0),which="loc"))
linfgumb<-lCIgumb[1]
lsupgumb<-lCIgumb[2]
dCIgumb<-confint(profile(fgev(vo2max,shape=0),which="scale"))
dinfgumb<-dCIgumb[1]
dsupgumb<-dCIgumb[2]
plot(profile(fgev(vo2max,shape=0),which="loc"))
abline(v=linfgumb,col="blue")
abline(v=lsupgumb,col="blue")
abline(v=b,col="red")
plot(profile(fgev(vo2max,shape=0),which="scale"))
abline(v=dinfgumb,col="blue")
abline(v=dsupgumb,col="blue")
abline(v=a,col="red")
par(mfrow=c(1,1))
```

A.18 Exceedance probability for the Gumbel model

```
exceedmaxml<-pgumbel(96,b,a,lower.tail=F)
exceedmaxpwm<-pgumbel(96,lgpwm,dgpwm,lower.tail=F)
cat(" [1] Maximum Likelihood: P(Y>96)=",exceedmaxml,"\n",
  "[2] Probability Weighted Moments: P(Y>96)=",exceedmaxpwm ,"\n")
```

A.19 ML and PWM estimation of the GEVd

```
fGev<-fgev(vo2max)
shapgev<-as.vector(fGev$param[3])
locgev<-as.vector(fGev$param[1])
scalgev<-as.vector(fGev$param[2])

library(fExtremes)
pwmgev_fit<-gevFit(vo2max,type="pwm")
lpwm<-as.vector(pwmgev_fit@fit$par.ests[2])
dpwm<-as.vector(pwmgev_fit@fit$par.ests[3])
gpwm<-as.vector(pwmgev_fit@fit$par.ests[1])
cat("[1] GEV ML estimates","\n"," gamma=",shapgev," lambda=",locgev,
  "  delta=",scalgev,"\n")
cat("[2] GEV PWM estimates","\n"," gamma=",gpwm," lambda=",lpwm,"  delta=",dpwm,"\n")
```

A.20 GEVd fit diagnosis

```
dGev<-function(x,g,a,b){ exp(-(1+g*(x-b)/a)^(-1/g))* (1+g*(x-b)/a)^(-1/g-1)/a}
pGev<-function(q,g,a,b){exp(-(1+g*(q-b)/a)^(-1/g))}
qGev<-function(p,g,a,b){b+a*((-log(p))^(-g)-1)/g}
b_Gev=80; a_Gev=6;c_Gev=-.2
fGevml<- fitdist(vo2max,"Gev",start=list(a=a_Gev,b=b_Gev,g=c_Gev))
plot(fGevml)
```

A.21 Profile likelihood CI's for the GEVd

```
confint(profile(fgev(vo2max)))

# profile likelihood plot
par(mfrow=c(2,2))
lCI<-confint(profile(fGev,which="loc"))
linf<-lCI[1]
lsup<-lCI[2]
dCI<-confint(profile(fGev,which="scale"))
dinf<-dCI[1]
dsup<-dCI[2]
gCI<-confint(profile(fGev,which="shape"))
ginf<-gCI[1]
gsup<-gCI[2]
plot(profile(fGev,which="loc"))
abline(v=linf,col="blue")
abline(v=lsup,col="blue")
abline(v=locgev,col="red")
plot(profile(fGev,which="scale"))
abline(v=dinf,col="blue")
abline(v=dsup,col="blue")
abline(v=scalgev,col="red")
plot(profile(fGev,which="shape"))
abline(v=ginf,col="blue")
abline(v=gsup,col="blue")
abline(v=shapgev,col="red")
par(mfrow=c(1,1))
```

A.22 Exceedance probability for the GEVd

```
gevexceedmaxml<-1-pGev(96,shapgev,scalgev,locgev)
gevexceedmaxpwm<-1-pGev(96,gpwm,dpwm,lpwm)
cat(" [1] Maximum Likelihood: P(Y>96)=",gevexceedmaxml,"\n",
  "[2] Probability Weighted Moments: P(Y>96)=",gevexceedmaxpwm ,"\n")
```

A.23 Endpoint estimation-Block Maxima

```
xF_ml<-locgev-scalgev/shapgev
xF_pwm<-lpwm-dpwm/gpwm
cat(" [1] Maximum Likelihood: xˆF=",xF_ml,"\n",
  "[2] Probability Weighted Moments: xˆF=",xF_pwm ,"\n")
```

A.24 Exponential QQ-plot

```
excess<-vo2max[which(vo2max>80)]-80
me<-length(excess)
ie<-c(1:me)
fun<-function(x) -log(1-x)
Qe<-fun(ie/(me+1))
plot(Qe,excess,pch=19,xlab=expression(-log(1-p[i])),ylab=expression(y[i:m]))
grid()
```

A.25 Correlation for the GPd QQ-plot

```
correlgpd<-function(g){
x<-((1-ie/(me+1))ˆ(-g)-1)/g
cor(excess,x)
}
(ggpdopt<-optimize(correlgpd,interval=c(-1,.5),maximum=T))

# Correlation plot
seqgpdg<-seq(-1.5,.5,.01)
correl2<-sapply(seqgpdg,correlgpd)
correl2[151]<-cor(Qe,excess)
plot(seqgpdg,correl2,type="l",xlab=expression(gamma),
  ylab=expression(corr(Q[list(gamma,1)](p[i]),y[i:m])),mgp=c(2.2,1,0))
grid()
abline(v=ggpdopt$maximum,lty=2,col="red")
abline(h=ggpdopt$objective,lty=2,col="blue")
```

A.26 GPd QQ-plot

