Modelling extremal events using Gnedenko distributions

1999 J. Phys. A: Math. Gen. 32 1099

(http://iopscience.iop.org/0305-4470/32/7/003)

J. Phys. A: Math. Gen. **32** (1999) 1099–1113. Printed in the UK PII: S0305-4470(99)99028-5

Modelling extremal events using Gnedenko distributions

Thierry Huillet†§ and Henri-Francois Raynaud†‡

† LIMHP-CNRS, Université Paris XIII, Institut Galilée, 93430 Villetaneuse, France
‡ L2TI, Université Paris XIII, Institut Galilée, 93430 Villetaneuse, France

Received 4 November 1998

Abstract. We address the problem of fitting a Gnedenko distribution to a realization of an IID sample, thereby deciding upon the heavy/light character of the upper tail of the phenomenon under study. An illustration based on earthquake magnitude data is supplied.

1. Introduction

The simplest way to apply a stochastic model to some random natural phenomenon E is to consider its occurrences as so many realizations of independent random variables with identical CPDF (cumulative probability distribution function) F_E. An important feature of this simple model is the rate at which the complementary distribution function $1 - F_E(z)$, which is the probability that the next occurrence of E exceeds the value z, decreases towards zero as z increases. Indeed, in some applications the whole effort of fitting F_E to data may be aimed mainly or solely at accurately identifying this extremal behaviour of the distribution—including the probability of exceeding values higher than the maximum of available data.

In the past two decades, it has been claimed (see, e.g., [18, 17, 1]) that a wide range of natural or social phenomena can be accurately modelled by using so-called 'heavy-tailed' power-law distributions in the form

$$F_E(z) = 1 - (z/z_0)^{-a} \qquad a > 0. \tag{1.1}$$

A simple distinctive feature of this class of distributions is that the log–log plot of the complementary cumulative distribution function is a straight line with negative slope $-a$. Thus, log–log plots of empirical cumulative distribution which look convincingly close enough to a straight line make up the main body of statistical evidence produced to support the claim that power-law distributions are indeed ubiquitous. However, empirical cumulative distribution functions necessarily exhibit at most a limited quasi-linear regime followed by significant curvature. In [16], Laherrère and Sornette argued that such departures from the power-law description should not necessarily be explained by the finite size of the data set, but could result from a deeper departure from the power-law hypothesis. Using rank-ordering statistics to back up their claim, they suggested that occurrences of numerous phenomena, ranging from earthquake death tolls and energies [13, 19] to radio light emissions in galaxies (to which could be added insurance claims or traffic load in communication networks), apparently fit the so-called 'stretched exponential', or sub-exponential, distribution

$$F_E(z) = 1 - \exp[-(z/z_0)^{\delta}] \qquad 0 < \delta < 1. \tag{1.2}$$

§ E-mail address: huillet@limhp.univ-paris13.fr

0305-4470/99/071099+15$19.50

In this paper, we discuss a simple two-parameter class of distribution functions, initially introduced by Gnedenko [14, 15], whose extremal behaviour can be similar to (1.1), (1.2) or 'over-exponential', i.e. similar to (1.2) but with $\delta \geqslant 1$. As a consequence, identifying the value of the model parameters which fits the data in the best way according to some sensible criterion would enable, in some restrictive yet objective sense, to decide whether the phenomenon under study exhibits power-law, subexponential or over-exponential extremal behaviour. This is an essential feature in the context of *risk theory* where E stands for, say, the annual claim amount and where *ruin* probabilities are known [7] to be highly dependent on the light (the Cramér–Lundberg theory) or heavy character of the claim size.

We next establish limit theorems for rank-ordering statistics which can then be used to test whether the identified model is compatible with the empirical cumulative distribution as a whole or with its upper tail only. It turns out that solving both the parameter identification problem and the compatibility tests is made a lot easier by using the transformation $X = \log E$, which therefore emerges as the 'natural' choice of coordinates for this class of problems.

2. Stochastic energy model

2.1. The Gnedenko model

Consider the class of random variables defined as

$$E = (s_0 S)^{1/\delta} \qquad \delta \neq 0 \qquad s_0 > 0 \tag{2.1}$$

where S is an exponentially distributed random variable with mean unity, i.e. with CPDF

$$F_S(s) = 1 - \exp(-s). \tag{2.2}$$

The variable E can be seen as the output of some deterministic 'machine', with parameters (δ, s_0), triggered by the stochastic source of disorder S [3]. Note that in the language of statistical physics, the source S is the random variable with maximum entropy under the constraint that its average value is equal to one.

While s_0 is simply a scaling factor, the parameter δ defines, roughly speaking, the way in which the disorder generated by the source S is spread or concentrated through the transformation (2.1) over the positive real axis. For positive z, the density function (DF) and CPDF of E are obtained easily by combining (2.1) with (1.1), yielding the *Gnedenko distribution:*

$$f_E(z) = \frac{|\delta|}{s_0} z^{\delta-1} \exp\left(-\frac{1}{s_0} z^{\delta}\right) \tag{2.3}$$

$$F_E(z) = \exp\left(-\frac{1}{s_0} z^{\delta}\right) \qquad \text{if} \quad \delta < 0 \tag{2.4}$$

$$F_E(z) = 1 - \exp\left(-\frac{1}{s_0} z^{\delta}\right) \qquad \text{if} \quad \delta > 0. \tag{2.5}$$

When $\delta < 0$, E is a *Fréchet* variable, also called an *exponentially truncated power law* [4] whereas when $\delta > 0$ it is a *Weibull* variable [7]. Depending on the sign of δ, this distribution will exhibit very different extremal behaviour. Let us recall [8] that a distribution is said to be *heavy-tailed* (or *slowly varying*) it there exists some finite strictly positive constant a such that

$$1 - F_E(z) \underset{z\to+\infty}{\sim} z^{-a} L(z) \tag{2.6}$$

where L is some function with regular variation, i.e. such that for all strictly positive t:

$$\lim_{z\to+\infty} \frac{L(tx)}{L(x)} = 1. \tag{2.7}$$

Such distributions only have moments of order less than a. Clearly, when $\delta < 0$,

$$1 - F_E(z) = 1 - \exp\left(-\frac{1}{s_0} z^\delta\right) \underset{z\to+\infty}{\sim} \frac{1}{s_0} z^\delta \tag{2.8}$$

so that the *Fréchet* distribution F_E is heavy-tailed, with only moments of order less than $-\delta$. Note that when $-1 < \delta < 0$, E does not even have a mean value, i.e., with $\mathbb{E}$ the symbol for mathematical expectation, $\mathbb{E}(E) = +\infty$. On the other hand, when $\delta > 0$, E is special case of the so-called *Von Mises* [7], whose complementary cumulative distributions can be written in the form

$$1 - F_E(z) = [1 - F_E(z_0)] \exp\left[-\int_{z_0}^{z} h_E(z)\,\mathrm{d}z\right] \tag{2.9}$$

where the *hazard energy density* h_E defined by this formula verifies

$$\lim_{z\to+\infty} z h_E(z) = +\infty. \tag{2.10}$$

The cumulative distribution of a *Von Mises* variable decreases towards zero faster than hyperbolically, so that these distributions are light-tailed (or rapidly varying). As a consequence, these variables have moments of arbitrary positive order. If in addition the function h_E verifies

$$\lim_{z\to+\infty} h_E(z) = 0 \tag{2.11}$$

the variable E is said to be sub-exponential; otherwise, it is over-exponential. When $\delta > 0$, we get

$$h_E(z) = \frac{\delta}{s_0} z^{\delta-1}. \tag{2.12}$$

Thus, when $0 < \delta < 1$ the *Weibull* variable E is sub-exponential, whereas for $\delta \geqslant 1$ it is over-exponential.

