The Mean/Max Statistic in Extreme Value Analysis

Paul Rochet and Isabel Serra

Abstract

Most extreme events in real life can be faithfully modeled as random realizations from a Generalized Pareto distribution, which depends on two parameters: the scale and the shape. In many actual situations, one is mostly concerned with the shape parameter, also called tail index, as it contains the main information on the likelihood of extreme events. In this paper, we show that the mean/max statistic, that is the empirical mean divided by the maximal value of the sample, constitutes an ideal normalization to study the tail index independently of the scale. This statistic appears naturally when trying to distinguish between uniform and exponential distributions, the two transitional phases of the Generalized Pareto model. We propose a simple methodology based on the mean/max statistic to detect, classify and infer on the tail of the distribution of a sample. Applications to seismic events and detection of saturation in experimental measurements are presented.

Keywords Generalized Pareto model; tail index; hypothesis testing
MSC (2000): 62G32; 62F03; 62G15

1 Introduction

Two fundamental results marked the starting point of Extreme Value Theory (EVT): Fisher-Tippet-Gnedenko and Pickands-Balkemax-de Haan Theorems. These theorems enabled to characterize the asymptotic behavior of extreme values by a real number k, called tail index (we refer to [15], [2], [4, 9, 10] and [16] for a survey). Ever since Gnedenko's original paper, estimation of the tail index has become a major concern in the literature, for which even the most used techniques can produce unsatisfactory results in many situations. For instance, several authors have highlighted the estimation problems that arise from the second approach, see [3, 13, 22].

Pickands-Balkemax-de Haan Theorem has widened the use of the Generalized Pareto Distribution (GPD) in extreme values theory as a model for tails. According to the behavior of the probability density functions of the GPD, we can distinguish three submodels corresponding to $k < 0$, $0 < k < 1$ and $k > 1$ respectively, separated by the exponential distribution ($k = 0$) and the uniform distribution ($k = 1$). Since no method is able to provide a satisfying estimation of the tail index regardless of the true model (see [7]), the parameter inference should be made subsequently to the choice of submodel, even more so for small samples. The first recommendation is to specify the submodel of the GPD a priori, in order to estimate the parameters. Therefore, we propose to use a test procedure to distinguish between the two transitional phases of the tail index in GPD models: the exponential and uniform distributions. Due to the monotonic behavior of the test statistic with respect to the tail index of the GPD, the test can be used for classification purposes.

Tests on separate families of hypotheses have been originally studied in [6] who proposed a generalization of the Neyman-Pearson principle, based on a likelihood ratio criterion. The so-called maximum-likelihood ratio test is not necessarily optimal, although it provides a practical methodology to test a wide range of hypotheses. Examples studied in the literature include invariant tests of exponential versus normal or uniform distributions in [20], normal versus Cauchy distribution in [12] and so on.

In this paper, we show that Cox's maximum likelihood ratio test for uniform versus exponential distributions is the most powerful among scale-free tests. It relies on the ratio τ_n between the empirical mean and sample maximum, whose behavior is highly dictated by the tail of the distribution, thus making it a main feature of the GPD framework. The mean/max statistic τ_n then provides a simple yet effective way to detect the tail behavior in a Generalized Pareto model. In fact, the tail classification method based solely on the value of τ_n produces extremely conclusive results that compare favorably with the more computationally expensive inference methods such as Zhang and Stephens [22], Song and Song [18] or maximum likelihood.

The problem of bad classification in GPD estimation was recently detected [7] but no satisfactory solution has been proposed so far. Summarizing, the quality of inference methods for the tail index k in the GPD highly depends on the underlying submode, whether it is Model A ($k > 1$), Model B ($0 \leq k \leq 1$) or Model C ($k < 0$). Because standard methods tend to specialize to certain regions of the values of the parameter, it is crucial to separate the GPD model into these three submodels reflecting the different behavior of the tails.

The test on separate families of hypotheses to distinguish between uniform and exponential distributions is described in Section 2. In Section 3, we show how to extend the test procedure for tail classification and detection in Generalized Pareto models or more general distributions. In Section 4, we propose a simple general recipe to infer on the tail of the distribution of a sample and an application on global seismic activity data is presented. Theoretical results regarding the distribution of the mean/max statistic are discussed in the Appendix.

2 Test to distinguish between uniform and exponential tail distributions

Let $X = (X_1, \ldots, X_n)$ be a sample of independent identically distributed variables drawn from a distribution on $[0, +\infty)$ assumed to be either uniform on some interval $[0, \theta]$ with $\theta > 0$ or exponential with parameter $\lambda > 0$. As it is usually the case when dealing with separate families of hypotheses, a uniformly most powerful test does not exist because the likelihood ratio statistic depends on the true value of the parameter. The maximum likelihood ratio statistic was proposed in [6] as a way to generalize Neyman-Pearson's principle to families of distributions, although the optimality of the test is no longer backed up by the theory. Nevertheless, we show that in this situation, Cox's maximum likelihood ratio test is the most powerful among a large class of tests whose distribution is invariant within each hypothesis. For such tests, both the level and power are determined by the common distributions of the statistic under the null hypothesis and the alternative. The unicity of these distributions within each hypothesis enables to deduce the most powerful test from a direct application of Neyman-Pearson's lemma.

We are interested in testing the null hypothesis

$\mathcal{H}_0$: "The X_i's are uniformly distributed on an interval $[0, \theta]$ with $\theta > 0$"

against the alternative

$\mathcal{H}_1$: "The X_i's are exponentially distributed."

In this particular situation, one can exploit the fact that both the uniform and exponential models are closed by positive scaling: if the sample X belongs to the uniform (resp. exponential) model, then so does tX for all $t > 0$. We consider the class of scale-free tests, that is binary valued functions $\Phi = \Phi(X)$ of the sample X satisfying

$$\Phi(X) = \Phi(tX) \ , \ \forall t > 0.$$

If we let $\Phi(X) = \mathbb{1}\{X \in \mathcal{R}\}$ for $\mathcal{R}$ a Borel set of $\mathbb{R}^n$, saying that Φ is scale-free simply means that $\mathcal{R}$ is a cone. Equivalently, a statistic is scale-free if it can be expressed as a function of the normalized sample $X/\|X\|$, for any norm $\|.\|$ on $\mathbb{R}^n$.

