On Tail Decay Rate Estimation of Loss Function Distributions

Etrit Haxholli, Marco Lorenzi

On Tail Decay Rate Estimation of Loss Function Distributions

Etrit Haxholli etrit.haxholli@inria.fr
Inria
Univesity Côte d'Azur
2004 Rte des Lucioles, 06902 Valbonne, France

Marco Lorenzi marco.lorenzi@inria.fr
Inria
Univesity Côte d'Azur
2004 Rte des Lucioles, 06902 Valbonne, France

Editor:

Abstract

The study of loss function distributions is critical to characterize a model's behaviour on a given machine learning problem. For example, while the quality of a model is commonly determined by the average loss assessed on a testing set, this quantity does not reflect the existence of the true mean of the loss distribution. Indeed, the finiteness of the statistical moments of the loss distribution is related to the thickness of its tails, which are generally unknown.

Since typical cross-validation schemes determine a family of testing loss distributions conditioned on the training samples, the total loss distribution must be recovered by marginalizing over the space of training sets. As we show in this work, the finiteness of the sampling procedure negatively affects the reliability and efficiency of classical tail estimation methods from the Extreme Value Theory, such as the Peaks-Over-Threshold approach. In this work we tackle this issue by developing a novel general theory for estimating the tails of marginal distributions, when there exists a large variability between locations of the individual conditional distributions underlying the marginal. To this end, we demonstrate that under some regularity conditions, the shape parameter of the marginal distribution is the maximum tail shape parameter of the family of conditional distributions. We term this estimation approach as *cross-tail estimation (CTE)*.

We test cross-tail estimation in a series of experiments on simulated and real data[1], showing the improved robustness and quality of tail estimation as compared to classical approaches, and providing evidence for the relationship between model performance and loss distribution tail thickness.

Keywords: Extreme Value Theory, Tail Modelling, Loss Function Distributions, Peaks-Over-Threshold, Cross-Tail-Estimation, Model Ranking

1. Introduction

Distributions of loss functions are central objects of study, as they represent the performance of machine learning models. For a given model and machine learning problem, the

1. The code is available at https://github.com/ehaxholli/CTE

true distribution of the loss function is generally not known, and we usually only posses a finite sample set generated from different choices of training and testing sets. In order to make comparisons between the performance of different models in terms of the underlying loss function distributions, different methods have been developed. Classical approaches are based on information criteria such as the Akaike Information Criterion (AIC) Akaike (1973, 1974), which is an asymptotic approximation of the KL divergence of the true distribution of the data and the fitting candidate, and its corrected version (AICc) Sugiura (1978); Hurvich and Tsai (1989), as well as the Bayesian Information Criterion (BIC) Schwarz (1978). The use of information criteria, particularly AIC, is often limited by the several underlying approximations and assumptions Burnham and Anderson (2007), making them impractical in some cases. Other strategies, known as splitting/resampling methods, have been developed in parallel, where a part of the data is used to test the performance of the trained model. The family of such methods is broad, and is based on a variety of partitioning and evaluation strategies adopted to mitigate data heterogeneity and unbalancedness. Neyman (1934); Cochran (2007).

In the case of cross validation methods, the quantity of interest that is usually estimated for model assessment is the mean of the distribution of the loss function. However, even though this estimation will always be a finite value, in reality, the first moment is not guaranteed to exist, and its existence as well as the existence of higher moments cannot be established through cross validation alone. From a theoretical standpoint, the highest existing moment of a distribution is strongly linked to the thickness of its tails, which highlights the importance of studying the behaviour and decay rate of tails of loss function distributions.

In order to proceed, we first must be able to model the tails of distributions and to quantify their "thickness". Extreme Value Theory (EVT) is an established field concerned with modelling the tails of distributions. One of the fundamental results in EVT is the Pickands–Balkema–De Haan Theorem, which states that the tails of a large class of distributions can be approximated with generalized Pareto ones Pickands (1975); de Haan and Ferreira (2007). In practice, the shape and scale parameter of the generalized Pareto are approximated from a finite sample, while its location parameter is always zero. It is the shape parameter which quantifies tail thickness, with larger values corresponding to heavier tails. The resulting estimation method is called Peaks-Over-Threshold (POT).

In the context of distributions of loss functions, for each training set, there is a corresponding conditional loss function distribution over points in the sample space. The actual total loss function distribution, the entity of our interest, is the weighted sum (integral) of all such conditional distributions, that is, it is the distribution created after marginalizing across the space of training datasets. In practice, we have a finite number of conditional distributions, as we have a finite number of training sets. Furthermore, for each of these conditional distributions, we only possess an approximation of them, derived from the samples in the testing set. The empirical approximation of the total loss function distribution therefore consists of the union of the sample sets of conditional distributions. Within this setting, the estimation of the tail shape of the total loss function distribution could be ideally carried out by applying POT on this union of samples.

In theory, as we show in this work, the role of the fattest conditional tails in determining the decay rate of the marginal is preserved, since the marginal and conditional distributions

are defined everywhere, which allows the assessment of tails at extreme locations. Unfortunately, in practice, the finiteness of the sampling affects the estimation of the tail of the marginal distribution, as the tails may be poorly or not even represented across different conditional distributions. To be more specific, during marginalization, samples from the tails of heavy tailed distributions can be overshadowed by the samples from the non-tail part of individual thin tailed ones. This suggests that modelling the tails of a marginal distribution by the usual application of POT can give inaccurate results in practice.

In this paper, we develop a general method to mitigate the issue of estimating the tails of marginal distributions, when there exists a large variability between locations of the individual conditional distributions underlying the marginal. To this end, we demonstrate that under some regularity conditions, the shape parameter of the marginal distribution is precisely the maximum tail shape parameter of the family of conditional distributions. We refer to the method constructed from this result as *cross tail estimation*, due to similarities that it shares with Monte Carlo cross validation. Furthermore, we show evidence of polynomial decay of tails of distributions of model predictions, and empirically demonstrate a relationship between the thickness of such tails and model performance. An additional benefit of using the approach proposed here instead of the standard POT, is the reduced computational time in the case that the marginal is estimated from many conditional distributions.

The following is a summary of the structure of the paper: In Section 2 we recall some of the main concepts and results from Extreme Value Theory. In Section 3, we state and generalize the main problem, which we tackle in Section 4, by building our theory. We conclude Section 4, by proving three statements which are useful for the experimental part, and by highlighting the relation between the tail of a distribution and its moments. In the final section, we show experimentally that our method can improve estimation in practice, as compared to the standard use of POT.

2. Related Work and Background

In the first part of this section we give a short review of Monte Carlo cross validation, as the method we develop, namely cross tail estimation, shares many similarities with it. In the second subsection, standard results and definitions from extreme value analysis are stated, as they are at the basis of the proofs in Section 4.

2.1 Monte Carlo Cross Validation

Let $D = \{(x_1, y_1), ...(x_n, y_n)\}$, be a set of data samples drawn form the same distribution. During each iteration i we sample k samples $D_i = \{(x_{\pi(1)}, y_{\pi(1)}), ..., (x_{\pi(k)}, y_{\pi(k)})\}$ without repetition from the original dataset D, and consider it as the training set for that iteration. The set $D \setminus D_i$ is then used as the testing set. The quantity of interest, during iteration i is the sample mean of the loss of the model trained on D_i, namely $\hat{f}_{D_i}$, over the points of the testing set:

$$\tilde{M}_i^L := \frac{1}{|D \setminus D_i|} \sum_{j \in D \setminus D_i} L(\hat{f}_{D_i}(x_j), y_j), \tag{1}$$

for a given loss function L.
We evaluate the total performance of the model, based on its average performance over different choices of the training/testing sets, that is, the true evaluation metric is:

$$\tilde{M}^L := \frac{1}{m} \sum_{i \in [m]} \tilde{M}_i^L = \frac{1}{m} \sum_{i \in [m]} \frac{1}{|D \setminus D_i|} \sum_{j \in D \setminus D_i} L(\hat{f}_{D_i}(x_j), y_j), \tag{2}$$

where m is the number of iterations.
We discuss cross validation in more detail in Subsection 3.4, where we show the similarities between the method we develop for tail estimation of marginalized distribution, i.e. cross tail estimation, and Monte Carlo cross validation.

2.2 Extreme Value Theory

Extreme value theory (EVT) or extreme value analysis (EVA) is a branch of statistics dealing with the extreme deviations from the median of probability distributions. Extreme value theory is closely related to failure analysis and dates back to 1923, when Richard von Mises discovered that the Gumbell distribution is the limiting distribution of the maximum of an iid sequence, sampled from a Gaussian distribution. In 1928, Ronald A. Fisher and Leonard H. C. Tippett in Fisher and Tippett (1928), characterized the only three possible non-degenerate limiting distributions of the maximum in the general case: Frechet, Gumbel and Weibull. In 1943, Boris V. Gnedenko, gave a rigorous proof of this fact in Gnedenko (1943). This result is known Fisher–Tippett–Gnedenko theorem, and forms the foundation of EVT. The three aforementioned limiting distributions of the maximum can be written in compact form and they are known as the class of extreme value distributions:

Definition 1 *The Generalized Extreme Value Distribution is defined as follows:*

$$G_{\xi,a,b}(x) = e^{-(1+\xi(ax+b))^{-\frac{1}{\xi}}}, \quad 1 + \xi(ax+b) > 0, \tag{3}$$

where $b \in \mathbb{R}$, $\xi \in \mathbb{R} \setminus \{0\}$ and $a > 0$. For $\xi = 0$, we define the generalized Extreme Value Distribution as the limit when $\xi \to 0$, that is

$$G_{0,a,b}(x) = e^{-e^{-ax-b}}. \tag{4}$$

Theorem 2 (Fisher–Tippett–Gnedenko) *: Let X be a real random variable with distribution F_X. Denote by $\{X_1, X_2, ..., X_n\}$ a set of iid samples from the distribution F_X, and define $M_n = max\{X_1, ..., X_n\}$. If there exist two sequences $\{c_i > 0\}_{i \in \mathbb{N}}$ and $\{d_i \in \mathbb{R}\}_{i \in \mathbb{N}}$, such that*

$$c_n^{-1}(M_n - d_n) \xrightarrow{d} F \text{ as } n \to \infty, \tag{5}$$

for some non-degenerate distribution F, then we must have $F(x) = G_{\xi,a,b}(x)$.

If X is a random variable as in Theorem 2, such that $F(x) = G_{\xi,a,b}(x)$, we say that F_X is in the Maximum Domain of Attraction of $G_{\xi,a,b}(x)$, and we write $F_X \in MDA(\xi)$. Depending on whether $\xi > 0$, $\xi = 0$, $\xi < 0$, we say that F_X is in the MDA of a Frechet, Gumbell, or

Weibull distribution respectively.
In 1974, J. Pickands in Pickands (1975), and Balkema, A. & De Haan, Laurens in Balkema and de Haan (1974) proved that the limiting distribution of samples larger than a threshold is a Generalized Pareto distribution, whose location parameter is zero.

Definition 3 *A Generalized Pareto distribution with location parameter zero is defined as below:*

$$G_{\xi,\sigma}(w) = \begin{cases} 1 - (1 + \xi\frac{w}{\sigma}))^{-\frac{1}{\xi}} & \text{for } \xi \neq 0 \\ 1 - e^{-\frac{w}{\sigma}} & \text{for } \xi = 0 \end{cases}, \tag{6}$$

where $w > 0$ when $\xi > 0$ and $0 < w < -\frac{\sigma}{\xi}$ for $\xi < 0$. The shape parameter is denoted by ξ, while the scale parameter by σ.

Theorem 4 (Pickands–Balkema–De Haan) *: Let X be a random variable with distribution F_X and $x_F \leq \infty$ such that $\forall x > x_F,\ \bar{F}_X(x) = 0$. Then $F_X \in MDA(\xi) \iff \exists g : (0,\infty) \to (0,\infty)$ such that*

$$\lim_{u \to x_F} \sup_{y \in [0, x_F - u]} |\bar{F}_u^X(y) - \bar{G}_{\xi, g(u)}(y)| = 0, \tag{7}$$

where $\bar{F}_u^X(y) = \frac{1 - F_X(y+u)}{1 - F_X(u)}$.

This result forms the basis of the well-known Peak-Over-Threshold (POT) method which is used in practice to model the tails of distributions. The shape parameter can be estimated via the Pickands Estimator:

Definition 5 *Let X_1, X_2,...,X_n be iid samples from the distribution F_X. If we denote with $X_{1,n}$, $X_{2,n}$,..., $X_{n,n}$ the samples sorted in descending order, then the Pickands estimator is defined as follows:*

$$\hat{\xi}_{k,n}^{(P)} = \frac{1}{\ln 2} \ln \frac{X_{k,n} - X_{2k,n}}{X_{2k,n} - X_{4k,n}}. \tag{8}$$

Another important result which we are going to use frequently in our proofs is Theorem 9, which can be found in Embrechts et al. (2013), and gives the connection between the maximum domain of attraction and slowly varying functions.

Definition 6 *A positive measurable function L is called slowly varying if it is defined in some neighborhood of infinity and if:*

$$\lim_{x \to \infty} \frac{L(ax)}{L(x)} = 1, \text{for all } a > 0. \tag{9}$$

Theorem 7 (Representation Theorem, see Galambos and Seneta (1973)) *: A positive measurable function L on $[x_0, \infty]$ is slowly varying if and only if it can be written in the form:*

$$L(x) = e^{c(x)} e^{\int_{x_0}^{x} \frac{u(t)}{t} dt}, \tag{10}$$

where $c(t)$ and $u(t)$, are measurable bounded functions such that $\lim_{x\to\infty} c(t) = c_0 \in (0,\infty)$ and $u(t) \to 0$ as $t \to \infty$.

Proposition 8 *Mikosch et al. (1999) If L is slowly varying then for every $\epsilon > 0$:*

$$\lim_{x\to\infty} x^{-\epsilon} L(x) = 0. \tag{11}$$

Proof We give a proof in the Appendix for the sake of completeness. ■

Theorem 9 *: If $X \in MDA(\xi)$ and x_F is such that $\forall x > x_F,\ \bar{F}_X(x) = 0$ then:*

- $\xi > 0 \iff \bar{F}_X(x) = x^{-\frac{1}{\xi}} L(x)$, *where L is slowly varying,*
- $\xi < 0 \iff \bar{F}_X(x_F - \frac{1}{x}) = x^{\frac{1}{\xi}} L(x)$, *where L is slowly varying,*
- $\xi = 0 \iff \bar{F}_X(x) = c(x)e^{-\int_w^x \frac{g(t)}{a(t)}dt} (= c(x)M(x)),\ w < x < x_F \leq \infty$, *where c and g are measurable functions satisfying $c(x) \to c > 0, g(x) \to 1$ as $x \uparrow x_F$, and $a(x)$ is a positive, absolutely continuous function (with respect to Lebesgue measure) with density $a'(x)$ having $\lim_{x\uparrow x_F} a'(x) = 0$.*

3. Setup and Problem Statement

In the first part of this section, we formalize the problem of tail modelling of total loss distributions as explained in the introduction. In the second part we expand this formulation to the case of general marginal distributions, and state the method we intend to develop in order to mitigate the weaknesses of POT in this setting, which were described in the introduction. We conclude this section by showing the similarity between our method and cross-validation.