```
Qgpd<-((1-ie/(me+1))ˆ(-ggpdopt$maximum)-1)/(ggpdopt$maximum)
plot(Qgpd,excess,xlab=expression(((1-p[i])ˆ-hat(gamma)-1)/hat(gamma)),
  ylab=expression(y[i:m]),pch=19)
grid()
```

A.27 Linear fit for the GPd QQ-plot

```
(linereggpd<-lm(excess~Qgpd-1))
abline(linereggpd,col="blue",lwd=2)
```

A.28 Gomes and van Monfort (1986) test

```
exc<-vo2max[which(vo2max>80)]
Gm<-exc[me]/exc[floor(me/2)+1]
Gmstar<-log(2)*Gm-log(me)
pvaluniGmstar<-pgumbel(Gmstar)
cat("[1] g_m=",Gm,"  g_m*=",Gmstar,"  p-value=",pvaluniGmstar,"\n")
```

A.29 Marohn (2000) test-POT

```
T_m<-0.5*((var(exc)*(me-1)/me)/(mean(exc)-80)^2-1)
T_mstar<-sqrt(me)*T_m
pvaluniTmstar<-pnorm(T_mstar)
cat("[1] One-sided Test","\n","t_m=",T_m,"  t_m*=",T_mstar,
  "  p-value=",pvaluniTmstar,"\n")
```

A.30 Exponential and GPd ML estimation

```
a_exp<-as.vector(1/linereggpd$coefficients)
fexp<-fitdist(excess,"exp",start=list(rate=a_exp))
aexp<-as.vector(1/fexp$estimate)

library(ismev)
fgpd<-gpd.fit(vo2max,threshold=80,show=F)
shapgpd<-as.vector(fgpd$mle[2])
scalgpd<-as.vector(fgpd$mle[1])
cat(" [1] Exponential ML estimates","\n"," sigma_u=",aexp,"\n",
  "[2] GPd ML estimates","\n"," gamma=",shapgpd,"  sigma_u=",scalgpd,"\n")
```

A.31 LRT-POT

```
loglikexp<-fexp$loglik
loglikgpd<--fgpd$nllh
LRgpd<--2*(loglikexp-loglikgpd)
LRgpdstar<-LRgpd/(1+4/me)
pvalueLRgpd<-pchisq(LRgpdstar,1,lower.tail=F)
cat("[1] l=",LRgpd,"  l*=",LRgpdstar,"  p-value=",pvalueLRgpd,"\n")
```

A.32 Kolmogorov-Smirnov test-POT

```
KS<-max(max(abs(pexp(excess,rate=1/aexp)-ie/me),
  abs(pexp(excess,rate=1/aexp)-(ie-1)/me)))
cat("Kolmogorov-Smirnov statistic:  ",KS,"\n")
```

A.33 Cramér-von Mises and Anderson-Darling tests-POT

```
pGP<-function(x,g,a){1-(1+g*x/a)^(-1/g)}
CVM<-sum((pGP(excess,shapgpd,scalgpd)-(2*ie-1)/(2*me))^2)+1/(12*me)
AD<--me-1/me*sum((2*ie-1)*log(pGP(excess,shapgpd,scalgpd))+
  (2*me+1-2*ie)*log(1-pGP(excess,shapgpd,scalgpd)))
cat(" Cramer-von Mises statistic:  ",CVM,"\n",
  "Anderson-Darling statistic:  ",AD,"\n")
```

A.34 GPd PWM estimation

```
fgpdpwm<-gpd(vo2max,threshold=80,method="pwm")
shapgpdpwm<-as.vector(fgpdpwm$par.ests[1])
scalgpdpwm<-as.vector(fgpdpwm$par.ests[2])
cat(" [1] GPd PWM estimates","\n","gamma=",shapgpdpwm,"  sigma_u=",scalgpdpwm,"\n")
```

A.35 GPd ML and PWM fit diagnosis

```
library(POT)
# ML fit diagnosis
fitml<-fitgpd(vo2max,threshold=80)
par(mfrow=c(2,2))
plot(fitml,which=1:3)
par(mfrow=c(1,1))

# PWM fit diagnosis
fitpwm<-fitgpd(vo2max,threshold=80,est="pwmu")
par(mfrow=c(2,2))
plot(fitpwm,which=1:3)
par(mfrow=c(1,1))
```

A.36 Profile likelihood CI's for the GPd

```
gpd.profxi(fgpd,xlow=-.9,xup=-.1,nint=8000)
```

A.37 Exceedance probability and endpoint estimation for the GPd

```
# exceedance probability
exceedmaxpotml<-me/m*(1-pGP(16,shapgpd,scalgpd))
exceedmaxpotpwm<-me/m*(1-pGP(16,shapgpdpwm,scalgpdpwm))
cat(" [1] Maximum Likelihood: P(X>96)=",exceedmaxpotml,"\n",
  "[2] Probability Weighted Moments: P(X>96)=",exceedmaxpotpwm ,"\n")

# endpoint estimation
xF_potml<-80-scalgpd/shapgpd
xF_potpwm<-80-scalgpdpwm/shapgpdpwm
cat(" [1] Maximum Likelihood: x^F=",xF_potml,"\n",
  "[2] Probability Weighted Moments: x^F=",xF_potpwm ,"\n")
```

A.38 Greenwood, Hasofer-Wang and Ratio sample paths (two-sided)