To compute the mean m_E, which as we have noted is defined when $\delta > 0$ or $\delta < -1$, let us introduce an auxiliary random variable V distributed according to a gamma DF with parameter $1 + \lambda/\delta$. Then, the DF for the variable $U = (s_0 V)^{1/\delta}$ is

$$f_U(u) = \frac{|\delta|}{s_0^{1+\lambda/\delta}\Gamma(1+\lambda/\delta)} u^{\lambda+\delta-1} \exp\left(-\frac{1}{s_0} u^\delta\right) \mathbf{1} \qquad (u > 0). \tag{2.13}$$

Using this relation, and noting that f_U is a normalized DF with unitary mass, we get

$$\begin{aligned}\mathbb{E}(E^\lambda) &= \int_0^\infty z^\lambda f_E(z)\,\mathrm{d}z = \int_0^\infty \frac{|\delta|}{s_0} z^{\delta+\lambda-1} \exp\left(-\frac{1}{s_0} z^\delta\right) \mathrm{d}z \\ &= s_0^{\lambda/\delta}\Gamma(1+\lambda/\delta) \int_0^\infty f_U(u)\,\mathrm{d}u = s_0^{\lambda/\delta}\Gamma(1+\lambda/\delta)\end{aligned} \tag{2.14}$$

where Γ is *Euler*'s function. Hence, when $\delta > 0$ or $\delta < -1$ and $\lambda = 1$, the condition $1 + \lambda/\delta > 0$ holds, and

$$m_E = s_0^{1/\delta}\Gamma(1 + 1/\delta). \tag{2.15}$$

The median value of E, say $\overline{m}_E$, defined as the solution of $F_E(\overline{m}_E) = \frac{1}{2}$, is

$$\overline{m}_E = (s_0 \log 2)^{1/\delta}. \tag{2.16}$$

Finally, the distribution of E has a non-zero mode only at the condition that $\delta(\delta - 1) > 0$; in this case, the mode m_E^* is

$$m_E^* = \left(s_0 \frac{\delta - 1}{\delta}\right)^{1/\delta}. \tag{2.17}$$

Remark. *A remarkable property of the* Gnedenko *distribution is the following: let* $E_{n:n} := \max(E_1, \ldots, E_n)$*,* $E_{1:n} := \min(E_1, \ldots, E_n)$ *stand for the maximum and minimum of n IID copies of* E*. Then, for any integer n, as* $\delta < 0$, $n^{1/\delta} E_{n:n} \overset{d}{=} E$*, whereas as* $\delta > 0$, $n^{1/\delta} E_{1:n} \overset{d}{=} E$*, where the symbol* $\overset{d}{=}$ *means that the random variables have the same distributions. Hence, Gnedenko distributions are* max–min *stable.*

2.2. The observable

In many physical situations, the random variable E is not directly observed. Rather, the observed variable is

$$X := \log E. \tag{2.18}$$

The distinctive feature of the logarithmic scale is that it measures the distance between two values through their ratio rather than their difference. Thus, the intensity of noise, as perceived by the human ear, is usually measured in decibels, i.e. using a logarithmic scale. Similarly, earthquake magnitude is determined from the logarithm of the amplitude of waves recorded by seismographs; adjustments are included in the magnitude formula to compensate for the variation in the distance between the various seismographs and the epicentre of the earthquake.

Actually, the observed variable should be $X = \log E + G$, where G is a centered additive measurement noise, e.g. Gaussian. Note that if $Y := e^X$, $Y = E.L$, where L has the lognormal distribution: were the energies to be reconstituted, they would be polluted by a multiplicative lognormal noise under such models. However, we shall suppose that the source of error in the data collecting process is negligible, so that $X = \log E$ is indeed the observable.

Another motivation for working with X rather than with E is that the logarithmic transformation has a regularizing effect on the distribution's tail. Most notably, as indicated above, E does not have a mean when $-1 < \delta < 0$, whereas, as we shall see, X always does—a fact which will be exploited in section 3 to construct an estimator for the distribution's parameters (δ, s_0).

Elementary calculations yield the DF and PDF for the variable X:

$$f_X(x) = \frac{|\delta|}{s_0} \exp\left(\delta x - \frac{1}{s_0} \mathrm{e}^{\delta x}\right) \tag{2.19}$$

$$F_X(x) = \exp\left(-\frac{1}{s_0} \mathrm{e}^{\delta x}\right) \qquad \text{if} \quad \delta < 0 \tag{2.20}$$

$$F_X(x) = 1 - \exp\left(-\frac{1}{s_0} e^{\delta x}\right) \qquad \text{if} \quad \delta > 0. \tag{2.21}$$

We observe that the density of X is invariant under the transformation $(\delta, x) \to (-\delta, -x)$. If $\delta < 0$, this is the PDF of a *Fisher–Tippett* random variable, which is a *Gumbel* distribution in the special case $\delta = -1$ and $s_0 = 1$ [11]. When $\delta > 0$, the authors are unaware of any previous mention of this distribution in the literature.

For all choice of (δ, s_0), the variable X is *Von Mises'*; in addition, it is over-exponential, which means that the tails of its PDF decrease towards zero at exponential rate or faster at both extremities, $\pm\infty$, of the support. Hence, the distribution is 'thin', although very asymmetric.