3.1 Problem Statement

Let $\boldsymbol{U} = (\boldsymbol{X}, \boldsymbol{Y})$ and $\boldsymbol{V} = (\boldsymbol{X_T}, \boldsymbol{Y_T})$, be random variables where $\boldsymbol{X}$ and $\boldsymbol{Y}$ stand respectively for a sample of features and labels in the test set, while $\boldsymbol{X_T}$, $\boldsymbol{Y_T}$ stand respectively for samples of features and labels in the training set. A model which tries to approximate the ground truth is denoted as $\hat{\boldsymbol{f}}_{\boldsymbol{V}}(\boldsymbol{X})$, while the prediction error on the testing data $\boldsymbol{U}$ conditioned on the training set $\boldsymbol{V}$ is denoted as $W_{\boldsymbol{V}}(\boldsymbol{U})$. We assume for simplicity that $W_{\boldsymbol{V}}(\boldsymbol{U}) > 0$ and notice that the probability density function of $W_{\boldsymbol{V}}(\boldsymbol{U})$ is

$$f_W(w) = \int f_{W,\boldsymbol{V}}(w, \boldsymbol{v})d\boldsymbol{v} = \int f_{\boldsymbol{V}}(\boldsymbol{v}) f(w|\boldsymbol{V} = \boldsymbol{v})d\boldsymbol{v} = \int f_{\boldsymbol{V}}(\boldsymbol{v}) h_{\boldsymbol{v}}(w)d\boldsymbol{v}, \tag{12}$$

therefore the distribution function of $W_{\boldsymbol{V}}(\boldsymbol{U})$ is:

$$F_W(w) = \int f_{\boldsymbol{V}}(\boldsymbol{v}) H_{\boldsymbol{v}}(w)d\boldsymbol{v}. \tag{13}$$

We would like to estimate the shape of the tails of $F_W(w)$, by estimating the shape of the tails of the distributions $H_{\boldsymbol{v}}(w)$ conditioned on the training set $\boldsymbol{v}$. It is important to notice

that we can estimate the shape of the tails of $W_{\boldsymbol{V}}(\boldsymbol{U})$ by also conditioning on the testing set:

$$f_W(w) = \int f_{W,\boldsymbol{U}}(w, \boldsymbol{u})d\boldsymbol{u} = \int f_{\boldsymbol{U}}(\boldsymbol{u})f(w|\boldsymbol{U} = \boldsymbol{u})d\boldsymbol{u} = \int f_{\boldsymbol{U}}(\boldsymbol{u})h_{\boldsymbol{u}}(w)d\boldsymbol{u} \tag{14}$$

$$F_W(w) = \int f_{\boldsymbol{U}}(\boldsymbol{u})H_{\boldsymbol{u}}(w)d\boldsymbol{u}. \tag{15}$$

3.2 The General Problem

Generalizing the problem stated in Section 3.1 requires to consider a one dimensional random variable of interest Z_i, dependent on other random variables $\{Z_1, Z_2, ..., Z_{i-1}, Z_{i+1}, ..., Z_n\}$, such that the probability density function of Z_i is

$$f_{Z_i}(z_i) = \int f(z_1, ..., z_{i-1}, z_i, z_{i+1}, ..., z_n)d_{z_1} \cdots d_{z_{i-1}}d_{z_{i+1}} \cdots d_{z_n} \tag{16}$$

$$= \int f(z_1, ..., z_{i-1}, z_{i+1}, ..., z_n)f(z_i|z_1, ..., z_{i-1}, z_{i+1}, ..., z_n)d_{z_1} \cdots d_{z_{i-1}}d_{z_{i+1}} \cdots d_{z_n} \tag{17}$$

$$= \int f(\boldsymbol{z_{-i}})f(z_i|\boldsymbol{z_{-i}})d\boldsymbol{z_{-i}} = \int f(\boldsymbol{z_{-i}})h_{\boldsymbol{z_{-i}}}(z_i)d\boldsymbol{z_{-i}}. \tag{18}$$

Integrating with respect to z_i we get

$$F_{Z_i}(z_i) = \int f(\boldsymbol{z_{-i}})F(z_i|\boldsymbol{z_{-i}})d\boldsymbol{z_{-i}} = \int f(\boldsymbol{z_{-i}})H_{\boldsymbol{z_{-i}}}(z_i)d\boldsymbol{z_{-i}}. \tag{19}$$

In this case, with regards to the previous section, we notice that $\boldsymbol{Z_{-i}} = \boldsymbol{V}$ is the training set on which we condition, while $Z_i = W$ is the random variable of interest. Ideally, we would like to find a relation between the shape of the tails of conditional $F(z_i|\boldsymbol{z_{-i}})$ and that of $F_{Z_i}(z_i)$. In Section 4, we show that under some regularity conditions, this relation exists, and the shape parameter of the tail of $F_{Z_i}(z_i)$, if positive, is the same as the largest shape parameter of the tails of the distributions $F(z_i|\boldsymbol{z_{-i}})$. Otherwise, if the shape parameter of the tail of $F_{Z_i}(z_i)$ is non-positive ($F_{Z_i}(z_i)$ has thin tails), then the largest shape parameter of the tails of distributions $F(z_i|\boldsymbol{z_{-i}})$ is non-positive (each $F(z_i|\boldsymbol{z_{-i}})$ is thin tailed), and vice-versa.

3.3 The need for cross tail estimation

Estimating the tails of marginal distributions via standard methods such as using POT directly, can give unsatisfactory results. In order to get a glimpse of this issue, let's assume that our variable of interest is $Z_1 > 0$, which in turns depends on the variable Z_2. For simplicity we can assume that Z_2 can be either 0 or 1, with equal probability, and if $Z_2 = 0$ then $f(z_1|Z_2 = 0)$ is a thick tailed distribution whose even first moment does not exist, while if $Z_2 = 1$ then $f(z_1|Z_2 = 1)$ is a Gaussian distribution, with a large

mean. Suppose we proceed with the standard POT approach, that is, we integrate out the random variable Z_2, and subsequently estimate the shape parameter of the tail of $f(z_1)$. In practice, when the number of samples is limited, it is possible that none of the samples of Z_1 from the fat tailed distributions exceeds those of the Gaussian due to the difference between their locations. Therefore, the sample tail of the marginal (mixture) distribution $f(z_1) = \frac{1}{2}(f(z_1|Z_2 = 0) + f(z_1|Z_2 = 1))$ is defined by the sample tail of the Gaussian $f(z_1|Z_2 = 1)$, while in reality, as we will show, the tail of $f(z_1)$ is defined by the fat tail of $f(z_1|Z_2 = 0)$. Of course in the ideal case where the sampling process is not finite, we would recover the true tail shape; however, for practical applications, the theory developed in this paper is necessary as it enables the incorporation of information provided by correlated variables to improve estimation.

3.4 Analogies between cross tail estimation and cross validation

During cross validation, for a given iteration, a training set $\boldsymbol{V}$ and a testing set $\boldsymbol{U}$ are selected. The following conditional expectation is then estimated:

$$\mathbb{E}[W_{\boldsymbol{V}}(\boldsymbol{U})|\boldsymbol{V} = \boldsymbol{v}] = \int w h_{\boldsymbol{v}}(w)dw. \tag{20}$$

The estimates of $\mathbb{E}[W_{\boldsymbol{V}}(\boldsymbol{U})|\boldsymbol{V}]$ received in each iteration are then averaged to get an estimation of the total expectation:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{U},\boldsymbol{V}}(W_{\boldsymbol{V}}(\boldsymbol{U})) = \int w f(w)dw = \int f_{\boldsymbol{V}}(\boldsymbol{v}) \int w h_{\boldsymbol{v}}(w)dwd\boldsymbol{v} = \\ = \int f_{\boldsymbol{V}}(\boldsymbol{v})\mathbb{E}[W_{\boldsymbol{V}}(U)|\boldsymbol{V} = \boldsymbol{v}]d\boldsymbol{v} = \mathbb{E}[\mathbb{E}[W_{\boldsymbol{V}}(\boldsymbol{U})|\boldsymbol{V} = \boldsymbol{v}]].\end{aligned} \tag{21}$$

In the language of Section 3.1, the mean of distribution $f_W(w)$ is the average of the means of the conditional distributions $h_{\boldsymbol{v}}(w) = f(w|\boldsymbol{V} = \boldsymbol{v})$.
This statement about sums stands parallel with our claim about extremes; that the shape parameter of the tail of $f_W(w)$, if positive, is the maximum of the shape parameters of the tails of the conditional distributions $h_{\boldsymbol{v}}(w) = f(w|\boldsymbol{V} = \boldsymbol{v})$.

4. Theoretical Results

In this section, we build our theory of modelling the tails of marginal distributions, which culminates with Theorem 13 and Theorem 20. Regarding Equation (19), Theorem 20 shows that the shape parameter of the tail of $F_{Z_i}(z_i)$, if positive, is the same as the largest shape parameter of the tails of distributions $F(z_i|\boldsymbol{z_{-i}})$. Otherwise, if the shape parameter of the tail of $F_{Z_i}(z_i)$ is non-positive, that is $F_{Z_i}(z_i)$ has thin tails, then the largest shape parameter of the tails of distributions $F(z_i|\boldsymbol{z_{-i}})$ is non-positive, that is each $F(z_i|\boldsymbol{z_{-i}})$, is thin tailed. We conclude this section by proving three statements which are useful in the experimental Section 5, and give the relation between the existence of the moments of a distribution and the thickness of its tails. Unless stated otherwise, the proofs of all the statements are given in Appendix A.

4.1 Tails of marginal distributions

For two given distributions, whose tails have positive shape parameters, we expect the one with larger tail parameter to decay slower. Indeed:

Lemma 10 *If $F_1 \in MDA(\xi_1)$ and $F_2 \in MDA(\xi_2)$, and if $\xi_1 > \xi_2 > 0$, then $\lim_{x\to\infty} \frac{\bar{F}_2(x)}{\bar{F}_1(x)} = 0$.*

In a similar fashion, regardless of the signs of the shape parameters, we expect the one with larger tail parameter to decay slower. In fact we have the following:

Lemma 11 *If $F_1 \in MDA(\xi_1)$ and $F_2 \in MDA(\xi_2)$ then:*

1. *If $\xi_1 > 0$ and $\xi_2 = 0$ then $\lim_{x\to\infty} \frac{\bar{F}_2(x)}{\bar{F}_1(x)} = 0$.*
2. *If $\xi_1 = 0, x_{F_1} = \infty$ and $\xi_2 < 0$ then $\lim_{x\to\infty} \frac{\bar{F}_2(x)}{\bar{F}_1(x)} = 0$.*
3. *If $\xi_1 > 0$ and $\xi_2 < 0$ then $\lim_{x\to\infty} \frac{\bar{F}_2(x)}{\bar{F}_1(x)} = 0$.*
4. *If $\xi_1 < 0$ and $\xi_2 < 0$ then $\lim_{x\to\infty} \frac{F_2(x)}{F_1(x)} = 1$.*

Despite the fact that a linear combination of slowly varying functions is not necessarily slowly varying, the following statement holds true:

Lemma 12 *If for $i \in \{1, ..., n\}$ we let $L_i(x)$ be slowly varying functions, and $\{a_1, ..., a_n\}$ be a set of positive real numbers, then*

$$L(x) = \sum_{i=1}^{n} a_i L_i(x)$$

is slowly varying.

Before we continue, we simplify the notation of Equation (19) by setting $\boldsymbol{Z} = \boldsymbol{Z_{-i}}$ and $X = Z_i$. In this case, Equation (19) becomes:

$$F_{Z_i}(z_i) = \int f(\boldsymbol{z_{-i}}) H_{\boldsymbol{z_{-i}}}(z_i) d\boldsymbol{z_{-i}} = \int f(\boldsymbol{z}) H_{\boldsymbol{z}}(x) d\boldsymbol{z} =: H(x). \tag{22}$$

In the following Theorem, we show that the tail shape parameter of a mixture of a finite number of distributions is the same as the maximal tail shape parameter of the conditional distributions.

Theorem 13 *Let $Z : \Omega \to A \subset \mathbb{R}^n$ be a random vector where $|A| < \infty$. At each point $\boldsymbol{z}_1, .., \boldsymbol{z}_n \in A$, we define a distribution $H_{\boldsymbol{z}_i}(x) \in MDA(\xi_i)$ and assume that $\xi_{max} := max(\xi_1 = \xi_{z_1}, ..., \xi_n = \xi_{z_n}) > 0$. If the set $\{p_1, ..., p_n\}$ is a set of convex combination parameters, that is $\sum_i p_i = 1$ and $p_i > 0$ then:*

$$H(x) = \sum_{i}^{n} p_i H_{\boldsymbol{z}_i}(x) \in MDA(\xi_{max}). \tag{23}$$

If $\xi_{max} \leq 0$ then if ξ_H exists we have $\xi_H \leq 0$.

From now on, we assume that the functions $H_S(x) = \int_S f_{\boldsymbol{Z}}(\boldsymbol{z}) H_{\boldsymbol{z}}(x) d\boldsymbol{z}$ defined on any element S of the Borel $\sigma - algebra$ induced by the usual metric are in the MDA of some extreme value distribution.

Proposition 8 states that every slowly varying function is sub-polynomial. That is for any $\delta > 0$ and any slowly varying function $L(x)$, if we are given any $\gamma > 0$, then we can find $x(L, \delta, \gamma) > 0$, such that for all $x > x(L, \delta, \gamma)$, the inequality $x^{-\delta} L(x) < \gamma$ holds. However, since $x(L, \delta, \gamma)$ depends on the function L, assuming that we have a family of $\{L_z | z \in A\}$, where A is a measurable set, the set $\{x(L_z, \delta, \gamma) | z \in A\}$ can be unbounded, suggesting that the beginning of the tail of $\bar{F}_z(x) = x^{-\frac{1}{\xi(z)}} L_z(x)$ can be postponed indefinitely across the family $\{F_z | z \in A\}$. These concepts are formalized in the following:

Definition 14 *For a set A, the family of sub-polynomial functions $\{L_z(x) | z \in A\}$ is called γ-uniformly sub-polynomial if for any fixed $\delta > 0$, there exists a $\gamma(\delta)$ so that the set $\{x(L_z, \delta, \gamma) | z \in A\}$ is bounded from above, where $x(L_z, \delta, \gamma)$ are chosen such that when $x > x(L_z, \delta, \gamma)$ we have $x^{-\delta} L_z(x) < \gamma$.*

Proposition 15 *Let $\boldsymbol{Z} : \Omega \to A \subset \mathbb{R}^n$ be a random vector where A is measurable and define a family of slowly varying functions $\{L_z(x) | z \in A\}$, which we assume is γ-uniformly sub-polynomial. Then for a probability density function $f_Z(z)$ on A induced by Z, the function $L_z(x) = \int_A f_Z(z) L_z(x) dz$ is sub-polynomial.*

In the following theorem, we assume that all conditional distributions have positive tail shape parameters, and we show that the marginal distribution cannot have a tail shape parameter larger (smaller) than the largest (smallest) tail shape parameter across conditional distributions. Furthermore, if the tail shape parameters vary continuously across the space of conditional distributions, then the tail shape parameter of the marginal is precisely the same as the maximal tail shape parameter of the conditional distributions.

