UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Considerations and Possible Solutions to the Problem of Estimating the Population Minimum with Applications to Earthquake Data

Matheus Henrique Junqueira Saldanha

Master's Dissertation for the Interinstitutional Graduate Program in Statistics (PIPGES)

ICMC USP
SÃO CARLOS

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: ______________________

Matheus Henrique Junqueira Saldanha

Considerations and Possible Solutions to the Problem of Estimating the Population Minimum with Applications to Earthquake Data

Master dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar, in partial fulfillment of the requirements for the degree of the Master Interagency Program Graduate in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Prof. Dr. Adriano Kamimura Suzuki

USP – São Carlos
May 2024

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Defesa de Dissertação de Mestrado do candidato Matheus Henrique Junqueira Saldanha, realizada em 29/04/2024.

Comissão Julgadora:

Prof. Dr. Adriano Kamimura Suzuki (ICMC/USP)

Prof. Dr. Marcio Luis Lanfredi Viola (UFSCar)

Prof. Dr. Diego Carvalho do Nascimento (UDA)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa Interinstitucional de Pós-Graduação em Estatística.

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

S162c

Saldanha, Matheus Henrique Junqueira
Considerations and Possible Solutions to the Problem of Estimating the Population Minimum with Applications to Earthquake Data / Matheus Henrique Junqueira Saldanha; orientador Adriano Kamimura Suzuki. -- São Carlos, 2024.
125 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2024.

1. estimação do endpoint. 2. estimação por máxima verossimilhança. 3. teoria do valor extremo. 4. estimação de quantis extremos. 5. terremotos. I. Suzuki, Adriano Kamimura, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Matheus Henrique Junqueira Saldanha

Considerações e Possíveis Soluções para o Problema da Estimação do Mínimo Populacional com Aplicações em Dados de Terremotos

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Adriano Kamimura Suzuki

USP – São Carlos
Maio de 2024

ABSTRACT

SALDANHA, M. H. J. **Considerations and Possible Solutions to the Problem of Estimating the Population Minimum with Applications to Earthquake Data**. 2024. 125 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

A myriad of physical, biological and other phenomena are better modeled with semi-infinite distribution families, in which case not knowing the populational minimum becomes a hassle when performing parametric inference. This problem has not been directly discussed in the literature thus far, but it is straightforward to devise a maximum likelihood solution (denoted hereafter as "pure MLE"), and endpoint estimators proposed in the literature could also be used. Although endpoint estimators are usually evaluated according to their bias and variance, in this project we argue that these are not adequate metrics, so we discuss and use alternatives. We then propose some solutions of our own, some of them aiming to achieve simplicity in terms of their computational cost, and one method (what we call "maximum likelihood estimation with parameter-dependent support," or MLEPDS) where we estimate the population minimum indirectly, by maximizing a modified likelihood function $L(\cdot \mid \boldsymbol{\theta})$ that shifts the sample by a certain amount depending on $\boldsymbol{\theta}$. Experiments demonstrate that the proposed MLEPDS method outperforms both the pure MLE method as well as the approaches that use endpoint estimators proposed in the literature. In particular, our method offers significantly better results in smaller samples, which will surely be of use to many practitioners out there who have to work with limited data. The dissertation is concluded with an application of the proposed MLEPDS method to predict the maximum magnitude of earthquakes. The probability distribution of earthquake magnitudes is subject to a lot of discussion in the literature; Kijko (2004) describes a few options, which we modify appropriately for use in the MLEPDS method, with which we estimate maximum magnitudes. The regions of Japan, New Zealand, Balkan peninsula and worldwide are analyzed. Experiments show that our method overall gives higher estimates for the maximum magnitude than two other methods inspired by the literature, and also displays an apparent sensitivity in the year-by-year analysis, indicating that it manages to better capture and understand the underlying changes in seismic activity.

Keywords: endpoint estimation, maximum likelihood estimation, extreme value theory, extreme quantile estimation, earthquakes.

RESUMO

SALDANHA, M. H. J. **Considerações e Possíveis Soluções para o Problema da Estimação do Mínimo Populacional com Aplicações em Dados de Terremotos**. 2024. 125 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2024.

Existem inúmeros fenômenos físicos e biológicos (e de outras áreas da ciência) que são inerentemente mais bem modelados se forem utilizadas famílias de distribuições com suporte semi-infinito, caso em que o desconhecimento sobre o mínimo populacional torna-se um obstáculo. Esse problema ainda não foi discutido de forma clara na literatura, mas é razoavelmente direta a construção de uma solução por máxima verossimilhança (denotada “método MLE puro”), e também há na literatura os estimadores de *endpoint*, que também podem ser utilizados. Porém, apesar da literatura comparar tais estimadores por meio de seus *biases* e variância, aqui argumenta-se que essas métricas não sejam adequadas ao propósito dessa dissertação, e portanto são discutidas algumas alternativas. Em seguida, são propostas soluções para o problema, algumas delas objetivando a simplicidade computacional, e uma, que chamamos de “*maximum likelihood estimation with parameter-dependent support*” (ou MLEPDS), que maximiza uma função modificada de máxima verossimilhança $L(\cdot \mid \boldsymbol{\theta})$. Experimentos demonstram que o método MLEPDS consegue resultados melhores do que o método MLE puro, assim como as abordagens que utilizam estimadores de valor extremo propostos na literatura. Em particular, o método MLEPDS oferece melhores resultados em pequenas amostras, o que é de grande utilidade para estatísticos que trabalham com quantidade limitada de dados. Essa dissertação é concluída com uma aplicação do método MLEPDS para predizer a máxima magnitude de terremotos. A distribuição de probabilidade da magnitude dos terremotos é sujeita a muita discussão na literatura; Kijko (2004) descreve algumas opções, as quais modificamos apropriadamente para uso com o método MLEPDS e possibilitar a estimação da magnitude máxima de uma dada região. As regiões do Japão, Nova Zelândia, península Balcânica e do globo terrestre como um todo são analisadas. Os experimentos demonstram que o método MLEPDS, em geral, retorna estimativas mais altas para a magnitude máxima, quando comparado com os dois outros métodos inspirados na literatura, e apresenta também uma sensibilidade maior na análise ano-a-ano, indicando que o método tem capacidade de melhor capturar e entender as mudanças subjacentes que eventualmente ocorrem na atividade sísmica de uma região.

Palavras-chave: estimação do *endpoint*, estimação por máxima verossimilhança, teoria do valor extremo, estimação de quantis extremos, terremotos.

LIST OF FIGURES

Figure 1 – Ideal scenario for performing inference simultaneously over the multiple data sets (in our particular case, sets of execution times). The experimenter has collected data from a number of different phenomena whose underlying probability distribution is believed to belong to a certain family $\mathscr{D}(\alpha, \beta, \gamma, \delta)$. We then would like to infer $\alpha_i, \beta_i, \gamma_i, \delta_i$ for each experiment. For computational maximization of the likelihood, a grid of initial parameters is necessary, and in the ideal scenario, the grid of initial parameters is the same, and works well, for all data sets. 21

Figure 2 – Example of convergence problems faced during our previous project. The histograms on the top show a bad inference result that happened when naively assuming the population minimum to be zero. On the bottom is shown a possibly more reasonable inference obtained after using a better estimate for the population minimum, though it might not be good that the estimated population minimum is so near the sample minimum. 23

Figure 3 – Generalized extreme value distribution for different values of the parameter γ. 29

Figure 4 – Example of the pinball loss function, with $\theta = 0.7$. 32

Figure 5 – (a) Density for the Weibull$(2, 1)$ used in the example. (b) Q–Q plot of the set of negative sample minima $-\overline{m}_i$ against the quantiles of a GEV(-0.72) distribution. The population maximum of the GEV (dashed vertical line) and the corresponding estimate for the negative population minimum (blue cross mark) are shown. 35

Figure 6 – Sample generated from a $20 + \text{Gamma}(3, 1/2)$ distribution, and the corresponding distributions obtained by maximum likelihood estimation by using different estimates m for the population minimum. 56

Figure 7 – Result of applying MLE to estimate the population minimum for a sample taken from a $20 + \text{Gamma}(3, 1/2)$ distribution. 58

Figure 8 – Second example of inference when the population minimum is unknown. The graph shows the area between the inferred distribution and the actual distribution. The areas have been displaced on the y-axis to ease visualization. 60

Figure 9 – The dashed red line shows the truncated version of the underlying actual distribution (black solid line), whereas the blue dotted line shows what can be achieved by MLE using the same family of distributions (gamma) to which the real distribution belongs. 63

Figure 10 – Comparison of the original distribution and its truncated version. 63

Figure 11 – Inferred distributions plotted against the histogram of the sample as well as the original density function that generated the sample. 69

Figure 12 – The low quantile estimators based on the data represented by the histogram in light gray. The data was generated from the density shown as a black line. 73

Figure 13 – Bias distribution of each method on each distribution used for the experiments (gamma above, Weibull on the bottom). Bias distribution here should be seen as the collection of differences $\hat{m} - m$ between the estimated population minimum and the actual minimum. 77

Figure 14 – Histogram of Hellinger (top) and Kantorovich (bottom) distances obtained in experiments concerning the gamma distribution. 79

Figure 15 – Histogram of Hellinger (top) and Kantorovich (bottom) distances obtained in experiments concerning the Weibull distribution. 80

Figure 16 – Histogram of Hellinger (top) and Kantorovich (bottom) distances obtained in experiments concerning the gamma distribution. 81

Figure 17 – Histogram of Hellinger (top) and Kantorovich (bottom) distances obtained in experiments concerning the Weibull distribution. 82

Figure 18 – Average difference between the Hellinger distance obtained by the proposed LIL method and Alves et al.' method, in the form $d_{\text{proposed}} - d_{\text{alves}}$, for each pair of distribution parameters used in the experiments. Negative values correspond to an advantage of our method. 82

Figure 19 – Signed biases of the modified MLEPDS method in the gamma case plotted against the signed biases of the non-modified version. The figure to the right zooms in a smaller region of the cartesian plane, which is not very visible in the figure to the left. 85

Figure 20 – Signed biases of the modified MLEPDS method in the Weibull case plotted against the signed biases of the non-modified version. The figure to the right zooms in a smaller region of the cartesian plane, which is not very visible in the figure to the left. 85

Figure 21 – Shapes of the generalized gamma distribution configurations selected for the experiment. 86

Figure 22 – Proportion of times that the MLEPDS method, with $q = 0.5$, performed better than the pure MLE method, in terms of the parameter squared error. 87

Figure 23 – Proportion of times that the MLEPDS method, with $q = 0.5$, performed better than the pure MLE method, in terms of the population minimum squared error. 87

Figure 24 – Proportion of times that the MLEPDS method, with $q = 0.25$, performed better than the pure MLE method, in terms of the parameter squared error. . 87

Figure 25 – Proportion of times that the MLEPDS method, with $q = 0.25$, performed better than the pure MLE method, in terms of the population minimum squared error. 87
Figure 26 – Illustration of the earthquake catalogs analyzed in this project. Earthquakes are shown as semi-transparent red dots, and plate boundaries are shown in black. 98
Figure 27 – Behavior of the catalogs before and after restricting the earthquakes to being above a certain magnitude threshold. If the dots form a straight line, that means the data follows the Gutenberg–Richter relation. 100
Figure 28 – Maximum magnitude estimates yielded by the proposed MLEPDS method, the method by Alves and Neves (2014), and the method of Kijko (2004) when they are fed with Japanese earthquake data that goes from the year 2000 until the year shown in the x-axis. The solid line shows the observed maximum magnitude on each year, whereas the dots represent the maximum magnitude estimates obtained by each method. 102
Figure 29 – Maximum magnitude estimates yielded by the proposed MLEPDS method, the method by Alves and Neves (2014), and the method of Kijko (2004) when they are fed with New Zealand earthquake data that goes from the year 2000 until the year shown in the x-axis. The solid line shows the observed maximum magnitude on each year, whereas the dots represent the maximum magnitude estimates obtained by each method. 104
Figure 30 – Maximum magnitude estimates yielded by the proposed MLEPDS method, the method by Alves and Neves (2014), and the method of Kijko (2004) when they are fed with Balkan peninsula earthquake data that goes from the year 2000 until the year shown in the x-axis. The solid line shows the observed maximum magnitude on each year, whereas the dots represent the maximum magnitude estimates obtained by each method. 106

LIST OF TABLES

Table 1 – Distance between the estimated distributions and the original underlying distribution for the example given in this chapter's introduction, rounded to three decimal places. 60
Table 2 – Distance between the estimated distributions and the original underlying distribution for the second example mentioned in the text. 61
Table 3 – Distance between the estimated distributions and the original underlying distribution, rounded to three decimal places. 69
Table 4 – Average rank of each inference method on samples simulated from a gamma distribution. "MLEPDS X" stands for the proposed MLEPDS method with its hyperparameter q set to value X, and "LIL" for the estimator based on the law of iterated logarithm. The "errors" rows show the proportion of times that each method failed to provide a valid population minimum; "–" means zero errors, and 0.00 represent that the error rate was rounded to zero, but was not equal to zero. 75
Table 5 – Average rank of each inference method on samples simulated from a Weibull distribution. Other details are as described in Table 4. 75
Table 6 – Interquartile range (IQR) and mean for the boxplots shown in Figure 13. . . . 78
Table 7 – Percentage of experimental trials where the MLEPDS method achieved a better model than the maximum likelihood method, considering the Hellinger and Kantorovich metrics. 78
Table 8 – Percentage of experimental trials where the modified MLEPDS method (with parameters $q \in \{0.25, 0.5, 0.75\}$) achieved a better model than the maximum likelihood method, considering the Hellinger and Kantorovich metrics. . . . 84
Table 9 – Parameters selected for the experiment with the generalized gamma distribution. 86
Table 10 – Overall results without disregarding convergence errors. The proposed methods, with different values of the hyperparameter q, are here compared with the pure MLE method in terms of the respective squared errors (SE). 88
Table 11 – Estimated maximum magnitudes for Japan using the proposed Maximum Likelihood Estimation with Parameter-Dependent Support (MLEPDS) method, the method by Alves and Neves (2014) and the method by Kijko (2004). Values rounded to the second decimal place. 101

Table 12 – Estimated maximum magnitudes for New Zealand using the proposed Maximum Likelihood Estimation with Parameter-Dependent Support (MLEPDS) method, the method by Alves and Neves (2014) and the method by Kijko (2004). Values rounded to the second decimal place. 103
Table 13 – Estimated maximum magnitudes for the Balkan region using the proposed Maximum Likelihood Estimation with Parameter-Dependent Support (MLEPDS) method, the method by Alves and Neves (2014) and the method by Kijko (2004). Values rounded to the second decimal place. 105

LIST OF ABBREVIATIONS AND ACRONYMS

DKW	Dvoretzky–Kiefer–Wolfowitz
ecdf	empirical cumulative distribution function
EVT	extreme value theory
FTG	Fisher–Tippett–Gnedenko
GEV	generalized extreme value (distribution)
iid	independent and identically distributed
IQR	interquartile range
jpd	joint probability distribution
KW	Kantorovich–Wasserstein
KW-CWG	Kumaraswamy complementary Weibull geometric
L-BFGS	limited-memory Broyden–Fletcher–Goldfarb–Shanno
LIL	law of iterated logarithm
MLE	Maximum Likelihood Estimation
OLL-GG	odd log-logistic generalized gamma
PWM	probability weighted moment

CONTENTS

1 INTRODUCTION 19
1.1 Motivation 21
1.2 Objectives 24
1.3 The Problem of Earthquake Prediction 25
1.4 Outline of this Document 25

2 THEORETICAL FOUNDATION 27
2.1 Extreme Value Theory 27
2.2 Quantile Estimation 30
2.3 Extreme Quantile Estimation 33
2.3.1 When There Is Access to Many Sample Maxima 34
2.3.2 When Multiple Sample Maxima Are Not Available 35

3 LITERATURE REVIEW 41
3.1 Endpoint Estimation 42
3.2 Extreme Quantile Estimation 51

4 METHODS TO DEAL WITH THE UNKNOWN POPULATION MINIMUM IN PARAMETRIC INFERENCE 55
4.1 Distance Between Probability Measures 58
4.2 A Few Preliminary Considerations 61
4.3 Proposed Methods for Performing the Inference Procedure 64
4.4 Experiments 73
4.5 A Way to Reduce Instability of the Method 82
4.6 Simulation Experiment With the Generalized Gamma Distribution 84

5 APPLICATION ON EARTHQUAKES 91
5.1 Fundamental Concepts 92
5.2 Data Sources and Configurations 96
5.3 Ensuring Reliability of the Datasets 99
5.4 Estimation of the Maximum Earthquake Magnitude 100
5.4.1 Estimating the Maximum Magnitude for Japan 101
5.4.2 Estimating the Maximum Magnitude for New Zealand 103
5.4.3 Estimating the Maximum Magnitude for the Balkan Region 105

6 CONCLUSION 107
6.1 Acknowledgements 109

BIBLIOGRAPHY 111

APPENDIX A COMPLEMENT TO THE PROOF OF THEOREM 3 123

CHAPTER 1

INTRODUCTION

In parametric inference, it is often the case that the underlying phenomenon is known to have some population minimum, that is, its support is bounded from below. In these cases, it becomes necessary to establish what will be considered as the population minimum, as is often done (somewhat arbitrarily) for phenomena whose population minimum is clearly zero. In more complex cases, this can be done by giving an educated guess for its value, or by adding the population minimum as a parameter to be optimized by maximum likelihood estimation (MLE).

However, the former can be counterproductive, especially when there are multiple datasets to be analyzed, as occurs in the case study that motivates this dissertation (see Section 1.1). The latter option, on the other hand, often estimates the population minimum as being the the sample minimum, which leads to poor performance; this could be seen as overfitting caused by adding an extra parameter to the distribution family (ANDERSON; BURNHAM, 2004).

It is also often the case that the underlying probability distribution is known to have population minimum zero, but with a long and shallow left tail; we demonstrate, further in this document, that this case also causes problems for parametric inference, so much that taking a low quantile as being the population minimum can actually increase the performance of inference, in terms of the difference between the inferred model and the actual model. This project is intended to investigate and attempt to bring better solutions to this problem.

In practice, complicated cases of the aforementioned problem happen frequently. In survival analysis, for example, even though the data is always supported on a semi-infinite interval, there is rarely sufficient evidence that the population minimum is indeed 0, and even if it was, it is likely that the left tail would be long (LAWLESS, 2003).

An exception is data concerning failure count, which most of the times do not fall into the problematic scenarios mentioned above; however, it holds for lifetime data, temperatures, sea and river levels and flow rates (OKUNO; IKEUCHI; AIHARA, 2021), material resistance, fraction apertures (MUSTAFAYEV; HAZLETT, 2019), execution time of programs (SALDANHA;

SOUZA, 2019) etc (LAWLESS, 2003).

For illustration, the time between failures in a supply chain might follow a Weibull with shape $\beta = 10$ and scale $\lambda = 80$, in which case there is almost zero probability ($9.5 \cdot 10^{-5}$, more precisely) of observing a sample minimum lower than 20 in a sample of size 100, so the statistician would never have certainty that the population minimum was indeed zero. Another example would be the time of a flight from Tokyo to Toronto, which clearly has a certain positive minimum value given by the natural limitations of airplane speed. These examples illustrate the two cases distinguished above: one when the underlying random variable has a long left tail; the other, when its support is $[m, \infty)$ for some unknown $m > 0$.

In both cases, one would perform inference using positively supported distributions (e.g., gamma, lognormal, Weibull), maybe after subtracting the experimental data by a certain value $\hat{m}$ that the statistician believes is the theoretical minimum of the underlying distribution. If the underlying distribution has a long left tail, then optimizing the likelihood is made difficult by the necessity of providing it with adequate initial conditions. Of course, simple models can be given initial conditions based on method of moments, but the same cannot be said about more complex models such as generalized versions of gamma and Weibull (STACY; MIHRAM, 1965; MUDHOLKAR; SRIVASTAVA, 1993), nested models (e.g., Kumaraswamy- and logistic-generalized distributions (CORDEIRO; CASTRO, 2011; TORABI; MONTAZERI, 2014)), mixture models (LINDSAY, 1995) etc.

A problem arises here when choosing models by maximizing likelihood or some information criteria for model selection (ANDERSON; BURNHAM, 2004). The chosen model might be far from the actual model due to an unrealistic assumption of the population minimum, or due to difficulties in the mathematical optimization caused by the necessity of initial conditions that could not be anticipated by the experimenter. One consequence would be biased conclusions in favour of simpler models, which are less prone to optimization issues due to bad initial conditions; in other words, it might effectively render usage of complex models useless. We thus argue that the sample should be modified in some way in order to avoid these issues.

In this project we attempt to devise methods with as much theoretical support we can leverage, using results from nonparametric statistics (e.g., law of iterated logarithm), extreme value theory (EVT) and properties of the maximum likelihood estimator. Nonetheless, we are also considering giving room to informal reasoning, in the style of how 25 (or 30) is accepted as a sufficient sample size for the central limit theorem to *usually* hold (WALPOLE *et al.*, 1993), or how the whiskers of a box-plot *usually* serve as a good detector of outliers (HOAGLIN, 2003). We argue that methods to cope with the problem in question must comply with the following objectives:

- avoid the need for case-by-case analysis to “guess” the population minimum;

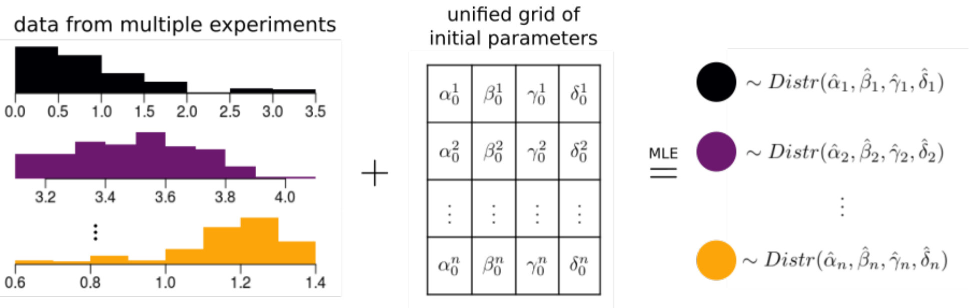

data from multiple experiments
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
3.2
3.4
3.6
3.8
4.0
⋮
0.6
0.8
1.0
1.2
1.4
+
unified grid of
initial parameters
α_0^1 β_0^1 γ_0^1 δ_0^1
α_0^2 β_0^2 γ_0^2 δ_0^2
⋮ ⋮ ⋮ ⋮
α_0^n β_0^n γ_0^n δ_0^n
MLE
=
$\sim Distr(\hat{\alpha}_1, \hat{\beta}_1, \hat{\gamma}_1, \hat{\delta}_1)$
$\sim Distr(\hat{\alpha}_2, \hat{\beta}_2, \hat{\gamma}_2, \hat{\delta}_2)$
⋮
$\sim Distr(\hat{\alpha}_n, \hat{\beta}_n, \hat{\gamma}_n, \hat{\delta}_n)$

Figure 1 – Ideal scenario for performing inference simultaneously over the multiple data sets (in our particular case, sets of execution times). The experimenter has collected data from a number of different phenomena whose underlying probability distribution is believed to belong to a certain family $\mathscr{D}(\alpha, \beta, \gamma, \delta)$. We then would like to infer $\alpha_i, \beta_i, \gamma_i, \delta_i$ for each experiment. For computational maximization of the likelihood, a grid of initial parameters is necessary, and in the ideal scenario, the grid of initial parameters is the same, and works well, for all data sets.

Source – Prepared by the author.

- make it possible to recycle the same grid of initial parameters for performing inference over multiple datasets;
- obtain higher overall performance over multiple models and datasets (performance metrics will be discussed in Section 4.1);
- serve as a good initial value for inferring the population minimum by MLE; and
- increase the computational complexity of MLE as little as possible.

There does not seem to exist approaches, for the problem outlined above, that manage to comply with these objectives. The estimators found either require assumptions in the underlying distribution of the random variable, such as in (VALK; CAI, 2018; DREES *et al.*, 2003; DEMOULIN; GUILLOU, 2018), or they are computationally expensive, often due to usage of resampling techniques (e.g., Dong and Nakayama (2017), Kala (2019), Minasny and McBratney (2006), Liu and Yang (2012)). A more comprehensive overview of related work is deferred to Chapter 3.

1.1 Motivation

In this chapter's opening, we mentioned various problematic scenarios to which we would like to make a contribution to. We ourselves have faced one of these problems in a previous project (SALDANHA, 2020) whose results were published recently (SALDANHA; SUZUKI, 2021). In that project, we were interested in determining the probability distribution for the execution time of computer programs. Obviously, the random variable in question here has a population minimum $m > 0$ that is unknown, since the time to execute each computer instruction is lower bounded by some value; if anything, one could use the speed of light to give a rough estimate of this value, which is not negligible.

The execution time of programs is a key element in certain fields of computer science; one of them is the now widely known field of cloud computing, which most users of smartphones and laptops use in some way or another (TANENBAUM; BOS, 2015). However, running the same program multiple times yields different time measurements, which calls for a probabilistic treatment of the problem.

This is heavily neglected in the literature concerning cloud computing, which often considers only the expected value of execution times (PANDA; JANA, 2015; RODRIGUEZ; BUYYA, 2014; ZUO *et al.*, 2015; XAVIER; ANNADURAI, 2019; TANG *et al.*, 2015), or conveniently assuming that they are normal, uniform or exponential (SUJANA *et al.*, 2019; HAIDRI; KATTI; SAXENA, 2019; SHESTAK *et al.*, 2008; CHEN *et al.*, 2016; ZHENG; SAKELLARIOU, 2013) without further discussing whether these are reasonable assumptions.

The objective of our previous project was to contribute to this field (the stochastic branch of cloud computing) by determining the distribution family that best models execution times. The Weibull distribution is likely the most frequently used for lifetime data (LAWLESS, 2003), because we know a posteriori that it is a suitable model for many lifetime scenarios; in a similar manner, why can we not have a distribution family that is reasonably suitable for most scenarios of execution times? In (SALDANHA; SUZUKI, 2021) we argue that the exponentiated Weibull (MUDHOLKAR; SRIVASTAVA, 1993) is one such family.

A number of problems were faced during the process of finding such a distribution. We had executed different algorithms in different machines, and obtained 37 sets of 1000 execution times, upon which parametric inference had to be performed. Naively considering the population minimum to be zero for all these sets was out of question, as it led the more complex distribution families to poorly fit the data due to convergence issues of the underlying optimization algorithm; this is illustrated in Figure 2. Namely, the generalized gamma (STACY, 1962), exponentiated Weibull (MUDHOLKAR; SRIVASTAVA, 1993), odd log-logistic generalized gamma (OLL-GG) (CORDEIRO *et al.*, 2017), and the Kumaraswamy complementary Weibull geometric (KW-CWG) (AFIFY *et al.*, 2017) distributions, which have 3 or more parameters, suffered from severe convergence problems.

The convergence problems could be solved by looking at each set of 1000 execution times individually, and handcrafting a suitable grid of initial parameters for each of them. We were not interested in such a solution, as it would take a lot of manual effort, and thus would go against the objectives posed in this chapter's opening. We wanted parametric inference to work well for all sets of execution times, using a single standard grid of initial parameters (see Figure 1). In order to do this, it was necessary to deal with the unknown population minimum m.

We could not find, in the literature, a theoretical discussion on this sort of "simultaneous parametric inference" with an existing mathematical language that we could use here. Therefore, in order to use existing concepts, we formulate the problem as multivariate parametric inference, although this might transmit a different idea as to what our objective really is.

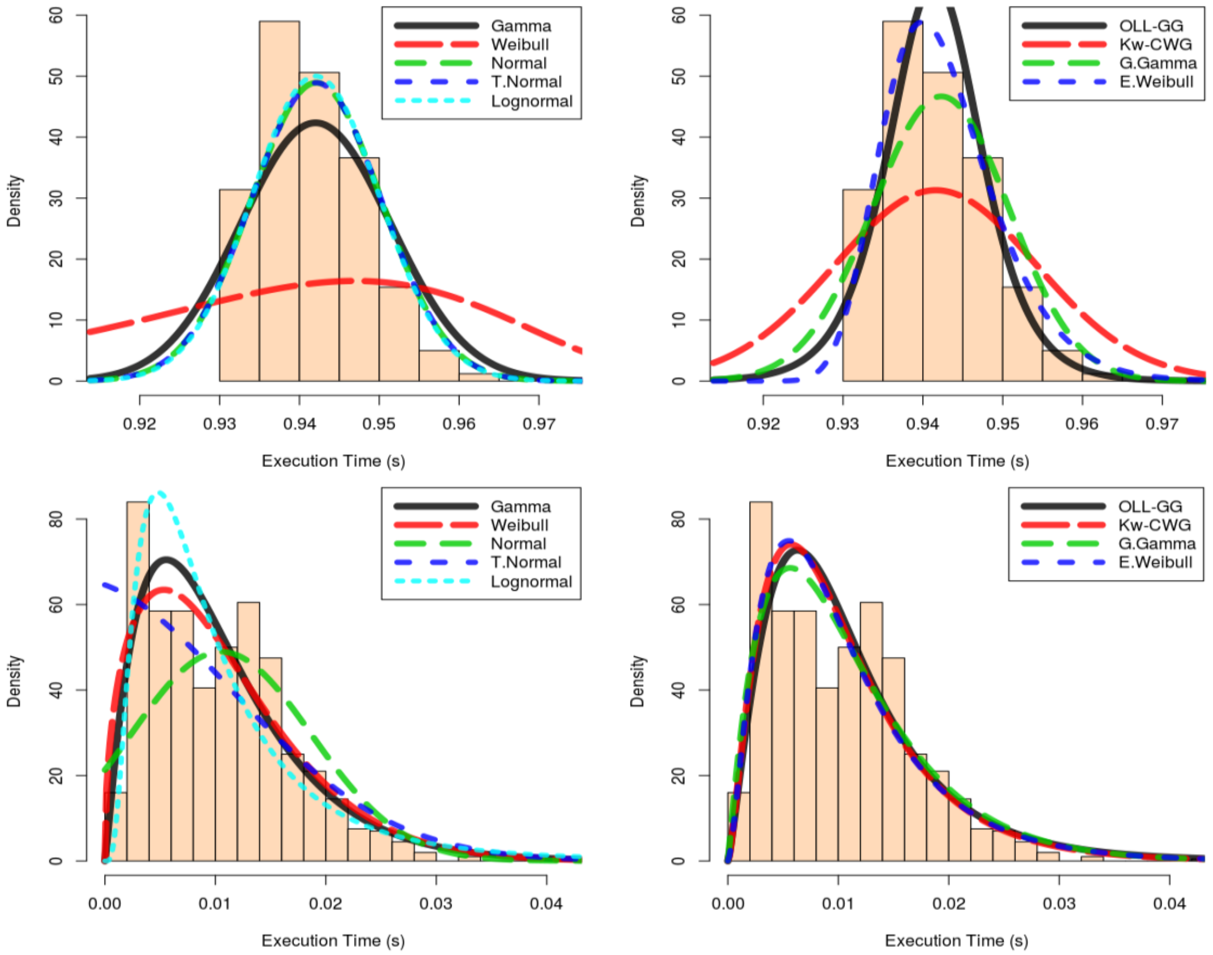


Figure 2 – Example of convergence problems faced during our previous project. The histograms on the top show a bad inference result that happened when naively assuming the population minimum to be zero. On the bottom is shown a possibly more reasonable inference obtained after using a better estimate for the population minimum, though it might not be good that the estimated population minimum is so near the sample minimum.

Source – Prepared by the author.

The 37 sets of 1000 execution times can be organized as a matrix with terms x_{ij} corresponding to the j-th execution time obtained for the i-th program experimented with. Each row i is thus a sample from a certain random variable X_i with density $f_i(x \mid \boldsymbol{\theta}_i)$, each $\boldsymbol{\theta}_i$ possibly having different dimensions.

By assuming that each experiment is independent from each other, and that the execution times obtained in each experiment are independent and identically distributed (iid), the log-likelihood then becomes:

$$
\begin{aligned}
l(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k) &= \log\left(\prod_{1\leq i\leq k} f_i\left(x_{i1},x_{i2},\ldots,x_{in} \mid \boldsymbol{\theta}_i\right)\right) \\
&= \sum_{i=1}^{k} \log\left(f_i\left(x_{i1} \mid \boldsymbol{\theta}_i\right)\cdot f_i\left(x_{i2} \mid \boldsymbol{\theta}_i\right)\cdot\ldots\cdot f_i\left(x_{in} \mid \boldsymbol{\theta}_i\right)\right)
\end{aligned}
$$

$$= \sum_{i=i}^{k} l\left(\boldsymbol{\theta}_i\right),$$

where n is the number of measurements made in each experiment, and k the number of experiments.

Note that in our case we want to use the same distribution family for all experiments, which is equivalent to dropping the subscript in the functions f_i (but keeping the subscripts of $\boldsymbol{\theta}_i$). Since it is true that:

$$\underset{(x,y)}{\arg\max}\,\{a(x)+b(y)\} = \underset{x}{\arg\max}\,\{a(x)\} \times \underset{y}{\arg\max}\,\{b(y)\},$$

then we have:

$$\underset{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k}{\arg\max}\,\{l(\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_k)\} = \underset{\boldsymbol{\theta}_1}{\arg\max}\,\{l(\boldsymbol{\theta}_1)\} \times \underset{\boldsymbol{\theta}_2}{\arg\max}\,\{l(\boldsymbol{\theta}_2)\} \times \cdots \times \underset{\boldsymbol{\theta}_k}{\arg\max}\,\{l(\boldsymbol{\theta}_k)\},$$

which gives the theoretical the basis to state:

1. the optimal parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ can be obtained by optimizing each experiment separately; and
2. the maximum likelihood for using the same distribution family on all experiments is equivalent to the sum of the maximum likelihood obtained for each experiment separately.

This gives a metric to compare the performance of different distribution families on all experiments; it suffices to sum the maximum likelihoods it obtained in each experiment.

It was by this reasoning that we argued in (SALDANHA; SUZUKI, 2021) that the exponentiated Weibull was superior to the other families tested. However, we had used an estimator for the population minimum that might be suboptimal, whereas the theoretical formulation above assume that the support of the distributions are correct. Therefore, having a good estimator for the population minimum is imperative, and in this project we aim to investigate this issue more thoroughly.

1.2 Objectives

Considering the discussion above, the general objective of this dissertation is thus to investigate the problem of performing parametric inference while (ideally) fulfilling two objectives simultaneously: (i) inferring the distribution of the data, and (ii) estimating a left endpoint that is as close as possible to the real population minimum; while also paying attention to the computational cost of each solution to the problem.

The specific objectives are: (i) devise solutions to the estimation problem mentioned in the general objective, (ii) find methods in the literature that could be used, even if adaptations are necessary, (iii) devise a way to compare different methods, preferably with a metric that can correctly classify the different methods in terms of how well they achieve the two objectives mentioned in the general objective,(iv) perform simulation experiments comparing the proposed solutions with methods taken from the literature, evaluating the results using the metrics devised in item (iii), and (v) use the methods that had best performance in the experiments to analyze and forecast earthquake magnitudes.

1.3 The Problem of Earthquake Prediction

Frequent seismic events of considerable magnitude lead to significant casualties and substantial economic damages. Even with thorough preparation, developed nations remain vulnerable to the devastating impacts of earthquakes measuring M8.5[1] or higher (STEIN; WYSESSION, 2003).

The inability to accurately forecast earthquakes stems from several factors. Primarily, our understanding of the mechanisms behind strain energy accumulation on faults remains limited. Fundamental processes like mantle convection are not fully comprehended, with only a few compelling geoscientific theories, such as the "plume tectonics" theory, offering some insight (YUEN *et al.*, 2007; LARSON, 1991).

Given the complex nature of earthquakes, a stochastic approach is essential for analysis and potential forecasting. This makes earthquakes one of the most important use-cases of methods to estimate extreme values while also inferring the underlying distribution. Therefore, this dissertation also aims to apply the investigated methods to earthquake data.

1.4 Outline of this Document

This document is organized as follows. We begin by giving an overview, in Chapter 2, of the main research fields related to our project, namely extreme value theory (Sections 2.1 and 2.3) and quantile estimation (Section 2.2), although they are both tightly interconnected. In Chapter 3 we present all the relevant related work we could find; although the problem of parametric inference with unknown population minimum has not been discussed directly yet (to the best of our knowledge), we were able to find a decent number of proposed techniques that could be used to solve the problem. Is any of these methods better than the natural solution to the problem, that is, better than inferring the population minimum by MLE? Also, what performance metric should we use for such comparison? We tackle these questions in Chapter 4, where we also discuss our own proposed solutions to the problem. More specifically, in Sections 4.1 and

---

[1] M stands for magnitude, with no specification of the particular method to calculate it.

