



Faculty of Natural and Agricultural Sciences
Department of Actuarial Science and Mathematical Statistics

Second-order estimation procedures for complete and incomplete heavy-tailed data

Author: Gaonyalelwe Maribe

Promoter: Dr Andréhette Verster

Submitted in accordance with the requirements for the degree of
MAGISTER SCIENTIAE in MATHEMATICAL STATISTICS
at the University of Free State
June 23, 2016

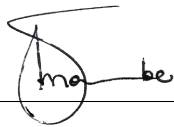
This work is based on the research supported wholly/in part by the National Research Foundation of South Africa (Grant Number 94762). The Grant-holder acknowledges that opinions, findings and conclusions or recommendations expressed in any publication generated by the NRF supported research is that of the author(s), and that the NRF accepts no liability whatsoever in this regard



Declaration

I hereby declare that this work, submitted for the degree Magister Scientiae, at the University of the Free State, is my own original work and has not previously been submitted, for degree purposes or otherwise, to any other institution of higher learning. I further declare that all sources cited or quoted are indicated and acknowledged by means of a comprehensive list of references. Copyright hereby cedes to the University of the Free State.

Signature: _____

A handwritten signature in black ink, appearing to read 'ma be', written over a horizontal line. The signature is stylized with a large, looped initial 'm'.

Date: **June 23, 2016**

Abstract

This thesis investigates the second-order refined peaks over threshold model called the Extended Pareto Distribution (EPD) introduced by Beirlant et al. (2009). Focus is placed on estimation of the Extreme Value Index (EVI). Firstly we investigate the effectiveness of the EPD in modelling heavy-tailed distributions and compare it to the Generalized Pareto Distribution (GPD) in terms of the bias, mean squared error and variance of the EVI. This is done through a simulation study and the Maximum Likelihood (ML) method of estimation is used to make the comparison.

In practice, data can be tampered by some arbitrary process or study design. We therefore investigate the performance of the EPD in estimating the EVI for heavy-tailed data under the assumption that the data is completely observable and uncontaminated, random right censored and contaminated respectively.

We suggest an improved ML numerical procedure in the estimation of EPD parameters under the assumption that data is completely observable and uncontaminated. We further propose a Bayesian EPD estimator of the EVI and show through a simulation study that this estimator leads to much improved results as the ML EPD estimator. A small case study is conducted to assess the performance of the Bayesian EPD estimator and the ML EPD estimator using a real dataset from a Belgian reinsurance firm.

We investigate the performance of some well known parametric and semi-parametric estimators of the EVI adapted for censoring by a simulation study and further illustrate their performance by applying them to a real survival dataset. A censored Bayesian EPD estimator for right censored data is then proposed through an altered expression of the posterior density. The censored Bayesian EPD estimator is compared with the censored ML EPD estimator through a simulation study.

Behaviour of the minimum density power divergence estimator (MDPDE) is assessed at uncontaminated and contaminated distributions respectively through an exhaustive simulation study including other EPD estimators mentioned in this thesis. The comparison is made in terms of the bias and mean squared error. EVI estimates from the different estimators are then used to estimate quantiles, the results are reported concurrently with the EVI estimates. We illustrate the performance of all mentioned estimators on a real dataset from geopedology, in which a few abnormal soil measurements highly influence the estimates of the EVI and high quantiles.

*I wish to dedicate this dissertation to my late mother
Mamosiako Emily Maribe–Moremi.*

Acknowledgements

I would like to express my gratitude to the following people:

First and foremost I would like to thank God. Thank you for being in control – and not me.

Dr Andréhette Verster, for her guidance throughout my postgraduate student life, for believing that I could do things for which I deemed impossible, for inspiring and pushing me to new levels and for her much appreciated mentorship.

A special thanks to Mr Sean van der Merwe, for his suggestions and constant review of my code.

My Family: My Father, Mother, Brothers, Sisters, Aunts and Uncles. Most importantly my two aunts Kelebogile Moremi and Disebo Moremi who instilled in me the ability to dream beyond my circumstances.

Pearl Kgabo Mphahlele, for her constant love and support. Despite the abstraction of my study, I could always run my ideas by her.

Simon Mosia, for introducing me to his precept of research, for supporting and encouraging me to pursue my postgraduate studies.

My Friends: Anthea Davids and Tokelo Fako for constantly having to comment on my never ending LaTeX graphics. Judy Mothibedi for her support and for allowing me to use her office throughout my dissertation.

Lerato Molisana, for hours of proofreading this dissertation and for her invaluable suggestions.

Contents

1	Introduction	1
1.1	Problem statement	2
1.2	Objectives and contribution of the study	2
1.3	Outline of study	3
2	Overview of Extreme Value Theory	5
2.1	Introduction	5
2.2	Historic developments of classical theory of extremes	5
2.2.1	Bayesian methods in extreme value theory	7
2.2.2	Tail estimation	7
2.2.3	Tail estimation under contamination	8
2.2.4	Tail estimation theory under censoring	8
2.3	Extremal limit problem	9
2.3.1	The generalized Pareto distribution	10
2.4	Graphical tools	11
2.5	Areas of Application	13
2.6	Second Order Conditions and Regular Variation	14
2.6.1	Fréchet-Pareto class	17
3	Extended Pareto Distribution	20
3.1	Introduction	20
3.2	Generalization of the EPD	21
3.3	Estimation procedures: Maximum likelihood	23
3.3.1	Estimation of the second order parameter	26
3.4	Simulation Study	27
3.5	Conclusion	32
4	Improved EPD maximum likelihood estimator for complete data	33
4.1	Introduction	33
4.2	Estimation procedure	34
4.3	Simulation study	34

5	EPD Bayesian estimator for complete data	38
5.1	Introduction	38
5.2	Overview of Bayesian Inference	39
5.3	Markov Chain Monte Carlo methods	40
5.3.1	Metropolis algorithm	40
5.3.2	Gibbs sampler	41
5.3.3	Implementation the EPD joint posterior in OpenBUGS . . .	41
5.3.4	Prior Choice	44
5.3.5	Convergence diagnostics	47
5.3.6	OpenBUGS Efficient Sampling	59
5.4	Asymptotic theory	59
5.5	Simulation study: MLE vs Bayes	60
5.5.1	Case 1: Estimating only two parameters	60
5.5.2	Case 2: Estimating all three parameters	66
5.6	Estimates of the Second Order Parameter	72
5.7	Case-Study	77
5.8	Conclusion	80
6	Estimation of the EVI under random censorship	81
6.1	Introduction	81
6.1.1	Censoring in Practice	82
6.2	Random right Censoring in EVT	83
6.2.1	Semi-Parametric Estimators	85
6.2.2	Parametric Estimators	86
6.3	Monte Carlo Simulation	88
6.3.1	Case 1: 25% Censoring	89
6.3.2	Case 2: 40% Censoring	90
6.3.3	Case 3: 55% Censoring	91
6.4	Application to Cancer Survival Data	92
6.5	An asymptotically unbiased Bayesian estimator for random right censored data	94
6.6	Monte Carlo Simulation: Bayes vs. MLE	95
6.6.1	Fréchet	96
6.6.2	Burr (XII)	99
6.6.3	Student-t	102
6.7	Conclusion	105
7	Estimation of the EVI for contaminated data	106
7.1	Introduction	106
7.2	Minimum Density Power Divergence Estimator	107
7.2.1	Estimation procedure	108

7.3	Simulation Study	108
7.3.1	Uncontaminated Distributions	109
7.3.2	Contaminated Distributions	116
7.4	Case Study	136
7.4.1	EVI and High Quantile estimates	140
7.5	Conclusion	144
8	Conclusion	145
8.1	Further research possibilities	146
	References	146

List of Tables

2.1	Q-Q plot coordinates for some distributions.	12
2.2	List of Pareto-type distributions.	19
5.1	Gibbs sampler: 1000 iterations of 3 chains	49
5.2	Metropolis algorithm: 1000 iterations of 3 chains	49
5.3	OpenBUGS: 1000 iterations of 3 chains	49
5.4	Gibbs sampler: 1000 iterations of 6 chains	50
5.5	Metropolis algorithm: 1000 iterations of 6 chains	50
5.6	OpenBUGS: 1000 iterations of 6 chains	50
5.7	Gibbs sampler: 5000 iterations, 3 chains	51
5.8	Metropolis algorithm: 5000 iterations, 3 chains	51
5.9	OpenBUGS: 5000 iterations, 3 chains	51
5.10	Gibbs sampler: 5000 iterations, 6 chains	52
5.11	Metropolis algorithm: 5000 iterations, 6 chains	52
5.12	OpenBUGS: 5000 iterations, 6 chains	52
5.13	1000 iterations, 3 chains (time in seconds)	53
5.14	1000 iterations, 6 chains (time in seconds)	53
5.15	5000 iterations, 3 chains (time in seconds)	53
5.16	5000 iterations, 6 chains	53
5.17	Sampling method hierarchy used by WinBUGS. Each method is only used if no previous method in the hierarchy is appropriate (Lunn et al., 2012).	59
5.18	Pareto type distributions with the second order parameter	73
7.1	Simulation setting for uncontaminated data	109
7.2	Simulation setting for contaminated data	117

List of Figures

2.1	Depiction of the method of Block-maxima (left) and Peaks over a threshold t (right)	11
3.1	Fréchet(α) samples: Mean Square Error (MSE) estimate (top), Variance estimate (middle) and Bias estimate (bottom). The sample size is $n = 1000$ and the plots are obtained by averaging out over 10,000 samples.	29
3.2	Burr(η, τ, λ) samples: Mean Square Error (MSE) estimate (top), Variance estimate (middle) and Bias estimate (bottom). The sample size is $n = 1000$ and the plots are obtained by averaging out over 10,000 samples.	30
3.3	Student-t(v) samples: Mean Square Error (MSE) estimate (top), Variance estimate (middle) and Bias estimate (bottom). The sample size is $n = 1000$ and the plots are obtained by averaging out over 10,000 samples.	31
4.1	Mean Square Error (MSE) estimate (top), Variance estimate (middle), and Bias estimate (bottom). The sample size is $n = 1000$ and the plots are obtained by averaging out over 10,000 samples.	35
4.2	Mean Square Error (MSE) estimate (top), Variance estimate (middle), and Bias estimate (bottom). The sample size is $n = 1000$ and the plots are obtained by averaging out over 10,000 samples.	36
5.1	Representation of the EPD model as a directed acyclic graph (DAG)	43
5.2	Traceplot of the EVI posterior simulations for 1000 iterations of 3 chains	55
5.3	Traceplot of the EVI posterior simulations for 1000 iterations of 6 chains	56
5.4	Traceplot of the EVI posterior simulations for 5000 iterations of 3 chains	57
5.5	Traceplot of the EVI posterior simulations for 5000 iterations of 6 chains	58

5.6	EVI estimates of a Fréchet sample of size $n=200$ (left), and size $n=500$ (right), and EVI of 1 (top) and 0.5 (bottom), 90 % HPD (shaded fill)	61
5.7	MSE of EVI estimates for a Fréchet sample of size $n=200$ (left), and size $n=500$ (right), and EVI of 1 (top) and 0.5 (bottom)	62
5.8	EVI estimates of a Burr sample of size $n=200$ (left), and size $n=500$ (right), and EVI of 1 (top) and 0.5 (bottom), 90 % HPD (shaded fill)	63
5.9	MSE of EVI estimates for a Burr sample of size $n=200$ (left), and size $n=500$ (right), and EVI of 1 (top) and 0.5 (bottom)	64
5.10	EVI estimates of a Student-t sample of size $n=200$ (left), and size $n=500$ (right), and EVI of 1 (top) and 0.5 (bottom), 90 % HPD (shaded fill)	65
5.11	MSE of EVI estimates for a Student-t sample of size $n=200$ (left), and size $n=500$ (right), and EVI of 1 (top) and 0.5 (bottom)	66
5.12	EVI estimates of a Fréchet sample of size $n=200$ (left), and size $n=500$ (right), and EVI of 1 (top) and 0.5 (bottom), 90 % HPD (shaded fill)	67
5.13	MSE of EVI estimates for a Fréchet sample of size $n=200$ (left), and size $n=500$ (right), and EVI of 1 (top) and 0.5 (bottom)	68
5.14	EVI estimates of a Burr sample of size $n=200$ (left), and size $n=500$ (right), and EVI of 1 (top) and 0.5 (bottom), 90 % HPD (shaded fill)	69
5.15	MSE of EVI estimates for a Burr sample of size $n=200$ (left), and size $n=500$ (right), and EVI of 1 (top) and 0.5 (bottom)	70
5.16	EVI estimates of a Student-t sample of size $n=200$ (left), and size $n=500$ (right), and EVI of 1 (top) and 0.5 (bottom), 90 % HPD (shaded fill)	71
5.17	MSE of EVI estimates for a Student-t sample of size $n=200$ (left), and size $n=500$ (right), and EVI of 1 (top) and 0.5 (bottom)	72
5.18	Simulated mean value of τ for a sample of sizes $n = 200$ (left) and $n = 500$ (right) from a Fréchet distribution (where true value $\tau = \rho = -1$), zoomed out view (top), zoomed in view (bottom) . . .	74
5.19	Simulated mean value of τ for a sample of sizes $n = 200$ (left) and $n = 500$ (right) from a Burr distribution (where true value $\tau = \rho = -1$), zoomed out view (top), zoomed in view (bottom) . . .	75
5.20	Simulated mean value of τ for a sample of sizes $n = 200$ (left) and $n = 500$ (right) from a Student-t distribution (where true value $\tau = \rho = -2$), zoomed out view (top), zoomed in view (bottom) . . .	76
5.21	Estimates of the Extreme value index (a) and the quantile estimate associated with $p = 0.001$ (b) for the Secura Belgian Re data	78
5.22	Estimates of the tail probabilities of X greater than an operational priority level $R = 5,000,000$ for the Secura Belgian Re data	79

6.1	A Fréchet ($\gamma_X = 0.25$) distribution censored by a Fréchet ($\gamma_Y = 0.75$), thus $\gamma_Z = \frac{0.25 \cdot 0.75}{0.25 + 0.75} = 0.1875$ and proportion of non-censored data $-p = 75\%$ (25% censoring in the right tail); (a) EVI estimates adapted for censoring and (b) EVI estimate not adapted for censoring, (c) RMSE	89
6.2	A Fréchet ($\gamma_X = 0.25$) distribution censored by a Fréchet ($\gamma_Y = 0.375$), thus $\gamma_Z = \frac{0.25 \cdot 0.375}{0.25 + 0.375} = 0.15$ and proportion of non-censored data $-p = 60\%$ (40% Censoring in the right tail); (a) EVI estimates adapted for censoring and (b) EVI estimate not adapted for censoring, (c) RMSE	90
6.3	A Fréchet ($\gamma_X = 0.25$) distribution censored by a Fréchet ($\gamma_Y = 0.2046$), thus $\gamma_Z = \frac{0.25 \cdot 0.2046}{0.25 + 0.2046} = 0.1125$ and proportion of non-censored data $-p = 45\%$ (65% Censoring in the right tail); (a) EVI estimates adapted for censoring and (b) EVI estimate not adapted for censoring, (c) RMSE	91
6.4	Pareto Quantile-Quantile plots for survival data related with cancer of the tongue	92
6.5	Estimates of the EVI for survival data related with cancer of the tongue	93
6.6	Estimates of γ_X (left) and MSE (right) for a Fréchet (γ_X) distribution censored by another Fréchet (γ_Y) at 25% censoring in the right-tail of X	96
6.7	Estimates of γ_X (left) and MSE (right) for a Fréchet (γ_X) distribution censored by another Fréchet (γ_Y) at 40% censoring in the right-tail of X	97
6.8	Estimates of γ_X (left) and MSE (right) for a Fréchet (γ_X) distribution censored by another Fréchet (γ_Y) at 55% censoring in the right-tail of X	98
6.9	Estimates of γ_X (left) and MSE (right) for a Burr (γ_X) distribution censored by another Burr (γ_Y) at 25% censoring in the right-tail of X	99
6.10	Estimates of γ_X (left) and MSE (right) for a Burr (γ_X) distribution censored by another Burr (γ_Y) at 40% censoring in the right-tail of X	100
6.11	Estimates of γ_X (left) and MSE (right) for a Burr (γ_X) distribution censored by another Burr (γ_Y) at 55% censoring in the right-tail of X	101
6.12	Estimates of γ_X (left) and MSE (right) for a Student-t (γ_X) distribution censored by another Student-t (γ_Y) at 25% censoring in the right-tail of X	102
6.13	Estimates of γ_X (left) and MSE (right) for a Student-t (γ_X) distribution censored by another Student-t (γ_Y) at 40% censoring in the right-tail of X	103

6.14	Estimates of γ_X (left) and MSE (right) for a Student-t (γ_X) distribution censored by another Student-t (γ_Y) at 55% censoring in the right-tail of X	104
7.1	Estimates of the EVI (top) and U(1-1/500) (bottom) of a Fréchet sample with $\text{EVI} = 0.5$	110
7.2	MSE of EVI estimates (top) and U(1-1/500) (bottom) of a Fréchet sample with $\text{EVI} = 0.5$	111
7.3	Estimates of the EVI (top) and U(1-1/500) (bottom) of a Burr (XII) sample with $\text{EVI} = 0.5$	112
7.4	MSE of EVI estimates (top) and U(1-1/500) (bottom) of a Burr (XII) sample with $\text{EVI} = 0.5$	113
7.5	Estimates of the EVI (top) and U(1-1/500) (bottom) of a Student-t sample with $\text{EVI} = 0.5$	114
7.6	MSE of EVI estimates (top) and U(1-1/500) (bottom) of a Student-t sample with $\text{EVI} = 0.5$	115
7.7	Estimates of the EVI (top) and U(1-1/500) (bottom) of a Fréchet sample with $\text{EVI} = 0.5$	118
7.8	MSE of EVI estimates (top) and U(1-1/500) (bottom) of a Fréchet sample with $\text{EVI} = 0.5$	119
7.9	Estimates of the EVI (top) and U(1-1/500) (bottom) of a Fréchet sample with $\text{EVI} = 0.5$	120
7.10	MSE of EVI estimates (top) and U(1-1/500) (bottom) of a Fréchet sample with $\text{EVI} = 0.5$	121
7.11	Estimates of the EVI (top) and U(1-1/500) (bottom) of a Fréchet sample with $\text{EVI} = 0.5$	122
7.12	MSE of EVI estimates (top) and U(1-1/500) (bottom) of a Fréchet sample with $\text{EVI} = 0.5$	123
7.13	Estimates of the EVI (top) and U(1-1/500) (bottom) of a Burr (XII) sample with $\text{EVI} = 0.5$	124
7.14	MSE of EVI estimates (top) and U(1-1/500) (bottom) of a Burr (XII) sample with $\text{EVI} = 0.5$	125
7.15	Estimates of the EVI (top) and U(1-1/500) (bottom) of a Burr (XII) sample with $\text{EVI} = 0.5$	126
7.16	MSE of EVI estimates (top) and U(1-1/500) (bottom) of a Burr (XII) sample with $\text{EVI} = 0.5$	127
7.17	Estimates of the EVI (top) and U(1-1/500) (bottom) of a Burr (XII) sample with $\text{EVI} = 0.5$	128
7.18	MSE of EVI estimates (top) and U(1-1/500) (bottom) of a Burr (XII) sample with $\text{EVI} = 0.5$	129