Because the uniform and exponential models are stable by positive scaling, the distribution of a scale-free statistic does not depend on the parameter, whether we are under the null hypothesis (uniform) or the alternative (exponential). Therefore, we know there exists a most powerful scale-free test for these hypotheses, which we can derive from the likelihood ratio statistic of the normalized sample. As we show below, the resulting test is function of the statistic

$$\tau_n := \frac{\overline{X}_n}{X_{(n)}}$$

where $X_{(n)} = \max\{X_1, \dots, X_n\}$ and $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, whose properties are often discussed in relations with the tail distribution, see [21].

Theorem 2.1. *The test $\Phi = \mathbb{1}\{\tau_n \le c_\alpha\}$ with c_α such that $\mathbb{P}(\tau_n \le c_\alpha | \mathcal{H}_0) = \alpha$ is the most powerful scale-free test of level $\alpha \in (0, 1)$ to test $\mathcal{H}_0$ against $\mathcal{H}_1$.*

Proof. Let $\Phi = \Phi(X)$ be a scale-free test of level $\alpha \in (0, 1)$, we can write almost surely (for $X_n \neq 0$)

$$\Phi(X_1, \dots, X_n) = \Phi\Big(\frac{X_1}{X_n}, \dots, \frac{X_{n-1}}{X_n}, 1\Big) := \Psi\Big(\frac{X_1}{X_n}, \dots, \frac{X_{n-1}}{X_n}\Big),$$

for $\Psi : \mathbb{R}^{n-1} \to \{0, 1\}$. Let $Y = (Y_1, \dots, Y_{n-1}) = (X_1/X_n, \dots, X_{n-1}/X_n)$. Since the distribution of Y is constant over $\mathcal{H}_0$ and over $\mathcal{H}_1$, Neyman-Pearson's Lemma tells us that the likelihood ratio test has maximal power among tests of significance level α. Both likelihood functions $\mathcal{L}_0$ and $\mathcal{L}_1$ of Y under $\mathcal{H}_0$ and $\mathcal{H}_1$ respectively can be calculated explicitly, yielding

$$\mathcal{L}_0(Y) = \frac{1}{n(\max\{1, Y_{(n-1)}\})^n} \quad \text{and} \quad \mathcal{L}_1(Y) = \frac{(n-1)!}{(1 + \sum_{i=1}^{n-1} Y_i)^n}.$$

Recall that $Y_i = X_i/X_n$, Neyman-Pearson's likelihood ratio is thus given by

$$\frac{1}{n!}\frac{(1 + \sum_{i=1}^{n-1} Y_i)^n}{(\max\{1, Y_{(n-1)}\})^n} = \frac{(n\tau_n)^n}{n!}.$$

Since it is an increasing function of τ_n, the likelihood ratio test writes as $\mathbb{1}\{\tau_n \leq c_\alpha\}$ for some suitable c_α. □

In this framework, the most powerful scale-free test actually recovers Cox's maximum likelihood ratio test, introduced in [6] in a more general context. Indeed, considering the likelihood functions $L_0(\theta, X) = \theta^{-n}\mathbb{1}\{X_i \leq \theta, i = 1, \ldots, n\}$ and $L_1(\lambda, X) = \lambda^n \exp(-\lambda \sum_{i=1}^n X_i)$, the maximum likelihood ratio statistic is given by

$$\frac{\sup_{\lambda>0} L_1(\lambda, X)}{\sup_{\theta>0} L_0(\theta, X)} = \left(\frac{X_{(n)}}{e\,\overline{X}_n}\right)^n = \left(\frac{1}{e\,\tau_n}\right)^n.$$

Thus, Cox's maximum likelihood ratio test rejects the null hypothesis $\mathcal{H}_0$ for sufficiently small values of τ_n. The threshold c_α corresponding to a test of significance level α can be computed easily by Monte-Carlo. Nevertheless, to get an analytical expression of the threshold and the corresponding power of the test, one needs to derive the true distribution of the statistic τ_n under the null hypothesis and under the alternative. This issue is discussed in the Appendix.

3 Tail classification in Generalized Pareto distribution

If there is no particular reason to favor the uniform model over the exponential one, the significance level can be chosen so that the probabilities of error under $\mathcal{H}_0$ and $\mathcal{H}_1$ are equal, thus inducing a minimal probability of error in the worst case scenario. In this purpose, one must choose the threshold c_n as the unique solution of

$$\mathbb{P}(\tau_n \geq c_n | \mathcal{H}_0) = \mathbb{P}(\tau_n \leq c_n | \mathcal{H}_1) := A(n).$$

Thus, when using the threshold c_n, the probability $A(n)$ of selecting the correct model between uniform and exponential no longer depends on the true distribution. The values of the threshold c_n and percentage of accurate selections $A(n)$ are computed in Table 1 for different sample sizes n. It appears that the test quickly reaches a near perfect accuracy as the sample size increases. For a sample of size $n = 5$, the method selects the correct model more than 71% of the time, while a sample of size $n = 20$ achieves a precision above 95%.

n	5	10	20	50	100
$A(n)$	0.7178	0.8437	0.9514	0.9984	1
c_n	0.531	0.460	0.420	0.392	0.380

Table 1: Optimal threshold c_n and percentage of accuracy $A(n)$ for sample sizes $n = 5, 10, 20, 50, 100$, computed by Monte-Carlo simulations with 5.10^4 replications.

Of course, the level of accuracy is only exact if the model is well-specified, which is rarely the case in practice. Nevertheless, the test procedure may provide substantial information on the tail of the distribution. To study the behavior of extreme events, one focuses generally on the tail observations, i.e. the data over a certain threshold μ, so as to obtain independent realizations X_i conditionally to $X_i > \mu$. Under regularity conditions, the tail observations (translated to the left by a factor μ so as to have their support starting at zero) converge in distribution as $\mu \to \infty$ towards independent realizations

