Athletic Events and Statistics of Extremes: Estimation of Useful Parameters*

M. Ivette Gomes Dinis Pestana
CEAUL, DEIO, Faculty of Science
University of Lisbon, Portugal

Lígia Henriques-Rodrigues
CEAUL, University of Lisbon
Instituto Politécnico de Tomar, Tomar, Portugal

May 1, 2009

Abstract: TV shows on any athletic event make clear that *statistics* cannot be dispensed by those who want *gold medals*. And the statistics more appealing to the champions are the *extremal order statistics*, and in particular *maximum* (or *minimum*) *values* and *records*. The models in *statistics of extremes* are usually semi-parametric or even non-parametric in nature, with the imposition of a few "regularity conditions" in the right-tail of the unknown model underlying the available data. The primordial parameter is the *extreme value index* or *tail index*, the shape parameter in the (unified) *extreme value* distribution. The estimation of the *extreme value index* is the basis for the estimation of other parameters of rare events, like the *right endpoint* of the model underlying the data, a *high quantile*, the *return period* and the *probability of exceedance* of a high level. In this paper, we are interested in an application of *statistics of extremes* to the best personal marks in a few athletic events. We begin with a parametric data analysis, but we pay special attention to the semi-parametric estimation of the extreme value index, as the basis of the estimation of the right endpoint whenever finite, the possible "world record", given the actual conditions. In order to achieve a better decision we consider a few alternative semi-parametric estimators available in the literature.

AMS 2000 subject classification. Primary 62G32, 62E20; Secondary 65C05.
Keywords and phrases. *Statistics of extremes; athletics; semi-parametric estimation; extreme value index; right endpoint.*

*Research partially financed by FCT / POCTI, POCI and PPCDT / FEDER.

1 Introduction

Statistical facts are quite commonly used by sports'commentators. We all have listened to excellent programs on different athletic events, showing that *statistics* is an instrument that champions' instructors need to use. It is without doubt a subject which cannot be dispensed by those who want *gold medals*, and the statistics more appealing to the champions are the *extreme order statistics* (o.s.'s), and in particular *maximum* (or *minimum*) *values* and *records*.

The models in *statistics of extremes* are usually semi-parametric or even non-parametric in nature, with the imposition of a few "regularity conditions" in the right-tail (or left-tail), $\overline{F}(x) := 1 - F(x)$, as $x \to +\infty$ (or $F(x)$, as $x \to -\infty$), of an unknown model F underlying the available data, whenever we are interested in large (or small) values. The primordial parameter is the *extreme value index* or *tail index*. For large values, the extreme value index is the shape parameter γ in the distribution function (d.f.)

$$G_\gamma(x) = \begin{cases} \exp\left(-(1+\gamma x)^{-1/\gamma}\right), \ 1+\gamma x > 0 & \text{if} \quad \gamma \neq 0 \\ \exp(-\exp(-x)), x \in \mathbb{R} & \text{if} \quad \gamma = 0, \end{cases} \tag{1.1}$$

the (unified) *extreme value* (EV) distribution. The *extreme value index* γ needs to be estimated in a "precise" way, because such an estimation is the basis for the estimation of other parameters of extreme and large events, like the *right endpoint* of the model F underlying the data,

$$x^F := \sup\{x : F(x) < 1\}, \tag{1.2}$$

a *high quantile* with probability $1-p$, p small, i.e., $\chi_{1-p} := \inf\{x : F(x) \geq 1-p\}$, $p < 1/n$, with n the available sample size, the *return period* and the *probability of exceedance* of a *high level*.

In this paper, we shall be interested in an application of *statistics of extremes* to the best personal marks in a few athletic events, in a context similar to the one used in Einmahl & Magnus (2008). We shall pay special attention to the estimation of the tail index γ in (1.1), based on a quite recent estimator, the *mixed moment* estimator (Fraga Alves *et al.*, 2009), as well as to the associated estimation of the right endpoint x^F in (1.2), whenever finite, and of an indicator of the "return period" of the level $x_{n:n}$. The right endpoint provides an estimate of the possible "world record" given the actual conditions, and the closer to one the "return period" indicator of the level $x_{n:n}$ is, the better is the actual world record. In Section 2, we present some preliminary results in extreme value theory. In Section 3, we shall refer a few details on the semi-parametric estimation of a few parameters of extreme events. In Section 4, we justify the choice of the mixed moment estimator on the basis of a heuristic choice of the threshold for an

adaptive semi-parametric estimation of the extreme value index. In Section 5, we analyze data related with seven athletic events and draw some final comments.

2 Preliminary results in extreme value theory

Let us consider any athletic event, like for instance the women marathon. Let us denote the best personal marks of n athletes by $X_1, X_2, \cdots, X_n$ and by $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$ the associated ascending o.s.'s. For simplicity, let us assume that $(X_1, X_2, \cdots, X_n)$ can be considered as independent, identically distributed (i.i.d.) from an underlying model F, obviously unknown. Let us also assume that, if necessary, data are transformed so that we can speak of maximum values (and not of minimum values). We shall thus work with upper o.s.'s.

Remark 2.1. *Note that any result for maxima can be reformulated for minima, due to the fact that* $\min(X_1, X_2, \cdots, X_n) = -\max(-X_1, -X_2, \cdots, -X_n)$.

One of the main results in *extreme value theory* is related with the possible limiting laws of the sequence $X_{n:n} := \max(X_1, X_2, \cdots, X_n)$, of maximum values, as $n \to \infty$. Since

$$\mathbb{P}(X_{n:n} \leq x) = \mathbb{P}\Big(\bigcap_{i=1}^{n} \{X_i \leq x\}\Big) = F^n(x) \underset{n\to\infty}{\longrightarrow} \begin{cases} 0 & \text{if} \quad F(x) < 1 \\ 1 & \text{if} \quad F(x) = 1, \end{cases}$$

we obviously have

$$X_{n:n} \underset{n\to\infty}{\overset{p}{\longrightarrow}} x^F,$$

with x^F given in (1.2).

In order to obtain a possible non-degenerate behaviour for $X_{n:n}$, we thus need to normalize it. Similarly to the *central limit theorem* for sums or means, we know that if the maximum $X_{n:n}$, linearly normalized, converges to a non-degenerate random variable (r.v.), then there exist real constants $\{a_n\}_{n\geq 1}$ $(a_n > 0)$ and $\{b_n\}_{n\geq 1}$, the so-called *attraction coefficients* of F, such that

$$\lim_{n\to\infty} \mathbb{P}\Big(\frac{X_{n:n} - b_n}{a_n} \leq x\Big) = \lim_{n\to\infty} F^n(a_n x + b_n) = G_\gamma(x), \tag{2.1}$$

for some $\gamma \in \mathbb{R}$ (Gnedenko, 1943; de Haan, 1970), with $G_\gamma(x)$ given in (1.1). We then say that F is in the *domain of attraction* (for maxima) of G_γ and we use the notation $F \in \mathcal{D}_{\mathcal{M}}(G_\gamma)$.

The extreme value index γ in (1.1) measures essentially the weight of the right-tail $\overline{F} = 1 - F$.

- If $\gamma < 0$, the right tail is *light*, i.e., F has a finite right endpoint $(x^F < +\infty)$.

- If $\gamma > 0$, the right-tail is *heavy*, of a negative polynomial type, i.e., F has an infinite right endpoint.
- If $\gamma = 0$, the right tail is of an *exponential* type and the right endpoint can be either finite or infinite.

In Figure 1, we represent graphically the probability density function (p.d.f.) associated with the d.f. G_γ in (1.1), i.e. $g_\gamma(x) = dG_\gamma(x)/dx$, for $\gamma = -1,\ 0$ and 1. We also picture the standard normal p.d.f., $\varphi(x) = \exp(-x^2/2)/\sqrt{2\pi}$, $x \in \mathbb{R}$, as well as a "zoom" of the right tails of these four models. It is clear the lightness of the right-tail of G_γ for $\gamma < 0$ (finite right endpoint), followed by the normal tail and next the Gumbel tail ($\gamma = 0$). It is also clear the heaviness of the right-tail of G_γ for $\gamma > 0$.

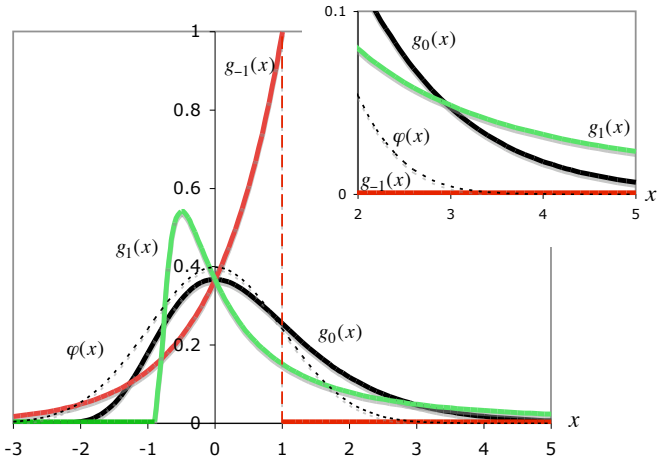

Figure 1: Extreme value p.d.f.'s, $g_\gamma(\cdot)$, $\gamma = -1, 0$ and 1, and normal p.d.f., $\varphi(\cdot)$.

