Journal of Machine Learning Research 25 (2024) 1-47 Submitted 7/22; Revised 7/23; Published 1/24

On Tail Decay Rate Estimation of Loss Function Distributions

Etrit Haxholli ETRIT.HAXHOLLI@INRIA.FR
Epione Research Group
Inria, Univesity Cte d'Azur
2004 Rte des Lucioles, 06902 Valbonne, France

Marco Lorenzi MARCO.LORENZI@INRIA.FR
Epione Research Group
Inria, Univesity Cte d'Azur
2004 Rte des Lucioles, 06902 Valbonne, France

Editor: Lorenzo Rosasco

Abstract

The study of loss-function distributions is critical to characterize a model's behaviour on a given machine-learning problem. While model quality is commonly measured by the average loss assessed on a testing set, this quantity does not ascertain the existence of the mean of the loss distribution. Conversely, the existence of a distribution's statistical moments can be verified by examining the thickness of its tails.

Cross-validation schemes determine a family of testing loss distributions conditioned on the training sets. By marginalizing across training sets, we can recover the overall (marginal) loss distribution, whose tail-shape we aim to estimate. Small sample-sizes diminish the reliability and efficiency of classical tail-estimation methods like Peaks-Over-Threshold, and we demonstrate that this effect is notably significant when estimating tails of marginal distributions composed of conditional distributions with substantial tail-location variability. We mitigate this problem by utilizing a result we prove: under certain conditions, the marginal-distribution's tail-shape parameter is the maximum tail-shape parameter across the conditional distributions underlying the marginal. We label the resulting approach as 'cross-tail estimation (CTE)'.

We test CTE 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.
Keywords: Extreme Value Theory, Tail Modelling, Peaks-Over-Threshold, Cross-Tail-Estimation, Model Ranking

1. Introduction

Loss function distributions form critical subjects of analysis, serving as barometers for machine learning model performance. In the context of a particular model and associated machine learning task, the true distribution of the loss function is typically elusive; we predominantly have access to a finite sample set, born from diverse choices of training and testing sets. To facilitate performance comparisons across different models based on the underlying loss function distributions, a spectrum of methodologies has been established.

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

Traditional strategies derive from information criteria such as the Akaike Information Criterion (AIC) (Akaike, 1973, 1974), an asymptotic approximation of the Kullback-Leibler divergence between the true data distribution and the fitting candidate, and its corrected version (AICc) (Sugiura, 1978; Hurvich and Tsai, 1989), in addition to the Bayesian Information Criterion (BIC) (Schwarz, 1978). The application of these information criteria, especially the AIC, is often constrained by the multiple inherent approximations and assumptions (Burnham and Anderson, 2007), making them less feasible in certain scenarios. However, it warrants mention that more recent penalized criteria have considerably expanded their suitability for realistic setups (Birge and Massart, 1995; Arlot and Massart, 2009). Simultaneously, other methodologies, termed splitting/resampling methods, have been devised, wherein a subset of the data is deployed to assess the performance of the trained model. This group of methodologies is expansive, predicated on a diverse range of partitioning and evaluation strategies addressing data heterogeneity and imbalance (Neyman, 1934; Cochran, 2007).

In the domain of cross-validation strategies (Allen, 1974; Stone, 1976, 1977), the common metric employed for gauging model performance is the sample mean of the loss function distribution. This practice, which invariably provides a finite numerical value, does not assure the existence of the first or higher order statistical moments. Moreover, this metric, in spite of its prevalence, should not necessarily be construed as a sole indicator of the model's performance, as it does not necessarily quantify its robustness to the underlying data distribution and model architecture. Furthermore, while it is true that aforementioned methods allow to rank models according to their relative performance on a given data set, these scores still have limited value in quantifying the overall stability of a model. From a theoretical perspective, there is a connection between the uppermost existing moment of a distribution and the thickness of its tail, which underscores the significance of examining the behavioural traits and decay rate of the tails of loss function distributions and their relation to the stability of the model.

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 PickandsBalkemaDe 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 data sets. 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 thickest 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. The proposed solution enables a reduction in the sample size requirements, in the experiments we conducted. In the context of model comparison, our theory establishes that, under some assumptions, we can estimate the shape of the total loss distribution, by simply investigating the models prediction, without the need for target data. 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 overfitting. 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

This section initially provides a succinct overview of Monte Carlo cross validation (Kuhn and Johnson, 2013), given its conceptual similarities with the proposed method, "cross tail estimation". The subsequent subsection outlines standard results and definitions from extreme value analysis, forming the bedrock of the statements presented in Section 4.

2.1 Monte Carlo Cross Validation

Let $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, 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 replacement from the original data set 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:

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

where m is the number of iterations (data partitions).
A detailed discussion on cross validation, elucidating its similarities with our proposed method for tail estimation in marginal loss function distributions, namely 'cross tail estimation', is presented in Subsection 3.2.

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 FisherTippettGnedenko 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 (FisherTippettGnedenko) *: Let X be a real random variable with distribution F_X. Denote by $\{X_1, X_2, \ldots, X_n\}$ a set of iid samples from the distribution F_X, and define $M_n = \max\{X_1, \ldots, X_n\}$. 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 $F(x) = G_{\xi,a,b}(x)$, for some $b, \xi \in \mathbb{R}, a > 0$.

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.

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

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

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

(Balkema and de Haan, 1974) and (Pickands, 1975) proved that the limiting distribution of samples larger than a threshold is a Generalized Pareto distribution, whose location parameter is zero.

Theorem 4 (PickandsBalkemaDe 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 different estimators such as the Pickands Estimator or the Deckers-Einmahl-de Haan Estimator (DEdH), (Dekkers et al., 1989).

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

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

Definition 6 *Let $X_1, X_2, \ldots, X_n$ be iid samples from the distribution F_X. If we denote with $X_{1,n}, X_{2,n}, \ldots, X_{n,n}$ the samples sorted in descending order, then the DEdH estimator is defined as follows:*

$$\hat{\xi}_{k,n}^{(H)} = 1 + H_{k,n}^{(1)} + \frac{1}{2}\left(\frac{(H_{k,n}^{(1)})^2}{H_{k,n}^{(2)}} - 1\right)^{-1}, \tag{9}$$

where

$$H_{k,n}^{(1)} = \frac{1}{k}\sum_{j=1}^{k}(\ln X_{j,n} - \ln X_{k+1,n}) \tag{10}$$

and

$$H_{k,n}^{(2)} = \frac{1}{k}\sum_{j=1}^{k}(\ln X_{j,n} - \ln X_{k+1,n})^2. \tag{11}$$

An important result which we are going to use frequently in our proofs is Theorem 10, which can be found in (Embrechts et al., 2013; de Haan and Ferreira, 2007), and gives the connection between the maximum domain of attraction and slowly varying functions.

Definition 7 *A positive measurable function L is called slowly varying if it is defined in some neighborhood of infinity and if:*

$$\lim_{x\to\infty}\frac{L(ax)}{L(x)} = 1, \textit{for all } a > 0. \tag{12}$$

Theorem 8 (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:*

$$L(x) = e^{c(x)}e^{\int_{x_0}^{x}\frac{u(t)}{t}dt}, \tag{13}$$

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

Proposition 9 *(Mikosch et al., 1999) If L is slowly varying then for every $\epsilon > 0$:*

$$\lim_{x\to\infty} x^{-\epsilon}L(x) = 0. \tag{14}$$

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

Theorem 10 *: 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,*
- $\xi = 0 \iff \bar{F}_X(x) = c(x)e^{-\int_w^x \frac{1}{a(t)}dt}$, $w < x < x_F \leq \infty$, *where c is a measurable function satisfying $c(x) \to c > 0$ as $x \uparrow x_F$, and $a(x)$ is a positive, absolutely continuous function (with respect to the Lebesgue measure) with density $a'(x)$ having $\lim_{x\uparrow x_F} a'(x) = 0$. If $x_F < \infty$ then $\lim_{x\uparrow x_F} a(x) = 0$ as well.*

3. Setup and Problem Statement

In the initial subsection, we establish a formal framework to address the problem of tail modelling for loss distributions, and elucidate how unsatisfactory results can arise from a naive application of the Peaks-Over-Threshold (POT) method. Subsequently, in the second subsection, we present Cross-Tail-Estimation (CTE), a novel methodology that addresses these shortcomings. A salient feature of this section is the illustration of the analogy between CTE and Cross-Validation, providing an intuitive understanding of CTE. In the concluding subsection, we lay the groundwork for the upcoming Section 4. In this forthcoming section, we provide the theoretical justification, in the form of Theorem 21, for the application of our introduced method, CTE.

3.1 Problem Statement

We assume that each data sample $(\boldsymbol{X}, \boldsymbol{Y})$ comes from distribution $\mathcal{D}$ and that the sampling is independent. We use the symbol $\boldsymbol{X}$ to denote the features and the symbol $\boldsymbol{Y}$ to denote the labels (targets). The training set will be defined as a random vector comprised of iid random vectors $(\boldsymbol{X}, \boldsymbol{Y})$ sampled from $\mathcal{D}$. More precisely, after fixing a natural number k, we define a training set as $\boldsymbol{V} = [(\boldsymbol{X}, \boldsymbol{Y})_1, (\boldsymbol{X}, \boldsymbol{Y})_2, \ldots, (\boldsymbol{X}, \boldsymbol{Y})_k]$, where each $(\boldsymbol{X}, \boldsymbol{Y})_i$ has distribution $\mathcal{D}$. On the other hand, a test point $\boldsymbol{U}$ naturally is defined as a sample from $\mathcal{D}$, i.e., $\boldsymbol{U} = (\boldsymbol{X}, \boldsymbol{Y})$. In practice, the realisation of $\boldsymbol{U}$ should not be an entry in $\boldsymbol{V}$.
A model which is trained on $\boldsymbol{V}$ to predict $\boldsymbol{Y}$ from $\boldsymbol{X}$ is denoted as $\hat{\boldsymbol{h}}_{\boldsymbol{V}}(\boldsymbol{X})$. The prediction error on the testing datum $\boldsymbol{U}$ of a model trained on $\boldsymbol{V}$ is denoted as $W_{\boldsymbol{V}}(\boldsymbol{U})$. For the remainder of the paper we assume that $W_{\boldsymbol{V}}(\boldsymbol{U}) > 0$ and notice that the probability density function of $W_{\boldsymbol{V}}(\boldsymbol{U})$ is

$$f_W(w) = \int f_{W,\boldsymbol{V}}(w, \boldsymbol{v}) d\boldsymbol{v} = \int f_{\boldsymbol{V}}(\boldsymbol{v}) f(w|\boldsymbol{V} = \boldsymbol{v}) d\boldsymbol{v} = \int f_{\boldsymbol{V}}(\boldsymbol{v}) f_{\boldsymbol{v}}(w) d\boldsymbol{v}, \tag{15}$$

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

$$F_W(w) = \int f_{\boldsymbol{V}}(\boldsymbol{v}) F_{\boldsymbol{v}}(w) d\boldsymbol{v}. \tag{16}$$

$F_{\boldsymbol{v}}(w)$ is the distribution of the prediction error (loss) of the model trained on training set $\boldsymbol{v}$, while $F_W(w)$ is the unconditional distribution of the loss.

Standard methods such as the Peaks-Over-Threshold (POT) approach, when employed directly for estimating the tails of general marginal distributions like $F_W(w)$ in Equation (16), may yield unsatisfactory outcomes.

To provide insight into the problem, let's simplify the scenario for a moment, by assuming that some random vector $\boldsymbol{V}$ can be either $\boldsymbol{v}_1$ or $\boldsymbol{v}_2$, each with an equal likelihood. In the situation where $\boldsymbol{V} = \boldsymbol{v}_1$, assume $F_{\boldsymbol{v}_1}(w)$ corresponds to a thick-tailed distribution, wherein even the first moment does not exist. Conversely, if $\boldsymbol{V} = \boldsymbol{v}_2$, suppose $F_{\boldsymbol{v}_2}(w)$ takes the form of a Gaussian distribution, characterized by a large mean. Under these conditions, Equation (16) simplifies to $F_W(w) = \frac{1}{2} F_{\boldsymbol{v}_1}(w) + \frac{1}{2} F_{\boldsymbol{v}_2}(w)$. It is known that the tail shape parameter of $F_W(w) = \sum_{i=1}^{n} p(\boldsymbol{v}_i) F_{\boldsymbol{v}_i}(w)$ is determined by the conditional distribution $F_{\boldsymbol{v}_i}(w)$ with

the thickest tail. In our case, n above is 2, and the tail of $F_W(w)$ is defined by the fat tail of $F_{\boldsymbol{v}_1}(w)$. Suppose we proceed with the standard POT approach, that is, we integrate out the random variable $\boldsymbol{V}$, and subsequently estimate the shape parameter of the tail of $F_W(w)$. In practical scenarios, this translates to merging the samples from both conditional distributions into a singular array. Given the finite nature of sample sizes in such cases, it's conceivable that none of the samples of W from the thick-tailed distributions surpass those from the Gaussian distribution, owing to the discrepancies in their locations. As a consequence, the sample tail of the marginal (mixture) distribution takes its shape from the sample tail of the Gaussian $F_{\boldsymbol{v}_2}(w)$, while in reality, the tail of $F_W(w)$ is dictated by the heavy tail of $F_{\boldsymbol{v}_1}(w)$. In the ideal scenario with limitless sampling, we would expect to determine the true tail shape. Yet, within the constraints of practical applications, it may be necessary to estimate the tail shape parameters of $F_{\boldsymbol{v}_1}(w)$ and $F_{\boldsymbol{v}_2}(w)$ individually.

A natural question that arises in the case that $\boldsymbol{V}$ has a continuous distribution as in Equation (16) is whether the tail of the marginal $F_W(w)$ is still determined by the largest tail of the conditional distributions $F_{\boldsymbol{v}}(w)$. As we will prove in Section 4, under some regularity conditions, the answer is in the affirmative.