From (2.14), the Laplace transform $\mathcal{L}_X(\lambda)$ of X is given by

$$\mathcal{L}_X(\lambda) := \mathbb{E}(e^{\lambda X}) = \mathbb{E}(E^{\lambda}) = s_0^{\lambda/\delta} \Gamma(1 + \lambda/\delta). \tag{2.22}$$

This function is thus defined on the range $\lambda > -\delta$, if $\delta > 0$, and $\lambda < -\delta$, if $\delta < 0$, therefore containing the origin $\lambda = 0$ in any case, as required. As a result, the variable X always has convergent moments of arbitrary integral order p, which can be obtained as the *Taylor* coefficients of $\mathcal{L}_X(\lambda)$ at $\lambda = 0$. Denoting as γ *Euler*'s constant, the mean of X is

$$m_X = \mathbb{E}(X) = \frac{1}{\delta}(\log s_0 - \gamma) \simeq \frac{1}{\delta}(\log s_0 - 0.5772). \tag{2.23}$$

The median of X is

$$\overline{m}_X = \frac{1}{\delta}(\log s_0 + \log\log 2) \simeq \frac{1}{\delta}(\log s_0 - 0.3665). \tag{2.24}$$

It should be emphasized that the mean and median have a simple expression in terms of the pair (s_0, δ). In addition, (see e.g. [10] for an exploitation of this fact) the distribution X is always unimodal, and even strongly unimodal, which means that the information function $I_X(x) := -\log f_X(x)$ is strictly convex; its mode is

$$m_X^* = \frac{1}{\delta}\log s_0. \tag{2.25}$$

An important property of the mean–median–mode trio is that for $\delta < 0$,

$$m_X > \overline{m}_X > m_X^* \tag{2.26}$$

whereas the order should be reversed as $\delta > 0$. Since the empirical mean and median provide almost surely (a.s.) convergent estimators for m_X and $\overline{m}_X$, they can be used to decide whether δ is negative or positive, i.e. whether the distribution of E is heavy or light-tailed. Also, since we are dealing with a two-parameters family of distributions, these a.s. convergent estimators will provide an estimator of the pair (δ, s_0). This estimation problem will be discussed in the next section.

3. Parameter estimation problem

Let us consider the problem of deciding whether, and for what choice of the parameter pair (s_0, δ), the distribution function F_X is a good statistical model for a particular data set $(x_1, x_2, \ldots, x_n)$; or equivalently, whether F_E is a good model for the data set $(\mathrm{e}^{x_1}, \mathrm{e}^{x_2}, \ldots, \mathrm{e}^{x_n})$. As mentioned in the introduction, the solution advocated here will be a two-step procedure. The first step, discussed in this section, is to identify the value of the parameter pair (s_0, δ) under the hypothesis that $(x_1, x_2, \ldots, x_n)$ is a realization of an IID sequence $X_1^n := (X_1, X_2, \ldots, X_n)$ with PDF F_X. The second step, discussed in section 4, is to decide whether the identified distribution fits the data.

Because the domain of interest for δ, and thus for the parameter pair (s_0, δ), is non-convex, one needs to distinguish between the two situations $\delta > 0$ and $\delta < 0$, and thus fit (s_0, δ) under the alternative hypothesis $\delta > 0$ and $\delta < 0$. One possible approach would be to compute the maximum likelihood estimator for (s_0, δ). This estimator is defined as the value of the parameter pair $(\hat{s}_n, \hat{\delta}_n)$ which maximizes the likelihood function $L_n := \prod_{m=1}^n f_X(x_m)$; the corresponding distribution, whenever it exists, is precisely the one for which the realization $(x_1, x_2, \ldots, x_n)$ is the more likely to occur [5]. In this case, the maximum likelihood estimator could be obtained as follows: compute the two maximum likelihood estimators $(\hat{s}_{0,n}^+, \hat{\delta}_n^+)$ and $(\hat{s}_{0,n}^-, \hat{\delta}_n^-)$ maximizing the likelihood function under the hypothesis $\delta > 0$ and $\delta < 0$, evaluate the corresponding candidate maxima $L_n^+ = L_n(\hat{s}_{0,n}^+, \hat{\delta}_n^+)$ and $L_n^- = L_n(\hat{s}_{0,n}^-, \hat{\delta}_n^-)$, and then retain the best of the two, i.e. $(\hat{s}_n, \hat{\delta}_n) = (\hat{s}_{0,n}^+, \hat{\delta}_n^+)$ if $L_n^+ > L_n^-$, $(\hat{s}_{0,n}, \hat{\delta}_n) = (\hat{s}_{0,n}^-, \hat{\delta}_n^-)$ otherwise. However, it turns out that we have no guarantee that the likelihood function L_n attains its maximum at some point inside its domain, which is unfortunately open, so that the very maximum likelihood approach may not make sense; furthermore, we cannot even guarantee that the likelihood function L_n is convex into the two regions $\delta > 0$ and $\delta < 0$, so that it could be difficult to compute in a reliable way the maxima of L_n in these two domains, should they exist.

An alternative approach is to substitute the empirical mean and median of the sample $(X_1, X_2, \ldots, X_n)$ in (2.23) and (2.24), and to solve these equations in (s_0, δ).

Denote as $\overline{X}_n$ the cumulative sum of the sample, i.e.

$$\overline{X}_n = \sum_{m=1}^{n} X_m \tag{3.1}$$

and as $X_{1:n}^{n:n} := (X_{1:n}, \ldots, X_{n:n})$ the ordered version of X_1^n, which means

$$X_{1:n} < \cdots < X_{m:n} < \cdots < X_{n:n} \tag{3.2}$$

so that the empirical mean and median are, respectively, $\frac{1}{n}\overline{X}_n$ and $X_{[n/2]:n}$.In this way, we obtain the estimator:

$$\hat{s}_{0,n} = \exp\left[\frac{\gamma X_{[n/2]:n} + \log\log 2\frac{1}{n}\overline{X}_n}{X_{[n/2]:n} - \frac{1}{n}\overline{X}_n}\right] \tag{3.3}$$

$$\hat{\delta}_n = \frac{\gamma + \log\log 2}{X_{[n/2]:n} - \frac{1}{n}\overline{X}_n}. \tag{3.4}$$

Note that the mode (most probable state) is therefore estimated to be located at

$$\frac{1}{\hat{\delta}_n}\log \hat{s}_{0,n} = \frac{\gamma X_{[n/2]:n} + \log\log 2.\frac{1}{n}\overline{X}_n}{\gamma + \log\log 2}. \tag{3.5}$$

Note that the sign of $\hat{\delta}_n$, which controls the heavy or light nature of the tail for the original variable E, depends on the relative position of the empirical mean and median. This provides a simple test for the extremal behaviour of E. However, the empirical median cannot be computed recursively, which is a minor drawback in ‘dynamical’ situations where the sample size increases.