Remark 2.2. *Note that to say that $F \in \mathcal{D}_{\mathcal{M}}(G_\gamma)$ is equivalent to say that for all x real and such that $0 < G_\gamma(x) < 1$,*

$$\lim_{n\to\infty} n \ln F(a_n x + b_n) = \ln G_\gamma(x) = -(1+\gamma x)^{-1/\gamma}.$$

Consequently, $F(a_n x + b_n) \to 1$ for those values of x. Since

$$\lim_{n\to\infty} \frac{-\ln F(a_n x + b_n)}{1 - F(a_n x + b_n)} = 1,$$

we equivalently have

$$\lim_{n\to\infty} n(1 - F(a_n x + b_n)) = -\ln G_\gamma(x) = (1+\gamma x)^{-1/\gamma}. \tag{2.2}$$

Let us define

$$U(t) := F^{\leftarrow}(1-1/t) \;\; (t>1), \qquad F^{\leftarrow}(x) := \inf\{y : F(y) \geq x\}, \tag{2.3}$$

with $F^{\leftarrow}$ denoting thus the generalised inverse function of F. It is reasonably easy to prove that, with G_γ^{-1} denoting the inverse function of the *extreme value* d.f. G_γ in (1.1),

$$\lim_{t\to\infty} \frac{U(tx)-b_t}{a_t} = G_\gamma^{-1}(\exp(-1/x)) = \frac{x^\gamma - 1}{\gamma}, \tag{2.4}$$

for all $x>0$, with $a_t \equiv a(t) := a_{[t]}$ and $b_t \equiv b(t) = b_{[t]}$, $[t]$ = integer part of t, and (a_n, b_n) the attraction coefficients in (2.1). Moreover, we can choose $b_t = U(t)$, with $U(\cdot)$ defined in (2.3) (see Theorem 1.1.2 of de Haan & Ferreira, 2006).

Remark 2.3. *When $\gamma = 0$, and by continuity arguments, the functions $-\ln G_\gamma(x) = (1+\gamma x)^{-1/\gamma}$ and $G_\gamma^{-1}(\exp(-1/x))$=$(x^\gamma - 1)/\gamma$, in* (2.2) *and* (2.4)*, should be interpreted as* $\exp(-x)$ *and* $\ln x$*, respectively.*

3 Semi-parametric estimation of a few relevant parameters of extreme events

On the basis of the available random sample, $(X_1, X_2, \cdots, X_n)$, let us see how to estimate the extreme value index γ, the primordial parameter in statistics of extremes, the scale a, the location b, the right endpoint x^F, a high quantile χ_{1-p}, with $p < 1/n$, small, and the return period of a high level x_H, usually defined as the expected number of exceedances of such a level.

3.1 Estimation of the extreme value index

For any integer $j \geq 1$, let us denote

$$L_{k,n}^{(j)} := \frac{1}{k}\sum_{i=1}^{k}\left\{1 - \frac{X_{n-k:n}}{X_{n-i+1;n}}\right\}^j \tag{3.1}$$

and

$$M_{k,n}^{(j)} := \frac{1}{k}\sum_{i=1}^{k}\left\{\ln X_{n-i+1:n} - \ln X_{n-k:n}\right\}^j. \tag{3.2}$$

These statistics have revealed to be fundamental in *statistics of extremes*. For the estimation of γ, we shall first refer three estimators, valid for all $\gamma \in \mathbb{R}$.

1. The *moment* (M) estimator (Dekkers *et al.*, 1989), with the functional form

$$\hat{\gamma}_{k,n}^{M} \equiv M_{k.n} := M_{k,n}^{(1)} + \tfrac{1}{2}\Big\{1 - \Big(\frac{M_{k,n}^{(2)}}{[M_{k,n}^{(1)}]^2} - 1\Big)^{-1}\Big\}, \tag{3.3}$$

$M_{k,n}^{(j)},\ j = 1, 2$ defined in (3.2).

2. The *generalized Hill* (GH) estimator introduced in Beirlant *et al.* (1996), further studied in Beirlant *et al.* (2005), and based on the Hill estimator (Hill, 1975), the statistic $M_{k,n}^{(1)}$ in (3.2), also denoted

$$\hat{\gamma}_{k,n}^{H} \equiv H_{k.n} := \frac{1}{k}\sum_{i=1}^{k}\left\{\ln X_{n-i+1:n} - \ln X_{n-k:n}\right\}, \tag{3.4}$$

and valid only for $\gamma \geq 0$. The GH-estimator, valid for all $\gamma \in \mathbb{R}$, has the functional form

$$\hat{\gamma}_{k,n}^{GH} \equiv GH_{k.n} := \hat{\gamma}_{k,n}^{H} + \frac{1}{k}\sum_{i=1}^{k}\left\{\ln \hat{\gamma}_{n,i}^{H} - \ln \hat{\gamma}_{k,n}^{H}\right\}. \tag{3.5}$$

3. The *mixed moment* (MM) estimator (Fraga Alves *et al.*, 2009), with the functional form

$$\hat{\gamma}_{k,n}^{MM} \equiv MM_{k.n} := \frac{\hat{\varphi}_{k,n} - 1}{1 + 2\min\left(\hat{\varphi}_{k,n} - 1, 0\right)}, \qquad \hat{\varphi}_{k,n} := \frac{M_{k,n}^{(1)} - L_{k,n}^{(1)}}{\left(L_{k,n}^{(1)}\right)^2}, \tag{3.6}$$

$L_{k,n}^{(1)}$ and $M_{k,n}^{(1)}$ defined in (3.1) and (3.2), respectively.

The three estimators in (3.3), (3.5) and (3.6) are consistent in all $\mathcal{D}_{\mathcal{M}}(G_\gamma)$, $\gamma \in \mathbb{R}$, provided that $k = k_n$ is an intermediate sequence, i.e., a sequence of integers such that

$$k = k_n \to \infty \quad \text{and} \quad k_n = o(n), \quad \text{as} \quad n \to \infty. \tag{3.7}$$

Due to the specificity of the data, we shall also consider a simple estimator,

4. the *location invariant* estimator (F) introduced in Falk (1995),

$$\hat{\gamma}_{k,n}^{F} \equiv F_{k.n} := \frac{1}{k}\sum_{i=1}^{k-1}\ln\frac{X_{n,n} - X_{n-i,n}}{X_{n,n} - X_{n-k,n}}, \tag{3.8}$$

valid only for $\gamma < 0$.

For intermediate k, i.e. if (3.7) holds, the estimator in (3.8) is consistent for γ in $\mathcal{D}_{\mathcal{M}}(G_{\gamma<0})$.

We still would like to refer the so-called "*maximum likelihood*" estimator, introduced in Smith (1987) and further studied in Drees *et al.* (2004). Such an estimator is valid for all $\gamma > -1/2$ and it is based on the application of the maximum likelihood methodology to the excesses $X_{n-i+1:n} - X_{n-k:n}$, $1 \le i \le k$. These excesses are approximately the k top o.s. in a sample of size k from a *generalized Pareto* (GP) model, strongly related with the *extreme value* d.f. G_γ in (1.1), through the relation,

$$GP(x;\gamma,\alpha) = 1 + \ln G_\gamma(\alpha x/\gamma) = 1 - (1+\alpha\ x)^{-1/\gamma}\,, \quad 1+\alpha x > 0,\ x > 0 \quad (\alpha,\gamma \in \mathbb{R}).$$

This is a re-parametrization due to Davison (Davison, 1984). Then, with such a re-parametrization, the *maximum likelihood* (ML) estimator of γ has an explicit expression as a function of the ML-estimator $\widehat{\alpha} = \widehat{\alpha}_{_{ML}}$ of α and the sample of the excesses. We have

$$\hat{\gamma}_{k,n}^{ML} = \widehat{\gamma}_{k,n,\hat{\alpha}}^{ML} \equiv ML_{k,n} := \frac{1}{k}\sum_{i=1}^{k} \ln\left(1 + \hat{\alpha}\ (X_{n-i+1:n} - X_{n-k:n})\right). \tag{3.9}$$

The associated estimates can be obtained only by numerical methods. This is the reason why we shall not consider this estimator in the Monte Carlo simulation in Section 4, related with a heuristic choice of the threshold. We shall however consider the ML estimator in the data analysis provided in Section 5.3.2.

For a large variety of models, and under mild second-order conditions, we can guarantee the asymptotic normality of all the above mentioned estimators and build approximate confidence intervals (CI's) for γ, as well as for all other parameters of extreme events discussed next in Section 3.2.