3.2 Cross Tail Estimation

We will denote with $\xi_{\boldsymbol{v}}$ the tail shape parameter of $F_{\boldsymbol{v}}(w)$ and with ξ the shape tail parameter of $F_W(w)$. Our goal in Section 4 is to prove that under some regularity conditions if $\exists \boldsymbol{v}$, such that $\xi_{\boldsymbol{v}} > 0$, then $\xi = \max\{\xi_{\boldsymbol{v}} | \boldsymbol{v}\}$, and if $\forall \xi_{\boldsymbol{v}} \leq 0$, then we have $\xi \leq 0$. This motivates

Algorithm 1 Naive Cross Tail Estimation

Input: Data $D = [(\boldsymbol{x},\boldsymbol{y})_1, (\boldsymbol{x},\boldsymbol{y})_2, \ldots, (\boldsymbol{x},\boldsymbol{y})_n]$; the Pickands or DEdH estimator
Define: $A = \{\}$
Fix the number of training sets (rounds): $m \in \mathbb{N}$
repeat
1. sample $(\boldsymbol{x},\boldsymbol{y})_{\pi(1)}, \ldots, (\boldsymbol{x},\boldsymbol{y})_{\pi(k)}$ from $(\boldsymbol{x},\boldsymbol{y})_1, (\boldsymbol{x},\boldsymbol{y})_2, \ldots, (\boldsymbol{x},\boldsymbol{y})_n$
2. train model $\hat{\boldsymbol{h}}_{\boldsymbol{v}}$ on $\boldsymbol{v} = [(\boldsymbol{x},\boldsymbol{y})_{\pi(1)}, \ldots, (\boldsymbol{x},\boldsymbol{y})_{\pi(k)}]$
3. calculate the prediction errors $W_{\boldsymbol{v}}(\boldsymbol{U})$ of model $\hat{\boldsymbol{h}}_{\boldsymbol{v}}$ on the testing set $D \setminus \boldsymbol{v}$
4. group the calculated prediction errors in the set $E_{\boldsymbol{v}}(D)$
5. apply the Pickands or DEdH estimator on $E_{\boldsymbol{v}}(D)$ to estimate $\xi_{\boldsymbol{v}}$
6. add $\hat{\xi}_{\boldsymbol{v}}$ to A

until $|A| = m$
return $\max A$ if $\max A > 0$, else return 'non-positive'

Algorithm 1 which we name "Naive Cross Tail Estimation" (NCTE). Since for each $\boldsymbol{v}$, the estimated $\hat{\xi}_{\boldsymbol{v}}$ is prone to estimation errors, taking the maximum $\hat{\xi}_{\boldsymbol{v}}$ over all $\boldsymbol{v}$ tends to cause NCTE to overestimate the true ξ, especially when the number of conditional distributions $F_{\boldsymbol{v}}(w)$ is large. For this reason we propose Algorithm 2, named 'Cross Tail Estimation' (CTE), where we split the samples from $F_{\boldsymbol{v}}(w)$ into p sets in order to get p estimates of the tail shape parameter of $F_{\boldsymbol{v}}(w)$, that is $\{\hat{\xi}^1_{\boldsymbol{v}}, \hat{\xi}^2_{\boldsymbol{v}}, ..., \hat{\xi}^p_{\boldsymbol{v}}\}$. Our final estimation of $\xi_{\boldsymbol{v}}$ is the average of the p estimations, i.e., $\frac{1}{p}\sum_{i=0}^{p} \hat{\xi}^i_{\boldsymbol{v}}$. A more detailed justification for utilizing

Algorithm 2 is given in Appendix E. We notice that Algorithm 2 is identical to Algorithm 1 when $p = 1$.

Algorithm 2 Cross Tail Estimation

Input: Data $D = [(\boldsymbol{x}, \boldsymbol{y})_1, (\boldsymbol{x}, \boldsymbol{y})_2, ..., (\boldsymbol{x}, \boldsymbol{y})_n]$; the Pickands or DEdH estimator
Define: $A = \{\}$
Fix the number of training sets (rounds): $m \in \mathbb{N}$
repeat
1. sample $(\boldsymbol{x}, \boldsymbol{y})_{\pi(1)}$, ..., $(\boldsymbol{x}, \boldsymbol{y})_{\pi(k)}$ from $(\boldsymbol{x}, \boldsymbol{y})_1, (\boldsymbol{x}, \boldsymbol{y})_2, ..., (\boldsymbol{x}, \boldsymbol{y})_n$
2. train model $\hat{\boldsymbol{h}}_{\boldsymbol{v}}$ on $\boldsymbol{v} = [(\boldsymbol{x}, \boldsymbol{y})_{\pi(1)}, ..., (\boldsymbol{x}, \boldsymbol{y})_{\pi(k)}]$
3. calculate the prediction errors $W_{\boldsymbol{v}}(\boldsymbol{U})$ of model $\hat{\boldsymbol{h}}_{\boldsymbol{v}}$ on the testing set $D \setminus \boldsymbol{v}$
4. group the calculated prediction errors in the set $E_{\boldsymbol{v}}(D)$
5. split $E_{\boldsymbol{v}}(D)$ into $\{E^1_{\boldsymbol{v}}(D), ..., E^p_{\boldsymbol{v}}(D)\}$
6. apply the Pickands or DEdH estimator on each $E^i_{\boldsymbol{v}}(D)$ to get an estimate $\hat{\xi}^i_{\boldsymbol{v}}$ of $\xi_{\boldsymbol{v}}$
7. average over p to get the final estimate $\hat{\xi}_{\boldsymbol{v}} = \frac{1}{p}\sum_{i=0}^{p} \hat{\xi}^i_{\boldsymbol{v}}$ of $\xi_{\boldsymbol{v}}$
8. add $\hat{\xi}_{\boldsymbol{v}}$ to A

until $|A| = m$
return $\max A$ if $\max A > 0$, else return 'non-positive'

Remark: Estimating particular statistics of $F_W(w)$ through the statistics of $F_{\boldsymbol{v}}(w)$ as in in Algorithm 1 and 2 is a key component of Cross Validation. During Cross Validation, a training set $\boldsymbol{v}$ and a testing set $D \setminus \boldsymbol{v}$ are selected in each iteration, during which the following conditional expectation is then estimated:

$$\mathbb{E}[W_{\boldsymbol{V}}(\boldsymbol{U})|\boldsymbol{V} = \boldsymbol{v}] = \int w f_{\boldsymbol{v}}(w) dw. \tag{17}$$

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:

$$\begin{aligned} \mathbb{E}_{\boldsymbol{U},\boldsymbol{V}}(W_{\boldsymbol{V}}(\boldsymbol{U})) = \int w f(w) dw = \int f_{\boldsymbol{V}}(\boldsymbol{v}) \int w f_{\boldsymbol{v}}(w) dw d\boldsymbol{v} = \\ = \int f_{\boldsymbol{V}}(\boldsymbol{v}) \mathbb{E}[W_{\boldsymbol{V}}(U)|\boldsymbol{V} = \boldsymbol{v}] d\boldsymbol{v} = \mathbb{E}[\mathbb{E}[W_{\boldsymbol{V}}(\boldsymbol{U})|\boldsymbol{V} = \boldsymbol{v}]]. \end{aligned} \tag{18}$$

In the language of Section 3.1, the mean of distribution $F_W(w)$ is the average of the means of the conditional distributions $F_{\boldsymbol{v}}(w)$.
This statement about sums stands parallel with our claim about extremes: 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 $F_{\boldsymbol{v}}(w)$.

3.3 The General Problem

Generalizing the problem stated in Section 3.1 requires considering a one dimensional random variable of interest X, dependent on other random variables $\{Z_1, Z_2, ..., Z_n\}$, such that the probability density function of X is

$$f_X(x) = \int f(z_1, ..., z_n, x) d_{z_1} \cdots d_{z_n} \tag{19}$$

$$= \int f(\boldsymbol{z}) f(x|\boldsymbol{z}) d\boldsymbol{z} = \int f(\boldsymbol{z}) f_{\boldsymbol{z}}(x) d\boldsymbol{z}. \tag{20}$$

Integrating with respect to x we get

$$F_X(x) = \int f(\boldsymbol{z}) F(x|\boldsymbol{z}) d\boldsymbol{z} = \int f(\boldsymbol{z}) F_{\boldsymbol{z}}(x) d\boldsymbol{z}. \tag{21}$$

In this case, with regards to the previous section, we notice that $\boldsymbol{Z} = \boldsymbol{V}$ is the training set on which we condition, while $X = W$ is the random variable of interest. In Section 4, we give several results which relate the tails of $F_X(x)$ and $F(x|\boldsymbol{z})$, culminating with Theorem 21 which justifies the usage of the CTE algorithm, by providing limiting behaviour guarantees.

4. Theoretical Results

In this section, we build our theory of modelling the tails of marginal distributions, which culminates with Theorem 21. 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 11 *If $F_1 \in MDA(\xi_1)$, $F_2 \in MDA(\xi_2)$, and $\xi_1 > \xi_2 > 0$, 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 12 *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 $\lim_{x\to\infty} \frac{\bar{F}_2(x)}{\bar{F}_1(x)} = 0$.*

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

Lemma 13 *If for $i \in \{1, \ldots, n\}$ we let $L_i(x)$ be slowly varying functions, and $\{a_1, \ldots, a_n\}$ be a set of positive real numbers, then*

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

is slowly varying.

In the case of a mixture of a finite number of distributions the following known result holds:

Theorem 14 *Let $Z : \Omega \to A \subset \mathbb{R}^n$ be a random vector where $|A| < \infty$. At each point $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n \in A$, we define a distribution $F_{\boldsymbol{z}_i}(x) \in MDA(\xi_i)$ and assume that $\xi_{\max} := \max(\xi_1 = \xi_{z_1}, \ldots, \xi_n = \xi_{z_n}) > 0$. If the set $\{p_1, \ldots, p_n\}$ is a set of convex combination parameters, that is $\sum_i p_i = 1$ and $p_i > 0$ then:*

$$F(x) = \sum_i^n p_i F_{\boldsymbol{z}_i}(x) \in MDA(\xi_{\max}). \tag{22}$$

If $\xi_{\max} \leq 0$ then if ξ_F exists we have $\xi_F \leq 0$.

Proof While this result is well known, we give an alternative proof in Appendix A, using the Pickands-Balkema-De Haan Theorem. ■

From now on, we assume that the functions $F_A(x) = \int_A f_{\boldsymbol{Z}}(\boldsymbol{z}) F_{\boldsymbol{z}}(x) d\boldsymbol{z}$ defined on any element A of the Borel $\sigma-$algebra induced by the usual metric are in the MDA of some extreme value distribution. Furthermore, we assume that the pdf $f_{\boldsymbol{Z}}(\boldsymbol{z})$ is strictly positive everywhere in its domain.

Proposition 9 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 $\{L_{\boldsymbol{z}} | \boldsymbol{z} \in A\}$, where A is a measurable set, the set $\{x(L_{\boldsymbol{z}}, \delta, \gamma) | \boldsymbol{z} \in A\}$ can be unbounded, suggesting that the beginning of the tail of $\bar{F}_{\boldsymbol{z}}(x) = x^{-\frac{1}{\xi_{\boldsymbol{z}}}} L_{\boldsymbol{z}}(x)$ can be postponed indefinitely across the family $\{F_{\boldsymbol{z}} | \boldsymbol{z} \in A\}$. These concepts are formalized in the following:

Definition 15 *For a set A, the family of sub-polynomial functions $\{L_{\boldsymbol{z}}(x) | \boldsymbol{z} \in A\}$ is called γ-uniformly sub-polynomial if for any fixed $\delta > 0$, there exists a $\gamma(\delta)$ so that the set $\{x_0 | \boldsymbol{z} \in A\}$ is bounded from above, where $x_0 = x_0(L_{\boldsymbol{z}}, \delta, \gamma)$ is the smallest value for which when $x > x_0$ we have $x^{-\delta} L_{\boldsymbol{z}}(x) < \gamma$.*

Proposition 16 *Let $\boldsymbol{Z} : \Omega \to A \subset \mathbb{R}^n$ be a random vector where A is measurable and define a family of sub-polynomial functions $\{L_{\boldsymbol{z}}(x) | \boldsymbol{z} \in A\}$, which we assume is γ-uniformly sub-polynomial. Then for a probability density function $f_{\boldsymbol{Z}}(\boldsymbol{z})$ on A induced by $\boldsymbol{Z}$, the function $L(x) = \int_A f_{\boldsymbol{Z}}(\boldsymbol{z}) L_{\boldsymbol{z}}(x) d\boldsymbol{z}$ 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 17 *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 $F_{\boldsymbol{z}}(x) \in MDA(\xi_{\boldsymbol{z}})$, and suppose there exist ξ_{lo}, ξ_{up} such that $\forall \boldsymbol{z} \in A,\ 0 < \xi_{lo} \leq \xi_{\boldsymbol{z}} \leq \xi_{up}$. Furthermore, let $L_{\boldsymbol{z}}(x)$ be the slowly varying function corresponding to $F_{\boldsymbol{z}}(x)$. If the family $\{L_{\boldsymbol{z}}(x)|\boldsymbol{z} \in A\}$ is γ-uniformly sub-polynomial, then for $F(x) = \int_A f_{\boldsymbol{Z}}(\boldsymbol{z}) F_{\boldsymbol{z}}(x) d\boldsymbol{z}$ we have $\xi_{lo} \leq \xi_F \leq \xi_{up}$. Furthermore, if $\xi_{\boldsymbol{z}}$ is continuous in $\boldsymbol{z}$, then $\xi_F = \xi_{\max}$, where $\xi_{\max} := \sup\{\xi_{\boldsymbol{z}}|\boldsymbol{z} \in A\}$.*

Similarly to the case when $F_{\boldsymbol{z}}(x)$ 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. We notice that if $F_z(x) \in MDA(\xi)$ for $\xi \leq 0$, then $\bar{F}_{\boldsymbol{z}}(x)$ itself is sub-polynomial, whether its support is bounded or not. This observation motivates the following:

Definition 18 *For a set A, define the family of distribution functions $\mathcal{F}_A = \{F_{\boldsymbol{z}}(x)|\boldsymbol{z} \in A\}$, and define $A^+ = \{\boldsymbol{z}|\xi_{\boldsymbol{z}} > 0\},\ A^- = \{\boldsymbol{z}|\xi_{\boldsymbol{z}} \leq 0\}$. We say family $\mathcal{F}_A$ has stable cross-tail variability if,*

- *$\{L_{\boldsymbol{z}}(x)|\boldsymbol{z} \in A^+\}$ is γ-uniformly sub-polynomial,*
- *$\{\bar{F}_{\boldsymbol{z}}(x)|\boldsymbol{z} \in A^-\}$ is γ-uniformly sub-polynomial.*

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

Lemma 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 $F_{\boldsymbol{z}}(x) \in MDA(\xi_{\boldsymbol{z}})$, 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 $F(x) = \int_A f_{\boldsymbol{Z}}(\boldsymbol{z}) F_{\boldsymbol{z}}(x) d\boldsymbol{z}$ we have $\xi_F \leq \epsilon$.*

Corollary 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 $F_{\boldsymbol{z}}(x) \in MDA(\xi_{\boldsymbol{z}})$, 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 $F(x) = \int_A f_{\boldsymbol{Z}}(\boldsymbol{z}) F_{\boldsymbol{z}}(x) d\boldsymbol{z}$ we have $\xi_F \leq 0$.*

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 $\xi_F \leq \epsilon, \forall \epsilon > 0$. ■

Finally, we prove the generalization of Theorem 17 in the case that the tail shape parameters $\xi_{\boldsymbol{Z}}$ of the conditional distributions are real numbers:

Theorem 21 *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 $F_{\boldsymbol{z}}(x) \in MDA(\xi_{\boldsymbol{z}})$, 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 $F(x) = \int_A f_{\boldsymbol{Z}}(\boldsymbol{z}) F_{\boldsymbol{z}}(x) d\boldsymbol{z}$ we have $\xi_F = \xi_{\max}$. In the case that $\xi_{\max} \leq 0$ then $\xi_F \leq 0$.*

Examples when the conditions of Theorem 21 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 22 *Let F_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 F_X, that is:*

$$F_{X_1}(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \frac{F(x)-F(0)}{1-F(0)} & \text{for } x > 0 \end{cases}. \tag{23}$$

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

$$F_{X_2}(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{F(0)-F(-x)}{F(0)} & \text{for } x \geq 0 \end{cases}. \tag{24}$$

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

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

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

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

Proof Since

$$\begin{aligned} F_{|X|}(x) = \mathbb{P}(|X| < x) &= \mathbb{P}(X < x | X > 0)\mathbb{P}(X > 0) + \mathbb{P}(-X < x | X \leq 0)\mathbb{P}(X \leq 0) \\ &= p_1 F_{X_1}(x) + p_2 F_{X_2}(x), \end{aligned} \tag{25}$$

Theorem 14 gives the desired conclusion. ■

Proposition 23 *Let X be a random variable such that $X \in MDA(\xi_X > 0)$. 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 $\xi_X \leq 0$ then $\xi_Y \leq 0$.*

It is important to notice that we can estimate the shape of the tail of $W_{\boldsymbol{V}}(\boldsymbol{U})$ by also conditioning on the test label $\boldsymbol{y}$:

$$f_W(w) = \int f_{W,\boldsymbol{Y}}(w, \boldsymbol{y}) d\boldsymbol{y} = \int f_{\boldsymbol{Y}}(\boldsymbol{y}) f(w|\boldsymbol{Y} = \boldsymbol{y}) d\boldsymbol{y} = \int f_{\boldsymbol{Y}}(\boldsymbol{y}) f_{\boldsymbol{y}}(w) d\boldsymbol{y} \tag{26}$$

$$F_W(w) = \int f_{\boldsymbol{Y}}(\boldsymbol{y}) F_{\boldsymbol{y}}(w) d\boldsymbol{y}. \tag{27}$$

We use this fact to prove the following:

Proposition 24 *Let the loss function be defined as* $W_{\boldsymbol{V}}(\boldsymbol{U}) = |Y - \hat{f}_{\boldsymbol{V}}(\boldsymbol{X})|^p$ *for some* $p \in \mathbb{R}^+$*, and let* $F_y(t)$ *be the distribution of* $\hat{f}_{\boldsymbol{V}}(\boldsymbol{X})$ *given* Y*. If we assume that the distribution of the labels* Y *has bounded support* S*, that the family* $\{F_y(t)|y \in S\}$ *has stable cross-tail variability, and that the shape parameters* ξ_y *of* $F_y(t)$ *change continuously, then the tail shape parameters of* $W_{\boldsymbol{V}}(\boldsymbol{U})$ *and* $|\hat{f}_{\boldsymbol{V}}(\boldsymbol{X})|^p$ *share the same sign, and are identical if either of them is positive.*

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 25 *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, for a model with a positive loss function whose distribution has a shape parameter that is bigger than one, 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. However, if every model has an infinite expected loss, it's not advisable to eliminate them all. An alternative approach could be to utilize the tail thickness and medians of the loss function distributions to guide decision-making about which models to keep.