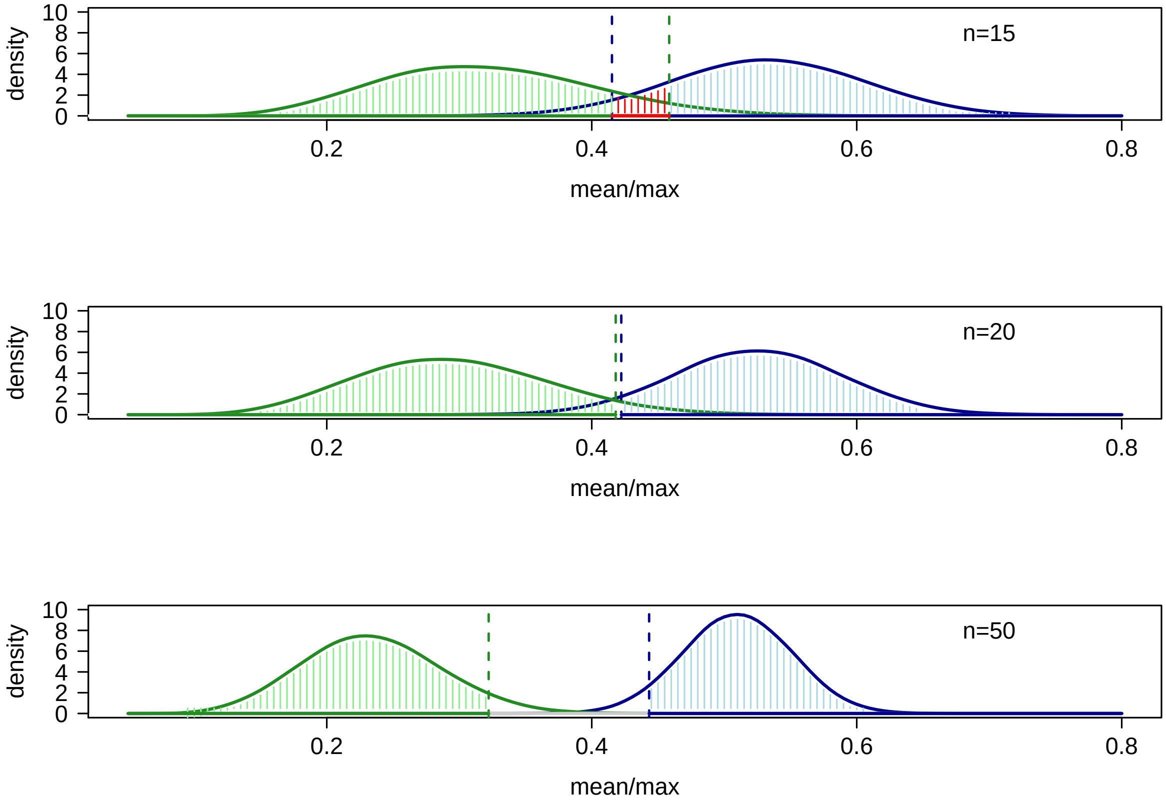

Figure 1: Monte-Carlo estimated densities of the mean/max statistic τ_n under $\mathcal{H}_0$ (blue) and $\mathcal{H}_1$ (green) for sample sizes $15, 20$ and 50 from top to bottom. The 5% quantile under $\mathcal{H}_0$ and 95% quantile under $\mathcal{H}_1$ are marked in vertical dotted lines, pointing out that a sample size of $n = 20$ is needed for a 95% precision.

of a Generalized Pareto Distribution (GPD), with density

$$f_{k,\sigma}(x) = \frac{1}{\sigma}\Big(1 - k\frac{x}{\sigma}\Big)^{\frac{1-k}{k}} \ , \ \begin{cases} x \geq 0 & \text{if } k \leq 0 \\ 0 \leq x \leq \frac{\sigma}{k} & \text{if } k > 0. \end{cases}$$

The scale $\sigma > 0$ and the shape $k \in \mathbb{R}$ (also called the tail index) are the two parameters used to describe the tail. The Generalized Pareto density can be extended by continuity to the values $k = 0$ and $k = 1$ recovering respectively the exponential and uniform distributions. Generally, the relevant information on the tail comes from the scale parameter k and less importance is given to the scale σ.

When working with tail observations, it is thus customary to assume that the data are drawn from a GPD. In this model, the behavior of the mean/max statistic τ_n is highly dictated by the tail index k while its distribution does not depend on the scale parameter σ (τ_n is scale-free). The mean/max statistic is thus perfectly suited to investigate the tail index of the GPD without having to concern about the scale. As a matter of fact, a crucial matter in Generalized Pareto models is to determine to what submodel the tail belongs, i.e. if the tail index k is greater than 1 (model A), between 0 and 1 (model B) or negative (model C). Because the frontiers between the different models are achieved for the exponential ($k = 0$) and uniform ($k = 1$) cases, the test procedure to distinguish between uniform and exponential distributions can be naturally extended for classification purposes.

The asymptotic distribution of τ_n (see Theorem A.3 in the Appendix) reveals that τ_n vanishes at a rate of n^{-k} when $k \in (0, 1)$, converges to $k/(k + 1)$ when $k > 0$ and is of the order $1/\log(n)$ in-between, for $k = 0$. In particular, the median of τ_n can be deduced from Theorem A.3,

$$\text{med}(\tau_n) = \begin{cases} -\dfrac{k}{k+1}\Big(\dfrac{\log(2)}{n}\Big)^{-k} + o(n^k) & \text{for } -1 < k < 0 \\[2ex] \dfrac{1}{\log(n)} + \dfrac{\log\log(2)}{\log^2(n)} + o\Big(\dfrac{1}{\log^2(n)}\Big) & \text{for } k = 0 \\[2ex] \dfrac{k}{k+1} + o\Big(\dfrac{1}{\sqrt{n}}\Big) & \text{for } k > 0. \end{cases} \tag{1}$$

The monotonic behavior of τ_n with respect to the tail index k makes it a usefool tool to learn to which submodel the distribution of the data belongs. The idea is simple: the practitioner chooses two reals numbers $0 < a_n < b_n < 1$ and concludes to the model A if $\tau_n \geq b_n$, the model B if $a_n < \tau_n < b_n$ and the model C if $\tau_n \leq a_n$. As suitable values of a_n and b_n, we use the theoretical median of τ_n under the exponential and uniform distributions respectively. By taking these values, the probability of selecting each model in the transitional cases $k = 0$ and $k = 1$ does not exceed $1/2$ so that none of the model A, B or C is favored. The actual value of $\text{med}(\tau_n)$ in the uniform case follows directly from Lemma A.1. Because we could not obtain an analytical expression of $\text{med}(\tau_n)$ in the exponential model, we use the asymptotic approximation given in Equation (1), which is in fact extremely accurate, even for small sample sizes. Thus, the bounds a_n and b_n used for the classification are

$$a_n = \frac{1}{\log(n)} + \frac{\log\log(2)}{\log^2(n)} \quad \text{and} \quad b_n = \frac{n+1}{2n}. \tag{2}$$

Although it is quite straight-forward to implement, the proposed procedure performs well in terms of classification compared to other estimation methods such as Zhang and Stephens (ZSE) [22], Song and Song (SSE) [18] or maximum likelihood (MLE). A comparative study is shown in Table 3.