Theorem 16 *Let $\boldsymbol{Z} : \Omega \to A \subset \mathbb{R}^n$ be a random vector where A is measurable. At each point $\boldsymbol{z} \in A$ define a distribution $H_{\boldsymbol{z}}(x) \in MDA(\xi_{\boldsymbol{z}})$, and suppose there exist ξ_{lo}, ξ_{up} such that $\forall \boldsymbol{z} \in A$, $0 < \xi_{lo} \leq \xi_{\boldsymbol{z}} \leq \xi_{up}$. If the family $\{L_{\boldsymbol{z}}(x) | \boldsymbol{z} \in A\}$ is γ-uniformly sub-polynomial, then for $H(x) = \int_A f_{\boldsymbol{Z}}(\boldsymbol{z}) H_{\boldsymbol{z}}(x) d\boldsymbol{z}$ we have $\xi_{lo} \leq \xi_H \leq \xi_{up}$. Furthermore, if $\xi_{\boldsymbol{z}}$ is continuous in $\boldsymbol{z}$, then $\xi_H = \xi_{max}$, where $\xi_{max} := \sup\{\xi_{\boldsymbol{z}} | \boldsymbol{z} \in A\}$.*

Similarly to the case when $H_{\boldsymbol{z}}(x)$ are in the $MDA(\xi_{\boldsymbol{z}})$ for $\xi_{\boldsymbol{z}} > 0$, if we wish to extend the results above, regularity conditions are required for the $\xi_{\boldsymbol{z}} \leq 0$ case. Similarly as before, we notice that if $F_z(x) \in MDA(0)$, that is $F(x) = c(x) M(x)$, then the corresponding $c(x)$, as stated in Theorem 9, converges and therefore is bounded. Furthermore, the corresponding $M(x)$ is sub-polynomial as seen in the proof of Lemma 11. These observations motivate the following:

Definition 17 *For a set A, define the family of distribution functions $\mathcal{F}_{\mathcal{A}} = \{F_{\boldsymbol{z}}(x) | \boldsymbol{z} \in A\}$, and define $A^+ = \{\boldsymbol{z} | \xi_{\boldsymbol{z}} > 0\}$, $A^- = \{\boldsymbol{z} | \xi_{\boldsymbol{z}} < 0\}$, and $A^0 = \{\boldsymbol{z} | \xi_{\boldsymbol{z}} = 0\}$. We say family $\mathcal{F}_{\mathcal{A}}$ has stable cross-tail variability if,*

- *$\{L_{\boldsymbol{z}}(x) | \boldsymbol{z} \in A^+\}$ is γ-uniform sub-polynomial,*

- $\{M_{\boldsymbol{z}}(x)|\boldsymbol{z} \in A^0\}$ *is* γ*-uniform sub-polynomial, and* $c_{\boldsymbol{z}}(x)$ *are uniformly bounded,*
- $\bar{F}_{\boldsymbol{z}}(x)$ *are uniformly bounded, where* $\boldsymbol{z} \in A^-$.

We notice that in the previous theorem, if for all $\boldsymbol{z}$ we have $0 < \xi_{\boldsymbol{z}} \leq \epsilon$, then $\xi_H \leq \epsilon$. If the corresponding family $\mathcal{F}_{\mathcal{A}} = \{F_{\boldsymbol{z}}(x)|\boldsymbol{z} \in A\}$ has stable cross-tail variability, this holds independently from the lower bound of $\{\xi_{\boldsymbol{z}}|\boldsymbol{z} \in A\}$. Indeed:

Lemma 18 *Let* $\boldsymbol{Z} : \Omega \to A \subset \mathbb{R}^n$ *be a random vector where* A *is measurable. At each point* $\boldsymbol{z} \in A$ *define a distribution* $H_{\boldsymbol{z}}(x) \in MDA(\xi_{\boldsymbol{z}})$*, and suppose that* $\forall \boldsymbol{z} \in A,\ \xi_{\boldsymbol{z}} \leq \epsilon$*. If the family* $\{F_{\boldsymbol{z}}(x)|\boldsymbol{z} \in A\}$ *has stable cross-tail variability, then for* $H(x) = \int_A f_{\boldsymbol{Z}}(\boldsymbol{z}) H_{\boldsymbol{z}}(x) d\boldsymbol{z}$ *we have* $\xi_H \leq \epsilon$.

Corollary 19 *Let* $\boldsymbol{Z} : \Omega \to A \subset \mathbb{R}^n$ *be a random vector where* A *is measurable. At each point* $\boldsymbol{z} \in A$ *define a distribution* $H_{\boldsymbol{z}}(x) \in MDA(\xi_{\boldsymbol{z}})$*, and suppose that* $\forall \boldsymbol{z} \in A,\ \xi_{\boldsymbol{z}} \leq 0$*. If the family* $\{F_{\boldsymbol{z}}(x)|\boldsymbol{z} \in A\}$ *has stable cross-tail variability, then for* $H(x) = \int_A f_{\boldsymbol{Z}}(\boldsymbol{z}) H_{\boldsymbol{z}}(x) d\boldsymbol{z}$ *we have* $\xi_H \leq 0$.

Proof We notice that for any $\epsilon > 0$, we have $\xi_{\boldsymbol{z}} < \epsilon$ for all $\boldsymbol{z} \in A$. Hence, from the previous Lemma we conclude that $\xi_H \leq \epsilon, \forall \epsilon > 0$. ■

Finally, we prove the generalization of Theorem 16 in the case that the tail shape parameters ξ_z of the conditional distributions are real numbers:

Theorem 20 *Let* $\boldsymbol{Z} : \Omega \to A \subset \mathbb{R}^n$ *be a random vector where* A *is measurable. At each point* $\boldsymbol{z} \in A$ *define a distribution* $H_{\boldsymbol{z}}(x) \in MDA(\xi_{\boldsymbol{z}})$*, where* $\xi_{\boldsymbol{z}}$ *is continuous and* $\xi_{max} > 0$*. If the family* $\{F_{\boldsymbol{z}}(x)|\boldsymbol{z} \in A\}$ *has stable cross-tail variability, then for* $H(x) = \int_A f_{\boldsymbol{Z}}(\boldsymbol{z}) H_{\boldsymbol{z}}(x) d\boldsymbol{z}$ *we have* $\xi_H = \xi_{max}$*. In the case that* $\xi_{max} \leq 0$ *then* $\xi_H \leq 0$.

Examples when the conditions of Theorem 20 hold, as well as when they are violated, can be found in Appendix C and B, respectively.

4.2 Useful propositions for the experimental part

In this subsection, we prove three statements which are useful in the experimental Section 5, and state the well-known relation between the existence of the moments of a distribution and the thickness of its tails.

Proposition 21 *Let* H_X *be the distribution of the random variable* X*. We define* X_1 *to be a random variable whose distribution is the normalized right tail of* H_X*, that is:*

$$H_{X_1}(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \frac{H(x) - H(0)}{\mathbb{P}(X > 0)} & \text{for } x > 0 \end{cases}. \tag{24}$$

Similarly we define X_2 *whose distribution is the normalized left tail of* H_X*,*

$$H_{X_2}(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{H(0) - H(-x)}{\mathbb{P}(X \leq 0)} & \text{for } x \geq 0 \end{cases}. \tag{25}$$

If $H_{X_1} \in MDA(\xi_1)$, $H_{X_2} \in MDA(\xi_2)$, and $max\{\xi_1, \xi_2\} > 0$, then:

$$\xi_{|X|} = max\{\xi_1, \xi_2\}.$$

If $H_{X_1} \in MDA(\xi_1)$, $H_{X_2} \in MDA(\xi_2)$, and $max\{\xi_1, \xi_2\} \leq 0$, then:

$$\xi_{|X|} \leq 0.$$

Proof Since

$$\begin{aligned} H_{|X|}(x) = \mathbb{P}(|X| < x) = \mathbb{P}(X < x | X > 0)\mathbb{P}(X > 0) + \mathbb{P}(-X < x | X \leq 0)\mathbb{P}(X \leq 0) \\ = p_1 H_{X_1}(x) + p_2 H_{X_2}(x), \end{aligned} \tag{26}$$

Theorem 13 gives the desired conclusion. ■

Proposition 22 *Let X be a random variable such that $X \in MDA(\xi_X)$. If we define Y to be equal to X^α, for some $\alpha \in \mathbb{R}^+$, then $Y \in MDA(\xi_Y)$ where $\xi_Y = \alpha \xi_X$.*

If we condition our model on the test set

$$f_W(w) = \int f_{W,U}(w, u)du = \int f_U(u) f(w|U = u)du = \int f_U(u) h_u(w)du, \tag{27}$$

then we can estimate the shape parameters of the distribution of W, without the need for target data in the testing set:

Theorem 23 *If we define the loss function as $W_{\boldsymbol{V}}(\boldsymbol{U}) = |Y - \hat{f}_{\boldsymbol{V}}(\boldsymbol{X})|^p$, then under the assumptions of Theorem 20, the distribution of $W_{\boldsymbol{V}}(\boldsymbol{U})$ has the same shape parameter of the tail as the distribution of $|\hat{f}_{\boldsymbol{V}}(\boldsymbol{X})|^p$, where $p \in \mathbb{R}^+$.*

There exists a strong connection between the Maximum Domain of Attraction of a distribution, and the existence of its moments (see Embrechts et al. (2013)):

Proposition 24 *If $F_{|X|}$ is the distribution function of a random variable $|X|$, and $F_{|X|} \in MDA(\xi)$ then:*

$$\textit{i) if } \xi > 0, \textit{ then } \mathbb{E}[|X|^r] = \infty, \forall r \in (\frac{1}{\xi}, \infty), \tag{28}$$

$$\textit{ii) if } \xi \leq 0, \textit{ then } \mathbb{E}[|X|^r] < \infty, \forall r \in (0, \infty). \tag{29}$$

This means that, if a model induces a loss function whose distribution has a shape parameter that is bigger than one, then even the first moment of that loss function distribution does not exist. Hence, we would expect that our model has an infinite mean, which would suggest that this model should be eliminated during model ranking.

In Theorem 23, we showed that if we condition on the testing set, we can estimate the shape of the total loss distribution, that is the distribution of $W_{\boldsymbol{V}}(\boldsymbol{U})$, by simply investigating the models prediction, without the need for target data. This can also be motivated from the moments of $W_{\boldsymbol{V}}(\boldsymbol{U})$ as shown in Appendix F.

5. Experiments

In this section, we demonstrate the significance of Theorem 20. In the first subsection, we show experimental evidence that the shape parameter of the estimated marginal distribution, under the assumption that we have an abundance of sample points, coincides with the maximal shape parameter of individual conditional distributions. In the second subsection, we show that when the sample size is small, as it is the case in the real world, the method proposed by Theorem 20 (cross tail estimation) can be necessary for proper tail shape parameter estimation of marginal distributions. Furthermore, in the third subsection, we compare the standard POT and cross tail estimation on real data. For the considered regression scenarios, we notice that when these shape parameters are calculated by cross tail estimation, there is a relationship between the ranking of models proposed by the MSE in the test set, and the magnitude of shape parameters of the distribution of model predictions. We also notice there that such a relationship does not appear in the case that we use directly the POT method to estimate the aforementioned shape parameters. Finally, in the fourth subsection, we discuss the computational advantages of using cross tail estimation.

5.1 Validity of Cross Tail Estimation in Practice

The main problem that we tried to tackle in the previous section, was estimating the shape parameters of the tail of distribution $H(x)$:

$$H(x) = \int f(\boldsymbol{z}) H_{\boldsymbol{z}}(x) d\boldsymbol{z}, \tag{30}$$

via tail shape estimation of the conditional distributions $H_{\boldsymbol{z}}(x)$. In what follows, we give two experiments showing that this is feasible in practice.

5.1.1 Cross tail estimation in the uniform case

For simplicity, we set $\boldsymbol{z}$ to be one dimensional, and thus denote the conditional distributions $H_{\boldsymbol{z}}$ as H_j, where $j \in \mathbb{R}$. In this case equation (30) becomes

$$H(x) = \int f(j) H_j(x) dj. \tag{31}$$

First, for a given fixed $i \in \mathbb{N}$, we fix $\xi^i_{max} \in [-5, 5]$. Then we set $H_j(x) = 1 - x^{-\frac{1}{\xi_j}}$, which has tail shape parameter ξ_j, as $\bar{H}_j(x) = x^{-\frac{1}{\xi_j}}$. We sample m points from $H_j(x)$, by sampling from a uniform distribution first and transforming these samples by H_j^{-1}. We repeat this task for conditional distributions $H_j(x)$ whose corresponding shape parameters ξ_j have the following values $\{-5, -5 + \frac{\xi}{i-1}, ..., -5 + j\frac{\xi}{i-1}, ..., \xi^i_{max} = -5 + (i-1)\frac{\xi}{i-1}\}$, where $\xi = \xi^i_{max} + 5$. Thus, i is number of equidistant separations of $[-5, \xi^i_{max}]$, that is $i = |\{-5, -5 + \frac{\xi}{i-1}, ..., \xi^i_{max}\}|$. We assume a uniform distribution over the choice of j, that is, $f(j) = \frac{1}{|\{-5, -5+\frac{\xi}{i-1}, ..., \xi^i_{max}\}|} = \frac{1}{i}$. In practice this means that when we marginalize over j, we simply join all i arrays of samples together. This creates an array of $i * m$ data points, denoted by $S_{i,m}$, representing $i*m$ samples from the distribution $H(x) \in MDA(\xi_i)$. Using Pickands' estimator on $S_{i,m}$, we estimate the shape parameter of the tail of $H(x)$,

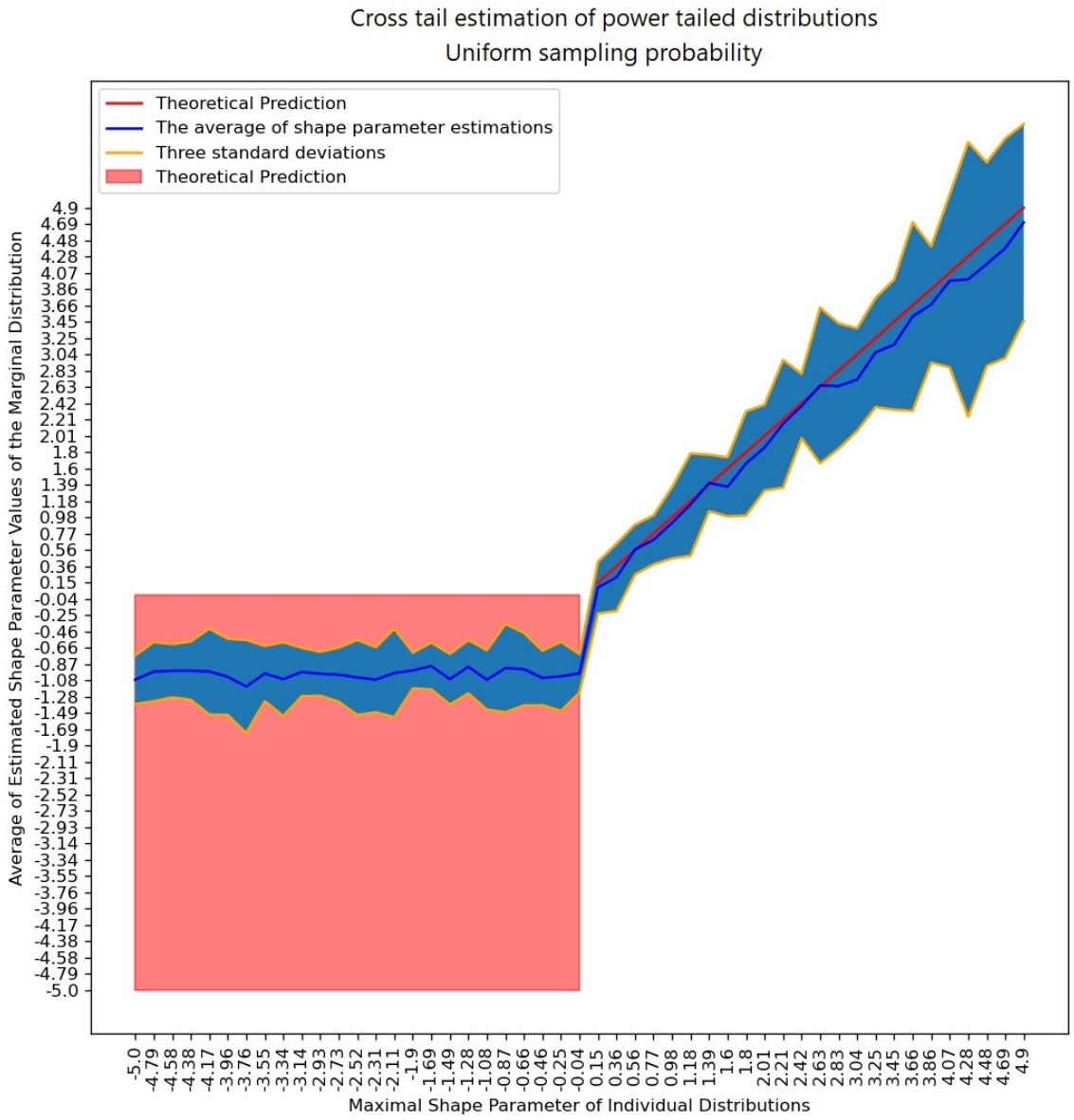

Figure 1: Equivalence of POT and CTE for tail shape estimation of the marginal distribution. In this case the conditional distributions have power tails, and the sampling probability between different conditional distributions is uniform.

and expect that this estimated parameter $\hat{\xi}^i_{POT}$ will be close to ξ_i. Based on our theoretical results for positive ξ^i_{max} we have $\xi^i_{max} = \xi_i$, therefore it is expected that $\xi^i_{max} = \hat{\xi}^i_{POT}$.