In Proposition 24, we showed that if we condition on the testing set, under some assumptions, 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 D.

5. Experiments

In this section, we demonstrate the significance of Theorem 21. In the first subsection, we show experimental evidence that the estimated shape parameter of the marginal distribution, coincides with the maximal shape parameter of individual conditional distributions. In the second subsection, we show that when the sample size is finite, as it is the case in practice, the method proposed by Theorem 21 (cross tail estimation) can be necessary to reduce the required sample size 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, the magnitude of shape parameters of the distribution of model predictions increases significantly when the model overfits. We also notice that such a relationship does not hold 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 $F(x)$:

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

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

5.1.1 Experimental Setting

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

$$F(x) = \int f(z) F_z(x) dz. \tag{31}$$

First, we define $f(z)$ as a mixture of Gaussian distributions. To do so we choose a mean μ_i from a uniform distribution in $[-5, 5]$ and then a standard deviation σ_i from a uniform distribution between $[0, 4]$, together defining a Gaussian distribution $g_i(z)$. We repeat this process for 30 Gaussian distributions and define $f(z) = \sum_{i=1}^{30} \frac{g_i(z)}{30}$.

Second, we define the function ξ_z as

$$\xi_z = \frac{\frac{(nz+2m^2+kz^3)e^{-|z|}+a}{b} + c}{d}, \tag{32}$$

where $n = 1$, $m = 2$, $k = 2$, $b = 5.76$, $a = -3b - 3.80$, $d = (\frac{7}{8}\xi_{\max} + \frac{29}{8})^{-1}$ and $c = d\xi_{\max} + 3$. The $\xi_{\max}$ in the variables c, d determines the maximum value that the function ξ_z takes as long as $\xi_{\max} \in [-4, 5]$. More details about the function ξ_z are provided in Appendix G.

Third, we define $F_z(x)$. If $\xi_z \leq 0$, then we define $F_z(x)$ as a Generalized Pareto distribution (GPD) where the scale parameter is set to 1 and the tail shape parameter is ξ_z. Otherwise we define $F_z(x)$ as $1 - x^{-\frac{1}{\xi_z}}$. The choice of $\xi_{\max}$ completely determines each ξ_z and hence each $F_z(x)$, thus it fully defines $F(x)$ in Equation (31).

We run the experiments for different values of the parameter $\xi_{\max}$, that is, $\xi_{\max}$ takes the following 45 values $\{-4, -4 + 0.2, -4 + 0.4, ..., 5\}$. We denote these $\xi_{\max}$ values as $\xi_j = -4 + \frac{2j}{10}$, where $j \in \{0, ..., 45\}$. Each choice of j defines a particular maximal value $\xi_{\max} = \xi_j$ and thus a marginal distribution $F_j(x)$ as on the left side of Equation (31). Also since the particular choice j of the maximum $\xi_{\max}$ determines all corresponding ξ_z in Equation (32) then we denote ξ_z as $\xi_{z,j}$.

For each j we repeat p times Algorithm 3. On repetition k, the algorithm returns $\hat{\xi}_j^k$ which is an estimation of ξ_j. As guided by the ideas laid in Appendix E, our final estimation of ξ_j after p repetitions of Algorithm 3 above is $\hat{\xi}_j = \frac{1}{p}\sum_{k=1}^{p} \hat{\xi}_j^k$.

Algorithm 3 Construction of a Continuous Mixture Distribution and Direct POT Usage

define: $J = \{\}$
fix the number of iterations: $M \in \mathbb{N}$
repeat
  a) Sample a z from the distribution $f(z)$
  b) For that z, calculate $\xi_{z,j}$ (given that $\xi_{\max} = \xi_j$)
  if $\xi_{z,j} \leq 0$ **then**
    c) Sample a point x from a GPD with location zero, $\sigma = 1$, shape parameter $\xi_{z,j}$.
    d) $J = J \cup \{x\}$
  else
    c) Sample x from $F_z(x) = 1 - x^{-\frac{1}{\xi_{z,j}}}$
    d) $J = J \cup \{x\}$
  end if
until $|J| = M$
apply the Pickands or DEdH estimator on J to estimate ξ_j, the shape parameter of $F_j(x)$.
return the estimated value of ξ_j

5.1.2 Experimental validation of CTE using the Pickands estimator

We show the results of the experiment described above, when the Pickands Estimator is applied. In this study, a comprehensive set of experimental outcomes has been illustrated in Figure 1. Here, the parameter M, delineated in the preceding subsection, is assigned values from the set $\{10^5, 10^6, 10^7, 10^8\}$. In the context of these experiments, p was set to 10 as a constant across all trials. The experiments were performed encompassing a total of 10 runs to capture potential variability and better reflect the stochastic nature of the process.

In order to acquire a more robust and representative understanding of the results, given the inherent variability of the experimental setup, statistical metrics including the mean and standard deviation were computed across these multiple experimental runs.

Upon examining the obtained results, they seem to align with our initial theoretical expectations. Specifically, when the maximum tail shape parameter within the mixture of conditional distributions is of positive value, the estimated shape parameter of the marginal distribution is that identical positive value. On the other hand, if the maximum tail shape parameter within the conditional distributions is negative, the estimated shape parameter of the marginal distribution duly returns a negative value. This symmetry in the estimations provides a degree of confidence in the validity of the conducted experiments and the consistency of the underlying theoretical framework.

Complementary results, pertaining to a replica of the above-described experiment, wherein the DEdH Estimator is utilized, can be found in Appendix I.

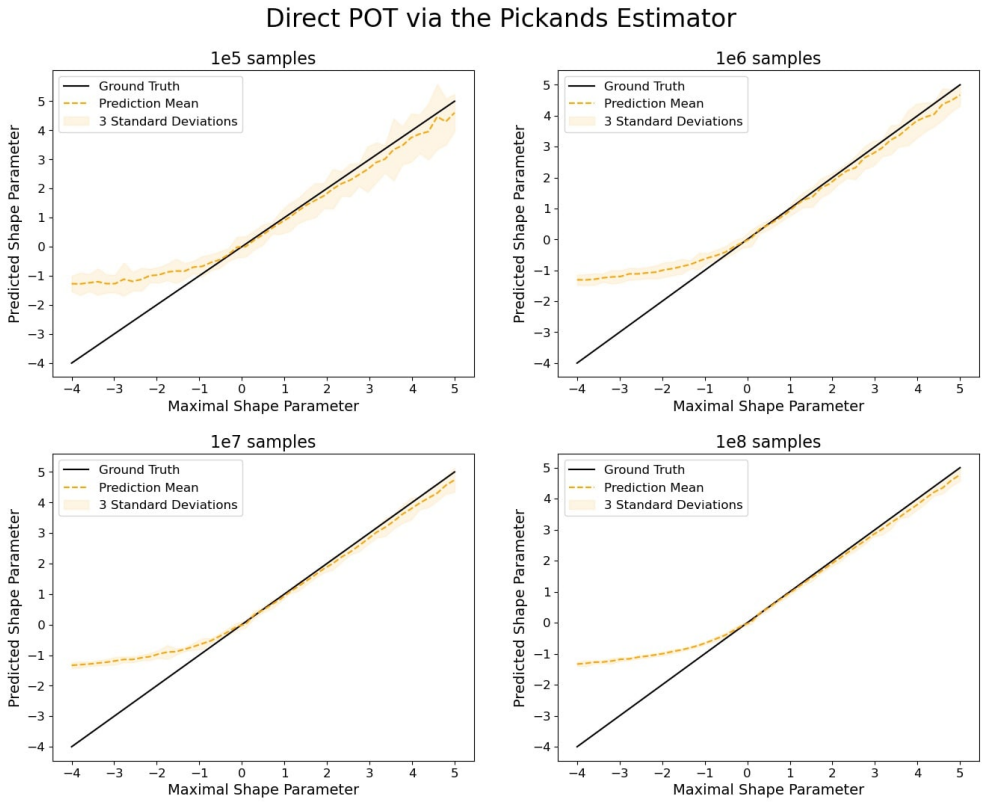

Figure 1: In cases where the maximum tail shape parameter in the mixture of conditional distributions is positive, the estimated shape parameter of the marginal is equal to this maximal value. If this maximum value is negative, the estimated shape parameter is negative. These results were obtained using the Pickands estimator.

5.2 Robustness to Variance in the Location of Conditional Distributions

In Subsection 5.1, we presented empirical evidence to substantiate Theorem 21. Notably, for computational expediency, we elected to set all conditional distributions with a location parameter of zero. This decision was motivated by the fact that, if location parameters were permitted to exhibit significant variability, the direct Peaks Over Threshold (POT) approach would necessitate an unfeasibly large sample size to verify our claims. This issue is addressed in the current subsection, wherein we illustrate that the CTE approach provides a suitable remedy. Specifically, in Subsection 5.2.1, we outline modifications to the experimental setup from Subsection 5.1 that allow for variation in the location parameter, and present the experimental results accordingly. In Subsection 5.2.2, we apply the CTE approach to the same marginal distributions as in Subsection 5.2.1, and demonstrate that it allows for correct estimation of shape parameters. Additional experiments, in more simplified settings, highlighting the necessity of CTE are provided in Appendix F.

5.2.1 Applying POT directly when the location of conditional distributions exhibits substantial variability

In order to ensure high variability of the location of conditional distributions $F_z(x)$, we modify step (c) in the *if* statement of Algorithm 3 into (c*) as delineated in Algorithm 4.

Algorithm 4 Modification of Algorithm 3 to Ensure High Location Variability

if $\xi_{z,j} \leq 0$ **then**
  c) Sample a point x from a GPD with location zero, $\sigma = 1$, shape parameter $\xi_{z,j}$.
  d) $J = J \cup \{x\}$
else
  c*) Sample x from $F_z(x) = 1 - x^{-\frac{1}{\xi_{z,j}}}$. Translate x by adding $\frac{1}{{\xi_{z,j}}^4}$, i.e., $x = x + \frac{1}{{\xi_{z,j}}^4}$.
  d) $J = J \cup \{x\}$
end if

This adaptation ensures that conditional distributions with lower positive shape parameters are situated at greater distances from the origin, thereby augmenting the probability that their tails will dominate over those that exhibit heavier tails.

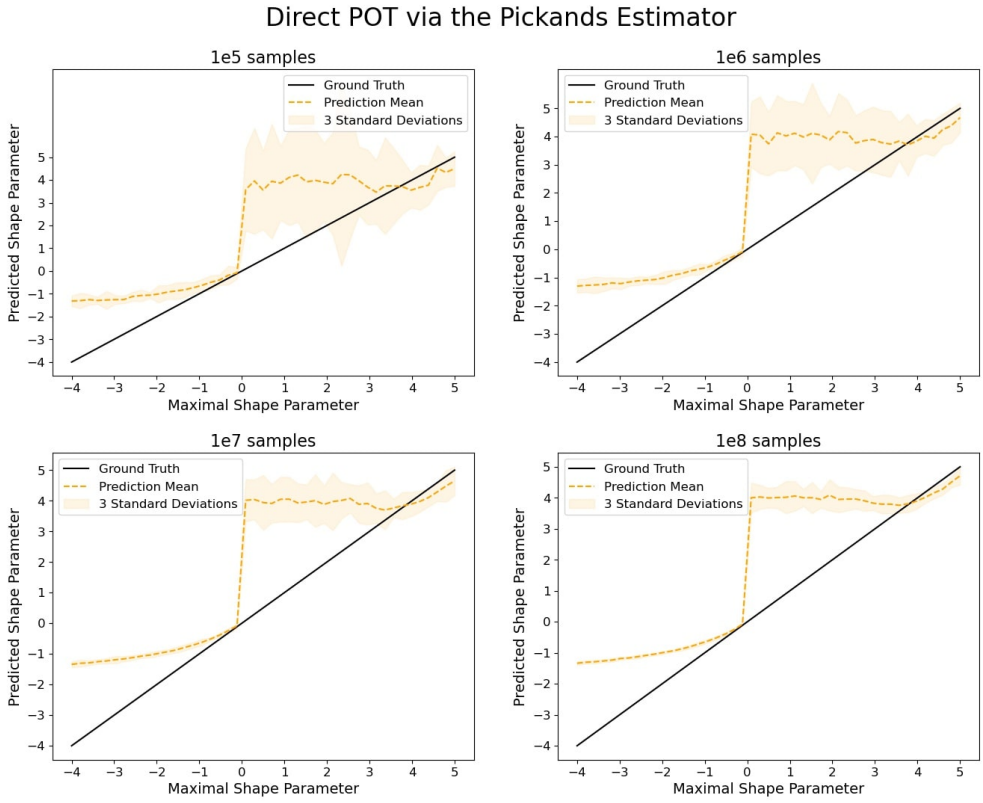

Figure 2: Estimation of the shape parameter of the marginal by direct application of POT. We utilize the Pickands estimator as our estimator of choice.

The results in Figure 2 show that the estimators predict that the shape parameter of the tail is constantly 4 as the tail of the marginal is determined by ${\xi_{z,j}}^{-4}$ instead of $1 - x^{-\frac{1}{\xi_{z,j}}}$ which merely becomes noise around $\xi_{z,j}^{-4}$. This changes once $\xi_{\max} = \xi_j$ becomes larger than 4, in which case the tails of the conditional distribution are once again determined by $1 - x^{-\frac{1}{\xi_{z,j}}}$.

Complementary results, pertaining to a replica of the above-described experiment, wherein the DEdH Estimator is utilized, can be found in Appendix I, Figure 15.

5.2.2 Enhancing parameter estimation accuracy through the CTE approach

We demonstrate that the CTE method can effectively recover the true shape of the tail of the marginal, even in cases where the conditional distributions exhibit highly varying locations, as observed in Subsection 5.2.1. To ensure objectivity, we define the functions $f(z)$, $\xi_{z,j}$, and $F_{z,j}(x)$ in a consistent manner as before, thereby ensuring that all marginal distributions under consideration are equivalent to those studied in previous cases. As per the definition of the CTE, the sampling and estimation procedure is described in detail in Algorithm 5.