4.2 we present and discuss the performance metrics we intend to use; our proposed methods are discussed in Section 4.3; and preliminary experiments are given in Section 4.4. Chapter 5 shows applications of the estimators discussed throughout the dissertation to the field of seismology. Finally, concluding remarks are given in Chapter 6.

CHAPTER 2

THEORETICAL FOUNDATION

This chapter gives an overview of theoretical topics and research fields that are related to the main objective of this dissertation which revolves mainly around extreme value theory, the field that investigates the asymptotic distribution of the sample minimum, which leads to interesting subfields such as extreme quantile estimation and even estimation of the population minimum (also called endpoint estimation).

2.1 Extreme Value Theory

Fisher and Tippett (1928) and Gnedenko (1943) brought forth important asymptotic results concerning sample minima and maxima. What is known as the Fisher–Tippett–Gnedenko (FTG) theorem gives an asymptotic distribution to both extreme order statistics, in a similar fashion as the central limit does to the sample mean. This theorem gives hope that we are able to study (and maybe foresee) extreme events before they are observed in practice, which is of particular importance when it comes to natural disasters such as earthquakes and floods. Extreme value theory (EVT) is the research field that pursues extensions to the FTG theorem and applications of it to practical applications.

Let $\{X_i\}, i \in \mathbb{N}$, be a sequence of independent and identically distributed (iid) random variables. Denote the sample maxima and sample minima as:

$$\begin{aligned} \overline{m}_n &= \min\{X_1, \ldots, X_n\} \text{ and} \\ \overline{\mathcal{M}}_n &= \max\{X_1, \ldots, X_n\}. \end{aligned}$$

The n subscript will only be used when we must emphasize that they form the sequences $n \mapsto \overline{m}_n$ and $n \mapsto \overline{\mathcal{M}}_n$.

If the sample $\{X_1, \ldots, X_n\}$ is iid with cdf $X_i \sim F_X(x)$, we define the population minimum and maximum as:

$$\begin{aligned} m &= \inf\{x : F_X(x) > 0\} \text{ and} \\ \mathcal{M} &= \sup\{x : F_X(x) < 1\}, \end{aligned}$$

which might not exist (be infinite). If we have $F_X(x)$, then we know the distribution of $\overline{m}_n$ and $\overline{\mathcal{M}}_n$ (MOOD, 1950):

$$\begin{aligned} F_{\overline{m}_n}(x) &= 1 - [1 - F_X(x)]^n, \text{ and} \\ F_{\overline{\mathcal{M}}_n}(x) &= F_X(x)^n. \end{aligned}$$

Since any cdf is bounded within $[0, 1]$ and is non-decreasing, by taking $n \to \infty$ we end up with a degenerate cdf. If the population minimum is finite, $F_{\overline{m}_n}$ tends to a step function $\mathrm{I}[x < m]$, and if the populational maximum exists, $F_{\overline{\mathcal{M}}_n}$ becomes $\mathrm{I}[x < \mathcal{M}]$ as the sample size increases.

The above result is not very informative, however. It states that at some point the sample maximum and minimum will converge to the population maximum and minimum, but gives no insight on what happens to $\overline{m}_n$ and $\overline{\mathcal{M}}_n$ in the interim, that is, during their route to fully converging to the population values. It is possible to find an asymptotic distribution for $\overline{m}_n$ and $\overline{\mathcal{M}}_n$, but they need to be conveniently shifted and rescaled with two sequence of constants a_n and b_n. This is where the FTG theorem comes in, and is stated in the following theorem (RESNICK, 2008; LEADBETTER; LINDGREN; ROOTZÉN, 1983; HAAN; FERREIRA, 2006; BEIRLANT *et al.*, 2004).

Theorem 1 (Fisher–Tippett–Gnedenko)**.** *Let $\{X_i\}, i \in \mathbb{N}$, be iid random variables with cdf $F_X(x)$, and let $\overline{\mathcal{M}}_n$ be the sample maximum for $X_1, \ldots, X_n$, for each $n \in \mathbb{N}$. If there are sequences of constants $a_n > 0$ and $b_n \in \mathbb{R}$ such that $(\overline{\mathcal{M}}_n - b_n)/a_n$ converges in distribution, then it converges to a distribution with one of the following cdf's:*

$$\begin{aligned} (\textit{Fréchet distribution}) \quad G_1(y \mid \gamma) &= \exp\{-y^{-\gamma}\}\,\mathrm{I}[y > 0], && \textit{with } \gamma > 0 \\ (\textit{reverse Weibull distr.}) \quad G_2(y \mid \gamma) &= \exp\{-|y|^{\gamma}\}\,\mathrm{I}[y \le 0] + \mathrm{I}[y > 0], && \textit{with } \gamma > 0 \\ (\textit{Gumbel distr.}) \quad G_3(y) \quad &= \exp\{-e^{-y}\}, \end{aligned}$$

except for an extra pair of scale-location parameters (a, b) which depend on the sequences $\{a_n\}$ and $\{b_n\}$.

The parametrization for the Fréchet distribution used above is the same as originally proposed in (FRÉCHET, 1927). The reader may see G_2 in the literature being mistakenly claimed to be a Weibull distribution, but it is actually the cdf for $-Y$ where Y has a Weibull distribution:

$$F_Y(y\,|\,\alpha,\beta) = \left[1 - \exp\left\{-\left(\frac{y}{\alpha}\right)^{\beta}\right\}\right] \mathrm{I}\,[y > 0] + 0 \cdot \mathrm{I}\,[y \leq 0]\,,$$

then the cdf of $-Y$ is $F_{-Y} = 1 - F_Y(-y)$, which results in:

$$F_{-Y}(y\,|\,\alpha,\beta) = \exp\left\{-\left(\frac{-y}{\alpha}\right)^{\beta}\right\} \mathrm{I}\,[y \leq 0] + 1 \cdot \mathrm{I}\,[y > 0]\,,$$

now since $-y = |y|$ whenever $y \leq 0$, and by doing the reparametrization $\alpha = 1, \beta = \gamma$, we end up with:

$$F_{-Y}(y\,|\,\gamma) = \exp\left\{-|y|^{\gamma}\right\} \mathrm{I}\,[y \leq 0] + \mathrm{I}\,[y > 0]\,,$$

which is the same as G_2 in Theorem 1.

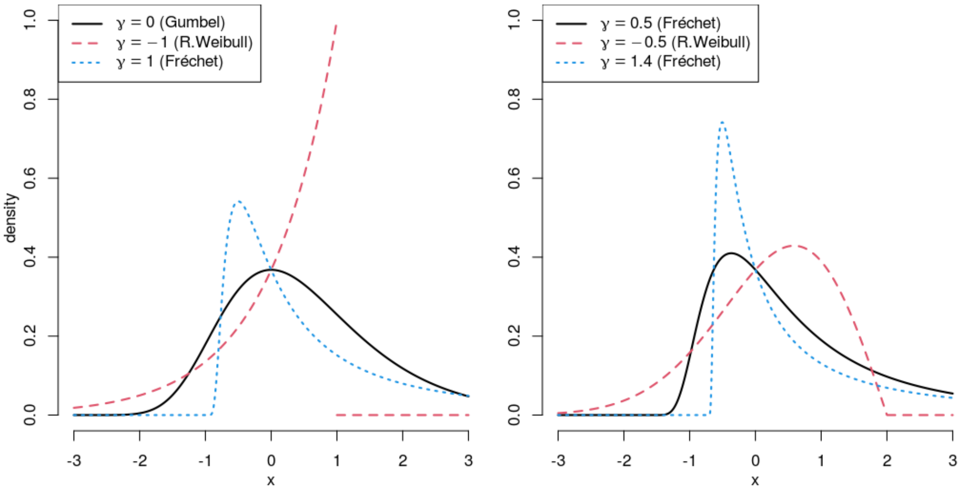


Figure 3 – Generalized extreme value distribution for different values of the parameter γ.

Source – Prepared by the author.

The confusion is only natural, because Theorem 1 could be stated using the Weibull distribution, instead of its reverse counterpart. This is often not done in the literature because the Fréchet, reverse Weibull and Gumbel distributions, in the parametrization given in Theorem 1, can be bundled up under one single-parameter distribution with cdf:

$$G_{\gamma,a,b}(x) = \exp\left\{-\left[1 + \gamma\left(\frac{x-a}{b}\right)\right]^{-1/\gamma}\right\}, \text{ with } 1 + \gamma\left(\frac{x-a}{b}\right) > 0, \tag{2.1}$$

which is called generalized extreme value (GEV) distribution. This is convenient at times, and it is also remarkable that the limiting distribution of the sample maximum has one parameter only, apart from the location and scale; it is thus just a little more complex than the limiting distribution given by the central limit theorem, which only has the location and scale parameters.

The GEV distribution is not defined for $\gamma = 0$, but the limit when $\gamma \longrightarrow 0$ exists and will be equal to the Gumbel distribution (HOSKING; WALLIS; WOOD, 1985; HAAN; FERREIRA, 2006). Figure 3 illustrates the GEV distribution for some values of its parameter. This common parametrization is due to Jenkinson (1955), although it might have been proposed earlier.[1] As can be seen above, the support of the distribution changes significantly with the parameter γ, which can be very cumbersome for inferring the parameter γ computationally.

As a brief comment on the proof of Theorem 1, we note that the proofs used to be complicated and given separately for each of the three types of limiting distributions, until the thesis of Laurens de Haan (HAAN, 1970), which unified and simplified the proof for all three types.

The proof uses a few important results: i) the convergence of a sequence of functions f_n to some g is equivalent to the convergence of the inverses f_n^{-1} to g^{-1}, under conditions that are always satisfied when dealing with cdfs; ii) a function is Riemann-integrable if and only if the set of discontinuities has measure zero; and iii) every monotonic function is Riemann-integrable. It follows immediately from items ii) and iii) that no cdf can have a set of discontinuities with measure larger than zero. These two items reside in the mathematical field of real analysis, and the reader is invited to check these results in Theorem 7.2.8 and Result 7.3.12 in (BARTLE; SHERBERT, 2011).

Theorem 1 concerns sample maxima only. It can be easily adapted to sample minima, which is of larger interest to us, by noting the relation:

$$\min\{X_1, \ldots, X_n\} = -\max\{-X_1, \ldots, -X_n\}.$$

Since the distribution of the maximum is given by Theorem 1, the distribution of the minimum must be the reverse of these distributions, but by manipulating their scale parameter b, we can safely state that the limiting distributions for the sample minimum are the same as the maximum.

2.2 Quantile Estimation

The population minimum can be found by estimating the q-quantile of the underlying continuous distribution, with $q \longrightarrow 0^+$. For extremely low quantiles, the field is actually called “extreme quantile estimation.” As the name implies, it has deep connections with extreme value theory. However, since we have a more modest goal than finding the population minimum per se, it is useful to also present the fundamentals of standard quantile estimation procedures; after all, knowledge of, say, the 0.05-quantile is also very helpful for loosely estimating the population minimum.

---

[1] According to Haan and Ferreira (2006), credits are also due to Mises (1936), but we could not confirm this, and the fact that this is not mentioned in (FALK; MAROHN, 1993), makes us uneasy about giving undue credits. It is clear, however, that Richard Edler von Mises made remarkable contributions to the field of extreme value theory.

A q-quantile of a continuous distribution is defined as $F^{-1}(q)$, where F^{-1} is the inverse of F. If F is not injective, then F^{-1} might return sets of quantiles, rather than unique values. When this is inconvenient for mathematical proofs, it is common to define the quantile as (SERFLING, 1980):

$$F^{-1}(q) := \sup \{x \mid F(x) \leq q\} .$$

One property here is that $F^{-1}(0)$ will yield the population minimum, which is more convenient to us. It is also possible to define it as $\inf\{x \mid F(x) \geq q\}$, in which case the relevant property is that $F^{-1}(1)$ yields the population maximum. In any case, in this project we will usually be working with strictly increasing cumulative distribution functions, so the inverse exists and need not be redefined as above.

Given an iid sample $X_1, \ldots, X_n$, define the empirical cumulative distribution function (ecdf) as:

$$F_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}\{X_i < x\} .$$

In other words, $F_n(x)$ returns the proportion of the sample that is lower than x, for each $x \in \mathbb{R}$. As the name suggests, the function $F_n(x)$ is an estimate of the underlying cdf $F(x)$, and we know it is consistent, thanks to the Glivenko–Cantelli theorem (VAPNIK, 1998).

Because of this, the most immediate estimator for a quantile is given by the inverse of F_n, so that the q-quantile would be given by $F_n^{-1}(q)$. Since F_n is a step function, we again have the problem of non-invertibility; in any case, any value within the inverse image set $F_n^{-1}(q)$ will be an estimator for q. This is too rudimentary, though, and it can be improved a lot by applying interpolation or regression techniques upon F_n before inverting it (AVRAMIDIS; WILSON, 1998; GLASSERMAN, 2013). This same idea is also applicable for extreme quantiles, but its performance can be very disappointing; fortunately, extreme value theory offers some results that are used to improve it, and they will be mentioned in a later opportunity.

There is one more fundamental topic in quantile estimation that must be mentioned. Recall that the sample mean is the value with the smallest squared distance from all sample points, that is:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \arg\min_{\xi} \sum_{i=1}^{n} (X_i - \xi)^2. \tag{2.2}$$

To see this, take the derivative of the function being minimized:

$$f'(\xi) = 2n\xi - 2\sum_{i=1}^{n} X_i,$$

which will be zero if and only if $\xi = (1/n)\sum_{i=1}^{n} X_i$. Since $f(\xi)$ is $+\infty$ when $\xi \to \pm\infty$, we conclude that the sample mean minimizes Equation (2.2).

In a similar fashion, the sample median minimizes the absolute distance from all points:

$$X_{((n+1)/2)} \vee \frac{\left(X_{(n/2)} + X_{((n/2)+1)}\right)}{2} = \arg\min_{\xi} \sum_{i=1}^{n} \left|X_i - \xi\right|, \tag{2.3}$$

which gives the peculiar derivative:

$$f'(\xi) = -\sum_{n=1}^{n} \mathrm{I}\left\{X_i < \xi\right\} + \sum_{n=1}^{n} \mathrm{I}\left\{X_i > \xi\right\},$$

where we are defining the derivative of $|x|$ at $x = 0$ as being zero. The derivative is the number of sample points X_i located to the right of ξ, minus the number of points to its left. As a consequence, the values of ξ that minimizes Equation (2.3) are those that divide the data points in equal parts to its left and right. If n is odd, then ξ is unique and equal the median (this is one convenient aspect of defining the derivative of $|x|$ as zero in its non-differentiable point). If n is even, there will be a range of possible ξ, but it will contain the sample median, so our original assertion is proved.

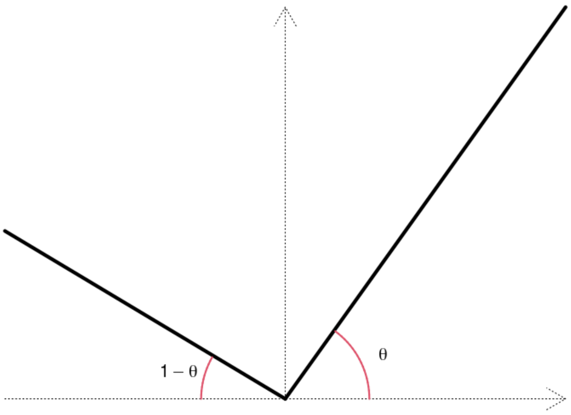


Figure 4 – Example of the pinball loss function, with $\theta = 0.7$.

Source – Prepared by the author.

The above problems can be generalized to minimizing:

$$\arg\min_{\xi} \sum_{i=1}^{n} \omega(X_i - \xi), \tag{2.4}$$

where ω is a weight function which assume non-negative values. In particular, and of most importance here, Koenker and Bassett (1978) introduced the *pinball loss function*, defined as:

$$\omega(y) = \begin{cases} (1-\theta)(-y), & \text{if } y < 0, \\ \theta y, & \text{if } y \geq 0, \end{cases}$$

whose general form is shown in Figure 4. One importance of this function is that, as the name suggests, it gives us a loss metric for quantile estimators. Minimizing $\sum_{i=1}^{n} \omega(X_i - \xi)$ is an estimate for the sample θ-quantile, and it is consistent for the population θ-quantile. It is mostly helpful in the area of quantile regression, where the quantile is a function of some dependent variables, so $X_i - \xi$ in Equation 2.4 is substituted by $Y_i - \xi(\boldsymbol{X}_i, \boldsymbol{\beta})$, and ξ is selected from the parametric family of functions with parameter $\boldsymbol{\beta}$ (KOENKER; HALLOCK, 2001).

2.3 Extreme Quantile Estimation

When it comes to extreme quantile estimation, methods based on the extreme value theory (EVT) seem to be the mainstream within the literature. Here, the most favorable scenario occurs when:

1. we are in possession of k samples $\mathcal{S}_i = \{X_{i,1}, X_{i,2}, \ldots, X_{i,n}\}$, $i \in \{1, \ldots, k\}$, each of which has a sample maximum $\overline{\mathcal{M}}_i = \max \mathcal{S}_i$, and k is sufficiently large to allow parametric inference over the set of sample maxima $\{\overline{\mathcal{M}}_i \mid i = 1, \ldots, k\}$; and

2. each sample $\mathcal{S}_i = \{X_{i,1}, X_{i,2}, \ldots, X_{i,n}\}$ is large enough for the asymptotic result given by the FTP theorem (Theorem 1) to hold.

This is often the case when the random variable in question concerns a phenomenon that is measured multiple times per year, and there is interest in determining the behaviour of its maximum value per year. One example is the intensity of river flow, whose maximum is used to guide measures to prevent floods (OKUNO; IKEUCHI; AIHARA, 2021). Naturally, if one has a large single sample, it can be split into k samples of equal size n, and depending on the particular values of k and n it will also meet the two conditions above.

2.3.1 When There Is Access to Many Sample Maxima

In this favorable situation, we can fit a generalized extreme value distribution (see Equation (2.1)) over the set of sample maxima $\{\overline{\mathcal{M}}_i\}$, thus finding the corresponding shape, scale and location parameters, with which the population maximum can be found. Even simpler, one might use Q–Q plots to achieve the same goal. Take the quantile–quantile correspondence:

$$\left(\overline{\mathcal{M}}_{(i)}, F_\gamma^{-1}\left(\frac{i}{k+1}\right)\right), \quad \text{with } i = 1, \ldots, k,$$

where $\overline{\mathcal{M}}_{(i)}$ are the order statistics for the set of sample maxima, and F_γ^{-1} is the quantile function for a GEV distribution with parameter γ.

All that must be done here is to find γ that makes all these points stand in a straight line, which can be done manually when convenient. The intercept and slope of this line will automatically give your estimates $\hat{a}$ and $\hat{b}$ for the normalizing constants mentioned in Theorem 1 (BEIRLANT *et al.*, 2004).

As an example, consider we have 20 samples from a Weibull distribution with shape $\alpha = 2$ and scale $\beta = 1$ (see Figure 5a), each sample having size 100. We are interested in finding extreme low quantiles and the population minimum, so we take the sample minimum $\overline{m}_i$ for each of these samples; actually, we consider their negatives $-\overline{m}_i$ so that we can continue working with the theory for sample maximum (see Section 2.1).

We then find γ that optimizes the Q–Q plot of the negative sample minima against the quantiles of GEV(γ), where we obtained $\gamma = -0.72$ (corresponding to a reverse Weibull distribution), as shown in Figure 5. Linear regression by least squares yields a line with intercept $\hat{a} = -0.108$ and slope $\hat{b} = 0.0659$, which are our estimates for the location and scale parameters, respectively, of the GEV distribution, in case they are needed. Note in Equation (2.1) that for $\gamma = -0.72$ the support of the distribution satisfies:

$$1 - 0.72\left(\frac{x-a}{b}\right) > 0 \implies x < \frac{b}{0.72} + a \approx 1.39b + a,$$

which gives the population maximum of the corresponding GEV distribution. Assuming $b = 1$ and $a = 0$, if we follow the regression line up to point $x = 1.39$, we obtain in the y-axis the negative estimate for the population minimum for the set of minima $\{\overline{m}_i\}$, which in this case is 0.0163 (with sign already reversed). Since the population minimum of the sample minimum is the same as the population minimum of the distribution, 0.0163 is also our estimate for the population minimum of the samples generated (the correct value would be zero).

Here we were fortunate that the γ we found corresponded to a reverse Weibull, which has a population maximum. However, if it happened to be a Fréchet or Gumbel, such a direct way to determine the population maximum would not be available. Also, note that the estimator was somewhat accurate.

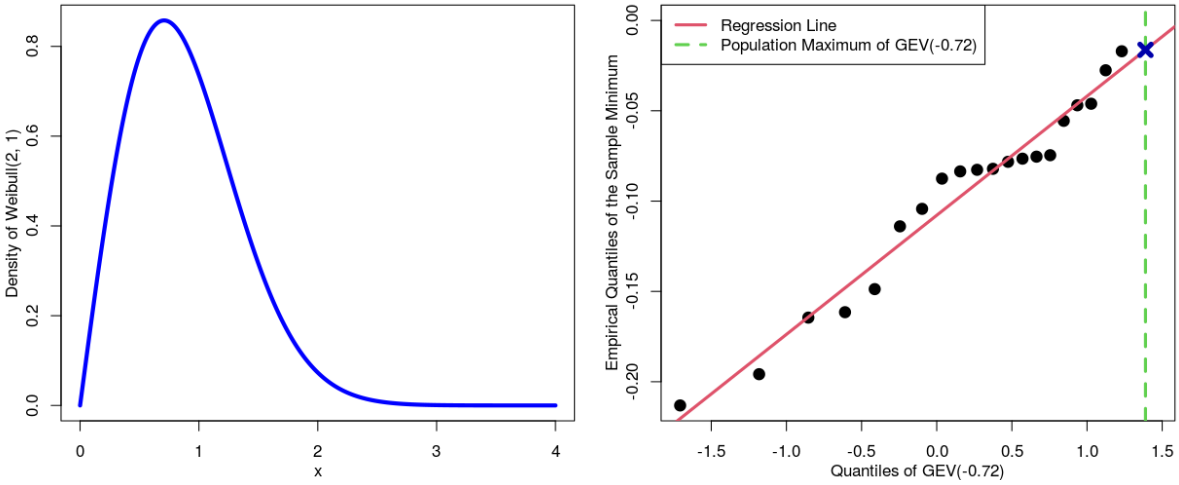
The Weibull$(2, 1)$ has cdf (this particular case is also a Rayleigh distribution):

$$F(x) = 1 - e^{-x^2},$$

so that the distribution of the minimum is:

$$1 - (1 - F(x))^n = 1 - e^{-nx^2} = 1 - e^{-(x\sqrt{n})^2},$$

which is the cdf for a Weibull$(2, \sqrt{n}^{-1})$, which would correspond to a GEV distribution with parameter $\gamma = -1/2 = -0.5$, so the -0.72 we obtained is seemingly a good estimate.



(a) (b)

Figure 5 – (a) Density for the Weibull$(2, 1)$ used in the example. (b) Q–Q plot of the set of negative sample minima $-\overline{m}_i$ against the quantiles of a GEV(-0.72) distribution. The population maximum of the GEV (dashed vertical line) and the corresponding estimate for the negative population minimum (blue cross mark) are shown.

Source – Prepared by the author.

Of course, the same procedure above can be used to estimate any quantile of the sample maximum or minimum. We might be interested, for example, in calculating the river flow τ such that there is an extremely low probability ϵ that, in a year, the river will ever surpass τ. This means we want $P(\max\{X_i : i = 1, \ldots, n\} > \tau) = \epsilon$, where X_i are the flow rates in each day, and this corresponds to finding the $(1 - \epsilon)$-quantile of the sample maximum.

2.3.2 When Multiple Sample Maxima Are Not Available

When a sample of sample maxima (i.e., $x_{(n),1}, x_{(n),2}, \ldots, x_{(n),k}$) is not available, extreme quantile estimation relies on different theoretical reasoning, which requires us to go deeper into extreme value theory (EVT). To avoid clogging the text with references, we state beforehand that the following discussion is heavily based on the two great books (HAAN; FERREIRA, 2006) and (BEIRLANT *et al.*, 2004).

Recall from Section 2.1 that we must find, for a given cdf F, sequences $a_n > 0$ and $b_n \in \mathbb{R}$ as well as the shape parameter γ such that the pointwise convergence holds:

$$\lim_{n\to\infty} F^n\left(a_n x + b_n\right) = G_\gamma(x). \tag{2.5}$$

However, the vast literature on EVT, which ran throughout most of the twentieth century, learned by experience that is more profitable to work with the following function:

$$U(x) := \left(\frac{1}{1-F(x)}\right)^{\text{inv}} = F^{-1}\left(1 - 1/x\right), \quad \text{with } x \in [1, \infty), \tag{2.6}$$

where the exponent inv denotes the inverse whenever the usual notation becomes ambiguous. Whenever F is not injective, we take the inverse as being the left-continuous inverse:

$$F^{-1}(x) = \inf\left\{y : F(y) \geq x\right\}, \tag{2.7}$$

and the results from EVT remain the same.

It can be shown that, under certain regularity conditions, if a sequence of functions $f_n \to g$ pointwise, then we also have the pointwise convergence of the inverse functions $f_n^{-1} \to g$. These regularity conditions always hold for f_n and g when f_n is a non-degenerate cdf and g is the cdf for the GEV distribution. With this result, Equations (2.5) and (2.6) can be manipulated to obtain the following theorem, adapted from (HAAN; FERREIRA, 2006).

Theorem 2. *Let F be a non-degenerate cdf, and G_γ be the cdf for the generalized extreme value distribution with parameter $\gamma \in \mathbb{R}$. Then the following statements are equivalent:*

(i) There exist sequences $a_n > 0$ and $b_n \in \mathbb{R}$ such that:

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G_\gamma(x), \tag{2.8}$$

for all x with $1 + \gamma x > 0$.

(ii) There exists a positive function $a(t)$ such that:

$$\lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} = G_\gamma^{-1}\left(e^{-1/x}\right) = \frac{x^\gamma - 1}{\gamma}, \tag{2.9}$$

for all $x > 0$ (when $\gamma = 0$, the left side is $\log x$).

(iii) There exists a positive function $a(t)$ (the same as in item 2) such that:

$$\lim_{t\to\infty} t\left[1 - F\left(a(t)x + U(t)\right)\right] = (1+\gamma x)^{-1/\gamma}, \tag{2.10}$$

whenever $1 + \gamma x > 0$ holds.

(iv) There exists a positive function $h(t)$ such that:

$$\lim_{t\to\mathcal{M}^-} \frac{1 - F(t + xh(t))}{1 - F(t)} = (1+\gamma x)^{-1/\gamma}, \tag{2.11}$$

whenever $1 + \gamma x > 0$ and with $\mathcal{M} = \sup\{x : F(x) < 1\}$ the population maximum.

Furthermore, Equation (2.8) holds with $b_n = U(n)$, and Equation (2.11) holds with $h(t) = a(1/(1 - F(t)))$.

It is worth noting that the function $a(t)$ that appears in Equations (2.9) and (2.10) are merely interpolations built upon the sequence a_n in Equation (2.8); in fact, the proof that $(i) \Rightarrow (iii)$ uses $a(t) = a_{\lfloor t \rfloor}$. Also, the theorem also highlights the importance of the function $U(x)$, as it can be used to define the sequence b_n.

The left term on Equation (2.11) can be rewritten as:

$$\begin{aligned}\frac{1 - F(t + xh(t))}{1 - F(t)} &= \frac{P\left(X \geq t + xh(t)\right)}{P\left(X \geq t\right)}\\ &= \frac{P\left(X \geq t + xh(t), X \geq t\right)}{P\left(X \geq t\right)}\\ &= P\left(\frac{X - t}{h(t)} \geq x \,\middle|\, X \geq t\right),\end{aligned}$$

where the second equality uses the fact that $h(t)$ is a positive function and assumes $x \geq 0$. If $x < 0$, the equality does not hold. Thus, for $x \geq 0$, Equation (2.11) says that the probability distribution of the high values of X (i.e., larger than t for large t) should follow a distribution with survival function $(1 + \gamma x)^{-1/\gamma}$, which corresponds to a standard generalized Pareto distribution, whose cdf is (EMBRECHTS; SCHMIDLI, 1994):

$$F(x \,|\, \xi, \mu, \sigma) = 1 - \left(1 + \xi\frac{x-\mu}{\sigma}\right)^{-1/\xi}, \; x > \mu, \tag{2.12}$$

and the standard version is given by placing $\mu = 0$ and $\sigma = 1$.

Unfortunately, the above intuition can only be made mathematically rigorous for the case $\gamma > 0$. The result used here is that:

$$\lim_{t\to\infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-1/\gamma}, \; \gamma > 0,$$

is a necessary and sufficient condition for $F^n(a_n x + b_n) \to G_\gamma(x)$, as long as $\gamma > 0$. In this case, the interpretation is that, for $x > 0$:

$$\lim_{t\to\infty} P\left(\frac{X}{t} \geq x \,\middle|\, X \geq t\right) = x^{-1/\gamma}, \tag{2.13}$$

so the largest values of the sample, scaled by the threshold t, follow a generalized Pareto distribution with parameters $\xi = \gamma$, $\sigma = \gamma$ and $\mu = 1$, whose cdf is $F(x \,|\, \gamma) = 1 - x^{-1/\gamma}$.

Now consider we have a sample $X_1, \ldots, X_n$ from a cdf F. Say we pick the k largest values of the sample, $X_{(n-k+1)}, X_{(n-k+2)}, \ldots, X_{(n)}$. Following the logic given in Equation (2.12), our threshold t here is $X_{(n-k)}$. If we assume that this threshold is large enough for the relation in Equation (2.12) to be approximately valid, then the random variables $X_{(n-k+1)}, X_{(n-k+2)}, \ldots, X_{(n)}$, divided by $X_{(n-k)}$, should form a sample from distribution with cdf $H(x \,|\, \gamma) = 1 - x^{-1/\gamma}$ and pdf:

$$h(x \,|\, \gamma) = \frac{1}{\gamma x^{1+1/\gamma}}, \; x > 1,$$

so that we can calculate the log-likelihood of the sample of the k largest values:

$$\begin{aligned} l(\gamma) &= \sum_{i=1}^{k} \left[-\log \gamma - \left(1 + \frac{1}{\gamma}\right) \log \frac{X_{(n-k+i)}}{X_{(n-k)}} \right] \\ &= -k \log \gamma - \left(1 + \frac{1}{\gamma}\right) \sum_{i=1}^{k} \left[\log X_{(n-k+i)} - \log X_{(n-k)} \right], \end{aligned}$$

which by differentiation and equating the derivative to zero leads to the equation:

$$\frac{1}{\gamma^2} \sum_{i=1}^{k} \left[\log X_{(n-k+i)} - \log X_{(n-k)} \right] = \frac{k}{\gamma},$$

which yields our maximum likelihood estimator for γ:

$$\hat{\gamma}_H = \frac{1}{k} \sum_{i=1}^{k} \left[\log X_{(n-k+i)} - \log X_{(n-k)} \right],$$

where we leave the verification of the boundaries to the literature, especially to Hill (1975) who first presented this estimator, which is why we call it **Hill's estimator** and denote it as $\hat{\gamma}_H$. It is a consistent estimator for γ (MASON, 1982).

As mentioned, the above estimator works only for $\gamma > 0$. For the general case $\gamma \in \mathbb{R}$, there is the Pickands' estimator (PICKANDS, 1975), which, surprisingly, was proposed in the same year as Hill's estimator:

$$\hat{\gamma}_P = \frac{1}{\log 2} \log \left(\frac{X_{(n-k)} - X_{(n-2k)}}{X_{(n-2k)} - X_{(n-4k)}} \right),$$

which is also a consistent estimator (HAAN; FERREIRA, 2006). This estimator will be used in Chapter 4 to compare with our proposed methods.

CHAPTER 3

LITERATURE REVIEW

This chapter is intended to present the state-of-the-art in areas most related to this dissertation, namely extreme quantile estimation and estimation of population maximum and minimum. Insofar as we could infer from the literature, the problem of estimating the population maximum or minimum is most commonly called “endpoint estimation”, dating back at least to (HALL, 1982). Much less frequently, it is called “boundary estimation” or “population minimum (or maximum) estimation.”

This is somewhat unfortunate because the sentence “endpoint estimation” appears in other scientific fields, which happened to make our search in scientific databases more difficult. Most notably, in biomedical sciences there exists the important concept of clinical endpoint, which corresponds to the point at which clinical experimentation ends for a specific individual that participated in the experiment.

For example, in testing a new drug for treating the COVID-19 disease, the endpoint can be when the individual tests negative twice, or when the viral load is below some threshold, or when the individual dies (HOSOGAYA *et al.*, 2021; ZHANG *et al.*, 2021). Needless to say, this subject is of extreme importance, so the choice of endpoints is studied in-depth by biomedical researchers and statisticians who aim to improve the power of hypothesis tests, reduce the necessary sample sizes or reduce risks for the experiments (MOLENBERGHS; BUYSE; BURZYKOWSKI, 2005; SAAD; LEE, 2020; MAO; KIM; MIAO, 2021).

It happens that the words “endpoint” and “estimation” occur frequently in these kinds of articles, which appeared in large quantities in our searches. Besides that, it also appear frequently in survival analysis, where the point where data is censored is sometime called *endpoint* (COOK; WANG, 2020; EDELMANN; WELCHOWSKI; BENNER, 2021), and also in confidence interval calculation, where the interval bounds are also called *endpoints* (FRANCO *et al.*, 2019; BURCH, 2017). In light of the discussed here, we argue that it might be better for the field to find itself a more discerning name, such as *population extrema estimation*.

3.1 Endpoint Estimation

To search the literature, we used the search engines Scopus (BOYLE; SHERMAN, 2006) and Web of Science (CLARIVATE ANALYTICS, 2021). For endpoint estimation, we used the search string:

```
TITLE-ABS-KEY ( "Endpoint"
                AND  ( "Estimation"  OR  "Estimator"  OR  "Estimating" )
                AND  ( "Distribution"  OR  "Probability"  OR  "Sample" ) ),
```

where the last `AND` statement is an attempt to exclude articles from biomedical areas. We are here making the reasonable assumption that all articles that touch the topic of endpoint estimation will use the word “distribution”, “probability” or “sample” in their abstracts. This raised 1.393 articles in the Scopus database, and 892 in Web of Science.

We also used:

```
TITLE-ABS-KEY ( ( "Boundary Estimation"  OR  "Boundary Estimator"
                  OR  "Estimator for the Boundary"  OR  "Estimate the Boundary" )
                AND  ( "Distribution"  OR  "Probability"  OR  "Sample" ) ),
```

which raised 166 in Scopus and 100 in Web of Science.

Finally, we used:

```
TITLE-ABS-KEY ( ( "population minimum"  OR  "population maximum"
                  OR  "population minima"  OR  "population maxima" )
                AND  ( "Estimation"  OR  "Estimator"  OR  "Estimate" )
                AND  ( "Distribution"  OR  "Probability"  OR  "Sample" ) ),
```

which yielded 25 results in Scopus and 12 in Web of Science.

Some degree of trial and error was involved in choosing these strings, and the whole method is not presented here. The titles of all the articles found were read, and those that were clearly unrelated to endpoint estimation were excluded, which was actually the majority of the articles found.Those that could possibly be related to endpoint estimation had their abstract and, when deemed necessary, their introductions read; this resulted in further filtering of the set of relevant articles to this dissertation, which went down to 31 articles, the most important of which are commented in the following.

The most recent work in the field is (WANG *et al.*, 2019), whose main purpose is to strengthen the endpoint estimator given in (HALL, 1982), which does not work well in some circumstances. Wang *et al.* (2019) begin by assuming the underlying distribution (with cdf F) to satisfy:

$$1 - F(x) = c\,(\mathcal{M} - x)^{\alpha} + o\left\{(\mathcal{M} - x)^{\alpha}\right\} \quad \text{as } x \longrightarrow \mathcal{M}^{-}, \tag{3.1}$$

where $\mathcal{M}$ is the right endpoint (population maximum), $c > 0$ is a constant and α is what they call the exponent of the distribution.

What they are doing here is to assume the distribution has the same tail as a generalized Pareto distribution in the limit as $x \longrightarrow \mathcal{M}^{-}$. This is not very immediate to see, so recall here that the generalized Pareto distribution has cdf (BEIRLANT *et al.*, 2004):

$$G(x \mid \gamma, \sigma, \mu) = 1 - \left(1 + \frac{\gamma(x-\mu)}{\sigma}\right)^{-1/\gamma}, \quad \text{with} \begin{cases} x \geq \mu, & \text{if } \gamma \geq 0 \\ \mu \leq x \leq \mu - \dfrac{\sigma}{\gamma}, & \text{if } \gamma < 0, \end{cases}$$

with $\gamma, \mu \in \mathbb{R}$ and $\sigma > 0$. By making the reparametrization $\gamma' = -1/\gamma$, $\sigma' = 1/\gamma$, $\mu' = \mu$ and requiring the new γ' to be positive, we obtain:

$$G(x \mid \gamma', \mu') = 1 - (1 + \mu' - x)^{\gamma'}, \quad \text{with } \mu' \leq x \leq \mu' + 1,$$

and the similarity with Equation (3.1) comes by easily by setting up the location parameter μ conveniently.