	Model A			Model B						Model C		
	$k=-0.1$			$k=0.1$			$k=0.9$			$k=1.1$		
	A	B	C	A	**B**	C	A	**B**	C	A	B	**C**
ZSE												
$n=15$	**65.8**	34.1	0.1	44.1	**55.5**	0.4	0.9	**76.5**	22.6	0.4	57.7	**41.9**
$n=100$	**84.2**	15.8	0.0	19.9	**80.1**	0.0	0.0	**90.7**	9.3	0.0	31.0	**69.0**
SSE												
$n=15$	**85.3**	14.5	0.3	73.8	**25.6**	0.7	20.7	**50.0**	29.4	14.7	40.7	**44.6**
$n=100$	**70.1**	30.0	0.0	29.0	**70.9**	0.0	0.0	**77.1**	22.9	0.0	27.3	**72.7**
MLE												
$n=15$	**41.4**	50.2	8.4	18.8	**64.6**	16.6	0.1	**11.2**	88.8	0.0	4.1	**95.9**
$n=100$	**75.0**	25.0	0.0	9.7	**90.3**	0.0	0.0	**49.8**	50.2	0.0	4.4	**95.6**
$n=15$	**59.7**	39.9	0.4	38.6	**60.3**	1.1	0.2	**59.6**	40.2	0.0	40.7	**59.3**
$n=100$	**74.4**	25.6	0.0	20.4	**79.6**	0.0	0.0	**79.8**	20.2	0.0	22.3	**77.7**

Table 2: Percentages of the model classifications obtained by Zhang and Stephens Estimation (ZSE), Song and Song Estimation (SSE) and Maximum Likelihood Estimation (MLE) compared to the proposed methodology (bottom lines). In bold are the percentages of correct classification.

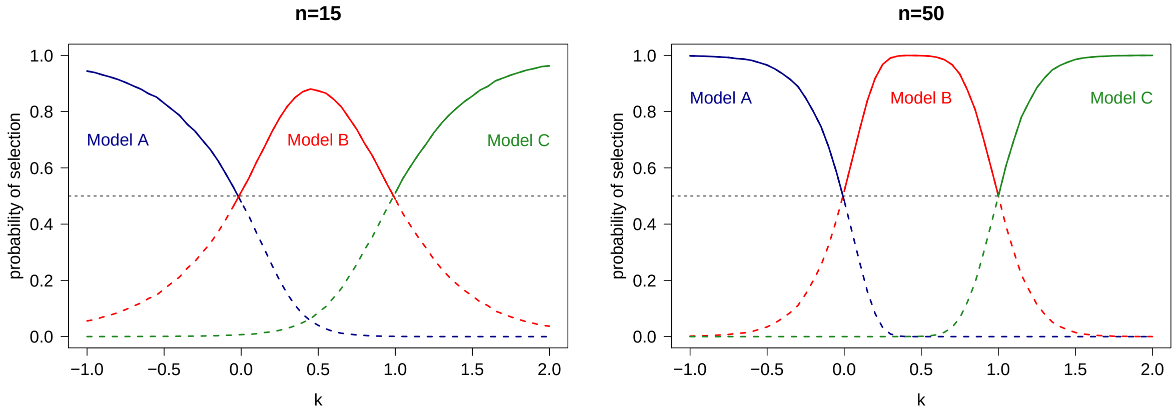

Figure 2: Monte-Carlo estimated probabilities of selecting the models A,B and C in function of the shape parameter k of the GPD, for sample sizes of $n = 15$ (left) and $n = 50$ (right). The straight line gives the probability of selecting the correct model.

Our classification method manages to not favor any model in particular, as the probability of correct selection depends essentially on the distance between the scale parameter k and the closest transitional phase. This is well illustrated in Figure 2 where the probability of accurate selection appears to be nearly symmetric locally around 0 and 1. In comparison, classification by MLE has a clear tendency to overestimate the scale parameter while SSE tends to underestimate it. For a small sample size, these biases are particularly clear for $k = 0.9$ where model C is selected 88.8% of the time and for $k = 0.1$

where model A is selected 73.8% of the time by SSE. In fact, for the small sample case $n = 15$, our method is almost objectively best compared to the three other methods. For the examples considered, the situation $n = 100, k \approx 0$ is the only scenario where our method is outperformed, ZSE being clearly best. This strongly suggests that the information on the tail of the distribution is not contained entirely in the mean/max statistic in this case. Nevertheless, the results of the method are overall quite satisfying given the small computational cost.

4 Applications

4.1 Fit a Generalized Pareto distribution

Due to the missclassification issue in Generalized Pareto models, inference methods tend to specialize in a specific region of the set of parameters, with none being uniformly best [7]. To solve the problem, we propose a simple recipe that involves a two-step procedure. We assume that the data $X = (X_1, \dots, X_n)$ are independent and identically distributed from a GPD and that no prior information is available on the parameters.

Step 1. Compute the mean/max statistic $\tau_n = \overline{X}_n / X_{(n)}$ and compare it to the values a_n and b_n obtained in Eq. (2). Select:

- Model A if $\tau_n < a_n$
- Model B if $a_n < \tau_n \leq b_n$
- Model C if $\tau_n \geq b_n$

Step 2. Consider the estimation method depending on the selected model and sample size by this recommendation based on the analysis in [7].

sample size	Model A	Model B	Model C
$n \leq 30$	SSE	ZSE	MLE
$n > 30$	ZSE	ZSE	MLE

In all situations, the method must be applied with the corresponding restriction on the set of parameters. Furthermore, when the sample is classified as Model B or C, a previous estimation of the endpoint can be recommended. Non-parametric approach shows satisfactory results, see [11].