Algorithm 5 Application of CTE on the Mixture Distribution Defined in Algorithm 4

sample K values z from $f(z)$
for each z **do**
  Calculate $\xi_{z,j}$ (given that $\xi_{\max} = \xi_j$)
  for $l = 1$ to p **do**
    if $\xi_{z,j} \leq 0$ **then**
      Sample a set S of N samples from a GPD with shape parameter $\xi_{z,j}$, scale $\sigma = 1$.
    else
      Sample a set S of N samples from $F_z(x) = 1 - x^{-\frac{1}{\xi_{z,j}}}$.
      Translate each sample x in S as follows: $x = x + \frac{1}{{\xi_{z,j}}^4}$.
    end if
    Apply the Pickands or DEdH estimator on S to get an estimate $\hat{\xi}^l_{z,j}$ of the shape parameter $\xi_{z,j}$ of $F_{z,j}(x)$
  end for
  As guided by the ideas laid in Appendix E, our final estimation of $\xi_{z,j}$ after p repetitions of the process above is $\hat{\xi}_{z,j} = \frac{1}{p}\sum_{l=1}^{p} \hat{\xi}^l_{z,j}$.
end for
We select the maximal $\hat{\xi}_{z,j}$ from the K predicted values (corresponding to the K sampled z). According to Theorem 21 the estimated $\hat{\xi}_j$ should be close to ξ_j.
return $\hat{\xi}_j$, the estimated value of ξ_j

We set $p = 10$ at all times. Furthermore, for the sake of fairness, we sample the same number of points from each marginal distribution as in the previous subsection, that is, we set $KN = M$. Since we set $K = 50$, in order for M to take values in $\{1e5, 1e6, 1e7, 1e8\}$, N needs to take values in $\{2e3, 2e4, 2e5, 2e6\}$. We execute the experiment 10 times, and to

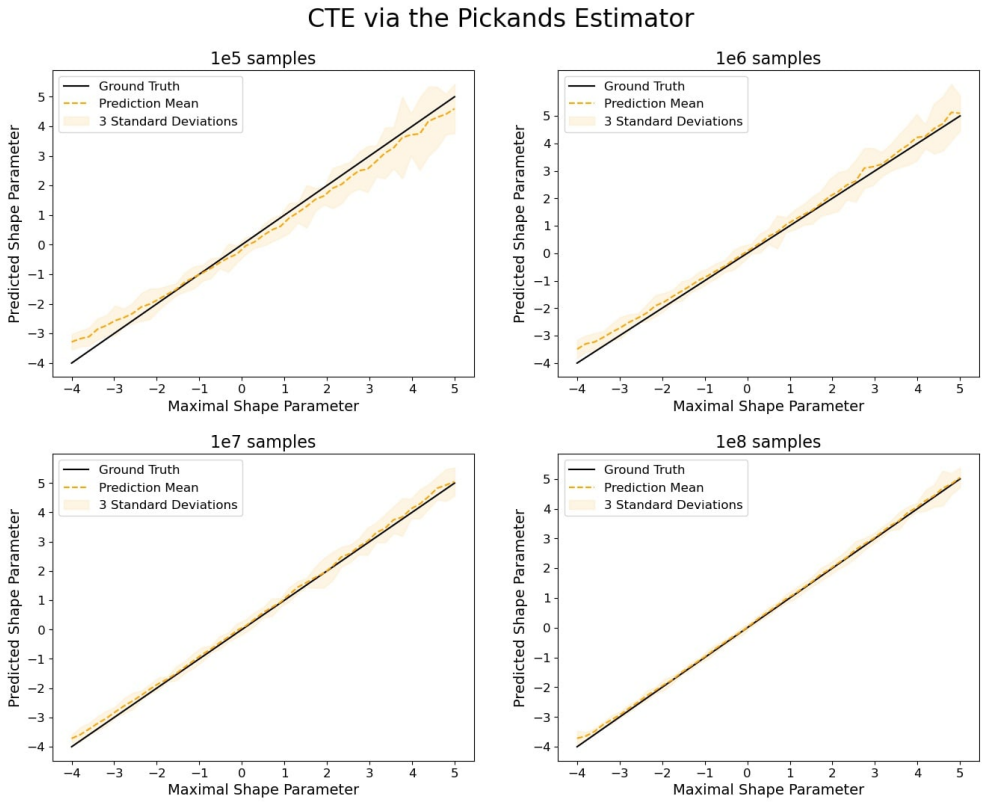

Figure 3: Estimation the shape parameter of the marginal using CTE. We utilize the Pickands estimator as our estimator of choice.

account for variability across the different runs, we compute the mean and standard deviation of the results, which are shown in Figure 3. Naturally, the more K is increased the more likely we are to sample the z corresponding to the conditional distribution with the maximal shape parameter. Theorem 21 provides assurance that as the value of K increases, our estimation progressively converges to the true shape parameter of the marginal distribution. The conducted experiments indicate that merging samples from various conditional distributions, which form the marginal, may potentially be detrimental when estimating the tail shape of the marginal.

Complementary results, pertaining to a replica of the above-described experiment, wherein the DEdH Estimator is utilized, can be found in Appendix I, Figure 16.

5.3 Model Performance Inference Improvements via Cross Tail Estimation, Relative to POT

In what follows, we show 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 positively correlated with over-fitting, 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 a one-dimensional time series taken from the UCR Time Series Anomaly Archive [2] (Wu and Keogh, 2020), which we reorganize in windows of size 2, and use each window to fit a Gaussian process (GP) model in order to predict the next value in the series. Our complete data set D is composed of $n = 1e4$ windows. On each run we randomly select 340 points of D for training (denote D_i), and then group

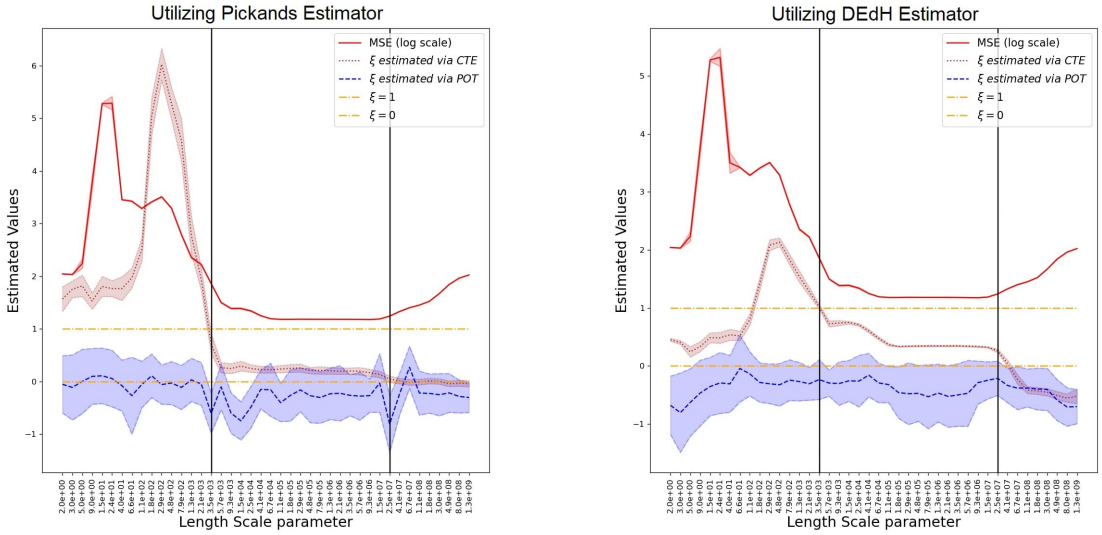

Figure 4: Experimental results in the case of testing Gaussian processes. Left: The Pickands estimator is used. Right: The DEdH estimator is used. In both cases we notice that CTE estimates larger shape parameters of the loss function distributions for models which overfit. This is not the case when POT is applied directly. The first black vertical line marks the first model with lower MSE than the model with the smallest length scale parameter (the point where the models stop overfitting). The second black vertical line marks the model in from which MSE starts growing again (the point when models begin underfitting). The MSE is presented in log scale and has been further linearly scaled to fit the plot. Shaded areas denote the standard deviation of the measurements across 200 independent runs.

the predictions of the model on the $1e4$ points of D into an array which we denote by $\hat{Y}_i$. Then we split $\hat{Y}_i$ into five equally sized subsets $\hat{Y}_{i,j}$. We proceed to estimate the shape parameter of the tails of the prediction of the model, for given training set D_i. This is done by applying the Pickands/DEdH estimator to $\hat{Y}_{i,j}$, receiving $\hat{\xi}_{i,j}$ and then as per Appendix

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

E, we get the estimate $\hat{\xi}_i = \frac{1}{5}\sum_{j=1}^{5}\hat{\xi}_{i,j}$ which corresponds to $\hat{Y}_i$. We repeat this process 1000 times (for 1000 choices of the training set D_i), 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 direct POT 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/DEdH estimator on $Y = \bigcup_{i=1}^{1000} \hat{Y}_i$.

These experiments are repeated for length scale parameters given in the $x-$axis of Figure 4 as well as in Appendix H. We repeat every experiment 200 times to account for variability across different runs, we compute the mean and standard deviation of the results.

In Figure 4 , we notice that when the CTE approach is used, the shape parameter is significantly larger for models which overfit. In Appendix H, we illustrate that the MSE is large for small scale parameters due to overfitting (Figure 9). Furthermore, the shape parameter only drops to (under) zero, when the model starts underfitting for length scale parameters bigger than $2.5e7$. In Appendix H (Figure 10), it is shown that for such large values of the length scale parameter, the predictions become roughly constant.

On the other hand, if POT is applied directly, then the estimated shape parameters are not significantly larger for models which overfit compared to those that do not. This is because conditioning on the training set, the predicted values on the test set vary significantly with regards to the their location. Hence, the tail that is estimated by the direct application of the POT approach is sometimes simply the one translated the furthest from the origin. Thus, if there is some inverse relationship between the magnitude of the location and the size of the estimated shape parameter across different conditional, then we expect POT to underestimate the true shape parameter of the marginal. This is shown in Appendix H, for the model with the highest estimated shape parameter (290). The variability (sorted) of the estimated shape parameters of the 1000 conditionals for each length scale parameter is given in Figure 12 of Appendix H, together with the corresponding $97th$ percentile (threshold) from each corresponding conditional distribution. We notice that indeed, quite often the difference between locations is large, and that the largest threshold often corresponds to conditional distributions with small, even negative shape parameters.

The outcomes presented herein are robust with regards to the choice of applying the method explicated in Appendix E in conjunction with the direct Peaks Over Threshold (POT) approach. Furthermore, the findings presented in Figure 4 demonstrate near equivalence in relation to the magnitude of the selected threshold (in this study, we evaluated 99.7 and 99.997 percentiles).

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 set 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 5.

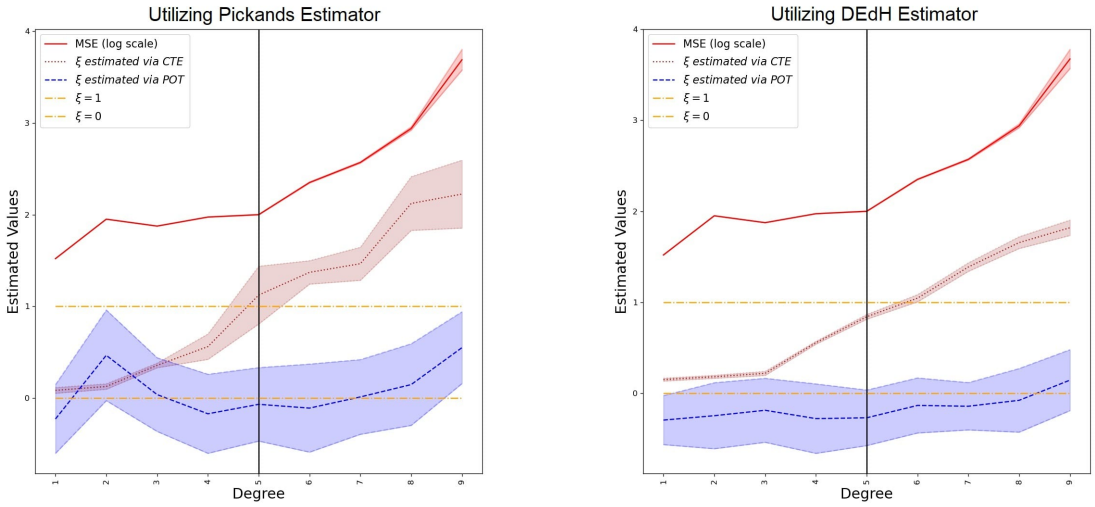

Figure 5: Experimental results in the case of testing polynomial kernels. Left: The Pickands estimator is used. Right: The DEdH estimator is used. In both cases we notice that CTE estimates larger shape parameters of the loss function distributions for models which overfit. This is not the case when POT is applied directly. The black vertical line marks the inflection point of the MSE. The MSE is presented in log scale and has been further linearly scaled to fit the plot. Shaded areas denote the standards deviation of the measurement across 200 independent runs.

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 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 conditional distributions composing the marginal. 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 phenomenon regarding the relationship between the overfitting 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.

Acknowledgments

This work has been supported by the French government, through the 3IA Cte dAzur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002. The authors are grateful to the OPAL infrastructure from Universit Cte d'Azur for providing resources and support.