```
n<-length(vo2max)

Mj<-function(k,r) {
y<-NULL
for(i in 1:k) {
y[i]<-(vo2max[n-i+1]-vo2max[n-k])^r
}
(1/k)*sum(y)
}



Rstar<-function(k){
R<-Mj(k,2)/(Mj(k,1))^2
stat<-sqrt(k/4)*(R-2)
}

Wstar<-function(k){
R<-Mj(k,2)/(Mj(k,1))^2
W<-1/(k*(R-1))
stat<-sqrt(k/4)*(k*W-1)
}

Tstar<-function(k){
T<-(vo2max[n]-vo2max[n-k])/Mj(k,1)
stat<-T-log(k)
}

x<-seq(2,n-1,1)
```

```
Rstarobs<-sapply(x,Rstar)
Wstarobs<-sapply(x,Wstar)
Tstarobs<-sapply(x,Tstar)
plot(Rstarobs,type="l",ylim=c(-2,6),xlab="k",ylab="observed statistic",
  cex=0.8,col="darkgreen")
grid()
lines(Wstarobs,col="red");lines(Tstarobs,col="blue");

abline(h=qnorm(.025),lty=2,lwd=2) ; abline(h=qnorm(.975),lty=2,lwd=2)
library(evd)
abline(h=qgumbel(.025),lty=2,lwd=2,col="blue")
abline(h=qgumbel(.975),lty=2,lwd=2,col="blue")

Text<-c(expression(R[n]^paste("*")~(k)),expression(W[n]^paste("*")~(k)),
  expression(T[n]^paste("*")~(k)),expression(list(z[.975],z[.025])),
  expression(list(g[.975],g[.025])))
legend(40,1.5,legend=Text,col=c("darkgreen","red","blue","black","blue"),
  lty=c(1,1,1,2,2),lwd=c(1,1,1,2,2),cex=0.8)
```

A.39 Greenwood, Hasofer-Wang and Ratio sample paths (one-sided)

```
plot(Rstarobs,type="l",ylim=c(-2,6),xlab="k",ylab="observed statistic",
  cex=0.8,col="darkgreen")
grid()
lines(Wstarobs,col="red");lines(x,Tstarobs,col="blue");

abline(h=qnorm(.05),col="darkgreen",lty=2,lwd=2)
abline(h=qnorm(.95),lty=2,lwd=2,col="red")
abline(h=qgumbel(.05),lty=2,col="blue",lwd=2)

Text<-c(expression(R[n]^paste("*")~(k)),expression(W[n]^paste("*")~(k)),
  expression(T[n]^paste("*")~(k)),expression(z[.05]),expression(z[.95]),
  expression(g[.05]))
legend(50,1.5,legend=Text,col=c("darkgreen","red","blue","darkgreen","red","blue"),
  lty=c(1,1,1,2,2,2),lwd=c(1,1,1,2,2,2),cex=0.8)
```

A.40 Pickands plot

```
gammahatp<-function(k){
1/log(2)*log((vo2max[n-floor((k+1)/4)+1]-vo2max[n-2*floor((k+1)/4)+1])/
  (vo2max[n-2*floor((k+1)/4)+1]-vo2max[n-4*floor((k+1)/4)+1]))
}
```

```
xp<-seq(1,n,1)
Pest<-sapply(xp,gammahatp)
plot(Pest,type="l",xlab="k",ylab=expression(hat(gamma)[list(n,k)]^P),mgp=c(2.2,1,0))
grid()
abline(h=0)
```

A.41 Negative Hill plot

```
gammahatnh<-function(k) {
ind<-NULL
ind<-seq(1,k-1,1)
f1<-function(x) log(vo2max[n]-vo2max[n-x])-log(vo2max[n]-vo2max[n-k])
f2<-sapply(ind,f1)
1/k*sum(f2)
}

x<-seq(2,n-1,1)
NHest<-sapply(x,gammahatnh)

plot(x,NHest,type="l",xlab="k",ylab=expression(hat(gamma)[list(n,k)]^NH),
  mgp=c(2.2,1,0))
grid()
```

A.42 PORT estimators sample paths

```
Xistar<-function(x){
y<-NULL
nq<-floor(n*x)+1
for(i in 1:n){
y[i]<-vo2max[i]-vo2max[nq]
}
return(y)
}

# PORT Moment
Mnrq<-function(k,r,q) {
y<-NULL
for(j in 1:k) {
y[j]<-(log(Xistar(q)[n-j+1])-log(Xistar(q)[n-k]))^r
}
(1/k)*sum(y)

}
```

```
gammahatmPORT<-function(k,q) Mnrq(k,1,q)+1-0.5*(1-(Mnrq(k,1,q))^2/Mnrq(k,2,q))^-1

PORTMest<-function(q){
nq<-floor(n*q)+1
x<-seq(1,n-nq-1,1)
sapply(x,gammahatmPORT,q=q)
}

# PORT Mixed Moment
Lnrq<-function(k,r,q) {
y<-NULL
for(j in 1:k) {
y[j]<-(1-(Xistar(q)[n-k]/Xistar(q)[n-j+1]))^r
}
(1/k)*sum(y)
}

phihatnkq<-function(k,q) (Mnrq(k,1,q)-Lnrq(k,1,q))/(Lnrq(k,1,q))^2

gammahatmmPORT<-function(k,q) (phihatnkq(k,q)-1)/(1+2*min(phihatnkq(k,q)-1,0))