We repeat this process for 50 different shape parameters ξ^i_{max} whose values are equidistant in the interval $[-5, 5]$, and are increasing in i. The results are shown in Figure 1. We see that the values of $\{\hat{\xi}^i_{POT}|i\}$ (in the y-axis), are close to those in $\{\xi^i_{max}|i\}$ (in the x-axis) when $\xi^i_{max} > 0$, for $m = 5 * 10^6$ as predicted by Theorem 13 and Theorem 20. When

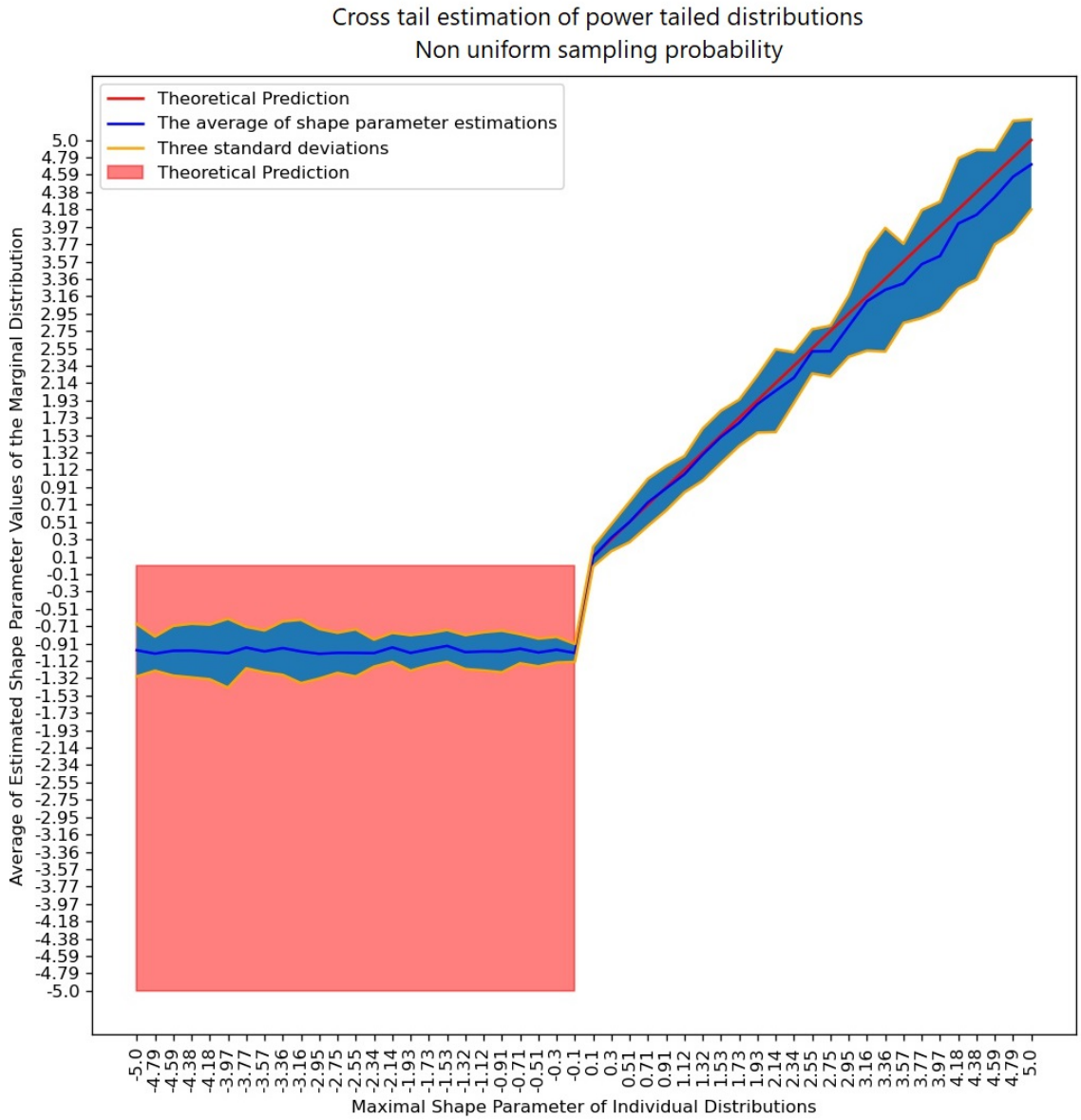

Figure 2: Equivalence of POT and CTE for tail shape estimation of the marginal distribution. In this case the conditional distributions have power tails, and the sampling probability between different conditional distributions is not uniform.

$\xi^i_{max} \leq 0$, we notice that $\hat{\xi}^i_{POT} \leq 0$ as well. In Appendix E, we show that the rate of convergence of $\hat{\xi}^i_{POT}$ to ξ^i_{max} is slow in terms of the size of m, which highlights the need of using cross tail estimation.

5.1.2 Cross tail estimation in the non-uniform case

In this experiment we perform a similar task, however in this case we assume a non-uniform distribution over the choice of j, that is,

$$f(j) = \frac{\frac{1}{\epsilon+|-k+\frac{j}{i-1}\xi|}}{\sum_j \frac{1}{\epsilon+|-k+\frac{j}{i-1}\xi|}}. \tag{32}$$

where $\epsilon > 0$, is a small value added to avoid division by zero. In practice, this implies that when we marginalize over j, we will pick less samples from arrays of size m corresponding to shape parameters with large absolute values. The size of the final array will be roughly $\sum_j \frac{m}{\epsilon+|-k+\frac{j}{i-1}\xi|}$, and we denote this sample set by $Z_{i,m}$, representing the $\sum_j \frac{m}{\epsilon+|-k+\frac{j}{i-1}\xi|}$ samples from the distribution $H(x) \in MDA(\xi_i)$. Using Pickands' estimator on $Z_{i,m}$, we estimate the shape parameter of the tail of $H(x)$, and expect that this estimated parameter $\hat{\xi}^i_{POT}$ will be close to ξ_i. Based on our theoretical results for positive ξ^i_{max} we have $\xi^i_{max} = \xi_i$, therefore it is expected that $\xi^i_{max} = \hat{\xi}^i_{POT}$.
We repeat this process for 50 different shape parameters ξ^i_{max} whose values are equidistant in the interval $[-5, 5]$, and are increasing in i. The results are shown in Figure 2. We see that the values of $\{\hat{\xi}^i_{POT}|i\}$ (in the y-axis), are close to those in $\{\xi^i_{max}|i\}$ (in the x-axis) when $\xi^i_{max} > 0$, for $m = 5 * 10^6$ as predicted by Theorem 13 and Theorem 20 . When $\xi^i_{max} \leq 0$, we notice that $\hat{\xi}^i_{POT} \leq 0$ as well.

5.2 The inadequacy of the direct POT usage on mixture distributions

In this section, we illustrate two cases where cross tail estimation is necessary for proper tail shape estimation.

5.2.1 Uniform Case

We sample with 50% probability from a distribution with power law tails with shape parameter 1, and with equal probability from a distribution with power law tails with shape parameter 0.5.

When we sample 10^3 points from each distribution, as seen in Figure 3 (left), we cannot properly estimate the tail if we join all the samples together in a common array and then apply Pickands Estimator. However, if we increase the sample size from 10^3 to $2 * 10^4$, we can retrieve the the true shape of the tail. In contrast, using our method, 10^3 samples are already sufficient to get a proper estimation Figure 3 (right).

5.2.2 Non-Uniform Case

Similarly, in the second experiment, we sample with 20% probability from a distribution with power law tails with shape parameter 1, and with 80% probability probability from a distribution with power law tails with shape parameter 0.5.

When sampling $5 * 10^3$ points from each distribution, Figure 4, we are not able to properly estimate the tail if we join all the samples together in a common array and then

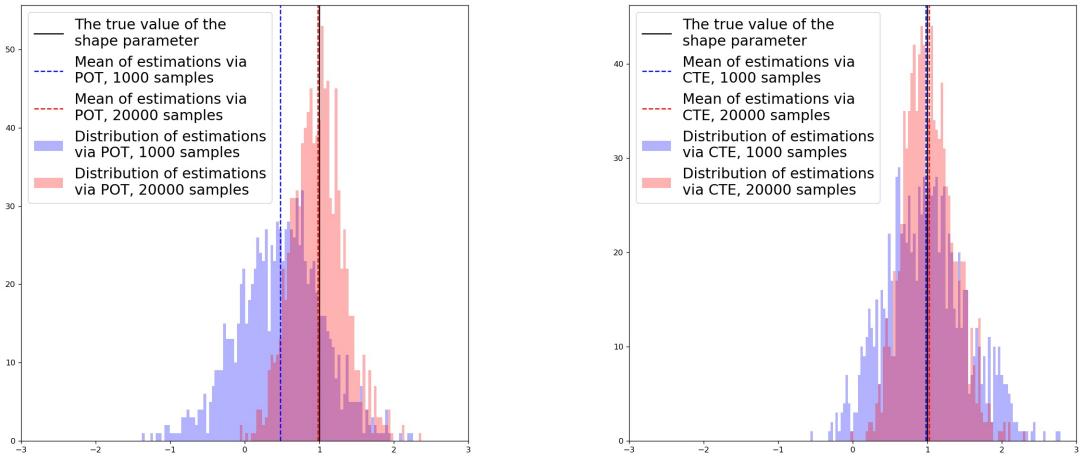

Figure 3: Standard estimation of the shape parameter of the tails by simply applying the Pickands' Estimator, on average, gives poor results on fewer data (left). Cross tail estimation (CTE) gives the correct estimation on average. (right).

apply the Pickands' Estimator. But, if we increase the sample size from $5 * 10^3$ to $5 * 10^7$, we manage to retrieve the the true tail shape of the mixture. However, using our method, $5 * 10^3$ samples are already sufficient to get a proper estimation.

5.3 Model performance inference improvements via cross tail estimation, relative to POT

In what follows, we show the results of two experiments, where we observe that cross tail estimation can improve the estimation of the shape of the tail in realistic settings. Furthermore, we observe that in these cases, the thickness of the tail is negatively correlated with performance, therefore inference regarding the performance of the model is improved when using CTE instead of POT.

5.3.1 Gaussian Processes

In this experiment, our data is composed of an one-dimensional time series taken from the UCR Time Series Anomaly Archive [2] Wu and Keogh (2020), which we reorganize in windows of size 50, and use each window to fit a Gaussian process (GP) model in order to predict the next value in the series. Our complete dataset D is composed of $n = 10^4$ windows. At each run, roughly 700 windows and their corresponding labels are chosen randomly to form a set we denote with D_f. We randomly select half of D_f for training

2. `https://www.cs.ucr.edu/~eamonn/time_series_data_2018/UCR_TimeSeriesAnomalyDatasets2021.zip`

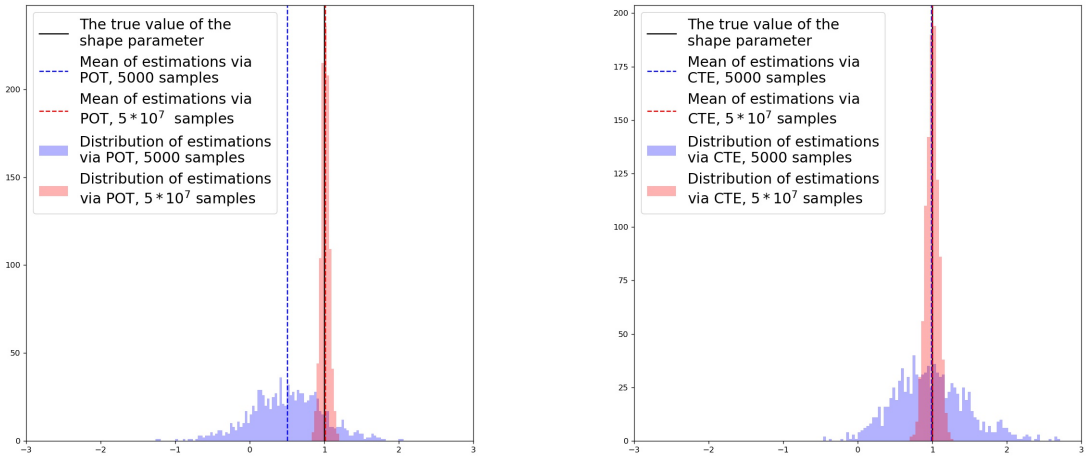

Figure 4: Standard estimation of the shape parameter of the tails by simply applying the Pickands' Estimator, on average, gives poor results on fewer data (left). Cross tail estimation (CTE) gives the correct estimation on average. (right).