3.2 Estimation of other parameters of interest

3.2.1 Estimation of location and scale

As mentioned before, we have $b_t = U(t)$, with $U(\cdot)$ defined in (2.3). On another side, the universal uniform transformation enables us to guarantee that $\forall\ F$, unknown and underlying the r.v. X, $X \stackrel{d}{=} U(Y)$, with Y a unit Pareto r.v., i.e. a r.v. with d.f. $F_{_Y}(y) = 1 - 1/y,\ y \ge 1$. Consequently,

$$X_{n-k:n} \stackrel{d}{=} U(Y_{n-k:n}), \quad \text{and since} \quad Y_{n-k:n} \stackrel{p}{\sim} n/k, \text{ as } n \to \infty,$$

it is sensible to consider

$$\hat{b} = \hat{b}_{k,n} = \hat{U}(n/k) = X_{n-k:n}.$$

And for any γ-estimator, $\hat{\gamma}^\bullet \equiv \hat{\gamma}^\bullet_{k,n}$, we can consider (de Haan & Ferreira, 2006)

$$\hat{a}^\bullet = \hat{a}^\bullet_{k,n} = X_{n-k:n} M^{(1)}_{k,n}\big(1 - \min(0, \hat{\gamma}^\bullet)\big),$$

with $M^{(1)}_{k,n}$ given in (3.2).

3.2.2 Estimation of the right endpoint for $\gamma < 0$

It is possible to prove (de Haan, 1984) that $F \in \mathcal{D}_{\mathcal{M}}(G_\gamma)$ if and only if, for all $x > 0$,

$$\lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} = \begin{cases} \frac{x^\gamma - 1}{\gamma} & \text{if} \quad \gamma \neq 0 \\ \ln x & \text{if} \quad \gamma = 0, \end{cases} \tag{3.10}$$

with $U(\cdot)$ defined in (2.3). For large values of t and $\gamma \neq 0$, we can write

$$U(tx) \;\approx\; U(t) + a(t)\,\frac{x^\gamma - 1}{\gamma}. \tag{3.11}$$

But $x^F = U(\infty)$ and for all $\gamma < 0$, $(x^\gamma - 1)/\gamma \to -1/\gamma$, as $x \to \infty$. If we consider $t = n/k$, with k intermediate, we can thus guarantee that, whenever $\hat{\gamma}^\bullet < 0$,

$$x^F \;\approx\; U(n/k) - a(n/k)/\gamma \quad \Longrightarrow \quad \hat{x}^F_\bullet := \hat{b} - \hat{a}^\bullet/\hat{\gamma}^\bullet.$$

3.2.3 Estimation of a high quantile χ_{1-p}, with p small

For large values of t, we can again use the approximation in (3.11). Since $\chi_{1-p} = U(1/p)$, it is enough to consider $tx = 1/p$ and $t = n/k$, i.e. $x = k/(np)$, and to use the relation

$$\chi_{1-p} = U(1/p) \;\approx\; U(n/k) + a(n/k)\,\frac{\big(k/(np)\big)^\gamma - 1}{\gamma}.$$

We then have

$$\hat{\chi}^\bullet_{1-p} := \hat{b} + \hat{a}^\bullet\,\frac{\big(k/(np)\big)^{\hat{\gamma}^\bullet} - 1}{\hat{\gamma}^\bullet}.$$

3.2.4 Estimation of the return period of a high level x_H

In a pure framework of i.i.d. observations, if we think on the number of observations N_H needed to reach a value higher than x_H, such a r.v. has support $\{1, 2, \cdots r, \cdots\}$ and $\mathbb{P}(N_H = r) = p_H(1 - p_H)^{r-1}$, $r \geq 1$, with $p_H = \mathbb{P}(X > x_H) = 1 - F(x_H)$, i.e. N_H is a *geometric* r.v. The *return period* of the high level x_H is usually defined as the mean value of N_H, being given by

$$R(x_H) := \frac{1}{1 - F(x_H)}.$$

In the framework of this paper, it is perhaps more sensible to restrict ourselves to the n athletes under consideration, and to define the return period of a level x_H as the mean number of athlets, among the n, who will have in the future a personal mark larger than x_H. We thus have the mean value of a Binomial$(n, p = 1 - F(x_H))$ r.v., given by

$$R^*(x_H) := n(1 - F(x_H)).$$

On the basis of the limiting relation in (2.2), we can then consider the estimators

$$\hat{R}(x_H) := \frac{n}{k}\Big(\min\Big(+\infty, 1 + \hat{\gamma}^\bullet\Big(\frac{x_H - \hat{b}}{\hat{a}^\bullet}\Big)\Big)\Big)^{1/\hat{\gamma}^\bullet}$$

and

$$\hat{R}^*(x_H) := k\Big(\max\Big(0, 1 + \hat{\gamma}^\bullet\big(\frac{x_H - \hat{b}}{\hat{a}^\bullet}\big)\Big)\Big)^{-1/\hat{\gamma}^\bullet}$$

of $R(x_H)$ and $R^*(x_H)$, respectively.

Remark 3.1. *For $x_H = x_{n:n}$, F absolutely continuous, and denoting $(U_1, \cdots, U_n)$ a random sample from a uniform d.f. in $(0,1)$, we have $R^*(X_{n:n}) = n(1 - F(X_{n:n})) \stackrel{d}{=} n\, U_{1:n}$, which converges weakly towards a unit exponential r.v. Consequently,* $\exp(-R^*(X_{n:n}))$ *converges weakly towards a uniform r.v. in $(0,1)$. In the data analysis provided in* Section 5.3.2 *we shall thus consider*

$$\hat{R}_n^{**} := \exp(-\hat{R}^*(x_{n:n})) \tag{3.12}$$

as the "return period" indicator of the world record $x_{n:n}$. The closer to 1 *this indicator is, the better is the actual world record.*

For further details on the subject of this Section, see Chapters 1 and 4 of de Haan & Ferreira (2006).

4 A heuristic choice of the threshold in the semi-parametric estimation of γ: a Monte-Carlo study

For any arbitrary estimator, $\hat{\gamma}^\bullet_{k,n}$, of γ, like the ones in (3.3), (3.4), (3.5), (3.6), (3.8) and (3.9), and under the validity of a second-order condition that measures the rate of convergence in (3.10) (see de Haan & Ferreira (2006), for details), we get an asymptotic distributional representation of the type:

$$\hat{\gamma}^\bullet_{k,n} \stackrel{d}{=} \gamma + \frac{\sigma_\bullet\, P_k^\bullet}{\sqrt{k}} + v_\bullet A(n/k)(1 + o_p(1)), \tag{4.1}$$

with $P_k^\bullet \overset{a}{\sim} Normal(0,1)$ and $A(t) \to 0$, as $t \to \infty$. More specifically (Geluk & de Haan, 1987), we have $|A| \in RV_\rho$, $\rho \le 0$, where RV_α denotes the class of regularly varying functions with an index of regular variation equal to α, i.e. positive measurable functions $g(\cdot)$ such that $g(tx)/g(t) \underset{t\to+\infty}{\to} x^\alpha$ for all $x > 0$. Consequently, for intermediate levels k such that $\sqrt{k}A(n/k) \to \lambda$, finite, $\exists v_\bullet \in \mathbb{R}$ and $\sigma_\bullet \in \mathbb{R}^+$ such that

$$\sqrt{k}(\hat{\gamma}_{k,n}^\bullet - \gamma) \underset{n\to\infty}{\overset{d}{\to}} \text{Normal}(\lambda v_\bullet, \sigma_\bullet^2). \tag{4.2}$$

The "*asymptotic mean squared error*" ($AMSE$) is defined as

$$AMSE[\hat{\gamma}_{k,n}^\bullet] := \frac{\sigma_\bullet^2}{k} + v_\bullet^2 A^2(n/k),$$

i.e. we get asymptotic bias and variance given by $BIAS_\infty[\hat{\gamma}_{k,n}^\bullet] := v_\bullet\ A(n/k)$ and $Var_\infty[\hat{\gamma}_{k,n}^\bullet] := \sigma_\bullet^2/k$, respectively. If $\lambda = 0$, the mean value of the limiting normal law in (4.2) is equal to zero.

Let us define $k_0^\bullet = k_0^\bullet(n)$:=$\arg\inf_k AMSE[\hat{\gamma}_{k,n}^\bullet]$, the level associated with a minimal $AMSE$, as the optimal level for the estimation of γ through $\hat{\gamma}_{k,n}^\bullet$, and let us denote $\hat{\gamma}_{n0}^\bullet := \hat{\gamma}_{k_0^\bullet,n}^\bullet$, the estimator computed at its optimal level. With the notation $A(t) = \beta t^\rho$, $\rho < 0$, the value $\sigma_\bullet$ is a function of γ and $v_\bullet$ is usually a function of β and ρ (possibly also of γ). We then get

$$k_0^\bullet = \left(\frac{\sigma^2}{v^2\beta^2(-2\rho)}\right)^{1/(1-2\rho)} n^{-2\rho/(1-2\rho)}. \tag{4.3}$$