Appendix A. Proofs

Proof of Proposition 9

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 11

From Theorem 10, 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 12

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 13

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 14

Since if $\xi_{\boldsymbol{z}_i} < 0$ then $\exists x_0 > 0$, such that $\forall x > x_0$ we have $F_{\boldsymbol{Z}_i}(x) = 0$, this means that the tail of the distribution is not affected by $F_{\boldsymbol{Z}_i}(x)$. In fact if $\xi_{\max} < 0$ then F will have finite support hence $\xi_F \leq 0$. Furthermore if $\xi_{\max} = 0$ from Lemma 12 we get that $\xi_F \leq 0$. Therefore for the case $\xi_{\max} > 0$ we only consider the setting where $\xi_i \geq 0$.

$$\bar{F}_u(w) = \frac{1 - F(u+w)}{1 - F(u)} = \frac{\sum_i^n p_i(1 - F_{\boldsymbol{z}_i}(u+w))}{\sum_i^n p_i(1 - F_{\boldsymbol{z}_i}(u))} = \sum_i^n \frac{\bar{F}_{\boldsymbol{z}_i}(u+w)}{\sum_j^n \frac{p_j}{p_i}\bar{F}_{\boldsymbol{z}_j}(u)} \tag{42}$$

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

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

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

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

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

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

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

The first expression,

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

goes to zero due to Pickands Theorem while the expression,

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

converges to 0 as well because from Lemma 11 we have $\lim_{u\to\infty} \frac{\bar{F}_{\boldsymbol{z}_j}(u)}{\bar{F}_{\boldsymbol{z}_{i(\max)}}(u)} = 0$ for every j. Finally the last expression,

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

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

In the derivation above we assumed that the $F_{\boldsymbol{z}_{i(\max)}}$ which corresponds to $\xi_{\max}$ is unique. In the case that this is not true we notice that for F_1 and F_2 which share the same corresponding parameter $\xi > 0$ we have $p_1F_1(x)+p_2F_2(x) = x^{-\frac{1}{\xi}}(p_1L_1(x)+p_2L_2(x)) = x^{-\frac{1}{\xi}}L(x)$, and since $L(x) > 0$, from Lemma 13 we have that $L(x)$ is slowly varying, therefore $p_1F_1(x)+p_2F_2(x) \in MDA(\xi)$. ∎

Proof of Proposition 16

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

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

Proof of Theorem 17

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

$$\frac{\bar{F}(x)}{x^{-\frac{1}{\xi_{lo}-\epsilon}}} = \frac{x^{-\frac{1}{\xi_F}} L_F(x)}{x^{-\frac{1}{\xi_{lo}-\epsilon}}} = \frac{\int_A f_{\boldsymbol{Z}}(\boldsymbol{z}) x^{-\frac{1}{\xi_{\boldsymbol{z}}}} L_{\boldsymbol{z}}(x) d\boldsymbol{z}}{x^{-\frac{1}{\xi_{lo}-\epsilon}}} = \int_A f_{\boldsymbol{Z}}(\boldsymbol{z}) x^{-\frac{1}{\xi_{\boldsymbol{z}}}+\frac{1}{\xi_{lo}-\epsilon}} L_{\boldsymbol{z}}(x) d\boldsymbol{z} = \int_A f_{\boldsymbol{Z}}(\boldsymbol{z}) x^{\alpha(\boldsymbol{z})} L_{\boldsymbol{z}}(x) d\boldsymbol{z}. \tag{52}$$

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

$$\lim_{x\to\infty} \frac{\bar{F}(x)}{x^{-\frac{1}{\xi_{lo}-\epsilon}}} = \lim_{x\to\infty} \int_A f_{\boldsymbol{Z}}(\boldsymbol{z}) x^{\alpha(\boldsymbol{z})} L_{\boldsymbol{z}}(x) dz, \tag{53}$$

by using Fatou's lemma:

$$\lim_{x\to\infty} \int_A f_{\boldsymbol{Z}}(\boldsymbol{z}) x^{\alpha(\boldsymbol{z})} L_{\boldsymbol{z}}(x) d\boldsymbol{z} \geq \int_A \lim_{x\to\infty} f_{\boldsymbol{Z}}(\boldsymbol{z}) x^{\alpha(\boldsymbol{z})} L_{\boldsymbol{z}}(x) d\boldsymbol{z} = \infty, \tag{54}$$

we get

$$\lim_{x\to\infty} \frac{x^{-\frac{1}{\xi_{lo}-\epsilon}}}{\bar{F}(x)} = 0, \tag{55}$$

implying

$$\lim_{x\to\infty} \frac{x^{-\frac{1}{\xi_{lo}-\epsilon}}}{x^{-\frac{1}{\xi_F}} L_F(x)} = \lim_{x\to\infty} \frac{x^{-\frac{1}{\xi_{lo}-\epsilon}+\frac{1}{\xi_F}}}{L_F(x)} = 0, \tag{56}$$

therefore

$$\xi_{lo} - \epsilon < \xi_F, \forall \epsilon > 0 \text{ thus } \xi_{lo} \leq \xi_F. \tag{57}$$

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

$$\frac{\bar{F}(x)}{x^{-\frac{1}{\xi_{up}+\epsilon}}} = \frac{x^{-\frac{1}{\xi_F}} L_F(x)}{x^{-\frac{1}{\xi_{up}+\epsilon}}} = \frac{\int_A f_{\boldsymbol{Z}}(\boldsymbol{z}) x^{-\frac{1}{\xi_{\boldsymbol{z}}}} L_{\boldsymbol{z}}(x) d\boldsymbol{z}}{x^{-\frac{1}{\xi_{up}+\epsilon}}} = \int_A f_{\boldsymbol{Z}}(\boldsymbol{z}) x^{-\frac{1}{\xi_{\boldsymbol{z}}}+\frac{1}{\xi_{up}+\epsilon}} L_{\boldsymbol{z}}(x) d\boldsymbol{z} = \int_A f_{\boldsymbol{Z}}(\boldsymbol{z}) x^{\beta(\boldsymbol{z})} L_{\boldsymbol{z}}(x) d\boldsymbol{z}. \tag{58}$$

We notice that $\xi_{\boldsymbol{z}} \leq \xi_{up} < \xi_{up} + \epsilon \implies -\frac{1}{\xi_{\boldsymbol{z}}} \leq -\frac{1}{\xi_{up}} < -\frac{1}{\xi_{up}+\epsilon}$ hence $\beta(\boldsymbol{z}) = -\frac{1}{\xi_{\boldsymbol{z}}} + \frac{1}{\xi_{up}+\epsilon} < -\delta < 0$. 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

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) x^{\beta(\boldsymbol{z})} L_{\boldsymbol{z}}(x) \leq f_{\boldsymbol{Z}}(\boldsymbol{z}) x^{-\delta} L_{\boldsymbol{z}}(x) \leq f_{\boldsymbol{Z}}(\boldsymbol{z}) \gamma, \tag{59}$$

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

$$\lim_{x\to\infty} \frac{\bar{F}(x)}{x^{-\frac{1}{\xi_{up}+\epsilon}}} = \lim_{x\to\infty} \int_A f_{\boldsymbol{Z}}(\boldsymbol{z}) x^{\beta(\boldsymbol{z})} L_{\boldsymbol{z}}(x) d\boldsymbol{z} \tag{60}$$

$$\lim_{x\to\infty}\int_A f_{\boldsymbol{Z}}(\boldsymbol{z})x^{\beta(\boldsymbol{z})}L_{\boldsymbol{z}}(x)d\boldsymbol{z}=\int_A\lim_{x\to\infty}f_{\boldsymbol{z}}(\boldsymbol{z})x^{\beta(\boldsymbol{z})}L_{\boldsymbol{z}}(x)d\boldsymbol{z}=0, \tag{61}$$

meaning

$$\lim_{x\to\infty}\frac{\bar{F}(x)}{x^{-\frac{1}{\xi_{up}+\epsilon}}}=0, \tag{62}$$

which implies

$$\lim_{x\to\infty}\frac{x^{-\frac{1}{\xi_F}}L_F(x)}{x^{-\frac{1}{\xi_{up}+\epsilon}}}=\lim_{x\to\infty}x^{\frac{1}{\xi_{up}+\epsilon}-\frac{1}{\xi_F}}L_F(x)=0, \tag{63}$$

therefore we get

$$\xi_{up}+\epsilon>\xi_F,\forall\epsilon>0 \text{ hence } \xi_F\leq\xi_{up}. \tag{64}$$

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

$$\frac{\bar{F}(x)}{\bar{G}_0(x)}=\frac{\bar{F}(x)}{x^{-\frac{1}{\epsilon}}}\frac{x^{-\frac{1}{\epsilon}}}{\bar{G}_0(x)}=\frac{\int_A f_{\boldsymbol{Z}}(\boldsymbol{z})x^{-\frac{1}{\xi_{\boldsymbol{z}}}}L_{\boldsymbol{z}}(x)d\boldsymbol{z}}{x^{-\frac{1}{\epsilon}}}\frac{x^{-\frac{1}{\epsilon}}}{\bar{G}_0(x)}. \tag{65}$$

As before we can prove that the first fraction $\frac{\bar{F}(x)}{x^{-\frac{1}{\epsilon}}}\to\infty$. The expression $\frac{x^{-\frac{1}{\epsilon}}}{\bar{G}_0(x)}$ goes to ∞ as well due to Lemma 12, thus

$$\lim_{x\to\infty}\frac{\bar{F}(x)}{\bar{G}_0(x)}=\infty. \tag{66}$$

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

$$\lim_{x\to\infty}\frac{\bar{F}(x)}{\bar{G}_0(x)}=\lim_{x\to\infty}1=1, \tag{67}$$

hence $\xi_F\neq 0$.

Finally we prove that, if $\xi_{\boldsymbol{z}}$ is continuous in $\boldsymbol{z}$ and $\xi_{\max}$ exists, then we have $\xi_F=\xi_{\max}$. 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,

$$\begin{aligned}\bar{F}(x)=\int_A f_{\boldsymbol{Z}}(\boldsymbol{z})\bar{F}_z(x)d\boldsymbol{z}&=p_1\int_{A_1}\frac{f_{\boldsymbol{Z}}(\boldsymbol{z})}{p_1}\bar{F}_{\boldsymbol{z}}(x)d\boldsymbol{z}+p_2\int_{A_2}\frac{f_{\boldsymbol{Z}}(\boldsymbol{z})}{p_2}\bar{F}_{\boldsymbol{z}}(x)d\boldsymbol{z}\\&=p_1\bar{F}_1(x)+p_2\bar{F}_2(x).\end{aligned} \tag{68}$$

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 $F_i\in MDA(\xi_i),\ i=1,2$. On the other hand Theorem 14 implies that $\xi_F=\xi_1$, therefore $\xi_F\in[\xi_{\max}-\lambda,\xi_{\max}]$ for all $\lambda>0$. We conclude that $\xi_F=\xi_{\max}$. ∎

Proof of Lemma 19

We assume that $\xi_F > \epsilon$. Then as in the earlier derivations, due to dominated convergence and Lemmas 11 and 12, for any $\delta > 0$, we get:

$$\begin{aligned}\lim_{x\to\infty} \frac{x^{-\frac{1}{\xi_F}} L_F(x)}{x^{-\frac{1}{\epsilon+\delta}}} &= \lim_{x\to\infty} \frac{\bar{F}(x)}{x^{-\frac{1}{\epsilon+\delta}}} = \lim_{x\to\infty} \int_A f_{\boldsymbol{Z}}(\boldsymbol{z}) \frac{\bar{F}_{\boldsymbol{z}}(x)}{x^{-\frac{1}{\epsilon+\delta}}} d\boldsymbol{z} \\ &= \lim_{x\to\infty} \int_{A^+} f_{\boldsymbol{Z}}(\boldsymbol{z}) \frac{\bar{F}_{\boldsymbol{z}}(x)}{x^{-\frac{1}{\epsilon+\delta}}} d\boldsymbol{z} + \lim_{x\to\infty} \int_{A^-} f_{\boldsymbol{Z}}(\boldsymbol{z}) \frac{\bar{F}_{\boldsymbol{z}}(x)}{x^{-\frac{1}{\epsilon+\delta}}} d\boldsymbol{z} \\ = \int_{A^+} \lim_{x\to\infty} f_{\boldsymbol{Z}}(\boldsymbol{z}) \frac{x^{-\frac{1}{\xi_{\boldsymbol{z}}}}}{x^{-\frac{1}{\epsilon+\delta}}} L_{\boldsymbol{z}}(x) d\boldsymbol{z} &+ \int_{A^-} \lim_{x\to\infty} f_{\boldsymbol{Z}}(\boldsymbol{z}) \frac{\bar{F}_{\boldsymbol{z}}(x)}{x^{-\frac{1}{\epsilon+\delta}}} d\boldsymbol{z} = 0.\end{aligned} \tag{69}$$

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

Proof of Theorem 21

The proof is similar to that of the last statement in Theorem 17. 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 $i \in \{1, 2\}$.

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

Based on Theorem 17 and Lemma 19: $\xi_1 = \xi_{\max}$, and $\xi_2 \in (-\infty, \xi_{\max} - \lambda]$, where $F_i \in MDA(\xi_i),\ i = 1, 2$. From Theorem 14, we conclude that $\xi_F = \xi_{\max}$. The last statement in the Theorem, that is, if $\xi_{\max} \leq 0$ then $\xi_F \leq 0$, is simply Corollary 20. ∎

Proof of Proposition 23

In the case that $\xi_X > 0$, based on our assumptions there exists $L(x)$ such that

$$\mathbb{P}(X > x) = \bar{F}_X(x) = x^{-\frac{1}{\xi_X}} L_1(x). \tag{71}$$

Therefore

$$\bar{F}_Y(x) = \mathbb{P}(Y > x) = \mathbb{P}(X^\alpha > x) = \mathbb{P}(X > x^{\frac{1}{\alpha}}) = (x^{\frac{1}{\alpha}})^{-\frac{1}{\xi_X}} L_1(x^{\frac{1}{\alpha}}) = x^{-\frac{1}{\alpha\xi_X}} L_2(x). \tag{72}$$

We conclude that $Y \in MDA(\alpha\xi_X)$. On the other hand if $\xi_X \leq 0$ then $\xi_Y \leq 0$, because if $\xi_Y > 0$, then from the first part we would have $\xi_X = \frac{1}{\alpha}\xi_Y > 0$. ∎

Proof of Proposition 24

We will first prove the case when $p = 1$. If we fix y and denote with ξ_y^{h-}, ξ_y^{h+} the shape parameters of the left and right tail of $p(\hat{f}_{\boldsymbol{V}}(\boldsymbol{X})|y)$, then assuming that at least one of them

is positive, from Proposition 21 we know that the tail shape parameter of $p(|\hat{f}_{\boldsymbol{V}}(\boldsymbol{X})||y)$ is $\xi_y^h = \max\{\xi_y^{h-}, \xi_y^{h+}\}$. We notice now that ξ_y^{h-}, ξ_y^{h+} are the right and left tail shape parameters of $p(-\hat{f}_{\boldsymbol{V}}(\boldsymbol{X})|y)$, therefore they are the right and left tail shape parameters of the distribution $p(y - \hat{f}_{\boldsymbol{V}}(\boldsymbol{X})|y)$. Due to this, if we denote with ξ_y^g the tail shape parameter of $p(|y - \hat{f}_{\boldsymbol{V}}(\boldsymbol{X})||y)$, using Proposition 21 once again we have that $\xi_y^g = \max\{\xi_y^{g+}, \xi_y^{g-}\} = \max\{\xi_y^{h-}, \xi_y^{h+}\} = \xi_y^h$, where ξ_y^{g-}, ξ_y^{g+} are the left and right shape parameters of $p(y - \hat{f}_{\boldsymbol{V}}(\boldsymbol{X})|y)$. If both ξ_y^{h-}, ξ_y^{h+} are non-positive then from Proposition 21, ξ_y^h is non-positive, and furthermore ξ_y^g is non-positive, otherwise we could go in the reverse direction and prove that $\xi_y^g > 0$ implies that either $\xi_y^{g-} = \xi_y^{h+}$ is positive, or that $\xi_y^{g+} = \xi_y^{h-}$ is positive.
Now, we denote by $G_y(s)$ the distribution of $|y - \hat{f}_{\boldsymbol{V}}(\boldsymbol{X})|$ given y, and prove that the family $\{G_y(s)|y \in S\}$ has stable cross-tail variability. For each y we denote with $t_0(y)$ the smallest value after which the sub-polynomial assumption is satisfied by $F_y(t)$. Similarly we define $s_0(y)$ for $G_y(s)$. Since the family $\{F_y(t)|y \in S\}$ has stable cross-tail variability, then each such $t_0(y)$ exists, and furthermore the set $\{t_0(y)|y \in S\}$ is bounded from above. Since each $s_0(y)$ is only displaced by a magnitude of $|y|$ from $t_0(y)$, and since the set S is bounded, then we can conclude that $\{s_0(y)|y \in S\}$ is bounded from above.
We denote ξ^g, ξ^h the tail shape parameters of $|Y - \hat{f}_{\boldsymbol{V}}(\boldsymbol{X})|$ and $|\hat{f}_{\boldsymbol{V}}(\boldsymbol{X})|$ respectively. Using Theorem 21 twice we get that if there is at least one $\xi_y^h = \xi_y^g > 0$ then $\xi^h = \max\{\xi_y^h|y \in S\} = \max\{\xi_y^g|y \in S\} = \xi^g > 0$, otherwise $\xi^h \leq 0, \xi^g \leq 0$.
Finally we finish the proof by applying Proposition 22 on $|Y - \hat{f}_{\boldsymbol{V}}(\boldsymbol{X})|$ and $|\hat{f}_{\boldsymbol{V}}(\boldsymbol{X})|$. ∎