This recipe is meant for data independently drawn from a Generalized Pareto distribution. For general problems related to tail detection and calibration, this assumes implicitly that the observations have been already extracted from the tail values of a larger dataset and suitably shifted so as to fit the GPD corresponding to the tail distribution. Nevertheless, the mean/max statistic can be used directly for tail detection. Suppose one is interested in the tail of the distribution of a sample $Y_1, ..., Y_N$, the idea is to calculate the mean/max statistic over the shifted tail observations $X_i := Y_{(N-n+i)} - Y_{(N-n)}, i = 1, \dots, n$ for various values of n. This way, one obtains a trajectory

$$\tau_n = \frac{\overline{X}_n}{X_{(n)}} = \frac{\sum_{i=1}^{n} (Y_{(N-n+i)} - Y_{(N-n)})}{n(Y_{(N)} - Y_{(N-n)})}, n = 1, 2, ...$$

that can be compared to the corresponding values of the bounds a_n, b_n. Ideally, only the values in the tail of the distribution are to be used for the classification so that n need not be too large. The optimal threshold for tail detection may appear more or less clearly (see Figure 4.1).

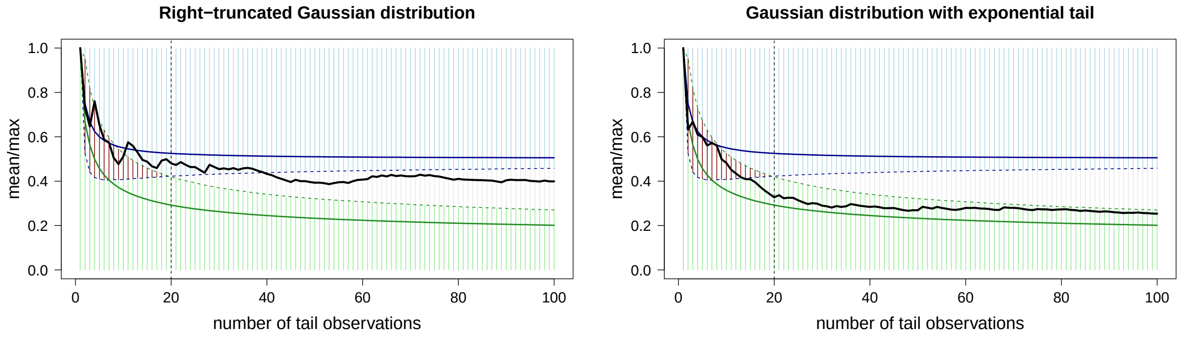

Figure 3: Examples of trajectories of the mean/max statistic τ_n from a sample of size 100 drawn from a standard Gaussian variable with modified tail. On the left, the distribution is truncated at the value 1.5 (left), on the right, it is extended to an exponential tail starting from 1.5. The bounds a_n, b_n from Eq. (2) are represented in straight lines along with their respective 95% confidence regions. The red-striped area corresponds to potential values of τ_n with more than 5% uncertainty to distinguish between uniform and exponential tails.

An overview of the mean/max trajectory can be helpful in order to select the correct model for the tail. In the examples displayed in Figure 4.1, the criterion is most conclusive around $n = 20$ even though a sample a this size may contain observations that do not obey the tail distribution. As a possible rule of thumb, the model for the tail can be decided based on the last value of τ_n that leaves the red-striped area, thus aiming for a precision of over 95%.

4.2 Distribution of extreme seismic events

The Gutenberg-Richter (GR) law states that the distribution of seismic moment corresponds a to power-law distribution [19, 5] with a density at a value M proportional to $1/M^{1+\beta}$ for a β evaluated to be approximately equal to 0.65. However, it is suspected that the collected data on seismic moments show a deviation of GR law for large values of M, entailing that the power-law model might have to be extended in order to add an exponential decay above a certain threshold [5]. In [14], Kagan enumerates the requirements that an extension of the GR law should fulfill. He also argues however that available seismic catalogs do not allow the reliable estimation of the threshold, except in the global case where the faster exponential decay may take place at the highest observed values of M, for which the available data are very poor [23].

The existence of a theoretical maximal magnitude earthquake entails that the true distribution of seismic moments must be right-truncated, thus displaying a uniform tail, see [7]. Nevertheless, the question remains as to know if the largest seismic events on record are not best modeled by an exponential decay. In order to infer on the underlying physical model, we compute the max/mean plot of the seismic moment collected data from the centroid moment tensor (CMT) catalog [8]. The analysis is restricted to shallow events (depth < 70 km) from 1976 to 2013, as recommended in [14]. The results are displayed in Figure 4 below.

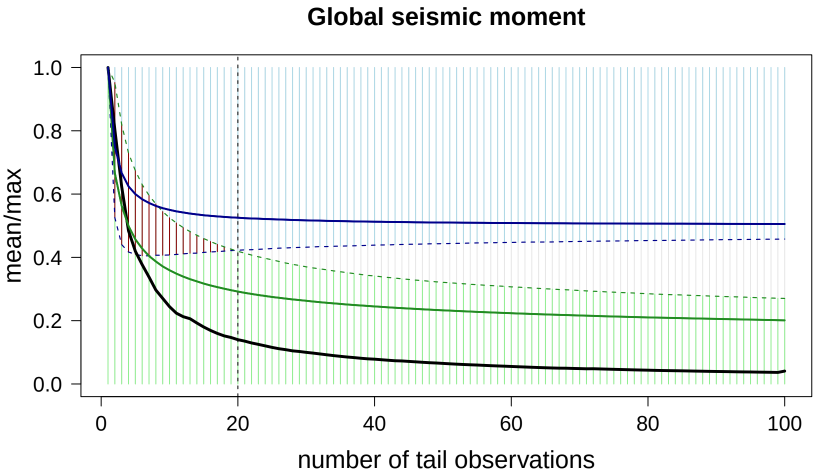

Figure 4: Trajectory of the mean/max statistic computed over the n largest (shifted) global seismic events, for $n = 1$ to $n = 100$.

This analysis shows no reason to suspect a change in the tail of the distribution, and confirms that the tail must be classified as Model A with $k < 0$. Consequently, the existence of a cut-off close to the maximal observed seismic moments has to be discarded.