In order to estimate $k_0^\bullet$ in (4.3), in a simple and precise way, we thus need to have "nice" estimates of the second order parameters (β, ρ). However, whereas such an estimation is reliable for $\gamma > 0$ (see, for instance, Caeiro *et al.*, 2005; Gomes & Pestana, 2007; Gomes *et al.*, 2007, 2008, among others), this is not the case for $\gamma \le 0$. Alternatively, we could also use, for instance, bootstrap methods (Draisma *et al.*, 1999; Danielson *et al.*, 2001; Gomes & Oliveira, 2001) for an optimal adaptive choice of k. Here, after deciding on a negative value for γ, as will be the case of the data in Section 5, we propose the following heuristic choice of the threshold k. Let us denote $\hat{\gamma}_{k,n}^{(i)}$, $i \in \mathcal{T} = \{1,2,3,4\}$ the set of alternative (and computationally simple to obtain) extreme value index estimators under consideration, i.e. the estimators in (3.3), (3.5), (3.6) and (3.8). Then, consider

$$k_{min}^* := \arg\min_k \sum_{(i,j)\in\mathcal{T}, i\neq j} \left(\hat{\gamma}_{k,n}^{(i)} - \hat{\gamma}_{k,n}^{(j)}\right)^2, \tag{4.4}$$

and work with the adaptive estimators

$$M_{min}^* := M_{k_{min}^*,n},\ GH_{min}^* := GH_{k_{min}^*,n},\ MM_{min}^* := MM_{k_{min}^*,n} \text{ and } F_{min}^* := F_{k_{min}^*,n}, \tag{4.5}$$

with $M_{k,n}$, $GH_{k,n}$, $MM_{k,n}$, $F_{k,n}$ and k^*_{min} given in (3.3), (3.5), (3.6), (3.8) and (4.4), respectively. In order to obtain distributional properties of the estimators in (4.5), we have performed a small-scale simulation study of size 5000×10 for sample sizes $n = 100,\ 200,\ 300,\ 400,\ 500,\ 1000,\ 2000$ and 5000, from the following underlying models:

- the *extreme value* (EV) model, with d.f. $F(x) = G_\gamma(x)$, with $G_\gamma(x)$ given in (1.1), for $\gamma = -0.1$, -0.2 and -0.3;
- the *generalized Pareto* (GP) model, with d.f. $F(x) = 1 + \ln G_\gamma(x) = 1 - \big(1 + \gamma x\big)^{-1/\gamma}$, $0 \le x < -1/\gamma$, also for $\gamma = -0.1, -0.2$ and -0.3.

For each value of n, and for each model, we have simulated the mean values and mean squared errors of the four estimators in (4.5). For underlying EV models, with $\gamma = -0.1, -0.2$ and -0.3, the estimates of the absolute bias ($|BIAS|$) and mean squared error (MSE) are presented in Figures 2, 3 and 4, respectively. We also present in these figures the corresponding values of one of the estimators at its simulated optimal level, the one with the lowest mean squared error for large values of n, denoted $T_{0,s}$, with $T = M, GH, MM$ or F.

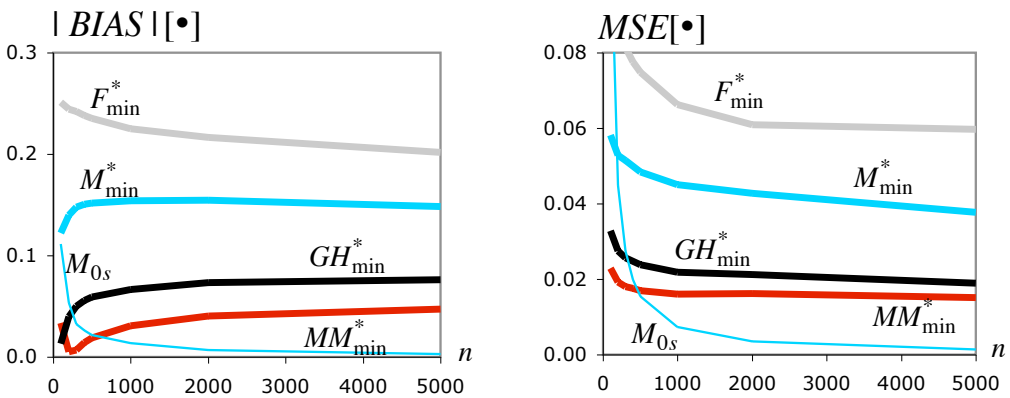

Figure 2: Absolute values of bias (*left*) and mean squared errors (*right*) of the adaptive extreme value index estimators in (4.5), for an *extreme value* model with $\gamma = -0.1$.

A few remarks related with the adaptive estimators in (4.5) for underlying EV models:

- For $\gamma = -0.1$, the absolute bias of MM^*_{min} is the smallest one, except for $n = 100$. For this sample size, and regarding absolute bias, GH^*_{min} beats MM^*_{min}. Regarding MSE, the best of the adaptive estimates is MM^*_{min}, for all n.
- For $\gamma = -0.2$, the absolute bias of MM^*_{min} is the smallest one, for $n \ge 1000$. But for

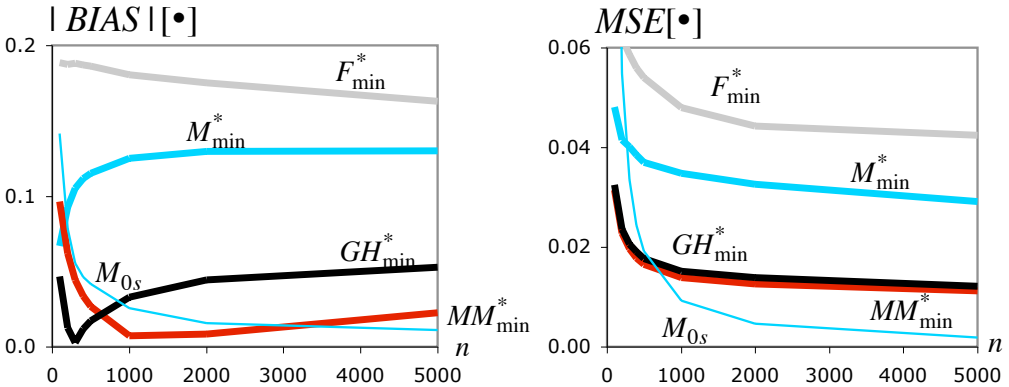

Figure 3: Absolute values of bias (*left*) and mean squared errors (*right*) of the adaptive extreme value index estimators in (4.5), for an *extreme value* model with $\gamma = -0.2$.

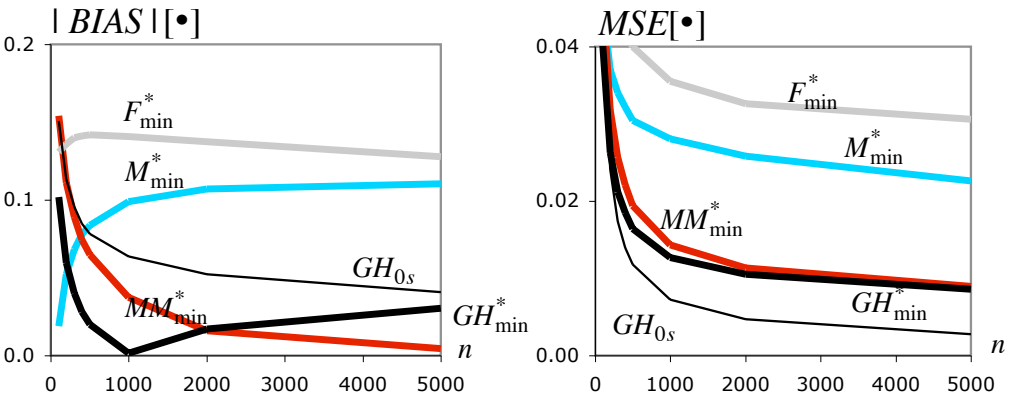

Figure 4: Absolute values of bias (*left*) and mean squared errors (*right*) of the adaptive extreme value index estimators in (4.5), for an *extreme value* model with $\gamma = -0.3$.

$n \leq 500$, GH^*_{min} beats MM^*_{min}. Regarding MSE, the best of the estimates is MM^*_{min}, quite close to GH^*_{min} for all n.

- For $\gamma = -0.3$, the absolute bias of MM^*_{min} is the smallest one, only for $n \geq 2000$. For $300 \leq n \leq 1000$, GH^*_{min} beats the other estimators. For $n \leq 200$, the smallest absolute bias is the one of M^*_{min}. Regarding MSE, the best of the estimates is GH^*_{min}, quite close to MM^*_{min} for all n.

Regarding the "potential" estimators T_0 at simulated optimal levels, with $T = M, GH, MM$ or F, we draw the following comments:

- At simulated optimal levels, GH_0 achieves the minimum MSE for all n, if $\gamma = -0.3$. For

the other values of γ, GH_0 is the best one for small n, but M_0 becomes the best for large n ($n \geq 1000$ for $\gamma = -0.1$ and $n \geq 2000$ for $\gamma = -0.2$).

- Regarding smallest absolute bias at simulated optimal levels, M_0 is the best for all n, if $\gamma = -0.1$. For the other values of γ, M_0 is the best for $n \geq 200$. For $n = 100$, GH_0 overpasses all other ones.

Similar patterns have been obtained for underlying GP parents. The unique difference lies in the fact that, at simulated optimal levels, the role of GH_0 is taken by F_0. Consequently, we see no reason to present those results.