The method proposed in (HALL, 1982) does not work very well when $\alpha \in (0, 1]$ since it estimates the population maximum as $X_{(n)}$, the sample maximum. The method consists of taking the $k + 1$ largest sample points and performing maximum likelihood estimation using the cdf F given in Equation (3.1), which yields:

$$L(\mathcal{M}, c, \alpha) = (c\alpha)^{k+1} \left[\prod_{j=0}^{k} \left(\mathcal{M} - X_{(n-k+j)}\right)^{\alpha-1}\right] \left[1 - c\left(\mathcal{M} - X_{(n-k)}\right)^{\alpha}\right]^{n-k-1}. \tag{3.2}$$

To solve the problem of returning the sample maximum as the estimate, Wang *et al.* (2019) proposes multiplying the above likelihood function by a penalty function, resulting in:

$$L(\mathcal{M}, c, \alpha) = c^{k+1}\alpha^{k} \left[\prod_{j=0}^{k} \left(\mathcal{M} - X_{(n-k+j)}\right)^{\alpha-1}\right] \left[1 - c\left(\mathcal{M} - X_{(n-k)}\right)^{\alpha}\right]^{n-k-1} \frac{\mathcal{M} - X_{(n)}}{\alpha\left(\mathcal{M} - X_{(n-k)}\right)}.$$

Wang *et al.* (2019) proceed to prove that there is a unique global maximum to the above problem when $\alpha > 2$ and $k \geq 2$. They also prove consistency of the estimator if the cdf F is continuous in a neighborhood of $\mathcal{M}$ and if k is fixed (actually, they generalize it for certain sequences k_n).

Slightly less recently, (LENG *et al.*, 2019) proposed an endpoint estimator for samples containing normal noise. Let $X_1, \ldots, X_n$ be a sample from a distribution with finite population maximum $\mathcal{M}$, and consider that what we actually observe is $Y_i = X_i + \epsilon_i$, where ϵ_i are iid normally distributed with mean 0 and variance σ^2.

Note that even though the variables X_i are limited from above, the Y_i are unbounded, having the real line as support. The problem of determining the underlying distribution when the noise is part of a known parametric distribution family belongs to the field of deconvolution (MEISTER; NEUMANN, 2010; STEFANSKI; CARROLL, 1990).

Goldenshluger and Tsybakov (2004) proposed an estimator for $\mathcal{M}$ that depends on σ, a quantity we do not have access to. Building upon these results, Leng *et al.* (2019) propose an estimator for σ^2 that, when plugged into the estimator of (GOLDENSHLUGER; TSYBAKOV, 2004), yields nice properties such as asymptotic normality and consistency. To derive their result, they require the underlying distribution of X_i to have a specific form, one which forces it to have the reverse Weibull as its asymptotic sample maximum distribution (see Section 2.1). Their estimator for σ is:

$$\hat{\sigma}_g = \frac{\sqrt{\log(n/k)}}{\sqrt{2}k} \sum_{i=1}^{k-1} g(i/k)(Y_{(n-i)} - Y_{(n-k)}),$$

where g is any function that satisfies $\int_0^1 -g(s)\log(s)\mathrm{d}s = 2$, though in (LENG *et al.*, 2019) the function $g(s) = -\log(s)$ is used the most. The function g affects the variance of the asymptotic distribution of the estimator, and there might exist better choices than the aforementioned one. With this, the resulting estimator for $\mathcal{M}$ is:

$$\hat{\mathcal{M}}_g = Y_{(n)} - \hat{\sigma}_g\sqrt{2\log n}.$$

Leng *et al.* (2019) prove, under some general but not very simple conditions, the asymptotic normality of $\sqrt{k_n}[(\hat{\sigma}_g/\sigma) - 1]$ and $\sqrt{k_n/\log n}(\hat{\mathcal{M}}_g - \mathcal{M})$. However, if the choice $g(s) = -\log(s)$ is made, then all that is needed is that k_n be a sequence such that $k_n \to \infty$ and $k_n/n \to 0$.

In (ALVES; HAAN; NEVES, 2013) and (ALVES; NEVES, 2014), an estimator is proposed for the case where the sample minimum converges to the Gumbel case of the extreme value theorem. This is one of the estimators that will be used in the experiments for comparison with our proposed methods. Recall the limit equation given in Equation (2.9):

$$\lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} = \frac{x^\gamma - 1}{\gamma},$$

and for $\gamma = 0$ it can be read as a convergence to the function $\log x$, so it ends up being a simpler expression than shown above. The authors in (ALVES; HAAN; NEVES, 2013; ALVES; NEVES, 2014) explore this limit equation to devise their estimator. They show, with the help of results from (HAAN; FERREIRA, 2006; DREES, 1998), that if the relation above holds with $\gamma = 0$, then the following also holds:

$$U(\infty) - U(t) = \int_t^\infty \frac{a(s)}{s} ds + o(a(t)), \quad \text{for } t \to \infty,$$

where $o(a(t))$ is any function b such that $b(t)/a(t) \to 0$. $U(\infty)$ should be read as $\lim_{x\to\infty} U(x)$ and is equal to the population maximum $\mathcal{M}$ by the definition of U (see Section 2.1). Because of this, the above expression can also be seen as:

$$\mathcal{M} = U(t) + \int_t^\infty \frac{a(s)}{s} ds + o(a(t)), \quad \text{for } t \to \infty,$$

so by estimating or approximating each quantity on the right-hand side, we obtain an estimator for $\mathcal{M}$. By following such procedures, the authors arrive at the estimator:

$$\hat{\mathcal{M}} = X_{(n-k)} + X_{(n)} + \frac{1}{\log 2} \sum_{i=0}^{k-1} \log\left(\frac{k+i}{k+i+1}\right) X_{(n-k-i)}, \tag{3.3}$$

where k is a very small number relative to n. The authors show and emphasize that the resulting estimate is always greater than the sample maximum, a property that is obviously desirable, if not mandatory, for any endpoint estimator.

The above estimator is further developed in (ALVES; NEVES; ROSÁRIO, 2017), where the same estimator is applied to cases where $\gamma \leq 0$. They show that even in this more general scenario, the estimator is still consistent. Here, the authors work mainly with the estimator in the form:

$$\hat{\mathcal{M}} = X_{(n)} + \sum_{i=0}^{k-1} a_{i,k} \left(X_{(n-k)} - X_{(n-k-i)}\right) \qquad \text{with } \sum_{i=0}^{k-1} a_{i,k} = 1.$$

Note that Equation (3.3) can be put in the above form by setting:

$$a_{i,k} = \log\left(\frac{k+i+1}{k+i}\right) \frac{1}{\log 2}.$$

Alves, Neves and Rosário (2017) prove the asymptotic distribution of the above estimator; specifically, they studied the quantity:

$$k^{1/2 - \max\{0, \gamma + 1/2\}} \left(\frac{\hat{\mathcal{M}} - \mathcal{M}}{b(n/2)} - h(\gamma)\right),$$

where b and h are conveniently defined functions, and found out that it is Weibull distributed if

$\gamma \in (-1/2, 0)$, normally distributed if $\gamma < -1/2$, and is the sum of mutually independent Weibull and normal distributions whenever $\gamma = -1/2$. Note, however, that to achieve this results we need to impose what is called "second order condition", which is common within the extreme value theory field (HAAN; FERREIRA, 2006; HAAN; STADTMÜLLER, 1996). The case $\gamma = 0$ is treated separately in (ALVES; NEVES, 2014).

Li and Peng (2012) demonstrate that bootstrapping works for constructing confidence intervals to estimates given by the estimator proposed in (HALL, 1982), which is the value of $\mathcal{M}$ that maximizes the likelihood shown in Equation (3.2). Li and Peng (2012) actually use the fact that the estimator is also the solution of the following equation (in $\mathcal{M}$):

$$\frac{1}{k+1} \sum_{i=1}^{k} \frac{\mathcal{M} - X_{(n-k)}}{\mathcal{M} - X_{(n-i+1)}} \left\{ \frac{1}{k+1} \sum_{i=1}^{k} \log \frac{\mathcal{M} - X_{(n-i+1)}}{\mathcal{M} - X_{(n-k)}} + 1 \right\} = 1, \tag{3.4}$$

and if we denote its solution as $\hat{\mathcal{M}}$, then the estimator for the extreme value index becomes:

$$\hat{\gamma} = \frac{1}{k+1} \sum_{i=1}^{k} \log \left\{ \frac{\hat{\mathcal{M}} - X_{(n-i+1)}}{\hat{\mathcal{M}} - X_{(n-k)}} \right\}. \tag{3.5}$$

The bootstrap estimator comes by substituting the order statistics above by corresponding order statistics $X^*_{(q)}$ of a bootstrapped sample. Li and Peng (2012) then give the estimators' asymptotic distribution using results discussed in (HAAN; STADTMÜLLER, 1996) and (HAAN; FERREIRA, 2006), with which deriving asymptotic confidence intervals become possible.

Out of an abundance of caution, we note that the authors took Equation (3.4) directly from a past work within their research group (LI; PENG, 2010), where the equation is presented without proof. It is not present in the original work where the estimator was proposed (HALL, 1982); in turn, Hall (1982) does say (and prove) that the estimator will be the solution in $\mathcal{M}$ of the equation:

$$(k+1) \left\{ \left(\sum_{i=1}^{k} \log \frac{\mathcal{M} - X_{(n-k)}}{\mathcal{M} - X_{(n-i+1)}} \right)^{-1} - \left(\sum_{i=1}^{k} \frac{X_{(n-i+1)} - X_{(n-k)}}{\mathcal{M} - X_{(n-i+1)}} \right)^{-1} \right\} = 1,$$

where we just adapt Hall's notation to agree with the notation we have been using thus far. Since we could not pinpoint the connection between maximizing the likelihood in Equation (3.2) with obtaining the estimator given in Equation (3.4), we recommend the interested reader to do their own careful assessment of the papers mentioned above, namely (HALL, 1982; LI; PENG, 2012; LI; PENG, 2010).

Since we mentioned (LI; PENG, 2010), we take the opportunity to present their work, albeit momentarily breaking the chronological order followed up to this point in the section. Li and Peng (2010) derive the asymptotic distribution for the Hall estimator when the underlying cdf has the form:

$$1 - F(x) = c(\mathcal{M} - x)^{\alpha} + d(\theta - x)^{\alpha+\beta} + o\left((\mathcal{M} - x)^{\alpha+\beta}\right), \tag{3.6}$$

and k_n is a sequence such that $\sqrt{k_n}(k_n/n)^{\beta/\alpha} \longrightarrow \lambda \in [0, \infty)$. Contrast the above equation with the similar condition shown in Equation (3.1).

Note also that Hall (1982) had also derived the asymptotic distribution, but his assumption was $k_n = o(n^{m/(m+1/2)})$ where m is a complex formula that we omit here. With the derived asymptotic distributions, Li and Peng (2010) conclude that fitting a generalized Pareto distribution to the largest data points of your sample may be a bad idea if the extreme value index is positive.

(GIRARD; GUILLOU; STUPFLER, 2012a) proposes a moment-based estimator for the endpoint, which will also be used in the experiments (Chapter 4). Let $\hat{\mu}_p$ be the sample moment of p-th order defined as:

$$\hat{\mu}_p = \frac{1}{n}\sum_{i=1}^{n} X_i^p,$$

then by taking a sequence p_n such that $p_n \longrightarrow \infty$, the proposed estimator is:

$$\frac{1}{\hat{\mathcal{M}}} = \frac{1}{a p_n}\left\{\left[(a+1)\, p_n + 1\right] \frac{\hat{\mu}_{(a+1)p_n}}{\hat{\mu}_{(a+1)p_n+1}} - (p_n + 1)\frac{\hat{\mu}_{p_n}}{\hat{\mu}_{p_n+1}}\right\}. \tag{3.7}$$

The estimator is constructed by taking the underlying distribution to be of a particular family, with survival function $S(y \mid \mathcal{M}, \alpha) = (1 - y/\mathcal{M})^{\alpha}$, which is the family of Pareto distributions also discussed previously in this Section.

With this assumption it is possible to derive the system of equations typical for finding estimators by the method of moments (MOOD, 1950), except that the resulting estimators depend on unknown quantities that must be estimated. Girard, Guillou and Stupfler (2012a) then substitute these unknowns by convenient estimators, and ends up with the above estimator for the endpoint. The authors prove consistency when the random variable is assumed to be positive, have all moments defined up to infinite order, and its endpoint to indeed be finite; they also prove asymptotic normality under a few more assumptions.

In (GIRARD; GUILLOU; STUPFLER, 2012b), the same authors follow a very similar construction procedure, but now doing in a way that does not require the underlying random variable to be positive, as is required in (GIRARD; GUILLOU; STUPFLER, 2012a). To do that, they look at the moments of the transformed random variable $Y = e^X$, which is supported on $[0, e^{\mathcal{M}}]$, so that finding the endpoint of Y also gives the endpoint of X. In this context, the proposed estimator is:

$$\hat{\mathcal{M}} = \frac{1}{a}\left\{\log\left(\frac{\hat{\mu}_{p_n}}{\hat{\mu}_{p_n+1}}\right) - \log\left(\frac{\hat{\mu}_{(a+1)p_n}}{\hat{\mu}_{(a+1)p_n+a+1}}\right)\right\},$$

where now the sample moments are taken from the transformed variable $Y = e^X$:

$$\hat{\mu}_p = \frac{1}{n}\sum_{i=1}^{n} e^{pX_i},$$

and $a > 0$ is an adjustable parameter. The authors also prove consistency here, but now only require that the endpoint be finite and all moments exist.

The biggest drawback in the two aforementioned works (GIRARD; GUILLOU; STUPFLER, 2012b; GIRARD; GUILLOU; STUPFLER, 2012a) is that there is no method to choose the hyperparameters a and p_n. Note that even though p_n is a sequence, in practice you only need to guess its value for the particular sample size n at hand. The authors perform experiments to test multiple values of these parameters, so the ones that performed best in their experiments could be used, albeit with no guarantee that they will be good choices for your particular problem.

(LI; PENG; XU, 2011) and (LI; PENG, 2009) are two works that focus on bias reduction for endpoint estimators. Both of them use the assumption given in Equation 3.6, and focus on estimating β in smarter ways. Assumptions here include, for example:

$$k_n^{1/2}\,|A(n/k_n)| \to \infty, \quad k_n^{1/2}A(n/k)^2 \to 0, \quad k_n^{1/2}A(n/k)B(n/k) \to 0,$$

where A and B are functions assumed to exist in the second order condition (HAAN; FERREIRA, 2006). The choice of β given in (LI; PENG, 2009) has more of a theoretical importance, and is difficult to find in practice; motivated by this, (LI; PENG; XU, 2011) propose a more practical bias reduction method, which involves optimizing the likelihood simultaneously over $\mathcal{M}, \alpha, \beta, c, d$ in the likelihood function that appears under the assumption in Equation (3.6). Using the same likelihood function, (LI; PENG; QI, 2011) derive confidence intervals for the endpoint estimator.

(CAI; HAAN; ZHOU, 2013) (preprint published in 2011) propose an improved version of the extreme value index estimator given in (HOSKING; WALLIS, 1987), called *probability weighted moment (PWM)* estimator, given by:

$$\hat{\gamma} = \frac{I_1 - 4I_2}{I_1 - 2I_2}, \tag{3.8}$$

where the moment functions I_m are defined as:

$$I_m = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{i}{k}\right)^{m-1}\left(X_{(n-i+1)} - X_{(n-k)}\right).$$

It happens that the above estimator has known asymptotic distribution, with a known non-zero expected value that involves an unknown quantity $\Lambda(n/k)$. Cai, Haan and Zhou (2013) propose an estimator $\hat{\Lambda}(n/k)$ for this quantity, so that it can be subtracted from the PWM estimator, thus obtaining an unbiased (or bias reduced) estimator. They also use this estimator to propose an improved estimator for the endpoint (see Section 3.3 of their work). The estimator by (HOSKING; WALLIS, 1987) will be used in Chapter 4.

In dealing with endpoints, it might be useful to test beforehand whether the underlying distribution has a finite point at all. This is particularly a problem when γ is near 0, because it corresponds to the Gumbel domain, where the underlying distribution can have finite or infinite endpoint. Neves and Pereira (2010) propose two hypothesis tests to aid in this problem, using the statistics:

$$T^{(1)}_{k,n} = \frac{1}{2}\sum_{i=1}^{k}\frac{X_{(n-i)} - X_{(n-k)} - \hat{a}(n/k)}{X_{(n)} - X_{(n-k)}}$$
$$T^{(2)}_{k,n} = \frac{1}{2}\sum_{i=1}^{k-1}\frac{X_{(n-i+1)} - X_{(n-i)}}{X_{(n-k)}},$$

where $\hat{a}$ is an estimator for the function a that appears in Equation (2.9); possible estimators $\hat{a}$ are given in (DEKKERS; EINMAHL; HAAN, 1989; HAAN; FERREIRA, 2006), for example.

More on the theoretical side, (PENG; QI, 2009) derives the asymptotic distribution for the likelihood estimator when using a cdf that is an adaptation of what is shown in Equation (2.11):

$$\lim_{t\to\infty}\frac{1 - F(\mathcal{M} - (tx)^{-1})}{1 - F(\mathcal{M} - t^{-1})} = x^{1/\gamma}, \quad \text{with } x > 0.$$

The authors do it for the case $\gamma \in (-1, -1/2]$, which is particularly difficult in terms of mathematical manipulations; in this sense, their work is a generalization of (DREES; FERREIRA; HAAN, 2004) and an alternative for the method proposed in (WOODROOFE, 1974).

(MÜLLER; HÜSLER, 2005) proposes an estimator based on the one proposed in (FALK, 1995), given by:

$$\hat{\gamma} = \frac{1}{k_n}\sum_{j=1}^{k_n}\log\left(\frac{\mathcal{M} - X_{(n-j+1)}}{\mathcal{M} - X_{(n-k_n)}}\right).$$

This estimator assumes the endpoint is known, which is not always practical. Because of that, Müller and Hüsler (2005) give an iterative procedure to estimate γ, which consists of taking an initial guess $\mathcal{M}^{(0)}$, estimating $\hat{\gamma}$ and then using this to get a new estimate $\mathcal{M}^{(1)}$ for the endpoint and repeat the process iteratively.

In the previous paragraphs, it was often mentioned the existence of a sequence k_n that determines how many tail samples we use for estimating the endpoint or high quantiles. We have not, however, discussed how to choose this sequence. Ferreira, Haan and Peng (2003) propose that the problem be formulated as a mathematical optimization, seeking the value:

$$k(n) = \arg\inf_{k} \lim_{n\to\infty} \mathrm{E}\left[\left(\hat{x}_q(k,n) - x_q\right)^2\right],$$

where $\hat{x}_q$ is an estimator for the q extreme quantile of the distribution, and the inner limit should be read as taking the expectation using the limit distribution of the estimator. In other words, take the expectation assuming the k largest samples is distributed according to a generalized Pareto distribution. In (FERREIRA; HAAN; PENG, 2003), the authors give a bootstrapping procedure to give a solution to the optimization problem formulated above.

Hall and Park (2002) propose two bias correction methods for endpoint estimation, one for the univariate case and one for multidimensional random variables. For the univariate case, which is of more interest to us here, they consider estimating the endpoint with the order statistic $X_{(n)}$, and they propose estimating the bias β as:

$$\hat{\beta}_m = \frac{\sum_{i=1}^{m}\left(X_{(n-i+1)} - X_{(n-i)}\right) K\left((i+\hat{\alpha})/m\right)}{\sum_{i=1}^{m} K(i/m)},$$

where $\hat{\alpha}$ is what they call a “translation correction” factor, and K can be seen as a (non-negative) weight function; in fact, if we set $\hat{\alpha} = 0$, it is easy to see the ratio as a weighted average with weights defined by the function K. The authors propose $\hat{\alpha}$ to be defined as:

$$\hat{\alpha} = -m\frac{\sum_{i=1}^{m-1} i\left(X_{(n-i+1)} - 2X_{(n-i)} + X_{(n-i-1)}\right) K(i/m)}{\sum_{i=1}^{m-1}\left(X_{(n-i+1)} - X_{(n-i)}\right) K'(i/m)},$$

and they prove that $\hat{\beta}_m$, using this translation correction factor, has a better convergence rate than the standard bias estimator (i.e., with $\hat{\alpha} = 0$).

We believe the works discussed above give a thorough overview of the state-of-the-art. Still, for completeness, we also give mention to a few older works of remarkable importance, but which are not described here (ATHREYA; FUKUCHI, 1997; HALL; WANG, 1999; LOH, 1984; FALK, 1995; SMITH, 1987; SMITH, 1985).

3.2 Extreme Quantile Estimation

Since our primary goal is focused on the endpoint, extreme quantiles end up being less important within the context of this dissertation. Because of this, we do not perform an equally detailed review of the literature of this field, but an overview is given in the following.

In contrast to endpoint estimation, extreme value estimation has a much broader practical applicability, which makes searching the literature somewhat more difficult. We highlight its application to, for example, cyclone intensity (JAGGER; ELSNER, 2009), windstorm prediction (FRIEDERICHS; THORARINSDOTTIR, 2012), rainstorm and flood forecasting (XU *et al.*, 2021; CHAUDHURI; SHARMA, 2021; DAS, 2021), electricity generation and consumption (BOULFANI *et al.*, 2021), measurement of market risk (PAUL; SHARMA, 2017; LIU *et al.*, 2018; CHAKRABORTY; CHANDRASHEKHAR; BALASUBRAMANIAN, 2021), large magnitude earthquake forecasting (DUTFOY, 2019) (usually long-term due to the difficulty of the problem (STEIN; WYSESSION, 2003)), analysis of the solar cycle (ACERO *et al.*, 2018; ACERO *et al.*, 2017; RAMOS, 2007), among many others.

Recall the asymptotic relation initially given in Equation (2.9):

$$\lim_{t\to\infty} \frac{U(tx)-U(t)}{a(t)} = \frac{x^\gamma - 1}{\gamma}, \tag{3.9}$$

for all $x > 0$ and some function a. This shows that, for a sufficiently large threshold t, we should have:

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

or equivalently:

$$U(x) \approx U(t) + a(t)\frac{(x/t)^\gamma - 1}{\gamma}.$$

Now say we are interested in a high quantile $F^{-1}(1-p)$ with a very small p. Since $F^{-1}(1-p) = U(1/p)$, we can substitute on the above equation to obtain the approximation:

$$F^{-1}(1-p) = U\left(\frac{1}{p}\right) \approx U(t) + a(t)\frac{(pt)^{-\gamma} - 1}{\gamma}.$$

We can then estimate the quantile by using suitable estimates for $U(t)$, γ and $a(t)$. Two estimators for γ were discussed in Section 2.1, namely those by Hill (1975) and Pickands (1975). Estimating $U(t)$ is relatively simple; since t must be "sufficiently large", it is common to choose it as being n/k (HAAN; FERREIRA, 2006; BEIRLANT *et al.*, 2004), in which case $U(n/k)$ is equivalent to the $1 - k/n$ quantile, whose simplest estimator is $X_{(n-k)}$.

A classical estimator for $a(n/k)$ is given as follows (DEKKERS; EINMAHL; HAAN, 1989; HAAN; FERREIRA, 2006; BEIRLANT *et al.*, 2004). Define:

$$M_n^{(j)} = \frac{1}{k}\sum_{i=0}^{k-1}\left(\log X_{(n-i)} - \log X_{(n-k)}\right)^j ,$$

then the estimator is defined as:

$$\hat{a}(n/k) = \hat{\sigma} = X_{(n-k)}M_n^{(1)}\left(1 - v\right)$$

$$\text{with } v = \min\left\{0, 1 - \frac{1}{2}\left(1 - \frac{\left(M_n^{(1)}\right)^2}{M_n^{(2)}}\right)^{-1}\right\}.$$

The classical methods discussed above and in Section 2.1 are still heavily influential and often studied in the literature. Valk (2016b) discusses and demonstrates that in some cases, the assumptions often made in extreme value theory (EVT) are too restrictive and sometimes not applicable to some simple probability models. In particular, they enumerate a few cases where assuming a generalized Pareto tail distribution (related to Equation (3.9)) fails. On that account, the author proposes an alternative to Equation (3.9):

$$\lim_{t\to\infty}\frac{\log U\left(e^{t\lambda}\right) - \log U\left(e^{t}\right)}{g(t)} = \frac{\lambda^\gamma - 1}{\gamma}, \tag{3.10}$$

which is called log generalized Weibull assumption (for the tail distribution), which generalizes the Weibull assumption discussed in (GARDES; GIRARD; GUILLOU, 2011). The idea here is to exploit the fact that Equation (3.9) is an existence statement on the function $a(t)$, which is a very broad statement and allows rewriting the limit equation in different ways; what changes is the particular function $a(t)$ (or $g(t)$ as above) that will satisfy the limit equation you write. In this sense, enveloping the function U with a pair of inverse functions (i.e. $f^{-1}\circ U \circ f$) as done in Equation (3.10) is a very natural modification to perform. A natural quantile estimator here is:

$$\log U\left(e^{t\lambda}\right) = \log U\left(e^{t}\right) + g(t)\frac{\lambda^\gamma - 1}{\gamma}$$

$$\Longrightarrow U\left(e^{\lambda}\right) = U\left(e^{t}\right)\exp\left\{g(t)\frac{(\lambda/t)^\gamma - 1}{\gamma}\right\}$$

$$\Longrightarrow U\left(e^{\lambda}\right) = X_{(n-k_n+1)}\exp\left\{g(t)\frac{(\lambda/t)^\gamma - 1}{\gamma}\right\},$$

where an estimator for each unknown component is given in (VALK, 2016b), and some improved ones are given in (VALK; CAI, 2018). This has been extended to the multivariate scenario (VALK, 2016a).

As for the Weibull assumption mentioned above, there seems to be a significant body of literature dedicated to it (e.g., see Beirlant, Bouquiaux and Werker (2006), Dierckx *et al.* (2009), Diebolt *et al.* (2008), Gardes, Girard and Guillou (2011), Girard, Guillou and Stupfler (2012b), Goegebeur, Beirlant and Wet (2010)). It focuses, however, on the case $\gamma = 0$, the Gumbel domain. Some important distributions whose extrema are asymptotically Gumbel include the normal, Weibull and gamma distributions (METHNI *et al.*, 2012; BEIRLANT *et al.*, 2004), so even though we are restricting γ to a single value, it can still be seen as comparably significant relative to the intervals $\gamma < 0$ or $\gamma > 0$.

We found (GARDES; GIRARD; GUILLOU, 2011) and (METHNI *et al.*, 2012) to be particularly interesting, since their results apply for $\gamma \geq 0$, so it has a broader applicability than the other cited works. Let $K_x(y) = \int_1^y u^{x-1} du$ for $x \in \mathbb{R}$. (GARDES; GIRARD; GUILLOU, 2011) models the tail distribution using the family defined by the survival function:

$$(1 - F(x)) = \exp\left(-K_\tau^{-1}(\log H(x))\right), \quad \text{for } x \geq x_0 > 0 \text{ and } \tau \in [0, 1],$$

where x_0 is the population minimum (depending on $H(x)$), $H(x)$ is an increasing function whose inverse is regularly varying with index θ, that is, $H^{-1}(x) = x^\theta L(x)$ with $L(x)$ slowly varying (for now consider it as $L(x) = 1$).

The parameters are $\tau \in [0, 1]$ and $\theta > 0$. Gardes, Girard and Guillou (2011) and Methni *et al.* (2012) give the estimators:

$$\hat{\theta}_n(k_n) = \frac{H(k_n)}{\mu(\log(n/k_n))}$$

$$\text{where } \mu(t) = \int_0^\infty (K_{\hat{\tau}}(x + t) - K_{\hat{\tau}}) e^{-x} dx$$

$$\hat{\tau}_n = \begin{cases} \psi^{-1}\left(\frac{H(k_n)}{H(k'_n)}; \log(n/k_n); \log(n/k'_n)\right), & \text{if } \frac{H(k_n)}{H_{(k'_n)}} < \frac{k'_n}{k_n} \\ \text{Unif}(0, 1), & \text{otherwise} \end{cases}$$

$$\psi(x, t, t') = \frac{\mu_x(t)}{\mu_x(t')},$$

where the cause where $\hat{\tau}$ is generated randomly almost never happens (probability zero) asymptotically. k'_n is another sequence such that $k'_n > k_n$. With these estimators, we get the high quantile estimator:

$$x_q = X_{(n-k_n+1)} \exp\left(\hat{\theta}_n(k_n)\left[K_{\hat{\tau}_n}(\log(1/q)) - K_{\hat{\tau}_n}(\log(n/k_n))\right]\right).$$

Finally, we mention that (BERANGER; PADOAN; SISSON, 2021) gives a contribution on the side of maximum likelihood approaches to extreme quantile estimation: i) they propose a bayesian version of the traditional approach, and ii) extend it to the multivariate case. Also, (DREES *et al.*, 2003) is an important work concerning samples that are not independent.

CHAPTER 4

METHODS TO DEAL WITH THE UNKNOWN POPULATION MINIMUM IN PARAMETRIC INFERENCE

Let U be a random variable supported on $[0, \infty)$, such that $X = m + U$ is supported on $[m, \infty)$. By support we mean the set on which the probability density is not zero, apart maybe from a subset of measure zero; hereafter, we consider all probability functions to be defined on the whole real line. Let $x_1, \ldots, x_n$ be a sample taken from X, and consider that the sample minimum is relatively high.

This is a typical case where, in practice, the practitioner must make a decision concerning how they will perform parametric inference: either i) use models supported on $[0, \infty)$, ii) guess the population minimum, iii) estimate it using an endpoint estimator, or iv) add the population minimum as a parameter to the models, and infer it by MLE together with the other parameters.

To the best of our knowledge, these are the only options one could find in the literature, but even so, there is no work discussing the specific problem of parametric inference with unknown population minimum. That is, all works will focus on estimation of the population minimum, and will use the bias or an estimated bias (obtained by simulation) to argue for the superiority (or not) of the proposed estimates. We show in this chapter that our ultimate goal is not to perfectly estimate the population minimum, and that biased or skewed estimates of the population minimum might be favorable for our objectives.

We begin with an example. Consider a sample of 20 points taken from a distribution $20 + \text{Gamma}(3, 1/2)$, as shown in the histogram in Figure 6, and assume the researcher that is investigating this data set does not know that the population minimum is 20, although they do suspect it comes from a gamma distribution.

In a first scenario, the researcher guesses the population minimum, so that they can continue to carry out their investigation. Figure 6 shows what would happen if their guess was any among $\{16, 19, 20.5\}$, and if they used maximum likelihood estimation to obtain the gamma distributions. Visually, none of the choices give a satisfactory result in recovering the true underlying distribution that generated the data, even though 19 and 20.5 are reasonable estimates considering that histogram. Note that 20.5 would be a good guess if the underlying distribution was a gamma of the one-tailed type, since in these cases the sample minimum tend to be very close to the population minimum.

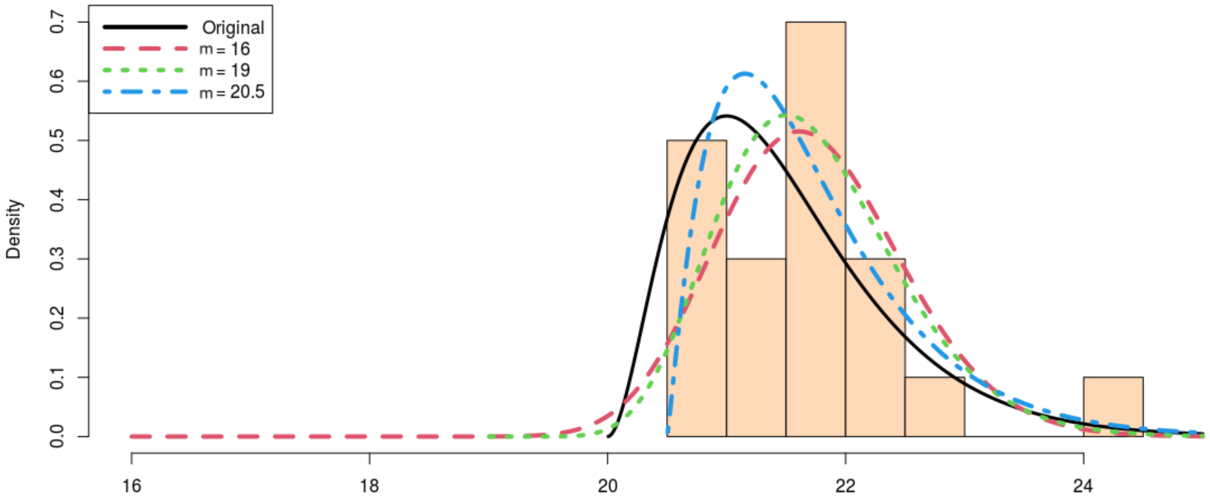


Figure 6 – Sample generated from a $20 + \text{Gamma}(3, 1/2)$ distribution, and the corresponding distributions obtained by maximum likelihood estimation by using different estimates m for the population minimum.

Source – Prepared by the author.

As an addendum, we note that even in this simple scenario we had to face one common source of frustration when dealing with population minima. The gamma distribution tends to yield very smooth likelihood functions, that can be easily maximized no matter the initial parameters are given. Nevertheless, in the case of Figure 6 when taking the population minimum as 19, the widely used limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) optimization algorithm (BYRD *et al.*, 1995; NOCEDAL; WRIGHT, 2006) would not converge when given the trivial initial parameters $\alpha = 1$ and $\beta = 1$ (shape and scale, respectively). The reader is invited to verify this by executing the R snippet shown in Algorithm 1; when the random seed is changed, about two thirds of the times will cause the same error.[1]

Fortunately, the gamma distribution is rather simple, all of its moments have closed mathematical forms, and we can easily select good initial conditions by setting $\beta = \overline{X}/\alpha$ based on the first moment. It is easy to see that as the distribution family gets more complicated, it gets increasingly difficult to find initial parameters that will lead MLE to converge. Not only the

---

[1] We have verified that setting the seed should cause the same random numbers to be generated in *most* operating systems and *most* versions of R. Exceptions do seem to exist; if that is the case, we suggest hard-coding the same sample we obtained: 21.71, 20.66, 20.7, 21.71, 22.06, 22.79, 21.89, 21.79, 20.87, 21.31, 21.47 21.61, 22.37, 20.89, 21.92, 21.68, 22.13, 20.99, 24.38, 21.4.

```
set.seed(74)
samp = round(rgamma(20, shape=3, scale=1/2) + 20, digits=2);
likelihood = function(p) -sum(dgamma(samp-19, shape=p[1], scale=p[2], log=T));
par = optim(c(1, 1), likelihood, method="L-BFGS");
```

Algorithm 1 – Snippet demonstrating that optimization failures happen even in simple cases when dealing with unknown population minimum.

number of degrees of freedom increases, but also the moment expressions get more complex (take the generalized gamma (STACY, 1962) and Weibull (MUDHOLKAR; SRIVASTAVA, 1993) as examples) and less useful in helping the researcher to devise initial parameter grids.

Back to Figure 6, what would happen if the practitioner decided to estimate the population minimum with MLE? This is likely the simplest and most general solution to the problem of unknown population minimum. However, for our example here, this results in the estimated population minimum being 20.62, resulting in the distribution shown in Figure 7. As can be seen, it looks even worse than any of the distributions obtained by guessing the population minimum (we will formally define "worse" later in this chapter). Recall that the log-likelihood for the gamma distribution is:

$$l(\alpha, \beta \,|\, x_1, \ldots x_n) = -n \log \Gamma(\alpha) - n\alpha \log \beta + (\alpha - 1) \sum_{i=1}^{n} \log(x_i) - \sum_{i=1}^{n} \frac{x_i}{\beta},$$

so that, if we add the population minimum as a parameter, becomes:

$$l(\alpha, \beta, m \,|\, x_1, \ldots x_n) = -n \log \Gamma(\alpha) - n\alpha \log \beta + (\alpha - 1) \sum_{i=1}^{n} \log(x_i - m) - \sum_{i=1}^{n} \left(\frac{x_i}{\beta}\right) + \frac{nm}{\beta},$$

and the derivative relative to m becomes:

$$\frac{\partial l}{\partial m} = (1 - \alpha) \sum_{i=1}^{n} \frac{1}{x_i - m} + \frac{n}{\beta}.$$

For $\alpha \in (0, 1]$, the derivative is clearly always positive, so that the MLE estimate $\hat{m}$ ends up being as large as possible (i.e., $m = x_{(1)}$). For $\alpha > 1$, we have:

$$(1 - \alpha) \sum_{i=1}^{n} \frac{1}{x_i - m} + \frac{n}{\beta} > 0 \iff \sum_{i=1}^{n} \frac{1}{x_i - m} < \frac{-n}{\beta(1 - \alpha)} = \frac{n}{\beta(\alpha - 1)},$$

where the inequality is inverted by dividing both sides by the negative quantity $(1 - \alpha)$. Now note in the sum that values $x_i - m$ are distances between the sample items and the population minimum; naturally, the largest term in the sum corresponds to $x_{(1)} - m$, the smallest distance.