(denote D_{fi}), and then group the predictions of the model on the 10^4 points of D into and array which we denote by $\hat{Y}_i$. We proceed to estimate the shape parameter of the tails of the prediction of the model for given training set D_{fi}, by applying Pickands' estimator to $\hat{Y}_i$, receiving $\hat{\xi}_i$. Keeping D_f fixed at all times, we repeat this process 10^3 times, and select the maximum individual estimated parameter, as our estimation of the shape parameter of the tail of the distribution of our loss function: $\hat{\xi}_{max} = \max\{\hat{\xi}_i | i \in [1000]\}$. On the other hand, we also calculate the MSE on the testing set $D \setminus D_{fi}$ after the model has been trained on D_{fi}. The set $D \setminus D_{fi}$ contains most of the points of D, as we wish to have enough testing points in order to get a good approximation of the tails. To check the difference of performance of the standard POT method of tail shape estimation and cross tail estimation, we also calculate the shape parameter of the overall distribution of prediction models via POT, by applying Pickands estimator on $Y = \bigcup_{i=1}^{1000} \hat{Y}_i$. The kernel used is the radial basis function (RBF), with length scale parameters varying from 1 to 30. We used the standard GP model implementation provided in Scikit-learn Pedregosa et al. (2011). We repeat this experiment 200 times. The results are shown in Figure 5. We notice the close behaviour between the MSE on test data, and the tail shape parameter of the distribution function estimated with cross tail estimation. It is visible that when the standard POT is used, we lose this relationship, as the shape parameter estimation for models with length scale greater than 28 is smaller than for those with lower MSE values on the test set.
A version of this experiment where $D_f = D$ and where the training set D_{fi} is created by

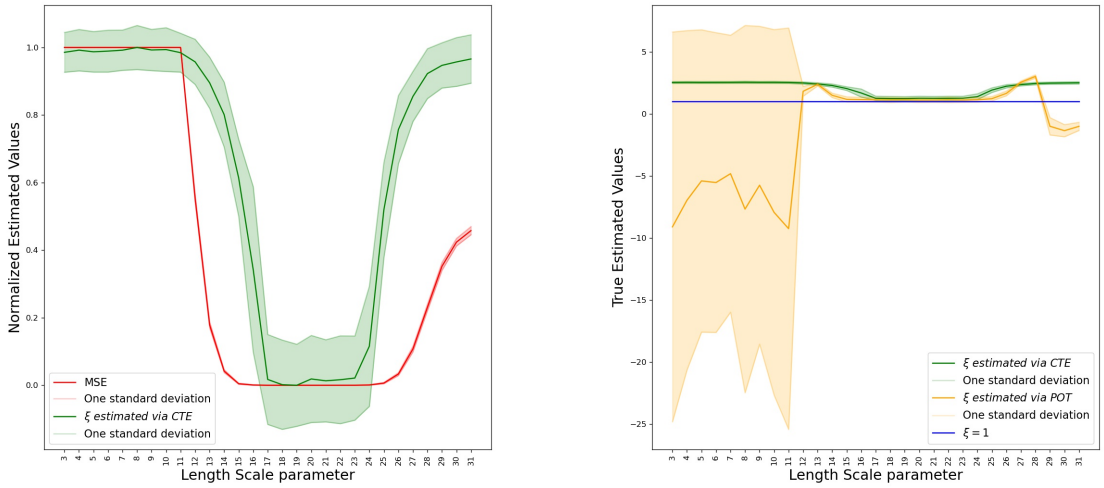

Figure 5: Experimental results in the case of testing Gaussian processes. Left: The relationship between MSE on test data, and the tail shape parameter of the distribution function estimated with cross tail estimation. Right: The tail shape parameter of the distribution function estimated with cross tail estimation and the standard POT approach.

randomly selecting 340 points from D, is given in Appendix D. The results displayed there support the ones presented here.

5.3.2 Polynomial Kernels

This experiment is almost identical to the previous one, with the only differences being that the models we test now are polynomial kernels, and the grid of possible candidate models in this case is defined by the degree of the polynomial kernel. We test polynomial kernels of degree from 1 to 9.

As before, we repeat this experiment 200 times. The results are shown in Figure 6. We notice the existence of a relationship between MSE on test data, and the tail shape parameter of the distribution function estimated with cross tail estimation. It is visible that when the standard POT method of estimating the tail is used, the tail shape estimates drop after degree 5 while the test MSE keeps increasing.
Similarly to the first experiment, a version of this experiment where $D_f = D$ and where the training set D_{fi} is created by randomly selecting 340 points from D, is given in Appendix D. The results displayed there support the ones presented here.

5.4 Computational Simplifications

Another benefit to using cross tail estimation is the reduction of computational time, as for a given number m of conditional distributions, with n samples for each, instead of

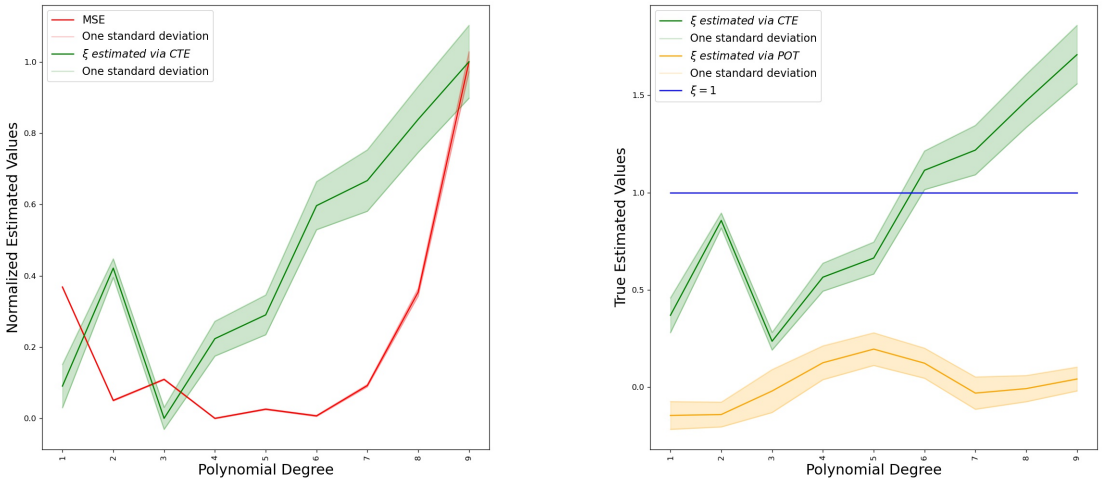

Figure 6: Experimental results in the case of testing polynomial kernels. Left: The relationship between MSE on test data, and the tail shape parameter of the distribution function estimated with cross tail estimation. Right: The tail shape parameter of the distribution function estimated with cross tail estimation and the standard POT approach. However, it is visible that when POT is used, the tail shape estimates drop after degree 5 while the test MSE keeps increasing.

joining all testing samples together in an array of size $m * n$, we perform calculations in m arrays of size n in parallel. This becomes useful in practice during shape parameter estimation, as using Pickands estimators requires sorted samples, where best algorithms for sorting require $n \log(n)$ operations for a vector of size n. Hence our method which requires $n \log(n)$ operations is much faster in practice than the standard POT approach which requires $mn \log(mn)$, in a setting where m and n are of approximately of the same order.

6. Conclusion

We study the problem of estimating the tail shape of loss function distributions, and explain the complications that arise in performing this task. We notice that such complications arise in general during the estimation of the tail shape of marginal distributions. In order to mitigate such shortcomings, we propose a new method of estimating the shape of the right tails of marginal distributions and give theoretical guarantees that the tail of the marginal distribution coincides with the thickest tail of the set of distributions defined on the points in range of the variable over which we integrate. We give experimental evidence that our method works in practice, and is necessary in applications with small sample sizes. Using the aforementioned method, we show experimentally that the tails of distribution functions in many cases can have non-exponential decay, as well as that it is possible that not even

their first moment exists. Furthermore, we discover an interesting phenomena regarding the relationship between the test MSE of a model, and the thickness of the tails of its prediction function distribution, in the experiments we conducted.

Potential additional applications of the method we develop include improving classic tail modelling, as well as the threshold selection for model comparison in anomaly detection Su et al. (2019). Furthermore, cross tail estimation could be used to estimate the existence of the moments of loss function distributions, and thus can be considered as a potential elimination criteria for models whose first moment does not exist.

Appendix A: Proofs

Proof of Proposition 8

We notice that if $L(x)$ converges the statement is trivial. However, if it does not then:

$$\begin{aligned}\lim_{x\to\infty} x^{-\epsilon}L(x) = \lim_{x\to\infty} \frac{L(x)}{x^{\epsilon}} = \lim_{x\to\infty} \frac{e^{c(x)}e^{\int_{x_o}^{x}\frac{u(y)}{y}dy}}{x^{\epsilon}} = \lim_{x\to\infty} \frac{e^{c(x)}e^{\int_{x_o}^{x}\frac{u(y)}{y}dy}}{e^{\epsilon\log(x)}} = \\ = \lim_{x\to\infty} e^{c(x)}e^{\int_{x_o}^{x}\frac{u(y)}{y}dy-\epsilon\log(x)} = \lim_{x\to\infty} e^{c(x)}e^{\log(x)(\frac{\int_{x_o}^{x}\frac{u(y)}{y}dy}{\log(x)}-\epsilon)}.\end{aligned} \tag{33}$$

Using L'Hopital's rule we get:

$$\lim_{x\to\infty} \frac{\int_{x_o}^{x}\frac{u(y)}{y}}{\log(x)} = \lim_{x\to\infty} \frac{\frac{u(x)}{x}}{\frac{1}{x}} = \lim_{x\to\infty} u(x) = 0, \tag{34}$$

therefore

$$\lim_{x\to\infty} e^{\log(x)(\frac{\int_{x_o}^{x}\frac{u(y)}{y}dy}{\log(x)}-\epsilon)} = 0. \tag{35}$$

Proof of Lemma 10

From Theorem 9, we get that

$$F_1 \in MDA(\xi_1) \iff \bar{F}_1(x) = x^{-\frac{1}{\xi_1}}L_1(x),$$

and

$$F_2 \in MDA(\xi_2) \iff \bar{F}_2(x) = x^{-\frac{1}{\xi_2}}L_2(x),$$

where $L_1(x)$ and $L_2(x)$ are slowly varying functions.
Therefore

$$\lim_{x\to\infty} \frac{\bar{F}_2(x)}{\bar{F}_1(x)} = \lim_{x\to\infty} x^{\frac{1}{\xi_1}-\frac{1}{\xi_2}} \frac{L_2(x)}{L_1(x)} = \lim_{x\to\infty} x^{\alpha}\frac{L_2(x)}{L_1(x)}, \tag{36}$$

since

$$\xi_1 > \xi_2 \implies -\frac{1}{\xi_1} > -\frac{1}{\xi_2} \implies \alpha := \frac{1}{\xi_1} - \frac{1}{\xi_2} < 0.$$

On the other hand $L(x) := \frac{L_2(x)}{L_1(x)}$ is defined in a neighborhood of infinity as $L_1(x) \neq 0$, and is also a slowly varying function as

$$\lim_{x\to\infty} \frac{L(ax)}{L(x)} = \lim_{x\to\infty} \frac{\frac{L_2(ax)}{L_1(ax)}}{\frac{L_2(x)}{L_1(x)}} = \lim_{x\to\infty} \frac{\frac{L_2(ax)}{L_2(x)}}{\frac{L_1(ax)}{L_1(x)}} = 1,$$

and since the quotient of positive measurable functions, is positive and measurable. Therefore, using Corollary 1, Equation (36) becomes

$$\lim_{x\to\infty} \frac{\bar{F}_2(x)}{\bar{F}_1(x)} = \lim_{x\to\infty} x^{\alpha}\frac{L_2(x)}{L_1(x)} = \lim_{x\to\infty} x^{\alpha}L(x) = 0. \tag{37}$$

Proof of Lemma 11

1. If $\xi_1 > 0$ and $\xi_2 = 0$ then

$$\lim_{x\to\infty} \frac{\bar{F}_2(x)}{\bar{F}_1(x)} = \lim_{x\to\infty} \frac{c(x)e^{-\int_w^x \frac{g(t)}{a(t)}dt}}{x^{-\frac{1}{\xi}}L(x)} = \lim_{x\to\infty} \frac{c(x)e^{-\log(x)(\frac{\int_w^x \frac{g(t)}{a(t)}dt}{\log(x)}-\frac{1}{\xi})}}{L(x)}, \tag{38}$$

using L'Hopital's rule:

$$\lim_{x\to\infty} \frac{\int_w^x \frac{g(t)}{a(t)}dt}{\log(x)} = \lim_{x\to\infty} \frac{\frac{g(x)}{a(x)}}{\frac{1}{x}} = \lim_{x\to\infty} \frac{x}{a(x)}, \tag{39}$$

we distinguish two cases:

if $\lim_{x\to\infty} a(x) \neq \infty$ then $\lim_{x\to\infty} \frac{x}{a(x)} = \infty$,

while if $\lim_{x\to\infty} a(x) = \infty$ then using L'Hopital's rule again, we obtain

$$\lim_{x\to\infty} \frac{x}{a(x)} = \lim_{x\to\infty} \frac{1}{a'(x)} = \infty. \tag{40}$$

Thus, in both cases

$$= \lim_{x\to\infty} \frac{c(x)e^{-\log(x)(\frac{\int_w^x \frac{g(t)}{a(t)}dt}{\log(x)}-\frac{1}{\xi})}}{L(x)} = \lim_{x\to\infty} \frac{c(x)x^{-(\frac{\int_w^x \frac{g(t)}{a(t)}dt}{\log(x)}-\frac{1}{\xi})}}{L(x)} = 0. \tag{41}$$

Statements 2. 3. and 4. are trivial.

Proof of Lemma 12

Since $L(x)$ is positive and measurable (linear combination of finite measurable functions), the only part left to prove is that

$$\lim_{x\to\infty} \frac{L(ax)}{L(x)} = 1, \forall a > 0.$$

First we prove that

$$\lim_{x\to\infty} \frac{L_1(ax) + L_2(ax)}{L_1(x) + L_2(x)} = 1, \forall a > 0.$$

Indeed, for each $\epsilon > 0$, there exist x_1, x_2 such that for $x > x_1$ we have $|\frac{L_1(ax)}{L_1(x)} - 1| < \epsilon$ and for $x > x_2$ we have $|\frac{L_2(ax)}{L_2(x)} - 1| < \epsilon$. Hence for $x_0 = max\{x_1, x_2\}$, $x > x_0$ implies $|L_1(ax) - L_1(x)| < L_1(x)\epsilon$ and $|L_2(ax) - L_2(x)| < L_2(x)\epsilon$ therefore $|L_1(ax) + L_2(ax) - (L_1(x)+L_2(x))| = |L_1(ax)-L_1(x)+L_2(ax)-L_2(x)| \leq |L_1(ax)-L_1(x)|+|L_2(ax)-L_2(x)| < (L_1(x) + L_2(x))\epsilon$ hence $|\frac{L_1(ax)+L_2(ax)}{L_1(x)+L_2(x)} - 1| < \epsilon$.

Now, we notice that for every $a_i > 0$, we get $\lim_{x\to\infty} \frac{a_i L_i(ax)}{a_i L_i(x)} = 1$, and $a_i L_i(x)$ is positive as well as measurable. This implies that $a_1 L_1$ and $a_2 L_2$ are slowly varying functions, and therefore based of the previous result we get

$$\lim_{x\to\infty} \frac{a_1 L_1(ax) + a_2 L_2(ax)}{a_1 L_1(x) + a_2 L_2(x)} = 1, \forall a > 0.$$

Using induction finishes the proof of the Lemma.