In Table 1, for GP underlying parents and for $T = M, GH, MM$ and F, we present a relative efficiency indicator, denoted $REFFI$, of T^*_{min}, in (4.5), relatively to T_{0s}, the T-estimator computed at its simulated optimal level. We have thus simulated

$$REFFI_{T^*} \equiv REFFI_{T^*_{min}|T_{0s}} := \sqrt{MSE_s(T_{0s})/MSE_s(T^*_{min})},$$

so that the higher than one this indicator is, the better the first estimator performs comparatively with the second one. For all n, the highest indicator is written in **bold** and <u>underlined</u>.

Table 1: Simulated $REFFI$'s of the estimators under study for *generalized Pareto* underlying parents, together with 95% CI's.

GP parent, $\gamma = -0.1$						
n	100	200	500	1000	2000	5000
$M^*_{min}\|M_{0s}$	1.023±0.014	0.744±0.006	0.539±0.008	0.439±0.004	0.366±0.005	0.278±0.003
$MM^*_{min}\|MM_{0s}$	**<u>1.269</u>**±0.020	**<u>1.154</u>**±0.018	**<u>1.007</u>**±0.015	**<u>0.892</u>**±0.011	**<u>0.799</u>**±0.014	**<u>0.698</u>**±0.006
$GH^*_{min}\|GH_{0s}$	1.078±0.014	0.945±0.010	0.766±0.012	0.653±0.008	0.565±0.010	0.469±0.007
$F^*_{min}\|F_{0s}$	0.759±0.004	0.657±0.003	0.550±0.004	0.489±0.002	0.440±0.002	0.390±0.001
GP parent, $\gamma = -0.2$						
$M^*_{min}\|M_{0s}$	**<u>1.190</u>**±0.016	0.871±0.008	0.622±0.010	0.498±0.005	0.407±0.005	0.296±0.004
$MM^*_{min}\|MM_{0s}$	1.187±0.014	**<u>1.184</u>**±0.010	**<u>1.162</u>**±0.015	**<u>1.116</u>**±0.010	**<u>1.060</u>**±0.016	**<u>0.968</u>**±0.007
$GH^*_{min}\|GH_{0s}$	1.168±0.014	1.084±0.012	0.934±0.013	0.811±0.011	0.708±0.013	0.587±0.008
$F^*_{min}\|F_{0s}$	0.743±0.004	0.625±0.004	0.496±0.004	0.421±0.002	0.362±0.002	0.302±0.002
GP parent, $\gamma = -0.3$						
$M^*_{min}\|M_{0s}$	**<u>1.302</u>**±0.017	0.973±0.011	0.697±0.008	0.551±0.005	0.450±0.006	0.320±0.005
$MM^*_{min}\|MM_{0s}$	1.079±0.011	**<u>1.130</u>**±0.007	**<u>1.209</u>**±0.012	**<u>1.251</u>**±0.009	**<u>1.264</u>**±0.020	**<u>1.216</u>**±0.016
$GH^*_{min}\|GH_{0s}$	1.134±0.008	1.091±0.012	1.006±0.013	0.922±0.012	0.829±0.015	0.700±0.016
$F^*_{min}\|F_{0s}$	0.710±0.004	0.577±0.003	0.435±0.003	0.353±0.002	0.291±0.002	0.229±0.002

Table 2 is equivalent to Table 1, but for underlying EV parents.

Table 2: Simulated $REFFI$'s of the estimators under study for *extreme value* underlying parents, together with 95% CI's.

EV parent, $\gamma = -0.1$						
n	100	200	500	1000	2000	5000
$M^*_{min}\|M_{0s}$	**1.403**±0.015	0.921±0.011	0.508±0.130	0.404±0.002	0.289±0.003	0.194±0.002
$MM^*_{min}\|MM_{0s}$	1.212±0.009	**1.116**±0.016	**0.973**±0.016	**0.874**±0.013	**0.771**±0.014	**0.684**±0.006
$GH^*_{min}\|GH_{0s}$	1.021±0.010	0.877±0.010	0.700±0.009	0.599±0.009	0.506±0.006	0.424±0.006
$F^*_{min}\|F_{0s}$	1.023±0.009	0.898±0.002	0.768±0.003	0.698±0.003	0.628±0.003	0.530±0.003
EV parent, $\gamma = -0.2$						
$M^*_{min}\|M_{0s}$	**1.697**±0.011	1.149±0.013	0.720±0.010	0.518±0.006	0.375±0.003	0.253±0.002
$MM^*_{min}\|MM_{0s}$	1.149±0.008	**1.166**±0.012	**1.149**±0.015	**1.110**±0.010	**1.042**±0.011	**0.958**±0.009
$GH^*_{min}\|GH_{0s}$	1.091±0.009	0.980±0.012	0.817±0.012	0.710±0.010	0.609±0.006	0.508±0.005
$F^*_{min}\|F_{0s}$	1.067±0.009	0.934±0.004	0.779±0.004	0.694±0.004	0.609±0.004	0.506±0.006
EV parent, $\gamma = -0.3$						
$M^*_{min}\|M_{0s}$	**1.932**±0.020	**1.355**±0.020	0.896±0.013	0.657±0.006	0.489±0.003	0.342±0.004
$MM^*_{min}\|MM_{0s}$	1.060±0.006	1.121±0.009	**1.214**±0.008	**1.245**±0.014	**1.246**±0.012	**1.195**±0.013
$GH^*_{min}\|GH_{0s}$	1.038±0.010	0.959±0.006	0.847±0.010	0.757±0.007	0.670±0.008	0.570±0.008
$F^*_{min}\|F_{0s}$	1.097±0.010	0.959±0.005	0.788±0.007	0.693±0.145	0.597±0.006	0.541±0.008

On the basis of the above mentioned results, and despite of the fact that it is not possible to claim that MM^*_{min} has, for all models in $\mathcal{D}_{\mathcal{M}}(G_\gamma)$, $\gamma < 0$, the best performance among the four adaptive estimators in (4.5), it is clear that if we have to elect one of these four adaptive estimators, we are inclined to the choice of MM^*_{min}, particularly if the model is not a long way from an EV model. And we have had a light indication for this underlying parent on the basis of the undertaken parametric data analysis in Section 5.1. This is the reason why in Section 5.3.2, we shall compute the final estimates of γ and other parameters of extreme events essentially on the basis of MM^*_{min}. Note however that, for small n, GH^*_{min} is also a serious alternative.

5 Data analysis of extreme indoor athletic events

The data under analysis are related with four running events, 60 Metres Hurdles, 200, 400 and 1500 Metres, all for men, and denoted 60MH, 200M, 400M and 1500M, respectively, and with three jumping events, also for men, high jump (HJ), long jump (LJ) and pole vault (PV). *Source:* http://www.iaaf.org/statistics/toplists/index.htmx. Data was collected until the end of 2007 and for any athlete only the best mark was taken into account.

5.1 Parametric data analysis

Prior to a semi-parametric analysis of the data, in the most common framework of *statistics of extremes*, we shall proceed to a parametric data analysis, in the lines of Robinson & Tawn (1995) and Barão & Tawn (1999), who considered the annual best times in the women's 3000m event. Also Smith (1988) has proposed a maximum likelihood method of fitting models to a series of records, and applied his method to athletics records for the mile and the marathon. The attempts made in these papers to predict an ultimate world record are based on the development of top performances over time. This is not the case in this paper. Here, as in Einmahl & Magnus (2008), we are not interested in predicting the world record in the future. We are using only the top performances associated with a set of n athletes, and consequently, our estimated ultimate record tells us what, in principle, is possible at this moment, given today's knowledge and material.

We first illustrate in Figures 5 and 6, the Gumbel QQ-plots associated to all data sets under analysis. In all figures we have thus plotted $\left(x_{i:n}, p_i^{\Lambda} = -\ln(-\ln(i/(n+1)))\right)$, $1 \le i \le n$, and proceeded to the fitting of a least-squares line.

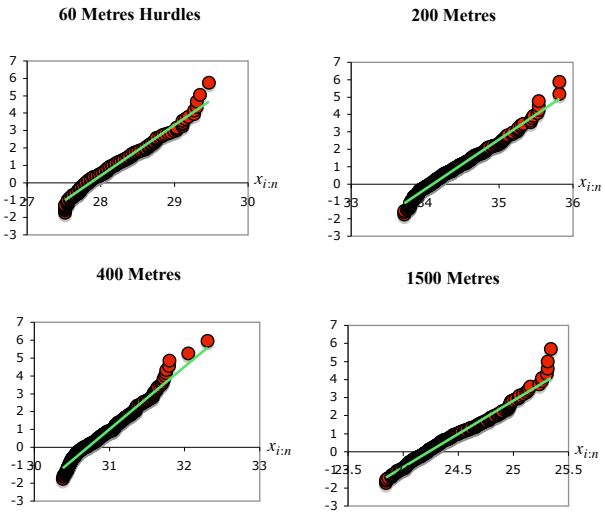

Figure 5: Gumbel QQ-plot related with the *running events* under analysis — 60 Metres Hurdles, 200, 400 and 1500 Metres.

Apart from the Long Jump event, where $\gamma = 0$ can perhaps provide a reasonable fit to the right tail, despite of a slight deviation of top o.s. smaller than the second largest value, all other events exhibit a light right tail, i.e. an extreme value index $\gamma < 0$ and consequently a finite right endpoint x^F.