We now write:

$$\frac{1}{x_{(1)} - m} < \sum_{i=1}^{n} \frac{1}{x_i - m} < \frac{n}{\beta(\alpha - 1)},$$

so that the derivative $\partial l / \partial m$ is positive if (but not if and only if):

$$\frac{1}{x_{(1)} - m} < \frac{n}{\beta(\alpha - 1)}$$

$$m < x_{(1)} - \frac{\beta(\alpha - 1)}{n},$$

which means that the MLE estimate $\hat{m}$ will be at least the value on the right hand side. Since an upper bound exists (namely, $x_{(1)}$), we can say:

$$x_{(1)} - \frac{\beta(\alpha - 1)}{n} \leq \hat{m} \leq x_{(1)}.$$

For the example in question (where $\alpha = 3$ and $\beta = 1/2$), the bound above is $(20.6149, 20.6649)$; contrast with the 20.62 obtained experimentally.

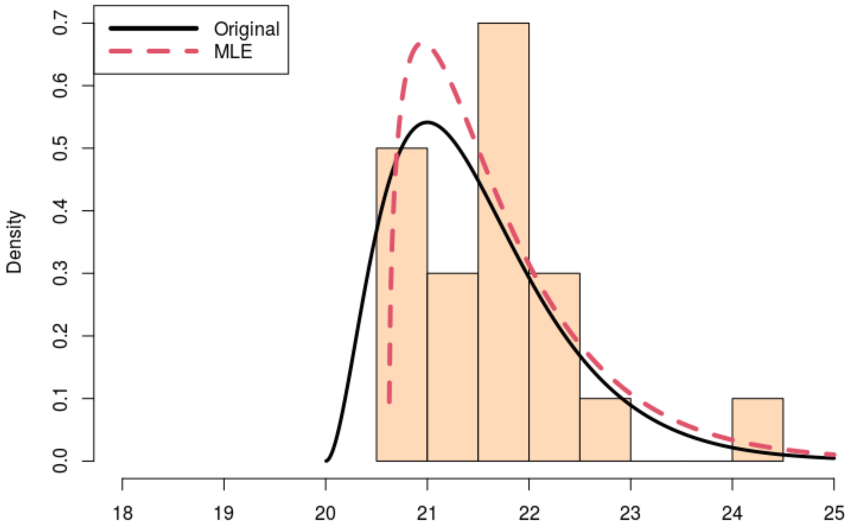


Figure 7 – Result of applying MLE to estimate the population minimum for a sample taken from a $20 + \text{Gamma}(3, 1/2)$ distribution.

Source – Prepared by the author.

4.1 Distance Between Probability Measures

In order to investigate means to fix the problems discussed above, we need to first establish ways to measure the quality of a solution, in a way that hopefully allows us to compare different solutions and tell which one is better. In the case of population minimum estimation (or endpoint estimation), one would use the bias or mean squared error, for example. Here, we consider a solution to be good if the inference process resulted in an estimated probability distribution that is as similar as possible to the actual generating process. Thus, we need to compare probability distributions. From here on, we let F, G be two distinct cdfs and f, g their corresponding pdfs.

A look into Figure 6 might tempt us to use the area between the curves as a measure of their divergence. Unfortunately, however, there is insufficient statistical ground to for that. A number of metrics for probability measures are given in (GIBBS; SU, 2002) and (CHUNG *et al.*, 1989). The Hellinger distance (HELLINGER, 1907) is defined as:

$$d_H(f, g) := \left[\frac{1}{2}\int_{\mathbb{R}} \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx\right]^{1/2}. \tag{4.1}$$

The original definition is actually much more general, but can be put in the above form if we conveniently choose the Lebesgue measure whenever an arbitrary measure is required. One property here is that $d_H(f, g) \in [0, 1]$. This distance will be related to the area between the curves $\sqrt{f(x)}$ and $\sqrt{g(x)}$; indeed, d_H is proportional to the L^2-norm (ROYDEN, 1988) of $\sqrt{f(x)} - \sqrt{g(x)}$, but precisely because of the power of two, this distance gives more weight to larger differences, and tends to tolerate smaller differences among the two functions.

Another distance of interest is the Kantorovich–Wasserstein (KW) distance (GIBBS; SU, 2002):

$$d_{KW}(F, G) := \int_{-\infty}^{\infty} |F(x) - G(x)|\, dx, \tag{4.2}$$

which is unbounded, $d_{KW} \in [0, \infty)$. This distance has the immediate interpretation of measuring how much the two distributions differ in terms of the probabilities they assign to the event $[X < x]$. d_{KW} captures the overall discrepancy among the two cdfs; alternatively, we could take the maximum discrepancy, in which case we would have the Kolmogorov metric (VAART, 2000):

$$d_K(F, G) := \sup_{x \in \mathbb{R}} |F(x) - G(x)|\,. \tag{4.3}$$

However, besides being computationally cumbersome to compute, it offers a somewhat smaller picture of the discrepancy between the distributions, so we do not use it here.

As for the example problem given in the introduction of this chapter (see Figures 6 and 7), the results in terms of the Hellinger and KW metrics are shown in Table 1. It can be seen that the best model according to the Hellinger metric was the one where we guessed the population minimum as being 19, whereas the KW distance says population minimum 20.5 is better, together with the MLE method. The area between the curve is only shown for comparison with the other

Model	**Area Between**	**Hellinger**	**Kantorovich**
$\hat{m} = 16$	0.397	0.061	0.263
$\hat{m} = 19$	0.360	0.049	0.248
$\hat{m} = 20.5$	0.283	0.103	0.217
MLE ($\hat{m} = 20.62$)	0.279	0.141	0.217

Table 1 – Distance between the estimated distributions and the original underlying distribution for the example given in this chapter's introduction, rounded to three decimal places.

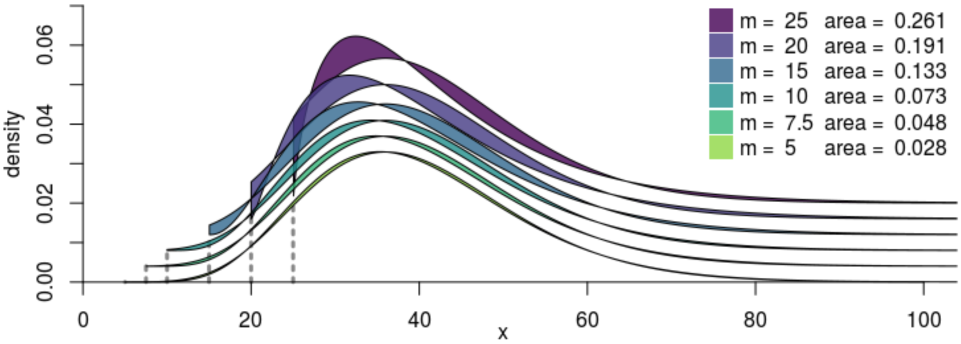


Figure 8 – Second example of inference when the population minimum is unknown. The graph shows the area between the inferred distribution and the actual distribution. The areas have been displaced on the y-axis to ease visualization.

Source – Prepared by the author.

metrics.

In order to further harness intuition about the metrics and the problem at hand, a second example is illustrated in Figure 8. This time, however, the real population minimum is zero; the challenge here is that the distribution has a long left tail, so that a sample from it may look as if the population minimum was far from zero. Figure 8 shows what happens when performing MLE after estimating the population minimum as being $5, 7.5, 10, 15, 20$ and 25. Table 2 shows the distance between the estimated distributions and the actual one.

There are two important points worth noticing here. First, both metrics follow the same overall trend as the area between the curves, so intuition holds strongly with these two metrics (unlike the Kolmogorov distance, for example). Second, even though the case $m = 5$ gives the best resuls, the metrics and the graph itself indicate that even $m = 7.5$ gives a fairly good estimated distribution, thus raising the question of whether it is really necessary to have a good estimate for the population minimum. It appears that the longer the left tail is, the less important it becomes to find the real population minimum, being sufficient to find a point that is located sufficiently to the left of the sample minimum.

Model	**Area Between**	**Hellinger**	**Kantorovich**
$\hat{m} = 5$	0.028	6.24E−4	0.229
$\hat{m} = 7.5$	0.048	1.75E−3	0.392
$\hat{m} = 10$	0.073	4.01E−3	0.605
$\hat{m} = 15$	0.133	1.53E−2	1.07
$\hat{m} = 20$	0.191	4.61E−2	1.38
$\hat{m} = 25$	0.261	1.17E−1	2.16

Table 2 – Distance between the estimated distributions and the original underlying distribution for the second example mentioned in the text.

4.2 A Few Preliminary Considerations

Let X be a random variable with pdf $f(x)$, cdf $F(x)$ and support $[0, \infty)$; this is without loss of generality since it is simple to extend the discussion here to supports of type $[m, \infty)$ and even $(-\infty, \mathcal{M}]$. Suppose, however, that the population minimum is not known to be zero. Thus, in an attempt to mitigate the problems caused by lack of knowledge of the left endpoint, we estimate it as $\hat{m}$, so the family of distributions used for inference has the form $g(x \,|\, \boldsymbol{\theta}), x > \hat{m}$. We are then interested in finding the distribution that minimizes the Hellinger distance $d_H(f, g)$ given in Equation (4.1). Assuming $\hat{m} > 0$ we have:

$$d_H(f,g) = \left[\frac{1}{2}\int_{\mathbb{R}} \left(\sqrt{f(x)} - \sqrt{g(x \,|\, \boldsymbol{\theta})}\right)^2 dx\right]^{1/2}$$

$$= \left[\frac{1}{2}\int_0^{\hat{m}} \left(\sqrt{f(x)} - \sqrt{g(x \,|\, \boldsymbol{\theta})}\right)^2 dx + \frac{1}{2}\int_{\hat{m}}^{\infty} \left(\sqrt{f(x)} - \sqrt{g(x \,|\, \boldsymbol{\theta})}\right)^2 dx\right]^{1/2}$$

But since $g(x \,|\, \boldsymbol{\theta}) = 0$ for every $x \leq \hat{m}$, the left hand term simplifies to $\sqrt{f(x)}^2 = f(x)$:

$$= \left[\frac{1}{2}\int_0^{\hat{m}} f(x)dx + \frac{1}{2}\int_{\hat{m}}^{\infty} \left(\sqrt{f(x)} - \sqrt{g(x \,|\, \boldsymbol{\theta})}\right)^2 dx\right]^{1/2}$$

$$= \left[\frac{F(\hat{m})}{2} + \frac{1}{2}\int_{\hat{m}}^{\infty} \left(\sqrt{f(x)} - \sqrt{g(x \,|\, \boldsymbol{\theta})}\right)^2 dx\right]^{1/2}, \tag{4.4}$$

whereas if $\hat{m} < 0$ we would analogously have:

$$d_H(f,g) = \left[\frac{G(0 \,|\, \boldsymbol{\theta})}{2} + \frac{1}{2}\int_0^{\infty} \left(\sqrt{f(x)} - \sqrt{g(x \,|\, \boldsymbol{\theta})}\right)^2 dx\right]^{1/2}.$$

The case $\hat{m} < 0$ is somewhat uninteresting because G is not predetermined, and MLE will often be able to make the term $G(0 \,|\, \boldsymbol{\theta})$ very small. In the case $\hat{m} > 0$, since the functions

f and F are defined by the phenomena being observed, the term $F(\hat{m})/2$ is of paramount importance to us. Equation (4.4) shows that the distance from f to g is lower bounded by $\sqrt{F(\hat{m})/2}$. If we want the lower bound to be very small, such as 0.01, then we need:

$$\sqrt{\frac{F(m)}{2}} = 0.01$$

$$m = F^{-1}(2 \cdot 0.01^2),$$

where the right hand side can be estimated, and since the desired quantile is very small, extreme quantile estimators should be used here. One problem here is that, if the sample size is large enough ($N \approx 5000$ in this particular case), then the estimate $\hat{m}$ will likely be larger than the sample minimum, which does not make sense. To solve that, it is necessary to further lower the threshold 0.01 chosen earlier.

Another insight given by Equation (4.4) concerns the second term in the sum. Clearly, that integral will be zero if we find g to be:

$$g(x \,|\, \boldsymbol{\theta}) = f(x), \ \text{for } x > \hat{m},$$

but then g is not a density since:

$$\int_{\mathbb{R}} g(x \,|\, \boldsymbol{\theta}) dx = \int_{\hat{m}}^{\infty} f(x) dx = 1 - F(\hat{m}) \neq 1,$$

so the tentative optimal solution must be modified to integrate to one. It immediately comes to mind to normalize the function:

$$g(x \,|\, \boldsymbol{\theta}) = \frac{f(x)}{1 - F(\hat{m})}, \ \text{for } x > \hat{m},$$

in which case g becomes the truncated version of the distribution f. We are tempted to believe that this g is the optimal solution among all densities supported on $[\hat{m}, \infty)$, but so far we have no reference for that, nor succeeded in proving this result. If this was true, the truncated version would serve as a benchmark to compare our experimental results against. Figure 9 illustrates how good the truncated distribution seems to be, especially when compared to the distribution inferred by maximum likelihood.

The approximation here consists of considering that the support of the random variable X begin at a certain m that is not the true one. We then would like to model the data under such a consideration; that is, find a model for Y. If we have knowledge about the distribution family of X and that its support begins at zero, then a good fit (asymptotically) would be achieved by selecting the distribution of Y as being a truncated version of the distribution of X (see Figure

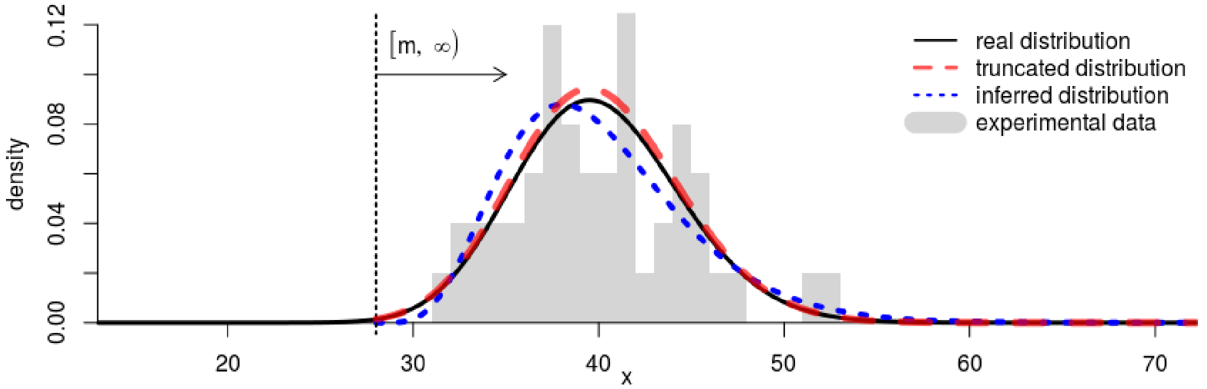


Figure 9 – The dashed red line shows the truncated version of the underlying actual distribution (black solid line), whereas the blue dotted line shows what can be achieved by MLE using the same family of distributions (gamma) to which the real distribution belongs.

Source – Prepared by the author.

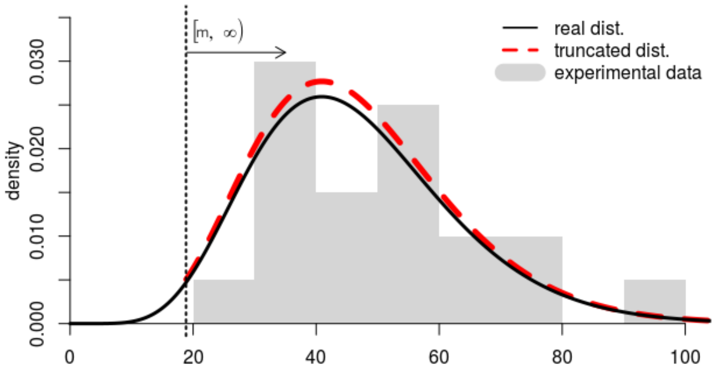


Figure 10 – Example of the first scenario analyzed. The underlying phenomenon is represented by a variable X whose distribution is shown as the solid line. From such a distribution we take a sample (light grey histogram). Then we choose m using methods to be discussed later, and fit a truncated distribution over $Y \approx X - m$ (dashed line).

Source – Prepared by the author.

10), given by

$$f_Y(y \mid \theta) = \frac{f_X(y + m \mid \theta)}{1 - F_X(m)}, \;\; y \in [0, \infty)$$

where f_Y, f_X are densities, F_x is a cumulative distribution function (cdf), θ is a parameter vector and m is given. Notice that $f_Y(y) = \lambda f_X(y + m)$ for a constant $\lambda = 1/(1 - F_X(m))$. Because of that, the likelihood over a sample $y_1, \ldots, y_n$ is

$$\mathcal{L}_{f_Y}(\theta \mid y_1, \ldots, y_n) = \prod_i f_Y(y_i \mid \theta) = \cdots$$

and if $y_i > 0$ for all i, then

$$\cdots = \prod_i \lambda f_X(y_i + m \mid \theta) = \prod_i \lambda f_X(x_i \mid \theta) = \lambda^n \mathcal{L}_{f_X}(\theta)$$

so that any θ maximizing the likelihood for f_X will also maximize $\mathcal{L}_{f_Y}$, proving our initial statement that the best asymptotic fit would be a truncated version of X's distribution. This happens regardless of m, so if we allowed m to also be optimized, then it would be chosen to maximize $\lambda^n = 1/(1 - F_X(m))$; from the monotonicity of F_X maximization happens when m approaches the sample minimum $x_{(1)}$. We cannot have $m \geq x_{(1)}$ because in such a case at least one of the y_i would be negative, thus making f_Y and the likelihood $\mathcal{L}_{f_Y}$ be zero.

The fact that m has no influence in the best parameter θ found by Maximum Likelihood Estimation (MLE) is actually a problem here. Although truncation allows us to shift the support origin, it does not help us with our original objective of making the space of initial parameters easier to design. This is related to the moments of Y and X being proportional to one another ($E[e^X] = e^m E[e^Y]$).

4.3 Proposed Methods for Performing the Inference Procedure

We are now in position to enumerate the methods that will be compared, including methods that follow directly from the literature as well as our proposals for solving the problem. We begin by the ones that involves no contribution of ours.

A) Estimating the population minimum using endpoint estimators. Endpoint estimators were thoroughly discussed and reviewed in Section 3.1. It is straightforward to apply them. If $x_1, \ldots, x_n$ is a random sample, $\hat{m}$ is the endpoint estimate and $f(\cdot \mid \boldsymbol{\theta})$ a family of pdfs supported on $[0, \infty)$, then inference consists of performing MLE on the modified samples $x_1 - \hat{m}, x_2 - \hat{m}, \ldots, x_n - \hat{m}$ using the family of functions $f(\cdot \mid \boldsymbol{\theta})$.

B) Inferring the population minimum by MLE. Let $x_1, \ldots, x_n$ be a random sample and $f(\cdot \mid \boldsymbol{\theta})$ a family of pdfs supported on $[0, \infty)$, then we can add a location parameter to this family of distributions:

$$g(x \mid \boldsymbol{\theta}, m) = \begin{cases} f(x - m \mid \boldsymbol{\theta}), & \text{if } x \geq m \\ 0, & \text{otherwise,} \end{cases}$$

so that we can perform MLE using the family $g(x \mid \boldsymbol{\theta}, m)$ over the sample $x_1, \ldots, x_n$, which will yield an MLE estimate $\hat{m}$ for the population minimum. Some problems of this approach were discussed in this chapter's introductory section. One caveat here is that the parameter m

should be constrained to the interval $\mathscr{m} < x_{(1)}$, a detail that should be told to the computational optimizer used, in order to perform MLE without problems.

The methods we propose are as follows, beginning by our flagship.

I) Maximum Likelihood with Parameter Dependent Support (MLEPDS). Inference by MLE begins with the assumption that the underlying distribution comes from a certain family. Under this assumption, we do have a lot of information about the underlying cdf and pdf. We intend to use this information to our advantage here.

Let X be a random variable with cdf $F(x - \mathscr{m} \mid \boldsymbol{\theta})$, so F itself is supported on $[0, \infty)$, but X is supported on $[\mathscr{m}, \infty)$. We know that the cdf of the sample minimum is given by (MOOD, 1950):

$$F_{\min}(x \mid \boldsymbol{\theta}) = 1 - [1 - F(x \mid \boldsymbol{\theta})]^n \,. \tag{4.5}$$

By inverting this equation we obtain the quantile function of the minimum (on $[0, \infty)$ rather than $[\mathscr{m}, \infty)$):

$$F^{-1}_{\min}(q \mid \boldsymbol{\theta}) = F^{-1}\left(1 - (1 - q)^{1/n} \,\middle|\, \boldsymbol{\theta}\right) \tag{4.6}$$

With this, the median and other quantiles of the sample minimum are given by $F^{-1}_{\min}(q \mid \boldsymbol{\theta})$. The median can be seen as a good “guess” for what the sample minimum would be if: (i) $\boldsymbol{\theta}$ was the true parameter and (ii) the random variable was shifted back to have support $[0, \infty)$ (i.e., take $X - \mathscr{m}$). Having determined the desired guess, the sample should be shifted so that the sample minimum coincides with it. That is, we want to find the constant m so that our sample is modified to $x_1 - m, x_2 - m, \ldots, x_n - m$ and so that the following holds $x_{(1)} - m = F^{-1}_{\min}(q \mid \boldsymbol{\theta})$; simple manipulation leads to $m = x_{(1)} - F^{-1}_{\min}(q \mid \boldsymbol{\theta})$. The hope here is that $X - m$ will be supported on $[0, \infty)$, which would effectively imply that:

$$m = x_{(1)} - F^{-1}_{\min}(q \mid \boldsymbol{\theta}) = \mathscr{m}.$$

If we were lucky and $x_{(1)}$ coincided precisely with the median of the sample minimum distribution, then the above equation holds exactly when $\boldsymbol{\theta}$ is the true parameter. However, this “estimate” will often be wrong because we do not have knowledge of the true parameter, and often even the distribution family used for inference will not perfectly contain the distribution that generated the sample. We argue, however, that once chosen a family of distributions to use for inference, this method will generate very good inferred distributions (better than through other methods) when it comes to the distance between the estimated distribution and the actual underlying distribution, using the metrics discussed in Section 4.1.

In any case, up to this point, then, we have the sample:

$$x_1 - x_{(1)} + F_{\min}^{-1}(q \,|\, \boldsymbol{\theta}),\ x_2 - x_{(1)} - F_{\min}^{-1}(q \,|\, \boldsymbol{\theta}),\ \ldots,\ x_n - x_{(1)} - F_{\min}^{-1}(q \,|\, \boldsymbol{\theta}). \tag{4.7}$$

Since $F_{\min}^{-1}(q \,|\, \boldsymbol{\theta})$ only depends on the parameter $\boldsymbol{\theta}$, we include this term within the family of distributions used for inference, as if it were a location parameter. If the set of functions originally considered was $\{f(x \,|\, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Omega\}$, now it becomes:

$$\left\{f\left(x + F_{\min}^{-1}(q \,|\, \boldsymbol{\theta}) \,\middle|\, \boldsymbol{\theta}\right) \,:\, \boldsymbol{\theta} \in \Omega\right\},$$

and fit this modified family onto the samples $x_1 - x_{(1)}, x_2 - x_{(1)}, \ldots, x_n - x_{(1)}$ instead. Thus, we manage to remove the terms in Equation (4.7) that depend on the parameter $\boldsymbol{\theta}$. Now define $Y_i = X_i - X_{(1)}$ so the sample becomes simply $y_1, y_2, \ldots, y_n$. If we can prove this sample is iid, important results about maximum likelihood estimates will also apply here. Identical distribution is straightforward, but independence is not. Given the lack of an immediate reference for the following result, we enunciate and prove the theorem below.

Theorem 3. *Let $X_1, \ldots, X_n$ be a random sample from a continuous random variable with pdf $f(x)$, cdf $F(x)$ and support $[\mathscr{m}, \infty)$; that is,* $\inf\{x : F(x) > 0\} = \mathscr{m}$. *Then the modified sample $X_1 - X_{(1)}, X_2 - X_{(1)}, ..., X_n - X_{(1)}$ is independent and identically distributed.*

Proof (Theorem 3). Identical distribution is straightforward. To prove independence, consider the iid sample $X_1, \ldots, X_n$ with the sample minimum $X_{(1)}$. Then we have:

$$P\left(\bigcup_{i=1}^{n} [X_i > a_i] \cup X_{(1)} > a_{n+1}\right)$$

and since every X_i is larger than or equal to $X_{(1)}$:

$$= P\left(\bigcup_{i=1}^{n} [X_i > \max\{a_i, a_{n+1}\}]\right)$$

by the independence of the sample, we have:

$$\begin{aligned} &= \prod_{i=1}^{n} P\left(X_i > \max\{a_i, a_{n+1}\}\right) \\ &= \prod_{i=1}^{n} \left[1 - F(\max\{a_i, a_{n+1}\})\right]. \end{aligned} \tag{4.8}$$

With analogous steps, we can obtain for each i:

$$\begin{aligned} &P(X_i > a_i \cup X_{(1)} > a_{n+1}) \\ &\qquad = P\Bigg(X_i > a_i \cup \bigcup_{i=1}^{n} [X_i > a_{n+1}]\Bigg) \\ &\qquad = P\Bigg(X_i > \max\{a_i, a_{n+1}\} \cup \bigcup_{j\neq i} \left[X_j > a_{n+1}\right]\Bigg) \\ &\qquad = P(X_i > \max\{a_i, a_{n+1}\}) \cdot \prod_{j\neq i} P\left(X_j > a_{n+1}\right) \\ &\qquad = [1 - F(\max\{a_i, a_{n+1}\})] \cdot [1 - F(a_{n+1})]^{n-1} \,. \end{aligned} \tag{4.9}$$

Now consider the variable transform:

$$\begin{cases} U_i & = X_1 - X_{(1)}, \text{ for } i = 1, \dots, n \\ U_{n+1} & = X_{(1)} \end{cases} \xRightarrow{\text{invert}} \begin{cases} X_i & = U_i + U_{n+1}, \text{ for } i = 1, \dots, n \\ X_{(1)} & = U_{n+1} \,. \end{cases}$$

Note that since $X_{(1)}$ is the smallest value in the sample, we have that $U_i > 0$ for $i = 1, \dots, n$. The above transform means that:

$$P\left(\bigcup_{i=1}^{n+1} [U_i > u_i]\right) = P\left(\bigcup_{i=1}^{n} [X_i > u_i + u_{n+1}] \cup X_{n+1} > u_{n+1}\right)$$

since we already obtained the joint distribution of $X_1, \dots, X_n, X_{(1)}$ in Equation (4.8):

$$= \prod_{i=1}^{n} [1 - F(\max\{u_i + u_{n+1}, u_{n+1}\})]$$

but as $u_i > 0$, the maximum is known:

$$= \prod_{i=1}^{n} [1 - F(u_i + u_{n+1})] \,.$$

Take the marginal of this cumulative probability function over U_{n+1}, noting that the limit of the event $U_{n+1} > u_{n+1}$ as $u_{n+1} \to -\infty$ is equivalent to $U_{n+1} > m$:

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} [U_i > u_i]\right) &= \lim_{u_{n+1}\to-\infty} P\left(\bigcup_{i=1}^{n+1} [U_i > u_i]\right) \\ &= P\left(\bigcup_{i=1}^{n} [U_i > u_i] \cup U_{n+1} > m\right) \end{aligned}$$

$$= \prod_{i=1}^{n} \left[1 - F(u_i + m)\right] . \tag{4.10}$$

Now consider the transform $U_i = X_i - X_{(1)}$ and $U_{n+1} = X_{(1)}$, we obtain the inverse $X_i = U_i + U_{n+1}$ and $X_{(1)} = U_{n+1}$, which means:

$$\begin{aligned}
P(U_i > u_i, U_{n+1} > u_{n+1}) &= P(X_i > u_i + u_{n+1}, X_{(1)} > u_{n+1}) \\
&= \left[1 - F(\max\{u_i + u_{n+1}, u_{n+1}\})\right] \left[1 - F(u_{n+1})\right]^{n-1} \text{ from Equation (4.9)} \\
&= \left[1 - F(u_i + u_{n+1})\right] \left[1 - F(u_{n+1})\right]^{n-1} ,
\end{aligned}$$

and taking the marginal over U_{n+1}:

$$\begin{aligned}
P(U_i > u_i) &= \lim_{u_{n+1} \to -\infty} P(U_i > u_i, U_{n+1} > u_{n+1}) \\
&= P(U_i > u_i, U_{n+1} > m) \\
&= \left[1 - F(u_i + m)\right] \cancel{\left[1 - F(m)\right]^{n-1}} \text{ since } F(m) = 0 \\
&= \left[1 - F(u_i + m)\right] . \qquad (4.11)
\end{aligned}$$

Finally, comparing Equations (4.10) and (4.11), we see that the joint cumulative distribution of the $\{U_i : i = 1, \ldots, n\}$ is the product of the individual cumulative distributions:

$$P\left(\bigcup_{i=1}^{n} [U_i > u_i] \right) = \prod_{i=1}^{n} \left[1 - F(u_i + m)\right] = \prod_{i=1}^{n} P(U_i > u_i)$$

concluding the proof of independence of the sample $U_1, \ldots, U_n$, that is equivalent to $X_1 - X_{(1)}, \ldots, X_n - X_{(1)}$. □

Recapitulating, we are given a random sample $X_1, \ldots, X_n$ from a distribution family $F(x \mid \boldsymbol{\theta})$, and we modify it by taking $Y_i = X_i - X_{(1)}$, thus obtaining the new sample:

$$X_1 - X_{(1)}, X_2 - X_{(1)}, \ldots, X_n - X_{(1)} .$$

Over this modified sample, we perform maximum likelihood estimation to find the best density within the family:

$$\mathcal{F} = \left\{ f\left(x + F_{\min}^{-1}(q \mid \boldsymbol{\theta}) \,\middle|\, \boldsymbol{\theta} \right) \;:\; \boldsymbol{\theta} \in \Omega \right\} ,$$

where q is the quantile we believe will make:

$$X_{(1)} \approx m + F_{\min}^{-1}(q \mid \boldsymbol{\theta})$$

for most values of $\boldsymbol{\theta}$, where we note that $q = 0.5$ seems to be a good, parsimonious value. Theorem 3 proves that the new sample $Y_1, \ldots, Y_n$ is iid, so that the rich theory on maximum likelihood estimators also apply here. That is, if the distribution that generate the sample $Y_1, \ldots, Y_n$ is contained within $\mathcal{F}$, then the MLE estimate is consistent, asymptotically normal and so on.

In the chapter introduction we had presented the performance of existing methods on an example (Table 1). Table 3 complements these results by also showing the performance of our proposed MLEPDS method, using $q = 0.5$. Figure 11 extends Figure 6 with the MLEPDS method. Both the table and the figure display a remarkable superiority of our proposal against the other approaches, indicating that we must be on the right path here. Further experiments are left to their own section (see Section 4.4).

Model	**Area Between**	**Hellinger**	**Kantorovich**
$\hat{m} = 16$	0.397	0.061	0.263
$\hat{m} = 19$	0.360	0.049	0.248
$\hat{m} = 20.5$	0.283	0.103	0.217
MLE ($\hat{m} = 20.62$)	0.279	0.141	0.217
MLEPDS ($\hat{m} = 19.998$)	0.114	0.006	0.093

Table 3 – Distance between the estimated distributions and the original underlying distribution, rounded to three decimal places.

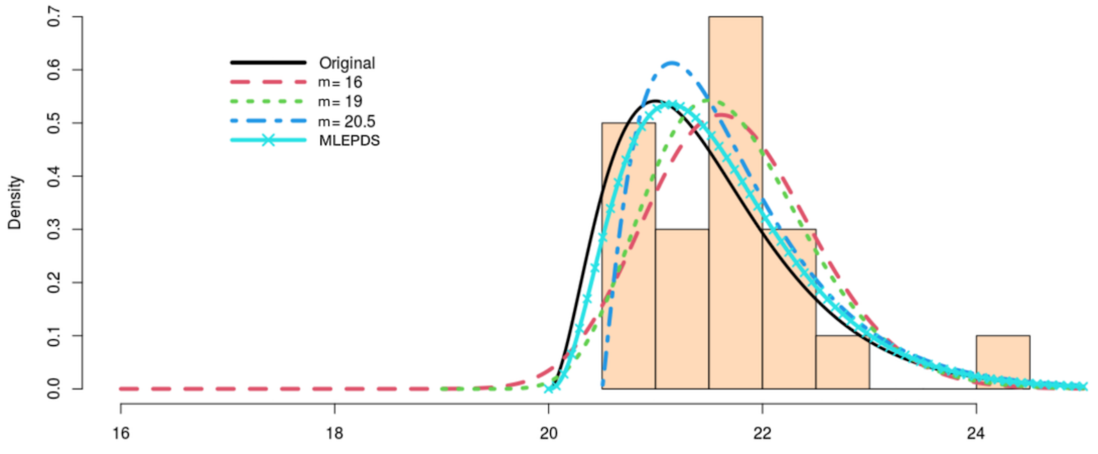


Figure 11 – Inferred distributions plotted against the histogram of the sample as well as the original density function that generated the sample.

Source – Prepared by the author.

II) Biased estimators for the left endpoint. Considering what we discussed in Sections 4.1 and 4.2, we tend to believe that using unbiased estimators of the population minimum is not optimal if we want to minimize the Hellinger or Kantorovich metrics. This will be a particular problem if the underlying distribution has a long left tail. In these cases, we would rather have a biased estimate $\hat{m}$ that is sufficiently near the sample minimum, but with a sufficiently small cumulative density $F(\hat{m})$. We also value computational complexity of the inference process, so in the following we propose computationally simple estimators.

We consider additive estimators, that is, those of the form $\hat{m}(x_1, \ldots, x_n) = x_{(1)} - h(x_1, \ldots, x_n)$. This is in contrast to multiplicative ones as found in (YAZIDI; HAMMER, 2017) for example. The reason is that it allows for some rather intuitive properties to hold. Define the population minimum:

$$m_X = \inf\{x \mid F(x) > 0\}, \tag{4.12}$$

then it is clear that for linear transformations of X we have:

$$m_{X+a} = \inf\{x \mid F(x-a) > 0\} = \inf\{x + a \mid F(x) > 0\} = m_X + a \tag{4.13}$$

$$m_{aX} = \inf\{x \mid F(x/a) > 0\} = \inf\{ax \mid F(x) > 0\} = a m_X. \tag{4.14}$$

The populational value follows these properties, so it is only natural to expect them to hold for the corresponding estimators, and it is particularly difficult to ensure location invariance when using a multiplicative estimator of type $\hat{c}(x_1, \ldots, x_n) = x_{(1)} \cdot h(x_1, \ldots, x_n)$.

Although there are many parametric approaches for estimating quantiles, estimating very low quantiles (< 0.05) is a problem that has not yet been solved in a sufficiently general way. That is, most parametric solutions rely on assumptions about the underlying distribution or quantile functions; constraints on the derivative of the pdf, for example (see Daouia and Simar (2007), Alexopoulos *et al.* (2019) and Section 2.1). To maintain generality, we opt for nonparametric approaches, using as main tool the empirical cdf $F_n(x)$, whose uniform convergence to $F(x \mid \theta)$ is given by the Glivenko-Cantelli theorem (for an iid sampling). Using the law of iterated logarithm (VAPNIK, 1998), which gives the rate of convergence of $F_n(x)$ to $F(x \mid \theta)$, and also using that $F_n(x_{(1)} - \epsilon)$ is zero for any $\epsilon > 0$, we must have for sufficiently large n:

$$F(x_{(1)} - \epsilon \mid \theta) \leq F_n(x_{(1)} - \epsilon) + \sqrt{\frac{\ln \ln n}{2n}} \longrightarrow \sqrt{\frac{\ln \ln n}{2n}}.$$

As this number decreases, the less we can expect the population minimum to be lower than the actual sample minimum. Hence, we propose the following estimator:

$$\hat{m}(x_1, \ldots, x_n) = \overline{m} - (X_{(0.5(n+1))} - \overline{m})\sqrt{\frac{\ln \ln n}{2n}}, \tag{4.15}$$

where we also embody the hope that the deviation between population minimum and sample minimum be proportional to the distance between sample minimum and median (where the appropriate interpolation is done when $0.5(n+1)$ is not integer). There are two advantages brought by using the distance between sample minimum and median: **(i)** it ensures that properties in Equations (4.13) and (4.14) hold, and **(ii)** it has same measurement unit as $\overline{m}$, so from a dimensional analysis point of view (GOLDBERG, 2006), it makes more physical sense than merely subtracting the dimensionless root term.