7.19	Estimates of the EVI (top) and U(1-1/500) (bottom) of a Student-t sample with $\text{EVI} = 0.5$	130
7.20	MSE of EVI estimates (top) and U(1-1/500) (bottom) of a Student-t sample with $\text{EVI} = 0.5$	131
7.21	Estimates of the EVI (top) and U(1-1/500) (bottom) of a Student-t sample with $\text{EVI} = 0.5$	132
7.22	MSE of EVI estimates (top) and U(1-1/500) (bottom) of a Student-t sample with $\text{EVI} = 0.5$	133
7.23	Estimates of the EVI (top) and U(1-1/500) (bottom) of a Student-t sample with $\text{EVI} = 0.5$	134
7.24	MSE of EVI estimates (top) and U(1-1/500) (bottom) of a Student-t sample with $\text{EVI} = 0.5$	135
7.25	Plot of Ca versus pH	137
7.26	Boxplots of the Ca measurements at different pH levels	138
7.27	Pareto quantile plots of the Ca measurements at different pH levels	139
7.28	EVI (left) and corresponding quantiles (right) of the Ca measurements as a function of k at different pH levels	140
7.29	EVI (left) and corresponding quantiles (right) of the Ca measurements as a function of k at different pH levels	141
7.30	EVI (left) and corresponding quantiles (right) of the Ca measurements as a function of k at different pH levels	142

Abbreviations and Acronyms

2ERV	2nd Order Extended Regular Variation
BFGS	Broyden-Fletcher-Goldfarb-Shanno
BUGS	Bayesian inference Under Gibbs sampling
GEV	Generalized Extreme Value
EPD	Extended Pareto Distribution
ERV	Extended Regular Variation
EVI	Extreme Value Index
EVT	Extreme Value Theory
GPD	Generalized Pareto Distribution
HPD	Higest Posterior Density
MCMC	Markov Chain Monte Carlo
MDI	Maximal Data Information
MDPDE	Minimum Density Power Divergence Estimate
ML	Maximum Likelihood
MSE	Mean square error
POT	Peaks Over Threshold
RMSE	Root Mean Square Error
RV	Regular Variation

Notation

γ	Extreme value index.
Γ	Gamma function.
$\mathcal{D}_{\mathcal{M}}(G_{\gamma})$	domain of attraction of G_{γ} .
F^{\leftarrow}	Left-continuous version of the function F .
$o(1)$	When considering the limit of $f(x)$ as $x \rightarrow c$, $f(x) = g(x)(1 + o(1)) \Rightarrow f(x)/g(x) \rightarrow 0$ as $x \rightarrow c$.
$o(g(x))$	When considering the limit of $f(x)$ as $x \rightarrow c$, we write $f(x) = o(g(x))$ if and only if $f(x)/g(x) \rightarrow 0$ as $x \rightarrow c$.
ω_i	Censoring indicator, 0 if the observation is right-censored and 1 if not.
▲	Indicates the end of a definition, proof, lemma or proposition.
ℓ_U & ℓ_U	Slowly varying functions.
$F(t) \overset{t \uparrow \infty}{\sim} G(t)$	$F(t)$ is approximately $G(t)$ as t goes to infinity.
$Q(p)$	The quantile function denoted by $Q(p) = \inf\{x F(x) \geq p\}$.
$H_{k,n}$	The Hill estimator based on k excesses above the threshold.
$ \delta \in \mathcal{R}_{\tau}$	Implies $\delta \rightarrow 0$ as $x \rightarrow \infty$ with $\tau < 0$.

Chapter 1

Introduction

“The top 1% of a population owns 40% of the wealth; the top 2% of Twitter users send 60% of all tweets; medical care for the most expensive one fifth of patients accounts for four fifths of total spending. These figures are always reported as shocking, as if anything but a nice bell curve were an aberration, but Pareto distributions pop up all over. Regarding them as anomalies prevents us from thinking clearly about the world”

—Clay Shirky, as quoted in Newsweek the Guardian

Heavy-tailed processes are found everywhere; financial meltdowns, political catastrophes, natural disasters, insurance claims, amongst others. We usually treat heavy-tailed events as somewhat surprising and mysterious events since they happen with such small probabilities but have such adverse consequences. Extreme events raise controversial issues in statistics in the sense that most commonly employed, simple and statistical appealing methods fail to capture many strange properties of the underlying distribution of such processes.

Extreme Value Theory (EVT) aims to describe the tails of the underlying distribution in order for us to understand its properties thereof. Through EVT we are thus able to now quantify the probabilities of occurrence of these extreme observations. In this thesis, primary focus is placed on distributions with heavy-tails i.e. heavier than exponential.

The aim of this chapter is to give a description of the research problem at hand and explain the scope of this study as well as its contribution to EVT. An outline of the entire dissertation is given at the end of this chapter.

1.1 Problem statement

A crucial parameter in EVT is the Extreme Value Index (EVI). This parameter measures the tail heaviness of the underlying distribution and is used to estimate other extreme quantities such as small tail probabilities, high quantiles and return periods. It is therefore a fundamental concern of EVT to develop new and accurate estimators of the EVI.

The development of new estimators of the EVI is usually anchored on bias reduction, variance reduction, efficiency and more recently robustness.

According to the Peaks Over Threshold (POT) methodology, the distribution of excesses over a threshold converges in distribution to a Generalized Pareto distribution (GPD), a conditional distribution function in the limit of some infinitely high threshold. There exists a number of methods which can be used to estimate parameters of the GPD.

Under the second order framework Beirlant et al. (2009) was able to refine the simple Pareto approximation thereby developing the extended Pareto distribution (EPD) which is also a limiting distribution of the relative excesses over a threshold. Estimating parameters of the EPD results in an asymptotically unbiased estimator of the EVI. In this thesis we give special attention to the EPD.

The main problem is that in reality we do not always have the luxury of modeling completely observable and uncontaminated data. The data can sometimes be censored, contaminated, truncated or simply just missing. In such cases, we need to either develop new estimators or alter currently existing ones to withstand the effects of such settings.

1.2 Objectives and contribution of the study

The Objective of this study is to firstly show through a simulation study, that the EPD is a better POT model to consider than the GPD in terms of the bias, mean squared error and variance.

Furthermore we investigate different estimation methods for estimating parameters of the EPD and assess the behavior of these estimators through a number of simulation studies in terms of the bias and mean squared error.

The simulation studies are set according to the following design:

1. Uncontaminated and non-censored data
2. Right censored and uncontaminated data
3. Non-censored and contaminated data

From these simulation studies, EVI estimates of the different estimators are drawn and used to further estimate extreme quantiles which will also be assessed at contaminated and uncontaminated distributions.

The main contributions of this thesis to EVT are as follows:

- An asymptotically unbiased EPD Bayesian estimator is proposed using a non-informative prior. Accuracy of this estimator can be heightened by incorporating expert opinion into the prior distribution. This can be very helpful, especially in EVT since extreme data is often scarce.
- We extend the censored EPD likelihood approach of Beirlant et al. (2016) for estimating the EVI for right censored data into a Bayesian paradigm. There exists very little literature on censoring EVT, moreover parametric Bayesian methods for censored data in EVT are almost non-existent.
- An exhaustive simulation study is conducted under various settings from which one can assess the behavior of estimators. This simulation study will aid practitioners in observing how the different methods of estimation perform when subjected to censorship and contamination.

1.3 Outline of study

In Chapter 2 a brief overview of EVT is given. The following literature is provided on historical developments in EVT in Section 2.2; Bayesian methods in EVT in Section 2.2.1; Tail estimation in EVT Section 2.2.2; Tail estimation under contamination in Section 2.2.3 and right censoring in Section 2.2.4. The extremal limit problem is discussed in Section 2.3. And a brief definition of the POT framework is given in Section 2.3.1.

In Chapter 3 the EPD is introduced and defined. Generalization of the EPD is shown along with some properties of the Pareto, the EPD and the GPD in Section 3.2. The ML estimation procedure is detailed in Section 3.3. In Section 3.4, further comparison is made between the EPD, the GPD and the Hill (1975) estimate, in terms of the mean square error, variance and bias by way of simulation. Some chapter concluding remarks are made in Figure 3.4 and Section 3.5.

In Chapter 4 an improved maximum likelihood estimation procedure is proposed and detailed in Section 4.2. In Section 4.3 a simulation study is conducted to compare this proposed estimation procedure to the previously applied one in Chapter 3. Some chapter concluding remarks are made in Figure 4.3.

In Chapter 5 the Bayesian EPD estimator is proposed. A short overview of Bayesian inference is given in Section 5.2. The estimation procedure is detailed in Section 5.3.3. An exhaustive Monte Carlo simulation study is conducted in Section 5.5 to assess the behavior of our estimate where the second order param-

eter is fixed (the first case) and where the second order parameter is estimated concurrently with other parameters (the second case). A case-study is conducted in Section 5.7 using a real dataset. Some chapter concluding remarks are made in Section 5.8.

In Chapter 6 performance of the EPD under random right censorship is investigated. Section 6.1.1 shows a closer look into some situations in practice in which random censoring can occur. A general treatment of random right censoring is given in Section 6.2. In Section 6.3 a simulation study is conducted to assess the behavior of the EVI estimators mentioned in Section 6.2. In addition to the Monte Carlo simulation conducted in Section 6.3, a case study is conducted in Section 6.4. In Section 6.5 we extend the EPD likelihood-adapted approach of Beirlant et al. (2016) into a Bayesian setting, and construct a censored posterior. In Section 6.6 an additional simulation study is conducted to assess the censored Bayesian estimator in comparison to the censored ML estimator. In Section 6.7 some chapter concluding remarks are made.

In Chapter 7 performance of the EPD is assessed when data is contaminated. The minimum density power divergence estimator (MDPDE) is also defined in this chapter. In Section 7.2 a short definition of the the MDPDE is given. In Section 7.2.1 the estimation procedures of the MDPDE are shown. In Section 7.3 an exhaustive simulation study is conducted, to assess the performance of our estimates and a brief case study is carried out in Section 7.4 to further assess the behavior of the MDPDE in comparison to other mentioned estimators when data is both contaminated and uncontaminated. Some chapter concluding remarks are made in Section 7.5

Finally an overall conclusion of this study is given in Chapter 8.

Chapter 2

Overview of Extreme Value Theory

2.1 Introduction

The aim of this chapter is to give a brief summary of EVT and discuss the basic result of the extremal limit problem. Section 2.3.1 extends the limit problem into modelling data conditional on some high threshold using the GPD. Section 2.4 covers some basic graphical tools used in EVT. In Section 2.6 we provide a short summary of some well known necessary and sufficient first and second order results from EVT which are used subsequently throughout this thesis.

The rest of this chapter is organized as follows. In Section 2.2 we discuss available literature on historical developments in EVT. We further provide some literature on Bayesian methods in EVT in Section 2.2.1. We provide further literature on the tail estimation in Section 2.2.2. In Section 2.2.3 we provide some literature on tail estimation under contamination and under right censoring in Section 2.2.4.

The extremal limit problem is discussed in Section 2.3. And a brief definition of the POT framework is given in Section 2.3.1. In Section 2.4 we discuss some well known graphical tools in EVT. This chapter ends with detailed explanations of first-order and second-order conditions in EVT and regular varying functions in Section 2.6.

2.2 Historic developments of classical theory of extremes

Extreme value theory is a branch of statistics mainly concerned with asymptotic distributions of the maximum (minimum) of independent, identically distributed random variables.

The first reference of extremes can be found in the Bible (Genesis): Methuselah, the man reported to have lived the longest at the age of 969. The Genesis flood narrative which makes up chapters 6–9 in the book of Genesis considers many great floods, including the preparation of Noah’s ark – the long rain, the large flood and the structural safety of the ark.

The field of EVT began with a paper by E.L Dodd in 1923. Dodd was able to relate the asymptotic growth of the maximum of a number of independent identically distributed random variables to the rate at which the right tail of the underlying probability density function f fell off to zero. Following this idea, in 1927 Fréchet introduced an idea that if there exists a limit law for maxima (or minima), then that law must be stable in the sense that the distribution F of the maximum (or minimum) of a number of independent identically distributed observations drawn from it must be of the same type, except for a linear transformation (Loynes, 1986).

In 1928 an English statistician by the name of Leonard Henry Caleb Tippett applied his statistical knowledge to the problem of yarn breakage rates in weaving in an effort to make cotton threads stronger (Stanton, 1987).

While at the Shirley Institute, Tippet realized that for a single cotton thread, the strength of its weakest fibers had more control over the overall strength of that thread. Tippet along with Sir Ronald Aylmer Fisher found the three asymptotic limits that described the distributions of extremes in their paper titled *On the estimation of the frequency distributions of the largest or smallest member of a sample*.

In 1936 von Mises established simple and sufficient conditions for the existence of a limit distribution of a sequence of partial maxima, linearly normalized, of an independent identically distributed sequence.

Gnedenko (1943) then developed the first mathematically rigorous treatment of the fundamental limit theorems of extremes, defined in Theorem 2.1 of Section 2.3. The roles played by E.L Dodd and M. Fréchet were instrumental in providing the conceptual framework needed to develop the asymptotic theory of extremes.

The evolution of regular variation also played a crucial role in EVT. Karamata (1930) studied Taubarian theorems and functions of regular variation. These functions were subsequently applied to the study of stationarity of distributions in probability and to asymptotic behavior of differential equations.

In his ground breaking doctoral thesis titled "On Regular Variation and its Application to the Weak Convergence of Sample extremes" (de Haan, 1970). Laurens de Haan’s work gave rise to a variant, or refinement, of Karamata’s regular variation and regular variation soon became the core primacy necessary for understanding extremes. Comprehensive literature on regular variation and probability can be found in Bingham et al. (1987).

2.2.1 Bayesian methods in extreme value theory

There exists very little literature that link Bayesian inference to EVT the framework. A common conceptual complication with Bayesian methods in EVT is the formulation of prior beliefs in an EVT context.

The Weibull distribution is a sub-family of the Generalized Extreme Value (GEV) distribution with a negative EVI ($\gamma < 0$). Early reference of Bayesian inference based on this 3-parameter Weibull class was made by Basu (1964), Holla (1966) and Bhattacharya (1967).

The first paper to illustrate the computational feasibility within extreme value models using prior distributions that were scientifically motivated was by Smith (1987). In this paper, Smith (1987) use the 3-parameter Weibull distribution and compared the maximum likelihood (ML) estimator with the Bayesian estimator using informative priors.

Sinha and Sloan (1988) proposed using Bayes linear estimates to approximate the posterior expectations for the Weibull distribution as well as the survival function \bar{F} . This method was however only able to yield the first and the second posterior moments instead of the full posterior distribution.

Engelund and Rackwitz (1992) considered a 2-parameter Gumbel distribution (also a sub-family of the GEV distribution) and compared the Bayesian estimators with the ML estimators for a simulated dataset. They reported that the Bayesian estimators were more efficient than those of the ML-based estimators.

Pickands III (1994) proposed a non-conjugate specification of the prior distribution for the GPD.

Coles and Tawn (1996) conducted a case study where they incorporated expert knowledge into prior distribution as a basis for Bayesian analysis of extreme rainfall.

For other references on Bayesian approaches in EVT, see for example, Coles and Powell (1996), Reiss and Thomas (1999), Coles et al. (2001), Reiss et al. (2001) and Beirlant et al. (2004)

2.2.2 Tail estimation

Let X_1, X_2, \dots, X_n be independent, identically distributed random variables, distributed according to some unknown distribution function F . A question of great interest is how to obtain a good estimator for the quantile

$$F^{\leftarrow}(1 - \epsilon) = \inf\{y : F(y) \geq 1 - \epsilon\}, \quad (2.1)$$

for ϵ very small, such that this quantile is situated at the border of, or beyond the range of the data. The study of such extreme quantities is directly linked to the

accurate estimation of the tail of the distribution function

$$\hat{F}(x) := 1 - F(x) = \mathbb{P}(X > x) \quad (2.2)$$

for some large threshold say x . In EVT the behavior of the tail distribution is known to be governed by the EVI, this parameter measures the tail heaviness of F and is thus of primordial importance. There exists a vast amount of literature devoted to the estimation of this parameter. In this thesis we mention for example, Hill (1975), Smith (1987), Dekkers et al. (1989), Drees (1996) and Beirlant et al. (1996).

2.2.3 Tail estimation under contamination

Model misspecification and contamination by outliers are common problems in statistics. Robust statistical analysis aims to construct methods that are able to withstand the effects of outliers. However by way of contrast, in EVT more attention is given to extreme observations whereas robust methods aim to reduce the influence that outlying observations have on statistical quantities. It is thus uncommon for one to lay claim of some conceptual contradiction. However robust methods for extreme values have already been considered in literature e.g. Brazauskas and Serfling (2000) considered robust estimation in the context of a strict Pareto distribution. Dupuis (1998), Peng and Welsh (2001) and Juárez and Schucany (2004) considered the problem of making inferences about extreme values when the underlying distribution is a GEV distribution, a GPD and light or heavy-tailed, respectively. Dell'Aquila and Embrechts (2006) noted that “Robust methods do not down-weight ‘extreme’ observations if they conform to the majority of the data”.

Basu et al. (1998) introduced the minimum density power divergence estimator (MDPDE) as a parametric estimator that balances the trade-off between efficiency and robustness. Kim and Lee (2008) obtained a robust estimator for $\gamma > 0$ by fitting the strict Pareto distribution to the top order statistics using the MDPDE criterion, however the estimators were not asymptotically unbiased. Vandewalle et al. (2007) fitted a partial density component involving a mixture of two Pareto distributions by minimizing a L_2 distance. Dierckx et al. (2013) introduced a robust and asymptotically unbiased estimator for γ by fitting the EPD to a sample of relative excesses by the MDPDE criterion. Goegebeur et al. (2014) then further derived a robust and asymptotically unbiased extreme quantile estimator. Other contributions of robust methods in EVT can be found in Dupuis and Victoria-Feser (2006) and Hubert et al. (2013) amongst others.