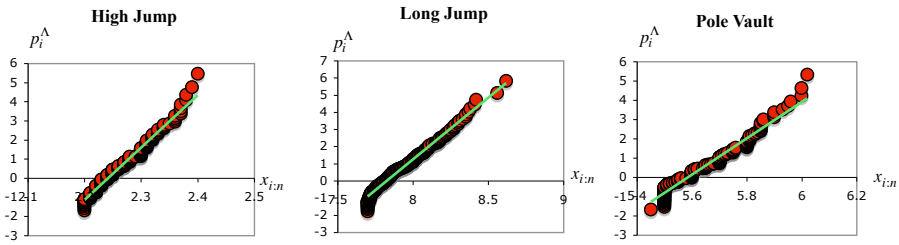

Figure 6: Gumbel QQ-plot related with the *jumping events* under analysis — Long Jump, High Jump and Pole Vault.

Due to the fact that the observed data considered are already maxima, possibly of a small and dependent number of marks associated with any of the n athletes, but the *extreme value* limiting law in (1.1) is robust to changes of the i.i.d. assumption, we have first tried the fitting, through maximum likelihood, of an EV model $F(x;\lambda,\delta,\gamma) = G_\gamma((x-\lambda)/\delta)$, with $G_\gamma(x)$ given in (1.1). We have used the EVIR package in the R-software. The estimate of the right endpoint is then provided by $\hat{x}^F = \max(x_{n:n}, \hat{\lambda} - \hat{\delta}/\hat{\gamma})$, with $(\hat{\lambda}, \hat{\delta}, \hat{\gamma})$ the maximum likelihood estimates of the vector $(\lambda, \delta, \gamma)$ of unknown parameters.

n	**Event**	$(x_{1:n}, x_{n:n})$	$\hat{\gamma}$ (I.C. a 95%)	$\hat{\delta}$ [st. error]	$\hat{\lambda}$ [st. error]	$\hat{x}^F$ [st. error]
312	60MH	$(27.52, 29.47)^*$	-0.21 $(-0.324, -0.085)$	0.28 [0.015]	27.84 [0.019]	29.47 [0.744]
352	200M	$(33.72, 35.82)^*$	-0.22 $(-0.336, -0.096)$	0.28 [0.015]	34.08 [0.018]	35.82 [0.651]
380	400M	$(30.38, 32.31)^*$	-0.15 $(-0.277, -0.024)$	0.25 [0.013]	30.70 [0.016]	32.36 [2.757]
296	1500M	$(23.84, 25.34)^*$	-0.05 $(-0.154, +0.058)$	0.26 [0.013]	24.23 [0.017]	29.51 [1.724]
235	HJ	$(2.20, 2.40)^{**}$	-0.08 $(-0.209, +0.059)$	0.034 [0.002]	2.24 [0.003]	2.69 [0.209]
339	LJ	$(7.70, 8.62)^{**}$	-0.24 $(-0.376, -0.110)$	0.11 [0.006]	7.81 [0.007]	8.62 [0.223]
205	PV	$(5.45, 6.02)^{**}$	-0.12 $(-0.315, +0.075)$	0.09 [0.007]	5.58 [0.009]	6.37 [0.466]

Table 3: Maximum likelihood estimates of $(\gamma, \delta, \lambda)$ and x^F in an underlying model $G_\gamma((x-\lambda)/\delta)$, with $G_\gamma(x)$ given in (1.1): *Km/h, **metres

As expected, all estimates of γ are negative. But for the 1500 Metres, High Jump and Pole Vault, the upper limits of the associated 95% CI's are positive, suggesting that the value $\gamma = 0$ could possibly be adequate. Slightly problematic is the estimation of the right endpoint, which is equal to the maximum value in the data, the value $x_{n:n}$, for three of the athletic events, 60 Metres Hurdles, 200 Metres and Long Jump.

5.2 Fitting the extreme value model

In Figure 7, we picture in *black* the critical points of the Kolmogorov-Smirnov (KS) statistic at a significance level $\alpha = 0.05$, equal to $1.36/\sqrt{n}$. The observed values of the KS statistic, $KS_n := \max_{1\le i\le n}(|G_{\hat{\gamma}}((x_{i:n} - \hat{\lambda})/\hat{\delta} - i/n|, |G_{\hat{\gamma}}((x_{i:n} - \hat{\lambda})/\hat{\delta} - (i-1)/n|)$ are pictured in *grey*.

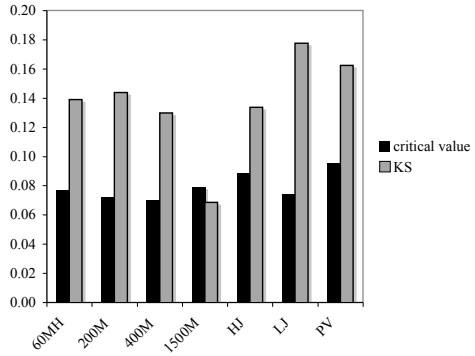

Figure 7: Critical values (*black*) and observed values of KS-statistic (*grey*) for all athletic events under analysis — 60 Metres Hurdles, 200 Metres, 400 Metres, 1500 Metres, High Jump, Long Jump and Pole Vault.

At the significance level $\alpha = 0.05$, the hypothesis of a (unified) *extreme value* model has thus been rejected by the Kolmogorov-Smirnov test for all data sets, but the 1500 Metres, as could also have been inferred graphically from Figure 8 and Figure 9, where we picture the empirical d.f. and the fitted extreme value d.f. Alternative parametric models have even provided worse fitting results. We thus claim for the need of a semi-parametric data analysis, developed in Section 5.3.

5.3 A semi-parametric data analysis

5.3.1 Testing the sign of the extreme value index

As mentioned before, whenever we place ourselves under a semi-parametric framework, we assume only that (3.10) holds, or equivalently, that $F \in \mathcal{D}_{\mathcal{M}}(G_\gamma)$, for a certain γ, being γ the primordial parameter of extreme events.

In many areas where extremes are relevant, as for instance in environmental statistics, the simplest case $\gamma = 0$ is often considered. Moreover, if we clearly think that $\gamma < 0$ or that $\gamma > 0$, we have specific procedures for the estimation of γ, often more reliable than the procedures valid for a general $\gamma \in \mathbb{R}$. Prior to a deeper analysis of the tail, it is thus sensible to test

$$H_0 : F \in \mathcal{D}_{\mathcal{M}}(G_0) \quad \text{versus} \quad H_1 : F \in \mathcal{D}_{\mathcal{M}}(G_{\gamma\neq 0}), \tag{5.1}$$

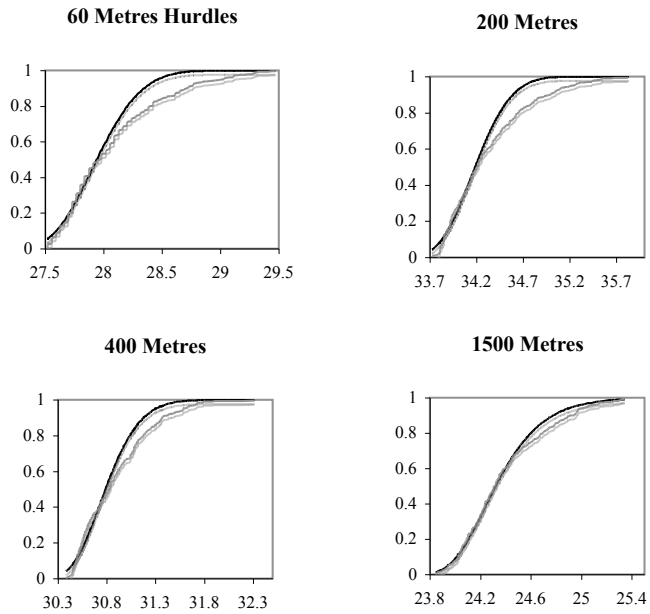

Figure 8: Empirical d.f. (*grey*) and fitted extreme value d.f. (*black*) for the running events under analysis — 60 Metres Hurdles, 200, 400 and 1500 Metres.

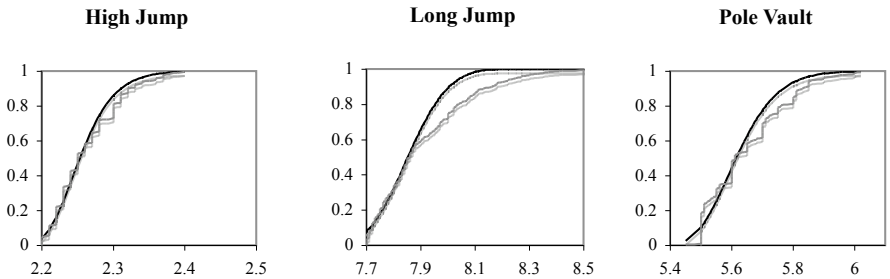

Figure 9: Empirical d.d. (*grey*) and fitted extreme value d.f. (*black*) for the jumping events under analysis — Long Jump, High Jump and Pole Vault.

or even against one-sided alternatives $F \in \mathcal{D}_{\mathcal{M}}(G_{\gamma>0})$ or $F \in \mathcal{D}_{\mathcal{M}}(G_{\gamma<0})$.