PORTMMest<-function(q){
nq<-floor(n*q)+1
x<-seq(1,n-nq-1,1)
return(sapply(x,gammahatmmPORT,q=q))
}

# PORT sample paths
q<-c(0,0.25,0.5)
PORTMestfin<-sapply(q,PORTMest)
PORTMMestfin<-sapply(q,PORTMMest)

par(mfrow=c(1,2))
plot(PORTMestfin[[1]],type="l",xlab="k",xlim=c(0,60),ylim=c(-3,0),
  ylab=expression(hat(gamma)[list(n,k)]^M(q)),
  main="PORT Moment estimator sample path",mgp=c(2.2,1,0),lwd=2)
lines(PORTMestfin[[2]],col="red")
lines(PORTMestfin[[3]],col="blue")
abline(h=0)
legend(15,-2.5,legend=c("q=0","q=0.25","q=0.5"),lty=c(1,1,1),
  col=c("black","red","blue"),cex=0.8,lwd=c(2,1,1))

plot(PORTMMestfin[[1]],type="l",xlab="k",xlim=c(0,60),ylim=c(-2,1),
  ylab=expression(hat(gamma)[list(n,k)]^MM(q)),
  main="PORT Mixed Moment estimator sample path", mgp=c(2.2,1,0),lwd=2)
```

```
lines(PORTMMestfin[[2]],col="red")
lines(PORTMMestfin[[3]],col="blue")
legend(15,-1,legend=c("q=0","q=0.25","q=0.5"),lty=c(1,1,1),
  col=c("black","red","blue"),cex=0.8,lwd=c(2,1,1))
abline(h=0)
par(mfrow=c(1,1))
```

A.43 Semi-parametric estimators plot

```
# Moment
Mnr<-function(k,r) {
y<-NULL
for(i in 1:k) {
y[i]<-(log(vo2max[n-i+1])-log(vo2max[n-k]))^r
}
(1/k)*sum(y)

}

gammahatm<-function(k) Mnr(k,1)+1-0.5*(1-(Mnr(k,1))^2/Mnr(k,2))^-1

# Generalized Hill
gammahath<-function(k) {
ind<-NULL
ind<-seq(1,k,1)
f1<-function(x) log(vo2max[n-x+1])-log(vo2max[n-k])
f2<-sapply(ind,f1)
1/k*sum(f2)
}

gammahatgh<-function(k){

ind<-NULL
ind<-seq(1,k,1)
f1<-function(x) log(gammahath(x))-log(gammahath(k))
f2<-sapply(ind,f1)
gammahath(k)+(1/k)*sum(f2)
}

# Mixed Moment
Lnr<-function(k,r) {
y<-NULL
for(i in 1:k) {
y[i]<-(1-(vo2max[n-k]/vo2max[n-i+1]))^r
}
```

```
(1/k)*sum(y)

}

phihatnk<-function(k) (Mnr(k,1)-Lnr(k,1))/(Lnr(k,1))^2

gammahatmm<-function(k) (phihatnk(k)-1)/(1+2*min(phihatnk(k)-1,0))

# plot
x<-seq(1,n-1,1)
Mest<-sapply(x,gammahatm)
GHest<-sapply(x,gammahatgh)
MMest<-sapply(x,gammahatmm)
plot(x,Mest,type="l",xlab="k",ylim=c(-2,0),ylab=expression(hat(gamma)[list(n,k)]^E),
  mgp=c(2.2,1,0),lwd=1)
abline(h=0)
lines(GHest,col="red")
lines(MMest,col="blue")
lines(PORTMest(.01),col="darkgreen")
lines(PORTMMest(.01),lwd=2,col="gray")
lines(Pest,lwd=2,col="violet")
legend(35,-1.25,c("Moment","Generalized Hill","Mixed Moment",
  "PORT-Moment (q=0.01)","PORT-Mixed Moment(q=0.01)","Pickands"),lwd=c(1,1,1,1,2,2),
  col=c("black","red","blue","darkgreen","gray","violet"),cex=0.9)
grid()
```

A.44 k heuristic choice

```
# for equal vector length
Mest1<-Mest[-length(Mest)]
GHest1<-GHest[-length(GHest)]
MMest1<-MMest[-length(MMest)]

sqdifest<-(Mest1-GHest1)^2+(Mest1-MMest1)^2+(Mest1-PORTMest1)^2+(GHest1-MMest1)^2+
  (GHest1-PORTMest1)^2+(MMest1-PORTMest1)^2
kopt<-which.min(sqdifest)
cat("[1] k opt=",kopt,"\n")
```

A.45 Distance function plot

```
plot(sqdifest,type="l",ylim=c(0,.6),xlab="k",
  ylab=expression(sum((hat(gamma)[list(n,k)]^(i)-hat(gamma)[list(n,k)]^(j))^2)),
  mgp=c(2.2,1,0))
grid()
```

```
abline(h=sqdifest[kopt],lty=2,col="blue"); abline(v=kopt,lty=2,col="red")
```

A.46 Second k heuristic choice

```
kopt2<-which.min(sqdifest[35:45])+34
cat("[1] k opt=",kopt2,"\n")
```

A.47 EVI semi-parametric estimation

```
# k=19
gmk1<-gammahatm(kopt)
gghk1<-gammahatgh(kopt)
gmmk1<-gammahatmm(kopt)
gmpk1<-gammahatmPORT(kopt,0.01)