2.2.4 Tail estimation theory under censoring

In extreme value analysis, very little work has been done in estimation of the tail distribution under censorship. Extreme value analysis for right censored Pareto-

type data was first considered by Beirlant et al. (2007) followed by Einmahl et al. (2008) and later Gomes and Neves (2011). Beirlant et al. (2007) used a generalization of the POT methodology using the GPD, and also proposed an adaptation of the moment estimator. Einmahl et al. (2008) later proved in detail, asymptotic normality results for various estimators of the EVI under random censorship. Worms and Worms (2014) presented some new approaches for estimating the EVI in the framework of random right censorship, based on the ideas of Kaplan-Meier integration and the synthetic data approach of Leugrands (1987). Beirlant et al. (2016) showed that the bias of the estimators of the right tail EVI can be substantial and proposed the adapted EPD likelihood approach by showing that it leads to much improved results, both in bias and MSE.

2.3 Extremal limit problem

Suppose the sample X_1, X_2, \dots, X_n is independent and identically distributed random variables from a distribution F . Let $X_{n,n} = \max\{X_1, \dots, X_n\}$ denote the maximum of the sample and let $S_n := X_1 + X_2 + \dots + X_n$ be the sum. Generally the classical central limit problem tries to find constants $a_n \geq 0$ and b_n such that

$$Y_n := a_n^{-1}(S_n - b_n) \tag{2.3}$$

tends in distribution to a non-degenerate distribution. The limit can then be used to approximate the distribution of the quantity Y_n (Beirlant et al., 2004). Typically for any distribution F the Normal distribution is attained as the limit of the S_n unless the F has heavy tails, in which case a stable distribution will appear as the limit. Particularly Pareto-type distributions which have an infinite variance and thereby generating non-normal/extreme limits that produce asymptotic behavior different from normal behavior.

In EVT the main concern is finding the possible limit distributions of the maximum $X_{n,n}$. That is, the search for non-degenerate distributions of $X \in \mathbb{R}$ for which there exist a sequence of numbers b_n and a_n where $n \geq 1$ such that

$$P\left(\frac{X_{n,n} - b_n}{a_n} \leq x\right) \rightarrow G(x), \quad n \rightarrow \infty. \tag{2.4}$$

The term non-degenerate implies that the distribution does not have all its mass centered at one point. The limit problem was first solved by Gnedenko (1943), Fisher and Tippett (1928) in what is known as the first Extreme value theorem and later revived by de Haan (1970).

Theorem 2.1. Fisher-Tippett-Gnedenko

Let X_1, X_2, \dots, X_n be a sequence of independent, identically distributed random variables from common distribution function F and denote

$$X_{n,n} = \max\{X_1, X_2, \dots, X_n\}$$

the sample maximum. If condition 2.4 holds for a non-degenerate distribution G , then G belongs in the GEV family of distributions and is of the form

$$G(x|\mu, \sigma, \gamma) = \begin{cases} \exp\left(-\left(1 + \gamma\frac{x - \mu}{\sigma}\right)^{-1/\gamma}\right), & 1 + \gamma\frac{x - \mu}{\sigma} > 0, \quad \gamma \neq 0, \\ \exp(-\exp(-\frac{x - \mu}{\sigma})), & x \in \mathbb{R}, \quad \gamma = 0 \end{cases} \quad (2.5)$$

where $\mu \in \mathbb{R}$ is the location parameter, $\sigma > 0$ the scale parameter and γ is the tail index also known as the EVI. \blacktriangle

Generalized extreme value distributions can be characterized by the following families conditional on γ .

Fréchet: $\gamma > 0$

$$G_1(x|\gamma) = \exp(-x^{-1/\gamma}) \quad (2.6)$$

Extreme Weibull: $\gamma < 0$

$$G_1(x|\gamma) = \exp(-|x|^{-1/\gamma}) \quad (2.7)$$

Gumbel: $\gamma = 0$

$$G_1(x|\gamma) = \exp(-\exp(-x)) \quad (2.8)$$

The GEV is most commonly applied to the method of Block maxima which is the traditional modelling approach in EVT. This method is implemented by grouping the data into equal blocks of equal length. Thereafter we make an assumption that the maximum in each block follows the asymptotic distribution $G(x)$ (De Zea Bermudez and Kotz, 2010). The sampled data for this approach is shown graphically in Figure 2.1a. A more recent alternative approach, known as the POT method has been developed by hydrologists over the last 46 years; to model exceedances over a high threshold (Dupuis, 1998).

The POT methodology offers a unifying approach to the modelling of heavy tailed distributions (McNeil and Saladin, 1997). Here, a sufficiently high threshold t is chosen, and the observations above this threshold are then modelled by a GPD as defined in Section 2.3.1. This approach is shown graphically in Figure 2.1b.

2.3.1 The generalized Pareto distribution

Within the POT framework, we consider the distribution F of the random variable X for the k^{th} largest order statistics. We are only interested in estimating the

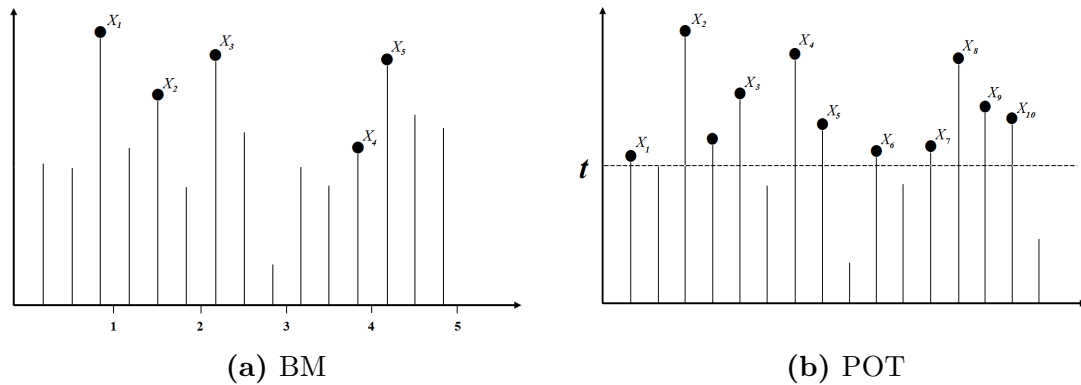


Figure 2.1: Depiction of the method of Block-maxima (left) and Peaks over a threshold t (right)

conditional distribution F_t of the variable X above a high threshold t . Which brings us to what is known as the second theorem in EVT describing the tail distribution of the random variable X .

Theorem 2.2. Pickands–Balkema–de Haan(Balkema and De Haan, 1974; de Haan, 1984; Pickands, 1975) Let X be a random variable with distribution function F . Then, for a large enough threshold t , the conditional excess distribution function of $Z = X - t$, given by

$$F_t(z) = P(X - t \leq z | X > t) = \frac{F(z + t) - F(t)}{1 - F(t)} \quad (2.9)$$

for $0 \leq z \leq x_F - t$, where x_F is the right endpoint of the underlying distribution F . For a large enough threshold t , $F_t(z)$ can be approximated by

$$G_{\gamma, \sigma}(z) = \begin{cases} 1 - \left(1 + \frac{\gamma z}{\sigma}\right)^{-\frac{1}{\gamma}}, & \text{if } \gamma \neq 0, \\ 1 - \exp(-z/\sigma), & \text{if } \gamma = 0 \end{cases} \quad (2.10)$$

where $\sigma > 0$, and $z \geq 0$ when $\gamma \geq 0$ and $0 \leq z \leq -\sigma/\gamma$ when $k < 0$. ▲

that is if and only if $G_{\gamma, \sigma}$ is in the maximum domain of attraction of the Extreme value distribution in Equation 2.5. This is often referred to as the domain of attraction condition. Equation 2.10 is called the Generalized Pareto distribution (GPD). In this thesis, we refer to the GPD under the restriction $\gamma > 0$.

2.4 Graphical tools

When presented with extreme data, a practitioner needs to first visualize and analyze the data in order to know which extreme value analysis (EVA) tools to

apply. We might have questions such as: What is the distribution of the data on its full support? Is the data heavy-tailed? (usually in comparison to the exponential) or is there reasonable plausibility for fitting some proposed model? The two most commonly used graphical tools in EVT are Q-Q plots (short for Quantile-Quantile plots) and the mean excess plots (also known as mean residual life plots).

Quantile-Quantile plots

A Q-Q plot is a probability plot used to compare two probability distributions by plotting their quantiles against each other to see if the two probability distributions are linearly related to each other. This linear relationship can be seen captured by eye. If the two distributions being compared seem to be linear, the points on the plot can be approximated by some line $y = x$. We mainly use Q-Q plots to compare the theoretical distributions with the empirical distribution of the data. We can further quantify the agreement of a fitted distribution with the observed data by means of a correlation coefficient. Since we are comparing the theoretical distribution and the empirical distribution of the data, we use plotting positions to form the Q-Q plots (see the following definition by Berning (2010)).

Definition 2.3. Quantile-Quantile Plot

A plot of $(Q(i/(n+1)); x_{i,n}), i = 1, 2, \dots, n$ should be approximately linear if x_1, x_2, \dots, x_n are from a distribution with quantile function Q . ▲

The following is a table of Q-Q plot coordinates for some distributions as given in Beirlant et al. (2004).

Table 2.1: Q-Q plot coordinates for some distributions.

Distribution	$F(x)$	Coordinates
Normal	$\int_{-\infty}^x \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) du$ $x \in \mathbb{R}; \mu \in \mathbb{R}, \sigma > 0$	$\Phi^{-1}(p_{i,n}, x_{i,n})$
Log-normal	$\int_0^x \frac{1}{\sqrt{2\pi}\sigma\mu} \exp\left(-\frac{(\log t - \mu)^2}{2\sigma^2}\right) du$ $x > 0; \mu \in \mathbb{R}, \sigma > 0$	$\Phi^{-1}(p_{i,n}, \log x_{i,n})$
Exponential	$1 - \exp(-\lambda x)$ $x > 0; \lambda > 0$	$(-\log(1 - p_{i,n}), x_{i,n})$
Pareto	$1 - x^{-\alpha}$ $x > 1; \alpha > 0$	$(-\log(1 - p_{i,n}), \log x_{i,n})$
Weibull	$1 - \exp(-\lambda x^\tau)$ $x > 0; \lambda, \tau > 0$	$(\log(-\log(1 - p_{i,n})), \log x_{i,n})$

Mean excess plots

One of the uses of the mean excess plot in EVT is to aid in making a choice of an extreme threshold t , for which the GPD offers a plausible approximation of the excess distribution F_t as specified by Equation 2.9. The mean excesses (e) function of a random variable X is defined as:

$$e(t) = E(X - t | X > t) \quad (2.11)$$

provided that $E(X) < \infty$. Estimates of the mean excesses for sample x_1, \dots, x_n are given by

$$e_{k,n} := \hat{e}_n(x_{n-k,n}) = \frac{1}{k} \sum_{j=1}^k x_{n-j+1,n} - x_{n-k,n} \quad (2.12)$$

where the empirical function \hat{e}_n is plotted at values $t = x_{n-k,n}$, $k = 1, \dots, n-1$, the $(k+1)$ —largest observation (Beirlant et al., 2004).

2.5 Areas of Application

From its inception, the area of EVT has been applied to a wide variety of fields. In this section we make mention of some common applications in EVT. We refer to Gumbel (2004), de Haan (1970), Galambos and Kotz (1978) and Beirlant et al. (2004) for a more elaborate body of literature devoted to the theory and applications of extremes.

Hydrology: EVT is very useful in flood frequency analysis. Practitioners may be interested in estimating the T -year flood discharge, which is the water level exceeded every T years on average. The parameter rainfall intensity is also of some interest in hydrology. This quantity is an essential component when modelling water course systems, urban drainage and water runoff.

Environmental research and meteorology: Meteorological data, such as ozone concentration, rainfall and wind speeds, all of which have tremendous impact on society and could result in negative consequences.

Finance applications: Risk management at a commercial bank is intended to guard against risk or loss due to fall in prices of financial assets held or issued by the bank. EVT is used to calculate the maximum (extreme) losses that can occur in any given time period. A very important quantity in finance is the high quantile estimate otherwise known as value at risk (VAR), which is the level below which the future portfolio can drop with a small specified probability.

Insurance applications: Non-life insurance is one of the most prominent areas of applications of EVT. Portfolios seem to have a tendency to include adversely large claims that could jeopardize the solvency of a portfolio and possibly a substantial

portion of the company. Another growing area for which EVT is exceedingly becoming useful is re-insurance.

Geological and seismic analysis: EVT can be used in unveiling some useful tail characteristics of data such as magnitudes of and losses resulting from earthquakes, and in diamond sizes and values. A case study is conducted in Chapter 7 using such data.

Metallurgy: In Metallurgy EVT can be used to model metal fatigue, which causes failure of a metallic element under extreme stress.

2.6 Second Order Conditions and Regular Variation

The study of regular variation of first and second order in EVT is motivated by the need to provide regularity conditions on the behavior of the tail of distribution functions. These conditions can be very technical and somewhat difficult to comprehend from an intuitive standpoint.

In EVT we are often only interested in the tail of a distribution thereby negating the need to make assumptions about the entire distribution. Focus then shifts to making assumptions about the tail of the distribution. First order regular variation is crucial to extreme value statistics in that it allows us to make these assumptions about the tail of a distribution function. More regularity, such as second order regular variation becomes necessary when we are interested in proving distributional results of estimators.

We first make the following preliminary setting. From Theorem 2.1, take G up to scale and location, then the Extreme Value distribution function, dependent on a shape parameter $\gamma \in \mathbb{R}$ becomes

$$G_\gamma(x) := \begin{cases} \exp(-(1 + \gamma x)^{-1/\gamma}), & 1 + \gamma x > 0 & \text{if } \gamma \neq 0 \\ \exp(-\exp(-x)), & x \in \mathbb{R} & \text{if } \gamma = 0 \end{cases} \quad (2.13)$$

We then say that F is in the domain of attraction for maxima of the distribution function G_γ in Equation 2.13 and write $F \in \mathcal{D}_M(G_\gamma)$.

Definition 2.4. Second order regular variation

A class of regular varying functions constitutes all functions ultimately zero, satisfying:

$$\lim_{t \rightarrow \infty} \frac{f(tx)}{f(t)} = x^\rho, \quad \text{for } x > 0, \text{ with some } \rho \in \mathbb{R} \quad (2.14)$$

We say, f is regular varying with index ρ and write $f \in RV_\rho$. Furthermore a function f is said to be slowly varying if for some $x > 0$:

$$\lim_{t \rightarrow \infty} \frac{f(tx)}{f(t)} = 1 \quad (2.15)$$

▲

Assume that for a measurable real-valued function f on $(0, \infty)$, there exists a positive auxiliary function $a(t)$ and some $\gamma \in \mathbb{R}$ such that

$$\phi(x) = \lim_{t \rightarrow \infty} (f(tx) - f(t))/a(t) = \int_1^x s^{\gamma-1} ds, \quad x > 0. \quad (2.16)$$

further assume that

$$\lim_{t \rightarrow \infty} (f(tx) - f(t) - a(t)\phi(x))/a(t) \quad (2.17)$$

exists non-trivially with some second auxiliary function $a_1(t)$. f is then said to be of generalized regular variation (see de Haan and Stadtmüller, 1996).

In EVT, by taking f to be a tail quantile function U and assuming the existence of some positive function $a(\cdot)$, we concern ourselves with functions $U(\cdot)$ for which:

$$\lim_{t \rightarrow \infty} \frac{U(tx) - U(t)}{a(t)} \text{ exists } \forall x > 0. \quad (2.18)$$

This leads us to the following extended regular variation property of de Haan (1984).

First-Order Condition

A general tail ($\gamma \in \mathbb{R}$)

The following extended regular variation property (de Haan, 1984), denoted ERV_γ , is a well-known necessary and sufficient condition for $F \in \mathcal{D}_{\mathcal{M}}(G_\gamma)$:

$$\lim_{t \rightarrow \infty} \frac{U(tx) - U(t)}{a(t)} = \int_1^x s^{\gamma-1} ds = \begin{cases} \frac{x^\gamma - 1}{\gamma} & \text{if } \gamma \neq 0 \\ \ln x & \text{if } \gamma = 0, \end{cases} \quad (2.19)$$

for all $x > 0$ and some positive measurable function $a > 0$. Then for heavy tails ($\gamma > 0$), for any real ρ where $\rho = -1/\gamma$,

$$F \in \mathcal{D}_{\mathcal{M}}(G_\gamma) \iff \bar{F} = 1 - F \in RV_{-1/\gamma} \quad (2.20)$$

Equivalently, and with U , the tail quantile associated with F , defined as:

$$U(t) := \left(\frac{1}{1-F} \right)^{\leftarrow} (t) = \inf \left\{ x : F(x) \geq 1 - \frac{1}{t} \right\} \quad (2.21)$$

For a much broader class of distributions with $\gamma > 0$ known as the Pareto-type class, we can choose

$$a(t) = \gamma x^\gamma \ell_U(t) = \gamma U(t) \quad (2.22)$$

where ℓ_U is a slowly varying function. A distribution function F is said to be in the domain of attraction of an extreme value distribution if and only if Equation 2.18 holds for the specified inverse function $(1/(1-F(\cdot)))$.

Second-Order Condition

A general tail ($\gamma \in \mathbb{R}$)

The second order condition specifies the rate of convergence in Equation 2.19, we strictly assume the existence of a function $A(t)$, possibly not changing in sign but tending to zero as $t \rightarrow \infty$, such that:

$$\lim_{t \rightarrow \infty} \frac{\frac{U(tx) - U(t)}{a_0(t)} - \frac{x^\gamma - 1}{\gamma}}{A(t)} = H_{\gamma, \rho}(x) := \frac{1}{\rho} \left(\frac{x^{\gamma+\rho} - 1}{\gamma + \rho} - \frac{x^\gamma - 1}{\gamma} \right) \quad (2.23)$$

for all $x > 0$ and $a_0 > 0$, with ρ , a non-positive second order parameter controlling the speed of convergence of the partial maxima towards the limit law in Equation 2.13. We say the function U is of second order regular variation, and we use the notation $U \in 2ERV(\gamma, \rho)$ noting also that, $|A| \in RV_\rho$ and tends to zero, hence larger values of ρ correspond to a higher rate of convergence. Written more specifically by Alves et al. (2007)

$$H_{\gamma, \rho}(x) = \begin{cases} \frac{1}{\rho} \left(\frac{x^\rho - 1}{\rho} - \ln x \right) & \text{if } \gamma = 0, \rho \neq 0 \\ \frac{1}{\gamma} \left(x^\gamma \ln x - \frac{x^\gamma - 1}{\gamma} \right) & \text{if } \gamma \neq 0, \rho = 0 \\ \frac{\ln^2 x}{2} & \text{if } \gamma = \rho = 0 \end{cases} \quad (2.24)$$

For heavy tails ($\gamma > 0$), in order to measure the rate of convergence in Equation 2.20, Alves et al. (2007) considered:

$$\begin{aligned} \lim_{t \rightarrow \infty} \frac{\frac{U(tx)}{U(t)} - x^\gamma}{\tilde{A}(t)} &= x^\gamma \frac{x^{\tilde{\rho}} - 1}{\tilde{\rho}} \iff \\ \lim_{t \rightarrow \infty} \frac{\ln U(tx) - \ln U(t) - \gamma \ln x}{\tilde{A}(t)} &= \frac{x^{\tilde{\rho}} - 1}{\tilde{\rho}} \end{aligned} \quad (2.25)$$

for all $x > 0$, where $\tilde{\rho} \leq 0$ is a second order parameter controlling the speed of convergence of maximum values, linearly normalized towards the limit law in Equation 2.13. We then say that the function U is of second order regular variation and write $U \in 2RV_{\gamma, \tilde{\rho}}$.