We shall consider here two test statistics of a similar type, i.e. both based on the excesses over a high random threshold $X_{n-k:n}$, with k satisfying (3.7). The first one was introduced by Greenwood (1946) and the second one by Hasofer & Wang (1992). These two statistics were further studied, under a semi-parametric framework, by Neves & Fraga Alves (2007). They are

given by

$$G_{k,n} := \frac{\frac{1}{k}\sum_{i=1}^{k}(X_{n-i+1:n} - X_{n-k:n})^2}{\left(\frac{1}{k}\sum_{i=1}^{k}(X_{n-i+1:n} - X_{n-k:n})\right)^2}$$

and

$$W_{k,n} := \frac{1}{k(G_{k,n}-1)}.$$

Under the null hypothesis H_0 in (5.1) and extra mild conditions on the right tail of F and on the growth of $k = k_n$, they both have an asymptotic normal behaviour. More specifically,

$$G^*_{k,n} := \sqrt{k/4}\,\left(G_{k,n} - 2\right)\Big|_{F\in\mathcal{D}_{\mathcal{M}}(G_0)} \underset{n\to\infty}{\overset{d}{\to}} N(0,1), \tag{5.2}$$

and

$$W^*_{k,n} := \sqrt{k/4}\,\left(kW_{k,n} - 1\right)\Big|_{F\in\mathcal{D}_{\mathcal{M}}(G_0)} \underset{n\to\infty}{\overset{d}{\to}} N(0,1). \tag{5.3}$$

Motivated by the important contribution of the maximum to the sum of the k excesses, $X_{n-i+1:n} - X_{n-k:n}$, $1 \le i \le k$, Neves *et al.* (2006) introduced the following complimentary statistic,

$$R_{k,n} := \frac{X_{n:n} - X_{n-k:n}}{\frac{1}{k}\sum_{i=1}^{k}(X_{n-i+1:n} - X_{n-k:n})},$$

also considered in the analysis of the data under study. The asymptotic behaviour of $R_{k,n}$ is provided by the Gumbel d.f., $\Lambda = G_0$, with G_γ given in (1.1). More specifically,

$$R^*_{k,n} := R_{k,n} - \ln k\Big|_{F\in\mathcal{D}_{\mathcal{M}}(G_0)} \underset{n\to\infty}{\overset{d}{\to}} Z \frown G_0. \tag{5.4}$$

As a function of k both $G^*_{k,n}$ and $R^*_{k,n}$ tend to have a slope with the sign of γ. The statistic $W^*_{k,n}$ works the other way round.

As an illustration, we present, in Figure 10, the sample paths of the three test statistics $G^*_{k,n}$, $W^*_{k,n}$ and $R^*_{k,n}$ in (5.2), (5.3) and (5.4), respectively, associated with the Long Jump and the 200 Metres. In this figure we also picture the quantiles $(\chi^{\bullet}_{0.025}, \chi^{\bullet}_{0.975})$ of the standard normal Φ, equal to $(-1.96, +1.96)$, and of the standard Gumbel $\Lambda \equiv G_0$, equal to $(-1.31, +3.68)$.

For all other data sets under analysis the graphs are similar, showing clearly a decreasing trend of $R^*_{k,n}$ and $G^*_{k,n}$ (with $G^*_{k,n}$ below $\chi^{\Phi}_{0.025}$ for a large number of k-values), as well as an increasing trend of $W^*_{k,n}$ (above $\chi^{\Phi}_{0.975}$ for moderate up to large values of k). This provides a strong suggestion of a negative *extreme value index*, as expected. Despite of that, notice that the

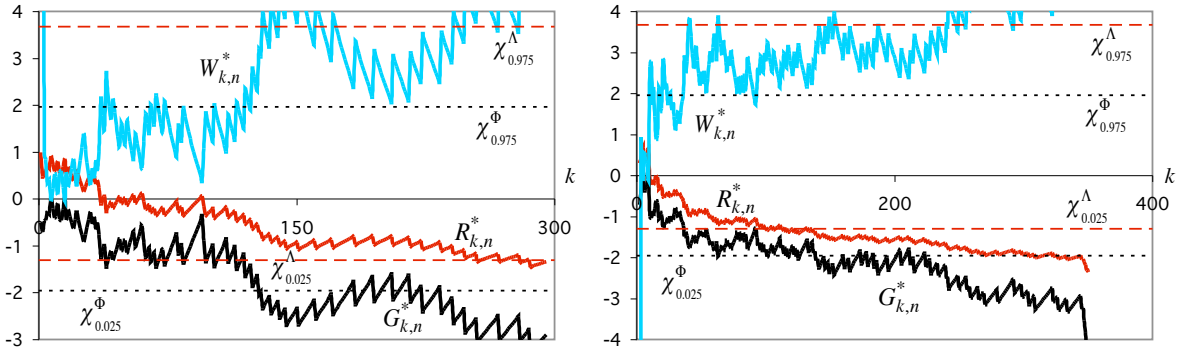

Figure 10: Test statistics for the LJ (*left*) and the 200M events (*right*)

sample path of $R^*_{k,n}$ is within the 95% CI for almost all k-values and the data set associated with the Long Jump event. This was also expected, because it is well known (see, for instance Neves & Fraga Alves, 2008) that $R^*_{k,n}$ tends to be a conservative test and the true value of γ is for sure close to zero.

5.3.2 Semi-parametric estimates of the extreme value index and the right endpoint

In Table 4 we present a summary of the performed data analysis, with estimates of γ and 95% CI's for the extreme value index γ. These estimates of γ were obtained through the mixed moment (MM) estimates, computed at the value k^*_{min}, in (4.4), i.e. the adaptive estimate MM^*_{min} in (4.5).

n	**Event**	$(x_{1:n}, x_{n:n})$	MM^*_{min} (95% CI)	k^*_{min}
312	60MH	$(27.52, 29.47)^*$	-0.36 $(-0.490, -0.230)$	304
352	200M	$(33.72, 35.82^*)$	-0.34 $(-0.627, -0.060)$	62
380	400M	$(30.38, 32.31)^*$	-0.26 $(-0.445, -0.080)$	128
296	1500M	$(23.84, 25.34)^*$	-0.43 $(-0.579, -0.286)$	279
235	HJ	$(2.20, 2.40)^{**}$	-0.40 $(-0.563, -0.240)$	214
339	LJ	$(7.70, 8.62)^{**}$	-0.29 $(-0.468, -0.121)$	150
205	PV	$(5.45, 6.02)^{**}$	-0.42 $(-0.618, -0.220)$	147

Table 4: Estimates of the extreme value index: *Km/h, **metros

However, in the data analysis, we have also considered the adaptive estimators MM^{**}_{min} and ML^{**}_{min}, the estimators in (3.6) and (3.9), respectively, computed at the value k^{**}_{min}, obtained through a minimization procedure of the type of the one in (4.4), but including also the ML

estimator. The reason for the consideration of the ML estimator lies on the fact that in the region $-1/2 < \gamma < 0$, where the estimates indeed lie, $\sigma^2_{ML} = (1+\gamma)^2$ is smaller than $\sigma^2_{MM} = \sigma^2_M = (1-\gamma)^2(1-2\gamma)(1-\gamma+6\gamma^2)/((1-3\gamma)(1-4\gamma))$ for all γ, with $\sigma_\bullet$ the asymptotic standard deviation in the asymptotic representation (4.1). These estimates are presented in Table 5. For the events where $MM^{**}_{min} \neq MM^{*}_{min}$ (400M and LJ), the estimates MM^{*}_{min} are written in *italic*. For those same athletic events, 400 Metres and Long Jump, we have got $ML^{*}_{min} = -.25(-.597, 0.095)$ and $ML^{*}_{min} = -.24(-.361, -0.117)$, respectively.

n	**Event**	ML^{**}_{min} (95% CI)	MM^{**}_{min} (95% CI)	k^{**}_{min}
312	60MH	-0.34 $(-0.412, -0.263)$	-0.36 $(-0.490, -0.230)$	304
352	200M	-0.38 $(-0.534, -0.224)$	-0.34 $(-0.627, -0.060)$	62
380	400M	-0.22 $(-0.351, -0.085)$	$\mathit{-0.26}$ $(-0.434, -0.077)$	133
296	1500M	-0.43 $(-0.495, -0.361)$	-0.43 $(-0.579, -0.286)$	279
235	HJ	-0.39 $(-0.475, -0.312)$	-0.40 $(-0.563, -0.240)$	214
339	LJ	-0.23 $(-0.315, -0.138)$	$\mathit{-0.24}$ $(-0.353, -0.118)$	295
205	PV	-0.42 $(-0.514, -0.326)$	-0.42 $(-0.618, -0.220)$	147

Table 5: Estimates of the extreme value index: *Km/h, **metros

Also as an illustration, we present, in Figure 11, the estimates $M \equiv M_{k,n}$, $GH \equiv GH_{k,n}$, $MM \equiv MM_{k,n}$, and $F \equiv F_{k,n}$ of γ, defined in (3.3), (3.5), (3.6) and (3.8), respectively, again for the Long Jump and the 200 Metres athletic events. We also picture the sample paths of the γ-estimator $ML \equiv ML_{k,n}$, in (3.9).