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

$$h(w) = \int_0^1 f_U(u) g_u(w) du = \int_0^1 u e^{-uw} du \tag{73}$$

the expectation is

$$\int_0^\infty \int_0^1 w f_U(u) g_u(w) du dw = \int_0^1 \int_0^\infty w u e^{-uw} dw du = \int_0^1 \frac{1}{u} du \tag{74}$$

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:

$$L_z(x) = \begin{cases} 1 + zx^{4-(z-\frac{1}{x})^2} & \text{for } x \in (1, \frac{1}{z}) \\ 1 + \frac{1}{z^3} & \text{for } x \in (\frac{1}{z}, \infty) \end{cases} \tag{75}$$

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:

$$F_z(x) = \begin{cases} x^{-1} + zx^{3-(z-\frac{1}{x})^2} & \text{for } x \in (1, \frac{1}{z}) \\ x^{-1} + \frac{1}{z^3}x^{-1} & \text{for } x \in (\frac{1}{z}, \infty) \end{cases} \tag{76}$$

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. Moment based motivation

In Proposition 24, we showed that under certain conditions, we could estimate the shape of the tail of the distribution of $W_{\boldsymbol{V}}(\boldsymbol{U})$ without using test labels. This can also be motivated from the moments of $W_{\boldsymbol{V}}(\boldsymbol{U})$. Indeed, conditioning on the test label y we have

$$\mathbb{E}[W^p_{\boldsymbol{V}}(\boldsymbol{U})|Y = y] = E_{\boldsymbol{V}}[(y - \hat{f}_{\boldsymbol{V}}(\boldsymbol{x}))^p|y] \tag{77}$$

$$= \sum_{k=0}^{p} \binom{p}{k} y^k(-1)^{p-k}E_{\boldsymbol{V}}[\hat{f}^{p-k}_{\boldsymbol{V}}(\boldsymbol{x})|y] \tag{78}$$

We can see that for test label y, if the moment p of $\hat{f}_{\boldsymbol{V}}(\boldsymbol{x})$ given y exists then the moment p of $W_{\boldsymbol{V}}(u)$ given y exists. If each $E_{\boldsymbol{V}}[\hat{f}^j_{\boldsymbol{V}}(\boldsymbol{x})|y]$, $j \in \{1, ..., p\}$ changes continuously with y then $\mathbb{E}[W^p_{\boldsymbol{V}}(\boldsymbol{U})|y]$ is continuous with respect to y. Further assuming that the support of Y is compact, then moment p of $W_{\boldsymbol{V}}(\boldsymbol{U})$, that is, $\mathbb{E}[W^p_{\boldsymbol{V}}(\boldsymbol{U})] = \mathbb{E}_y\mathbb{E}[W^p_{\boldsymbol{V}}(\boldsymbol{U})|Y = y]$ will exist as well.

Under these conditions, if $\hat{f}_{\boldsymbol{V}}(\boldsymbol{x})$ is a non-negative function, then the existence of $\mathbb{E}[\hat{f}^p_{\boldsymbol{V}}(\boldsymbol{x})] = \mathbb{E}_y\mathbb{E}[\hat{f}^p_{\boldsymbol{V}}(\boldsymbol{x})|y]$ guarantees the existence of $\mathbb{E}[\hat{f}^p_{\boldsymbol{V}}(\boldsymbol{x})|y]$ for almost all y, thus it ensures the existence of $\mathbb{E}[W^p_{\boldsymbol{V}}(\boldsymbol{U})]$.

Appendix E. Reducing the variability of the estimated shape parameters

It is proven in (Dekkers and Haan, 1989), that under certain conditions on k (in particular that $\frac{k(n)}{n} \to 0$ as $n \to \infty$) the Pickands Estimator has an asymptotically Gaussian distribution: $\sqrt{k(n)}(\hat{\xi}^{(P)}_{k,n} - \xi) \xrightarrow{d} \mathcal{N}(0, \sigma^2(\xi))$. This implies that for large n, we roughly have $\hat{\xi}^{(P)}_{k,n} \sim \mathcal{N}(\xi, \frac{\sigma^2(\xi)}{k(n)})$. Minding the size of n, we can split the n samples into m groups such that $n = m\frac{n}{m}$, and such that we still have roughly $\hat{\xi}^{(P)}_{k,\frac{n}{m}} \sim \mathcal{N}(\xi, \frac{\sigma^2(\xi)}{k(\frac{n}{m})})$. Since we can estimate $\hat{\xi}^{(P)}_{k,\frac{n}{m}}$ for each of the m groups we can define the average estimation as $\hat{\xi}^{(P),avg}_{k,\frac{n}{m}} = \frac{1}{m}\sum_{i=1}^{m}\hat{\xi}^{(P),i}_{k,\frac{n}{m}}$. Under the assumption that samples from such groups are independent, we get that $\hat{\xi}^{(P),avg}_{k,\frac{n}{m}} \sim \mathcal{N}(\xi, \frac{\sigma^2(\xi)}{mk(\frac{n}{m})})$. Since $k(n) = o(n)$, we can choose to reduce the variance 'linearly' by keeping $\frac{n}{m}$ constant and increasing m, instead of increasing the sub-linear $k(n)$. This becomes quite apparent if we set $k(n) = \log n$ or $k(n) = \sqrt{n}$. Indeed, for $k(n) = \log n$, the ratio between the variances of the direct approach and our approach is

$$\frac{m\log\frac{n}{m}}{\log n} = \frac{m\log\frac{n}{m}}{\log m + \log\frac{n}{m}} = \frac{mC}{\log m + C} \to \infty \tag{79}$$

as $m \to \infty$.

Similarly for $k(n) = \sqrt{n}$,

$$\frac{m\sqrt{\frac{n}{m}}}{\sqrt{n}} = \sqrt{m} \to \infty \tag{80}$$

as $m \to \infty$. Here we can see that even if we fix m and then allow each group with size $\frac{n}{m}$ to grow as n increases, the variance is still $\sqrt{m}$ times smaller using our approach.

The asymptotically Gaussian distribution property holds in the case of the DEdH estimator if one knows that $\xi > 0$ (Hill estimator, (Davis and Resnick, 1984)). Furthermore, both estimators $H^{(1)}_{k,n}$ and $H^{(2)}_{k,n}$ in Definition 6 jointly possess this property, (Dekkers et al., 1989).

Appendix F. 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.

Uniform Case

In our experimental procedure, we randomly select samples adhering to two distinct power law distributions. Each of these distributions has a unique characteristic shape parameter

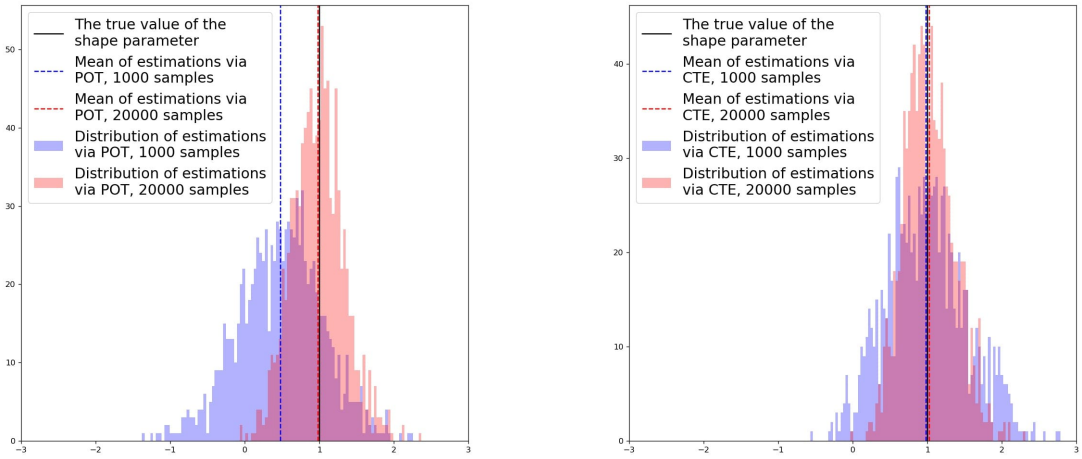

Figure 6: Standard estimation of the shape parameter of the tails by simply applying the Pickands' Estimator, on average, gives poor results on fewer data (left). Cross tail estimation (CTE) gives the correct estimation on average. (right).

- one has a shape parameter of 1, while the other possesses a shape parameter of 0.5. For our random sampling process, we afford equal probability, precisely 50%, to both these distributions. This means there is an identical chance of picking a sample from either of these power law distributions, each with their respective shape parameters.

When we examine an experimental set of 10^3 sampled points from each of these distributions, the resulting pattern becomes apparent as shown in Figure 6 (left). We find that if we amalgamate all the sampled data points from both distributions into a unified array, and subsequently apply Pickands Estimator on this consolidated data set, the process yields a sub-optimal estimation of the distribution tail. The outcome is unsatisfactory as it fails to reveal the accurate shape of the tail, thereby defeating the purpose of the estimation.

However, we discover that there is a noticeable enhancement in the quality of the estimation when we bolster the sample size from the initial 10^3 to a considerably larger size of $2 * 10^4$. This increase in sample size permits us to retrieve the true shape of the distribution tail.

Using CTE however, we find that a sample size of just 10^3 proves to be adequate in obtaining a satisfactory estimation of the distribution tail. As illustrated in Figure 6 (right), this method leads to an accurate estimation with a substantially smaller sample size. Therefore, our method introduces an efficient pathway towards achieving accurate estimations with fewer resources, thereby demonstrating its potential superiority over the traditional Pickands Estimator.

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.

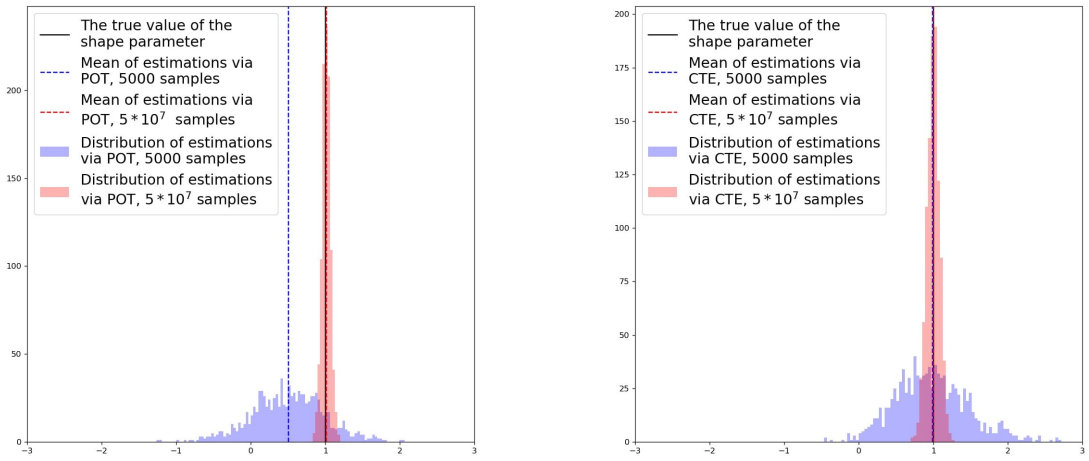

Figure 7: Standard estimation of the shape parameter of the tails by simply applying the Pickands' Estimator, on average, gives poor results on fewer data (left). Cross tail estimation (CTE) gives the correct estimation on average. (right).

When sampling $5 * 10^3$ points from each distribution, Figure 7, we are not able to properly estimate the tail if we join all the samples together in a common array and then 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.

Appendix G. Additional details with regards to Section 5.1

Below we provide Figure 8 which illustrates how ξ_z evolves depending on the $\xi_{\max}$ which is given as input. The parameter $\xi_{\max}$ takes the following 45 values $\{-4, -4 + 0.1, -4 + 0, 2, ..., 5\}$.

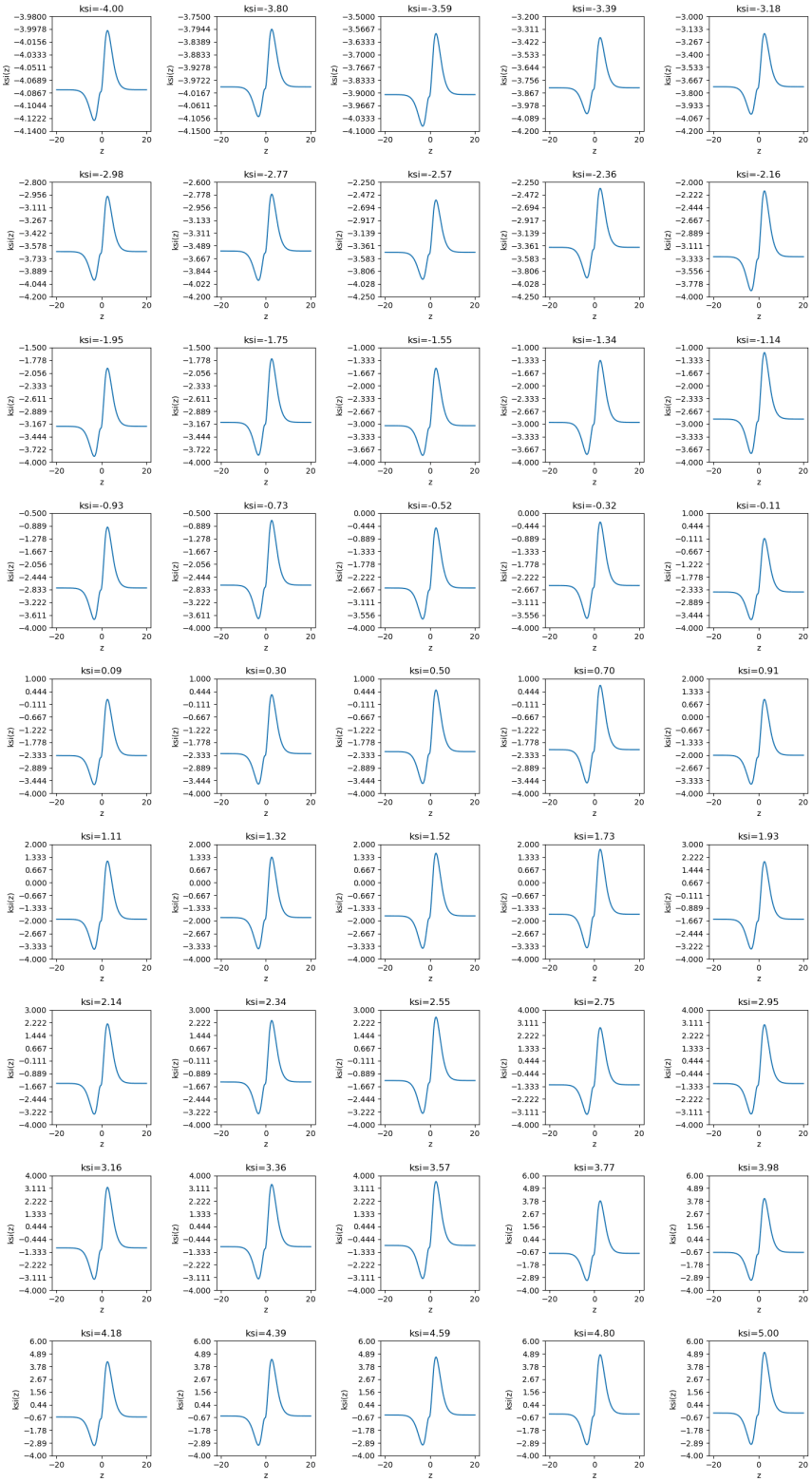

Figure 8: The evolution of ξ_z depending on the value of $\xi_{\max}$.