Proof of Theorem 13

Since if $\xi_{\boldsymbol{z}_i} < 0$ then $\exists x_0 > 0$, such that $\forall x > x_0$ we have $H_{\boldsymbol{Z}_i}(x) = 0$, this means that the tail of the distribution is not affected by $H_{\boldsymbol{Z}_i}(x)$. In fact if $\xi_{max} < 0$ then H will have finite support hence $\xi_H \leq 0$. Furthermore if $\xi_{max} = 0$ from Lemma 11 we get that $\xi_H \leq 0$. Therefore for the case $\xi_{max} > 0$ we only consider the setting where $\xi_i \geq 0$.

$$\bar{H}_u(w) = \frac{1 - H(u+w)}{1 - H(u)} = \frac{\sum\limits_i^n p_i(1 - H_{\boldsymbol{z}_i}(u+w))}{\sum\limits_i^n p_i(1 - H_{\boldsymbol{z}_i}(u))} = \sum_i^n \frac{\bar{H}_{\boldsymbol{z}_i}(u+w)}{\sum\limits_j^n \frac{p_j}{p_i}\bar{H}_{\boldsymbol{z}_j}(u)} \tag{42}$$

$$= \sum_i^n \frac{\bar{H}_{\boldsymbol{z}_i}(u+w)}{\bar{H}_{\boldsymbol{z}_i}(u)} \frac{\bar{H}_{\boldsymbol{z}_i}(u)}{\sum\limits_j^n \frac{p_j}{p_i}\bar{H}_{\boldsymbol{z}_j}(u)} = \sum_i^n \frac{\bar{H}_{\boldsymbol{z}_i}(u+w)}{\bar{H}_{\boldsymbol{z}_i}(u)} \frac{1}{\sum\limits_j^n \frac{p_j}{p_i}\frac{\bar{H}_{\boldsymbol{z}_j}(u)}{\bar{H}_{\boldsymbol{z}_i}(u)}}. \tag{43}$$

We denote with $i(max)$ the index corresponding to ξ_{max} and finish our proof using Pickand's theorem:

$$\lim_{u\to\infty} \sup_{w\in[0,\infty]} |\bar{H}_u(y) - \bar{G}_{\xi_{max},g(u)}| = \lim_{u\to\infty} \sup_{w\in[0,\infty]} |\sum_i^n \frac{\bar{H}_{\boldsymbol{z}_i}(u+w)}{\bar{H}_{\boldsymbol{z}_i}(u)} \frac{1}{\sum\limits_j^n \frac{p_j}{p_i}\frac{\bar{H}_{\boldsymbol{z}_j}(u)}{\bar{H}_{\boldsymbol{z}_i}(u)}} - \bar{G}_{\xi_{max},g(u)}| \tag{44}$$

$$= \lim_{u\to\infty} \sup_{w\in[0,\infty]} |\sum_i^n \frac{\bar{H}_{\boldsymbol{z}_i}(u+w)}{\bar{H}_{\boldsymbol{z}_i}(u)} \frac{1}{1 + \sum\limits_{j\neq i}^n \frac{p_j}{p_i}\frac{\bar{H}_{\boldsymbol{z}_j}(u)}{\bar{H}_{\boldsymbol{z}_i}(u)}} - \bar{G}_{\xi_{max},g(u)}| \tag{45}$$

$$\begin{aligned} &\leq \lim_{u\to\infty} \sup_{w\in[0,\infty]} |\frac{\bar{H}_{\boldsymbol{z}_{i(max)}}(u+w)}{\bar{H}_{\boldsymbol{z}_{i(max)}}(u)} \frac{1}{1 + \sum\limits_{j\neq i(max)}^n \frac{p_j}{p_{i(max)}}\frac{\bar{H}_{\boldsymbol{z}_j}(u)}{\bar{H}_{\boldsymbol{z}_{i(max)}}(u)}} - \bar{G}_{\xi_{max},g(u)}| \\ &+ \lim_{u\to\infty} \sup_{w\in[0,\infty]} |\sum_{i\neq i(max)}^n \frac{\bar{H}_{\boldsymbol{z}_i}(u+w)}{\bar{H}_{\boldsymbol{z}_i}(u)} \frac{1}{1 + \sum\limits_{j\neq i}^n \frac{p_j}{p_i}\frac{\bar{H}_{\boldsymbol{z}_j}(u)}{\bar{H}_{\boldsymbol{z}_i}(u)}}| \end{aligned} \tag{46}$$

$$\begin{aligned} &\leq \lim_{u\to\infty} \sup_{w\in[0,\infty]} |\frac{\bar{H}_{\boldsymbol{z}_{i(max)}}(u+w)}{\bar{H}_{\boldsymbol{z}_{i(max)}}(u)} - \bar{G}_{\xi_{max},g(u)}| \\ &+ \lim_{u\to\infty} \sup_{w\in[0,\infty]} |\frac{1}{1 + \sum\limits_{j\neq i(max)}^n \frac{p_j}{p_{i(max)}}\frac{\bar{H}_{\boldsymbol{z}_j}(u)}{\bar{H}_{\boldsymbol{z}_{i(max)}}(u)}} - 1||\frac{\bar{H}_{\boldsymbol{z}_{i(max)}}(u+w)}{\bar{H}_{\boldsymbol{z}_{i(max)}}(u)}| \\ &+ \lim_{u\to\infty} \sup_{w\in[0,\infty]} \sum_{i\neq i(max)}^n |\frac{\bar{H}_{\boldsymbol{z}_i}(u+w)}{\bar{H}_{\boldsymbol{z}_i}(u)}||\frac{1}{1 + \sum\limits_{j\neq i}^n \frac{p_j}{p_i}\frac{\bar{H}_{\boldsymbol{z}_j}(u)}{\bar{H}_{\boldsymbol{z}_i}(u)}}| \end{aligned} \tag{47}$$

$$
\begin{aligned}
&\leq \lim_{u\to\infty} \sup_{w\in[0,\infty]} |\frac{\bar{H}_{\boldsymbol{z}_{i(max)}}(u+w)}{\bar{H}_{\boldsymbol{z}_{i(max)}}(u)} - \bar{G}_{\xi_{max},g(u)}| \\
&+ \lim_{u\to\infty} |\frac{1}{1+\sum\limits_{j\neq i(max)}^{n} \frac{p_j}{p_{i(max)}} \frac{\bar{H}_{\boldsymbol{z}_j}(u)}{\bar{H}_{\boldsymbol{z}_{i(max)}}(u)}} - 1| \\
&+ \lim_{u\to\infty} \sum_{i\neq i(max)}^{n} |\frac{1}{1+\sum\limits_{j\neq i}^{n} \frac{p_j}{p_i} \frac{\bar{H}_{\boldsymbol{z}_j}(u)}{\bar{H}_{\boldsymbol{z}_i}(u)}}|.
\end{aligned} \tag{48}
$$

The first expression,

$$
\lim_{u\to\infty} \sup_{w\in[0,\infty]} |\frac{\bar{H}_{\boldsymbol{z}_{i(max)}}(u+w)}{\bar{H}_{\boldsymbol{z}_{i(max)}}(u)} - \bar{G}_{\xi_{max},g(u)}| \tag{49}
$$

goes to zero due to Pickands Theorem while the expression,

$$
\lim_{u\to\infty} |\frac{1}{1+\sum\limits_{j\neq i(max)}^{n} \frac{p_j}{p_{i(max)}} \frac{\bar{H}_{\boldsymbol{z}_j}(u)}{\bar{H}_{\boldsymbol{z}_{i(max)}}(u)}} - 1| \tag{50}
$$

converges to 0 as well because from Lemma 10 we have $\lim_{u\to\infty} \frac{\bar{H}_{\boldsymbol{z}_j}(u)}{\bar{H}_{\boldsymbol{z}_{i(max)}}(u)} = 0$ for every j. Finally the last expression,

$$
\lim_{u\to\infty} \sum_{i\neq i(max)}^{n} |\frac{1}{1+\sum\limits_{j\neq i}^{n} \frac{p_j}{p_i} \frac{\bar{H}_{\boldsymbol{z}_j}(u)}{\bar{H}_{\boldsymbol{z}_i}(u)}}| \tag{51}
$$

equals 0 since in each sum $\sum\limits_{j\neq i}^{n} \frac{p_j}{p_i} \frac{\bar{H}_{\boldsymbol{z}_j}(u)}{\bar{H}_{\boldsymbol{z}_i}(u)}$, there exists an index j such that $\bar{H}_{\boldsymbol{z}_j}(u) = \bar{H}_{\boldsymbol{z}_{i(max)}}(u)$, implying that $\sum\limits_{j\neq i}^{n} \frac{p_j}{p_i} \frac{\bar{H}_{\boldsymbol{z}_j}(u)}{\bar{H}_{\boldsymbol{z}_i}(u)} \to \infty$.

In the derivation above we assumed that the $H_{\boldsymbol{z}_{i(max)}}$ which corresponds to ξ_{max} is unique. In the case that this is not true we notice that for H_1 and H_2 which share the same corresponding parameter $\xi > 0$ we have

$$
p_1 H_1(x) + p_2 H_2(x) = x^{-\frac{1}{\xi}}(p_1 L_1(x) + p_2 L_2(x)) = x^{-\frac{1}{\xi}} L(x), \tag{52}
$$

and since $L(x) > 0$, from Lemma 12 we have that $L(x)$ is slowly varying, therefore $p_1 H_1(x) + p_2 H_2(x) \in MDA(\xi)$.

Proof of Proposition 15

First, we fix $\delta > 0$. We can find a $x(\gamma, \delta) > 0$, such that for $x > x(\gamma, \delta)$, we can bound $x^{-\delta} L_z(x) < \gamma$ for all $z \in A$ simultaneously. This implies that $f_Z(z) x^{-\delta} L_z(x)$ is bounded by

$f_z(z)\gamma$. Since $\int_z f_z(z)\gamma dz = \gamma < \infty$, by dominated convergence we get

$$\lim_{x\to\infty} x^{-\delta}\int_A f_Z(z)L_z(x)dz = \lim_{x\to\infty}\int_A f_Z(z)x^{-\delta}L_z(x)dz = \int_A \lim_{x\to\infty} f_Z(z)x^{-\delta}L_z(x)dz = 0. \tag{53}$$

Proof of Theorem 16

We will first assume that $\xi_H > 0$.
Since $\bar{H}(x) = x^{-\frac{1}{\xi_H}}L_H(x)$, for every $\epsilon > 0$:

$$\frac{\bar{H}(x)}{x^{-\frac{1}{\xi_{lo}-\epsilon}}} = \frac{x^{-\frac{1}{\xi_H}}L_H(x)}{x^{-\frac{1}{\xi_{lo}-\epsilon}}} = \frac{\int_A f_{\boldsymbol{Z}}(\boldsymbol{z})x^{-\frac{1}{\xi_{\boldsymbol{z}}}}L_{\boldsymbol{z}}(x)d\boldsymbol{z}}{x^{-\frac{1}{\xi_{lo}-\epsilon}}} = \int_A f_{\boldsymbol{Z}}(\boldsymbol{z})x^{-\frac{1}{\xi_{\boldsymbol{z}}}+\frac{1}{\xi_{lo}-\epsilon}}L_{\boldsymbol{z}}(x)d\boldsymbol{z}\int_A f_{\boldsymbol{Z}}(\boldsymbol{z})x^{\alpha(\boldsymbol{z})}L_{\boldsymbol{z}}(x)d\boldsymbol{z}. \tag{54}$$

We notice that $\xi_{\boldsymbol{z}} \geq \xi_{lo} > \xi_{lo} - \epsilon \implies -\frac{1}{\xi_{\boldsymbol{z}}} \geq -\frac{1}{\xi_{lo}} > -\frac{1}{\xi_{lo}-\epsilon}$ hence $\alpha(\boldsymbol{z}) = -\frac{1}{\xi_{\boldsymbol{z}}} + \frac{1}{\xi_{lo}-\epsilon} > 0$.
Considering that

$$\lim_{x\to\infty}\frac{\bar{H}(x)}{x^{-\frac{1}{\xi_{lo}-\epsilon}}} = \lim_{x\to\infty}\int_A f_{\boldsymbol{Z}}(\boldsymbol{z})x^{\alpha(\boldsymbol{z})}L_{\boldsymbol{z}}(x)dz, \tag{55}$$

by using Fatou's lemma:

$$\lim_{x\to\infty}\int_A f_{\boldsymbol{Z}}(\boldsymbol{z})x^{\alpha(\boldsymbol{z})}L_{\boldsymbol{z}}(x)d\boldsymbol{z} \geq \int_A \lim_{x\to\infty} f_{\boldsymbol{Z}}(\boldsymbol{z})x^{\alpha(\boldsymbol{z})}L_{\boldsymbol{z}}(x)d\boldsymbol{z} = \infty, \tag{56}$$

we get

$$\lim_{x\to\infty}\frac{x^{-\frac{1}{\xi_{lo}-\epsilon}}}{\bar{H}(x)} = 0, \tag{57}$$

implying

$$\lim_{x\to\infty}\frac{x^{-\frac{1}{\xi_{lo}-\epsilon}}}{x^{-\frac{1}{\xi_H}}L_H(x)} = \lim_{x\to\infty}\frac{x^{-\frac{1}{\xi_{lo}-\epsilon}+\frac{1}{\xi_H}}}{L_H(x)} = 0, \tag{58}$$

therefore

$$\xi_{lo} - \epsilon < \xi_H, \forall\epsilon > 0 \text{ thus } \xi_{lo} \leq \xi_H. \tag{59}$$

Now we turn to prove that $\xi_H \leq \xi_{up}$. As before,

$$\frac{\bar{H}(x)}{x^{-\frac{1}{\xi_{up}+\epsilon}}} = \frac{x^{-\frac{1}{\xi_H}}L_H(x)}{x^{-\frac{1}{\xi_{up}+\epsilon}}} = \frac{\int_A f_{\boldsymbol{Z}}(\boldsymbol{z})x^{-\frac{1}{\xi_{\boldsymbol{z}}}}L_{\boldsymbol{z}}(x)d\boldsymbol{z}}{x^{-\frac{1}{\xi_{up}+\epsilon}}} = \int_A f_{\boldsymbol{Z}}(\boldsymbol{z})x^{-\frac{1}{\xi_{\boldsymbol{z}}}+\frac{1}{\xi_{up}+\epsilon}}L_{\boldsymbol{z}}(x)d\boldsymbol{z} = \int_A f_{\boldsymbol{Z}}(\boldsymbol{z})x^{\beta(\boldsymbol{z})}L_{\boldsymbol{z}}(x)d\boldsymbol{z}. \tag{60}$$