2.6.1 Fréchet-Pareto class

The context by which Pareto-type tails are defined subsets all models with tails heavier than exponential for which $\gamma > 0$. Such models can be defined for \bar{F} as:

$$U(x) = x^\gamma \ell_U(x), \quad x \uparrow \infty \quad (2.26)$$

for $\gamma = 1/\alpha > 0$, with the tail quantile function: $U(x) = Q(1 - 1/x)$ where Q is the quantile function. The up arrow notation $x \uparrow \infty$ is used to show fast convergence of x to infinity. The arithmetic operation symbol $\{\uparrow\}$ represents an exponential sequence which grows faster than addition and multiplication operations. It is shown in Beirlant et al. (2004) that the condition Equation 2.18 for $\gamma > 0$ is equivalent to:

$$\frac{1 - F(ut)}{1 - F(t)} = Pr \left(\frac{X}{t} > u | X > t \right) \rightarrow u^{-1/\gamma} \quad \text{as } t \rightarrow \infty. \quad (2.27)$$

This is essentially saying that $x^{1/\gamma}(1 - F(x))$ is slowly varying, and therefore, there exists a slowly varying function $\ell_F(x)$ such that $1 - F(x) = x^{-1/\gamma} \ell_F(x)$. Pareto-type distributions can therefore be formulated in terms of the distribution F as well as in terms of the tail quantile function. i.e.

$$1 - F(x) = x^{-1/\gamma} \ell_F(x) \quad \text{and} \quad U(x) = x^\gamma \ell_U(x), \quad x \uparrow \infty \quad (2.28)$$

where the two slowly varying functions ℓ_U and ℓ_F are linked through the Bruyn conjugation. Next we define a special class of Pareto type distributions, called the Hall class. This is a wide class of models containing most useful heavy-tailed parents such as those used in this thesis, namely: the Fréchet, the Burr XII and the Student-t.

Definition 2.5. Hall class of distributions

Assuming the underlying distribution F satisfies $F(0) = 0$, the survival function of a distribution in the Hall class (Hall, 1982; Hall and Welsh, 1985) can be written as:

$$\bar{F}(t) = C^{1/\gamma} t^{-1/\gamma} [1 + \gamma^{-1} D C^{\rho/\gamma} t^{-\rho/\gamma} + o(1)], \quad t \uparrow \infty, \quad (2.29)$$

where $\gamma > 0, C > 0, \rho > 0$ and $D \in \mathbb{R}$. ▲

For any $\tau = \rho/\gamma$ and $C^{1/\gamma} \propto C$, we can write Equation 2.29 as:

$$1 - F(x) = C x^{-\alpha} [1 + D x^\tau + o(x^\tau)], \quad x \uparrow \infty, \quad (2.30)$$

where $\alpha > 0, C > 0, \tau < 0$ and $D \in \mathbb{R}$.

Using the tail quantile function, Equation 2.30 can also be written as:

$$U(t) = Ct^\gamma \left(1 + \frac{A(t)}{\rho} + o(t^\rho) \right), \quad A(t) = \gamma\beta t^\rho \quad (2.31)$$

Table 2.2 shows a list of the Pareto type distributions, together with their corresponding EVI and slowly varying functions as seen in Beirlant et al. (2004).

Table 2.2: List of Pareto-type distributions.

Distribution	$1 - F(x)$	Extreme value index (EVI)	$\ell_F(x)$
Pareto (α)	$x^{-\alpha}$, $x > 1; \alpha > 0$	$\frac{1}{\alpha}$	1
GP (γ, σ)	$\left(1 + \frac{\gamma x}{\sigma}\right)^{-\frac{1}{\gamma}}$, $x > 0; \sigma, \gamma > 0$	γ	$\left(\frac{\sigma}{\gamma}\right) \left(1 + \frac{\sigma}{\gamma x}\right)^{-\frac{1}{\gamma}}$
Burr (XII) (η, τ, λ) (type(XII))	$\left(\frac{\eta}{\eta + x^\tau}\right)^\lambda$, $x > 0; \eta, \tau, \lambda > 0$	$\frac{1}{\lambda\tau}$	$\left(\frac{\eta}{\eta + x^\tau}\right)^\lambda$
Burr (XII) (η, τ, λ) (type(III))	$1 - \left(\frac{\eta}{\eta + x^{-\tau}}\right)^\lambda$, $x > 0; \eta, \tau, \lambda > 0$	$\frac{1}{\tau}$	$\frac{\lambda}{\eta} \left(1 - \frac{1}{2} \frac{\lambda + 1}{\eta} x^{-\tau} + o(x^{-\tau})\right)$
$F(m, n)$	$\int_x^\infty \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \left(\frac{m}{n}\right)^{m/2} w^{m/2-1} \left(1 + \frac{m}{n}w\right)^{-(m+n)/2} dw$, $x > 0; m, n > 0$	$\frac{2}{n}$	$\frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2}+1)} \left(\frac{m}{n}\right)^{m/2} \left(\frac{m}{n} + \frac{1}{x}\right)^{-(m+n)/2} (1 + o(1))$
$\text{inv}\Gamma(\lambda, \alpha)$	$\int_x^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} \exp(-\lambda/w) w^{-\alpha-1} dw$, $x > 0; \lambda, \alpha > 0$	$\frac{1}{\alpha}$	$\frac{\lambda^\alpha}{\Gamma(\alpha+1)} \exp(-\lambda/x)(1 + o(1))$
$\log\Gamma(\lambda, \alpha)$	$\int_x^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} w^{-\lambda-1} (\log w)^{\alpha-1} dw$, $x > 1; \lambda, \alpha > 0$	$\frac{1}{\lambda}$	$\frac{\lambda^{\alpha-1}}{\Gamma(\alpha)} (\log x)^{\alpha-1} \left(1 + \frac{\alpha-1}{\lambda} \frac{1}{\log x} + o\left(\frac{1}{\log x}\right)\right)$
Fréchet (α)	$1 - \exp(-x^{-\alpha})$, $x > 0; \alpha > 0$	$\frac{1}{\alpha}$	$1 - \frac{x^{-\alpha}}{2} + o(x^{-\alpha})$
$ T_n $	$\int_x^\infty \frac{2\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \left(1 + \frac{w^2}{n}\right)^{-\frac{n+1}{2}} dw$, $x > 0; n > 0$	$\frac{1}{n}$	$\frac{n-1}{2} \frac{2\Gamma(\frac{n+1}{2})n}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \left(1 - \frac{n^2(n+1)}{2(n+2)} x^{-2} + o(x^{-2})\right)$

Chapter 3

Extended Pareto Distribution

3.1 Introduction

The POT approach entails modeling clusters of excesses above a threshold with a Poisson process and fitting the GPD to the excesses above some high enough threshold. The point process characterization of high-level excesses makes the GPD a natural selection in fitting excesses over a sufficiently high threshold, in fact Pickands (1975) showed that the GPD arises as a limiting distribution for excesses over a high threshold if and only if the parent distribution is in the domain of attraction of the extreme value distributions.

Another motivation of the GPD arising as a limiting distribution in Peaks over Threshold (POT) is the 'threshold stability' property, as given by Davison and Smith (1990): if Y is a GPD and $t > 0$, then the conditional distribution of $Y - t$ given $Y > t$ is also a GPD. In the case of heavy tailed distributions, when fitting a Pareto distribution (PD) to excesses over a high positive thresholds, this threshold stability is sometimes not visible as a result of the slow rate of convergence in the Pickands–Balkema–de Haan theorem.

Beirlant et al. (2009) proposed an extension of the simple PD called the Extended Pareto distribution (EPD), an asymptotically unbiased estimator of the EVI.

We fit the EPD to relative excesses X_i/t where $i = 1, \dots, n$ conditional on $X_i > t$. The choice of the threshold has received much attention in literature however the best method is still to be found. Choosing a data-adaptive threshold $t = t_n = X_{n-k:n}$ within the ordered sequence $X_{1,n} \leq X_{2,n} \dots \leq X_{n,n}$ removes the burden of having to chose an optimum choice of k . One can therefore plot estimates of the EVI against k , and choose the EVI to be the point where the estimator is more horizontal. This becomes a problem in the presence of a large bias and leads to poor coverage probabilities of confidence intervals, which is usually the case with the Hill estimate, as will be seen in this chapter.

The rest of this chapter is organized as follows: In Section 3.2, Generalization of the EPD is shown along with some properties of the Pareto, the EPD and the GPD, to see whether the EPD can improve approximations of absolute and relative excess distributions with some order of magnitude.

The ML estimation procedure will be detailed in Section 3.3 as given by Beirlant et al. (2009). In Section 3.4, further comparison is made between the EPD, the GPD and the Hill, in terms of the mean square error, variance and bias by way of simulation, to see whether the EPD results in biased-reduced estimates of the EVI. Some concluding remarks are made in Figure 3.4 and Section 3.5.

3.2 Generalization of the EPD

Suppose we have a sequence of independent and identically distributed observations X_1, X_2, \dots, X_n , from an unknown distribution function F . By definition of the Peaks-Over-Threshold method, we are interested in the distribution function of excesses over a high threshold t :

$$Pr\left(\frac{X}{t} > y | X > t\right) \Rightarrow Pr(X > ty | X > t), \quad y > 1. \quad (3.1)$$

Note, for convenience, we mainly make use of relative excesses X/t instead of absolute excesses $X - t$.

Now let x_0 be the finite or infinite right endpoint of the distribution F . Then the distribution function of the excesses in Equation 3.1 can be written as:

$$\frac{Pr(X > ty \cap X > t)}{Pr(X > t)} = \frac{Pr(X > ty)}{Pr(X > t)} = \frac{\bar{F}(ty)}{\bar{F}(t)} \quad (3.2)$$

Using Definition 2.5, the deviation between the true excess distribution and the asymptotic model for heavy-tailed distributions, can be parametrized using a power series expansion, and thus Equation 3.2 can be written as:

$$\frac{\bar{F}(ty)}{\bar{F}(t)} = \frac{C(ty)^{-\alpha}(1 + Dt^\tau y^\tau + \dots)}{C(t)^{-\alpha}(1 + Dt^\tau + \dots)}, \quad y \uparrow \infty \quad (3.3)$$

$$= y^{-\alpha} \frac{1 + Dt^\tau y^\tau + \dots}{1 + Dt^\tau + \dots}, \quad y \uparrow \infty \quad (3.4)$$

If $y \uparrow \infty$, then $Dy^\tau \rightarrow 0$ and therefore the Maclaurin series $(1 + Dt^\tau + o(t^\tau))^{-1}$ will converge to the geometric series:

$$(1 - Dt^\tau + \dots) \quad \text{where} \quad |Dt^\tau| < 1. \quad (3.5)$$

We can therefore write equation Equation 3.4 as:

$$\frac{\bar{F}(ty)}{\bar{F}(t)} = y^{-\alpha}(1 + Dt^\tau y^\tau + \dots)(1 - Dt^\tau + \dots) \text{ as } y \uparrow \infty \quad (3.6)$$

$$= y^{-1/\gamma}(1 + Dt^\tau y^\tau - Dt^\tau + \dots) \quad (3.7)$$

By letting $D\gamma t^\tau = \delta_t = \delta$, Equation 3.7 can be written as:

$$\frac{\bar{F}(ty)}{\bar{F}(t)} \stackrel{y \uparrow \infty}{\sim} y^{-1/\gamma} \left(1 + \frac{1}{\gamma} \delta y^\tau - \frac{1}{\gamma} \delta + \dots \right) \quad (3.8)$$

Using the same power series expansion property as in Equation 3.5 it can be shown that the second term in parenthesis, i.e. $\left(1 + \frac{1}{\gamma} \delta y^\tau - \frac{1}{\gamma} \delta + \dots \right)$ converges from the Maclaurin series:

$$\sim (1 + \delta - \delta y^\tau)^{-1/\gamma}.$$

Therefore:

$$\frac{\bar{F}(ty)}{\bar{F}(t)} \stackrel{t \uparrow \infty}{\sim} \bar{G}_{\gamma, \delta, \tau}(y) = \{y(1 + \delta - \delta y^\tau)\}^{-1/\gamma}, \quad y > 1 \quad (3.9)$$

which defines the EPD with parameter vector (γ, δ, τ) in the range $\tau < 0 < \gamma$ and $\delta > \max(-1, 1/\tau)$. The EGPD is then given by the distribution function:

$$H_{\gamma, \delta, \tau}(x) = G_{\gamma, \delta, \tau}(1 + x), \quad x \in \mathbb{R}. \quad (3.10)$$

Note, $H_{\gamma, \delta, \tau}(x)$ is just a reparametrization of the GPD if $\tau = -1$.

Condition (\mathcal{R}). (Beirlant et al., 2009) Let $\gamma > 0$ and $\tau < 0$ be constants. A distribution function F is said to belong to the class $\mathcal{F}(\gamma, \tau)$ if $x^{1/\gamma} \bar{F}(x) \rightarrow C \in (0, \infty)$ as $x \rightarrow \infty$ and if the function δ defined via:

$$\bar{F}(x) = Cx^{-\frac{1}{\gamma}} \{1 + \gamma^{-1} \delta(x)\} \quad (3.11)$$

is eventually nonzero and of constant sign and such that $|\delta| \in \mathcal{R}_\tau$. ▲

The following proposition shows that for $F \in \mathcal{F}(\gamma, \tau)$, the EPD and E(G)PD improve the PD and GPD approximations to the excess distribution with an order of magnitude, for the conditional distributions of relative and absolute excesses of X over t :

$$Pr(X/t > y | X > t) = \frac{\bar{F}(ty)}{\bar{F}(t)} \quad \text{and} \quad Pr(X - t > x | X > t) = \frac{\bar{F}(t+x)}{\bar{F}(t)}$$

Proposition 1. (Beirlant et al., 2009) If $F \in \mathcal{F}(\gamma, \tau)$, then as $t \rightarrow \infty$,

$$\sup_{y \geq 1} \left| \frac{\bar{F}(ty)}{\bar{F}(t)} - \bar{G}_{\gamma, \delta(t), \tau}(y) \right| = o\{|\delta(t)|\} \quad (3.12)$$

$$\sup_{x \geq 0} \left| \frac{\bar{F}(t+x)}{\bar{F}(t)} - \bar{H}_{\gamma, \delta(t), \tau}(x/t) \right| = o\{|\delta(t)|\} \quad (3.13)$$

▲

Proof. Given the conditional distribution of the relative and absolute excesses of X over t , we have:

$$\frac{\bar{F}(ty)}{\bar{F}(t)} = y^{-1/\gamma} \frac{1 + \gamma^{-1}\delta(ty)}{1 + \gamma^{-1}\delta(t)} = y^{-1/\gamma} \left(1 - \gamma^{-1}\delta(t) \frac{1 - \frac{\delta(ty)}{\delta(t)}}{1 + \gamma^{-1}\delta(t)} \right)$$

on the other hand, since $0 \leq 1 - y^\tau \leq 1$ for $y \geq 1$ and since $\delta(t) \rightarrow 0$,

$$[y\{1 + \delta(t) - \delta(t)y^\tau\}]^{-1/\gamma} = y^{-1/\gamma} \{1 - \gamma^{-1}\delta(t)(1 - y^\tau)\} + o(\{|\delta(t)|\}), \quad t \rightarrow \infty,$$

Uniformly in $y \geq 1$. It thus follows that:

$$\begin{aligned} \frac{\bar{F}(ty)}{\bar{F}(t)} - [y\{1 + \delta(t) - \delta(t)y^\tau\}]^{-1/\gamma} = \\ -\gamma^{-1}y^{-1/\gamma}\delta(t) \left(\frac{1 - \frac{\delta(ty)}{\delta(t)}}{1 + \gamma^{-1}\delta(t)} - (1 - y^\tau) \right) + o(\{|\delta(t)|\}), \quad t \rightarrow \infty. \end{aligned}$$

Now using the fact that $|\delta| \in \mathcal{R}_\tau$, the asymptotic relation Equation 3.12 then follows from the uniform convergence theorem for regularly varying functions with negative index (Bingham et.al., 1987, Theorem 1.5.2). \blacktriangle

If in Equation 3.12 and Equation 3.13, the EPD tail function $\bar{G}_{\gamma,\delta(t),\tau}(y)$ was replaced by the PD tail function $(y^{-1/\gamma})$, and the EGPD tail function $\bar{H}_{\gamma,\delta(t),\tau}(x/t)$ was replaced by the GPD tail function $(1 + \gamma x/\sigma)^{-1/\gamma}$ for some $\sigma = \sigma(t)$, and $\tau \neq -1$, then the rate of convergence would in both cases be $O\{|\delta(t)|\}$ only. Therefore by approximating the distribution of the POT's X/t given that $X > t$ by the EPD or EGPD rather than by Pareto or GPD, lifts up the first bias term. We retain our focus on the EPD.

3.3 Estimation procedures: Maximum likelihood

Assume the second-order assumption on F in terms of the tail quantile function U defined by

$$U(y) = Q(1 - 1/y) \text{ with } Q(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}, \quad (3.14)$$

where $y \in (1, \infty)$ and $p \in (0, 1)$.

Assume further that $F \in \mathcal{F}(\gamma, \tau)$ with $\lim_{x \rightarrow \infty} x^{1/\gamma} \bar{F}(x) = C \in (0, \infty)$, then

$$\lim_{y \rightarrow \infty} y^{-\gamma} U(y) = C^\gamma,$$

and the function a , eventually non zero and of a constant sign $|a| \in \mathcal{R}_\rho$, with $\rho = \gamma\tau$, defined implicitly by

$$U(y) = C^\gamma y^\gamma \{1 + a(y)\}, \quad (3.15)$$

satisfying

$$a(y) = \delta(U(y))\{1 + o(1)\} = \delta(C^\gamma y^\gamma)\{1 + o(1)\}, \quad (3.16)$$

as $y \rightarrow \infty$, with δ as in Equation 3.11. Note also that if F is not continuous, then $y\bar{F}(U(y)) = 1 + o\{|a(y)|\}$ as $y \rightarrow \infty$.