Appendix H. Additional details with regards to Section 5.3

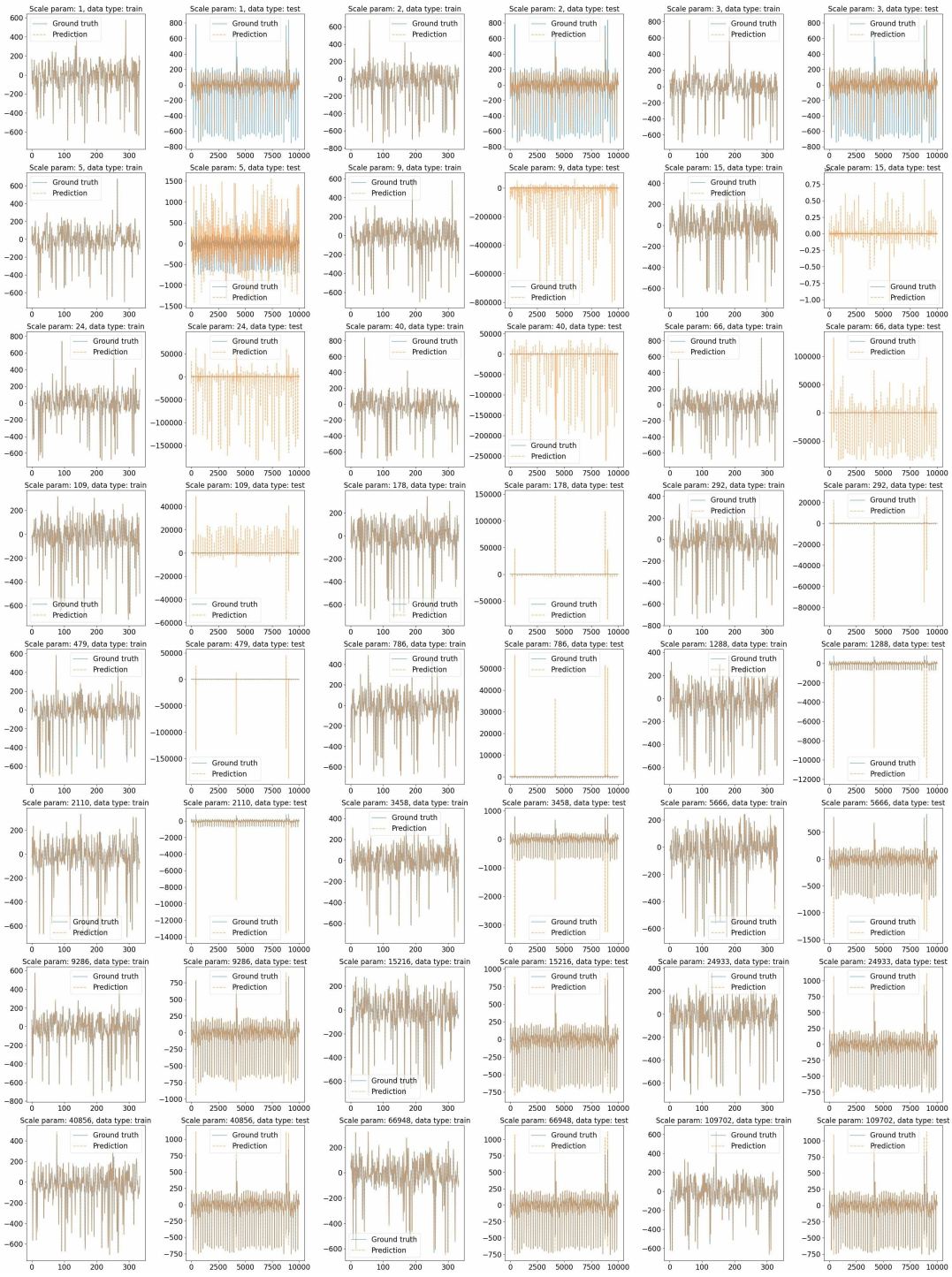

Figure 9: The performance of Gaussian process on train and test data depending on the length scale parameter. First half of the cases.

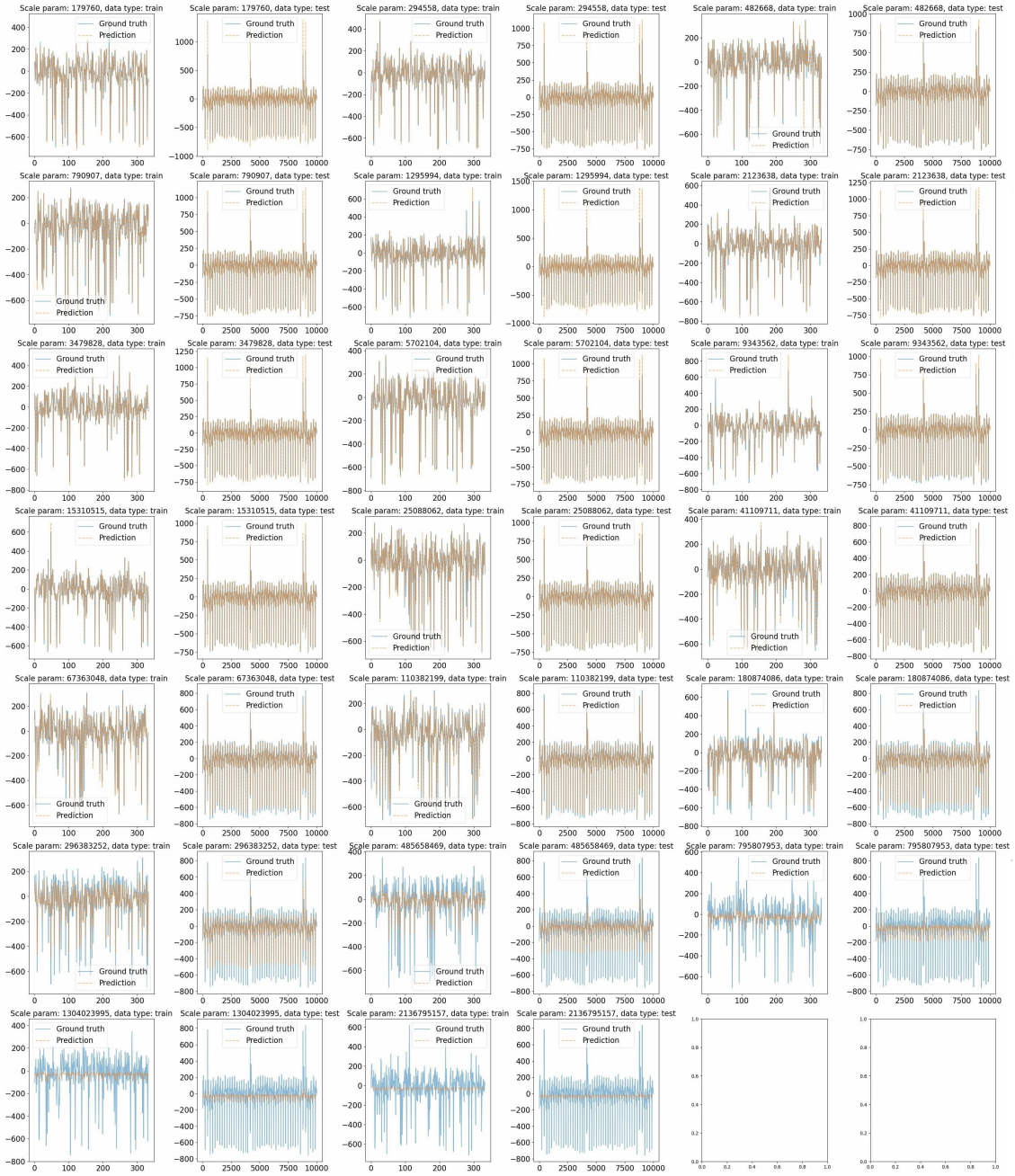

Figure 10: The performance of Gaussian process on train and test data depending on the length scale parameter. Second half of the cases.

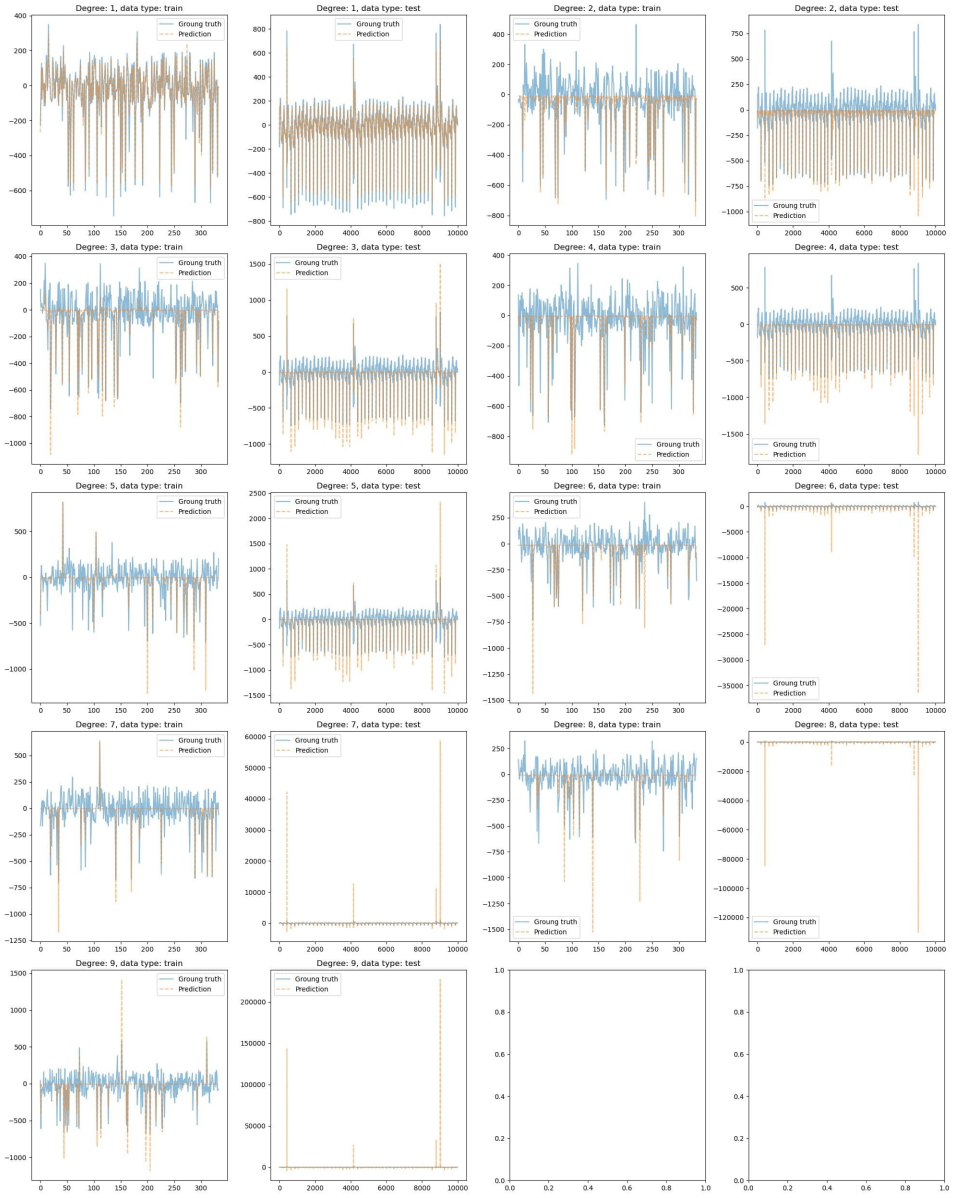

Figure 11: The performance of polynomial kernels on train and test data depending on the degree.

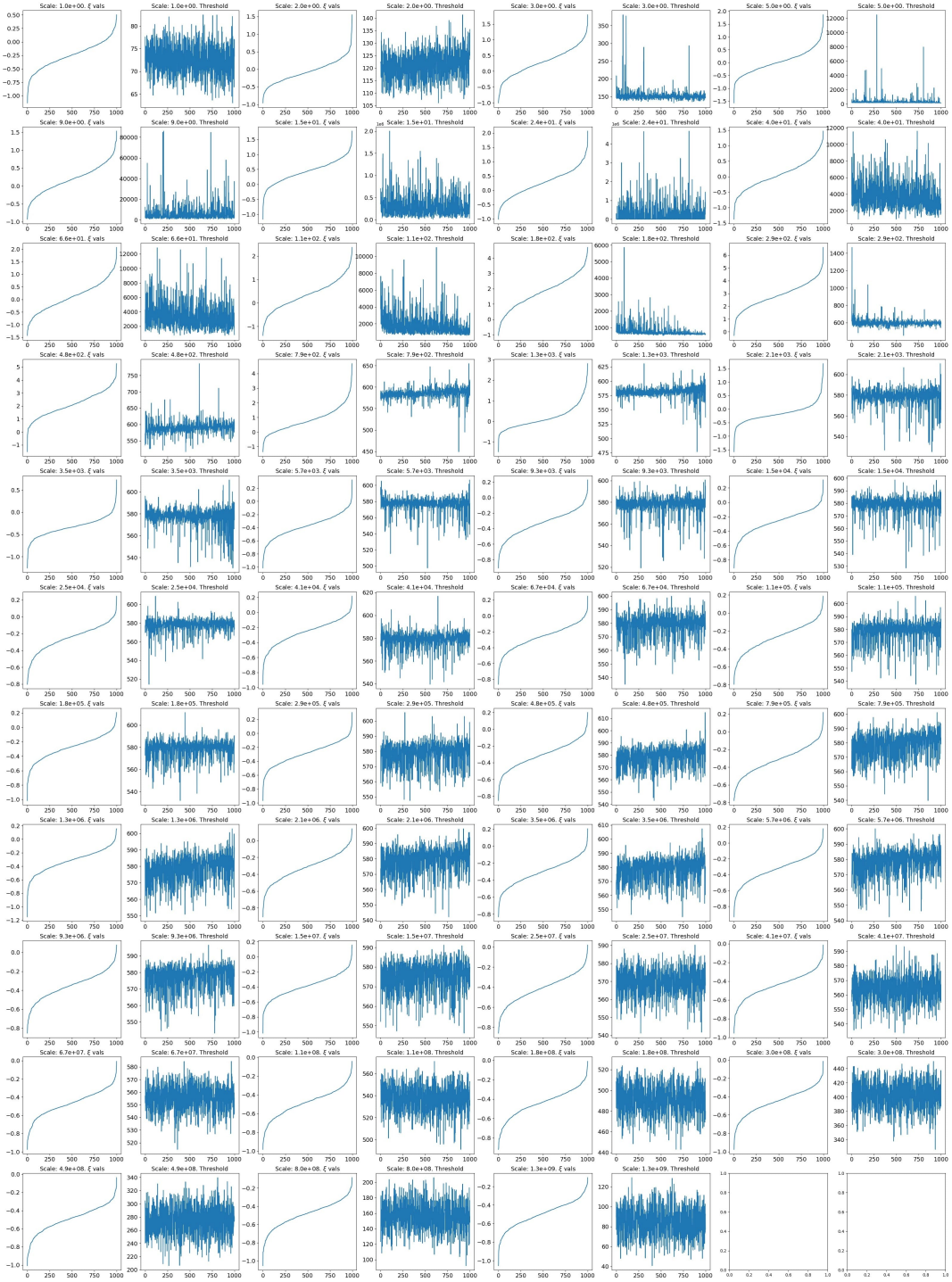

Figure 12: For each length scale parameter of the Gaussian Process, we present the variability (sorted) of the estimated shape parameters across 1000 conditional distributions (defined by the choice of training sets). Jointly, we also present the 97th percentile of the conditional distributions corresponding to each estimated shape parameter.