4.3 Detection of experimental limitations

We apply this methodology in the analysis of acoustic emission data in failure under compression experiments of a nanoporous quartz sample [1, 17]. A cylindrical sample is placed between to plates that approach each other at a constant velocity. Compression is done with no lateral confinement and the experiment finishes when the sample has experienced a big failure and has literally disappeared. In this compressive process, two transductors are embedded in both plates in order to detect acoustic emission activity. These signals are pre-amplified in order to measure properly different magnitudes: amplitude, duration, energy, etc. The discrete measure in dB of the signal amplitude is defined as $A = 20\log_{10}(A_v/1\mu V)$, where A_v is the highest voltage value of each acoustic emission signal and $1\mu V$ is a reference voltage. Remark that the exponential distribution for the tail, the largest values, is the most natural model from a physical point of view. Regarding this physical magnitude, one must take into account that the signal pre-amplification can lead to a saturation for the largest acoustic emission events. Therefore, this saturation must be understood as a fictitious cut-off which is inherent to the experimental set-up. Actually, this experimental limitations are present in all measuring devices since all of them have a certain measurement range. The mean/max statistic τ_n enables to detect this experimental artifact in most samples, as we see Figure 5 where V.Navas and E.Vives (private communication) experiments are shown.

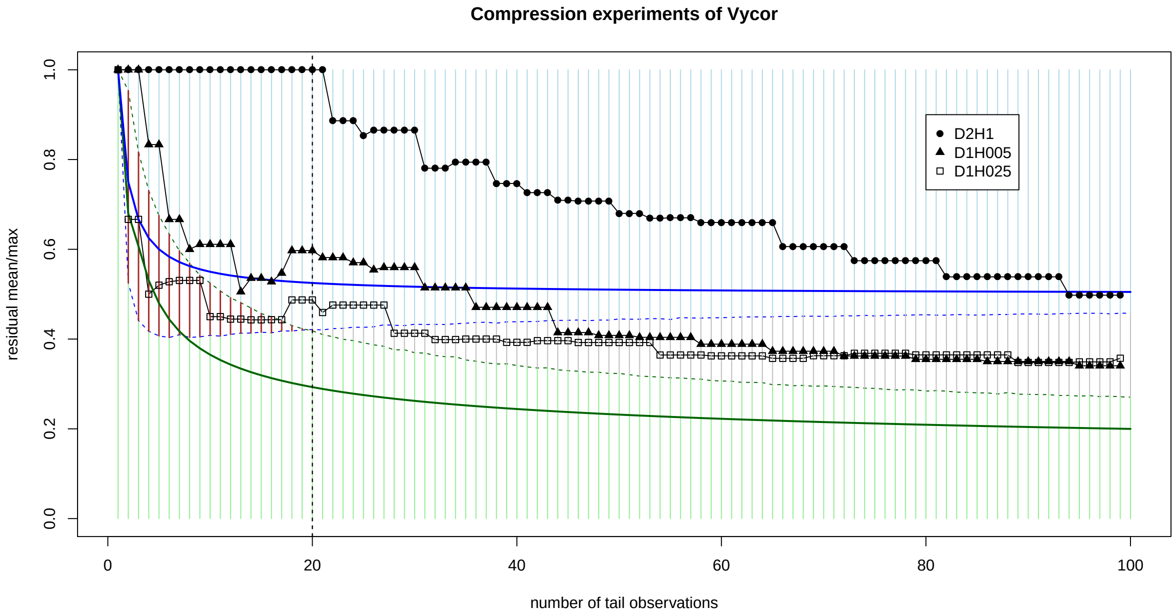

Figure 5: Trajectories of the mean/max statistic for the 100 largest values obtained from three compression experiments of nanoporous quartz samples.

According to the experimental intuition, one expects that acoustic emission signals do not reach saturation if the contact area between the plates is small. The mean/max methodology provides a simple and automatic way to identify if the data extracted from any experimental measurement are saturated or not. Figure 5 shows the trajectories of the mean/max statistic for the 100 largest values obtained from three compression experiments of nanoporous quartz samples. The circle trajectories correspond to the greatest material, 2mm diameter. This mean/max plot shows saturated experiment, since for the 20 largest emission signals the exponential decay is rejected. In fact, the tail of the dataset is classified as a compact support, Model C. In this case, a descriptive analysis reveals a clear suspicion of saturation experiment, since the largest value is repeated too many times. However, for small materials, 1mm diameter, it is difficult to detect if the saturation occurs. The mean/max test (for the 20 largest) reveals that the triangle trajectories was affected by the saturation, as in the previous case, but the trajectory of squares is classified as Model B, therefore a non-saturation can not be rejected, since Model B includes the exponential distribution for the tail.

Acknowledgement

We are grateful to V. Navas and E. Vives for their feedback. Research expenses were founded by projects MTM2012-31118 from Spanish MINECO, 2014SGR-1307 from AGAUR, and the Collaborative Mathematics Project from La Caixa Foundation.

A Distribution of the mean/max statistic

We investigate the distribution of the mean/max statistic τ_n when the sample is drawn from a GPD. The two transitional phases corresponding to uniform and exponential distributions are given a particular interest as they are directly linked to the level and power of the test discussed in Section 2.

A.1 The uniform case

The threshold c_α corresponding to a test of significance level $\alpha \in (0,1)$ in Theorem 2.1 follows from calculating the distribution of τ_n under the null hypothesis of a uniformly distributed sample $X_1, \dots, X_n$. Noticing that

$$\tau_n = \frac{1}{n}\Big(1 + \sum_{i=1}^{n-1} \frac{X_{(i)}}{X_{(n)}}\Big),$$

it appears that $n(\tau_n - 1)$ is distributed as the sum of $n-1$ independent standard uniform random variables. This distribution is known as the Irwin-Hall distribution of parameter $n-1$, whose density is given by

$$I_{n-1}(t) = \frac{n}{2}\sum_{k=0}^{n-1} \operatorname{sgn}(t-k)\frac{(-1)^k (t-k)^{n-2}}{k!(n-k)!}, \; t \in [0, n-1],$$

where $\operatorname{sgn}(y)$ is the sign of y, equal to 1 if y is positive, -1 if y is negative and 0 if y is zero. Geometrically, $I_{n-1}(t)$ represents the volume of the intersection of the ℓ^1-sphere of radius t in $\mathbb{R}^{n-1}$ with the hypercube $[0,1]^{n-1}$. For sake of completeness, we derive the distribution of τ_n under the null hypothesis in the next lemma. The density of a random variable Z will be denoted by f_Z.

Lemma A.1. *If $X = (X_1, \dots, X_n)$ is an iid sample from a uniform distribution on $[0, \theta]$, with $\theta > 0$, then τ_n has density*

$$f_{\tau_n}(t) = nI_{n-1}(nt-1) \ , \ t \in \Big[\frac{1}{n}, 1\Big].$$

In particular, $\mathbb{E}(\tau_n) = \operatorname{med}(\tau_n) = \dfrac{n+1}{2n}$.