# k=43
gmk2<-gammahatm(kopt2)
gghk2<-gammahatgh(kopt2)
gmmk2<-gammahatmm(kopt2)
gmpk2<-gammahatmPORT(kopt2,0.01)

cat("[1] k opt=19","\n","Moment: gamma=",gmk1,"   Generalized Hill: gamma=",
  gghk1,"\n","Mixed Moment: gamma=",gmmk1," PORT-Moment (q=0.01): gamma=",
  gmpk1,"\n")
cat(" [2] k opt=43","\n"," Moment: gamma=",gmk2,"   Generalized Hill: gamma=",
  gghk2,"\n"," Mixed Moment: gamma=",gmmk2,"  PORT-Moment (q=0.01): gamma=",
  gmpk2,"\n")
```

A.48 Semi-parametric asymptotic variances

```
## k=19
# Moment
s2mk1<-(1-gmk1)^2*(1-2*gmk1)*(1-gmk1+6*gmk1^2)/((1-3*gmk1)*(1-4*gmk1))
# Generalized Hill
s2ghk1<-(1-gghk1)*(1+gghk1+2*gghk1^2)/(1-2*gghk1)
# Mixed Moment
s2mmk1<-(1-2*gmmk1)^4*(1-gmmk1)^2*(6*gmmk1^2-gmmk1+1)/
  ((1-2*gmmk1)^3*(1-3*gmmk1)*(1-4*gmmk1))
# PORT-Moment
s2mpk1<-s2mk1

## k=43
# Moment
```

```
s2mk2<-(1-gmk2)^2*(1-2*gmk2)*(1-gmk2+6*gmk2^2)/((1-3*gmk2)*(1-4*gmk2))
# Generalized Hill
s2ghk2<-(1-gghk2)*(1+gghk2+2*gghk2^2)/(1-2*gghk2)
# Mixed Moment
s2mmk2<-(1-2*gmmk2)^4*(1-gmmk2)^2*(6*gmmk2^2-gmmk2+1)/
  ((1-2*gmmk2)^3*(1-3*gmmk2)*(1-4*gmmk2))
# PORT-Moment
s2mpk2<-s2mk2

cat("[1] k opt=19","\n","Moment: s2_M=",s2mk1,"   Generalized Hill: s2_GH=",
  s2ghk1,"\n","Mixed Moment: s2_MM=",s2mmk1," PORT-Moment (q=0.01): s2_M(q)=",
  s2mpk1,"\n")
cat(" [2] k opt=43","\n"," Moment: s2_M=",s2mk2,"   Generalized Hill: s2_GH=",
  s2ghk2,"\n"," Mixed Moment: s2_MM=",s2mmk2,"  PORT-Moment (q=0.01): s2_M(q)=",
  s2mpk2,"\n")
```

A.49 Location and scale coefficients semi-parametric estimates

```
# location
bnk1<-vo2max[n-kopt]
bnk2<-vo2max[n-kopt2]
cat(" [1] k opt=19","\n","b(n/k)=",bnk1,"\n","[2] k opt=43","\n"," b(n/k)=",
  bnk2,"\n")
# scale
negmom<-function(k) 1-0.5*(1-(Mnr(k,1))^2/Mnr(k,2))^-1
ank<-function(k) vo2max[n-k]*Mnr(k,1)*(1-negmom(k))
ank1<-ank(19)
ank2<-ank(43)
cat(" [1] k opt=19","\n","a(n/k)=",ank1,"\n","[2] k opt=43","\n"," a(n/k)=",
  ank2,"\n")
```

A.50 Endpoint semi-parametric estimation

```
endpm<-function(k) max(max(vo2max),vo2max[n-k]-ank(k)/gammahatm(k))
endpgh<-function(k) max(max(vo2max),vo2max[n-k]-ank(k)/gammahatgh(k))
endpmm<-function(k) max(max(vo2max),vo2max[n-k]-ank(k)/gammahatmm(k))
endpmp<-function(k) max(max(vo2max),vo2max[n-k]-ank(k)/gammahatmPORT(k,0.01))

cat("[1] k opt=19","\n","Moment: xF=",endpm(19),"   Generalized Hill: xF=",
  endpgh(19),"\n","Mixed Moment: xF=",endpmm(19),
  " PORT-Moment (q=0.01): xF=",endpmp(19),"\n")
cat(" [2] k opt=43","\n","Moment: xF=",endpm(43),"   Generalized Hill: xF=",
  endpgh(43),"\n","Mixed Moment: xF=",endpmm(43),
  " PORT-Moment (q=0.01): xF=",endpmp(43),"\n")
```

A.51 Endpoint semi-parametric estimators sample paths

```
x<-seq(1,n-1,1)
Mestendp<-sapply(x,endpm)
GHestendp<-sapply(x,endpgh)
MMestendp<-sapply(x,endpmm)
MPestendp<-sapply(x,endpmp)

# for equal vector length
Mestendp1<-Mestendp[-length(Mestendp)]
GHestendp1<-GHestendp[-length(GHestendp)]
MMestendp1<-MMestendp[-length(MMestendp)]
MPestendp1<-MPestendp[-length(MPestendp)]



plot(x,Mestendp,type="l",ylim=c(90,115),xlab="k",
  ylab=expression(hat(x)^F), mgp=c(2.2,1,0))
grid()
lines(GHestendp,col="red")
lines(MMestendp,col="blue")
lines(MPestendp,col="darkgreen")
legend(40,115,c("Moment","Generalized Hill","Mixed Moment",
  "PORT-Moment (q=0.01)"),lty=c(1,1,1,1),
  col=c("black","red","blue","darkgreen"))
```

A.52 Heuristic procedure for endpoint estimate