4. Model-data fitness

The adequacy of the identified distribution corresponding to $(\hat{s}_{0,n}, \hat{\delta}_n)$ with the empirical distribution can be tested from three different point of views, depending on what practical questions the model is supposed to answer: globally (*Kolmogorov–Smirnov* test); in the central body of the data (order statistics); for its tails (statistics of extremes). The statistics associated with these tests is based on the ordered sample $(X_{1:n}, \ldots, X_{n:n})$ and on the quantile distribution function (QDF) of X:

$$F_X^{-}(p) := \inf(x : F_X(x) > p). \tag{4.1}$$

This QDF is easily computed from (2.20) and (2.21):

$$F_X^{-}(p) = \frac{1}{\delta}\log[-s_0\log(p)] \qquad \text{if} \quad \delta < 0 \tag{4.2}$$

$$F_X^{-}(p) = \frac{1}{\delta}\log[-s_0\log(1-p)] \qquad \text{if} \quad \delta > 0. \tag{4.3}$$

4.1. Kolmogorov–Smirnov test

The *Kolmogorov–Smirnov* test [2] enables one to decide whether an IID sample $X_1^n = (X_1, X_2, \ldots, X_n)$ has been generated with some guessed (theoretical) probability distribution function F_X. Denote respectively as F_n and F_n^- the empirical PDF and QDF of the sample, i.e.

$$F_n(x) := \frac{1}{n}\sum_{m=1}^{n} \mathbf{1}(X_{m:n} \leqslant x) \tag{4.4}$$

$$F_n^-(p) := \inf(x : F_n(x) > p). \tag{4.5}$$

The *Kolmogorov–Smirnov* test is based on the random variable

$$\sup_x |F_n(x) - F_X(x)| \tag{4.6}$$

which measures some distance between the empirical and theoretical PDFs. Using the transformation $x = F_X^-(p)$, this is also

$$\sup_{p\in[0,1]} |F_n^U(p) - p| \tag{4.7}$$

where F_n^U is the empirical PDF of an IID uniform sequence on the interval (0, 1), so that

$$F_n^U(p) := F_n(F_X^-(p)) = \frac{1}{n}\sum_{m=1}^{n} \mathbf{1}(U_{m:n} \leqslant x) \tag{4.8}$$

with $U_{m:n} := F_X(X_{m:n})$. Using this notation, it is shown that

$$\sqrt{n} \sup_{p\in[0,1]} |F_n^U(p) - p| \underset{n\uparrow\infty}{\overset{d}{\rightarrow}} M \tag{4.9}$$

where the variable M is the absolute supremum of a *Brownian* bridge which admits the PDF

$$F_M(z) = 1 - 2\sum_{k\geqslant 1}(-1)^{k-1}\exp(-2k^2z^2). \tag{4.10}$$

Hence, searching for the level value $\gamma_n(\alpha)$ such that

$$P\left\{\sup_{p\in[0,1]} |F_n^U(p) - p| > \gamma_n(\alpha)\right\} = \alpha \tag{4.11}$$

for small α (say $\alpha = 0.05$) yields

$$\gamma_n(\alpha) \simeq \frac{1}{\sqrt{n}}\left[\frac{\log(2/\alpha)}{2}\right]^{1/2}. \tag{4.12}$$

Thus, the *Kolmogorov–Smirnov* test works as follows: (a) transform the original sample into $U_1^n := (F_X(X_1), \ldots, F_X(X_n))$; (b) compute

$$\max_{m=1,\ldots,n} |m/n - U_{m:n}| = \sup_{p\in[0,1]} |F_n^U(p) - p| \tag{4.13}$$

(c) if this number exceeds $\gamma_n(\alpha)$, reject the hypothesis that the sample has been generated with the theoretical PDF $F_X(x)$, otherwise accept it. α is the probability to decide that the sample is not a realization of the distribution F_X when it really is.

Related tests, such as the range test and the energy test, are also available [9].

4.2. Rank-ordering statistics

As the sample size n increases, one would expect that the mth largest component, the ordered sample $(X_{1:n}, \ldots, X_{n:n})$ converges, towards $F_X^-(m/n)$. In fact, this holds true only on the condition that $\min(m, n-m) \to \infty$, which in practical terms means the 'central body' of the sample, excluding the smallest and largest quantiles for which different asymptotic results hold (see below). It is well known [6] that the joint probability density of the ordered sample $(X_{1:n}, \ldots, X_{n:n})$ is

$$f_{X_{1:n},\ldots,X_{n:n}}(x_1, \ldots, x_n) = n! \prod_{m=1}^{n} f_X(x_m) \mathbf{1}(x_1 < \cdots < x_n). \tag{4.14}$$

The marginal distribution of $X_{m:n}$ can be derived from the observation that the events $\{X_{m:n} \leqslant x\}$ and $\{\overline{B}_n(x) \geqslant m\}$ coincide, where $\overline{B}_n(x) := \sum_{m=1}^{n} B_m(x)$ is the binomial cumulative sum of the IID Bernoulli series with general term

$$B_m(x) := \mathbf{1}(X_m \leqslant x) \qquad m = 1, \ldots, n. \tag{4.15}$$

Consequently, $\overline{B}_n(x)$ is a binomial variable with mean $nF_X(x)$ and variance $nF_X(x)(1 - F_X(x))$. As a result, the PDF of $X_{m:n}$ is

$$F_{X_{m:n}}(x) = \sum_{l=m}^{n} \binom{n}{l} F_X(x)^l (1 - F_X(x))^{n-l}. \tag{4.16}$$

From the central limit theorem

$$\frac{\overline{B}_n(x) - nF_X(x)}{\sqrt{nF_X(x)(1 - F_X(x))}} \underset{n\uparrow\infty}{\overset{d}{\to}} \mathcal{N}(0, 1) \tag{4.17}$$

one can easily deduce that, as $\min(m, n-m) \to \infty$,

$$\tilde{X}_{m:n} := \frac{\sqrt{n} f_X(F_X^-(m/n))}{\sqrt{m/n(1 - m/n)}} [X_{m:n} - F_X^-(m/n)] \overset{d}{\to} \mathcal{N}(0, 1) \tag{4.18}$$

where $\mathcal{N}(0, 1)$ denotes the Gaussian distribution with zero mean and unitary variance. Thus, there exists a rescaled adjusted version of $X_{m:n}$, which converges in distribution to a normal variable. Here the scaling function (or local fluctuation) is

$$\frac{\sqrt{m/n(1 - m/n)}}{\sqrt{n} f_X(F_X^-(m/n))} \tag{4.19}$$

which tends to zero under the considered asymptotics. A special case of interest is $m := [n/2]$. For large values of n, $F_X^-([n/2]/n) \sim F_X^-(\frac{1}{2})$ is the theoretical median, so that the empirical median $X_{[n/2]:n}$ verifies

$$2\sqrt{n} f_X(F_X^-(\tfrac{1}{2}))[X_{[n/2]:n} - F_X^-(\tfrac{1}{2})] \underset{n\uparrow\infty}{\overset{d}{\to}} \mathcal{N}(0, 1). \tag{4.20}$$