To approximate $\bar{F}(x)$ for $x \geq t$ in terms of $\bar{F}(t)$, Equation 3.13 is written as:

$$\bar{F}(ty) = \bar{F}(t)\bar{G}_{\gamma, \delta(t), \tau}(y) + o\{\bar{F}|\delta(t)|\} \quad (3.17)$$

and the remainder term is omitted. Let a random sample X_1, X_2, \dots, X_n from F , be ordered as:

$$X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n}$$

with relative excesses X/t taken over the data-adaptive threshold $\mu = \mu_n = X_{n-k:n}$ where $k = k_n \in \{1, \dots, n-1\}$ is an intermediate sequence of integers, i.e, $k \rightarrow \infty$ and $k/n \rightarrow 0$ as $n \rightarrow \infty$ and taken such that $X_i > t, \forall i \in \{1, \dots, n-1\}$. The estimators of γ and δ_n are obtained by maximizing the EPD likelihood function given the sample of k relative excesses $X_{n-k+i:n}/X_{n-k:n}, i \in \{1, \dots, k\}$, over the random threshold $X_{n-k:n}$ (Beirlant et al., 2009).

The density function of the EPD is given by:

$$g_{\gamma, \delta, \tau}(x) = \frac{d}{dx}[G_{\gamma, \delta, \tau}(x)] = \frac{1}{\gamma} x^{-1/\gamma-1} \{1 + \delta(1 - x^\tau)\}^{-1/\gamma-1} [1 + \delta\{1 - (1 + \tau)x^\tau\}] \quad (3.18)$$

The score functions are given by Beirlant et al. (2009) as:

$$\frac{\partial}{\partial \gamma} \log g_{\gamma, \delta, \tau}(x) = -\frac{1}{\gamma} + \frac{1}{\gamma^2} \log x + \frac{\delta}{\gamma^2} (1 - x^\tau) + O(\delta^2), \quad (3.19)$$

$$\begin{aligned} \frac{\partial}{\partial \delta} \log g_{\gamma, \delta, \tau}(x) = & \frac{1}{\gamma} \{(1 - \gamma\tau)x^\tau - 1\} + \{1 - 2(1 - \gamma\tau)x^\tau + \\ & (1 - 2\gamma\tau - \gamma\tau^2)x^{2\tau}\} \frac{\delta}{\gamma} + O(\delta^2) \end{aligned} \quad (3.20)$$

Solving these Linearized score equations yields:

$$\hat{\gamma}_{k,n} = H_{k,n} + \hat{\delta}_{k,n} \{1 - E_{k,n}(\tau)\}, \quad (3.21)$$

$$\begin{aligned}
 (\hat{\gamma}_{k,n}\tau - 1)E_{k,n}(\tau) + 1 = \\
 \{1 - 2(1 - \hat{\gamma}_{k,n}\tau)E_{k,n}(\tau) + (1 - 2\hat{\gamma}_{k,n}\tau - \hat{\gamma}_{k,n}\tau^2)E_{k,n}(2\tau)\}\hat{\delta}_{k,n}
 \end{aligned} \tag{3.22}$$

where

$$H_{k,n} = \frac{1}{k} \sum_{i=1}^k \log(X_{n-k+i:n}/X_{n-k:n}), \tag{3.23}$$

is the Hill (1975) estimate

$$E_{k,n}(s) = \frac{1}{k} \sum_{i=1}^k (X_{n-k+i:n})^s, \quad s \leq 0 \tag{3.24}$$

Lastly, Beirlant et al. (2009) gives the simplified estimators as:

$$\hat{\delta}_{k,n} = H_{k,n}(1 - 2\hat{\rho}_n)(1 - \hat{\rho}_n)^3 \frac{1}{\hat{\rho}_n^4} (E_{k,n}(\hat{\rho}_n/H_{k,n}) - \frac{1}{1 - \hat{\rho}_n}), \tag{3.25}$$

$$\hat{\gamma}_{k,n} = H_{k,n} - \hat{\delta}_{k,n} \frac{\hat{\rho}_n}{1 - \hat{\rho}_n} \tag{3.26}$$

where $\hat{\delta}_{k,n}$ can be expected to be of order $O_p(k^{-1/2})$, $n \rightarrow \infty$ and $\rho = \gamma\tau$.

In order to estimate $\tau = \hat{\tau}_{k,n} = \hat{\rho}_n/H_{k,n}$, the unknown second-order parameter ρ is then replaced by a consistent estimator $\hat{\rho}_n$ which will not affect the asymptotic distribution of the other estimators. This estimator is estimated externally in Section 3.3.1.

Due to the Maximum Likelihood estimation's asymptotic efficiency property, Beirlant et al. (2009) was able to prove that the EPD EVI estimator ($\hat{\gamma}_{k,n}^{EPD}$), has in fact reduced bias in comparison to that of the GPD ($\hat{\gamma}_{k,n}^{GPD}$), and Pareto ($\hat{\gamma}_{k,n}^{PD}$), expressed explicitly under the second-order condition:

$$\sqrt{k}(\gamma_{k,n}^{\hat{EPD}} - \gamma) \sim N\left(0, \gamma^2 \frac{(1 - \rho)^2}{\rho^2}\right), \quad n \rightarrow \infty \tag{3.27}$$

$$\sqrt{k}(\gamma_{k,n}^{PD} - \gamma) \sim N\left(\lambda \frac{\rho}{1 - \rho}, \gamma^2\right), \quad n \rightarrow \infty \tag{3.28}$$

$$\sqrt{k}(\gamma_{k,n}^{\hat{GPD}} - \gamma) \sim N(\lambda b(\gamma, \rho), (1 + \gamma)^2), \quad n \rightarrow \infty \tag{3.29}$$

where

$$b(\gamma, \rho) = \frac{\rho(1 + \gamma)(\gamma + \rho)}{\gamma(1 - \rho)(1 + \gamma - \rho)},$$

and λ is defined as follows

$$\sqrt{k}a(n/k) \rightarrow \lambda \in \mathbb{R}, \quad n \rightarrow \infty. \tag{3.30}$$

Beirlant et al. (2009) showed also that by expressing δ_n as $\delta_n = \delta(t_n) = \delta(X_{n-k:n})$ Equation 3.30 implies that:

$$\sqrt{k}\delta_n = \lambda + o_p(1), \quad n \rightarrow \infty. \quad (3.31)$$

As can be seen from Equation 3.27, the EPD has a significantly reduced asymptotic bias term, as compared to both the GPD and the PD. From Equation 3.31 we know that the estimators are asymptotically unbiased in the sense that whatever the value of λ , the mean of the normal limiting distribution will always be equal to zero. Also note that with the EPD estimate, the mean of the limit distribution is not proportional to λ and therefore it is possible to use a larger k and lower thresholds. Of the three estimators however, the Hill (1975) estimator has the smallest asymptotic variance.

3.3.1 Estimation of the second order parameter

In this chapter, we estimate the second order parameter ρ by using Fraga's estimator (Fraga Alves et al., 2003b). In Chapter 5 however, a comparison is made between estimating parameters of the EPD with ρ fixed and estimating all three parameters of the EPD simultaneously. Other possible ways to deal with the second order parameter is to fix it at some value, usually -1 , this is however not investigated in this study.

Gomes et al. (2010) notes that to estimate ρ it is necessary to employ many more upper order statistics than in the first order parameter γ .

Algorithm to estimate ρ

1. Given a sample (X_1, X_2, \dots, X_n) , plot the estimates of

$$\hat{\rho}_{k,r} := \min\left\{0, \frac{3(T_{k,n}^{(\tau)} - 1)}{T_{k,n}^{(\tau)} - 3}\right\}, \quad (3.32)$$

for $\tau = 0$, where

$$T_{k,n} := \frac{\left(M_{k,n}^{(1)}\right)^\tau - \left(\frac{M_{k,n}^{(2)}}{2}\right)^{\tau/2}}{\left(\frac{M_{k,n}^{(2)}}{2}\right)^{\tau/2} - \left(\frac{M_{k,n}^{(3)}}{6}\right)^{\tau/3}} \quad (3.33)$$

and

$$M_{k,n}^{(j)} := \frac{1}{k} \sum_{i \leq k} (\log X_{n-i+1:n} - \log X_{n-k:n})^j, \quad j \geq 1. \quad (3.34)$$

since $\tau = 0$, the notation $a^{b\tau} = b \ln a$ is used and therefore Equation 3.33 is written as

$$T_{k,n} := \frac{\log(M_{k,n}^{(1)}) - \frac{1}{2} \log\left(\frac{M_{k,n}^{(2)}}{2}\right)}{\frac{1}{2} \log\left(\frac{M_{k,n}^{(2)}}{2}\right) - \frac{1}{3} \log\left(\frac{M_{k,n}^{(3)}}{6}\right)} \quad (3.35)$$

2. Consider $\{\hat{\rho}_{k,n}\}_{k \in \mathcal{K}}$ as given in Equation 3.32 for large k , say $k \in \mathcal{K} = ([n^{0.990}], [n^{0.999}])$, compute their median, denoted ρ_τ and work with $\hat{\rho} \equiv \hat{\rho}_{\tau_0} := \hat{\rho}_{k_1}$, with

$$k_1 = [n^{0.995}] \quad (3.36)$$

The choice of the level k_1 is not very important, we can consider any reasonably large value of k of order $n^{1-\epsilon}$ from some intermediate sequence of integers $k = k(n)$ satisfying $\lim_{n \rightarrow \infty} \sqrt{k}A(n/k) = \infty$.

In the next section, a simulation study will be conducted to assess the behavior of the EPD ML estimator. This estimator will be compared with other estimators in terms of the bias, MSE and variance.

3.4 Simulation Study

To illustrate the behavior of all three estimators $\hat{\gamma}^{EPD}$, $\hat{\gamma}^{PD}$ and $\hat{\gamma}^{GPD}$ using the ML method of estimation, 10,000 samples of size $n = 1000$ are generated from three different Pareto-type distributions: the Fréchet, the Burr (XII) and the absolute Student-t distribution. The $\hat{\gamma}^{EPD}$, $\hat{\gamma}^{PD}$ and $\hat{\gamma}^{GPD}$ are then estimated using the ML method of estimation, where the number of observations above a threshold varies with k .

For each distribution the Monte Carlo estimates of the mean square error, variance and bias are estimated by averaging out over the 10,000 samples. The second-order parameter ρ of the EPD is estimated using the Fraga Alves et al. (2003b) method as explained earlier in Section 3.3.1. This parameter is considered as fixed, while the other two parameters are estimated with each simulation.

The Monte Carlo estimates of the asymptotic variance, are given in Beirlant et al. (2009) as:

$$\text{Var}(\hat{\gamma}^{EPD}) = \frac{1}{k} \left(\hat{\gamma}^2 \frac{(1 - \hat{\gamma}\hat{\tau})^2}{\hat{\gamma}\hat{\tau}^2} \right), \quad (3.37)$$

$$\text{Var}(\hat{\gamma}^{PD}) = \hat{\gamma}^2 \quad (3.38)$$

and

$$\text{Var}(\hat{\gamma}^{GPD}) = (1 + \hat{\gamma})^2, \quad (3.39)$$

respectively. The Monte Carlo estimates of the Mean Square Errors (MSE) for all three estimators as:

$$\text{MSE}(\hat{\gamma}) = [(\hat{\gamma} - \gamma)^2] \quad (3.40)$$

while the bias for all three estimators is given by

$$\text{Bias} = (\hat{\gamma} - \gamma). \quad (3.41)$$

Estimates of the Hill (1975) are given as in Equation 3.23, while for the GPD an **R** package ('evir') is used to fit the GPD to the simulated excesses above a threshold, by way of defining the argument: `'gpd(data, threshold = NA, nextremes = NA, method = c("ml", "pwm"), information = c("observed", "expected"), ...)`, see, appendix A, (Pfaff et al., 2012).

Next we present results of the simulation study.

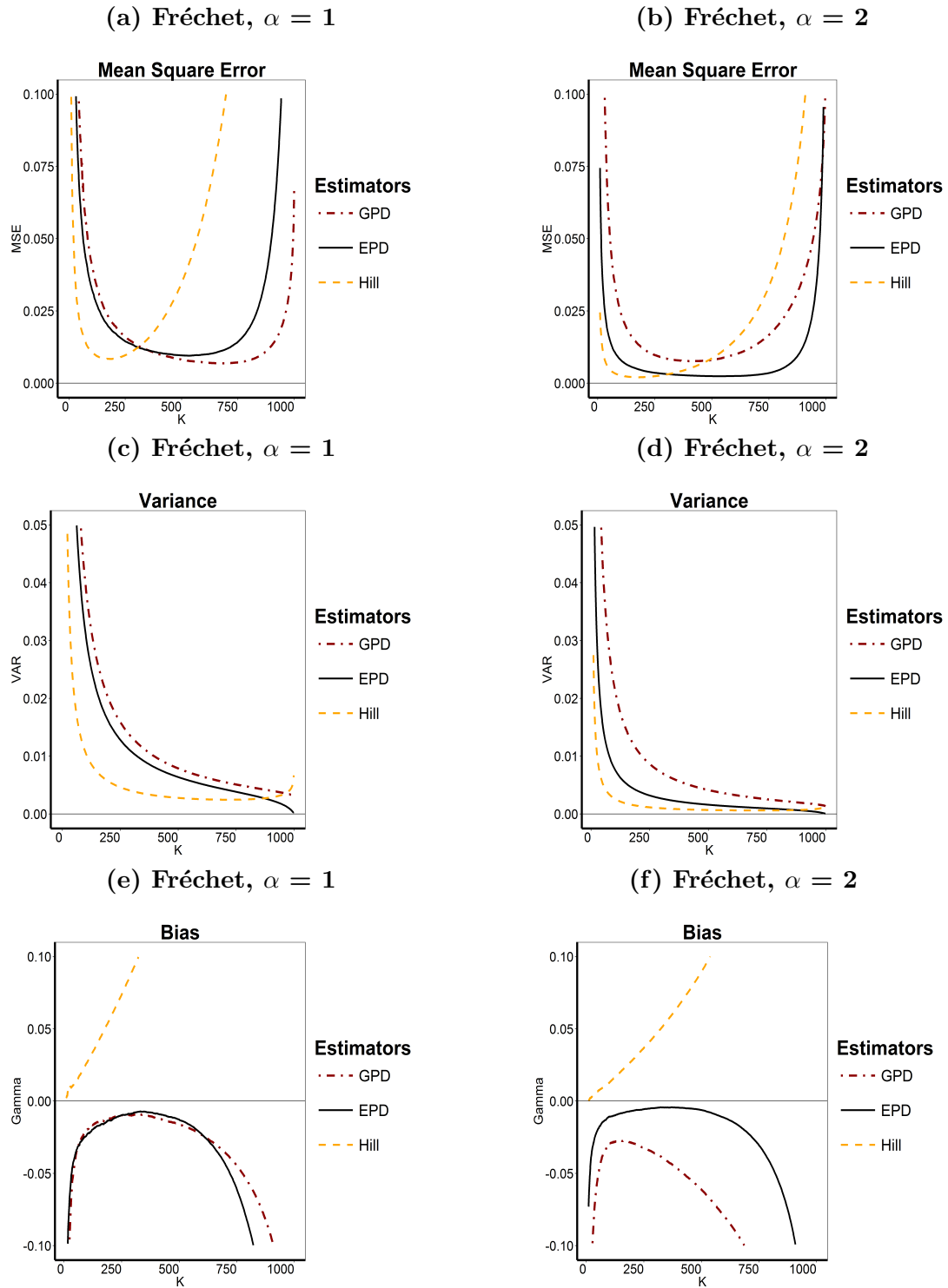


Figure 3.1: Fréchet(α) samples: Mean Square Error (MSE) estimate (top), Variance estimate (middle) and Bias estimate (bottom). The sample size is $n = 1000$ and the plots are obtained by averaging out over 10,000 samples.

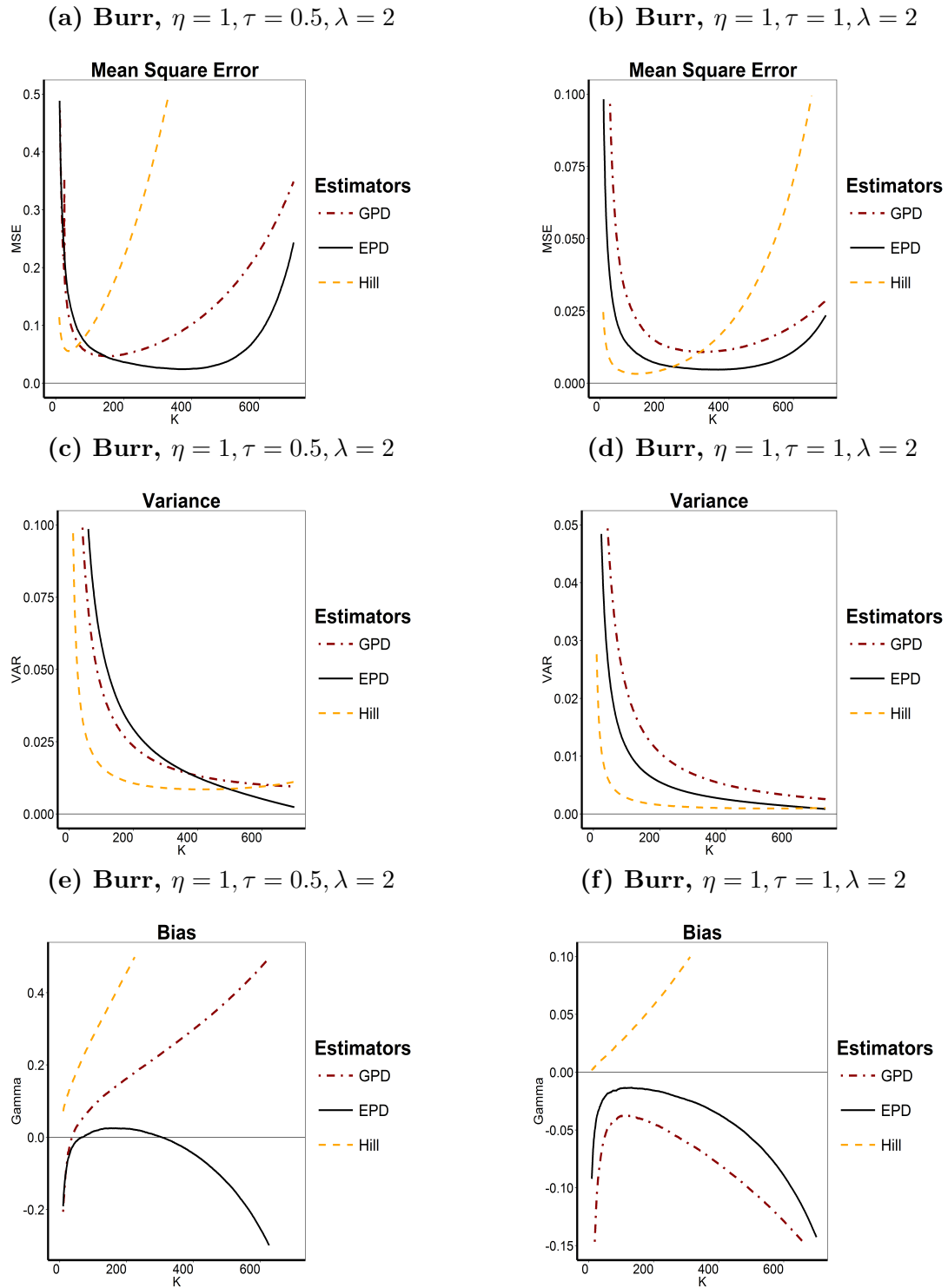


Figure 3.2: Burr(η, τ, λ) samples: Mean Square Error (MSE) estimate (top), **Variance** estimate (middle) and **Bias** estimate (bottom). The sample size is $n = 1000$ and the plots are obtained by averaging out over 10,000 samples.