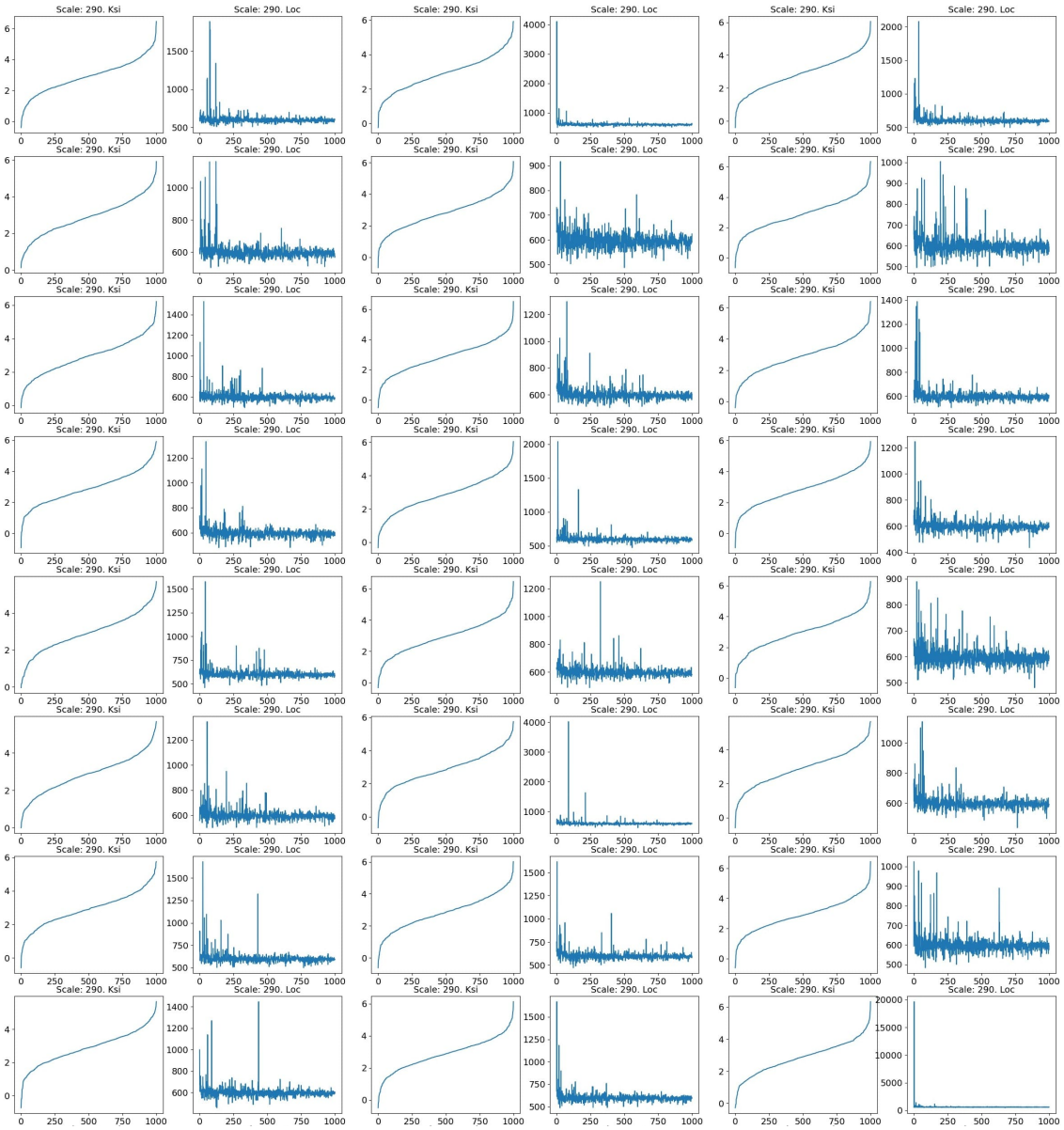

Figure 13: We run 32 times the Gaussian Process experiment for length scale parameter value of 290. On each run, we calculate thresholds (sorted) of the 1000 conditional distributions determined by the 1000 choices of the training set, as well as their corresponding shape tail parameters. We see that higher thresholds correspond to lower shape parameters

Appendix I. Additional DEdH Tail Shape Estimator Experiments

Experimental validation of CTE using the DEdH estimator

A comprehensive set of experimental outcomes has been illustrated in Figure 14.

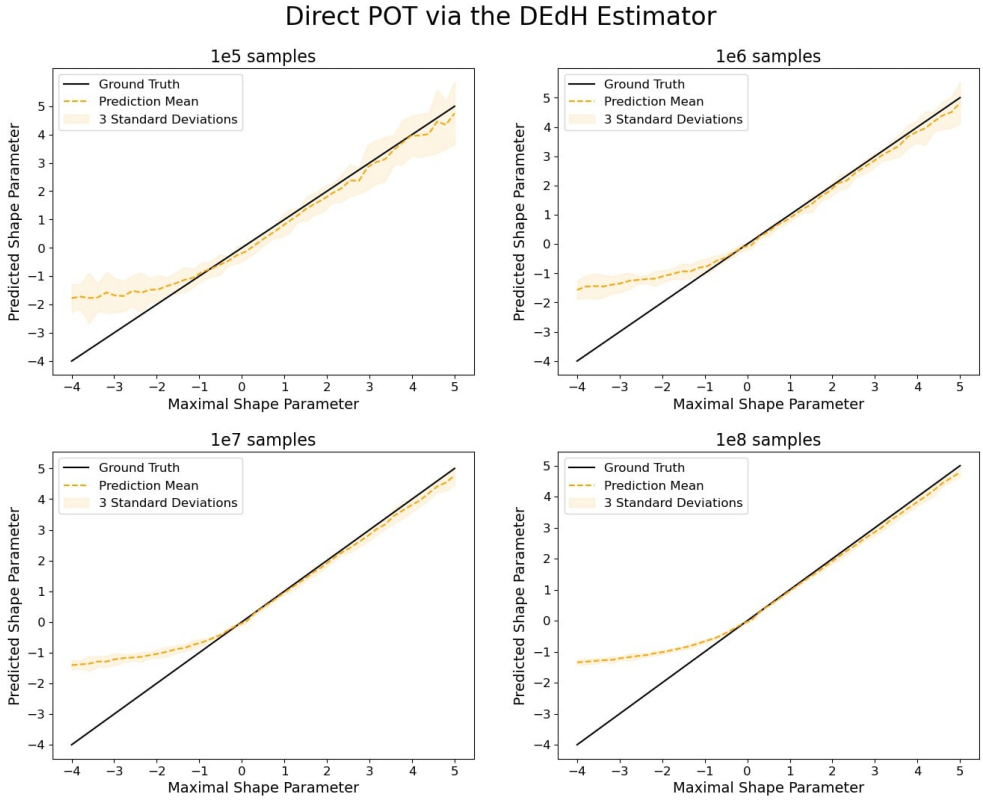

Figure 14: In cases where the maximum tail shape parameter in the mixture of conditional distributions is positive, the estimated shape parameter of the marginal is also positive and equal to this maximal value. However, if this maximum value is negative, the estimated shape parameter is also negative. We utilize the DEdH estimator as our estimator of choice.

Just as in Subsection 5.1.2, the parameter M, delineated in the Subsection 5.1.1, is assigned values from the set $\{10^5, 10^6, 10^7, 10^8\}$. In the context of these experiments, p was set to 10 as a constant across all trials. The experiments were performed repetitively, encompassing a total of 10 runs to capture potential variability and better reflect the stochastic nature of the process.

Upon examining the obtained results, they again seem to align with our initial theoretical expectations. This symmetry in the estimations provides a degree of confidence in the validity of the conducted experiments and the consistency of the underlying theoretical framework.

Applying POT directly when the location of conditional distributions exhibits substantial variability

In the scope of our investigation, we executed experiments akin to those detailed in Subsection 5.2.1, but in this case we utilized the DEdH estimator. Our observations reaffirm that the tail shape of the marginal is subject to incorrect estimations, which can be attributed to the substantial variability in the location of the conditional distributions constituting the marginal.

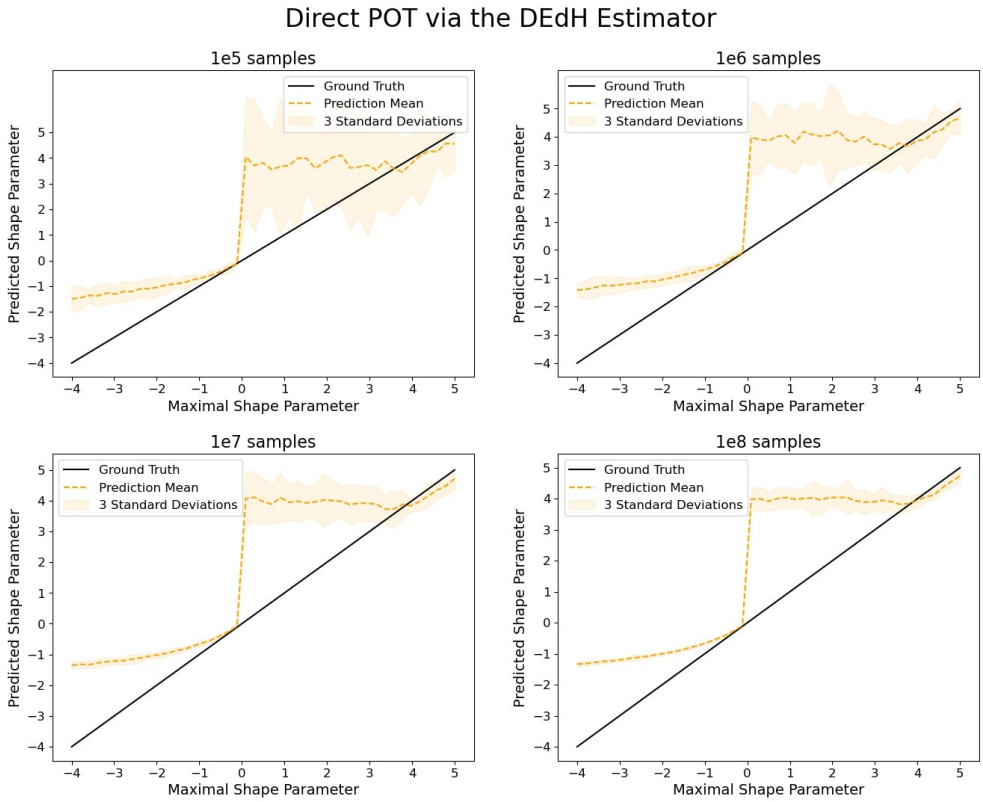

Figure 15: Estimation of the shape parameter of the marginal by direct application of POT. We utilize the DEdH estimator as our estimator of choice.

Enhancing parameter estimation accuracy through the CTE approach

Analogous to the approach taken in Section 5.2.2, we demonstrate here that the Cross Tail Estimation (CTE) effectively alleviates the issue associated with pronounced variation in the locations of conditional distributions that make up the marginal. It should be noted, however, that in this instance, we employ the DEdH estimator.

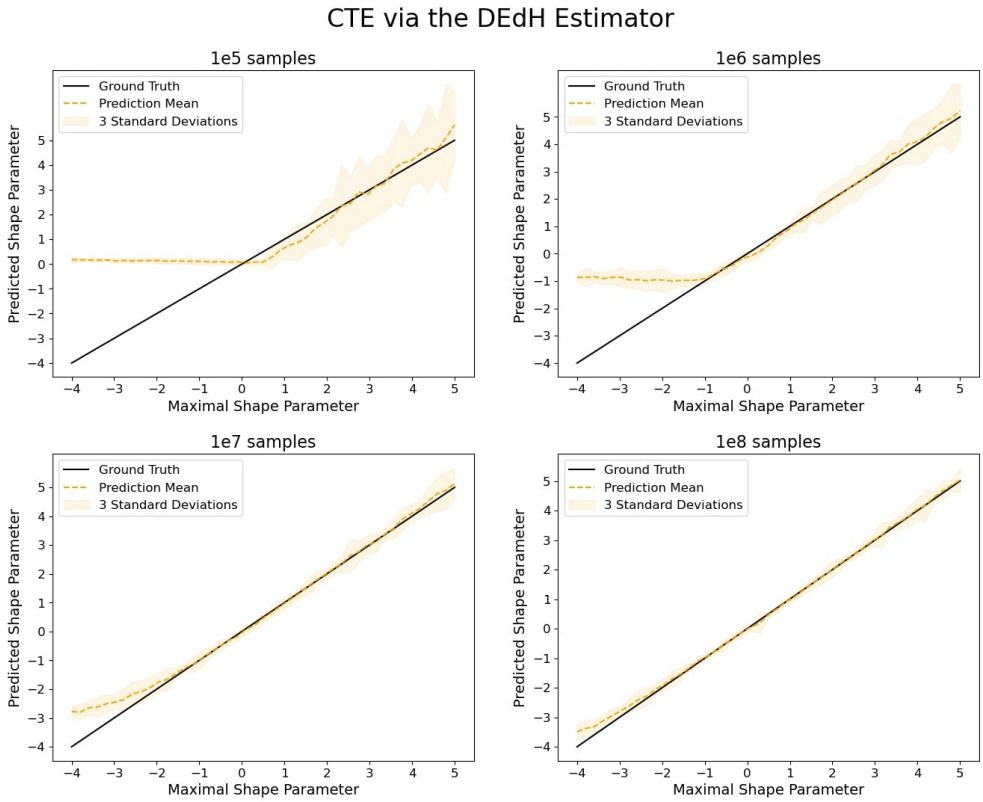

Figure 16: Estimation the shape parameter of the marginal using CTE. We utilize the DEdH estimator as our estimator of choice.

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.

David M. Allen. The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, 16(1):125–127, 1974. ISSN 00401706. URL `http://www.jstor.org/stable/1267500`.

Sylvain Arlot and Pascal Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, 10(10):245–279, 2009. URL `http://jmlr.org/papers/v10/arlot09a.html`.

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

Lucien Birge and Pascal Massart. Estimation of Integral Functionals of a Density. *The Annals of Statistics*, 23(1):11 – 29, 1995. doi: 10.1214/aos/1176324452. URL https://doi.org/10.1214/aos/1176324452.

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.

Richard Davis and Sidney Resnick. Tail Estimates Motivated by Extreme Value Theory. *The Annals of Statistics*, 12(4):1467 – 1487, 1984. doi: 10.1214/aos/1176346804. URL https://doi.org/10.1214/aos/1176346804.

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.

A. L. M. Dekkers, J. H. J. Einmahl, and L. De Haan. A Moment Estimator for the Index of an Extreme-Value Distribution. *The Annals of Statistics*, 17(4):1833 – 1855, 1989. doi: 10.1214/aos/1176347397. URL https://doi.org/10.1214/aos/1176347397.

Arnold L. M. Dekkers and Laurens De Haan. On the Estimation of the Extreme-Value Index and Large Quantile Estimation. *The Annals of Statistics*, 17(4):1795 – 1832, 1989. doi: 10.1214/aos/1176347396. URL https://doi.org/10.1214/aos/1176347396.

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.

Max Kuhn and Kjell Johnson. *Applied predictive modeling*. Springer, New York, NY, 2013. ISBN 9781461468493 1461468493 1461468485 9781461468486. URL http://www.amazon.com/Applied-Predictive-Modeling-Max-Kuhn/dp/1461468485/.

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.

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.

M. Stone. An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):44–47, 1977. ISSN 00359246. URL `http://www.jstor.org/stable/2984877`.

Mervyn Stone. Crossvalidatory choice and assessment of statistical predictions. *Journal of the royal statistical society series b-methodological*, 36:111–133, 1976. URL `https://api.semanticscholar.org/CorpusID:62698647`.

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 28282837, 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.