These results can be extended to a multi-dimensional ordered sequence of given size k belonging to the central body of the data. Let $1 \leqslant m_1 < \cdots < m_k \leqslant n$ be some increasing sequence of integers. Reasoning along the same lines as above, the joint PDF of the vector $\boldsymbol{X}_{m_k:n} := (X_{m_1:n}, \ldots, X_{m_k:n})'$ can be derived from the multinomial character of the vector $\overline{\boldsymbol{B}}_n(\boldsymbol{x}_k) := (\overline{B}_n(x_1), \ldots, \overline{B}_n(x_k))'$, and is given by

$$\begin{aligned} F_{X_{m_1:n},\ldots,X_{mk:n}}(x_1, \ldots x_k) &= \sum_{l_1=m_1}^{n} \cdots \sum_{l_k=m_k}^{n} \frac{n!}{\prod_{j=1}^{k} l_j!(n - \sum_{j=1}^{k} l_j)!} \\ &\quad \times \prod_{j=1}^{k} F_X(x_j)^{l_j} \left(1 - \sum_{j=1}^{k} F_X(x_j)\right)^{n - \sum_{j=1}^{k} l_j}. \end{aligned} \tag{4.21}$$

On the condition that $\min_{1\leqslant j\leqslant k}[\min(m_j, n-m_j)] \to \infty$, the vector $\boldsymbol{X}_{\boldsymbol{m}_k:n} := (X_{m_1:n}, \ldots, X_{m_k:n})'$ converges towards the multi-dimensional QDF $\boldsymbol{F}_X^-(\boldsymbol{m}_k/n) := (F_X^-(m_1/n), \ldots, F_X^-(m_k/n))'$. More precisely, we get the following asymptotic behaviour:

$$\tilde{\boldsymbol{X}}_{\boldsymbol{m}_k:n} := \sqrt{n}[\operatorname{diag} f_X(\boldsymbol{F}_X^-(\boldsymbol{m}_k/n))]H(\boldsymbol{m}_k/n)^{-1/2}(\boldsymbol{X}_{\boldsymbol{m}_k:n} - \boldsymbol{F}_X^-(\boldsymbol{m}_k/n)) \underset{n\uparrow\infty}{\overset{d}{\to}} \mathcal{N}(0, \mathbf{1}_k) \tag{4.22}$$

where $\mathbf{1}_k$ the $k \times k$ identity matrix and diag $f_X(\boldsymbol{F}_X^-(\boldsymbol{m}_k/n))$ is the diagonal matrix with $f_X(F_X^-(m_i/n))$ as $(i \times i)$—entry, $i = 1, \ldots, k$.

In addition, the covariance matrix $H(\boldsymbol{m}_k/n)$ which appears in the previous equation is defined by

$$H(\boldsymbol{m}_k/n)_{i,j} := -(m_i/n)\cdot(m_j/n) \qquad \text{if} \quad i \neq j \tag{4.23}$$

$$H(\boldsymbol{m}_k/n)_{i,i} := (m_i/n)\cdot(1 - m_i/n). \tag{4.24}$$

Thus, under the hypothesis that $X_1^n := (X_1, \ldots, X_n)$ is an IID sequence with common PDF F_X,

$$\|\tilde{\boldsymbol{X}}_{\boldsymbol{m}_k:n}\|_2^2 \underset{n\uparrow\infty}{\overset{d}{\to}} \chi_k^2 \tag{4.25}$$

where χ_k^2 is a Chi-2 variable with k degrees of freedom.

Using (4.22), one can test whether the sample is compatible with the 'central body' of the distribution F_X as follows: (a) select a sequence $1 \leqslant m_1 < \cdots < m_k \leqslant n$; (b) evaluate the rescaled statistics $\tilde{\boldsymbol{X}}_{\boldsymbol{m}_k:n}$; (c) select a level α and determine the level value $\gamma_n(\alpha)$ for which $P\{\chi_k^2 > \gamma_n(\alpha)\} = (\alpha$; c) if $\|\tilde{\boldsymbol{X}}_{\boldsymbol{m}_k:n}\|_2^2 > \gamma_n(\alpha)$, reject the hypothesis that the bulk of the sample has been generated with the theoretical PDF $F_X(x)$, otherwise accept it.

Remark. *It follows from these formulae that tests which are based on such rank ordering statistics only concern the 'central body' of the data, not the extremes. Indeed, the central limit theorem holds in the asymptotics* $\inf(m, n-m) \to \infty$. *Thus, both n and the rank m should tend to infinity, for example $n \to \infty$, $m \to \infty$, with the ratio m/n held fixed at* $\alpha \in (0, 1)$. *These results are invalid in the extreme ends of the sample, as nothing has been said so far, concerning the asymptotics $n \to \infty$, $m \to \infty$, with $m + p = n$ for p a* fixed *integer (in particular* $p = 0$*).*

For this part of the data, the Fisher–Tippett theorem is the key tool for limit theorems. We examine the part of this problem of interest to our purposes in the following.

4.3. Statistics of extremes

First observe the obvious fact that $X_{n:n} \overset{\text{a.s.}}{\to} +\infty$, as $n \uparrow \infty$. This observation does not enclose too much information and one would like a deeper insight on how the order of magnitude of the maximum evolves, as $n \uparrow \infty$. This can be done by defining the level value $x_n^*(\gamma)$ associated with a small positive γ (say $\gamma = 0.05$) as the solution of

$$n[1 - F_X(x_n^*(\gamma))] = \gamma. \tag{4.26}$$

Equivalently, $x_n^*(\gamma)$ can be defined as the solution of

$$P\{X_{n:n} > x_n^*(\gamma)\} = 1 - \mathrm{e}^{-\gamma}. \tag{4.27}$$

Knowledge of $x_n^*(\gamma)$ can be important in applications. Take the example of the hydrologist (or telecommunication engineer) whose problem is to dimension a dam (or a buffer) when the random inputs to their system are assumed to form an IID sequence. They might be interested by designing a dam (a buffer) whose height (size), $x_n^*(\gamma)$, is such that the overflow probability $\rho := 1 - \mathrm{e}^{-\gamma}$ over the laps of time n is small. Thus $x_n^*(\gamma)$ is an excessively high value for the n-sample.

When F_X is a *Von Mises* distribution, it is well known [12] that the asymptotic behaviour for the fluctuation of the difference between $x_n^*(\gamma)$ and the sample maximum $X_{n:n} = \max(X_1^n)$ when $n \uparrow \infty$ is given by

$$h_X[x_n^*(\gamma)][X_{n:n} - x_n^*(\gamma)] \underset{n\uparrow\infty}{\overset{d}{\to}} G_\gamma \tag{4.28}$$

where h_X is the hazard energy density of the *Von Mises* distribution of X and G_γ is the *Gumbel* variable with DF

$$f_{G_\gamma}(t) = \gamma \exp[-(t + \gamma \mathrm{e}^{-t})]. \tag{4.29}$$

Note that from (2.20) and (2.21), the explicit form of the hazard energy density h_X is

$$h_X(x) = -\frac{\delta \exp\left(\delta x - \frac{1}{s_0}\mathrm{e}^{\delta x}\right)}{s_0\left[1 - \exp\left(-\frac{1}{s_0}\mathrm{e}^{\delta x}\right)\right]} \qquad \text{if} \quad \delta < 0 \tag{4.30}$$

$$h_X(x) = \frac{\delta}{s_0}\exp(\delta x) \qquad \text{if} \quad \delta > 0 \tag{4.31}$$

for which it can be checked the *Von Mises* property that $x h_X(x) \to \infty$ when $x \uparrow \infty$; in addition, h_X does not vanish at infinity, in any case (X is over-exponential).