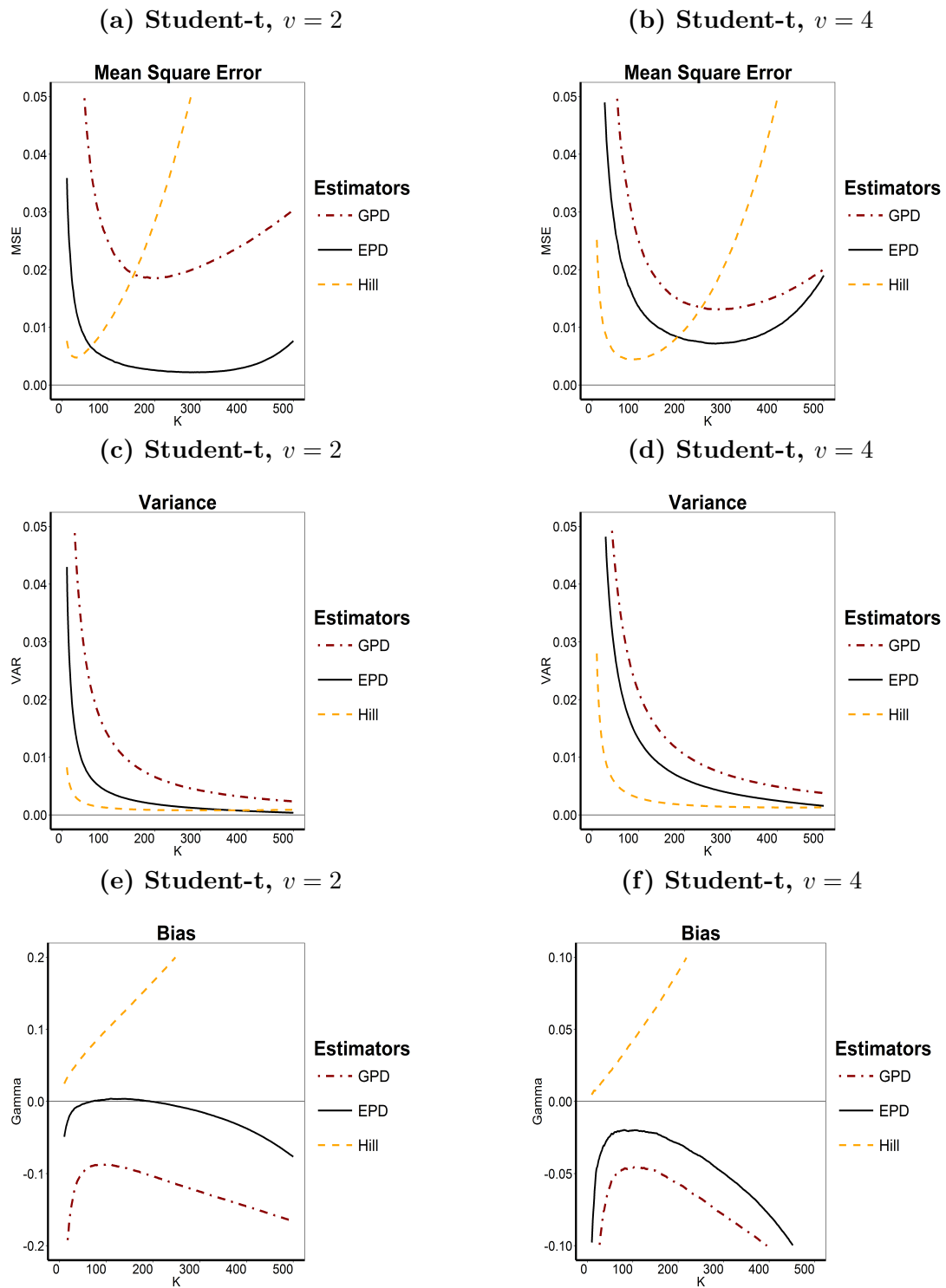


Figure 3.3: Student- $t(v)$ samples: Mean Square Error (MSE) estimate (top), Variance estimate (middle) and Bias estimate (bottom). The sample size is $n = 1000$ and the plots are obtained by averaging out over 10,000 samples.

Remarks

Figure 3.1 shows EVI estimates from a **Fréchet**(α) distribution for case 1 ($\alpha = 1 \rightarrow \gamma = 1$) and case 2 ($\alpha = 1 \rightarrow \gamma = 1/2$): in both cases, the asymptotic distributions of both the EPD and the GPD estimators have a zero asymptotic bias while the Hill (1975) has a nonzero asymptotic bias. The EPD in both cases seems to have reduced biased over a larger range of k 's, regardless of the tailed-heaviness of the sampled distribution, this can also explain why it is possible to use larger k 's and lower thresholds for the EPD.

Figure 3.2 shows the EVI estimates from a **Burr (XII)**(η, τ, λ) distribution for case 1 ($\eta = 1, \tau = 0.5, \lambda = 2 \rightarrow \gamma = 1$) and case 2 ($\eta = 1, \tau = 1, \lambda = 2 \rightarrow \gamma = 1/2$): For all three estimators, the EPD seems to be the only one that is asymptotically unbiased, and stable across a larger range of k 's.

Figure 3.3 shows the EVI estimates from a **Student-t**(v) for both case 1 ($v = 2 \rightarrow \gamma = 0.5$) and case 2 ($v = 4 \rightarrow \gamma = 0.25$): Again the EPD estimator is the only one which is asymptotically unbiased, as compared to the GPD and Hill (1975) estimator, regardless of the tail-heaviness of the sampled distribution.

We can thus see that the EPD can be fitted to a larger portion of the data, and results in reduction of bias and a good fit to excesses over a larger range of thresholds. This model will therefore be used in later chapters to estimate robust and asymptotically unbiased extreme quantiles.

3.5 Conclusion

In this section we discussed the motivation for the use of the EPD as a limiting distribution for excesses over a threshold. A generalization of the EPD was discussed in Section 3.2 as well as propositions as given by Beirlant et al. (2009) showing that the EPD improves approximations to the excess distribution with an order of magnitude.

A detailed simulation estimation procedure was explicated in section Section 3.3 and a simulation study was conducted in Section 3.4. From this study, we can see that the EPD is a better POT model to consider than the GPD. The rest of this thesis is focused on investigating other different estimation procedures for the EPD, and also applying the different methods of estimation to real datasets in order to assess their performance.

Chapter 4

Improved EPD maximum likelihood estimator for complete data

4.1 Introduction

In Chapter 3 we used pseudo-maximum likelihood parameter estimates as given in Beirlant et al. (2009). The main advantage of these estimates is their relative ease of implementation. However, analytically solving the score functions of the maximized EPD likelihood function, means we have to neglect some small order approximation $O(\delta^2)$, doing this gives rise to a small bias term. An alternative to maximizing the EPD likelihood function approach is to optimize the likelihood function directly and relaxing the assumption of $\delta \rightarrow 0$. This can be achieved by employing a numerical optimization technique. **R** (R Core Team, 2014) has a General-purpose Optimization function which makes use of the following methods to perform box-constrained optimization and simulated annealing: the Nelder-Mead (Nelder and Mead, 1965), the Conjugate gradients (Fletcher and Reeves, 1964), the SANN, a variant of simulated annealing (Bélisle, 1992), the Brent (Brent, 2013) and the limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS-B) method (Broyden (1970); Fletcher (1970); Goldfarb (1970); Shanno (1970)).

We make use of the L-BFGS-B quasi-Newton algorithm, to estimate parameters of a bounded and constrained function. This method takes advantage of the limited amount of computer memory used to approximation the original BFGS and thus implements the algorithm more efficiently. This algorithm is used to mainly solve large non-linear optimization problems with simple bounds. Byrd et al. (1994) compared the L-BFGS method against some well known conjugate gradient methods in terms of execution times, and found the L-BFGS-B method to be faster and more efficient.

In the following Section 4.2 we detail the estimation procedure as applied in the

statistical software **R**, and conduct a simulation study in Section 4.3. Some concluding remarks are made in Figure 4.3.

4.2 Estimation procedure

In this chapter we implement the L-BFGS-B algorithm on the EPD's negative log likelihood (objective function), to estimate the parameters γ and δ . Estimators of γ^* and δ^* are found by maximizing the EPD log likelihood or minimizing the negative log likelihood i.e $(\gamma^*, \delta^*) = \operatorname{argmin}\{-\log(\prod_{i=1}^n g_{\gamma,\delta,\tau}(x_i))\}$. The likelihood function of the EPD is given by:

$$\prod_{i=1}^n g_{\gamma,\delta,\tau}(x_i) = \frac{1}{\gamma^n} \prod_{i=1}^n x_i^{-1/\gamma-1} \prod_{i=1}^n \{1 + \delta(1 - x_i^\tau)\}^{-1/\gamma-1} \prod_{i=1}^n \{1 + \delta[1 - (1 + \tau)x_i^\tau]\} \quad (4.1)$$

By taking the log of Equation 4.1 and multiplying by -1, we get the Negative log-likelihood of the EPD, which is the objective function to be optimized, given as;

$$\begin{aligned} n \log \gamma + \left(\frac{1}{\gamma} + 1\right) \sum_{i=1}^n \log(x_i) + \left(\frac{1}{\gamma} + 1\right) \sum_{i=1}^n \log(1 + \delta(1 - x_i^\tau)) \\ - \sum_{i=1}^n \log(1 + \delta(1 - (1 + \tau)x_i^\tau)) \end{aligned} \quad (4.2)$$

The L-BFGS-B algorithm solves this objective function with bounds $lower \leq x \leq upper$ on the variables, where $g : \mathbb{R}^n \rightarrow \mathbb{R}$. For an intermediate sequence of integers $k \in \{1, \dots, n - 1\}$, every next $(k + 1) - th$ vector of parameter estimates (γ, δ) is bounded by the following lower and upper bound vector: $lower = [1e-6, -T * |\hat{\delta}_k|]$ and $upper = [T * |\hat{\gamma}_k|, T * |\hat{\delta}_k|]$, where T is an arbitrary scaler tuning parameter. The second order parameter ρ is again estimated by the Fraga Alves et al. (2003b) method and fixed for all k (see Section 3.3.1).

4.3 Simulation study

Estimators $\hat{\gamma}^{pseudo-EPD}$, $\hat{\gamma}^{PD}$ and $\hat{\gamma}^{GPD}$ are estimated the same way as in Chapter 3, Section 3.4 using the MLE method of estimation, however in this Section an additional ML estimator is added, we call it here the optim-EPD $\hat{\gamma}^{optim-EPD}$. We generate 10,000 samples of size $n = 1000$ from four different Pareto-type distributions; the Fréchet, the Burr (XII), the absolute Student-t distribution and the GPD. For each distribution the Monte Carlo estimates of the mean square error, variance and bias are estimated by averaging out over the 10,000 samples.

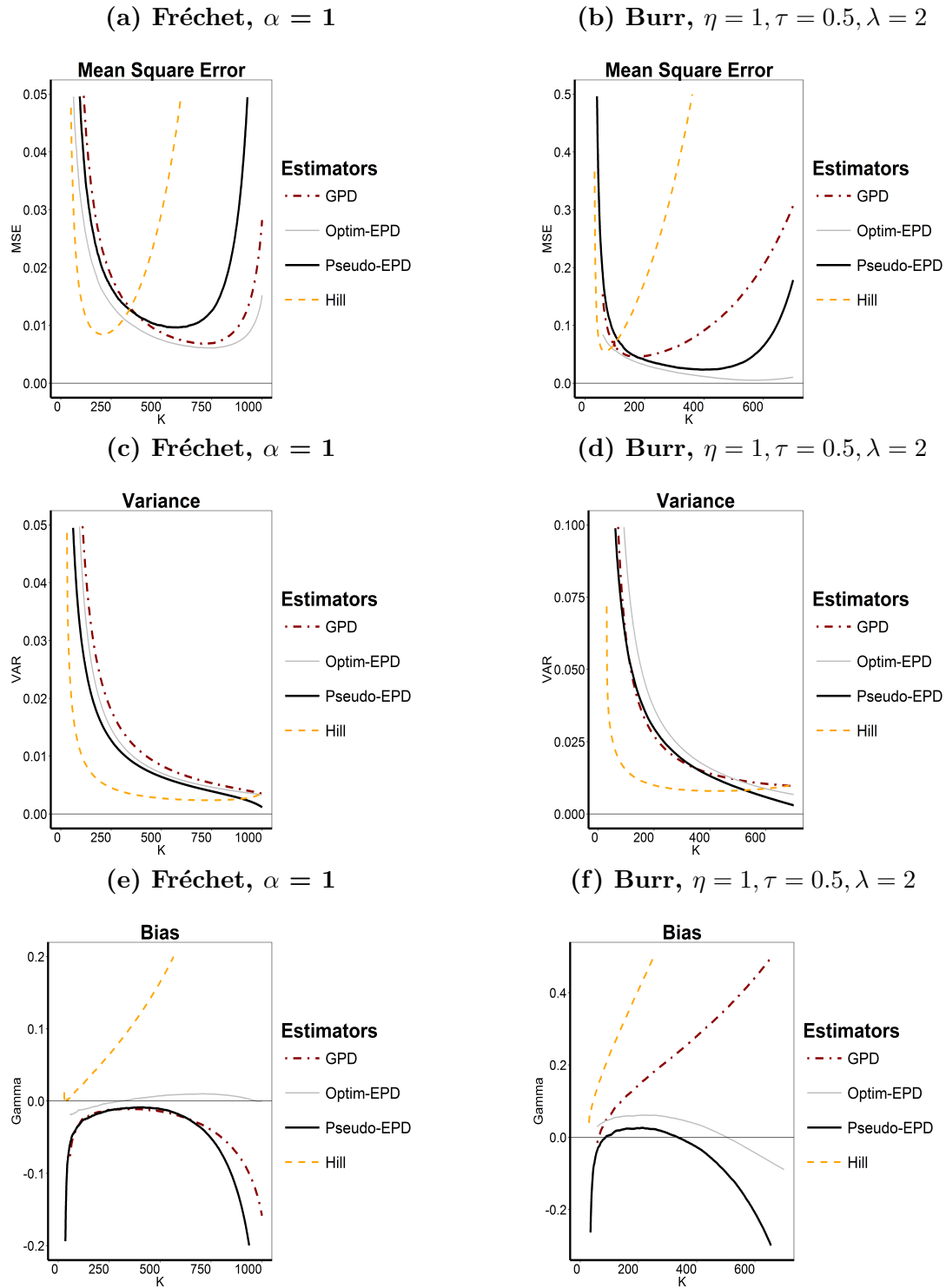


Figure 4.1: Mean Square Error (MSE) estimate (top), **Variance** estimate (middle), and **Bias** estimate (bottom). The sample size is $n = 1000$ and the plots are obtained by averaging out over 10,000 samples.

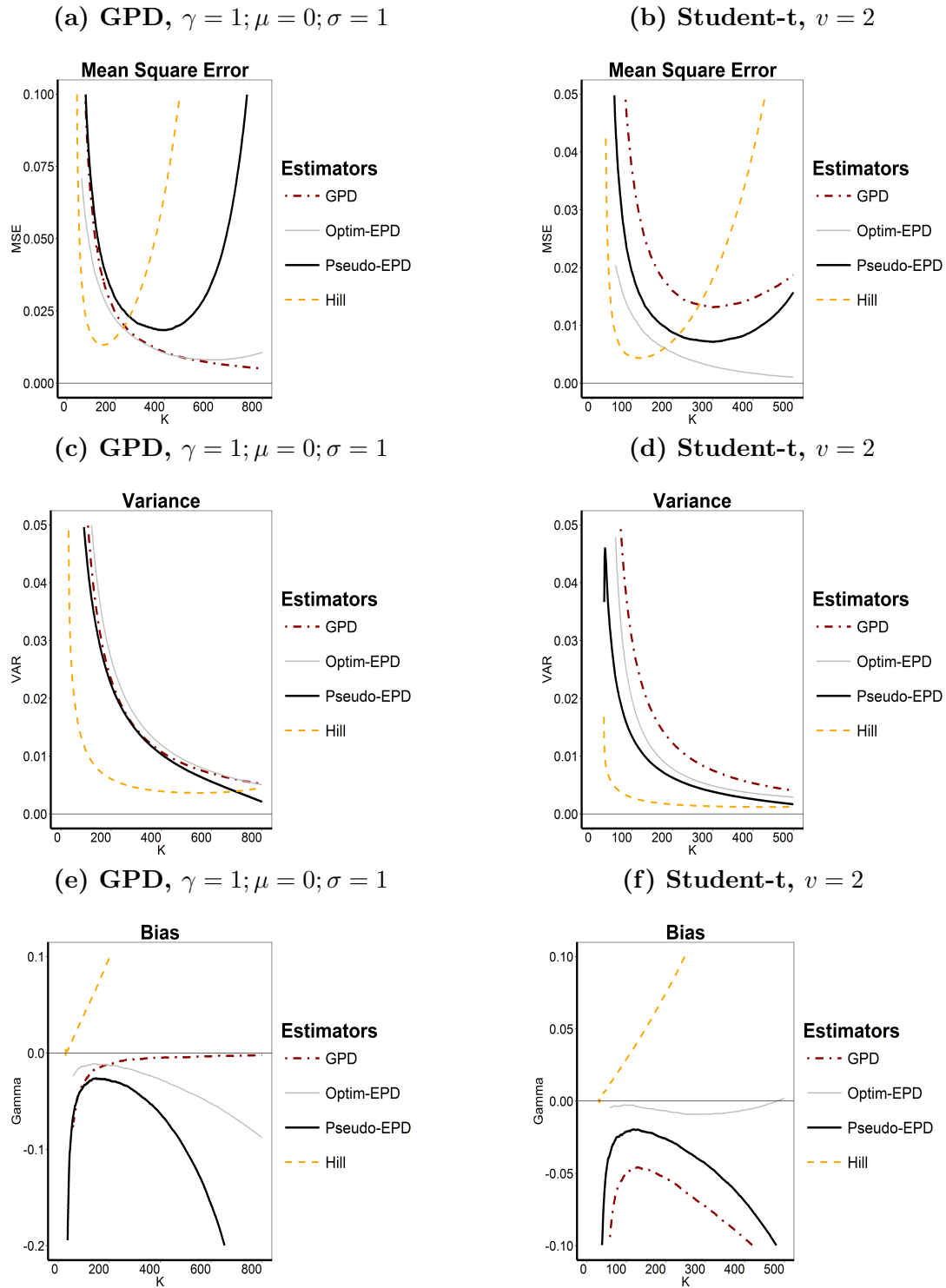


Figure 4.2: Mean Square Error (MSE) estimate (top), **Variance** estimate (middle), and **Bias** estimate (bottom). The sample size is $n = 1000$ and the plots are obtained by averaging out over 10,000 samples.