Notice that the cdf for the sample minimum, shown in Equation (4.5), can be promptly used to show that the sample minimum, as a random variable, follows $\overline{m}_{aX} = a\overline{m}_X$, and $\overline{m}_{X+a} = \overline{m}_X + a$, and this applies to any order statistic, although it is slightly more difficult to prove due to their more complex pdfs and cdfs (MOOD, 1950).[2] So, letting $Y = X + a$ and $Z = aX$, it follows that (consider $a > 0$ for simplicity):

$$\begin{aligned}
\hat{m}_Y &= \overline{m}_Y - (X_{(0.5(n+1)),Y} - \overline{m}_Y)\sqrt{\frac{\ln\ln n}{2n}} \\
&= \overline{m}_X + a - (X_{(0.5(n+1)),X} + \cancel{a} - (\overline{m}_X + \cancel{a}))\sqrt{\frac{\ln\ln n}{2n}} = \hat{m}_X + a, \text{ and} \\
\hat{m}_Z &= \overline{m}_Z - (X_{(0.5(n+1)),Z} - \overline{m}_Z)\sqrt{\frac{\ln\ln n}{2n}} \\
&= a\overline{m}_X - (aX_{(0.5(n+1)),X} - a\overline{m}_X)\sqrt{\frac{\ln\ln n}{2n}} = a\hat{m}_X.
\end{aligned}$$

We now prove the following theorem.

Theorem 4. *Let X be a continuous random variable in $\mathbb{R}$ with population minimum m and cdf $F(x)$ strictly increasing in $[m, \infty)$. Let $X_1, \dots, X_n$ be a random sample from X. Then, the proposed estimator $\hat{m}(X_1, \dots, X_n)$ is a consistent estimator for the population minimum.*

Proof. Let $A_n = X_{(0.5(n+1))} - \overline{m}$. By definition, the above holds if and only if for every $\epsilon > 0$ we have:

$$\lim_{n\to\infty} P\left(\left|\overline{m} - A_n\sqrt{\frac{\ln\ln n}{2n}} - m\right| < \epsilon\right) = 1. \tag{4.16}$$

By triangular inequality and the fact that $\overline{m} > m$, we have the implication:

$$\overline{m} - m < \frac{\epsilon}{2} \quad \& \quad A_n\sqrt{\frac{\ln\ln n}{2n}} < \frac{\epsilon}{2} \quad \Longrightarrow \quad \left|\overline{m} - A_n\sqrt{\frac{\ln\ln n}{2n}} - m\right| < \epsilon,$$

which implies that:

$$P\left(\overline{m} - m < \frac{\epsilon}{2}\,,\ A_n\sqrt{\frac{\ln\ln n}{2n}} < \frac{\epsilon}{2}\right) \quad \leq \quad P\left(\left|\overline{m} - A_n\sqrt{\frac{\ln\ln n}{2n}} - m\right| < \epsilon\right),$$

---

[2] In this section, let $x_{(1),h(X)}$ be the first order statistic relative to random variable $h(X)$.

so it suffices to prove that the left side goes to 1. The main difficulty arises from the lack of independence between A_n and the sample minimum. We rewrite it as:

$$
\begin{aligned}
&\mathrm{P}\left(A_n\sqrt{\frac{\ln\ln n}{2n}} < \frac{\epsilon}{2}\ ,\ \overline{m} < \frac{\epsilon}{2}+m\right)\\
&= \mathrm{P}\left(X_{(0.5(n+1))} < \frac{\epsilon}{2}\sqrt{\frac{2n}{\ln\ln n}}+\overline{m}\ ,\ \overline{m} < \frac{\epsilon}{2}+m\right)
\end{aligned}
$$

and since $\overline{m}$ is lower bounded by m:

$$
\geq \mathrm{P}\left(X_{(0.5(n+1))} < \frac{\epsilon}{2}\sqrt{\frac{2n}{\ln\ln n}}+m\ ,\ \overline{m} < \frac{\epsilon}{2}+m\right),
$$

but in the limit $n \longrightarrow \infty$, the term $\epsilon/2\sqrt{2n/\ln\ln n}\ + m$ goes to ∞. Since the sample median does not diverge (it converges in distribution due to the hypothesis that F is strictly increasing in $[m,\infty)$), we are left with the marginal cdf with just the right coordinate $\mathrm{P}(\overline{m} < \epsilon/2+m)$, which goes to 1 since the sample minimum converges in probability to the population minimum. To see this, just consider the cdf of the sample minimum $F_{\overline{m}}(x) = 1-[1-F_X(x)]^n$ which goes to 0 for $x > m$ and to 1 when $x \leq m$. Since the probability above goes to 1, the proof is concluded. □

Besides the law of iterated logarithm, the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality (MASSART, 1990) can also be invoked, which provides a different way to view the estimator. The inequality is:

$$
P(\sqrt{n}\sup_x|F_n(x)-F(x)| > \lambda) \leq 2\exp(-2\lambda^2)
$$

and by doing the necessary manipulations, we derive that the following will hold with probability of at least $1-\nu$:

$$
\sup_x|F_n(x)-F(x)| \leq \sqrt{\frac{-\ln(\nu/2)}{2n}}
$$

so if we choose ν to be very low, we can expect $F(x_{(1)}-\epsilon)$ to be lower than or equal to the right-side of the above equation. Following the same logic as previously, we define another estimator:

$$
\hat{m}(x_1,\ldots,x_n) = \overline{m} - (X_{(0.5(n+1))}-\overline{m})\cdot\sqrt{\frac{-\ln(\nu/2)}{2n}} \tag{4.17}
$$

which offers a probabilistic view, instead of the previous asymptotic view given by the law of iterated logarithm; besides that, it also includes a configurable parameter ν, which offers more flexibility than the previous estimator. Figure 12 illustrates these estimators.

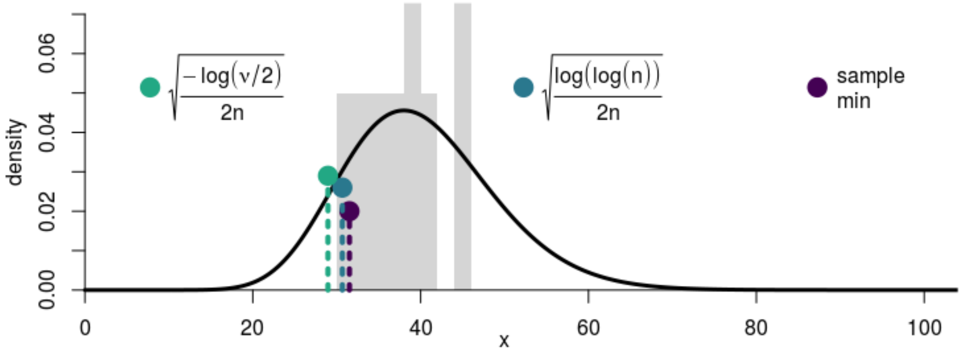


Figure 12 – The low quantile estimators based on the data represented by the histogram in light gray. The data was generated from the density shown as a black line.

Source – Prepared by the author.

III) Adapting extreme quantile estimators. Following our statement in item II, where we argued in favor of biased estimators, in this section we propose a way to use extreme quantile estimators to obtain biased estimates of the population minimum. We propose estimating a low quantile (say, the 0.05-quantile) of the sample minimum, which is related to the quantiles of the random variable according to the following simple theorem.

Theorem 5. *Let $X_1, \dots, X_n$ be a random variable with continuous and bijective cdf $F(x)$. Then, the q-quantile of the sample minimum corresponds to the $(1-(1-q)^{1/n})$-quantile of the random variable.*

Proof. Recall that the cdf for the sample minimum is $F_{\text{min}}(x) = 1-(1-F(x))^n$, so that its inverse gives the quantiles of the sample minimum, $F_{\text{min}}^{-1}(q) = F^{-1}(1-(1-q)^{1/n})$, which concludes the proof. □

Consider that our estimate is $\hat{m} = F_{\text{min}}^{-1}(0.05) = F^{-1}(1-0.95^{1/n})$, where $F^{-1}(1-0.95^{1/n})$ is calculated using one of the extreme quantile estimator presented in Section 3.2. Since the minimum converges in probability to the population minimum, any quantile is consistent, so this estimator is consistent. Also, since there is always 0.05 of cumulative density to the left of the estimator, long left tails tend not to be a problem, since they will be ignored. We were surprised that this result did not appear in any of the works reviewed in Chapter 3.

4.4 Experiments

In this section, we present experiments to validate our proposals, as well as compare them to the existing solutions. We compare our proposed methods — namely, the proposed MLEPDS method and the proposed estimator that is based on the law of iterated logarithm

(LIL)[3] — with the straightforward MLE solution discussed in Section 4.3 (item B), the Pickands' estimator (PICKANDS, 1975), the estimator proposed by Alves and Neves (2014) and shown in Equation (3.3), the one proposed by Girard, Guillou and Stupfler (2012a) (see Equation (3.7)), and the estimator of Hosking and Wallis (1987), which can be found in Equation (3.8).

We performed a simulation study considering the gamma and Weibull distributions, where samples of various sizes were generated for various combinations of distribution parameters. For each simulated sample, we perform inference using all the methods considered, which results in inferred distributions that are subsequently compared to the real underlying distribution, mainly by calculating the Hellinger and Kantorovich distances between the distributions.[4] We also collect the population minimum found by each method.

For both distributions considered, the collection of shape parameters was taken to be the set $\{0.5, 0.8, 1, 1.5, 4, 10, 20, 40\}$, and the set of scale parameters was defined as $\{0.5, 0.8, 1, 1.5, 4, 10, 40\}$; the cartesian product of these two sets defines the distributions simulated in this experiment. Besides that, experiments were replicated for simulated samples of sizes 20, 50, 100 and 500, each with a different population minimum initialized randomly over interval $[-100, 100]$ (this is to show that the methods work regardless of the particular location of the data). The experiment was repeated 50 times (i.e., the number of trials) for each combination of the triad: scale parameter, shape parameter and sample size.

Recall from Section 2.1 that Pickand's estimator operates on three of the lowest values of the sample, separated in a regular manner; in the experiments, we consider the values $x_{(1)}$, $x_{(2)}$ and $x_{(4)}$. We also set the estimator of Hosking and Wallis (1987) to operate on the 20% lowest values of the sample. As for the estimator of Girard, Guillou and Stupfler (2012a), the parameters (a, p_n) are set to $(2, 0.4)$, $(2, 5)$, $(10, 0.4)$ and $(10, 5)$ (values taken from the authors' paper), and only the best result is collected.

Table 4 shows initial experiments considering the gamma distribution, and Table 5 concerns the Weibull case. To obtain these tables, we collect the rank of each method on each setting of shape, scale and sample size; that is, the best method ranks first, the second best ranks second, and so forth. Recall that 50 trials were executed for each such combination, so we rank the methods based on their average Hellinger and Kantorovich distance in all fifty trials.

If the method did not work in a certain trial, we merely disconsider it without any effect on the mean; because of this, we also show in the table the proportion of times each method failed. The tables show the average rank over the different experimental settings, and also shows the average rank over the experimental settings where the simulated sample size was $N = 500$, which is intended to illustrate how the methods would behave in a hypothetical asymptotic case.

---

[3] We do not include the results achieved with the estimator based on the DKW inequality, since it was very similar than the LIL-based one.

[4] As a technical note, the distances were calculated using numerical integration by the Gauss–Kronrod quadrature method (KAHANER; MOLER; NASH, 1989), using a hundred million subdivisions, the largest number supported by the machine used.

		Pure MLE	MLEPDS .25	MLEPDS .5	MLEPDS .75	LIL	Pickands	Alves	Girard	Hosking
All N	**rank**	3.41	2.79	2.18	3.91	4.62	NA	5.41	5.71	NA
Hellinger	**errors**	0.01	0.08	0.05	0.03	–	0.80	–	0.33	0.88
$N = 500$	**rank**	2.11	3.09	2.13	3.84	5.87	NA	5.39	5.66	NA
Hellinger	**errors**	0.01	0.02	0.01	0.01	–	0.84	–	0.41	0.71
All N	**rank**	2.82	5.27	4.09	4.23	2.86	NA	3.43	5.34	NA
Kantorovich	**errors**	–	–	–	–	–	0.80	–	0.33	0.88
$N = 500$	**rank**	1.80	3.18	2.73	4.04	5.75	NA	5.39	5.20	NA
Kantorovich	**errors**	–	–	–	–	–	0.84	–	0.41	0.71

Table 4 – Average rank of each inference method on samples simulated from a gamma distribution. “MLEPDS X” stands for the proposed MLEPDS method with its hyperparameter q set to value X, and “LIL” for the estimator based on the law of iterated logarithm. The “errors” rows show the proportion of times that each method failed to provide a valid population minimum; “–” means zero errors, and 0.00 represent that the error rate was rounded to zero, but was not equal to zero.

		Pure MLE	MLEPDS .25	MLEPDS .5	MLEPDS .75	LIL	Pickands	Alves	Girard	Hosking
All N	**rank**	2.52	2.70	1.96	3.89	5.04	NA	6.43	5.46	NA
Hellinger	**errors**	0.01	0.05	0.03	0.02	–	0.81	–	0.50	0.86
$N = 500$	**rank**	1.96	2.96	2.02	3.66	6.25	NA	6.14	5.05	NA
Hellinger	**errors**	0.02	0.02	0.01	0.01	–	0.84	–	0.57	0.65
All N	**rank**	3.07	2.73	2.89	5.41	3.73	NA	4.96	5.20	NA
Kantorovich	**errors**	0.00	0.04	0.02	0.01	–	0.81	–	0.50	0.86
$N = 500$	**rank**	1.73	2.79	2.75	4.57	5.91	NA	5.95	4.36	NA
Kantorovich	**errors**	0.00	0.00	0.00	0.00	–	0.84	–	0.57	0.65

Table 5 – Average rank of each inference method on samples simulated from a Weibull distribution. Other details are as described in Table 4.

It can be observed in Tables 4 and 5 that, when considering all sample sizes, the pure MLE estimator often has a worse performance than the proposed MLEPDS estimator, usually ranking from 3rd to 4th in the gamma case, and from 2nd to 3rd in the Weibull case, according to the Hellinger distance.

One exception is the gamma case and considering the Kantorovich distance, which leads to the MLEPDS method having a worse performance than the pure MLE one. This is possibly due to inaccuracies of the integration algorithm. It might be better, for example, to divide the integration into the intervals $(-\infty, q_1]$, $[q_1, q_2)$, $[q_2, q_3)$, $[q_3, q_4)$ and $[q_4, \infty)$, where q_1, q_2, q_3, q_4 are the ordered 5% and 95% quantiles of both distributions involved in the calculation, but we leave that as future work. In any case, the results for $N = 500$ are very stable, so this data in the table should be reliable.

The pure MLE method displays better performance in all cases than the methods where the population minimum is estimated prior to performing MLE, and the same can be said of the MLEPDS method with hyperparameter $q = 0.5$. Besides that, we observe a significant improvement concerning the pure MLE when considering just the case where $N = 500$, which confirms its asymptotic strength. In this case, the pure MLE method competes somewhat closely with the the proposed MLEPDS method with $q = 0.5$, especially when using the Hellinger distance, and it is superior to the other configurations of the MLEPDS method.

Overall, the MLEPDS method with hyperparameters 0.75 showed a very bad performance, and the case $q = 0.25$ was not better than $q = 0.5$, but still showed interesting results and often competing closely with the pure MLE method. Finally, the best configuration of the MLEPDS method was with $q = 0.5$, but although it sometimes displayed better performance than the pure MLE method, most often it did not.

As it might already have been noticed, the right portions of the tables, beginning on column "LIL", show the methods that merely give an estimate for the population minimum prior to performing estimation by maximum likelihood. They are simpler in the sense that either they do not increase parameter count (as done by the MLE method), or they do not contribute to making the likelihood function more difficult to optimize (as done by the MLEPDS method). Because of this, we believe there is great value to be attained from also studying these estimators.

Among these estimators, it is immediately noticeable that some of them display a large number of errors. This is because their formulation are not made in a way that prevents the estimate for the population minimum to be lower than or equal to the sample minimum. Whenever the estimate is larger than the sample minimum, inference is made impossible as all parameters in the parameter space will assign probability 0 to the sample minimum and any other samples that happen to be smaller than the estimated population minimum.

In the tables, we observe a large number of errors from the estimators of Pickands, Hosking et al. and Girard et al. Since Pickands' and Hosking et al.'s estimators did not display a good performance, it seems more than reasonable to disregard them hereafter. As for Girard et al.'s estimator, it must be noted that it achieved a fairly good average rank in some cases. For example, in the Weibull scenario considering all sample sizes, Girard et al.'s estimator achieved an average rank of 5.46, which is better than the Alves et al. estimator, and for $N = 500$ it shows better performance than both Alves et al. and LIL estimators.

One the other hand, it did not provide a valid estimate in astounding ~40% of the cases; if these cases were considered to calculate the average ranks, it would certainly cause Girard et al.'s estimator to have a much worse result in the tables. This, together with the fact that we used four different hyperparameter configurations for their estimator and took only the best result, leads us to not suggest the use of this estimator.

As a last overall comparison, we show in Figure 13 the "bias distribution" resulting from each method, that is, the differences between their estimated population minimum and the actual one (i.e., $\hat{m} - m$). First we notice that there is a large number of outliers, no matter the method. For the gamma distribution, it is noticeable that the estimators on the right part (from LIL and

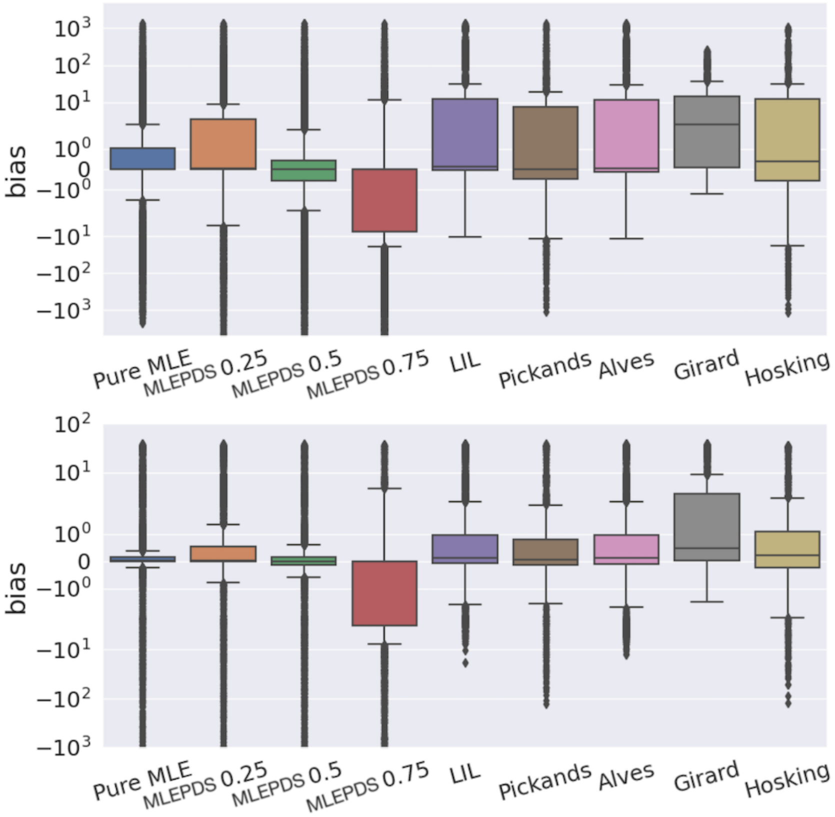


Figure 13 – Bias distribution of each method on each distribution used for the experiments (gamma above, Weibull on the bottom). Bias distribution here should be seen as the collection of differences $\hat{m} - m$ between the estimated population minimum and the actual minimum.

Source – Prepared by the author.

on) tend to overestimate the population minimum, whereas the MLE and MLEPDS estimators are more spread on both sides of the median; this also happens in the Weibull case, but to a smaller extent. The MLE and MLEPDS estimator with $q = 0.5$ display very small interquartile range (IQR), and with $q = 0.25$ it has a clear tendency to overestimate the population minimum, whereas the $q = 0.75$ case tends to underestimate it heavily.

The smaller IQR appears to be a clear advantage from these two kinds of estimators, but we do note here that their bias distributions have heavier tails, that is, their outliers go up to very large values: for the pure MLE, we have values up to −2200 (gamma) and −2E6 (Weibull), and for the MLEPDS with $q = 0.5$ we have −5E4 (gamma) and −3E7. Besides these, it is also remarkable that the overall biases found for the gamma distribution are that much larger than for the Weibull distribution. We believe it is likely due to the big difference in the behavior of the left-tail of these two distributions.

Table 6 complements the information given in Figure 13 by showing the medians and interquartile ranges of the boxes in the boxplots. Overall, we see that the LIL and the Alves et al.'s estimators were more biased than the pure MLE and MLEPDS estimators, which is unfavorable. This is not to say, however, that we are aiming to obtain an unbiased estimator, since as discussed in Section 4.1 one of our objectives is to minimize the distance among the inferred distribution

		Pure MLE	**MLEPDS .25**	**MLEPDS .5**	**MLEPDS .75**	**LIL**	**Pickands**	**Alves**	**Girard**	**Hosking**
Gamma	**median**	0.013	0.033	0.000	-0.016	0.136	0.005	0.058	2.454	0.365
	IQR	1.01	3.50	0.99	7.61	12.32	7.76	11.86	14.68	12.90
Weibull	**median**	0.008	0.028	0.000	-0.021	0.114	0.038	0.108	0.465	0.218
	IQR	0.15	0.53	0.30	3.17	1.03	0.94	1.07	3.60	1.31

Table 6 – Interquartile range (IQR) and mean for the boxplots shown in Figure 13.

and the actual one. Of course, if we are able to minimize such distance *and* obtain an unbiased low-variance estimate, that would be the most favorable scenario.

We now compare the proposed MLEPDS method against the maximum likelihood one. We begin by showing in Table 7 the proportion of experimental trials in which the MLEPDS method performed better than the MLE method, considering the Hellinger and Kantorovich distances. These results are, of course, closely related to those shown in Tables 4 and 5, but here they show more directly the differences among the two methods.

We see that, for the gamma case, the MLEPDS method with hyperparameter $q = 0.5$ performed better in about 51% of the experimental trials, according to the Hellinger distance. This proportion reduces to 44% when considering the Kantorovich distance, so we see that the MLEPDS method competes very closely with the pure MLE method, but is still worse. These values change to about 49% and 45%, respectively, for the Weibull case, which is slightly better. On the other hand, considering that the parameter space of the MLE method is one dimension larger, we believe that these are still very good results, and the fact that there is a lot of room for improvement is encouraging.

		$q = 0.25$	$q = 0.5$	$q = 0.75$
Gamma	**Hellinger**	47.01	51.19	42.45
	Kantorovich	39.56	44.21	40.97
Weibull	**Hellinger**	44.33	48.93	40.15
	Kantorovich	41.30	45.47	40.24

Table 7 – Percentage of experimental trials where the MLEPDS method achieved a better model than the maximum likelihood method, considering the Hellinger and Kantorovich metrics.

A more complete view of the distances obtained by each method is shown in Figures 14 and 15, where the histograms of the distances are shown. In general, the MLEPDS method manages to achieve distances very similar to those of the pure MLE method. Besides that, the fact that the histograms have very similar overall shapes also indicates that the performance of the two methods are not very different.

For the gamma case, using Hellinger distance (Figure 14, top part), we observe a small lump in the $[0.5, 0.75)$, which represent cases where the MLEPDS method achieves significantly worse results than the pure MLE. It appears that this lump is smaller for the $q = 0.75$ case, which

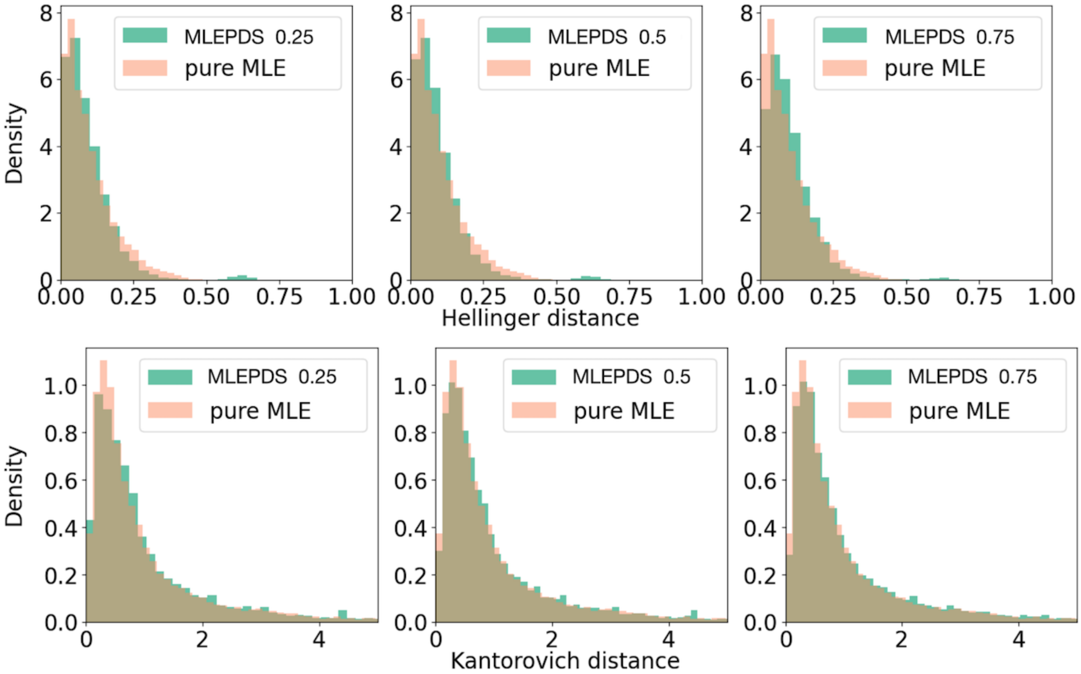


Figure 14 – Histogram of Hellinger (top) and Kantorovich (bottom) distances obtained in experiments concerning the gamma distribution.

Source – Prepared by the author.

is counter-intuitive since the previous discussion indicated that this case was outperformed by the other configurations of the MLEPDS method.

One conclusion we can take from this is that if we can pinpoint the reason for this lump and manage to modify the method to solve it, the results obtained could be improved by large amounts. The isolated slope is not present in the Kantorovich distance histogram, but this is likely because these “outliers” are more spread out. For the Weibull case, we observe the same pattern where the histograms for the MLEPDS method display the same overall shape, but no isolated slope was observed here.

We now proceed to comparing the methods that estimate the population minimum prior to performing MLE. We saw in Tables 4 and 5 that the proposed LIL estimator showed a somewhat variable result, achieving extremely good performance in the general scenario, but degrading considerably when $N = 500$, becoming on par with Alves et al.’s estimator. Also recall that in Figure 13 there appears to be little to no difference between the proposed method and Alves et al.’s estimator.

In Figures 16 and 17 we compare the three most promising estimator-type methods (based on the results shown previously), namely our proposed estimator as well as Alves et al.’s and Girard et al’s estimators, and we take the pure MLE method as basis of comparison. The histogram of the Hellinger and Kantorovich distances obtained throughout the experiments are shown. This way, it is possible to have a better view of how much the three estimators are worse than the pure MLE method.

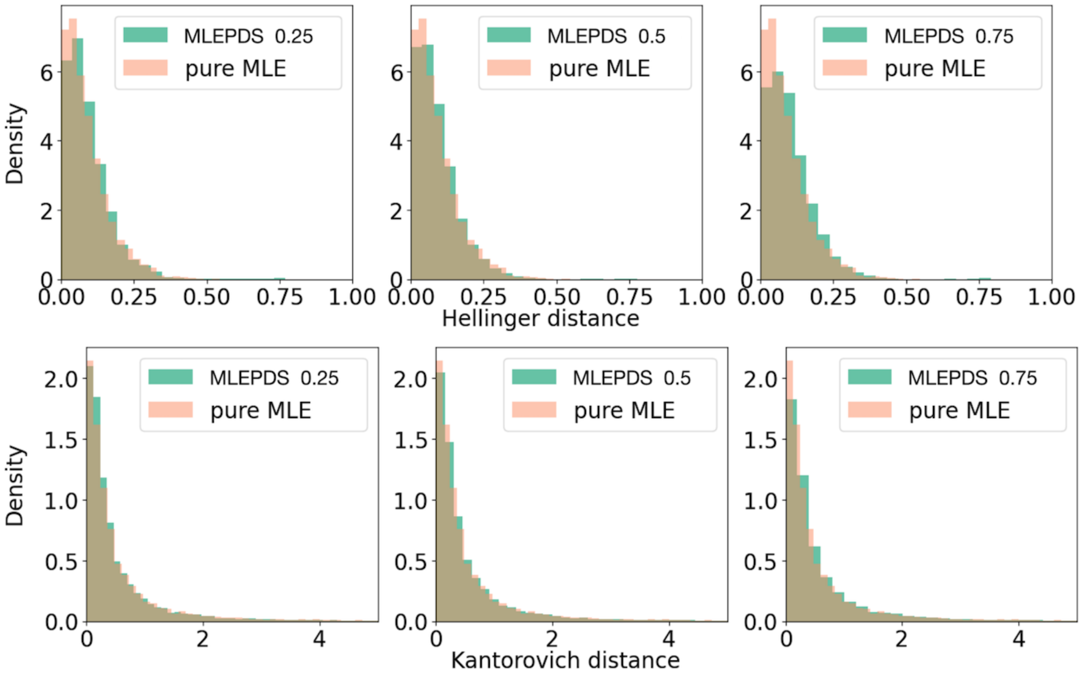


Figure 15 – Histogram of Hellinger (top) and Kantorovich (bottom) distances obtained in experiments concerning the Weibull distribution.

Source – Prepared by the author.

In Figure 16, notice that all three estimators show a remarkably different pattern of Hellinger distances than the pure MLE estimator, with histograms displaying a clearer left tail. Girard et al.'s estimator displays a surprisingly better shape, but it also has a much thicker right-tail, which actually makes it perform worse, on average, than our estimator and Alves et al.'s estimator. Between the latter two methods, it appears our method achieves smaller Hellinger distances overall, as can be seen in Figure 16; in addition to that, there is an abrupt increase in the amount of Hellinger distances in the interval $[0.275, 0.325]$ for Alves et al.'s estimator, which is clearly an undesirable feature in this kind of histogram.

The differences in terms of Kantorovich distances are less visually detectable, but it is also clear here that Girard et al's method results in larger distances, since its histogram has a heavier right-tail, as can be seen by the large margin that interposes the two plotted histograms on the portion that corresponds to the right-tail.

Very similar conclusions can be drawn from Figure 17, which considers the Weibull simulations. Interestingly, we observe here the same abrupt increase in the number of Hellinger distances for the Alves et al.'s estimator, and overall the histogram for LIL looks better in this case too.

The only exception is that Girard et al.'s estimator achieved a histogram shape even more similar to the pure MLE method, and its histogram might actually be marginally better than the other two estimators' histograms. Recall, however, the multiple drawbacks of the Girard et al.'s method mentioned previously, such as the large number of errors and also the unfairness of using

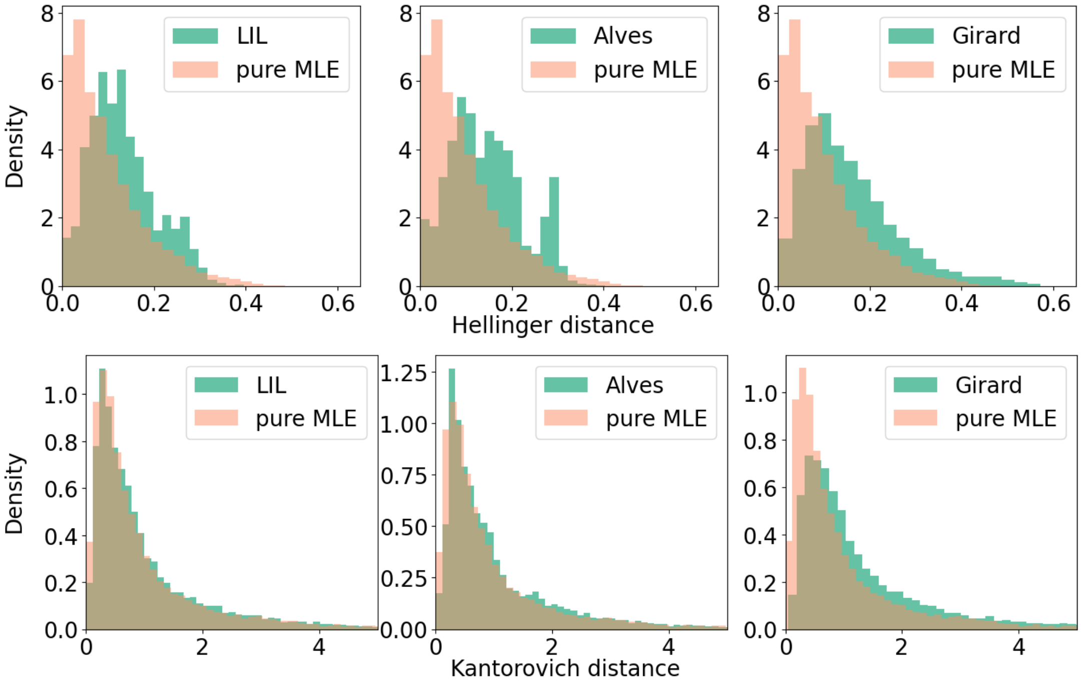


Figure 16 – Histogram of Hellinger (top) and Kantorovich (bottom) distances obtained in experiments concerning the gamma distribution.

Source – Prepared by the author.

four different configurations and taking only the best result.

As for the Kantorovich distances, Girard et al.'s estimator display a heavy right-tail in a similar fashion as in the gamma case. In any case, it seems clear that Girard et al.'s estimator did not show any significant advantage when compared to our method and Alves et al.'s estimator, so it seems reasonable to exclude it from analysis hereafter due to its high rate of failure, as shown in Tables 4 and 5.

In order to further compare the proposed LIL estimator with of Alves et al.'s estimator, since they displayed similar performances above, we show in Figure 18 a comparison for each pair of distribution parameters used in the experiments. We see that, in both cases, shape parameters $\alpha \leq 1.5$ tend lead to a larger advantage from using our method. These values of shape correspond to gamma and Weibull distributions with either a one-tailed shape or at least a very positively skewed shape.

These figures show the overall results for all values of sample size N. If we consider just $N = 500$, we continue to observe a significant superiority of our method for shapes of 1.5 and smaller, but for shapes greater than or equal to 4, the Hellinger distances yielded by Alves et al.'s estimator are slightly better, though this advantage is four to ten times smaller than the overall advantage given by our estimator on the other half of the parameter space. The above results appear to tip the scales in favor of our estimator.

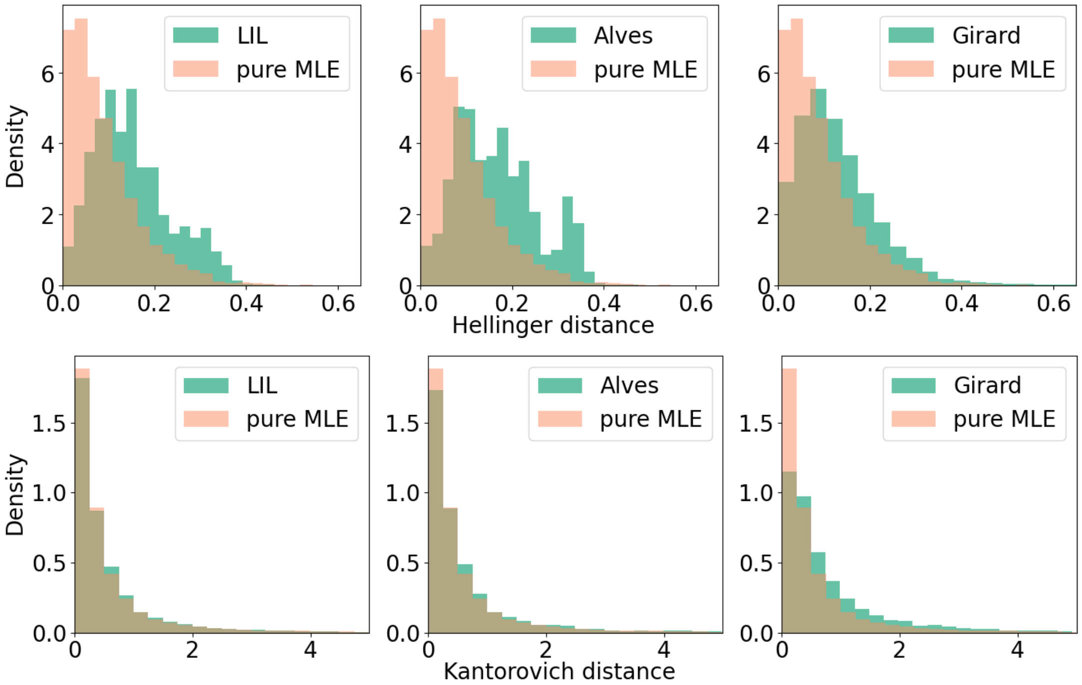


Figure 17 – Histogram of Hellinger (top) and Kantorovich (bottom) distances obtained in experiments concerning the Weibull distribution.