The proof for this result can easily be extended to obtain the asymptotic behaviour for the $(n-p)$th largest value $X_{n-p:n}$. The motivation for using such statistics is that the maximum is highly sensitive to 'outliers', i.e. abnormally high values in the sample resulting from errors in the data collection process such as typing errors. In practice, one should therefore select a value of p such that $(n-p)/n \simeq 1$ and that p/n is higher than the occurrence probability of an outlier.

Define $x_{n,p}^*(\gamma)$ as the solution of

$$(n-p)[1 - F_X(x_{n,p}^*(\gamma))] = \gamma. \tag{4.32}$$

Then, as $n \uparrow \infty$

$$h_X[x_{n,p}^*(\gamma)][X_{n-p:n} - x_{n,p}^*(\gamma)] \underset{n\uparrow\infty}{\overset{d}{\to}} G_{\gamma,p} \tag{4.33}$$

where $G_{\gamma,p}$ has the density

$$f_{G_{\gamma,p}}(t) = \frac{\gamma^{p+1}}{p!}\exp(-[(p+1)t + \gamma \mathrm{e}^{-t}]). \tag{4.34}$$

We now briefly indicate how these results can be used in practice. Fix a small real number, say $\alpha = 0.05$. We would like to compute the number $\epsilon_n(\alpha, \gamma)$ defined by $P\{|X_{n-p:n} - x_{n,p}^*(\gamma)| > \epsilon_{n,p}(\alpha,\gamma)\} = \alpha$. The number $\epsilon_n(\alpha, \gamma)$ is therefore the radius of the ball centred at $x_{n,p}^*(\gamma)$ which is likely (at level α) to enclose the $(n-p)$th largest value $X_{n-p:n}$.

From (4.33), $\epsilon_{n,p}(\alpha, \gamma)$ can be effectively computed by

$$P\{|G_{\gamma,p}| > \epsilon_{n,p}(\alpha,\gamma) h_X[x_{n,p}^*(\gamma)]\} = \alpha. \tag{4.35}$$

This construction therefore yields an approximation of the width of the confidence interval of the maximum around $x_{n,p}^*(\gamma)$. These constructions exhibit two parameters under control of the modeller. The first one, γ, which appears in the definition of $x_{n,p}^*(\gamma)$ is needed to decide what

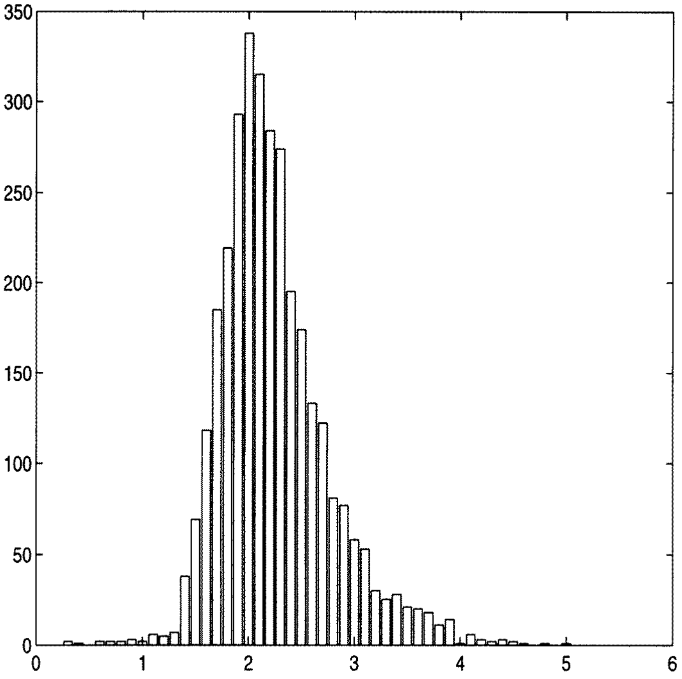

Figure 1. Histogram of magnitude measurements—Richter scale.

an excessively high value for the data X_1^n is. The second, α, is needed to design a confidence interval around $x^*_{n,p}(\gamma)$.

To summarize, the test works as follows: (a) select p and evaluate $X_{n-p:n}$; (b) select γ and α, and determine the level value $\epsilon_{n,p}(\alpha,\gamma)$ for which (4.35) holds; (c) if $|X_{n-p:n} - x^*_{n,p}(\gamma)| > \epsilon_{n,p}(\alpha,\gamma)$, reject the hypothesis that the tail of the sample has been generated with the theoretical PDF $F_X(x)$, otherwise accept it.

5. A parochial experiment

The procedure described above was tested on earthquake magnitude data obtained from the Northern California Earthquake Data Center (NCEDC). The data set comprised all earthquakes recorded from November 1995 to October 1998 in a polygon corresponding roughly to the boundaries of metropolitan France, excluding Corsica. November 1995 was chosen as a starting point because prior to this date, this catalogue apparently assigned a magnitude of one to all recorded small earthquakes. Records of earthquake magnitude are well suited to our purposes because of the logarithmic basis of the scale. However, on the Richter scale, magnitude is expressed in whole numbers and decimal fractions. This round-up effect makes the raw data inconsistent with any continuous model of the probability distribution function. To overcome this minor obstacle, we regularized the data by adding to each recorded value an IID random noise uniformly distributed in the range $(-0.05$–$0.05\,)$.