Remarks

For both the Fréchet and the Burr(XII) samples, from Figure 4.1 we can see that the MSE of the optim-EPD estimator is lower than both the GPD and the pseudo-EPD. However the optim-EPD estimator has a slightly bigger asymptotic variance than the pseudo-EPD estimator although in the case of the Fréchet the asymptotic variance is still smaller than the asymptotic variance of the GPD estimator. Both EPD estimators have zero asymptotic bias. The optim-EPD estimator is the most stable of all the four estimators.

For the GPD and the Student-t samples, from Figure 4.2 the optim-EPD has a smaller MSE than all the other estimators. The optim-EPD has an asymptotic variance smaller than the GPD estimator but higher than the pseudo-EPD estimator. The optim-EPD estimator gives an overall better performance than all mentioned estimators for estimating the EVI of Student-t samples. For the GPD samples the optim-EPD and the GPD estimators have a coinciding MSE better than the pseudo-EPD estimator. However the variance of the optim-EPD estimator is higher than all other mentioned estimators. Since the data is from a GPD we can expect the GPD estimator to give the best performance. The optim-EPD estimator performs better than the pseudo-EPD estimator.

The **R** statistical software contains a general-purpose optimization package that works very well for such maximization (minimization) problems. As can also be seen from the conducted simulations. Following this approach leads to improved estimates of the EVI. In the next section we introduce a new Bayesian EPD Estimator for the EVI and show that this estimator also leads to much improved results compared to the ML EPD estimator.

In the next chapter we introduce the Bayesian EVI estimator of the EVI and discuss some methods of approximating the joint posterior density in Bayesian inference. The Bayesian estimator is also compared with the ML estimator.

Chapter 5

EPD Bayesian estimator for complete data

“...it is not asserted that a belief is an idea which does actually lead to action, but one which would lead to action in suitable circumstances; just as a lump of arsenic is called poisonous not because it actually has killed or will kill anyone, but because it would kill anyone if he ate it” (Ramsey, 1931)

5.1 Introduction

In reality perfect information is never available, there are always some uncertain aspects of any situation. For this reason, beliefs about the uncertainty of a situation can be an input into the decision making process of any practical course of action. This concept of uncertainty and rationality embodies the framework of Bayesian theory. The well known Bayes' theorem describes the mechanisms of updating prior beliefs about our observables with new information in light of changing evidence.

Bayesian theory offers an additional set of interesting tools in EVT that may very well serve as an alternative to some of the classical methods of estimation such as the ML method. However complications of Bayesian inference stems from multidimensional integration problems involved in posterior calculations, amongst other things. Also, Markov Chain Monte Carlo (MCMC) methods can be “computationally demanding and much harder to implement, using non-conventional software that is not widely available among researchers and practitioners in the field” (Fridman and Harris, 1998). However, through advances in computing power and efficiency of MCMC algorithms, it has become easier to implement MCMC methods to some common statistical problems.

In EVT we are usually faced with data scarcity, due to the fact that extreme data

points are scarce for any system, for instance a natural disaster can happen only 5 times in 100 years. One could then ask; can we use an expert's opinion in our inference, or do we just rely on the few data points that we have? A pragmatic motivation for a Bayesian paradigm in EVT is quite clear, that any knowledge that can help us make better approximations/predictions, should surely be incorporated when making inference about our model parameters.

In this chapter, we will illustrate the ease of implementing Bayesian estimation procedures to estimate parameters of the EPD using **R** and the OpenBUGS software package (Sturtz et al., 2010) and also showing that this approach does indeed result in more stable and less biased estimates of the EVI, in comparison to the classical MLE.

In Section 5.2 we give a short overview of Bayesian inference. In Section 5.3 we show how to sample from the joint EPD posterior distribution using the Metropolis and Gibbs sampling algorithms. In Section 5.3.3 we detail the estimation procedure through specifying and implementing the EPD model in OpenBUGS, here we show how we specify the joint posterior and how we derive the prior distributions for all parameters. In Section 5.3.6, we explain how OpenBUGS chooses its sampling method. An exhaustive Monte Carlo simulation study is conducted in Section 5.5 to assess the behaviour of our estimate where the second order parameter is fixed (the first case) and where the second order parameter is estimated concurrently with other parameters (the second case).

We further demonstrate the benefits of our Bayesian approach with a small case-study in Section 5.7 using a real dataset and using our estimated parameters to compute the following quantities:

- Probability of exceedance of a high level x , $p_x := \mathbb{P}(X > x) = 1 - F(x)$,
- High Quantile of probability $1 - p$, p small, defined as $\chi_{1-p} := \inf(\{x : F(x) \geq 1 - p\}) =: F^{\leftarrow}(1 - p)$, $p < 1/n$,

where F^{\leftarrow} indicates the left-continuous inverse function of F . Some concluding remarks are made in Section 5.8.

5.2 Overview of Bayesian Inference

“Sometimes a lot of data can be meaningless: at other times one single piece of information can be very meaningful” (Taleb, 2010). Bayesian inference has a great advantage of incorporating in a unified way, any meaningful piece of information in describing our model parameters. This added advantage can be very useful in EVT, where data is often not enough to make any reliable inference.

Let $\mathbf{x} = (x_1, \dots, x_n)$ denote the observed data of a random variable X distributed according to a distribution with density function $p(x|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of un-

known parameters in the parameter space Ω . Assuming we have prior knowledge about the distribution of the parameter vector $\boldsymbol{\theta}$, the density of the prior distribution for $\boldsymbol{\theta}$ can be denoted as $p(\boldsymbol{\theta})$. We can write the likelihood as $p(\mathbf{x}|\boldsymbol{\theta})$. According to Bayes' theorem, the distribution of $\boldsymbol{\theta}|\mathbf{x}$, called the posterior distribution of $\boldsymbol{\theta}$, can be written as

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\Omega} p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \propto p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \quad (5.1)$$

The marginal posterior distribution function of the parameter θ_i can be obtained by integrating out the remaining parameters. If $\boldsymbol{\theta}$ is p dimensional, the marginal posterior distribution will be

$$p_i(\theta_i|\mathbf{x}) = \frac{\int_{\theta_1} \dots \int_{\theta_{i-1}} \int_{\theta_{i+1}} \dots \int_{\theta_p} p(\boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})d\theta_p \dots d\theta_{i+1}d\theta_{i-1} \dots d\theta_1}{\int_{\Omega} p(\boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})d\boldsymbol{\theta}}. \quad (5.2)$$

We then rely on modern MCMC methods to approximate the posterior distributions. Examples of these are: Metropolis-Hastings algorithm, Gibbs sampling, Slice sampling, rejection sampling and Adaptive rejection sampling. In order to make inference on $\boldsymbol{\theta}$ we take the mean of the posterior samples, the accuracy of this inference is described by the posterior distribution itself through the Highest Posterior Density (HPD) region according to a certain probability $1 - \alpha$.

5.3 Markov Chain Monte Carlo methods

In this section we will illustrate some practical considerations when fitting the EPD using Bayesian inference. We consider three ways of implementing the EPD model: programming the model in OpenBUGS by calling it in **R** using the "R2OpenBUGS" package; programming the EPD model directly in **R** using the Metropolis algorithm and also using the Gibbs sampler. Coding the algorithms directly in **R** gives us a greater level of control. However it can get very computationally intensive.

Next we define the Metropolis algorithm as given by Gelman et al. (2014) and the Gibbs sampler as given by (Gelfand, 2000)

5.3.1 Metropolis algorithm

1. Draw a starting point θ^0 where $p(\theta^0|y) > 0$, from a starting distribution $p_0(\theta)$ (we use here $\gamma_0 = H_{k,n}$, $\delta_0 = 0.01$ and $\tau_0 = -0.5$ as starting points, where $H_{k,n}$ is the Hill estimator)
2. For $t=1,2,\dots$:
 - Sample a proposal θ^* from a jumping distribution at time t , $J_t(\theta^*|\theta^{t-1})$ (We use here the Normal distribution as a jumping distribution since it is symmetric)

- Calculate the ratio of the densities,

$$r = \frac{p(\theta^*)}{p(\theta^{t-1}|y)}. \quad (5.3)$$

- Set

$$\theta^t = \begin{cases} \theta^* & \text{with probability } \min(r, 1) \\ \theta^{t-1} & \text{otherwise} \end{cases} \quad (5.4)$$

Given the current value θ^{t-1} , the transition distribution $T_t(\theta^t|\theta^{t-1})$ of the Markov chain thus becomes the mixture of a point mass at $\theta^t = \theta^{t-1}$, as well as the weighted version of the jumping distribution, $J_t(\theta^t|\theta^{t-1})$, that adjusts for the acceptance/rejection rate. This latter step requires the generation of a uniform random number. If $\theta^t = \theta^{t-1}$ the jump is not accepted. The algorithm then repeats until a certain number of jumps have been accepted.

5.3.2 Gibbs sampler

Suppose that the parameter vector $\boldsymbol{\theta}$ is partitioned into r blocks i.e. $\boldsymbol{\theta} = (\theta_1, \dots, \theta_r)$. If the current state of $\boldsymbol{\theta}$ is $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \dots, \theta_r^{(t)})$, then assume we make some transition $\boldsymbol{\theta}^{(t+1)}$ as follows:

draw $\theta_1^{(t+1)}$ from $h(\theta_1|\theta_2^{(t)}, \dots, \theta_r^{(t)})$,
draw $\theta_2^{(t+1)}$ from $h(\theta_2|\theta_1^{(t+1)}, \dots, \theta_3^{(t)}, \dots, \theta_r^{(t)})$,
 \vdots
draw $\theta_r^{(t+1)}$ from $h(\theta_r|\theta_1^{(t+1)}, \dots, \theta_{r-1}^{(t+1)})$.

The distribution of $h(\theta_i|\theta_1, \dots, \theta_{i-1}, \dots, \theta_{i+1}, \dots, \theta_r)$ are known as the full conditional distributions, the complete iterations are then produced by updating the entire vector $\boldsymbol{\theta}$. There are numerous ways to sample from θ_i . In this thesis we make use of a technique called discretization. We break the continuous log-posterior distribution of the EPD into discrete counterparts and use the *sample* function in **R** to sample from the sampling interval with a probability.

5.3.3 Implementation the EPD joint posterior in OpenBUGS

There are two main versions of BUGS (Spiegelhalter et al., 1995), namely WinBUGS and OpenBUGS. OpenBUGS is the open source version of WinBUGS and can run on both Windows and Linux. In this paper we use the word ‘BUGS’

and ‘OpenBUGS’ interchangeably. Although it is possible to apply this approach using the OpenBUGS graphical user interface, we use the “R2OpenBUGS” package in order to call OpenBUGS from the **R** (R Core Team, 2014) statistical software.

The EPD distribution function is given in Section 3.2 as

$$\bar{G}_{\gamma,\delta,\tau}(y) = \{y(1 + \delta - \delta y^\tau)\}^{-1/\gamma}, \quad y > 1 \quad (5.5)$$

where the parameter vector (γ, δ, τ) is restricted by $\tau < 0 < \gamma$ and $\delta > \max(-1, 1/\tau)$, with $\delta = \delta_t = -D\gamma^{-1}C^{\rho/\gamma}t^{-\rho/\gamma}$ and $\tau = -\rho/\gamma$. The density function is defined by:

$$g_{\gamma,\delta,\tau}(y) = \frac{1}{\gamma}y^{-1/\gamma-1}\{1 + \delta(1 - y^\tau)\}^{-1/\gamma-1}[1 + \delta\{1 - (1 + \tau)y^\tau\}] \quad (5.6)$$

This last piece of information about parameters of the EPD parametric model is going to be useful in specifying our Bayesian model. We approximate the posterior distribution in two ways;

- In the **first case** we estimate only γ and δ , and use an external estimator given in Section 3.3.1 to estimate the second order parameter ρ , where $\tau = \rho/\gamma$. This is shown in Figure 5.1a
- In the **second case** we estimate all three parameters of the EPD γ, δ and τ simultaneously. This is shown in Figure 5.1b

Graphical Model

Graphical models aid us in understanding the qualitative conditional independence structures of a model, allowing us to reduce globally complex models into smaller local components. In OpenBUGS they are known as ‘Directed Acyclic Graphs’ (DAG) or more commonly as ‘DoodleBUGS’. Below is a graphical model of the EPD expressing the joint relationship between all the parameters in the model. From this representation we are able to design the essential structures of the model for computation.

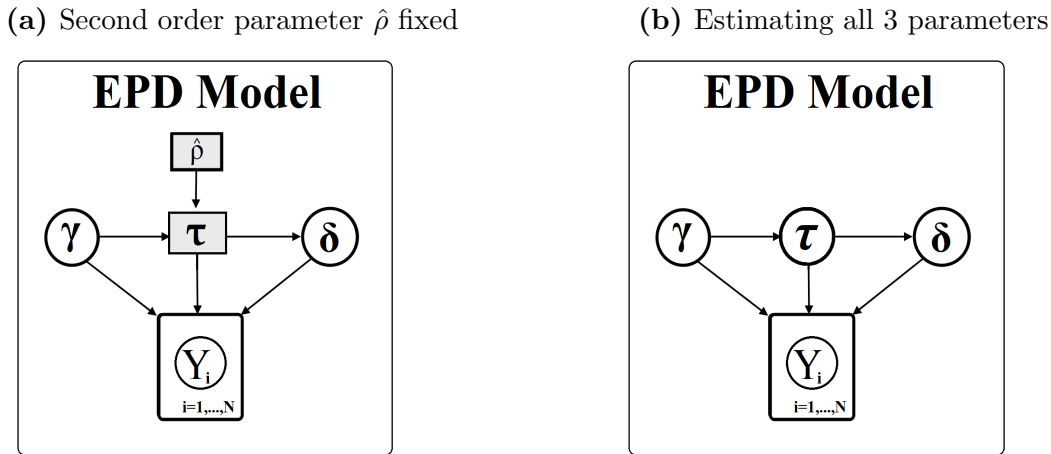


Figure 5.1: Representation of the EPD model as a directed acyclic graph (DAG)

This graphical model representation is called a directed acyclic graph because every node in this graph is directed and there are no cycles. This graphical model helps us to focus on the essential model structure. In this graphical model, the vertices are called nodes and represent quantities in the EPD model.

Conditional independence is shown by the directed edges or arrows which run into the nodes from the **parent nodes** to the **child nodes**. The three types of nodes seen in the DAGs above can be of the following types (see Lunn et al., 2012):

- **Constants** are fixed by the design of the study, they are always founder nodes (do not have parents), and are denoted as rectangles in the DAG. They must be specified in the data file;
- **Stochastic nodes** are variables following some distribution, and are denoted by a circle in the DAG. They may be parents or children (or both). Stochastic nodes can be observed as being the data or unobservable and hence parameters or unobservable due to the presence of censoring or missing values;
- **Deterministic nodes** are logical functions of other nodes.

The loops in the model are represented by a plate. This is a rectangle containing all the nodes and arrows of the model part which needs to be repeated (in this case, the likelihood specification of the EPD). From Figure 5.1 it is then easy to construct a joint probability distribution for all stochastic nodes from the graphical description of the conditional independence assumption.

Likelihood

For an intermediate sequence of integers $k \in 1, \dots, n-1$, where $k \rightarrow \infty$, we estimate only γ and δ . For each sample, we use a fixed estimate of ρ to obtain $\hat{\tau}_{k+1} = \hat{\rho}/\hat{\gamma}_k$ where $\hat{\tau}_1 = \hat{\rho}/H_{k,n}$. We define a single EPD log likelihood term for a sample of relative excesses $\mathbf{X} = (x_1, \dots, x_N)$ as:

$$L(x_i|\gamma, \tau, \delta) = -\log(\gamma) - \left(\frac{1}{\gamma} + 1\right) [\log x_i + \log(1 + \delta\{1 - x_i^\tau\})] + \log(1 + \delta\{1 - (1 + \tau)x_i^\tau\}) \quad (5.7)$$

The EPD is not included in the standard distributions list available in OpenBUGS, and we therefore manually implement the EPD model using a generic sampling distribution which applies the ‘zero poisson’ method (see Lunn et al., 2012). The sampling distribution contributes a likelihood term $L[i]$ which is a function of $x[i]$. Then we use the generic ‘loglik’ distribution *dloglik*, as a dummy observed variable by specifying the following in our model specification:

$$\text{dummy}[i] < -0 \quad \text{dummy}[i] \sim \text{dloglik}(\text{logLike}[i])$$

where *logLike* is given by Equation 5.7 and represents the contribution to the log-likelihood of the i^{th} observation.