Proof. First remark that the distribution of τ_n does not depend on θ so that we can assume that $\theta = 1$ without loss of generality. Let $W_1 = X_{(1)}$, $W_i = X_{(i)} - X_{(i-1)}$ for $i = 2, \dots, n$ and $W_{n+1} = 1 - X_{(n)}$. We use that $W = (W_1, \dots, W_{n+1})$ has a standard Dirichlet distribution $\mathrm{Dir}(n+1)$, which means that

$$W \overset{d}{=} \left(\frac{Y_1}{\sum_{i=1}^{n+1} Y_i}, \dots, \frac{Y_{n+1}}{\sum_{i=1}^{n+1} Y_i}\right)$$

where $Y_1, \dots, Y_{n+1}$ is an iid sample with exponential distribution. Thus, the vector

$$W' = \left(\frac{X_{(1)}}{X_{(n)}}, \dots, \frac{X_{(n-1)} - X_{(n-2)}}{X_{(n)}}\right) \overset{d}{=} \left(\frac{Y_1}{\sum_{i=1}^{n} Y_i}, \dots, \frac{Y_n}{\sum_{i=1}^{n} Y_i}\right)$$

has $\mathrm{Dir}(n)$ distribution and $Z_i = X_{(i)}/X_{(n)} \overset{d}{=} \sum_{j=1}^{i} Y_j / \sum_{j=1}^{n} Y_j$, $i = 1, \dots, n-1$ has the same distribution as an ordered iid sample of $n-1$ uniform random variables on $[0,1]$. We deduce that $\sum_{i=1}^{n-1} Z_i = n(\tau_n - 1)$ has Hirwin-Hall distribution with parameter $n-1$ and the result follows. □

A.2 The exponential case

Calculating the power of the test requires to know the distribution of τ_n under $\mathcal{H}_1$. Here again, this distribution can be computed explicitly.

Lemma A.2. *If $X = (X_1, \dots, X_n)$ is an iid sample from an exponential distribution of parameter $\lambda > 0$, then τ_n has density*

$$f_{\tau_n}(t) = \frac{n!}{n^{n-1}} \, \frac{I_{n-1}(nt-1)}{t^n}, \; t \in \Big[\frac{1}{n}, 1\Big].$$

Proof. The distribution of τ_n does not depend on λ so that we can assume that $\lambda = 1$ without loss of generality. We know that the ordered sample has density on $\mathbb{R}^n$ given by

$$(X_{(1)}, \dots, X_{(n)}) \sim n! \; \exp\big(-\textstyle\sum_{i=1}^n x_i\big) \mathbb{1}\{0 \le x_1 \le \dots \le x_n\}.$$

Let $Y_i = X_{(1)}/X_{(n)}$ for $i = 1, \dots, n-1$. By the change of variable $y_i = x_i/x_n$, we get

$$(Y_1, \dots, Y_{n-1}, X_{(n)}) \sim n! \; x_n^{n-1} \exp\big(-x_n(1 + \textstyle\sum_{i=1}^{n-1} y_i)\big) \mathbb{1}\{0 \le y_1 \le \dots \le y_{n-1} \le 1, x_n > 0\}.$$

Integrating the density over x_n, we find

$$Y = (Y_1, \dots, Y_{n-1}) \sim \frac{n!(n-1)!}{(1 + \sum_{i=1}^{n-1} y_i)^n} \; \mathbb{1}\{0 \le y_1 \le \dots \le y_{n-1} \le 1\}$$

To compute the density of $S_n := n\tau_n - 1 = \sum_{i=1}^{n-1} Y_i$, we now need to integrate the joint density over the level sets of the ℓ^1-norm $D_s = \{y \in \mathbb{R}^{n-1} : \sum_{i=1}^{n-1} y_i = s\}$. We obtain

$$\begin{aligned} f_{S_n}(s) &= \int_{D_s} \frac{n!(n-1)!}{(1 + \sum_{i=1}^{n-1} y_i)^n} \; \mathbb{1}\{0 \le y_1 \le \dots \le y_{n-1} \le 1\} \; dy_1 \dots dy_{n-1} \\ &= \frac{n!}{(1+s)^n} \int_{D_s} (n-1)! \; \mathbb{1}\{0 \le y_1 \le \dots \le y_{n-1} \le 1\} \; dy_1 \dots dy_{n-1}. \end{aligned}$$

Remark that $(n-1)! \; \mathbb{1}\{0 \le y_1 \le \dots \le y_{n-1} \le 1\}$ is the density of an ordered sample $(U_{(1)}, \dots, U_{(n-1)})$ of independent uniform random variables on $[0, 1]$. Thus,

$$\int_{D_s} (n-1)! \; \mathbb{1}\{0 \le y_1 \le \dots \le y_{n-1} \le 1\} \; dy_1 \dots dy_{n-1} = I_{n-1}(s)$$

to which we deduce that $S_n = n\tau_n - 1$ has density $n! I_{n-1}(s)/(1+s)^n$ over $[0, n-1]$. The result follows by a simple change of variable. □

A.3 Asymptotic distribution of τ_n in the GPD model

Finally, we investigate the distribution of τ_n when the observations are drawn from generalized Pareto distributions with shape parameter k. Because τ_n is scale-free, the following results do not depend on the scale parameter σ of the GPD, which can be taken equal to one without loss of generality. The actual distribution of τ_n being difficult to compute as a function of k, we only discuss the asymptotic distribution.

Theorem A.3. *Let $X = (X_1, \dots, X_n)$ be an iid sample drawn from a Generalized Pareto distribution with scale parameter k. Then,*

- *If $k > 0$,*
$$\tau_n \sim \frac{k}{k+1} + \frac{N_k}{\sqrt{n}} + o_P\Big(\frac{1}{\sqrt{n}}\Big),$$
where N_k has normal distribution $\mathcal{N}\Big(0, \dfrac{k^2}{(1+k)^2(1+2k)}\Big)$.

- *If $k = 0$ (exponential case),*
$$\tau_n \sim \frac{1}{\log(n)} - \frac{G}{\log^2(n)} + o_P\Big(\frac{1}{\log^2(n)}\Big),$$
where G has standard Gumbel distribution.

- *If $-1 < k < 0$,*
$$\tau_n \sim \frac{W_k}{n^{-k}} + o_P(n^k)$$
where W_k has Weibull distribution with shape parameter $-1/k$ and scale parameter $-k/(k+1)$.