Source – Prepared by the author.

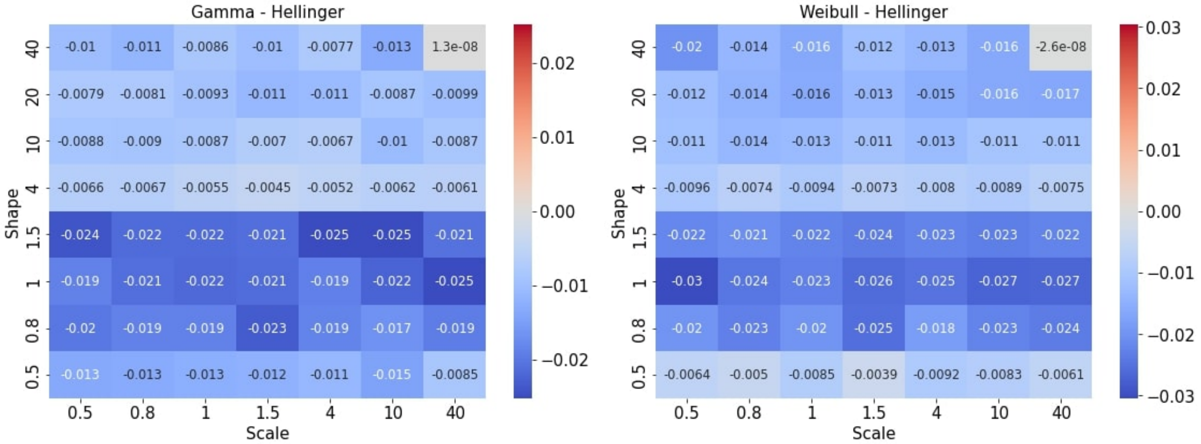


Figure 18 – Average difference between the Hellinger distance obtained by the proposed LIL method and Alves et al.’ method, in the form $d_{\text{proposed}} - d_{\text{alves}}$, for each pair of distribution parameters used in the experiments. Negative values correspond to an advantage of our method.

Source – Prepared by the author.

4.5 A Way to Reduce Instability of the Method

One major problem with the proposed MLEPDS method is that the optimization surface of the likelihood function is increased when compared to fixing the population minimum at a fixed value. Recall that the family of densities used has the form:

$$f\left(x + F_{\min}^{-1}(q \mid \boldsymbol{\theta}) \,\middle|\, \boldsymbol{\theta}\right),$$

so the derivatives of the log-likelihood would become:

$$\frac{\partial}{\partial\theta_i}\left(\sum_{j=1}^{n}\log f\left(x_j+F_{\min}^{-1}(q\,|\,\boldsymbol{\theta})\,\middle|\,\boldsymbol{\theta}\right)\right)=$$

$$=\sum_{j=1}^{n}\frac{\dfrac{\partial f\left(x_j+x_{q,\boldsymbol{\theta}}\,\middle|\,\boldsymbol{\theta}\right)}{\partial\theta_i}\left[\dfrac{\partial}{\partial\theta_i}F_{\min}(x_{q,\boldsymbol{\theta}}\,|\,\boldsymbol{\theta})\right]^{-1}}{f\left(x_j+x_{q,\boldsymbol{\theta}}\,\middle|\,\boldsymbol{\theta}\right)}$$

$$=\frac{1}{\left[n\,(1-q)^{n-1}\,f(x_{q,\boldsymbol{\theta}}\,|\,\boldsymbol{\theta})\right]}\sum_{j=1}^{n}\frac{\dfrac{\partial f\left(x_j+x_{q,\boldsymbol{\theta}}\,\middle|\,\boldsymbol{\theta}\right)}{\partial\theta_i}}{f\left(x_j+x_{q,\boldsymbol{\theta}}\,\middle|\,\boldsymbol{\theta}\right)}$$

where we substitute $x_{q,\boldsymbol{\theta}}=F_{\min}^{-1}(q\,|\,\boldsymbol{\theta})$ to clean up the notation. Compared to the derivatives for the usual log-likelihood, we have here an extra term $[n\,(1-q)^{n-1}\,f(x_{q,\boldsymbol{\theta}}\,|\,\boldsymbol{\theta})]$, which is always positive, so it should not be particularly a source of optimization difficulties here. The terms inside the sum are roughly the same, except for the $x_{q,\boldsymbol{\theta}}$ term, which is what adds to the complexity of the optimization surface.

One way we have found to control the instability of the method is as follows. Beginning with an initial parameter $\boldsymbol{\theta}^{(0)}$, we fix the value of $x_{q,\boldsymbol{\theta}^{(0)}}$, and maximize the likelihood iteratively by doing:

$$\boldsymbol{\theta}^{(n+1)}=\underset{\tilde{\boldsymbol{\theta}}}{\arg\max}\sum_{j=1}^{n}\log f\left(x_j+x_{q,\boldsymbol{\theta}^{(n)}}\,\middle|\,\tilde{\boldsymbol{\theta}}\right),$$

which gives a sequence of estimates for $\boldsymbol{\theta}$, and we stop once a stopping condition is met; for now, we stop it when the relative error between two consecutive estimates is lower than 0.001.

We have repeated the experiments using the MLEPDS method with the above modification, using similar experimental settings described in Section 4.4. Table 8 shows the proportion of times that the modified MLEPDS method performed better than the pure MLE method, and it can be seen here that a significant improvement was achieved when compared to Table 7, where the proportions were most often around 40% and rarely above 50%. Here, the MLEPDS method achieves a better performance in more than half of the experimental trials (in most cases), and can reach up to 62.08% when considering the Hellinger distance and the gamma distribution.

Figures 19 and 20 show the signed biases (i.e., $\hat{m}-m$) of the modified MLEPDS method plotted against those of the non-modified one. To obtain these figures, we take the set of biases obtained by the non-modified MLEPDS method in the experiments described in Section 4.4, and take the set of biases obtained by the modified version in the experiments that were newly performed. Then, we sort these sets and then plot the ordered vectors against each other; this is

		$q = 0.25$	$q = 0.5$	$q = 0.75$
Gamma	**Hellinger**	61.00	62.08	44.93
	Kantorovich	50.69	52.96	46.56
Weibull	**Hellinger**	54.87	54.06	42.04
	Kantorovich	46.67	46.94	46.56

Table 8 – Percentage of experimental trials where the modified MLEPDS method (with parameters $q \in \{0.25, 0.5, 0.75\}$) achieved a better model than the maximum likelihood method, considering the Hellinger and Kantorovich metrics.

why some of the noise is damped and the figures tend to display a smoother manifold instead. In these figures, if the points are located more between the black line ($x = y$ line) and the y-axis, then this means that the modified method tends to achieve smaller biases. On the other hand, if the points are between the black line and the x-axis, then it would be showing a superiority of the non-modified version. As such, it is very visible in both figures that, overall, the outliers are much less aberrant in the modified method.

In Figure 19, note that the modified method with $q = 0.25$ tends to underestimate the population minimum to a smaller extent, followed by the case $q = 0.5$ and then by $q = 0.75$ which tends to underestimate more intensely. These results make sense considering the interpretation of the parameter q. On the other hand, if we look at the first quadrant, we see that the case $q = 0.75$ tends to avoid overestimating the population minimum, which causes it to have a better performance in the first quadrant, followed by $q = 0.5$ which appears to be located exactly upon the black line, showing that the modified version performs equally well in the first quadrant as the non-modified version.

Figure 20, which concerns the Weibull case, display similar features. Outliers are less aberrant for the modified method, as seen in the third quadrant; here, the ordering of the methods according to the parameter q is not very clear, but we can still observe that the modified method with $q = 0.25$ appears to have the best behaving outliers. In the zoomed figure, we can see again the same ordering according to the parameter q, both in the first and in the third quadrants. That is, the method with $q = 0.25$ tends to overestimate the population minimum with more intensity, and the case $q = 0.75$ tends to underestimate it more intensely. Overall, in both figures, there are very few points in the regions that would represent a superiority of the old method.

4.6 Simulation Experiment With the Generalized Gamma Distribution

As pointed out multiple times in the previous sections, the increase in complexity of the optimization surface when using the MLEPDS method has been one of our greatest concerns. In order to try to verify that the above results carry on to more complicated distributions, here we

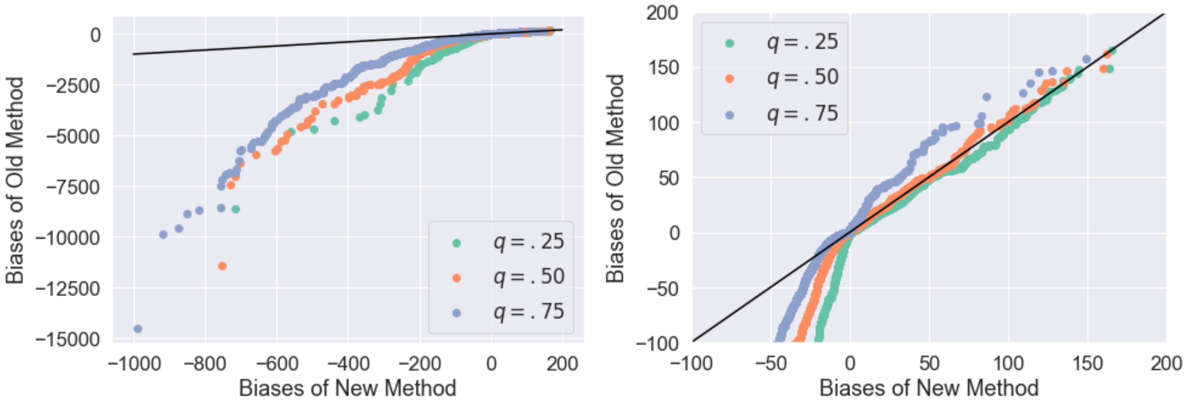


Figure 19 – Signed biases of the modified MLEPDS method in the gamma case plotted against the signed biases of the non-modified version. The figure to the right zooms in a smaller region of the cartesian plane, which is not very visible in the figure to the left.

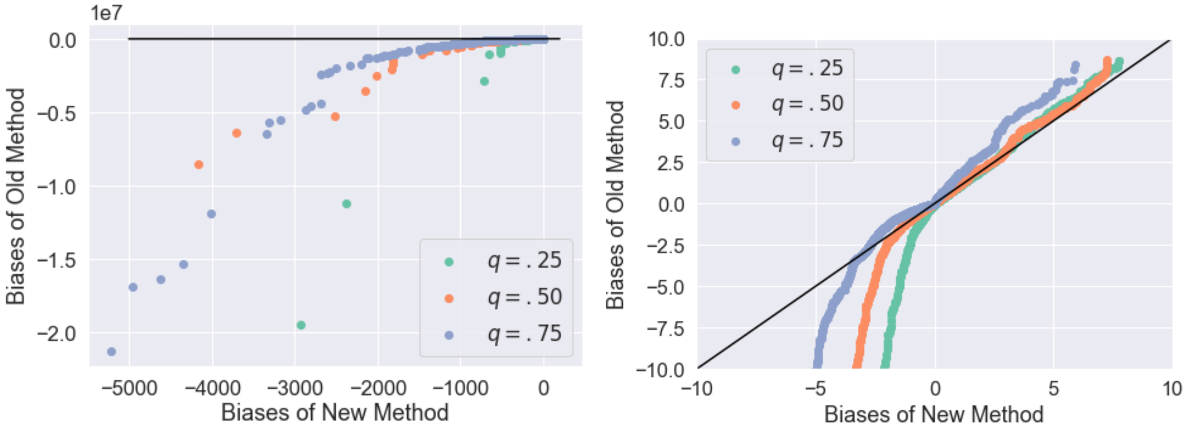


Figure 20 – Signed biases of the modified MLEPDS method in the Weibull case plotted against the signed biases of the non-modified version. The figure to the right zooms in a smaller region of the cartesian plane, which is not very visible in the figure to the left.

perform a simulation experiment with a three-parameter distribution, namely the generalized gamma (STACY, 1962). In (STACY, 1962) the density function is defined with parameters a, d and p, but here we use the common reparametrization $a = a$, $b = p$ and $k = d/p$. With this, the probability density and the cumulative density functions are defined as:

$$f(x \mid a, b, k) = \frac{b x^{bk-1} \exp\left[-(x/a)^b\right]}{a^{bk}\Gamma(k)} \quad \text{and}$$

$$F(x \mid a, b, k) = \frac{\gamma\left(k, (x/a)^b\right)}{\Gamma(k)},$$

where $\gamma(\cdot, \cdot)$ is the lower incomplete gamma function.

In this parametrization, when $k = 1$, the distribution corresponds to a Weibull distribution; if $b = k = 1$, it is the exponential distribution; and if $b = 1$, it is the gamma distribution. Thus, seeing the values $k = 1$ and $b = 1$ loosely as some sort of critical points (and including $a = 1$ as

a	b	k
0.5	0.8	0.8
0.5	0.8	2
0.5	2	0.8
0.5	2	2
1.2	0.8	2
1.2	1.5	0.8
2	0.8	0.8
2	2	2

Table 9 – Parameters selected for the experiment with the generalized gamma distribution.

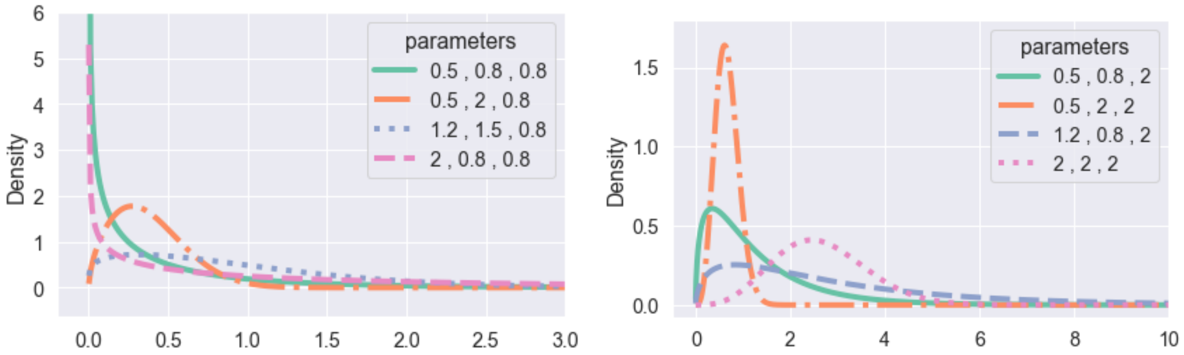


Figure 21 – Shapes of the generalized gamma distribution configurations selected for the experiment.

well), we selected a set of parameters that cover values in each of the eight quadrants implicitly defined by the $(1, 1, 1)$ critical point in $\mathbb{R}^3$. The parameters selected and the shapes of the corresponding distributions are shown in Figure 21.

We believe that enough comparisons have already been made between our proposed method and the one by Alves and Neves (2014), so now we focus on comparing only with the pure MLE method. This time, for each combination of parameters, 50 trials are performed for each sample size in the set $\{50, 100, 500\}$. We found that performing the numerical integrations, necessary to calculate the Hellinger and Kantorovich distances, was impossible for some choices of the parameters, because numerical errors were causing the integral to diverge or be undefined frequently; it is possible that the heavier tails make numerical integration more difficult. Therefore, we instead show the squared error of the parameters (excluding the population minimum in the pure MLE method) and the squared error of the population minimum. The separation is done for two reasons. First because the population minimum is not a parameter of the proposed method, although it does return an estimate for it. Second, the overall error on the population minimum is often larger, which could shadow differences in the shape and scale parameters.

Figures 22 and 24 show, for each parameter setting, how many times the stable proposed method with $q = 0.5$ and $q = 0.25$, respectively, achieves lower parameter squared error than the pure MLE. Figures 23 and 25, on the other hand, show the same information, but considering the squared error of the population minimum instead. In both figures, we disregard any trial where

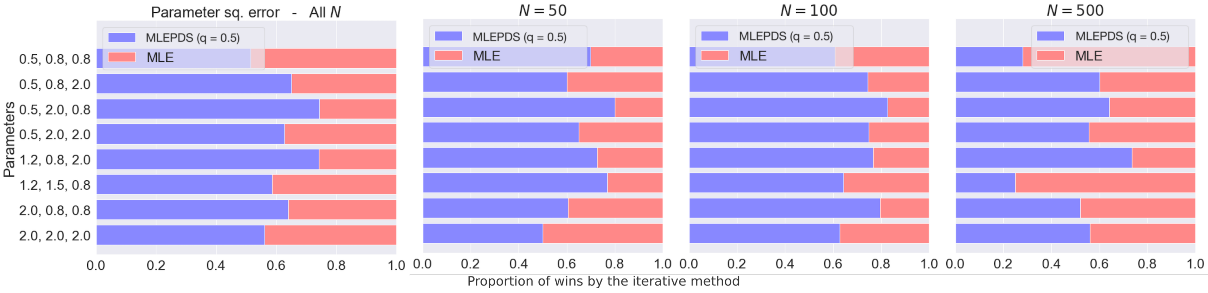


Figure 22 – Proportion of times that the MLEPDS method, with $q = 0.5$, performed better than the pure MLE method, in terms of the parameter squared error.

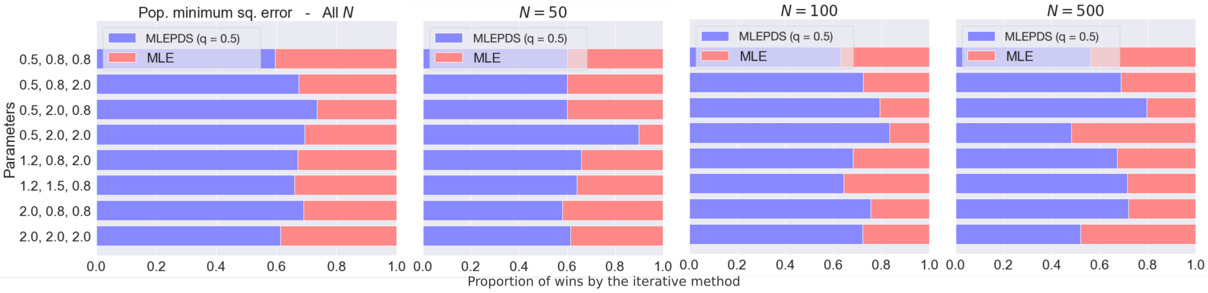


Figure 23 – Proportion of times that the MLEPDS method, with $q = 0.5$, performed better than the pure MLE method, in terms of the population minimum squared error.

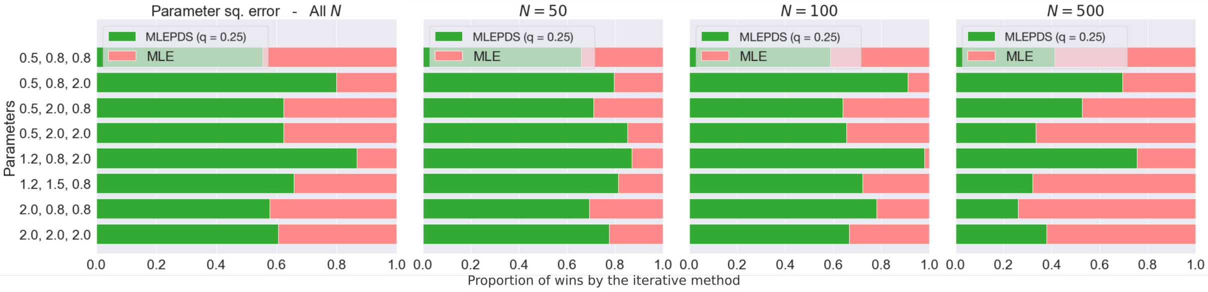


Figure 24 – Proportion of times that the MLEPDS method, with $q = 0.25$, performed better than the pure MLE method, in terms of the parameter squared error.

either method suffered of convergence problems in the underlying optimization algorithm; thus, we remark that the MLEPDS method with $q = 0.5$ had 14.58% trials that resulted in convergence problems, followed by 7.25% when $q = 0.25$, and 6.67% when using the pure MLE method. Overall results when the convergence errors are not ignored are introduced later (Table 10).

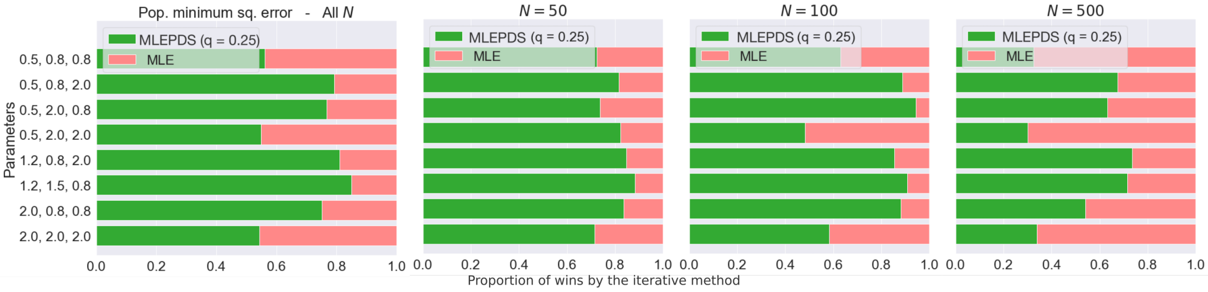


Figure 25 – Proportion of times that the MLEPDS method, with $q = 0.25$, performed better than the pure MLE method, in terms of the population minimum squared error.

	MLEPDS Win Rate	
	parameter SE	pop. minimum SE
$q = 0.25$	64.25%	67.58%
$q = 0.5$	58.83%	59.25%
$q = 0.75$	40.58%	34.08%

Table 10 – Overall results without disregarding convergence errors. The proposed methods, with different values of the hyperparameter q, are here compared with the pure MLE method in terms of the respective squared errors (SE).

In Figure 22, we notice that in the overall scenario, the proposed method achieves better parameter estimates than the pure MLE method in all distribution configurations, except maybe on the configuration 0.5-0.8-0.8, where it is roughly equal. However, note a slight tendency of our method to perform better when $N = 50$ and $N = 100$, and perform worse when $N = 500$. At this sample size, however, the difference in the performances becomes very small; this just shows the already well known asymptotic power of the pure MLE method. We observe similar conclusions from Figure 23, which considers the squared error on the estimate for the population minimum. We also see here that the MLEPDS method is better overall in all parameter configurations, and this time the degradation of performance on $N = 500$ cannot be observed as visibly as in Figure 22. This might be pointing out that the asymptotic power of the pure MLE method is more limited when it comes to estimating the left endpoint.

In Figure 24 we see that the proposed method with $q = 0.25$ has an outstanding performance,[5] better than when $q = 0.5$, for smaller samples. On the other hand, for $N = 500$ it is again outperformed by the pure MLE method, although now the proposed method suffers more on other parameter configurations (especially 2.0-0.8-0.8). Similar behavior is observed in Figure 25; here, when $N = 500$, it appears that the proposed method with $q = 0.25$ performs worse than with $q = 0.5$. It could be the case that the hyperparameter q should be different depending on the sample size, but we prefer to consider this as an effect of choosing the generalized gamma as the distribution used in the simulations, and expect the relative performance among the different values of q to differ for other distributions.

If we do not disregard the cases where convergence errors occurred, we get the overall results shown in Table 10. This means that if one method failed to converge but the other succeeded, then the latter is considered as "winning." By doing this, we notice that the proposed method with $q = 0.25$ estimates better parameters than the pure MLE method in about 64% of the trials, and gives better left endpoint estimates in about 68% of the times, which represents a significant improvement over the well-established method. When $q = 0.5$, the proposed method is better in about 59% of the trials in both scenarios, so it stays behind the method when $q = 0.25$, but we emphasize that this could be a characteristic of the distribution chosen for the simulation study. Overall, the experiments demonstrate the superiority of the proposed approach, and show that the iterative approach to the MLEPDS method served the intended purpose of increasing its stability, leading to better results and enabling it to be used in more complex distributions as well.

---

[5] We did not collect the results for $q = 0.75$ as it did not display good results in Figure 13.

CHAPTER 5

APPLICATION ON EARTHQUAKES

Frequent earthquakes of larger magnitudes cause numerous deaths and huge monetary losses to the affected countries. This is particularly concerning for developing nations, where comparatively poor construction standards result in numerous fatalities even from lower magnitude earthquakes (~ M6.5)[1]. However, no amount of preparation can safeguard developed countries from the damages caused by earthquakes of magnitude M8.5 or higher (STEIN; WYSESSION, 2003).

There are various reasons why we are still unable to perform earthquake forecasting well enough to prevent hazards. To begin with, thus far we only have a (relatively) superficial understanding of the mechanisms underlying the build up of strain energy on the faults. Even the basic processes involved in mantle convection have not yet been fully understood. What we have are only a couple of convincing (but not yet proven) geoscientific theories for it, such as the “plume tectonics” theory (YUEN *et al.*, 2007; LARSON, 1991).

On top of that, there is also the possibility that external factors (unrelated to mantle convection) might also play a significant role in seismic activity. To name one example, the literature often mentions the possibility of an influence by the tidal force exerted by the Moon on the Earth (IDE; YABE; TANAKA, 2016; HAGEN; AZEVEDO, 2017), which might be responsible for initiating the rupture process that releases seismic energy and can give rise to an earthquake.

Thus, earthquakes are a very complex phenomenon, with multiple factors involved, which makes it necessary to adopt a stochastic approach to analysing and potentially forecasting earthquakes. In this dissertation, we attempt to apply the main proposed method on earthquake data, aiming to contribute to the particular area of mid- to long-term earthquake forecasting. In particular, we study the value of the maximum possible earthquake that can occur within a time window of at least a month, and investigate whether the application of the method on subsequent time windows can give an insight about the trend of seismicity within a region, and if it can be used to help with forecasting large earthquakes.

---

[1] M stands for magnitude, with no specification of the particular method to calculate it.

5.1 Fundamental Concepts

Even though in this master's project we make heavy use of readily available earthquake catalogs, with which it would be possible to perform learning and forecasting without having much knowledge specific to seismology, we believe it is important to have at least a basic understanding of earthquakes. Besides helping with reading this text, which will involve some terms specific from the seismology field, it also enables an easier interpretation of the results, an easier identification of aspects that could be improved in our devised methods, and also help understanding why this physical phenomenon must be treated as a phenomenon with a large degree of stochasticity, despite being the consequence of many deterministic systems (such as mantle convection, sea level changes and Sun electromagnetic bursts). Therefore, we provide in this section a brief overview about the earthquake generating process.

The Earth is an irregular sphere with a radius of about 6371 km, a solid core estimated to be as hot as approximately 5700 K (ALFÈ; GILLAN; PRICE, 2007), and a relatively cold solid surface. In the interim, there is liquid material, called the mantle. As is well known, a liquid confined between solid surfaces with largely different temperatures leads to convection flows, and this movement is constantly pushing and pulling the Earth's crust in different directions, giving rise to plate tectonics, a phenomenon that thus far has not been found in any other celestial body other than the Earth (MCCALL, 2010).

The constant and long-lasting pressure on the Earth crust has divided it into roughly 14 large plates (plus 38 smaller ones) (BIRD, 2003) whose movement is governed by the complex dynamics taking place in the mantle layer. These plates form a tessellation of the Earth's surface, and each and every intersection between these plates is rich in seismic activity, be it hazardous or not. We introduce below some terminology from the field of seismology, as it will be needed later in the text.

A fault is a discontinuity in the Earth's surface caused by the relative movement of the surrounding tectonic plates. There are at least four main cases that give rise to a fault:

1. two plates collide against each other (i.e. a convergent boundary), creating a vertical deformation of the land;

2. two plates move away from each other (i.e. a divergent boundary), but at an angle that is not sufficient to create a ridge (to be defined later), resulting in one side of the fault being more elevated than the other, while the land remains connected by a ramp or, more precisely, a fault scarp;

3. a divergent boundary where two plates move away from each other, creating a rift extending throughout a long portion of the plate boundary (i.e. a ridge), making way for magma to well up all the way from the asthenosphere, which cools down and forms new seabed, and the newly formed rocks slowly move towards the other side of the plate, where they will eventually be recycled in the corresponding subduction zone; and

4. two plates move laterally relative to each other, resulting in a visual discontinuity in the soil, clearly showing that the rock masses were shifted horizontally.

Note that oblique faults, which combine characteristics of the faults described above, also exist.

Subduction is the seismological phenomenon where a block of crust descends relative to another block of crust (creating a “trench”), and “subduction zone” refers to the elongated region where such phenomenon occurs. Subduction zones appear in convergent boundaries that are sufficiently active.

The process that gives rise to faults is not a continuous one, due to friction and the rigidity of the rocks involved. There are many kinds of forces that oppose the relative motion of two plates, but for simplicity let us put them all under the name of “friction.” Relative movement occurs, pressing the soil and rocks of the two plates onto each other, thus building up friction, and eventually there will be enough friction to put a stop to all motion. At this point, we say the two plates are locked onto each other.

Although movement at the boundary is put to a halt, the inner parts of the plate continue to move as usual, which increases the stress in the fault. Stress builds up continuously, and when it exceeds a certain threshold, a rupture occurs in the fault, and strain energy is released in the form of sudden movement of the formerly locked plates; this is an earthquake. After the movement ends and the plates lock onto each other again, the process restarts itself. This simplified model for explaining the occurrence of earthquakes is called “Elastic–Rebound theory” (STEIN; WYSESSION, 2003).

There are some quantitative aspects of earthquakes that are paramount to understand the variables used for analysis, in particular the magnitude, which is relate to the amount of energy released by an earthquake. Such energy follows the expression (PANAKKAT; ADELI, 2007; STEIN; WYSESSION, 2003):

$$E = 10^{11.8+1.5M} \text{ ergs}, \tag{5.1}$$

where M is the calculated magnitude, and 11.8 and 1.5 are empirically determined constants that can differ a little among different papers in the literature: 12.24 and 1.44 in (BÅTH, 1966), 11.3 and 1.8 in (GUTENBERG; RICHTER, 1942), a quadratic term is added to the expression above in (GUTENBERG; RICHTER, 1956), and so on.

An important result by Beno Gutenberg and Charles Richter (GUTENBERG; RICHTER, 1944) is the proposal of a simple relation that exists between the frequency and magnitude of earthquakes, which incidentally describes remarkably well the occurrence of earthquakes, no matter the region or the time span chosen to analyze. Such relation, called the Gutenberg–Richter relation, dictates that the amount of earthquakes above a certain threshold M follows:

$$N(M) = 10^{a-bM}, \tag{5.2}$$

where a and b are parameters depending on the location; these are sometimes called the a-value and b-value. Note that, although Equations (5.1) and (5.2) appear to be similar, they represent two very different entities in the context of seismology. We use the above relation to convert a predicted amount of earthquakes into a prediction for the maximum magnitude among such set of earthquakes.

As commented previously, despite being the result of a combination of deterministic phenomena, seismicity is most often treated as a highly stochastic phenomenon. Since the essence of this Chapter is the statistical approach to analyzing earthquakes, it seems appropriate to introduce the main approach used in the literature to design a statistical model for earthquakes, namely a point process model.

Earthquakes are often called an “event series,” which is a sequence of samples $(t_i, \boldsymbol{x}_i)$ observed at non-equidistant points in time $t_1, t_2, \ldots, t_n$. This kind of data is often modeled using point processes (ROSS, 2014), and it is not different in the field of seismology. In the following, we briefly introduce the point process model often used in seismology.

Let $(T_i)_{i\in\mathbb{N}}$ be a sequence of non-negative random variables representing the time of occurrence of the i-th event (i.e. earthquake), where the following order restriction holds:

$$T_1 < T_2 < T_3 < \cdots < T_n < \ldots.$$

Formally, the model is completely specified by a joint probability distribution (jpd) $P(T_1, T_2, \ldots, T_n, \ldots)$. The jpd of an infinite sequence of random variables is not very easy to work with, and also not very intuitive, but it can be decomposed as:

$$P(T_1, T_2, \ldots, T_n, \ldots) = \prod_{i\in\mathbb{N}} P(T_{i+1} \mid T_i, T_{i-1}, \ldots, T_1) = \prod_{i\in\mathbb{N}} P(T_{i+1} \mid \mathcal{H}_i),$$

which indicates that it suffices to determine the probability of each event occurrence T_i conditioned on the history of previous occurrences (i.e. $\mathcal{H}_i = (T_i, T_{i-1}, \ldots, T_1)$), and this is a useful way to define a temporal point process. In practice, since each random variable can depend on a very large number of other variables, this specification is often unfeasible to work with, and sometimes may not even reflect reality very well. For example, for some phenomena, it might be more reasonable that each event occurrence depend on only a few previous ones. On the limit of this line of thought are the temporal point processes where each event depends only on the previous one, so that the inter-event times $T_i - T_{i-1}$ are independent and the above jpd decomposes as:

$$P(T_1, T_2, \ldots, T_n, \ldots) = \prod_{i \in \mathbb{N}} P(T_{i+1} \mid T_i).$$

In general, earthquakes are modeled with point processes where there are complex dependency relationships between an event T_i and the events that preceded it (i.e., $\mathcal{H}_{i-1}$), and the above tools are not sufficient to properly describe such generating processes. In such cases, it is more common to use a conditional intensity function, which is defined as follows. Let $f(t \mid \mathcal{H}_j)$ be the probability density function for the event $T_{j+1} \mid \mathcal{H}_j$, and $F(t \mid \mathcal{H}_j)$ the corresponding cumulative density function. Then, the conditional intensity function is:

$$\lambda^*(t) = \frac{f(t \mid \mathcal{H}_j)}{1 - F(t \mid \mathcal{H}_j)}, \tag{5.3}$$

where the star superscript is commonly used as a reminder of the fact that it depends on the history $\mathcal{H}_t$ up until time t (DALEY; VERE-JONES *et al.*, 2003). The interpretation is that it gives the instantaneous rate of occurrence of events at time t given the history up until that point.

One immediate generalization of the above is to make $\mathcal{S}$, the space where the points in the point process belong, include not only time, but also space (i.e., a spatio-temporal point process), in which case we have many dimensions to deal with. In this case, the conditional intensity function is $\lambda(t, x, y)$, representing the intensity of point occurrence in time t and position (x, y) (DALEY; VERE-JONES *et al.*, 2003), which clearly suits the case of earthquakes very well. In fact, a very popular model for earthquakes is the Epidemic Type Aftershock Sequence (ETAS) model described in (OGATA, 1988), where a suitable conditional intensity function is defined using existing principles from seismology, as well as some heuristic reasoning. The conditional intensity function assumes here the general form:

$$\lambda(t) = \mu + \sum_{t_i < t} g(t - e_i^{\mathrm{ts}}) e^{\beta(e_i^{\mathrm{mag}} - \mu)}, \tag{5.4}$$

where μ is a “natural” base rate for earthquake occurrence, which is the first assumption made in this model. That is, the underlying earthquake generating process is assumed to generate earthquakes stochastically with a base rate of μ. The inner sum in the above equation describes how past earthquakes increase the earthquake occurrence rate at each time t, so the sum goes over all earthquakes with time t_i lower than t.

The exponential term is a means to include magnitude information into the rate, and its main purpose here is to account for the fact that earthquakes with lower magnitude have lower influence in causing earthquakes, thus the exponential multiplicative term e_i^{mag} decreases with the magnitude, and we note that τ is a reference magnitude that needs to be adjusted for the particular dataset being modeled. Finally, $g(x)$ describes how the occurrence of an earthquake makes it more likely that more earthquakes will occur. In (OGATA, 1988), two forms of g are studied:

$$g_1(t) = ae^{-\alpha t} \tag{5.5}$$

$$g_2(t) = \frac{K}{(t+c)^p}, \tag{5.6}$$

where a, α, c, p are all parameters to be adjusted. Note that both functions are monotonically decreasing, implying that the influence of earthquakes vanish over time.

As the name implies, this is a model aimed at modeling mainshocks (major earthquakes) and aftershocks (which occur as a consequence of the mainshock). This is why the functions end up describing a process where each event has a probability of triggering further events, which thus might trigger grandchild events, and so forth, and the process at some point settles due to the vanishing of the second term in the conditional intensity function.

The mainshocks are then attributed to a completely stochastic component described by an underlying probability distribution with rate μ, which ends up being the weakness of the model, since these events and their magnitude are not entirely random, and in fact would depend on how much stress has built up over time on the respective fault. Despite this limitation, it does a very decent job at modeling earthquakes when compared to other approaches used in the literature. This should be enough to give the reader a good idea of the stochasticity contained in the earthquake generating process, and how complex it can be.