We notice that $\xi_z \leq \xi_{up} < \xi_{up} + \epsilon \implies -\frac{1}{\xi_z} \leq -\frac{1}{\xi_{up}} < -\frac{1}{\xi_{up}+\epsilon} = -\frac{1}{\xi_{up}} + \delta$ hence $\beta(z) = -\frac{1}{\xi_z} + \frac{1}{\xi_{up}+\epsilon} < -\delta$. This last inequality, combined with the fact that the family $\{L_{\boldsymbol{z}}(x)|x \in \mathbb{R}\}$ is γ-uniformly sub-polynomial, implies that

$$f_{\boldsymbol{Z}}(\boldsymbol{z})x^{\beta(\boldsymbol{z})}L_{\boldsymbol{z}}(x) \leq f_{\boldsymbol{Z}}(\boldsymbol{z})x^{-\delta}L_{\boldsymbol{z}}(x) \leq f_{\boldsymbol{Z}}(\boldsymbol{z})\gamma, \tag{61}$$

for some $\gamma > 0$. Since $\int_{\boldsymbol{z}} f_{\boldsymbol{Z}}(\boldsymbol{z})\gamma d\boldsymbol{z} = \gamma < \infty$, by dominated convergence

$$\lim_{x\to\infty} \frac{\bar{H}(x)}{x^{-\frac{1}{\xi_{up}+\epsilon}}} = \lim_{x\to\infty} \int_A f_{\boldsymbol{Z}}(\boldsymbol{z}) x^{\beta(\boldsymbol{z})} L_{\boldsymbol{z}}(x) d\boldsymbol{z} \tag{62}$$

$$\lim_{x\to\infty} \int_A f_{\boldsymbol{Z}}(\boldsymbol{z}) x^{\beta(\boldsymbol{z})} L_{\boldsymbol{z}}(x) d\boldsymbol{z} = \int_A \lim_{x\to\infty} f_{\boldsymbol{z}}(\boldsymbol{z}) x^{\beta(\boldsymbol{z})} L_{\boldsymbol{z}}(x) d\boldsymbol{z} = 0, \tag{63}$$

meaning

$$\lim_{x\to\infty} \frac{\bar{H}(x)}{x^{-\frac{1}{\xi_{up}+\epsilon}}} = 0, \tag{64}$$

which implies

$$\lim_{x\to\infty} \frac{x^{-\frac{1}{\xi_H}} L_H(x)}{x^{-\frac{1}{\xi_{up}+\epsilon}}} = \lim_{x\to\infty} x^{\frac{1}{\xi_{up}+\epsilon}-\frac{1}{\xi_H}} L_H(x) = 0, \tag{65}$$

therefore we get

$$\xi_{up} + \epsilon > \xi_H, \forall \epsilon > 0 \text{ hence } \xi_H \leq \xi_{up}. \tag{66}$$

Now we prove that indeed $\xi_H > 0$. It is simple to show that ξ_H cannot be negative. Indeed, if ξ_H is negative, it means that H has finite support which is not possible as for each fixed x, we have $H_{\boldsymbol{z}}(x) > 0, \forall \boldsymbol{z} \in A$, therefore $\forall x \in \mathbb{R}, H(x) > 0$.
Proving that $\xi_H \neq 0$ is slightly less trivial. For every distribution $G_0 \in MDA(0)$ and for $\epsilon < \xi_{lo}$

$$\frac{\bar{H}(x)}{\bar{G}_0(x)} = \frac{\bar{H}(x)}{x^{-\frac{1}{\epsilon}}} \frac{x^{-\frac{1}{\epsilon}}}{\bar{G}_0(x)} = \frac{\int_A f_{\boldsymbol{Z}}(\boldsymbol{z}) x^{-\frac{1}{\xi_{\boldsymbol{z}}}} L_{\boldsymbol{z}}(x) d\boldsymbol{z}}{x^{-\frac{1}{\epsilon}}} \frac{x^{-\frac{1}{\epsilon}}}{\bar{G}_0(x)}. \tag{67}$$

As before we can prove that the first fraction $\frac{\bar{H}(x)}{x^{-\frac{1}{\epsilon}}} \to \infty$. The expression $\frac{x^{-\frac{1}{\epsilon}}}{\bar{G}_0(x)}$ goes to ∞ as well due to Lemma 11, thus

$$\lim_{x\to\infty} \frac{\bar{H}(x)}{\bar{G}_0(x)} = \infty. \tag{68}$$

If ξ_H was 0, then for some $G_0 \in MDA(0)$ we would have

$$\lim_{x\to\infty} \frac{\bar{H}(x)}{\bar{G}_0(x)} = \lim_{x\to\infty} 1 = 1, \tag{69}$$

hence $\xi_H \neq 0$.

Finally we prove that, if $\xi_{\boldsymbol{z}}$ is continuous in $\boldsymbol{z}$ and ξ_{max} exists, then we have $\xi_H = \xi_{max}$. We will first separate A in two sets A_1, A_2, where $A_1 = \{\boldsymbol{z} | \xi_{max} - \lambda \leq \xi_{\boldsymbol{z}} \leq \xi_{max}\}$ and $A_2 = \{\boldsymbol{z} | \xi_{lo} \leq \xi_{\boldsymbol{z}} < \xi_{max} - \lambda\}$. Since $\xi_{\boldsymbol{z}}$ is continuous, then the pre-image of each of the measurable sets $[\xi_{max} - \lambda, \xi_{max}], [\xi_{lo}, \xi_{max} - \lambda)$ will be measurable. In addition, since $[\xi_{max} - \lambda, \xi_{max}]$ and $[\xi_{lo}, \xi_{max} - \lambda)$ contain an open set, then so will A_1 and A_2, implying that $p_i = \mathbb{P}(A_i) > 0$, where $i \in \{1, 2\}$. Thus,

$$\bar{H}(x) = \int_A f_{\boldsymbol{Z}}(\boldsymbol{z})\bar{H}_z(x)d\boldsymbol{z} = p_1 \int_{A_1} \frac{f_{\boldsymbol{Z}}(\boldsymbol{z})}{p_1}\bar{H}_{\boldsymbol{z}}(x)d\boldsymbol{z} + p_2 \int_{A_2} \frac{f_{\boldsymbol{Z}}(\boldsymbol{z})}{p_2}\bar{H}_{\boldsymbol{z}}(x)d\boldsymbol{z} \\ = p_1\bar{H}_1(x) + p_2\bar{H}_2(x). \tag{70}$$

From the first part of the Theorem: $\xi_1 \in [\xi_{max} - \lambda, \xi_{max}]$, and $\xi_2 \in [\xi_{lo}, \xi_{max} - \lambda]$, where $H_i \in MDA(\xi_i),\ i = 1, 2$. On the other hand Theorem 13 implies that $\xi_H = \xi_1$, therefore $\xi_H \in [\xi_{max} - \lambda, \xi_{max}]$ for all $\lambda > 0$. We conclude that $\xi_H = \xi_{max}$.

Proof of Lemma 18

We assume that $\xi_H > \epsilon$. Then as in the earlier derivations, due to dominated convergence and Lemmas 10 and 11, for any $\delta > 0$, we get:

$$\lim_{x\to\infty} \frac{x^{-\frac{1}{\xi_H}} L_H(x)}{x^{-\frac{1}{\epsilon+\delta}}} = \lim_{x\to\infty} \frac{\bar{H}(x)}{x^{-\frac{1}{\epsilon+\delta}}} = \lim_{x\to\infty} \int_A f_{\boldsymbol{Z}}(\boldsymbol{z}) \frac{\bar{H}_{\boldsymbol{z}}(x)}{x^{-\frac{1}{\epsilon+\delta}}} d\boldsymbol{z} \\ = \int_A \lim_{x\to\infty} f_{\boldsymbol{Z}}(\boldsymbol{z}) \frac{x^{-\frac{1}{\xi_{\boldsymbol{z}}}}}{x^{-\frac{1}{\epsilon+\delta}}} L_{\boldsymbol{z}}(x) d\boldsymbol{z} = 0. \tag{71}$$

therefore $\xi_H < \epsilon + \delta, \forall\delta > 0$, contradicting our assumption $\xi_H > \epsilon$.

Proof of Theorem 20

The proof is similar to that of the last statement in Theorem 16. We will first separate A in two sets A_1, A_2, where $A_1 = \{\boldsymbol{z} | \xi_{max} - \lambda \leq \xi_{\boldsymbol{z}} \leq \xi_{max}\}$ and $A_2 = \{\boldsymbol{z} | \xi_{\boldsymbol{z}} < \xi_{max} - \lambda\}$. Since $\xi_{\boldsymbol{z}}$ is continuous, then the pre-image of each of the measurable sets $[\xi_{max} - \lambda, \xi_{max}], (-\infty, \xi_{max} - \lambda)$, will be measurable. In addition, since $[\xi_{max} - \lambda, \xi_{max}]$ and $(-\infty, \xi_{max} - \lambda)$ contain an open set, then so will A_1 and A_2, implying that $p_i = \mathbb{P}(A_i) > 0$, where $i \in \{1, 2\}$.

$$\bar{H}(x) = \int_A f_{\boldsymbol{Z}}(\boldsymbol{z})\bar{H}_z(x)d\boldsymbol{z} = p_1 \int_{A_1} \frac{f_{\boldsymbol{Z}}(\boldsymbol{z})}{p_1}\bar{H}_{\boldsymbol{z}}(x)d\boldsymbol{z} + p_2 \int_{A_2} \frac{f_{\boldsymbol{Z}}(\boldsymbol{z})}{p_2}\bar{H}_{\boldsymbol{z}}(x)d\boldsymbol{z} \\ = p_1\bar{H}_1(x) + p_2\bar{H}_2(x). \tag{72}$$

Based on Theorem 16 and Lemma 18: $\xi_1 = \xi_{max}$, and $\xi_2 \in (-\infty, \xi_{max} - \lambda]$, where $H_i \in MDA(\xi_i),\ i = 1, 2$. From Theorem 13, we conclude that $\xi_H = \xi_{max}$. The last statement in the Theorem, that is, if $\xi_{max} \leq 0$ then $\xi_H \leq 0$, is simply Corollary 19.

Proof of Proposition 20

Based on our assumptions there exists $L(x)$ such that

$$\mathbb{P}(X > x) = \bar{F}_X(x) = x^{-\frac{1}{\xi_X}} L_1(x). \tag{73}$$

Therefore

$$\bar{F}_Y(x) = \mathbb{P}(Y > x) = \mathbb{P}(X^\alpha > x) = \mathbb{P}(X > x^{\frac{1}{\alpha}}) = (x^{\frac{1}{\alpha}})^{-\frac{1}{\xi_X}} L_1(x^{\frac{1}{\alpha}}) = x^{-\frac{1}{\alpha\xi_X}} L_2(x), \tag{74}$$

however we also have

$$\bar{F}_Y(x) = x^{-\frac{1}{\xi_Y}} L_3(x). \tag{75}$$

Hence we conclude that $\xi_Y = \alpha \xi_X$.

Proof of Theorem 23

We will first prove the case when $p = 1$. If denote with W_1 the distribution of $Y - \hat{f}_{\boldsymbol{V}}(\boldsymbol{X})$ conditional that it is positive, and respectively with W_2 in the negative case, then without loss of generality if we suppose that W_1 is the thickest tail, we conclude from Proposition 21 that $|Y - \hat{f}_{\boldsymbol{V}}(\boldsymbol{X})|$ has the same shape parameter as $Y - \hat{f}_{\boldsymbol{V}}(\boldsymbol{X})$. Now, due to the fact that for a fixed test set $(\boldsymbol{x}, y)$, the variable $y - \hat{f}_{\boldsymbol{V}}(\boldsymbol{x})$ has the same shape parameter as $\hat{f}_{\boldsymbol{V}}(\boldsymbol{x})$, since a change in location parameter does not change the tail of the distribution, we conclude that $|y - \hat{f}_{\boldsymbol{V}}(\boldsymbol{x})|$ and $|\hat{f}_{\boldsymbol{V}}(\boldsymbol{x})|$, have tails of the same shape. Applying Theorem 20, we reach the desired conclusion for $|Y - \hat{f}_{\boldsymbol{V}}(\boldsymbol{X})|$ and $|\hat{f}_{\boldsymbol{V}}(\boldsymbol{X})|$. By applying Proposition 22 twice, this result can be generalized in the case of $|Y - \hat{f}_{\boldsymbol{V}}(\boldsymbol{X})|^p$ and $|\hat{f}_{\boldsymbol{V}}(\boldsymbol{X})|^p$.

Appendix B: Examples where the regularity conditions do not hold

Below we give examples where the regularity conditions do not hold:

Example 1: Let $f_U(u)$ be a uniform distribution, and $g_u(w)$ an exponential distribution with parameter $\frac{1}{u}$. Clearly, the expectation of $g_u(w)$ at each $u \in (0, 1)$ exists. However for

$$h(w) = \int_0^1 f_U(u) g_u(w) du = \int_0^1 u e^{-uw} du \tag{76}$$

the expectation is

$$\int_0^\infty \int_0^1 w f_U(u) g_u(w) du dw = \int_0^1 \int_0^\infty w u e^{-uw} dw du = \int_0^1 \frac{1}{u} du \tag{77}$$

In this example, we can see that even though all the distributions $g_u(w)$ have shape parameter 0, the shape parameter of $h(w)$ is bigger or equal to one. This is because the beginning of the exponential behaviour of the tail is delayed indefinitely across the elements of the family, violating the γ-uniform sub-polynomial assumption.

Below we give an example of a family of slowly-varying functions $\{L_z(x) | z \in A\}$, where A is compact and $L_z(x)$ is continuous in x and z, but $\{L_z(x) | z \in A\}$ is not γ-uniformly sub-polynomial. In this case, the non slowly-varying behaviour (non sub-polinomiality) of $L_z(x)$, or in other words, the tail of $F_z(x)$, is postponed indefinitely across the family of $\{F_z(x) | z \in A\}$

Example 2: Let $L_z(x)$, for $z \in [0, 1]$, be defined as below:

$$L_z(x) = \begin{cases} 1 + z x^{4-(z-\frac{1}{x})^2} & \text{for } x \in (1, \frac{1}{z}) \\ 1 + \frac{1}{z^3} & \text{for } x \in (\frac{1}{z}, \infty) \end{cases} \tag{78}$$

when $z \neq 0$ and $L_0 = 1$ for $x \in (\frac{1}{z}, \infty)$. For x^{-1} we define $F_z(x) = x^{-1}L_z(x)$, that is:

$$F_z(x) = \begin{cases} x^{-1} + zx^{3-(z-\frac{1}{x})^2} & \text{for } x \in (1, \frac{1}{z}) \\ x^{-1} + \frac{1}{z^3}x^{-1} & \text{for } x \in (\frac{1}{z}, \infty) \end{cases} \tag{79}$$

when $z \neq 0$ and $F_0 = x^{-1}$ for $x \in (\frac{1}{z}, \infty)$. One can check that $F_z(x)$ and $L_z(x)$ are continuous in z. On the other hand for a given z, $F_z(\frac{1}{z}) = z + z^{-2}$, meaning that $F_z(\frac{1}{z})$ tends to infinity, when z tends to zero. Therefore $\{L_z(x)|z \in A\}$ is not γ-uniformly sub-polynomial.

Appendix C: Examples where the regularity conditions hold

Below we give examples where the regularity conditions do hold:

Example 3: Let $\bar{F}_z(x) = x^{-z} = x^{-\frac{1}{\frac{1}{z}=\xi_z}}$ for $z \in (1, \infty)$, and let $\bar{F}(x) = e\int_1^\infty e^{-z}\bar{F}_z(x)dz$. Then $\bar{F}(x) = x^{-1}\frac{1}{1+\ln x} = x^{-1}L(x)$, where $L(x) = \frac{1}{\ln x}$ is slowly varying as both 1 and $\ln x$ are slowly varying.

Example 4: Let $\bar{F}_z(x) = x^{-z}\ln x^z$ for $z \in (1, 2)$, and let $\bar{F}(x) = \int_1^2 \bar{F}_z(x)dz$. Then $\bar{F}(x) = x^{-1} - 2x^{-2} + x^{-1}\frac{1}{\ln x} - x^{-2}\frac{1}{\ln x} = x^{-1}(1 - 2x^{-1} + \frac{1}{\ln x} - x^{-1}\frac{1}{\ln x}) = x^{-1}L(x)$, where $L(x) = 1 - 2x^{-1} + \frac{1}{\ln x} - x^{-1}\frac{1}{\ln x}$ is slowly varying.