The sample contained $n = 3245$ recorded events ranging from 0.3 to 5.0 on the Richter scale. There seems to be roughly 1000 registered events per year. The histogram for the raw data is presented in figure 1, showing an asymmetrical distribution which seems consistent

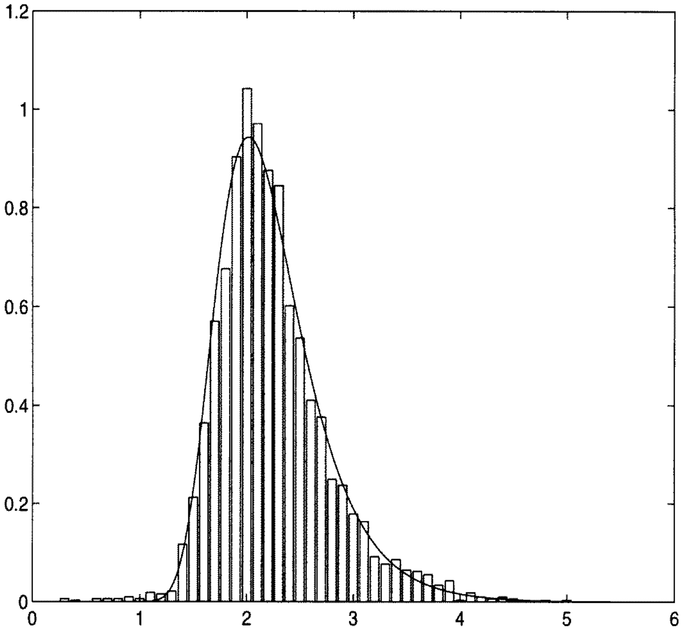

Figure 2. Estimated DF and normalized empirical histogram.

with a *Gumbel* model. The empirical mean and median for the regularized sample X_1^n are:

$$\frac{1}{n}\overline{X}_n \simeq 2.2399 \qquad X_{[n/2]:n} \simeq 2.1577. \tag{5.1}$$

In addition, the estimated most probable value for X is from (3.5)

$$\frac{1}{\hat{\delta}_n} \log \hat{s}_{0,n} \simeq 2.0148. \tag{5.2}$$

Applying the estimation procedure in section 3, we get using (3.3) and (3.4)

$$\hat{s}_{0,n} \simeq 5.7103 \times 10^{-3} \qquad \hat{\delta}_n = -2.5638. \tag{5.3}$$

As a result, there *is* statistical evidence that the estimated distribution for the energy ratio $E = \exp(X)$ is heavy-tailed, but with convergent moments of order less than 2.5638 (including the mean and variance).

Figure 2 shows the estimated DF versus the normalized empirical histogram. The fit appears good, except, maybe, for the lowest values of X. In our opinion, this reflects the fact that earthquakes of very small magnitude (microearthquakes are of magnitude lower than one on the Richter scale) may elude detection.

Proceeding to the model-data fitness tests of section 4, we first compute the *Kolmogorov–Smirnov* statistics in (4.13)

$$\max_{m=1,\ldots,n} |m/n - U_{m:n}| = \sup_{p\in[0,1]} |F_n^U(p) - p| \simeq 1.6952 \times 10^{-2} \tag{5.4}$$

to be compared, for the risk $\alpha = 0.05$, with the level value

$$\gamma_n(\alpha) \simeq \frac{1}{\sqrt{n}} \left[\frac{\log(2/\alpha)}{2}\right]^{1/2} \simeq 2.3841 \times 10^{-2}. \tag{5.5}$$

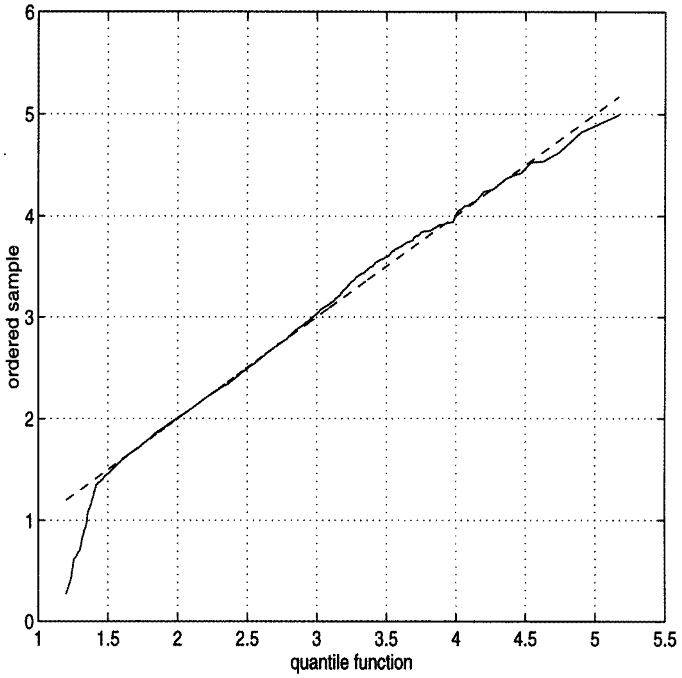

Figure 3. Ordered sample versus quantile function.

Therefore, the hypothesis that the sample has been generated, as a whole, with the theoretical PDF $F_X(x)$ in (2.20) with parameters

$$(\hat{s}_{0,n}, \hat{\delta}_n) = (5.7103 \times 10^{-3}, -2.5638) \tag{5.6}$$

should be accepted.

We now perform a rank-ordering test for the central body of the sample, following section 4.2. We extracted a subsample of size $k = 80$ with indices m_j ranging from $m_1 = 500$ to $m_{80} = 2870$, with a step of $m_j–m_{j-1} = 30$. Using (4.22)–(4.24), we computed the statistics

$$\|\tilde{\boldsymbol{X}}_{m_k:n}\|_2^2 \simeq 43.667. \tag{5.7}$$

These statistics are to be compared with the level which a Chi-2 variable with $k = 80$ degrees of freedom has probability $\alpha = 0.05$ of exceeding. We found

$$P\{\chi_{80}^2 > 101.9\} \simeq 0.05. \tag{5.8}$$

Therefore, the hypothesis that the ordered subsample $\boldsymbol{X}_{m_k:n} := (X_{m_1:n}, \ldots, X_{m_k:n})'$ has been generated with the theoretical PDF $F_X(x)$ with parameters $(\hat{s}_{0,n}, \hat{\delta}_n)$ should be accepted. To support this result, the ordered sample has been plotted against the quantile distribution function $\{F_X^{-}(m/n), m = 1, \ldots, n\}$ (figure 3). This plot reveals, as expected, a misfit for both the lowest and highest quantiles. On the contrary, the fitness seems reasonably good for the central part of the data.

To test the goodness of the fit for the upper tail of the distribution, according to the procedure in section 4.3, we choose to consider the statistics for the empirical maximum $(p = 0)$, assuming the probability of occurrence of an outlier in the data to be zero. In this case, the presence of outliers would mean that the database includes nonexistent comparatively