5.2 Data Sources and Configurations

Before discussing the results obtained by using the above prediction framework, we first clarify the sources of the catalogs that were analyzed. For worldwide earthquakes, the data were obtained from the United States Geological Survey,[2] the New Zealand catalog of earthquakes was obtained from the GeoNet project,[3], for Japan it was obtained from the Japanese Meteorological Agency,[4] and for Greece and surroundings it was downloaded from the website of the Department of Geophysics–Geothermics of the University of Athens.[5] Figure 26 illustrates these earthquake catalogs, with the intention to give the reader a more concrete view of how earthquakes are geographically located in each of the regions considered, and how their locations are determined (to a certain extent) by the organization of the surrounding tectonic plates.

---

[2] See <https://earthquake.usgs.gov/earthquakes/search/>.

[3] See <https://www.geonet.org.nz/>.

[4] See <https://www.data.jma.go.jp>.

[5] See <http://www.geophysics.geol.uoa.gr/>.

The specific regions analyzed are quite rich in seismic activity, with emphasis in Japan, as the most destructive earthquakes occur there. As can be seen in Figure 26, the seismological activity surrounding Japan is heavily influenced by the convergence of several tectonic plates, including the Pacific Plate, the North American Plate, the Eurasian Plate, and the Philippine Sea Plate (BIRD, 2003). The Pacific Plate, which lies to the east of Japan, is subducting beneath the North American Plate, and this subduction generates large megathrust earthquakes, such as the devastating 2011 Tohoku earthquake and tsunami. To the west, the Philippine Sea Plate is also subducting beneath the Eurasian Plate, contributing to seismic activity in Okinawa and surrounding islands. Finally, the Philippine Sea Plate is also moving somewhat parallel (but not perfectly parallel) to the Eurasian Plate along the Nankai Trough, where it produces frequent earthquakes in densely populated areas such as Tokyo and Kyoto. The complex interactions of these tectonic plates result in a high frequency of earthquakes, tsunamis, and volcanic activity throughout Japan, emphasizing the need for robust seismic monitoring and preparedness measures in the region.

As for New Zealand, we note that the convergence of the Pacific Plate and the Australian Plate is the main source of seismic activity in the region (BIRD, 2003), causing intense seismic activity and potential for large earthquakes and tsunamis, particularly in the North Island. Along the South Island's western edge, the Alpine Fault accommodates the lateral movement between the Pacific and Australian Plates, generating significant strike-slip earthquakes (STEIN, 1987).

Finally, the Aegean plate (around the Balkan peninsula) is nestled amid the Eurasian, African, and Anatolian plates (BIRD, 2003). To the north, the Eurasian Plate presses against the Aegean Plate. the African Plate to the south converges with the Aegean Plate primarily through a combination of subduction (i.e. one plate goes under the other) along the Hellenic Arc and collision (i.e. one plate collides directly with the other) along the Cyprus Arcl. To the east, the Anatolian Plate exhibits a lateral motion relative to the Aegean Plate along the North Anatolian Fault, which originates a smaller number of earthquakes.

The period covered by the datasets was chosen to be from January 1, 2000 to August 31, 2021. The starting date cannot be too far in the past, because due to changes in sensor technology, the characteristics of the data change drastically if data that is decades old is used. The end date was the most recent date that is covered by all the catalogs. For the New Zealand catalog, we imposed a minimum magnitude threshold of 2.5, whereas for the Japanese and worldwide data, we considered all earthquakes provided by the data provider, where we remark that the minimum magnitude was 2 for Japanese earthquakes, 1.5 for Greece and 3.38 for worldwide ones.

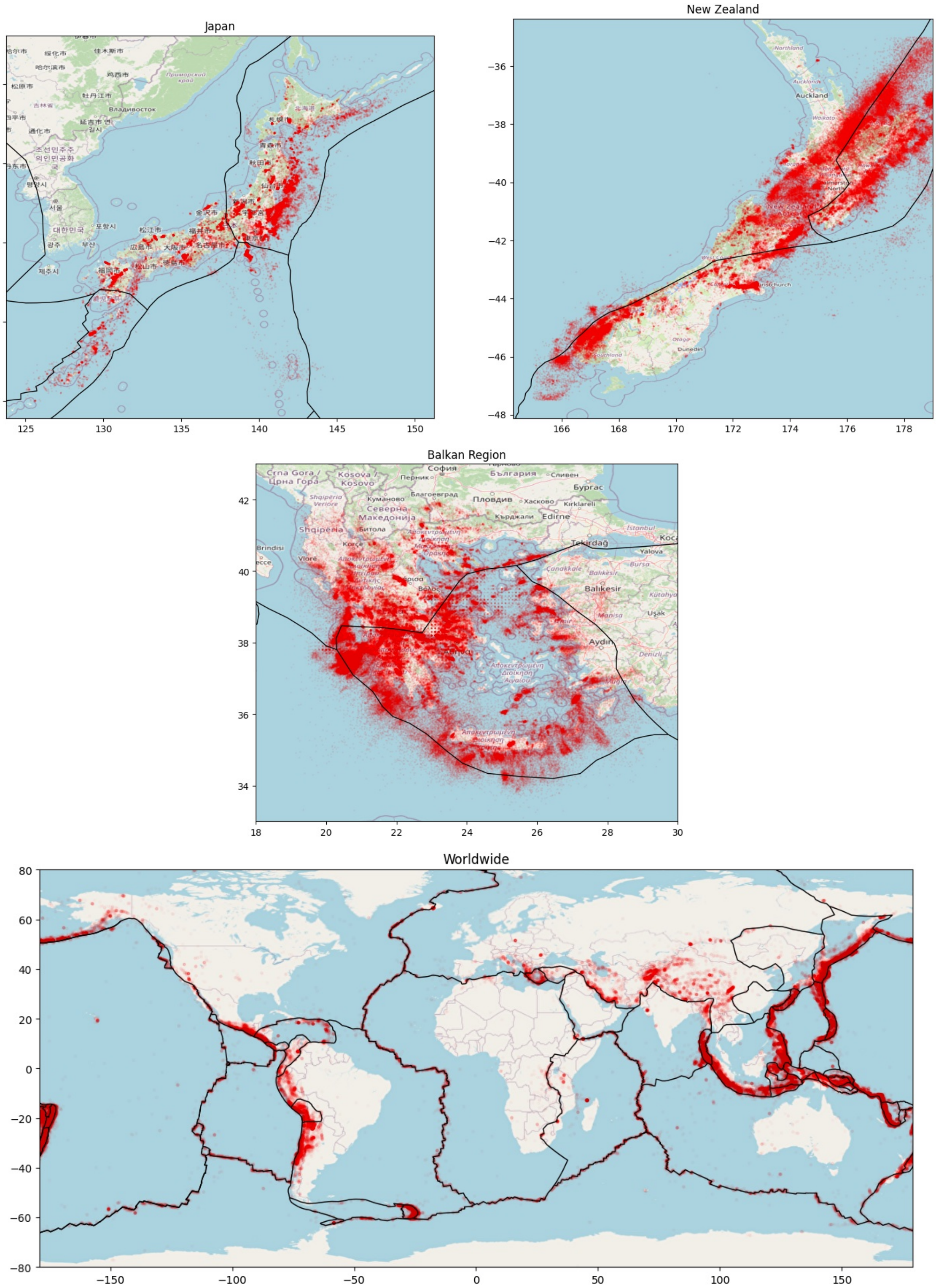

Japan
45
40
35
30
25
125
130
135
140
145
150
New Zealand
−36
−38
−40
−42
−44
−46
−48
166
168
170
172
174
176
178
Balkan Region
42
40
38
36
34
18
20
22
24
26
28
30
Worldwide
80
60
40
20
0
−20
−40
−60
−80
−150
−100
−50
0
50
100
150

Figure 26 – Illustration of the earthquake catalogs analyzed in this project. Earthquakes are shown as semi-transparent red dots, and plate boundaries are shown in black.

Source – Prepared by the author using earthquake catalogs that were obtained as described in the text. The maps were obtained from the OpenStreetMap project (see <https://www.openstreetmap.org/>).

5.3 Ensuring Reliability of the Datasets

The Gutenberg-Richter (G-R) relation is a fundamental principle in seismology that describes the frequency-magnitude distribution of earthquakes. It states that the logarithm of the number of earthquakes with a magnitude greater than or equal to a certain value is inversely proportional to the magnitude itself. In other words, smaller earthquakes occur much more frequently than larger ones, following a logarithmic scale. This relation provides valuable insights into the overall seismic activity of a region and forms the basis for seismic hazard assessment, risk analysis, and earthquake forecasting.

When dealing with earthquake datasets, it is often observed that the magnitudes do not follow the G-R relation. The main reason for this is that the earthquake catalogs often cover a long period of time, within which the sensors used to detect earthquakes were constantly improved. This means that older data results from less accurate sensors, that can fail to detect smaller earthquakes, or fail to distinguish earthquakes from other sources of seismic waves, such as explosions in cave mines. Usually, the reliability of the dataset increases if we restrict it to only earthquakes with magnitude above a certain threshold α. In particular, for data obtained in year 2000 or later, it is generally accepted that the threshold $\alpha = 5.1$ is adequate for most datasets with data from 1970 or older (KAGAN; SCHOENBERG, 2001). However, this is a conservative threshold, and if considering a particular earthquake catalog, the threshold that ensures reliability of the data can be a lot lower.

Thus, an important step of earthquake analysis is to select only earthquakes above a magnitude threshold that ensures adherence to the Gutenberg-Richter relation. To address this, we utilize the Kolmogorov test, which assesses whether the seismic moment (i.e. the magnitudes) distribution conforms to a power-law distribution with a specific exponent, denoted as β (the null hypothesis).

A challenge arises in applying the standard Kolmogorov test due to the need for precise knowledge of the distribution used in the null hypothesis. When determining the β parameter from observed data, adjustments to the test become necessary. It has been proposed by Kagan (2002) that a universal value of approximately 0.63 holds for β across all earthquakes. Assuming this conjecture allows for the use of the standard Kolmogorov test.

Figure 27 shows the behavior of the catalogs before and after restricting the earthquakes to being above a certain magnitude threshold. Here, we used a significance level of 0.01 for the Kolmogorov test, and obtained the thresholds (rounded down to one decimal place): 4.7 for Japan, 3.7 for New Zealand, 3.5 for Balkan region and 4.5 for worldwide. Note how the plots become significantly more linear after the thresholds are applied, which show very clearly how much missing data exists within the smaller earthquakes. In the experiments, we always perform the analyses after restricting the catalogs to the appropriate magnitude thresholds, as listed above.

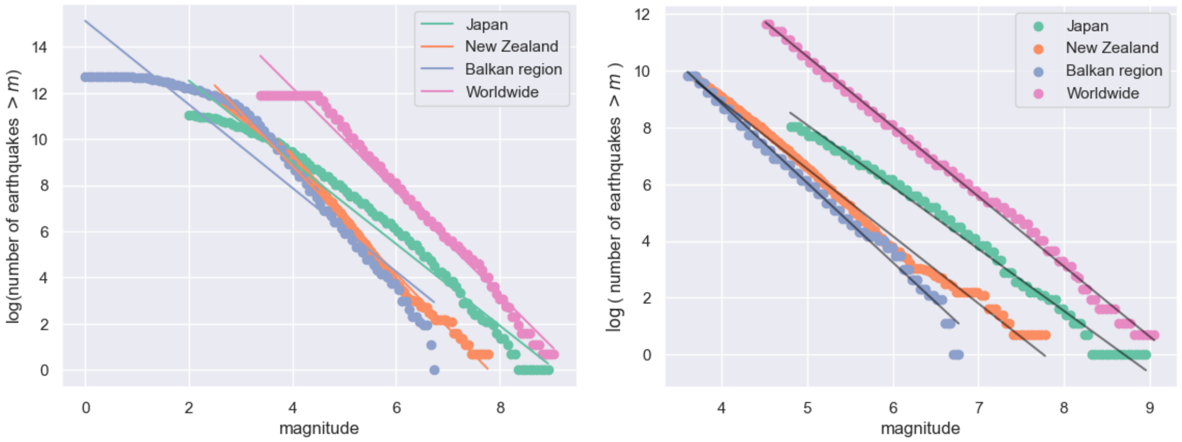


Figure 27 – Behavior of the catalogs before and after restricting the earthquakes to being above a certain magnitude threshold. If the dots form a straight line, that means the data follows the Gutenberg–Richter relation.

Source – Prepared by the author.

5.4 Estimation of the Maximum Earthquake Magnitude

In this section, we apply our proposed method for estimating extreme values in order to analyze earthquakes in the regions presented above. In general, earthquake data contains dependence, especially among samples close in time, but here we consider data collected over a large period of time, and the methods do not rely on the particular order of the samples, so they can be shuffled, which makes the difference between $P(X_i \mid X_j)$ and $P(X_i)$ negligible. We compare our method to the method of Alves and Neves (2014) that was presented and also used for experiments in Chapter 4, and also the method of (KIJKO, 2004).

The method of Kijko (2004) consists of applying some basic concepts of extreme value theory to the context of earthquakes. It starts with the observation that, if the earthquake magnitudes $m_1, m_2, \ldots, m_n$ are sorted in ascending order, the largest observed magnitude (m_n, also denoted as $m_{\max}$) will have cdf:

$$F_{M_n}(m) = \begin{cases} 0, & \text{for } m < m_{\min} \\ [F_M(m)]^n, & \text{for } m_{\min} \leq m \leq m_{\max} \\ 1, & \text{for } m > m_{\max}, \end{cases}$$

where F_M is the cdf of earthquake magnitudes in general.

Kijko (2004) analyzes three different cases, using different functions for F_M, but we consider the first case only, where the Gutenberg–Richter distribution is used. With this, it is possible to integrate:

$$E(M_n) = \int_{m_{\min}}^{m_{\max}} m \, \mathrm{d}F_{M_n}(m) = m_{\max} - \int_{m_{\min}}^{m_{\max}} F_{M_n}(m) \, \mathrm{d}m,$$

and by appropriately substituting $E(M_n)$ with an estimator, and calculating the integral using Cramér's approximation, it yields an estimate for $m_{\max}$.

For our method, we use the generalized extreme value (GEV) distribution to fit the set of largest magnitudes, as suggested in Pisarenko *et al.* (2008). We choose to use the 20% largest magnitudes for the calculations (as is common in the EVT field (BEIRLANT *et al.*, 2004)), or all the earthquakes above the completeness magnitude discussed in Section 5.3. Of these two, we take whichever set has less earthquakes to perform the estimation.

5.4.1 Estimating the Maximum Magnitude for Japan

In this experiment, we applied the three methods mentioned above to the earthquake catalogs from Japan. The earthquake catalog was filtered to include only events with magnitudes above certain thresholds: the completeness threshold discussed in the previous section and the thresholds $m > 5$ and $m > 6$. Since the amount of earthquakes decreases exponentially according to the threshold, we were not able to do the analysis beyond threshold 6 for all datasets.

Table 11 shows the estimated maximum magnitudes for such experimental setting. The maximum magnitude observed in the dataset is 9.0, which is different than the often reported 9.1 due to the Japanese Meteorological Agency using a different kind of seismic moment to calculate magnitude values, as well as the fact that they round the values to the first decimal place.

Table 11 – Estimated maximum magnitudes for Japan using the proposed Maximum Likelihood Estimation with Parameter-Dependent Support (MLEPDS) method, the method by Alves and Neves (2014) and the method by Kijko (2004). Values rounded to the second decimal place.

Magnitude Threshold	**MLEPDS**	**Alves and Neves**	**Kijko**
none	9.39	9.08	9.21
5	9.37	9.10	9.20
6	9.33	9.10	9.27

In the Table, we can immediately observe that the method by Alves and Neves (2014) is always located very close to the observed maximum magnitude, which is certainly a problem. This could be a lack of suitability of the method to the specific kind of seismic activity in Japan, or due to deficiencies inherent to the method; if the latter is the case, we can expect that it will be observed in the other experimental settings as well.

The other two methods offer rather close estimates to each other, with the proposed MLEPDS method offering slightly higher estimated values. For the magnitude threshold of 6, the two methods got a lot closer to each other, which exposes the sensitivity of the methods to the particular selection of earthquakes used to perform the estimation. It is difficult to reason about these results as we do not know the real right endpoint (in contrast with the toy models and experiments performed in Chapter 4), but based on our knowledge of earthquakes and the largest earthquakes that happened in the past, a value of around 9.35 is quite acceptable, albeit still seemingly low. In contrast, the method Alves and Neves (2014) displays far more stability over the different magnitude thresholds, which could be seen as an advantage.

Now let us view it from another perspective, by trying to verify the variance of the estimators, as well as verify how they behave if we make an estimate using a slice of the dataset, and see if this estimate agrees with the actual maximum magnitude observed on another slice of the same dataset.

In this experiment, we simulate the situation where we are constantly observing earthquakes and trying to calculate the maximum magnitude in a certain region with all the data available at that point. To reproduce this with the catalog we have in hand, we use the earthquakes from the beginning of the dataset up to the data of year N, in order to obtain an estimate for the year N+1, with N going from 2011 to 2020. The results are shown in Figure 28.

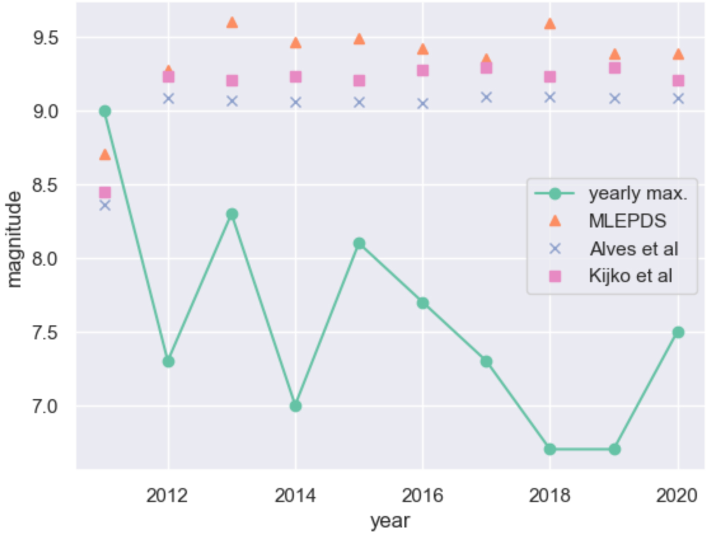


Figure 28 – Maximum magnitude estimates yielded by the proposed MLEPDS method, the method by Alves and Neves (2014), and the method of Kijko (2004) when they are fed with Japanese earthquake data that goes from the year 2000 until the year shown in the x-axis. The solid line shows the observed maximum magnitude on each year, whereas the dots represent the maximum magnitude estimates obtained by each method.

Source – Prepared by the author.

In Figure 28, it is noticeable, in the year of 2011, that none of the methods succeed in correctly covering the maximum earthquake magnitude that occurred in 2011, namely the Tōhoku-Oki earthquake of magnitude 9.0 that happened in Japan in that year. However, the proposed MLEPDS method was the one closest to covering such magnitude, so we see this as an advantage of our proposal in this particular case. After 2011, we notice that all methods adjust their estimators to be above 9.0, as that is now the historical maximum observed magnitude.

The method by Alves and Neves (2014) exhibits the same behavior as observed previously, yielding estimates very close to the historical observed maximum magnitude, which is problematic. On the other hand, the method by Kijko (2004) give slightly larger estimates, and it seems to enjoy low variability as the years pass.

The proposed MLEPDS method, in contrast, tends to give higher estimates for the maximum magnitude, and it seems to have larger variance throughout the years. This larger variance could be an inherent characteristic of the method, resulting from how it is formulated, which might not be a positive characteristic thereof. Otherwise, it could mean that after feeding the method with one extra year of data, it manages to properly adjust its estimate based on that, which would then be a positive characteristic. Overall, given how all the methods underestimated the 2011 earthquake, it seems adequate that the new estimate be significantly distant from the newly observed historical maximum. In fact, since the 2011 earthquake is quite larger than all other earthquakes, any estimation method should adjust their expectations of the underlying distribution so that its right tail is projected as being longer.

5.4.2 Estimating the Maximum Magnitude for New Zealand

For New Zealand we use the same experimental configuration as used for Japan, so we compare the same methods, and consider the same time period. Here we again filter the catalog to include only events with magnitudes above certain thresholds: the completeness threshold discussed previously and the thresholds $m > 5$ and $m > 6$.

Table 12 – Estimated maximum magnitudes for New Zealand using the proposed Maximum Likelihood Estimation with Parameter-Dependent Support (MLEPDS) method, the method by Alves and Neves (2014) and the method by Kijko (2004). Values rounded to the second decimal place.

Magnitude Threshold	**Proposed Method**	**Alves and Neves**	**Kijko**
none	8.14	8.00	8.36
5	8.12	7.98	8.33
6	8.12	7.94	8.40

Table 12 shows the estimated maximum magnitudes for such experimental setting. The maximum magnitude observed in the dataset is 7.82 in this case. The Table shows that, similar than the Japanese case, the method by Alves and Neves (2014) is always closer to the observed maximum magnitude. This reinforces that this could be due to deficiencies inherent to the method.

The other two methods again offer higher estimates, with Kijko's method yielding significantly higher estimates; since the magnitude is related to the logarithm of the energy released, a difference of 0.2 is significant. But compared to Kijko's method, the proposed MLEPDS method yields lower estimated values in all cases. Similar to the Japanese case, our

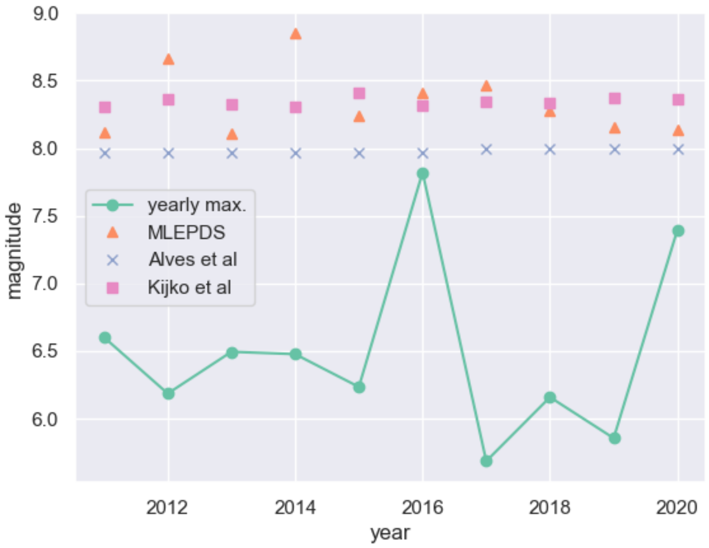


Figure 29 – Maximum magnitude estimates yielded by the proposed MLEPDS method, the method by Alves and Neves (2014), and the method of Kijko (2004) when they are fed with New Zealand earthquake data that goes from the year 2000 until the year shown in the x-axis. The solid line shows the observed maximum magnitude on each year, whereas the dots represent the maximum magnitude estimates obtained by each method.

Source – Prepared by the author.

knowledge of earthquakes and the largest earthquakes that happened in the past would indicate that both the estimates given by the proposed MLEPDS and Kijko's method are convincing, although it would likely be more cautious to choose Kijko's method here over ours. Like with the Japan catalog, when taking a higher magnitude threshold, our method decreased the estimated maximum magnitude; if this always happens, it certainly counts as a disadvantage of our method.

Now let us try again to view these estimators in the situation where we perform an estimate for year N+1 using the catalog until a certain year N, with N ranging from 2011 to 2020. The results are shown in Figure 29.

Since New Zealand is not nearly as seismically active as Japan, the earthquake magnitudes can be seen as more "predictable," so here we do not run into the problem of underestimating the maximum magnitude, no matter the method. The conclusions drawn here are, for the most part, the same as with the Japanese case. Here, the proposed MLEPDS method displayed a somewhat larger variance, with a peak in the year of 2014 that we could not find a clear reason for. However, considering the large earthquake that occurred in 2016, there is a possibility that the proposed MLEPDS method managed to somewhat foresee an increase in seismic activity, and thus raised the maximum magnitude estimate in response to that. But if this is not the case, then the large variance displayed by our proposed method is a negative characteristic thereof. We note that here, again, the method of Alves and Neves (2014) gave estimates that are too close to the historical maximum observed magnitude, and the method of Kijko gave more stable estimates (less variance over the years) than our method.

5.4.3 Estimating the Maximum Magnitude for the Balkan Region

We now present the results concerning the earthquake catalog for the region surrounding the Balkan peninsula. We use here the same experimental configuration as used for the two previous regions, so we compare the same methods, and consider the same time period. Here we again filter the catalog to include only events with magnitudes above certain thresholds: the completeness threshold discussed previously and the thresholds $m > 5$ and $m > 6$.

Table 13 – Estimated maximum magnitudes for the Balkan region using the proposed Maximum Likelihood Estimation with Parameter-Dependent Support (MLEPDS) method, the method by Alves and Neves (2014) and the method by Kijko (2004). Values rounded to the second decimal place.

Magnitude Threshold	Proposed Method	Alves and Neves	Kijko
none	7.26	6.82	7.01
5	7.10	6.88	7.07
6	7.23	6.90	7.11

Table 13 shows the estimated maximum magnitudes, and we note that the maximum magnitude observed in the dataset is 6.8 in this case. The Table shows that, once again, the method by Alves and Neves (2014) is always close to the observed historical maximum magnitude 6.8, although the estimate goes up as we increase the magnitude threshold.

In this experiment we observe that the method by Kijko (2004) in general gives lower estimates than the proposed MLEPDS method, similar to the case of Japan earthquakes. The MLEPDS had a large decrease in the estimate for the magnitude threshold $m > 5$, which exposes the larger variance of our method when compared to the other two methods.

Now let us try again to view these estimators in the situation where we obtain an estimate for year N+1 using the catalog until a certain year N, with N ranging from 2011 to 2020, as shown in Figure 30.

It is noticeable that the seismic activity in the region increased from the years of 2016 to 2020, but the estimators of Alves and Neves (2014) and Kijko (2004) do not seem to react to such changes. In contrast, the proposed MLEPDS method seems to increase the estimate in these last years. Not only that, the higher estimate for year 2012 also preceded a rise in seismicity in the following years. Therefore, this could be indicating that our method manages to perceive this change in seismic activity, which would be a remarkable feature of the method. However, since this was not observed consistently in the other regions, we cannot ascertain for sure if this behavior was not obtained by chance.

Overall, we consider the experiments to show that the method of Alves and Neves (2014) is not very reliable, as it tends to give too modest estimates of the maximum magnitude, which the existent knowledge about earthquakes would quickly disagree with. The method of Kijko (2004) gave a lot more convincing estimates of the maximum magnitudes, and it also displayed lower variance, but it showed very little sensitivity to the variation of seismicity across the year

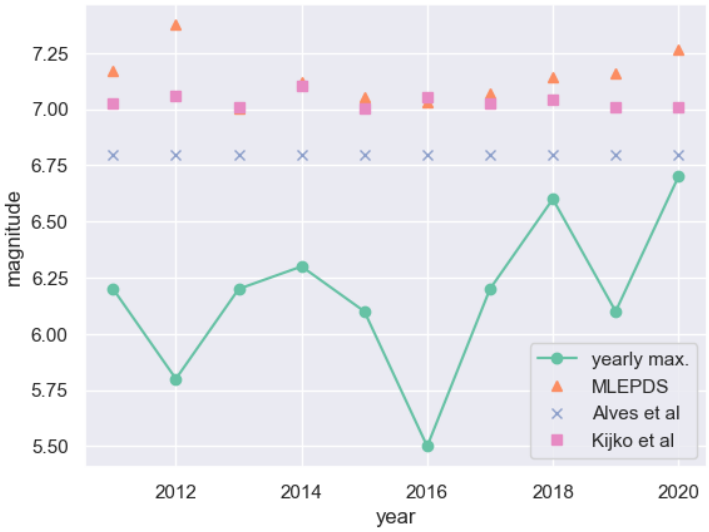


Figure 30 – Maximum magnitude estimates yielded by the proposed MLEPDS method, the method by Alves and Neves (2014), and the method of Kijko (2004) when they are fed with Balkan peninsula earthquake data that goes from the year 2000 until the year shown in the x-axis. The solid line shows the observed maximum magnitude on each year, whereas the dots represent the maximum magnitude estimates obtained by each method.

Source – Prepared by the author.

in each region, and major changes in the estimate were only observed when a new historical maximum was observed.

Our proposed MLEPDS method also gave quite convincing estimates, sometimes higher than the method of Kijko, and sometimes lower. It displayed, however, a large variance, giving results that are harder to interpret. In particular, when analyzing earthquakes above a given magnitude threshold, the method displayed unexpected variability, which makes the interpretation difficult.

When given new data, our method did seem to make an effort to use that extra data to hone the final maximum magnitude estimate, sometimes correctly predicting the trend of earthquake seismicity ahead of time. It would, however, be careless to state, without extensive experiments, that our method manages to predict the trend accurately.

So, overall, we believe that the proposed MLEPDS method competes quite well with the method of Kijko, and it seems to understand and capture different important characteristics of the underlying problem. Because of this, we believe future work could use our method to help with analyzing earthquake activity and making projections about the maximum earthquake magnitude on a given region, even if used as a second opinion.

CHAPTER 6

CONCLUSION

In this project, we showed that it is necessary to deal with the unknown population minimum when performing inference. Based on the literature review done in Chapter 3, it seems safe to assume that nobody has investigated this problem directly, raising questions as to how researchers have dealt with this problem thus far. In this dissertation we described the problem and laid out a basic theoretical framework that we believe is suitable for the problem. Namely, the discussion should be aimed in reducing the difference between the distribution inferred and the actual distribution, consequently placing the bias of endpoint estimators in the background, without ignoring it completely. As a measure of difference between distributions, we have considered the Hellinger and Kantorovich–Wasserstein distances for our experiments, but other choices exist, as commented in Section 4.1.

We have investigated the most immediate method, and likely the one researchers have used the most until now, which is to include the population minimum as a model parameter, and infer it by maximum likelihood estimation (MLE), together with the other parameters. This has the disadvantage of increasing model complexity (ANDERSON; BURNHAM, 2004), but we assessed this method on the gamma and Weibull distributions, and found out that it displays reasonably good results.

Another possibility to deal with the unknown population minimum is to use an endpoint estimator. Section 3.1 discusses the best endpoint estimators we could find in the literature. Some of these methods consist of fitting a distribution on the lowest values of your data, which involves an additional optimization procedure, and that goes against our original objective of not increasing computational cost *that* much. We thus selected the simpler endpoint estimators and investigated their behavior in practice. Section 4.4 shows that none of these estimators give better results than the pure MLE method (with $N = 500$).

Most of the simpler endpoint estimators tested were far from being an ideal way to deal with the unknown population minimum. However, the endpoint estimator of Alves and Neves

(2014) proved to be a very good solution, with satisfactory results, and our proposed estimator based in the law of iterated logarithm (LIL) proved to be better on average than such estimator. These two did not surpass the pure MLE method or our proposed MLEPDS method in terms of minimizing the Kantorovich or Hellinger distances, and they do have a tendency to overestimate the population minimum; but their variance is slightly smaller, and being simpler solutions we expect that they will likely continue to work well even with more complicated probability distributions, which we cannot exactly state for the MLEPDS method. Overall, although we are satisfied that the proposed estimator based in the LIL tends to perform better than Alves et al.'s method, it cannot be ignored that experiments were limited to only two families of distributions, and also that their their method has a more solid theoretical basis.

The maximum likelihood estimation with parameter-dependent support (MLEPDS) method was proposed to perform parametric inference in situations where we seek simultaneously the parameters of the distribution and the left endpoint. As such, it is not enough to simply estimate the left endpoint as done by our proposed method based in the LIL or the method of (ALVES; NEVES, 2014). It is necessary to specify a distribution family (from which the statistician believes the data came from) and estimate its parameters together with the endpoint. In this sense, the proposed MLEPDS method is nearer to the pure MLE method than the endpoint estimates found in the literature. Our results demonstrate that the population minimum found by the MLEPDS method is in most cases better than those given by the simpler endpoint estimates, and is also quite competitive, despite having one less parameter to estimate, when compared with the pure MLE method.

We diagnosed the initially proposed MLEPDS method and realized that the formulation can significantly increase the complexity of the optimization surface for maximizing the likelihood, which leads to many numerical exceptions that ultimately limited how our proposed method performs relative to the pure MLE. Thus, we proposed an iterative procedure to deal with such instability, and the experiments demonstrate that the modified MLEPDS method can outperform the pure MLE method in over 60% of the experimental trials, a proportion that increases for small sample sizes, which we believe makes the method all the more useful for practitioners out there, who often have to deal with limited amounts of data ($<$ 100 samples).

Overall, investigating the usage of existing tools was one of our original objectives, and we believe to have fulfilled it. We not only tested existing ones, but also aimed for finding novel methods, seeking one that would potentially surpass all the others. Our proposed MLEPDS method, described in Section 4.3, displayed very good results in the experiments when it comes to minimizing the Hellinger and Kantorovich distance from the inferred model to the real one, and the modified MLEPDS method managed to be better than the pure MLE method in most of the cases. We believe it was also an important feat to prove Theorem 4, as it served to prove that the same properties of an MLE estimator also applies for our MLEPDS method, and this is a significant advantage.

To further assess the practical usability of the estimators investigated throughout the text, we have tested some of them with data from the field of seismology. Overall, we found that the experiments show that the method of Kijko (2004) gives reasonably convincing estimates of the maximum magnitudes when compared to the other two methods experimented with, but it also showed a remarkable lack of sensitivity to the varying seismicity in each region over the years, which could indicate a lack of capability of capturing essential features of the seismic activity in a given region.

On the other hand, our proposed MLEPDS method also obtained convincing estimates, and it showed some sensitivity to the varying seismicity, although it also showed an undesirable higher variance than the method of Kijko. Therefore, we overall believe that the proposed MLEPDS method manages to understand and capture important characteristics of the underlying problem, maybe in a complementary manner to the Kijko's method, for example. Thus, we strongly suggest that future work also use our method to help with analyzing earthquake activity and making projections about the maximum earthquake magnitude on a given region.

6.1 Acknowledgements

The student thanks the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for the financial support.

BIBLIOGRAPHY

ACERO, F.; CARRASCO, V.; GALLEGO, M.; GARCÍA, J.; VAQUERO, J. Extreme value theory and the new sunspot number series. **The Astrophysical Journal**, IOP Publishing, v. 839, n. 2, p. 98, 2017. Citation on page 51.

ACERO, F.; GALLEGO, M.; GARCÍA, J.; USOSKIN, I.; VAQUERO, J. Extreme value theory applied to the millennial sunspot number series. **The Astrophysical Journal**, IOP Publishing, v. 853, n. 1, p. 80, 2018. Citation on page 51.

AFIFY, A. Z.; CORDEIRO, G. M.; BUTT, N. S. *et al.* A new lifetime model with variable shapes for the hazard rate. **Brazilian Journal of Probability and Statistics**, Brazilian Statistical Association, v. 31, n. 3, p. 516–541, 2017. Citation on page 22.

ALEXOPOULOS, C.; GOLDSMAN, D.; MOKASHI, A. C.; TIEN, K.-W.; WILSON, J. R. Sequest: A sequential procedure for estimating quantiles in steady-state simulations. **Operations Research**, INFORMS, v. 67, n. 4, p. 1162–1183, 2019. Citation on page 70.

ALFÈ, D.; GILLAN, M.; PRICE, G. Temperature and composition of the earth's core. **Contemporary Physics**, Taylor & Francis, v. 48, n. 2, p. 63–80, 2007. Citation on page 92.

ALVES, I. F.; HAAN, L. de; NEVES, C. How far can man go? In: **Advances in Theoretical and Applied Statistics**. [S.l.]: Springer, 2013. p. 187–197. Citation on page 44.

ALVES, I. F.; NEVES, C. Estimation of the finite right endpoint in the gumbel domain. **Statistica Sinica**, JSTOR, v. 24, n. 4, p. 1811–1835, 2014. Citations on pages 11, 13, 14, 44, 46, 74, 86, 100, 101, 102, 103, 104, 105, 106, and 108.

ALVES, I. F.; NEVES, C.; ROSÁRIO, P. A general estimator for the right endpoint with an application to supercentenarian women's records. **Extremes**, Springer, v. 20, n. 1, p. 199–237, 2017. Citation on page 45.

ANDERSON, D.; BURNHAM, K. Model selection and multi-model inference. **Springer**, 2004. Citations on pages 19, 20, and 107.

ATHREYA, K. B.; FUKUCHI, J.-i. Confidence intervals for endpoints of a cdf via bootstrap. **Journal of Statistical Planning and Inference**, Elsevier, v. 58, n. 2, p. 299–320, 1997. Citation on page 50.