Appendix D: Monte Carlo Cross Tail Estimation Experiments

Below we perform experiments similar to those in subsection 5.3, however in this case D_f is the entire set of data D. In this case cross tail estimation is the tail estimation homologue of Monte Carlo cross validation.

Gaussian Processes

In this experiment, our data is composed of an one-dimensional time series, which we reorganize in windows of size 50, and try to use each window to predict the next value. Our complete dataset D is made up of $n = 10000$ windows. On each run we randomly select 340 points of D for training (denote D_i), and then group the predictions of the model on the 10000 points of D into and array which we denote by $\hat{Y}_i$. We proceed to estimate the shape parameter of the tails of the prediction of the model, for given training set D_i by applying Pickands' estimator to $\hat{Y}_i$, receiving $\hat{\xi}_i$. We repeat this process 1000 times, and select as our estimation of the shape parameter of the tail of the distribution of our loss function, the maximum individual estimated parameter: $\hat{\xi}_i = \max\{\hat{\xi}_i | i \in [1000]\}$. On the other hand, we also calculate the MSE on the testing set $D \setminus D_i$ after the model has been trained on D_i. To check the difference of performance of the standard method of tail shape estimation and cross tail estimation, we also calculate the shape parameter of the overall distribution of prediction models, through the standard method, by applying Pickands estimator on $Y = \bigcup_{i=1}^{1000} \hat{Y}_i$. This experiment is repeated for length scale parameters from 1 to 30. We repeat this experiment 200 times. The results are shown in Figure 7.

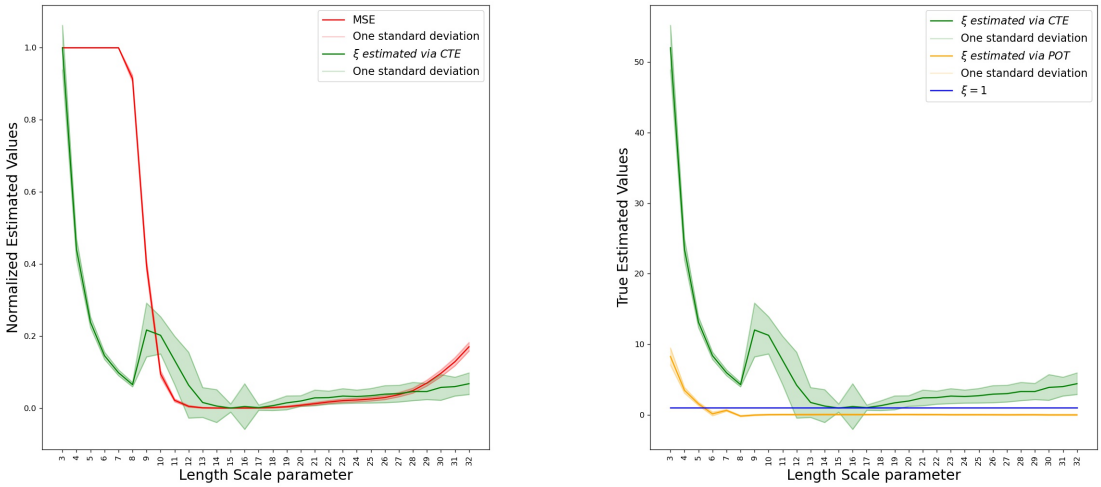

Figure 7: Experimental results in the case of testing Gaussian processes. Left: The correlation between MSE on test data, and the tail shape parameter of the distribution function estimated with cross tail estimation. Right: The tail shape parameter of the distribution function estimated with cross tail estimation and the standard approach. It is visible that when the standard method of estimating the tail is used, we lose this correlation, as the shape parameter estimation remains roughly constant for most models. Furthermore, if we use the criterion of existence of the first moment ($\xi < 1$), then we select the best model when cross tail estimation is used. On the other hand, if the standard method is applied almost all models models satisfy the criteria.

Polynomial Kernels

This experiment is almost identical to the previous one, with the only differences being that the models we test now are polynomial kernels, and the grid of possible candidate models instead of being made up of different length scale parameters from 1 to 30, in this case the hyperparameter that varies is the degree of the polynomial kernel. We test polynomial kernels of degree from 1 to 11. As before, we repeat this experiment 200 times. The results are shown in Figure 8.

Appendix E: Evolution of $\hat{\xi}^i_{POT}$ as a function of the sample size

In this section, we show the evolution of the convergence of $\hat{\xi}^i_{POT}$ to ξ^i_{max}, in the settings described in Section 5.1, as a function of the sample size m. See figure 9.

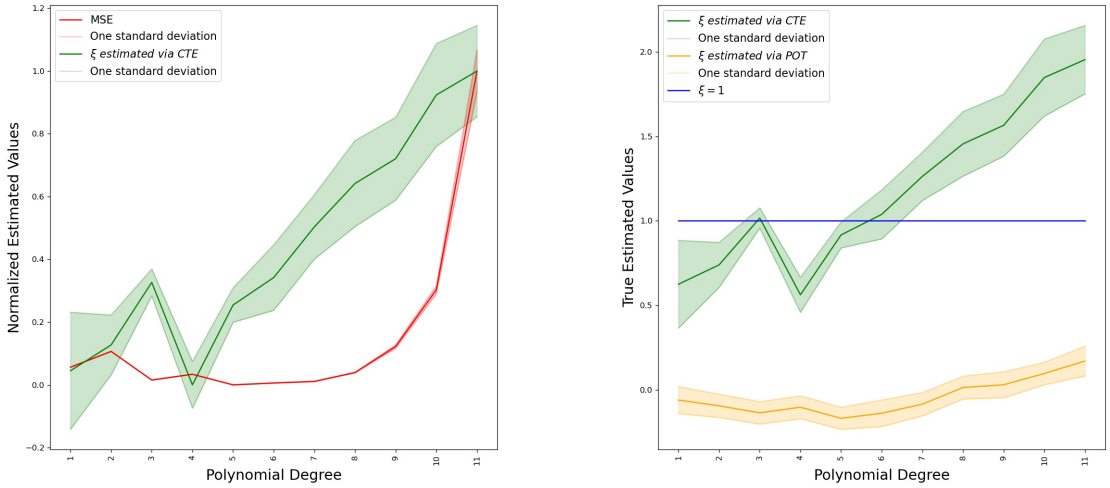

Figure 8: Experimental results in the case of testing polynomial kernels. Left: The correlation between MSE on test data, and the tail shape parameter of the distribution function estimated with cross tail estimation. Right: The tail shape parameter of the distribution function estimated with cross tail estimation and the standard approach. If we use the criterion of existence of the first moment ($\xi < 1$), then most non optimal models are eliminated. On the other hand, if the standard method is applied, all models satisfy the criteria.

Appendix F: Motivation from bias-variance decomposition

In Theorem 4.14, we showed that if we conditioned on the testing set, we could estimate the shape of the distribution of $W_{\boldsymbol{V}}(\boldsymbol{U})$ without labels in the testing set. This can also be motivated from the moments of $W_{\boldsymbol{V}}(\boldsymbol{U})$. Indeed, let us denote with f_{gt} the ground truth model which receives features as input and outputs labels. Let $A(\boldsymbol{X})$ be the bias $\mathbb{E}_{\boldsymbol{V}}(\hat{f}_{\boldsymbol{V}}(\boldsymbol{X})) - f_{gt}(\boldsymbol{X})$ and $B_{\boldsymbol{V}}(\boldsymbol{X}) = \mathbb{E}_{\boldsymbol{V}}(\hat{f}_{\boldsymbol{V}}(\boldsymbol{X})) - \hat{f}_{\boldsymbol{V}}(\boldsymbol{X})$ From the bias-variance decomposition theorem we notice that

$$W_{\boldsymbol{V}}(\boldsymbol{U})^2 = (A(\boldsymbol{X}) + B_{\boldsymbol{V}}(\boldsymbol{X}) + \epsilon)^2 \tag{80}$$

hence

$$\mathbb{E}_{\boldsymbol{V}}(W_{\boldsymbol{V}}(\boldsymbol{U}))^2 = A(\boldsymbol{X})^2 + E_{\boldsymbol{V}}(B_{\boldsymbol{V}}(\boldsymbol{X}))^2 \tag{81}$$

Similarly

$$\mathbb{E}[W^p_{\boldsymbol{V}}(\boldsymbol{U})|\boldsymbol{U} = \boldsymbol{u}] = E_{\{\boldsymbol{X}_T, \boldsymbol{Y}_T\}}|\boldsymbol{y} - \hat{f}_{\boldsymbol{X}_T, \boldsymbol{Y}_T}(\boldsymbol{x})|^p \tag{82}$$

$$= \sum_{k=0}^{p} \binom{p}{k} \sum_{i=0}^{k} \binom{k}{i} A(\boldsymbol{x})^i E_{\{\boldsymbol{X}_T, \boldsymbol{Y}_T\}}[B_{\boldsymbol{V}}(\boldsymbol{x})^{k-i}] E[\varepsilon^{p-k}] \tag{83}$$

We can see that at point $u = (\boldsymbol{x}, \boldsymbol{y})$, $W_{\boldsymbol{V}}(u)$ has all the moments if and only if $f_{\boldsymbol{V}}(\boldsymbol{x})$ has all its moments at $\boldsymbol{x}$. From this we can notice that the existence of $\mathbb{E}[W^p_{\boldsymbol{V}}(\boldsymbol{U})|\boldsymbol{U} = u]$ does not depend on the label $\boldsymbol{Y}$, as $u = (\boldsymbol{x}, \boldsymbol{y})$ is fixed. Hence, if we have conditions which guarantee that the existence of $\mathbb{E}[W^p_{\boldsymbol{V}}(\boldsymbol{U})|\boldsymbol{U} = u]$ for each u in the test set implies that existence of $\mathbb{E}[W^p_{\boldsymbol{V}}(\boldsymbol{U})]$, we would conclude that the existence of $\mathbb{E}[W^p_{\boldsymbol{V}}(\boldsymbol{U})]$ only depends

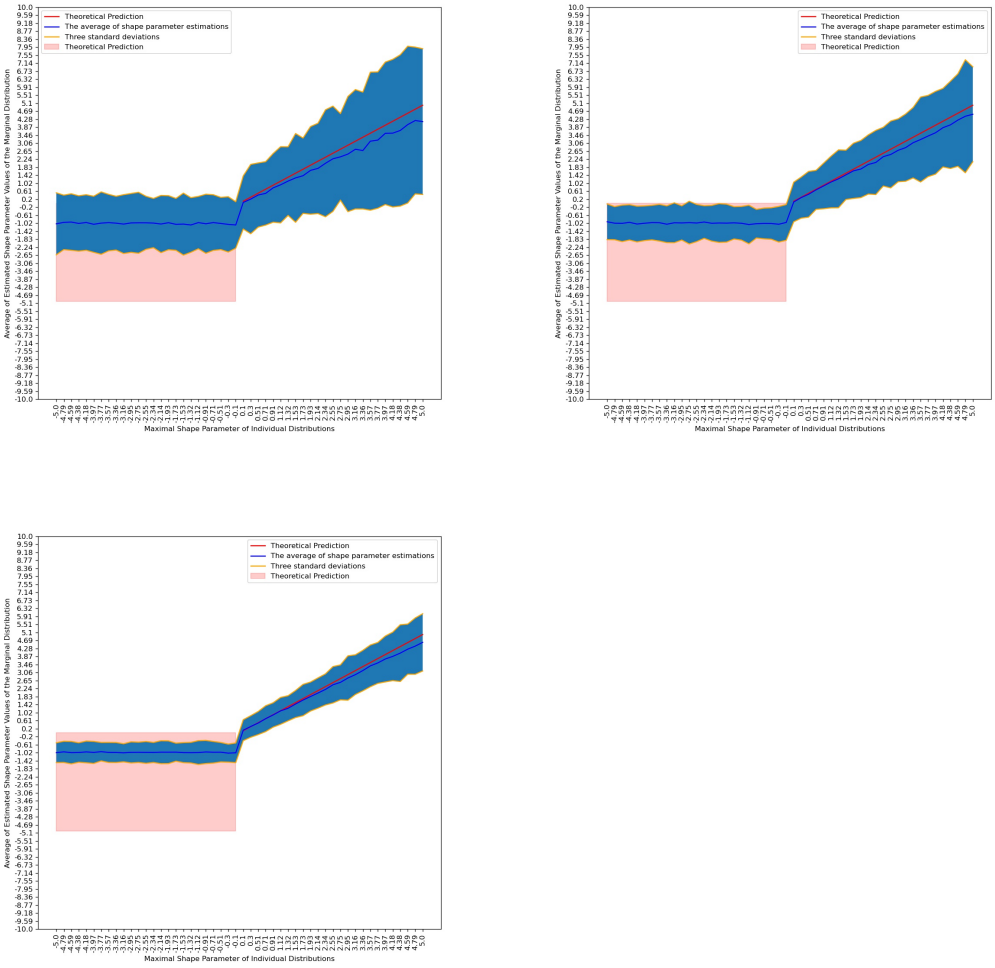

Figure 9: First row left: sample size 10^4. First row right: sample size 10^5. Second row: sample size 10^6.

on the existence of the moments of $f_{\boldsymbol{V}}(\boldsymbol{x})$.

References

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

Hirotogu Akaike. *Information Theory and an Extension of the Maximum Likelihood Principle*, pages 199–213. Springer New York, New York, NY, 1973.

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

K.P. Burnham and D.R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer New York, 2007. ISBN 9780387224565.

W.G. Cochran. *Sampling Techniques, 3Rd Edition*. A Wiley publication in applied statistics. Wiley India Pvt. Limited, 2007. ISBN 9788126515240.

L. de Haan and A. Ferreira. *Extreme Value Theory: An Introduction*. Springer Series in Operations Research and Financial Engineering. Springer New York, 2007. ISBN 9780387344713.

P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events: for Insurance and Finance*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, 2013. ISBN 9783540609315.

R. A. Fisher and L. H. C. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, 24 (2):180, January 1928.

J. Galambos and E. Seneta. Regularly varying sequences. *Proceedings of the American Mathematical Society*, 41(1):110–116, 1973. ISSN 00029939, 10886826.

B. Gnedenko. Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics*, 44(3):423–453, 1943. ISSN 0003486X.

Clifford M. Hurvich and Chih-Ling Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307, 06 1989. ISSN 0006-3444.

T. Mikosch, Operations Research EURANDOM European Institute for Statistics, Probability, and their Applications. *Regular Variation, Subexponentiality and Their Applications in Probability Theory*. EURANDOM report. Eindhoven University of Technology, 1999.

Jerzy Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625, 1934. ISSN 09528385.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

James Pickands. Statistical Inference Using Extreme Order Statistics. *The Annals of Statistics*, 3(1):119 – 131, 1975.

Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461 – 464, 1978.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2828–2837, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016.

Nariaki Sugiura. Further analysts of the data by Akaike' s information criterion and the finite corrections. *Communications in Statistics - Theory and Methods*, 7(1):13–26, 1978.

Renjie Wu and Eamonn J. Keogh. Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. *CoRR*, abs/2009.13807, 2020.