Proof. We tackle each case separately. For $k > 0$, we have by the central limit theorem,

$$\sqrt{n}\Big(\overline{X}_n - \frac{1}{1+k}\Big) \xrightarrow[n\to\infty]{d} \mathcal{N}\Big(0, \frac{1}{(1+k)^2(1+2k)}\Big)$$

while $X_{(n)}$ converges a.s. towards $1/k$, the upper bound of the support of the GPD. The result follows easily by Slutsky's lemma. In the exponential case $k = 0$, $\overline{X}_n$ converges a.s. towards 1 while for $t > -\log(n)$,

$$\mathbb{P}\Big(X_{(n)} - \log(n) \leq t\Big) = \Big(1 - e^{-(\log(n)+t)}\Big)^n = \Big(1 - \frac{1}{n^{1+\frac{t}{\log(n)}}}\Big)^n \xrightarrow[n\to\infty]{} e^{-e^{-t}}.$$

Thus $X_{(n)} - \log(n)$ converges towards a standard Gumbel distribution. The result follows in view of

$$\frac{1}{X_{(n)}} = \frac{1}{\log(n) - (X_{(n)} - \log(n))} = \frac{1}{\log(n)} + \frac{X_{(n)} - \log(n)}{\log^2(n)} + o_P\Big(\frac{1}{\log^2(n)}\Big).$$

For the final case $-1 < k < 0$, we have by the strong law of large numbers

$$\overline{X}_n = \frac{1}{1+k} + o_P(1)$$

For $t > n^k$, write

$$\mathbb{P}\Big(n^k(1 - kX_{(n)}) \leq t\Big) = \mathbb{P}\Big(X_{(n)} \leq \frac{1}{k}\big(1 - \frac{t}{n^k}\big)\Big) = \Big(1 - \frac{t^{1/k}}{n}\Big)^n \xrightarrow[n\to\infty]{} e^{-t^{1/k}}.$$

We deduce that $n^k(1-kX_{(n)}) \sim -kn^kX_{(n)}$ converges towards a Fréchet distribution with shape parameter $-1/k$ as $n \to \infty$. Thus, $-n^{-k}/kX_{(n)}$ converges to the inverse of a Fréchet variable whose distribution is Weibull. The result follows by Slutsky's lemma. □

References

[1] Jordi Baró, Álvaro Corral, Xavier Illa, Antoni Planes, Ekhard K. H. Salje, Wilfried Schranz, Daniel E. Soto-Parra, and Eduard Vives. Statistical similarity between the compression of a porous material and earthquakes. *Phys. Rev. Lett.*, 110:088702, Feb 2013.

[2] Jan Beirlant, Yuri Goegebeur, Johan Segers, and Jozef Teugels. *Statistics of extremes: theory and applications.* John Wiley & Sons, 2006.

[3] Enrique Castillo and Ali S Hadi. Fitting the generalized pareto distribution to data. *Journal of the American Statistical Association*, 92(440):1609–1620, 1997.

[4] Stuart Coles, Joanna Bawa, Lesley Trenner, and Pat Dorazio. *An introduction to statistical modeling of extreme values*, volume 208. Springer, 2001.

[5] A Corral. Scaling and universality in the dynamics of seismic occurrence and beyond. *Acoustic Emission and Critical Phenomena*, pages 225–244, 2008.

[6] David R Cox. Tests of separate families of hypotheses. 1:105–123, 1961.

[7] Joan del Castillo and Isabel Serra. Likelihood inference for generalized pareto distribution. *Computational Statistics & Data Analysis*, 83:116–128, 2015.

[8] G Ekström, M Nettles, and AM Dziewoński. The global cmt project 2004–2010: centroid-moment tensors for 13,017 earthquakes. *Physics of the Earth and Planetary Interiors*, 200:1–9, 2012.

[9] Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling extremal events: for insurance and finance*, volume 33. Springer, 1997.

[10] Barbel Finkenstadt and Holger Rootzén. *Extreme values in finance, telecommunications, and the environment.* CRC Press, 2003.

[11] Isabel Fraga Alves, Cláudia Neves, and Pedro Rosário. A general estimator for the right endpoint with an application to supercentenarian women's records. *Extremes*, pages 1–39, 2016.

[12] Wallace E Franck. The most powerful invariant test of normal versus cauchy with applications to stable alternatives. *Journal of the American Statistical Association*, 76(376):1002–1005, 1981.

[13] Jonathan RM Hosking and James R Wallis. Parameter and quantile estimation for the generalized pareto distribution. *Technometrics*, 29(3):339–349, 1987.

[14] Yan Y Kagan. Seismic moment distribution revisited: I. statistical results. *Geophysical Journal International*, 148(3):520–541, 2002.

[15] Natalia Markovich. *Nonparametric analysis of univariate heavy-tailed data : research and practice.* Wiley series in probability and statistics. John Wiley & Sons, Chichester, England, 2007.

[16] Alexander J McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative risk management: concepts, techniques, and tools.* Princeton university press, 2010.

[17] Victor Navas-Portella, Álvaro Corral, and Eduard Vives. Avalanches in displacement-driven compression of porous glasses. *unpublished*, 2016.

[18] Jongwoo Song and Seongjoo Song. A quantile estimation for massive data with generalized pareto distribution. *Computational Statistics & Data Analysis*, 56(1):143–150, 2012.

[19] Didier Sornette, Leon Knopoff, YY Kagan, and Christian Vanneste. Rank-ordering statistics of extreme events: Application to the distribution of large earthquakes. *Journal of Geophysical Research: Solid Earth*, 101(B6):13883–13893, 1996.

[20] Vincent A Uthoff. An optimum test property of two weil-known statistics. *Journal of the American Statistical Association*, 65(332):1597–1600, 1970.

[21] José A Villaseñor-Alva and Elizabeth González-Estrada. A bootstrap goodness of fit test for the generalized pareto distribution. *Computational Statistics & Data Analysis*, 53(11):3835–3841, 2009.

[22] Jin Zhang and Michael A Stephens. A new and efficient estimation method for the generalized pareto distribution. *Technometrics*, 51(3):316–325, 2009.

[23] Gert Zöller. Convergence of the frequency-magnitude distribution of global earthquakes: Maybe in 200 years. *Geophysical Research Letters*, 40(15):3873–3877, 2013.