```
sqdifestendp<-(Mestendp1-GHestendp1)^2+(Mestendp1-MMestendp1)^2+
  (Mestendp1-MPestendp1)^2+(GHestendp1-MMestendp1)^2+
  (GHestendp1-MPestendp1)^2+(MMestendp1-MPestendp1)^2
koptendp<-which.min(sqdifestendp)
cat("[1] k opt=",koptendp,"\n")

plot(sqdifestendp,type="l",ylim=c(0,100),xlab="k",
  ylab=expression(sum((hat(x)^F(i)-hat(x)^F(j)))^2),mgp=c(2.2,1,0))
grid()
abline(h=sqdifestendp[koptendp],lty=2,col="blue")
abline(v=koptendp,lty=2,col="red")
```

A.53 Semi-parametric CI's for endpoint

```
xFCIsupM19<-endpm(19)+pnorm(.95)*ank1/(gmk1)^2*sqrt(s2mk1/19)
xFCIsupM43<-endpm(43)+pnorm(.95)*ank2/(gmk2)^2*sqrt(s2mk2/43)
xFCIsupGH19<-endpgh(19)+pnorm(.95)*ank1/(gghk1)^2*sqrt(s2ghk1/19)
```

```
xFCIsupGH43<-endpgh(43)+pnorm(.95)*ank2/(gghk2)^2*sqrt(s2ghk2/43)
xFCIsupMM19<-endpmm(19)+pnorm(.95)*ank1/(gmmk1)^2*sqrt(s2mmk1/19)
xFCIsupMM43<-endpmm(43)+pnorm(.95)*ank2/(gmmk2)^2*sqrt(s2mmk2/43)
xFCIsupPM19<-endpmp(19)+pnorm(.95)*ank1/(gmpk1)^2*sqrt(s2mpk1/19)
xFCIsupPM43<-endpmp(43)+pnorm(.95)*ank2/(gmpk2)^2*sqrt(s2mpk2/43)
```

A.54 Semi-parametric exceedance probability