All semi-parametric γ-estimates at $k = k^{**}_{min}$ are within the CI's provided in Table 4 and based on MM^{*}_{min}. Similarly, all estimates in Table 4 are within the CI's provided in Table 5. However, apart from the parametric estimates of γ associated with the 200 Metres, 400 Metres and Long Jump events, the parametric estimates in Table 3 are outside the CI's provided in Table 4, as well as the other way round. The parametric estimates are above the semi-parametric estimates for the seven events considered.

Next, in Table 6 we present the estimates of the right endpoints of the models underlying the different data sets under study, as well as of the "return period" indicators of the levels $x_H = x_{n:n}$, provided in (3.12). We base these estimates on the MM and ML-estimates in Table 5.

The results in Tables 4 and 6 mean that, under the present conditions, there are finite upper

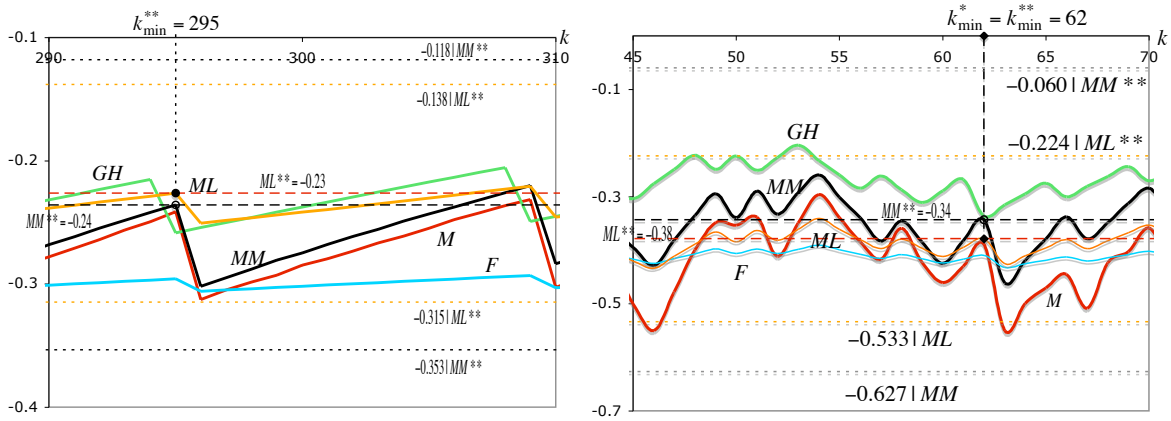

Figure 11: Long Jump (*left*) and 200 Metres (*right*)

Event	$x_{n:n}$	$\hat{x}^F_{ML}$ / $\hat{x}^F_{MM}$	$\hat{R}^{**}_{n\|ML}$ / $\hat{R}^{**}_{n\|MM}$
60MH	29.47* (00:07.33)•	29.75 / 29.64* (00:07.26 / 00:07.29)•	0.52 / 0.74
200M	35.82* (00:20.10)•	36.01 / 36.11* (00:19.99 / 00:19.93)•	0.68 / 0.54
400M	32.31* (00:44.57)•	32.67 / 32.45* (00:44.07 / 00:44.36)•	0.89 / 0.98
1500M	25.34* (03:33.08)•	25.45 / 25.44* (03:32.16 / 03:32.25)•	0.53 / 0.58
HJ	2.40**	2.41 / 2.41**	0.83 / 0.89
LJ	8.62**	8.81 / 8.66**	0.86 / 0.99
PV	6.02**	6.05 / 6.05**	0.81 / 0.80

Table 6: Estimates of the right endpoint, whenever finite, and of the "return period" indicator of the level $x_{n:n}$: *km/h, •minutes, **metres.

limits for all jumping events under analysis, as well as finite lower limits in the times associated with all running events under analysis. From the return period indicators of the world records, we can say that the current Long Jump and 400M are very good (above 85%). The indicators associated with 200 Metres and the 1500 Metres are clearly below 70%.

References

[1] Barão, M.I.; Tawn, J. (1999). Extremal analysis of short series with outliers: sea-levels and athletic records. *Applied Statistics* **48**, 469-487.

[2] Beirlant, J.; Dierckx, G.; Guillou, A. (2005). Estimation of the extreme-value index and generalized quantile plots. *Bernoulli* **11**:6, 949-970.

[3] Beirlant, J.; Vynckier, P.; Teugels, J. (1996). Excess functions and estimation of the extreme-value index. *Bernoulli* **2**, 293-318.

[4] Caeiro, F.; Gomes, M.I.; Pestana, D.D. (2005). Direct reduction of bias of the classical Hill estimator. *Revstat* **3** (2), 113-136.

[5] Danielsson, J.; Haan, L. de; Peng, L.; de Vries, C.G. (2001). Using a bootstrap method to choose the sample fraction in the tail index estimation. *J. Mult. Anal.* 76, 226-248.

[6] Davison, A. C. (1984). Modelling excesses over high thresholds. In *Statistical Extremes and Applications*, ed. J. Tiago de Oliveira, Dordrecht: D. Reidel, 461-482.

[7] Dekkers, A.; Einmahl, J.; de Haan, L. (1989). A moment estimator for the index of an extreme-value distribution. *Annals of Statistics* **17**, 1833–1855.

[8] Draisma, G.; Haan, L. de; Peng, L.; Pereira, T.T. (1999). A bootstrap-based method to achieve optimality in estimating the extreme-value index. *Extremes* **2**, 367-404.

[9] Drees, H.; Ferreira, A.; de Haan, L. (2004). On maximum likelihood estimation of the extreme value index. *Ann. Appl. Probab.* **14**, 1179-1201.

[10] Einmahl, J.; Magnus, J.R. (2008). Records in athletics through extreme-value theory. *Journal of the American Statistical Association*, **103**, 1382-1391.

[11] Falk, M. (1995). Some best parameter estimates for distributions with finite endpoint. *Statistics* **27**(1-2), 115-125.

[12] Fraga Alves, M.I.; Gomes, M.I.; de Haan, L.; Neves, C. (2009). The mixed moment estimator and location invariant alternatives. *Extremes* **12**, 149-185. DOI: 10.1007/s10687-008-0073-3, 2008.

[13] Geluk, J.; de Haan, L. (1987). *Regular Variation, Extensions & Tauberian Theorems.* CWI Tract 40, Center for Mathematics & Computer Science, Amsterdam, Netherlands.

[14] Gnedenko, B.V. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Ann. Math.* **44**, 423-453.

[15] Gomes, M.I.; de Haan, L.; Henriques Rodrigues, L. (2008). Tail index estimation through accommodation of bias in the weighted log-excesses. *J. Royal Statistical Society* **B70**, Issue 1, 31-52.

[16] Gomes, M.I.; Martins, M.J.; Neves, M. (2007). Improving second order reduced bias extreme value index estimation. *Revstat* **5**(2), 177-207.

[17] Gomes, M.I.; Oliveira, O. (2001). The bootstrap methodology in Statistics of Extremes: choice of the optimal sample fraction. *Extremes* **4**:4, 331-358.

[18] Gomes, M.I.; Pestana, D. (2007). A sturdy reduced-bias extreme quantile (VaR) estimator. *J. American Statistical Association*, Vol. **102**, No. 477, 280-292.

[19] Greenwood, M. (1946). The statistical study of infectious diseases. *J. Roy. Statist. Soc.* **A109**, 85-109.

[20] de Haan, L. (1970). *On Regular Variation & its Application to the Weak Convergence of Sample Extremes*. Mathematical Centre Tract 32, Amsterdam.

[21] de Haan, L. (1984). Slow variation & characterization of domains of attraction. In Tiago de Oliveira, ed., *Statistical Extremes & Applications*, 31-48, D. Reidel, Dordrecht, Holland.

[22] de Haan, L.; Ferreira, A. (2006). *Extreme Value Theory: an Introduction*. Springer Science+Business Media, LLC, New York.

[23] Hasofer, A.; Wang, J.Z. (1992). A test for extreme value domain of attraction. *J. Amer. Statist. Assoc.* **87**, 171-177.

[24] Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.* **3**, 1163-1174.

[25] Neves, C.; Fraga Alves, M.I. (2007). Semi-parametric approach to Hasofer-Wang and Greenwood statistics in extremes. *Test* **16**, 297-313.

[26] Neves, C.; Fraga Alves, M.I. (2008). Testing extreme value conditions — an overview and recent approaches. *Revstat* **6**(1), 83-100.

[27] 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. Statist. Planning & Inference* **136**(4), 1281-1301.

[28] Robinson, M.E.; Tawn, J. (1995). Statistics for exceptional athletic records. *Applied Statistics* **44**, 499-511.

[29] Smith, R.L. (1987). Estimating tails of probability distributions. *Ann. Statist.* **15**:3, 1174-1207.

[30] Smith, R.L. (1988). Forecasting records by maximum likelihood. *J. Amer. Statist. Assoc.* **83**, 331-338.