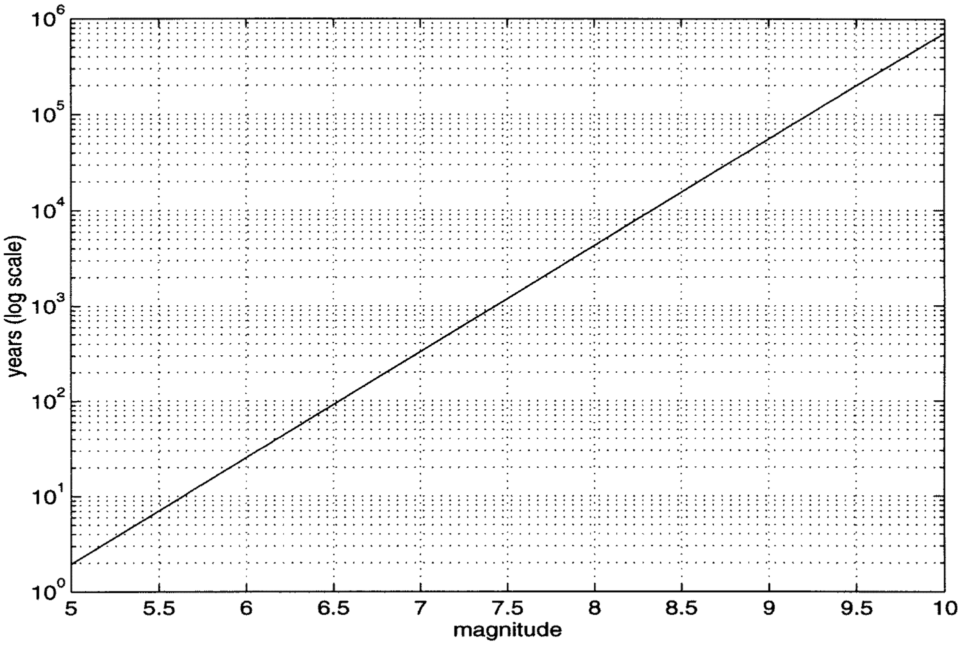

Figure 4. Estimated average time before next earthquake of magnitude exceeding x.

large earthquakes, which appears highly unlikely. We first need to compute, for $\gamma = 0.05$, the value $x_n^*(\gamma)$ defined by (4.26), and find

$$x_n^*(\gamma) \simeq 6.3367. \tag{5.9}$$

Recall that the maximum of an IID sample with distribution F_X and size n has probability $1-\exp(-\gamma) \simeq 0.05$ of exceeding this 'excessively high' value for the n-sample. The empirical maximum $X_{n:n}$ of the sample is 5.0, so that

$$|X_{n:n} - x_n^*(\gamma)| \simeq 1.3367. \tag{5.10}$$

This statistics should be compared with the solution $\epsilon_{n,0}(\alpha, \gamma)$ of

$$P\{|G_\gamma| > \epsilon_{n,0}(\alpha, \gamma)h_X[x_n^*(\gamma)]\} = \alpha = 0.05 \tag{5.11}$$

with

$$P\{|G_\gamma| > t\} = 1 - \exp(-\gamma \mathrm{e}^{-t}) + \exp(-\gamma \mathrm{e}^{t}). \tag{5.12}$$

For this value of γ, we find

$$\epsilon_{n,0}(\alpha, \gamma) \simeq 1.6 \tag{5.13}$$

so that the hypothesis that the tail of the sample has been generated with the theoretical PDF $F_X(x)$ with parameters $(\hat{s}_{0,n}, \hat{\delta}_n)$ should also be accepted.

An obvious question of interest is: how long should one wait before the next earthquake of magnitude greater than a given level? In [20], this problem is addressed through a detailed statistical study of the time separating two consecutive events. A more pedestrian approach is to compute the 'mean time between failure', which is defined as

$$N_x := \inf(n : X_n > x) \tag{5.14}$$

for some magnitude level x. Assuming that the number of earthquakes per year is constant, equal to $n/3 \simeq 1082$, the estimated average number of years one would have to wait before the next earthquake of magnitude greater than x is therefore

$$\frac{\mathbb{E}\{N_x\}}{n/3} = \frac{3}{n[1 - F_X(x)]}. \tag{5.15}$$

Figure 4 presents this estimated average time as a function of the magnitude, zooming in the extreme range $5 \leqslant x \leqslant 9$. In order to interpret such plots, one should recall that magnitudes exceeding $x = 8$ correspond to very large and rare events (on average, one earthquake of such size occurs somewhere in the world each year). Interestingly, an earthquake of magnitude $x = 6.4$, which is very unlikely to occur in France according to (4.26), should occur on average every 70 years or so.

Acknowledgments

The authors are grateful to the northern california earthquake data center NCEDC and to the Members Networks of the Council of the National Seismic System (CNSS) for supplying the raw data on earthquakes magnitudes through their internet site†.

References

[1] Aharony A and Feder J (ed) 1989 Fractals in physics *Physica* D **38** 1–3

[2] Billingsley P 1968 *Convergence of Probability Measures* (New York: Wiley)

[3] Bouchaud J P and Mézard M 1997 Universality classes for extreme value statistics *J. Phys. A: Math. Gen.* **30** 7997–8015

[4] Cont R and Bouchaud J-P 1997 Herd behaviour and aggregate fluctuations in financial markets *Preprint* Cond-Mat/ week: 48–97 no 9712318

[5] Cramér H 1946 *Mathematical Methods of Statistics* (Princeton, NJ: Princeton University Press)

[6] David H A 1981 *Order Statistics* 2nd edn (New York: Wiley)

[7] Embrechts P, Klüppelberg C and Mikosh T 1997 Modelling extremal events *Application of Mathematics* vol 33 (Berlin: Springer)

[8] Feller W 1971 *An Introduction to Probability Theory and its Applications* vol 2 (New York: Wiley)

[9] Feller W 1951 The asymptotic distribution of the range of sums of independent random variables *Ann. Math. Stat.* **22** 427–32

[10] Frisch U and Sornette D 1997 Extreme deviations and applications *J. Physique* I **7** 1155–71

[11] Gumbel E J 1958 *Statistics of Extremes* (Columbia: Columbia University Press)

[12] Galambos J 1978 *The Asymptotic Theory of Extreme Order Statistics* (New York: Wiley)

[13] Knopoff L and Sornette D 1995 Earthquake death tolls *J. Physique* I **5** 1681–8

[14] Koroliouk, V, Portenko N, Skorokhod A and Tourbine A 1983 *Aide-mémoire de Théorie des Probabilités et de Statistique Mathématique* (Moscow: MIR)

[15] Gnedenko B V 1943 Sur la distribution limite du maximum d'une série aléatoire *Ann. Math.* **44** 423–53

[16] Laherrère J and Sornette D 1998 Stretched exponential distributions in nature and economy: fat tails with characteristic scales *Eur. Phys. J.* B **2** 525–39

[17] Mandelbrot B B 1983 *The Fractal Geometry of Nature* (New York: Freeman)

[18] Pareto V 1896 *Cours d'Economie Politique*
1965 Reprinted as a volume of *Oeuvres Complètes* (Geneva: Droz)

[19] Sornette D, Knopoff L, Kagan Y Y and Vanneste C 1996 Rank-ordering statistics of extreme events–application to the distribution of large earthquakes *J. Geophys. Res.* **101** 13 883–93

[20] Sornette D and Knopoff L 1997 The Paradox of the expected time until the next earthquake *Bull. Seismo. Soc. Am.* **87** 789–98

† webpage http://quake.geo.berkeley.edu/cnss/cnss.catalog.html