```
excedprobM19<-19/n*(1-pGP(96-bnk1,gmk1,ank1))
excedprobM43<-43/n*(1-pGP(96-bnk2,gmk2,ank2))
excedprobGH19<-19/n*(1-pGP(96-bnk1,gghk1,ank1))
excedprobGH43<-43/n*(1-pGP(96-bnk2,gghk2,ank2))
excedprobMM19<-19/n*(1-pGP(96-bnk1,gmmk1,ank1))
excedprobMM43<-43/n*(1-pGP(96-bnk2,gmmk2,ank2))
excedprobPM19<-19/n*(1-pGP(96-bnk1,gmpk1,ank1))
excedprobPM43<-43/n*(1-pGP(96-bnk2,gmpk2,ank2))
```

Appendix B

References

Bibliography

Araújo Santos, P., Fraga Alves, M. I., Gomes, M. I., 2006. Peaks Over Random Threshold Methodology for Tail Index and High Quantile Estimation. *Revstat* **4** (3), 227–247.

Åstrand, P.-O., Ryhming, I., 1954. A nomogram for calculation of aerobic capacity (physical fitness) from pulse rate during submaximal work. *J. Appl. Physiol.* **7** (2), 218–221.

Balkema, A. A., de Haan, L., 1974. Residual life time at great age. *Ann. Probab.* **2** (5), 792–804.

Bangsbo, J., Larsen, H. B., 2001. *Running & Science - in an Interdisciplinary Perspective*. Institute of Exercise and Sport Sciences, University of Copenhagen.

Bassett, D. R. J., Howley, E. T., 2000. Limiting factors for maximum oxygen uptake and determinants of endurance performance. *Med. Sci. Sport. Exer.* **32** (1), 70–84.

Beirlant, J., Dierckx, G., Guillou, A., 2005. Estimation of the extreme-value index and generalized quantile plots. *Bernoulli* **11** (6), 949–970.

Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., 2004. *Statistics of Extremes: Theory and Applications*. Wiley, England.

Beirlant, J., Vynckier, P., Teugels, J., 1996. Excess functions and estimation of the extreme-value index. *Bernoulli* **2** (4), 293–318.

Caeiro, F., Gomes, M. I., 2006. A new class of estimators of a "scale" second order parameter. *Extremes* **9** (3-4), 193–211.

Cerretelli, P., Di Prampero, P. E., 1987. *Gas exchange in exercise*. Vol. 1. American Physiological Society, Bethesda, Maryland, Ch. Supplement 13: Handbook of Physiology, The Respiratory System, Gas Exchange, pp. 297–339.

Chandra, M., Singpurwalla, N. D., Stephens, M. A., 1981. Kolmogorov Statistics for Tests of Fit for the Extreme Value and Weibull Distributions. *J. Amer. Statist. Assoc.* **76** (375), 729–731.

Choulakian, V., Stephens, M. A., 2001. Goodness-of-fit tests for the generalized Pareto distribution. *Technometrics* **43** (4), 478–484.

Cox, D. R., Hinkley, D. V., 1974. *Theoretical Statistics*. Chapman and Hall, London.

Davison, A. C., 1984. *Statistical Extremes and Applications*. D. Reidel, Dordrecht, Holland, Ch. Modelling excesses over high thresholds, pp. 461–482.

Davison, A. C., Smith, R. L., 1990. Models for exceedances over high thresholds. *J. Roy. Statist. Soc. Ser. B* **52** (3), 393–442.

de Haan, L., 1970. *On Regular Variation and its Applications to the Weak Convergence of Sample Extremes*. Mathematical Centre Tract,32, Amsterdam.

de Haan, L., 1976. Sample extremes: an elementary introduction. *Stat. Neerl.* **30**, 161–172.

de Haan, L., 1984. *Statistical Extremes and Applications*. D. Reidel, Dordrecht, Holland, Ch. Slow Variation and Characterization of Domains of Attraction, pp. 31–48.

de Haan, L., Ferreira, A., 2006. *Extreme Value Theory - An Introduction*. Springer, New York.

Dekkers, A. L. M., de Haan, L., 1989. On the Estimation of the Extreme-Value Index and Large Quantile Estimation. *Ann. Stat.* **17** (4), 1795 – 1832.

Dekkers, A. L. M., Einmahl, J. H. J., de Haan, L., 1989. A Moment Estimator for the Index of an Extreme-Value Distribution. *Ann. Stat.* **17** (4), 1833–1855.

Drees, H., 1995. Refined Pickands estimators of the extreme value index. *Ann. Stat.* **23** (6), 2059–2080.

Drees, H., de Haan, L., Resnick, S., 2000. How to Make a Hill Plot. *Ann. Stat.* **28** (1), 254–274.

Einmahl, J. H. J., Magnus, J. R., 2008. Records in athletics through extreme-value theory. *J. Am. Stat. Assoc.* **103** (484), 1382–1391.

Einmahl, J. H. J., Smeets, S. G. W. R., 2011. Ultimate 100-m world records through extreme-value theory. *Stat. Neerl.* **65** (1), 32–42.

Embrechts, P., Klüppelberg, C., Mikosch, T., 1997. *Modelling Extremal Events for Insurance and Finance*, 1st Edition. Springer, Berlin.

Falk, M., 1995. Some best estimators for distributions with finite endpoint. *Statistics* **27** (1-2), 115–125.

Ferreira, A., de Haan, L., Peng, L., 2003. On optimizing the estimation of high quantiles of a probability distribution. *Statistics* **37** (5), 401–434.

Fisher, R. A., Tippett, L. H. C., 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proc. Camb. Phil. Soc.* **24**, 180–190.

Fraga Alves, M. I., 1995. Estimation of the tail parameter in the domain of attraction of an extremal distribution. *J. Stat. Plan. Infer.* **45** (1-2), 143–173.

Fraga Alves, M. I., 2001. A location invariant Hill-type estimator. *Extremes* **4**, 199–217.

Fraga Alves, M. I., de Haan, L., Lin, T., 2003. Estimation of the parameter controlling the speed of convergence in extreme value theory. *Math. Methods Statist.* **12** (2), 155–176.

Fraga Alves, M. I., Gomes, M. I., de Haan, L., Neves, C., 2009. The mixed moment estimator and location invariant alternatives. *Extremes* **12**, 14–185.

Gnedenko, B., 1943. Sur La Distribution Limite Du Terme Maximum D'Une Série Aléatoire. *Ann. Math.* **44** (3), 423–453.

Gomes, M. I., de Haan, L., Peng, L., 2002. Semi-parametric estimation of the second order parameter – asymptotic and finite sample behaviour. *Extremes* **5** (4), 387–414.

Gomes, M. I., Fraga Alves, M. I., Araújo Santos, P., 2007. *PORT Hill and Moment Estimators for Heavy-Tailed Models*. Tech. Rep. 15/07, CEAUL.

Gomes, M. I., Martins, M. J., 2002. Asymptotically unbiased estimators of the tail index based on external estimation of the second order parameter. *Extremes* **5** (1), 5–31.

Gomes, M. I., Oliveira, O., 2001. The bootstrap methodology in statistical extremes – the choice of the optimal sample fraction. *Extremes* **4** (4), 351-358.

Gomes, M. I., van Monfort, M. A. J., 1986. Exponentiality versus Generalized Pareto, quick tests. In: *Proc. III Internat. Conf. Statistical Climatology*. pp. 185–195.