AVRAMIDIS, A. N.; WILSON, J. R. Correlation-induction techniques for estimating quantiles in simulation experiments. **Operations Research**, INFORMS, v. 46, n. 4, p. 574–591, 1998. Citation on page 31.

BARTLE, R. G.; SHERBERT, D. R. **Introduction to real analysis**. 4. ed. [S.l.]: Wiley, 2011. Citation on page 30.

BÅTH, M. Earthquake energy and magnitude. **Physics and Chemistry of the Earth**, Elsevier, v. 7, p. 115–165, 1966. Citation on page 93.

BEIRLANT, J.; BOUQUIAUX, C.; WERKER, B. J. Semiparametric lower bounds for tail index estimation. **Journal of Statistical Planning and Inference**, Elsevier, v. 136, n. 3, p. 705–729, 2006. Citation on page 53.

BEIRLANT, J.; GOEGEBEUR, Y.; SEGERS, J.; TEUGELS, J. L. **Statistics of extremes: theory and applications**. [S.l.]: John Wiley & Sons, 2004. Citations on pages 28, 34, 36, 43, 51, 52, 53, and 101.

BERANGER, B.; PADOAN, S. A.; SISSON, S. A. Estimation and uncertainty quantification for extreme quantile regions. **Extremes**, Springer, v. 24, n. 2, p. 349–375, 2021. Citation on page 54.

BIRD, P. An updated digital model of plate boundaries. **Geochemistry, Geophysics, Geosystems**, Wiley Online Library, v. 4, n. 3, 2003. Citations on pages 92 and 97.

BOULFANI, F.; GENDRE, X.; RUIZ-GAZEN, A.; SALVIGNOL, M. A statistical approach for sizing an aircraft electrical generator using extreme value theory. **CEAS Aeronautical Journal**, Springer, p. 1–12, 2021. Citation on page 51.

BOYLE, F.; SHERMAN, D. Scopus™: The product and its development. **The Serials Librarian**, Taylor & Francis, v. 49, n. 3, p. 147–153, 2006. Citation on page 42.

BURCH, B. D. Distribution-dependent and distribution-free confidence intervals for the variance. **Statistical Methods & Applications**, Springer, v. 26, n. 4, p. 629–648, 2017. Citation on page 41.

BYRD, R. H.; LU, P.; NOCEDAL, J.; ZHU, C. A limited memory algorithm for bound constrained optimization. **SIAM Journal on scientific computing**, SIAM, v. 16, n. 5, p. 1190–1208, 1995. Citation on page 56.

CAI, J.-J.; HAAN, L. de; ZHOU, C. Bias correction in extreme value statistics with index around zero. **Extremes**, Springer, v. 16, n. 2, p. 173–201, 2013. Citations on pages 48 and 49.

CHAKRABORTY, G.; CHANDRASHEKHAR, G.; BALASUBRAMANIAN, G. Measurement of extreme market risk: Insights from a comprehensive literature review. **Cogent Economics & Finance**, Taylor & Francis, v. 9, n. 1, p. 1920150, 2021. Citation on page 51.

CHAUDHURI, R. R.; SHARMA, P. An integrated stochastic approach for extreme rainfall analysis in the national capital region of india. **Journal of Earth System Science**, Springer, v. 130, n. 1, p. 1–15, 2021. Citation on page 51.

CHEN, H.; ZHU, X.; QIU, D.; LIU, L. Uncertainty-aware real-time workflow scheduling in the cloud. In: IEEE. **2016 IEEE 9th International Conference on Cloud Computing (CLOUD)**. [S.l.], 2016. p. 577–584. Citation on page 22.

CHUNG, J.; KANNAPPAN, P.; NG, C.; SAHOO, P. Measures of distance between probability distributions. **Journal of mathematical analysis and applications**, Elsevier, v. 138, n. 1, p. 280–292, 1989. Citation on page 59.

CLARIVATE ANALYTICS. **Web of science**. 2021. Access: July 2021. Available: <https://www.webofscience.com>. Citation on page 42.

COOK, K.; WANG, R. Estimation of conditional power for cluster-randomized trials with interval-censored endpoints. **Biometrics**, Wiley Online Library, 2020. Citation on page 41.

CORDEIRO, G. M.; ALIZADEH, M.; OZEL, G.; HOSSEINI, B.; ORTEGA, E. M. M.; ALTUN, E. The generalized odd log-logistic family of distributions: properties, regression models and applications. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 87, n. 5, p. 908–932, 2017. Citation on page 22.

CORDEIRO, G. M.; CASTRO, M. de. A new family of generalized distributions. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 81, n. 7, p. 883–898, 2011. Citation on page 20.

DALEY, D. J.; VERE-JONES, D. *et al.* **An introduction to the theory of point processes: volume I: elementary theory and methods**. [S.l.]: Springer, 2003. Citation on page 95.

DAOUIA, A.; SIMAR, L. Nonparametric efficiency analysis: a multivariate conditional quantile approach. **Journal of Econometrics**, Elsevier, v. 140, n. 2, p. 375–400, 2007. Citation on page 70.

DAS, S. Extreme rainfall estimation at ungauged locations: Information that needs to be included in low-lying monsoon climate regions like bangladesh. **Journal of Hydrology**, Elsevier, v. 601, p. 126616, 2021. Citation on page 51.

DEGROOT, M. H.; SCHERVISH, M. J. **Probability and statistics**. [S.l.]: Pearson, 2012. Citation on page 124.

DEKKERS, A. L.; EINMAHL, J. H.; HAAN, L. D. A moment estimator for the index of an extreme-value distribution. **The Annals of Statistics**, JSTOR, p. 1833–1855, 1989. Citations on pages 49 and 52.

DEMOULIN, V. C.; GUILLOU, A. Extreme quantile estimation for β-mixing time series and applications. **Insurance: Mathematics and Economics**, Elsevier, v. 83, p. 59–74, 2018. Citation on page 21.

DIEBOLT, J.; GARDES, L.; GIRARD, S.; GUILLOU, A. Bias-reduced estimators of the weibull tail-coefficient. **Test**, Springer, v. 17, n. 2, p. 311–331, 2008. Citation on page 53.

DIERCKX, G.; BEIRLANT, J.; WAAL, D. D.; GUILLOU, A. A new estimation method for weibull-type tails based on the mean excess function. **Journal of Statistical Planning and Inference**, Elsevier, v. 139, n. 6, p. 1905–1920, 2009. Citation on page 53.

DONG, H.; NAKAYAMA, M. K. Quantile estimation with latin hypercube sampling. **Operations Research**, INFORMS, v. 65, n. 6, p. 1678–1695, 2017. Citation on page 21.

DREES, H. On smooth statistical tail functionals. **Scandinavian Journal of Statistics**, Wiley Online Library, v. 25, n. 1, p. 187–210, 1998. Citation on page 45.

DREES, H.; FERREIRA, A.; HAAN, L. D. On maximum likelihood estimation of the extreme value index. **Annals of Applied Probability**, JSTOR, p. 1179–1201, 2004. Citation on page 49.

DREES, H. *et al.* Extreme quantile estimation for dependent data, with applications to finance. **Bernoulli**, Bernoulli Society for Mathematical Statistics and Probability, v. 9, n. 4, p. 617–657, 2003. Citations on pages 21 and 54.

DUTFOY, A. Estimation of tail distribution of the annual maximum earthquake magnitude using extreme value theory. **Pure and Applied Geophysics**, Springer, v. 176, n. 2, p. 527–540, 2019. Citation on page 51.

EDELMANN, D.; WELCHOWSKI, T.; BENNER, A. A consistent version of distance covariance for right-censored survival data and its application in hypothesis testing. **Biometrics**, Wiley Online Library, 2021. Citation on page 41.

EMBRECHTS, P.; SCHMIDLI, H. Modelling of extremal events in insurance and finance. **Zeitschrift für Operations Research**, Springer, v. 39, n. 1, p. 1–34, 1994. Citation on page 37.

FALK, M. Some best parameter estimates for distributions with finite endpoint. **Statistics: A Journal of Theoretical and Applied Statistics**, Taylor & Francis, v. 27, n. 1-2, p. 115–125, 1995. Citations on pages 49 and 50.

FALK, M.; MAROHN, F. Von mises conditions revisited. **The Annals of Probability**, JSTOR, p. 1310–1328, 1993. Citation on page 30.

FERREIRA, A.; HAAN, L. d.; PENG, L. On optimising the estimation of high quantiles of a probability distribution. **Statistics**, Taylor & Francis, v. 37, n. 5, p. 401–434, 2003. Citation on page 50.

FISHER, R. A.; TIPPETT, L. H. C. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In: CAMBRIDGE UNIVERSITY PRESS. **Mathematical Proceedings of the Cambridge Philosophical Society**. [S.l.], 1928. v. 24, n. 2, p. 180–190. Citation on page 27.

FRANCO, C.; LITTLE, R. J.; LOUIS, T. A.; SLUD, E. V. Comparative study of confidence intervals for proportions in complex sample surveys. **Journal of survey statistics and methodology**, Oxford University Press, v. 7, n. 3, p. 334–364, 2019. Citation on page 41.

FRÉCHET, M. Sur la loi de probabilité de l'écart maximum. **Annales de la Société Polonaise de Mathématique**, v. 6, p. 93–116, 1927. Citation on page 28.

FRIEDERICHS, P.; THORARINSDOTTIR, T. L. Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. **Environmetrics**, Wiley Online Library, v. 23, n. 7, p. 579–594, 2012. Citation on page 51.

GARDES, L.; GIRARD, S.; GUILLOU, A. Weibull tail-distributions revisited: a new look at some tail estimators. **Journal of Statistical Planning and Inference**, Elsevier, v. 141, n. 1, p. 429–444, 2011. Citations on pages 52 and 53.

GIBBS, A. L.; SU, F. E. On choosing and bounding probability metrics. **International statistical review**, Wiley Online Library, v. 70, n. 3, p. 419–435, 2002. Citation on page 59.

GIRARD, S.; GUILLOU, A.; STUPFLER, G. Estimating an endpoint with high-order moments. **test**, Springer, v. 21, n. 4, p. 697–729, 2012. Citations on pages 47, 48, and 74.

____. Estimating an endpoint with high order moments in the weibull domain of attraction. **Statistics & Probability Letters**, Elsevier, v. 82, n. 12, p. 2136–2144, 2012. Citations on pages 47, 48, and 53.

GLASSERMAN, P. **Monte Carlo methods in financial engineering**. [S.l.]: Springer Science & Business Media, 2013. Citation on page 31.

GNEDENKO, B. Sur la distribution limite du terme maximum d'une serie aleatoire. **Annals of mathematics**, p. 423–453, 1943. Citation on page 27.

GOEGEBEUR, Y.; BEIRLANT, J.; WET, T. D. Generalized kernel estimators for the weibull-tail coefficient. **Communications in Statistics—Theory and Methods**, Taylor & Francis, v. 39, n. 20, p. 3695–3716, 2010. Citation on page 53.

GOLDBERG, D. E. **Fundamentals of chemistry**. [S.l.]: McGraw-Hill, 2006. Citation on page 71.

GOLDENSHLUGER, A.; TSYBAKOV, A. Estimating the endpoint of a distribution in the presence of additive observation errors. **Statistics & probability letters**, Elsevier, v. 68, n. 1, p. 39–49, 2004. Citation on page 44.

GUTENBERG, B.; RICHTER, C. F. Earthquake magnitude, intensity, energy, and acceleration. **Bulletin of the Seismological society of America**, The Seismological Society of America, v. 32, n. 3, p. 163–191, 1942. Citation on page 93.

____. Frequency of earthquakes in california. **Bulletin of the Seismological society of America**, Seismological Society of America, v. 34, n. 4, p. 185–188, 1944. Citation on page 94.

____. Earthquake magnitude, intensity, energy, and acceleration: (second paper). **Bulletin of the seismological society of America**, The Seismological Society of America, v. 46, n. 2, p. 105–145, 1956. Citation on page 93.

HAAN, L.; STADTMÜLLER, U. Generalized regular variation of second order. **Journal of the Australian Mathematical Society**, Cambridge University Press, v. 61, n. 3, p. 381–395, 1996. Citation on page 46.

HAAN, L. d. **On regular variation and its applications to the weak convergence of sample extremes**. Phd Thesis (PhD Thesis) — University of Amsterdam, 1970. Citation on page 30.

HAAN, L. d.; FERREIRA, A. **Extreme value theory: an introduction**. [S.l.]: Springer, 2006. Citations on pages 28, 30, 36, 39, 45, 46, 48, 49, 51, and 52.

HAGEN, M.; AZEVEDO, A. Sun-moon-earth interactions, external factors for earthquakes. **Natural Science**, Scientific Research Publishing, v. 9, n. 6, p. 162, 2017. Citation on page 91.

HAIDRI, R.; KATTI, C.; SAXENA, P. Cost-effective deadline-aware stochastic scheduling strategy for workflow applications on virtual machines. **Concurrency and Computation: Practice and Experience**, Wiley, 2019. Citation on page 22.

HALL, P. On estimating the endpoint of a distribution. **The Annals of Statistics**, JSTOR, p. 556–568, 1982. Citations on pages 41, 42, 43, 46, and 47.

HALL, P.; PARK, B. U. New methods for bias correction at endpoints and boundaries. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 30, n. 5, p. 1460–1479, 2002. Citation on page 50.

HALL, P.; WANG, J. Z. Estimating the end-point of a probability distribution using minimum-distance methods. **Bernoulli**, JSTOR, p. 177–189, 1999. Citation on page 50.

HELLINGER, E. **Die Orthogonalinvarianten quadratischer Formen von unendlichvielen Variabelen**. Phd Thesis (PhD Thesis) — University of Göttingen, 1907. Citation on page 59.

HILL, B. M. A simple general approach to inference about the tail of a distribution. **The annals of statistics**, Institute of Mathematical Statistics, v. 3, n. 5, p. 1163–1174, 1975. Citations on pages 38 and 51.

HOAGLIN, D. C. John w. tukey and data analysis. **Statistical Science**, JSTOR, p. 311–318, 2003. Citation on page 20.

HOSKING, J. R.; WALLIS, J. R. Parameter and quantile estimation for the generalized pareto distribution. **Technometrics**, Taylor & Francis, v. 29, n. 3, p. 339–349, 1987. Citations on pages 48, 49, and 74.

HOSKING, J. R. M.; WALLIS, J. R.; WOOD, E. F. Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. **Technometrics**, Taylor & Francis, v. 27, n. 3, p. 251–261, 1985. Citation on page 30.

HOSOGAYA, N.; MIYAZAKI, T.; FUKUSHIGE, Y.; TAKEMORI, S.; MORIMOTO, S.; YAMAMOTO, H.; HORI, M.; KUROKAWA, T.; KAWASAKI, Y.; HANAWA, M. *et al.* Efficacy and safety of nelfinavir in asymptomatic and mild covid-19 patients: a structured summary of a study protocol for a multicenter, randomized controlled trial. **Trials**, BioMed Central, v. 22, n. 1, p. 1–4, 2021. Citation on page 41.

IDE, S.; YABE, S.; TANAKA, Y. Earthquake potential revealed by tidal influence on earthquake size-frequency statistics. **Nature Geoscience**, Macmillan Publishers Limited, part of Springer Nature, v. 9, p. 834–837, 2016. Citation on page 91.

JAGGER, T. H.; ELSNER, J. B. Modeling tropical cyclone intensity with quantile regression. **International Journal of Climatology: A Journal of the Royal Meteorological Society**, Wiley Online Library, v. 29, n. 10, p. 1351–1361, 2009. Citation on page 51.

JENKINSON, A. F. The frequency distribution of the annual maximum (or minimum) values of meteorological elements. **Quarterly Journal of the Royal Meteorological Society**, Wiley Online Library, v. 81, n. 348, p. 158–171, 1955. Citation on page 30.

KAGAN, Y. Y. Seismic moment distribution revisited: I. statistical results. **Geophysical Journal International**, Blackwell Publishing Ltd Oxford, UK, v. 148, n. 3, p. 520–541, 2002. Citation on page 99.

KAGAN, Y. Y.; SCHOENBERG, F. Estimation of the upper cutoff parameter for the tapered pareto distribution. **Journal of Applied Probability**, Cambridge University Press, v. 38, n. A, p. 158–175, 2001. Citation on page 99.

KAHANER, D.; MOLER, C.; NASH, S. **Numerical methods and software**. [S.l.]: Prentice-Hall, 1989. Citation on page 74.

KALA, Z. Quantile-oriented global sensitivity analysis of design resistance. **Journal of Civil Engineering and Management**, v. 25, n. 4, p. 297–305, 2019. Citation on page 21.

KIJKO, A. Estimation of the maximum earthquake magnitude, m max. **Pure and Applied Geophysics**, Springer, v. 161, p. 1655–1681, 2004. Citations on pages 5, 7, 11, 13, 14, 100, 101, 102, 103, 104, 105, 106, and 109.

KOENKER, R.; BASSETT, G. Regression quantiles. **Econometrica: journal of the Econometric Society**, JSTOR, p. 33–50, 1978. Citation on page 33.

KOENKER, R.; HALLOCK, K. F. Quantile regression. **Journal of economic perspectives**, v. 15, n. 4, p. 143–156, 2001. Citation on page 33.

LARSON, R. L. Latest pulse of earth: Evidence for a mid-cretaceous superplume. **Geology**, Geological Society of America, v. 19, n. 6, p. 547–550, 1991. Citations on pages 25 and 91.

LAWLESS, J. F. **Statistical models and methods for lifetime data**. [S.l.]: Wiley, 2003. Citations on pages 19, 20, and 22.

LEADBETTER, M. R.; LINDGREN, G.; ROOTZÉN, H. **Extremes and related properties of random sequences and processes**. [S.l.]: Springer, 1983. Citation on page 28.

LENG, X.; PENG, L.; WANG, X.; ZHOU, C. Endpoint estimation for observations with normal measurement errors. **Extremes**, Springer, v. 22, n. 1, p. 71–96, 2019. Citations on pages 43 and 44.

LI, D.; PENG, L. Does bias reduction with external estimator of second order parameter work for endpoint? **Journal of statistical planning and inference**, Elsevier, v. 139, n. 6, p. 1937–1952, 2009. Citation on page 48.

____. Comparing extreme models when the sign of the extreme value index is known. **Statistics & probability letters**, Elsevier, v. 80, n. 7-8, p. 739–746, 2010. Citations on pages 46 and 47.

LI, D.; PENG, L.; QI, Y. Empirical likelihood confidence intervals for the endpoint of a distribution function. **Test**, Springer, v. 20, n. 2, p. 353–366, 2011. Citation on page 48.

LI, D.; PENG, L.; XU, X. Bias reduction for endpoint estimation. **Extremes**, Springer, v. 14, n. 4, p. 393–412, 2011. Citation on page 48.

LI, Z.; PENG, L. Bootstrapping endpoint. **Sankhya A**, Springer, v. 74, n. 1, p. 126–140, 2012. Citation on page 46.

LINDSAY, B. G. Mixture models: theory, geometry and applications. In: JSTOR. **NSF-CBMS regional conference series in probability and statistics**. [S.l.], 1995. p. i–163. Citation on page 20.

LIU, G.; WEI, Y.; CHEN, Y.; YU, J.; HU, Y. Forecasting the value-at-risk of chinese stock market using the harq model and extreme value theory. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 499, p. 288–297, 2018. Citation on page 51.

LIU, J.; YANG, X. The convergence rate and asymptotic distribution of the bootstrap quantile variance estimator for importance sampling. **Advances in Applied Probability**, Cambridge University Press, v. 44, n. 3, 2012. Citation on page 21.

LOH, W.-Y. Estimating an endpoint of a distribution with resampling methods. **The Annals of Statistics**, JSTOR, p. 1543–1550, 1984. Citation on page 50.

MAO, L.; KIM, K.; MIAO, X. Sample size formula for general win ratio analysis. **Biometrics**, Wiley Online Library, 2021. Citation on page 41.

MASON, D. M. Laws of large numbers for sums of extreme values. **The annals of probability**, Institute of Mathematical Statistics, p. 754–764, 1982. Citation on page 39.

MASSART, P. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. **The Annals of Probability**, p. 1269–1283, 1990. Citation on page 72.

MCCALL, G. New paradigm for the early earth: did plate tectonics as we know it not operate until the end of the archean? **Australian Journal of Earth Sciences**, Taylor & Francis, v. 57, n. 3, p. 349–355, 2010. Citation on page 92.

MEISTER, A.; NEUMANN, M. H. Deconvolution from non-standard error densities under replicated measurements. **Statistica Sinica**, JSTOR, p. 1609–1636, 2010. Citation on page 44.

METHNI, J. E.; GARDES, L.; GIRARD, S.; GUILLOU, A. Estimation of extreme quantiles from heavy and light tailed distributions. **Journal of Statistical Planning and Inference**, Elsevier, v. 142, n. 10, p. 2735–2747, 2012. Citation on page 53.

MINASNY, B.; MCBRATNEY, A. B. A conditioned latin hypercube method for sampling in the presence of ancillary information. **Computers & Geosciences**, Elsevier, v. 32, n. 9, p. 1378–1388, 2006. Citation on page 21.

MISES, R. E. v. La distribution de la plus grande de n valeurs. **Revue mathématique de l'Union interbalkanique**, v. 1, p. 141–160, 1936. Citation on page 30.

MOLENBERGHS, G.; BUYSE, M.; BURZYKOWSKI, T. The history of surrogate endpoint validation. In: **The evaluation of surrogate endpoints**. [S.l.]: Springer, 2005. p. 67–82. Citation on page 41.

MOOD, A. M. **Introduction to the Theory of Statistics.** [S.l.]: McGraw-hill, 1950. Citations on pages 28, 47, 65, and 71.

MUDHOLKAR, G. S.; SRIVASTAVA, D. K. Exponentiated weibull family for analyzing bathtub failure-rate data. **IEEE Transactions on Reliability**, IEEE, v. 42, n. 2, p. 299–302, 1993. Citations on pages 20, 22, and 57.

MÜLLER, S.; HÜSLER, J. Iterative estimation of the extreme value index. **Methodology and Computing in Applied Probability**, Springer, v. 7, n. 2, p. 139–148, 2005. Citations on pages 49 and 50.

MUSTAFAYEV, R.; HAZLETT, R. Simulation of fluid flow in induced fractures in shale by the lattice boltzmann method. In: SPRINGER. **International Conference on Computational Science**. [S.l.], 2019. p. 575–589. Citation on page 19.

NEVES, C.; PEREIRA, A. Detecting finiteness in the right endpoint of light-tailed distributions. **Statistics & probability letters**, Elsevier, v. 80, n. 5-6, p. 437–444, 2010. Citation on page 49.

NOCEDAL, J.; WRIGHT, S. **Numerical optimization**. [S.l.]: Springer, 2006. Citation on page 56.

OGATA, Y. Statistical models for earthquake occurrences and residual analysis for point processes. **Journal of the American Statistical association**, Taylor & Francis, v. 83, n. 401, p. 9–27, 1988. Citations on pages 95 and 96.

OKUNO, S.; IKEUCHI, K.; AIHARA, K. Practical data-driven flood forecasting based on dynamical systems theory. **Water Resources Research**, Wiley Online Library, v. 57, n. 3, 2021. Citations on pages 19 and 33.

PANAKKAT, A.; ADELI, H. Neural network models for earthquake magnitude prediction using multiple seismicity indicators. **International journal of neural systems**, World Scientific, v. 17, n. 01, p. 13–33, 2007. Citation on page 93.

PANDA, S. K.; JANA, P. K. Efficient task scheduling algorithms for heterogeneous multi-cloud environment. **J supercomput**, Springer, v. 71, n. 4, p. 1505–1533, 2015. Citation on page 22.

PAUL, S.; SHARMA, P. Improved var forecasts using extreme value theory with the realized garch model. **Studies in Economics and Finance**, Emerald Publishing Limited, 2017. Citation on page 51.

PENG, L.; QI, Y. Maximum likelihood estimation of extreme value index for irregular cases. **Journal of Statistical Planning and Inference**, Elsevier, v. 139, n. 9, p. 3361–3376, 2009. Citation on page 49.

PICKANDS, J. Statistical inference using extreme order statistics. **The annals of statistics**, Institute of Mathematical Statistics, p. 119–131, 1975. Citations on pages 39, 51, and 74.

PISARENKO, V.; SORNETTE, A.; SORNETTE, D.; RODKIN, M. New approach to the characterization of m max and of the tail of the distribution of earthquake magnitudes. **Pure and Applied Geophysics**, Springer, v. 165, p. 847–888, 2008. Citation on page 101.

RAMOS, A. A. Extreme value theory and the solar cycle. **Astronomy & Astrophysics**, EDP Sciences, v. 472, n. 1, p. 293–298, 2007. Citation on page 51.

RESNICK, S. I. **Extreme values, regular variation and point processes**. [S.l.]: Springer, 2008. Citation on page 28.

RODRIGUEZ, M. A.; BUYYA, R. Deadline based resource provisioning and scheduling algorithm for scientific workflows. **IEEE Trans Cloud Comput**, IEEE, v. 2, n. 2, 2014. Citation on page 22.

ROSS, S. M. **Introduction to probability models**. [S.l.]: Academic press, 2014. Citation on page 94.

ROYDEN, H. L. **Real analysis**. [S.l.]: Prentice Hall, 1988. Citation on page 59.

SAAD, M.; LEE, I. H. Leveraging hybrid biomarkers in clinical endpoint prediction. **BMC medical informatics and decision making**, Springer, v. 20, n. 1, p. 1–12, 2020. Citation on page 41.

SALDANHA, M.; SOUZA, P. High performance algorithms for counting collisions and pairwise interactions. In: SPRINGER. **19th International Conference on Computational Science**. [S.l.], 2019. p. 182–196. Citation on page 20.

SALDANHA, M. H. J. Probabilistic models for the execution time in stochastic scheduling. **University of São Paulo Repository of Thesis and Dissertations**, 2020. Citation on page 21.

SALDANHA, M. H. J.; SUZUKI, A. K. Determining the probability distribution of execution times. In: IEEE. **IEEE Symposium on Computers and Communications**. [S.l.], 2021. Upcoming. Citations on pages 21, 22, and 24.

SERFLING, R. J. **Approximation theorems of mathematical statistics**. [S.l.]: Wiley, 1980. Citation on page 31.

SHESTAK, V.; SMITH, J.; MACIEJEWSKI, A. A.; SIEGEL, H. J. Stochastic robustness metric and its use for static resource allocations. **J Parallel Dist Com**, Elsevier, v. 68, n. 8, p. 1157–1173, 2008. Citation on page 22.

SMITH, R. L. Maximum likelihood estimation in a class of nonregular cases. **Biometrika**, Oxford University Press, v. 72, n. 1, p. 67–90, 1985. Citation on page 50.

____. Estimating tails of probability distributions. **The annals of Statistics**, JSTOR, p. 1174–1207, 1987. Citation on page 50.

STACY, E. W. A generalization of the gamma distribution. **The Annals of Mathematical Statistics**, Institute of Mathematical Statistics, v. 33, n. 3, p. 1187–1192, 1962. Citations on pages 22, 57, and 85.

STACY, E. W.; MIHRAM, G. A. Parameter estimation for a generalized gamma distribution. **Technometrics**, Taylor & Francis, v. 7, n. 3, p. 349–358, 1965. Citation on page 20.

STEFANSKI, L. A.; CARROLL, R. J. Deconvolving kernel density estimators. **Statistics**, Taylor & Francis, v. 21, n. 2, p. 169–184, 1990. Citation on page 44.

STEIN, M. Large sample properties of simulations using latin hypercube sampling. **Technometrics**, Taylor & Francis Group, v. 29, n. 2, p. 143–151, 1987. Citation on page 97.

STEIN, S.; WYSESSION, M. **An introduction to seismology, earthquakes, and earth structure**. [S.l.]: John Wiley & Sons, 2003. Citations on pages 25, 51, 91, and 93.

SUJANA, J. A. J.; GEETHANJALI, M.; RAJ, R. V. *et al.* **Trust model based scheduling of stochastic workflows in cloud and fog computing**. [S.l.]: Springer, 2019. Citation on page 22.

TANENBAUM, A. S.; BOS, H. **Modern operating systems**. [S.l.]: Pearson, 2015. Citation on page 22.

TANG, Z.; JIANG, L.; ZHOU, J.; LI, K.; LI, K. A self-adaptive scheduling algorithm for reduce start time. **Future Generation Computer Systems**, Elsevier, v. 43, p. 51–60, 2015. Citation on page 22.

TORABI, H.; MONTAZERI, N. H. The logistic-uniform distribution and its applications. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, v. 43, n. 10, p. 2551–2569, 2014. Citation on page 20.

VAART, A. W. V. d. **Asymptotic statistics**. [S.l.]: Cambridge university press, 2000. Citation on page 59.

VALK, C. d. Approximation and estimation of very small probabilities of multivariate extreme events. **Extremes**, Springer, v. 19, n. 4, p. 687–717, 2016. Citation on page 53.

____. Approximation of high quantiles from intermediate quantiles. **Extremes**, Springer, v. 19, n. 4, p. 661–686, 2016. Citations on pages 52 and 53.

VALK, C. d.; CAI, J. A high quantile estimator based on the log-generalized weibull tail limit. **Econometrics and Statistics**, Elsevier, v. 6, p. 107–128, 2018. Citations on pages 21 and 53.

VAPNIK, V. N. **Statistical learning theory**. [S.l.]: John Wiley and Sons, 1998. Citations on pages 31 and 70.

WALPOLE, R. E.; MYERS, R. H.; MYERS, S. L.; YE, K. **Probability and statistics for engineers and scientists**. [S.l.]: Macmillan New York, 1993. Citation on page 20.

WANG, F.; PENG, L.; QI, Y.; XU, M. Maximum penalized likelihood estimation for the endpoint and exponent of a distribution. **Statistica Sinica**, JSTOR, v. 29, n. 1, p. 203–224, 2019. Citations on pages 42 and 43.

WOODROOFE, M. Maximum likelihood estimation of translation parameter of truncated distribution ii. **The Annals of Statistics**, JSTOR, p. 474–488, 1974. Citation on page 49.

XAVIER, V. A.; ANNADURAI, S. Chaotic social spider algorithm for load balance aware task scheduling in cloud computing. **Cluster Computing**, Springer, v. 22, n. 1, p. 287–297, 2019. Citation on page 22.

XU, P.; WANG, D.; WANG, Y.; QIU, J.; SINGH, V. P.; JU, X.; ZHANG, A.; WU, J.; ZHANG, C. Time-varying copula and average annual reliability-based nonstationary hazard assessment of extreme rainfall events. **Journal of Hydrology**, Elsevier, v. 603, p. 126792, 2021. Citation on page 51.

YAZIDI, A.; HAMMER, H. Multiplicative update methods for incremental quantile estimation. **IEEE Transactions on Cybernetics**, IEEE, v. 49, n. 3, p. 746–756, 2017. Citation on page 70.

YUEN, D. A.; KARATO, S.-I.; MARUYAMA, S.; WINDLEY, B. F. Superplumes: beyond plate tectonics. Springer, 2007. Citations on pages 25 and 91.

ZHANG, Q.; ZHU, J.; JIA, C.; XU, S.; JIANG, T.; WANG, S. Epidemiology and clinical outcomes of covid-19 patients in northwestern china who had a history of exposure in wuhan city: Departure time-originated pinpoint surveillance. **Frontiers in Medicine**, Frontiers, v. 8, p. 711, 2021. Citation on page 41.

ZHENG, W.; SAKELLARIOU, R. Stochastic dag scheduling using a monte carlo approach. **J parallel distr com**, Elsevier, v. 73, n. 12, p. 1673–1689, 2013. Citation on page 22.

ZUO, L.; SHU, L.; DONG, S. *et al.* A multi-objective optimization scheduling method based on the ant colony algorithm in cloud computing. **Ieee Access**, IEEE, v. 3, 2015. Citation on page 22.

APPENDIX A

COMPLEMENT TO THE PROOF OF THEOREM 3

The following is a weaker form of Theorem 3, which we originally used as definite proof that our proposed method has the same properties as any maximum likelihood estimator, but was pointed out by a reviewer of one of our papers as insufficient. We reproduce the previous version of the theorem and its proof for documentation purposes of this research project and how it unfolded over time.

Theorem 6. *Let $X_1, \ldots, X_n$ be a random sample from a continuous random variable with pdf $f(x)$, cdf $F(x)$ and support $[m, \infty)$; that is,* $\sup\{x : F(x) > 0\} = m$*. Then for every $i, j \in \{1, \ldots, n\}$, $i \neq j$, we have that $Y_i = X_i - X_{(1)}$ and $Y_j = X_j - X_{(1)}$ are independent.*

Proof. We analyze X_1 and X_2, the rest follows analogously. First consider the joint distribution The distribution of X_1, X_2 and $X_{(1)}$:

$$P(X_1 > a, X_2 > b, X_{(1)} > c) = P\left(X_1 > a, X_2 > b, X_1 > c, X_2 > c, \ldots, X_n > c\right),$$

where X_1 and X_2 appear two times, so by taking the intersection of the events they appear into, we obtain:

$$\begin{aligned} &= P\left(X_1 > \max\{a, c\}, X_2 > \max\{b, c\}, X_3 > c, \ldots, X_n > c\right) \\ &= P\left(X_1 > \max\{a, c\}\right) \cdot P\left(X_2 > \max\{b, c\}\right) \cdot \ldots \cdot P\left(X_n > c\right) \\ &= \left[1 - F(\max\{a, c\})\right]\left[1 - F(\max\{b, c\})\right]\left[1 - F(c)\right]^{n-2}. \end{aligned} \tag{A.1}$$

With analogous steps, we can obtain:

$$P(X_1 > a, X_{(1)} > c) = \left[1 - F(\max\{a, c\})\right]\left[1 - F(c)\right]^{n-1} \tag{A.2}$$

$$P(X_2 > a, X_{(1)} > c) = [1 - F(\max\{b, c\})]\,[1 - F(c)]^{n-1}\,. \tag{A.3}$$

Now consider the variable transform:

$$\begin{cases} U &= X_1 - X_{(1)} \\ V &= X_2 - X_{(1)} \\ W &= X_{(1)} \end{cases} \xRightarrow{\text{invert}} \begin{cases} X_1 &= U + W \\ X_2 &= V + W \\ X_{(1)} &= W. \end{cases}$$

Note that since $X_{(1)}$ is the smallest value in the sample, we have that $U > 0$ and $V > 0$. The above transform means that:

$$P(U > u, V > v, W > w) = P(X_1 > u + w, X_2 > v + w, X_{(1)} > w)$$

since we already obtained the joint distribution of $X_1, X_2, X_{(1)}$ in Equation (A.1):

$$= [1 - F(\max\{u + w, w\})]\,[1 - F(\max\{v + w, w\})]\,[1 - F(w)]^{n-2}$$

now since $u > 0$ and $v > 0$, the maximum operators always unfold into $u + w$ or $v + w$:

$$= [1 - F(u + w)]\,[1 - F(v + w)]\,[1 - F(w)]^{n-2}\,.$$

To take the marginal off a cumulative probability function we can take the limit of the event $W > w$ for $w \to -\infty$ (DEGROOT; SCHERVISH, 2012), which is equivalent to the event $W > m$, as the population minimum m is the lower bound of the support:

$$P(U > u, V > v) = \lim_{w \to -\infty} P(U > u, V > v, W > w) = P(U > u, V > v, W > m)$$

$$= [1 - F(u + m)]\,[1 - F(v + m)]\,[1 - \cancel{F(m)}]^{n-2}$$

$$= [1 - F(u + m)]\,[1 - F(v + m)] = P(X_1 - X_{(1)} > u, X_2 - X_{(1)} > v). \tag{A.4}$$

Now if we use the transform $U = X_1 - X_{(1)}$ and $W = X_{(1)}$, we obtain the inverse $X_1 = U + W$ and $X_{(1)} = W$, which means:

$$P(U > u, W > w) = P(X_1 > u + w, X_{(1)} > w)$$

$$= [1 - F(\min\{u + w, w\})]\,[1 - F(w)]^{n-1} \quad \text{[from Equation (A.2)]}$$

$$= [1 - F(u + w)]\,[1 - F(w)]^{n-1}\,,$$

and using the same trick as before, we have:

$$P(U > u) = P(X_1 - X_{(1)} > u) = P(U > u, W > m) = [1 - F(u + m)]\,. \tag{A.5}$$

It is easy to see that by following a very similar procedure we obtain:

$$P(V > v) = P(X_2 - X_{(1)} > u) = P(V > v, W > m) = [1 - F(v + m)]\,. \tag{A.6}$$

Comparing Equations (A.4), (A.5) and (A.6), we see that:

$$\begin{aligned} P(X_1 - X_{(1)} > u, X_2 - X_{(1)} > v) &= [1 - F(u + m)]\,[1 - F(v + m)] \\ &= P(X_1 - X_{(1)} > u) \cdot P(X_2 - X_{(1)} > v), \end{aligned}$$

concluding the proof of independence. □

ICMC USP
SÃO CARLOS