Henriques-Rodrigues, L., Gomes, M. I., Pestana, D., 2011. Statistics in Athletics. *Revstat* **9** (2), 127–153.

Hill, A. V., Long, C. N. H., Lupton, H., 1924. Muscular exercise, lactic acid and the supply and utilisation of oxygen: Parts VII-VIII. *P. R. Soc. B* **97** (682), 155–176.

Hill, A. V., Lupton, H., 1923. Muscular exercise, lactic acid, and the supply and utilization of oxygen. *QJM-Int. J. Med.* **16** (62), 135–171.

Hill, B., 1975. A simple general approach to inference about the tail of a distribution. *Ann. Stat.* **3** (5), 1163–1174.

Hosking, J. R. M., 1984. Testing whether the shape parameter is zero in the generalized extreme value distribution. *Biometrika* **71** (2), 367–374.

Hosking, J. R. M., 1985. Algorithm AS 215: Maximum likelihood estimation of the parameters of the generalized extreme value distribution. *Appl. Statist.* **34**, 301–310.

Hosking, J. R. M., Wallis, J. R., 1987. Parameter and Quantile Estimation for the Generalized Pareto Distribution. *Technometrics* **29** (3), 339–349.

Hosking, J. R. M., Wallis, J. R., Wood, E. F., 1985. Estimation of the Generalized Extreme-Value Distribution by the Method of Probability-Weighted Moments. *Technometrics* **27** (3), 251–261.

Jenkinson, A. F., 1955. The frequency distribution of the annual maximum (or minimum) values of metereological elements. *Q. J. Roy. Meteor. Soc.* **81**, 158–171.

Kozubowski, T. J., Panorska, A. K., Qeadan, F., Gershunov, A., Rominger, D., 2009. Testing exponentiality versus Pareto distribution via likelihood ratio. *Commun. Stat.-Simul. C.* **38** (1), 118–139.

Kravitz, L., Dalleck, L. C., 2002. The physiological limitations to endurance exercise capacity. *IDEA Health & Fitness Source* **20** (4), 40–49.

Landwehr, J. M., Matalas, N. C., Wallis, J. R., 1979. Probability weighted moments compared with some traditional techniques in estimating Gumbel parameters and quantiles. *Water Resour. Res.* **15**, 1055–1064.

Lilliefors, H. W., 1969. On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown. *J. Amer. Statist. Assoc.* **64** (325), 387–389.

Macleod, A. J., 1989. A remark on the algorithm AS 215: Maximum likelihood estimation of the parameters of the generalized extreme value distribution. *Appl. Statist.* **38**, 198–199.

Marohn, F., 2000. Testing extreme value models. *Extremes* **3** (4), 363–384.

McArdle, W. D., Katch, F. I., Katch, V. L., 2009. *Exercise Physiology: Nutrition, Energy and Human Performance*, 7th Edition. Lippincott Williams & Wilkins.

Neves, C., Fraga Alves, M. I., 2007. Semi-parametric approach to the Hasofer-Wang and Greenwood statistics in extremes. *Test* **16**, 297–313.

Neves, C., Fraga Alves, M. I., 2008. Testing extreme value conditions – An overview and recent approaches. *Revstat* **6** (1), 83–100.

Neves, C., Picek, J., Fraga Alves, M. I., 2006. The contribution of the maximum to the sum of excesses for testing max-domains of attraction. *J. Stat. Plan. Infer.* **136** (4), 1281–1301.

Noakes, T., 2003. *Lore of Running*, 4th Edition. Human Kinetics Publisher.

Peng, L., 1998. Asymptotically unbiased estimators for extreme value index. *Stat. Probabil. Lett.* **38**, 107–115.

Pereira, T. T., 1994. Second order behaviour of domains of attraction and the bias of generalized Pickands' estimator. In: eds J. Galambos, J. L., Simiu, E. (Eds.), *Extreme Value Theory and Applications III*. Proceedings of the Gaithersburg Conference. NIST, pp. 165–177.

Pickands III, J., 1975. Statistical inference using extreme order statistics. *Ann. Stat.* **3** (1), 119–131.

Prescott, P., Walden, A. T., 1980. Maximum likelihood estimation of the parameters of the generalized extreme-value distribution. *Biometrika* **67**, 723–724.

Prescott, P., Walden, A. T., 1983. Maximum likelihood estimation of the parameters of the three-parameter generalized extreme-value distribution from censored samples. *J. Stat. Comput. Sim.* **16**, 241–250.

Reiss, R.-D., Thomas, M., 2007. *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields*, 3rd Edition. Birkhäuser Verlag, Basel-Boston-Berlin.

Saltin, B., Åstrand, P.-O., 1967. Maximal oxygen uptake in athletes. *J. Appl. Physiol.* **23** (3), 353–358.

Segers, J., 2005. Generalized Pickands Estimators for the Extreme Value Index. *J. Stat. Plan. Infer.* **128** (2), 381–396.

Smith, R. L., 1985. Maximum likelihood estimation in a class of nonregular cases. *Biometrika* **72** (1), 67–90.

Smith, R. L., 1987. Estimating Tails of Probability Distributions. *Ann. Stat.* **15** (3), 1174–1207.

Stephens, M. A., 1976. Asymptotic results for goodness-of-fit statistics with unknown parameters. *Ann. Stat.* **4**, 357–369.

Stephens, M. A., 1977. Goodness-of-fit for the extreme value distribution. *Biometrika* **64** (3), 583–588.

Stephens, M. A., 1986. *Goodness-of-Fit Techniques*. Vol. 68. Marcel Dekker, Inc.

Tiago de Oliveira, J., 1981. *Statistical distributions in Scientific Work*. Vol. 6. D.Reidel, Dordrecht, Ch. Statistical choice of univariate extreme models, pp. 367–387.

Tiago de Oliveira, J., Gomes, M. I., 1984. *Statistical Extremes and Applications*. D. Reidel, Dordrecht, Holland, Ch. Two test statistics for choice of univariate extreme models, pp. 651–668.

Van Montfort, M. A. J., 1970. On testing that the distribution of extremes is of type I when type II is the alternative. *J. Hydrol.* **11**, 421–427.

von Mises, R., 1936. La distribution de la plus grande de n valeurs. Reprinted in Selected Papers Volumen II, *Amer. Math. Soc.*, Providence, R.I., 271–294.

Warpeha, J., 2003. Limitation of Maximal Oxygen Consumption: The Holy Grail of Exercise Physiology or Fool's Gold? *Professionalization of Exercise Physiology* **6** (9).

Weissman, I., 1978. Estimation of parameters and large quantiles based on the k largest observations. *J. Amer. Statist. Assoc.* **73** (364), 812–815.

Zhou, C., 2009. Existence and consistency of the maximum likelihood estimator for the extreme value index. *J. Multivariate Anal.* **100** (4), 794–815.

Zhou, C., 2010. The extent of the maximum likelihood estimator for the extreme value index. *J. Multivariate Anal.* **101** (4), 971–983.