5.3.4 Prior Choice

Model specification of the prior distribution component can seem somewhat arbitrary. Bernardo and Smith (2009) revealed a striking qualitative difference in posterior means where different distributions are chosen as priors. Let x denote the mean of n independent observables from a normal distribution with mean θ and precision λ , the idea is illustrated by considering the form of the posterior mean for the parameter θ when $p(x|\theta) = N(x|\theta, n\lambda)$ is Normally distributed and $p(\theta)$ is of arbitrary form. It was shown by Pericchi and Smith (1992) that

$$E(\theta|x) = x - n^{-1}\lambda^{-1}s(x) \quad (5.8)$$

where

$$p(x) = \int p(x|\theta)p(\theta)d\theta, \quad (5.9)$$

$$s(x) = \frac{\partial \log p(x)}{\partial x} \quad (5.10)$$

We are interested in how the behavior of the posterior mean depends on the assumed mathematical form of $p(\theta)$. Bernardo and Smith (2009) considered the following cases:

1. Assume $p(\theta)$ to be Normally with $p(\theta) \sim N(\theta|\mu, \lambda_0)$, then the posterior will be a product of two Normal distributions and hence $p(x)$ will be normal,

making $s(x)$ a linear combination of x and the prior mean. In this case

$$E(\theta|x) = (n\lambda + \lambda_0)^{-1}(n\lambda x + \lambda_0\mu) \quad (5.11)$$

2. Assume $p(\theta)$ to be a Student-t distribution with $p(\theta) \sim St(\theta|\mu, \lambda_0, \alpha)$. In this case

$$E(\theta|x) = x - \frac{(\alpha + 1)(x - \mu)}{n\lambda[\alpha\lambda^{-1} + (x - \mu)^2]} \quad (5.12)$$

(Pericchi and Smith, 1992)

3. Lastly, assume $p(\theta)$ to be a double-exponential distribution where

$$p(\theta) = \frac{1}{v\sqrt{2}} \exp\left(-\frac{\sqrt{2}}{v}|\theta - \mu|\right) \quad (5.13)$$

for some $v > 0, u \in \mathbb{R}$. Pericchi and Smith (1992) proved that if $b = n^{-1}v^{-1}\lambda^{-1}\sqrt{2}$,

$$E(\theta|x) = w(x)(x + b) + [1 - w(x)](x - b), \quad (5.14)$$

where $w(x)$ is a weight function, $0 \leq w(x) \leq 1$, so that

$$x - b \leq E(\theta|x) \leq x + b \quad (5.15)$$

By examining the three forms for $E(\theta|x)$ we can see the qualitative differences. In case of the Normal distribution, the posterior mean is unbounded in $x - \mu$. In the case of the Student-t, for very small $x - \mu$ the posterior mean is approximately linear in $x - \mu$, and for $x - \mu$ very large the posterior mean approaches x . In the case of the double exponential, the posterior mean is bounded, with limits equal to $x \pm c$, where c is a constant. We therefore have to be very aware of our prior choice in our model specification.

Clearly the way in which a prior is selected reflects a subjective opinion concerning the distribution of the parameter in question, prior to observing the data. This subjective selection may vary from one person to another, and hence make it difficult to justify our choice of a specific prior. Objective priors assume no subjective prior information and are derived from the assumed probability density function of the data, definitions of these two priors are given below (Beirlant et al., 2004):

Definition 5.1. Jeffreys' prior

Jeffreys' prior Jeffreys (1961) is defined as

$$\mathbf{J}(\boldsymbol{\theta}) \propto \sqrt{\mathbf{I}(\boldsymbol{\theta})}$$

where $I(\theta)$ is the Fisher's information matrix with (i, j) -th element

$$I_{ij}(\boldsymbol{\theta}) = E \left\{ -\frac{\partial^2 \log f(Y|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right\}, \quad i, j = 1, \dots, p \quad (5.16)$$

where p denotes the dimension of $\boldsymbol{\theta}$. ▲

The Jeffreys prior is invariant under one-to-one transformations and takes into account the dependence between the parameters. Bernardo and Smith (2009) noted that when applied to models appearing in extreme value methodology, Jeffreys' prior leads to the same confinements as the maximum likelihood approach.

The MDI prior was defined by Zellner (1997) to provide a maximal average data information on $\boldsymbol{\theta}$. The MDI prior is however not invariant under reparametrization, but is usually easier to implement.

Definition 5.2. Maximal Data Information (MDI) prior

The MDI prior is defined as

$$\pi(\boldsymbol{\theta}) \propto E\{\log f(Y|\boldsymbol{\theta})\}. \quad (5.17)$$

▲

In this study we assign an MDI prior to the parameter γ and a vague truncated Normal prior to δ . Beirlant et al. (2004) showed that for a Pareto distribution, the MDI prior is given by:

$$\pi(\gamma) \propto \frac{e^{-\gamma}}{\gamma}. \quad (5.18)$$

The EPD is a second order generalization of the simple Pareto distribution, and this justifies our choice of selecting the Pareto MDI prior for the EPD parameter γ . This prior can be approximated by a Gamma(α, λ) distribution as follows:

$$\pi(\gamma) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \gamma^{\alpha-1} e^{-\lambda\gamma} \propto \frac{e^{-\gamma}}{\gamma}, \text{ if } (\alpha, \lambda) \rightarrow (0, 1), \gamma > 0 \quad (5.19)$$

In OpenBUGS this approximation can be specified by choosing α close enough to zero and λ equal to 1.

The EPD parameters δ and τ have the following constraint $\delta > \max\{-1, \frac{1}{\tau}\}$ and $\tau > 1/\delta$ if and only if $\delta < 0$, the last constraint is not mentioned in Beirlant et al. (2009) but is implied. In order to enforce these constraints, we assign vague Normal priors, truncated with a lower bound. Also, in order to reflect our relative ignorance about the true parameter values we set a large prior variance of 10^6 . Only the Normal distribution can be truncated in OpenBUGS, making it a convenient choice.

Joint posterior

The joint posterior for the case where only δ and τ are estimated is given by

$$p(\gamma, \delta | x_1, \dots, x_n) \propto p(x_1, \dots, x_n | \gamma, \delta) p(\gamma) p(\delta) \quad (5.20)$$

and for the case where all three parameters (γ, δ, τ) are estimated it is given by

$$p(\gamma, \delta, \tau | x_1, \dots, x_n) \propto p(x_1, \dots, x_n | \gamma, \delta, \tau) p(\gamma) p(\delta) p(\tau) \quad (5.21)$$

where both $p(\delta)$ and $p(\tau)$ follow a Normal distribution truncated at the specified constraints.

5.3.5 Convergence diagnostics

In this section, we assess the mixing and convergence of the Gibbs, Metropolis and OpenBUGS Markov chains through the following measures:

- Traceplots: A traceplot that oscillates around a particular value and has small fluctuations indicates that the chain might have reached the correct distribution.
- Effective Sample Size (ESS) (Kass et al., 1998) ESS is defined as:

$$ESS = \frac{n}{1 + 2 \sum_{k=1}^{\infty} \rho_k(\theta)} \quad (5.22)$$

where n is the total sample size and $\rho_k(\theta)$ is the autocorrelation of lag k for θ . A low ESS indicates bad mixing of the Markov chain; we extend this measure to also look at the effective sampling speed (the rate at which the effective number of samples are generated).

- The potential scale reduction factor \hat{R} : At convergence $\hat{R} = 1$ and a high \hat{R} means we have to let our chains run longer in order to improve convergence to the stationary distribution (Gelman et al., 2014). The potential scale reduction factor is given by:

$$\hat{R} = \sqrt{\frac{\hat{Var}(\theta)}{W}} \quad (5.23)$$

where W is the mean of the variances of all chains.

It is not the aim of this section to make a generalization concerning the performance of MCMC methods described here, on a general class of Pareto-type distributions. Instead we use these diagnostic tools to help us make informed programming decisions such as the choice of the MCMC algorithm to employ and ways of making the simulating algorithm computationally efficient. A more general simulation study regarding Bayesian inference on heavy tailed distributions is conducted in Section 5.5.

A sample of size $n = 5000$ is generated from a Fréchet distribution with $EVI=0.5$ and stored into a file, `FrchetSample.csv`. We consider relative excesses with $k = 500$. We use the algorithms to estimate all parameters of the EPD (γ, δ, τ) where the true parameters are $\gamma = 0$, $\delta \rightarrow 0$ as $x \rightarrow \infty$ and $\tau = -1/0.5 = -2$.

“Keep in mind that OpenBUGS is an MCMC engine that makes use of various sampling methods including the Metropolis” (see, Section 5.3.6).

The simulation draws from each algorithm are stored as a 3-dimensional array. From these simulation draws we investigate mixing of the chains and convergence using traceplots and an **R** package called `monitor` which performs multiple chain diagnostics. More attention is placed on making inference about the EVI (γ). From Table 5.1 to Table 5.12 we present a 2-dimensional summary for the input samples, each row represents one parameter; the columns show the mean, standard deviation, quantiles, \hat{R} (potential scale reduction factor) and `n_eff` (estimate of ESS)

Simulation draws on the EVI (γ) are further then presented using traceplots to assess their speed of convergence.

Summary of results

Table 5.1: Gibbs sampler: 1000 iterations of 3 chains

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	\hat{R}
γ	0.54	0.01	0.08	0.41	0.49	0.52	0.58	0.74	46.00	1.08
δ	3.07	0.78	5.77	-0.58	-0.01	0.03	3.53	21.11	54.00	1.04
τ	-51.29	14.97	454.48	-382.94	-12.07	-0.51	-0.06	-0.01	922.00	1.00

Table 5.2: Metropolis algorithm: 1000 iterations of 3 chains

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	\hat{R}
γ	0.50	0.00	0.03	0.45	0.48	0.50	0.52	0.55	1496.00	1.00
δ	-0.01	0.00	0.00	-0.01	-0.01	-0.00	-0.00	-0.00	16.00	1.04
τ	-73.46	20.93	26.00	-107.11	-99.24	-77.54	-43.67	-32.23	2.00	6.00

Table 5.3: OpenBUGS: 1000 iterations of 3 chains

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	\hat{R}
γ	0.51	0.00	0.06	0.40	0.47	0.51	0.54	0.64	1500.00	1.00
δ	0.07	0.01	0.45	-0.77	-0.11	0.01	0.16	1.31	1500.00	1.00
τ	-0.93	0.02	0.94	-3.79	-1.22	-0.72	-0.24	-0.03	1500.00	1.00

From Table 5.1, the Gibbs sampler does not seem to converge well since $\hat{R} \neq 1$, however τ seems to converge ($\hat{R} = 1$), and has the highest ESS than all other parameters. Contrary to the Gibbs sampler, the Metropolis algorithm in Table 5.2 seems to have more convergence for γ than for all other parameters, in terms of both \hat{R} and ESS. From Table 5.3 OpenBUGS seems to have an overall better convergence of all parameters with $\hat{R} = 1$ and having an overall high ESS than the two coded algorithms. All estimators seem to result in the average simulation draws quite close to the true EVI.

Table 5.4: Gibbs sampler: 1000 iterations of 6 chains

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	\hat{R}
γ	0.53	0.01	0.07	0.39	0.49	0.52	0.57	0.72	74.00	1.04
δ	3.07	1.38	6.12	-0.79	-0.01	0.02	2.57	21.88	20.00	1.15
τ	-81.98	22.53	949.49	-465.13	-19.04	-0.66	-0.06	-0.01	1777.00	1.00

Table 5.5: Metropolis algorithm: 1000 iterations of 6 chains

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	\hat{R}
γ	0.49	0.00	0.03	0.44	0.47	0.49	0.52	0.55	96.00	1.05
δ	-0.02	0.02	0.09	-0.28	-0.01	-0.01	-0.00	-0.00	15.00	1.31
τ	-37.10	9.58	18.72	-67.71	-56.12	-33.38	-24.62	-0.43	4.00	3.77

Table 5.6: OpenBUGS: 1000 iterations of 6 chains

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	\hat{R}
γ	0.51	0.00	0.07	0.35	0.47	0.51	0.55	0.66	3000.00	1.00
δ	0.10	0.01	0.54	-0.87	-0.16	0.02	0.24	1.44	3000.00	1.00
τ	-1.00	0.04	2.04	-6.48	-1.00	-0.55	-0.23	-0.03	2879.00	1.00

When using more chains we can see from Table 5.4 that the Gibbs sampler is showing better convergence for τ than other parameters. From Table 5.5 we can see that convergence of the Metropolis algorithm worsens ($\hat{R} > 1$ for all parameters). From Table 5.6 OpenBUGS seems to maintain convergence of all parameters, notice also that the mean of simulated γ draws remains 0.51.

Table 5.7: Gibbs sampler: 5000 iterations, 3 chains

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	\hat{R}
γ	0.54	0.01	0.08	0.42	0.49	0.52	0.57	0.72	161.00	1.01
δ	2.39	0.30	5.26	-0.71	-0.01	0.02	1.69	20.47	303.00	1.00
τ	-87.46	36.04	1712.91	-391.23	-22.74	-0.86	-0.11	-0.01	2259.00	1.00

Table 5.8: Metropolis algorithm: 5000 iterations, 3 chains

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	\hat{R}
γ	0.50	0.00	0.03	0.45	0.48	0.50	0.52	0.55	6837.00	1.00
δ	-0.01	0.00	0.00	-0.01	-0.01	-0.00	-0.00	-0.00	225.00	1.02
τ	-88.23	12.40	23.81	-127.64	-105.48	-93.37	-70.41	-40.32	4.00	2.08

Table 5.9: OpenBUGS: 5000 iterations, 3 chains

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	\hat{R}
γ	0.52	0.00	0.07	0.36	0.48	0.51	0.55	0.68	7500.00	1.00
δ	0.22	0.01	0.74	-0.89	-0.09	0.02	0.36	2.35	7500.00	1.00
τ	-2.68	0.08	6.74	-24.36	-1.44	-0.50	-0.20	-0.03	7500.00	1.00

We then revert to using 3 chains but increasing the number of iterations. From Table 5.7, convergence of the Gibbs sampler improves with δ and τ having $\hat{R} = 1$. From Table 5.8, the Metropolis algorithm seems to be performing similarly as when we run 1000 iterations. From Table 5.9 OpenBUGS maintains similar performance with a slight change in the average of the simulated γ draws.

Table 5.10: Gibbs sampler: 5000 iterations, 6 chains

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	\hat{R}
γ	0.53	0.00	0.07	0.40	0.49	0.52	0.56	0.72	344.00	1.01
δ	2.76	0.28	5.76	-0.74	-0.01	0.01	1.99	21.28	428.00	1.02
τ	-179.20	100.82	7804.01	-447.54	-22.08	-0.74	-0.06	-0.01	5992.00	1.00

Table 5.11: Metropolis algorithm: 5000 iterations, 6 chains

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	\hat{R}
γ	0.50	0.00	0.03	0.45	0.48	0.50	0.52	0.55	14484.00	1.00
δ	-0.00	0.00	0.00	-0.01	-0.00	-0.00	-0.00	-0.00	11.00	1.18
τ	-170.25	43.16	77.66	-322.58	-225.92	-176.53	-104.02	-40.57	3.00	5.41

Table 5.12: OpenBUGS: 5000 iterations, 6 chains

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	\hat{R}
γ	0.51	0.00	0.08	0.34	0.47	0.51	0.56	0.69	14968.00	1.00
δ	0.22	0.01	0.78	-0.91	-0.15	0.03	0.42	2.36	15000.00	1.00
τ	-1.47	0.03	3.21	-12.24	-1.09	-0.46	-0.21	-0.03	14790.00	1.00

We further run more chains and more iterations. From Table 5.10 and Table 5.11 we can see that apart from the δ posterior estimates of the Gibbs Sampler, convergence of the Gibbs sampler and the Metropolis algorithm remain the same as when we used 3 chains. From Table 5.12 we can see that OpenBUGS continues to converge more than other considered algorithms, maintaining the average value of the simulated γ draws of 0.5 and with the average of the simulated τ draws being closer to the true value of $\tau = -2$.

We now consider computational efficiency in terms of computing time. This consideration will be very important in Section 5.5 when we estimate parameters of the EPD for $k = 1 : 1,000$ and over 1,000 sample i.e. running the MCMC algorithm over 1000,000 times.

Computation time for the EVI posterior samples

Table 5.13: 1000 iterations, 3 chains (time in seconds)

	N.eff (N)	Computing time (t) in seconds	Eff sampling speed (N/t)
Metropolis	1496.00	244.02	6.13
Gibbs	46.00	832.28	0.06
OpenBUGS	1500.00	10.96	136.86

Table 5.14: 1000 iterations, 6 chains (time in seconds)

	N.eff (N)	Computing time (t) in seconds	Eff sampling speed (N/t)
Metropolis	1926.00	372.31	5.17
Gibbs	74.00	1578.97	0.05
OpenBUGS	3000.00	17.48	171.62

Table 5.15: 5000 iterations, 3 chains (time in seconds)

	N.eff (N)	Computing time (t) in seconds	Eff sampling speed (N/t)
Metropolis	6837.00	1650.36	4.14
Gibbs	161.00	5012.48	0.03
OpenBUGS	7500.00	49.88	150.36

Table 5.16: 5000 iterations, 6 chains

	N.eff (N)	Computing time (t) in seconds	Eff sampling speed (N/t)
Metropolis	14484.00	5697.22	2.54
Gibbs	344.00	9954.44	0.03
OpenBUGS	14968.00	94.67	158.11

We compute the ESS and the computation time in seconds and use this to then calculate the effective sampling speed (the rate at which the algorithm generates

an effective number of samples). As one would expect, computation time of all MCMC algorithms increase when we increase the number of iterations of number of chains. This is also the case with the ESS. However the effective sampling speed of these MCMC algorithms unfold an interesting tendency worthy of note.

In Table 5.13 to Table 5.16 we note that the effective sampling speed of the Gibbs sampler remains the same regardless of the increase in number of iterations of number of chains used. However the effective sample speed of the Metropolis algorithm decreases while that of OpenBUGS slightly increases when we increase the number of iterations of the number of chains ran. Also, when we run 1000 iterations of 3 chains OpenBUGS is $244.02/10.96 \approx 22$ times faster than the Metropolis algorithm and $832.28/10.96 \approx 76$ times faster than the Gibbs sampler; when we run 5000 iterations of 6 chains, OpenBUGS is $5697.22/94.67 \approx 60$ times faster than the Metropolis algorithm and $9954.44/94.67 \approx 105$ times faster than the Gibbs sampler. We therefore conclude that OpenBUGS is more computationally efficient.

We further assess convergence using traceplots.

Traceplots

From Figure 5.2 The Gibbs sampler exhibits marginal mixing, we can see that the chain takes relatively small steps and seems to not traverse to the stationary distribution quickly. The Metropolis algorithm seems to be mixing well, although the first few observations might have to be discarded in order to make a more accurate inference. OpenBUGS seems to vary more than the Metropolis algorithm, but eventually also mixes well.

From Figure 5.3, the Gibbs sampler still exhibits some marginal mixing when we increase the number of chains. The Metropolis algorithm also exhibiting some marginal mixing but eventually seems to converge to the stationary distribution. We would have to discard more initial values than in the case where only 3 chains were ran.

From Figure 5.4, the bad mixing of the Gibbs sampler is more apparent, while the mixing of the Metropolis algorithm seems to be converging faster to the stationary distribution. We would thus only discard a few initial values in order to make an inference. OpenBUGS seems to have a good mixing, with no need to discard initial values. A similar conclusion can be made for both Figure 5.4 and Figure 5.5

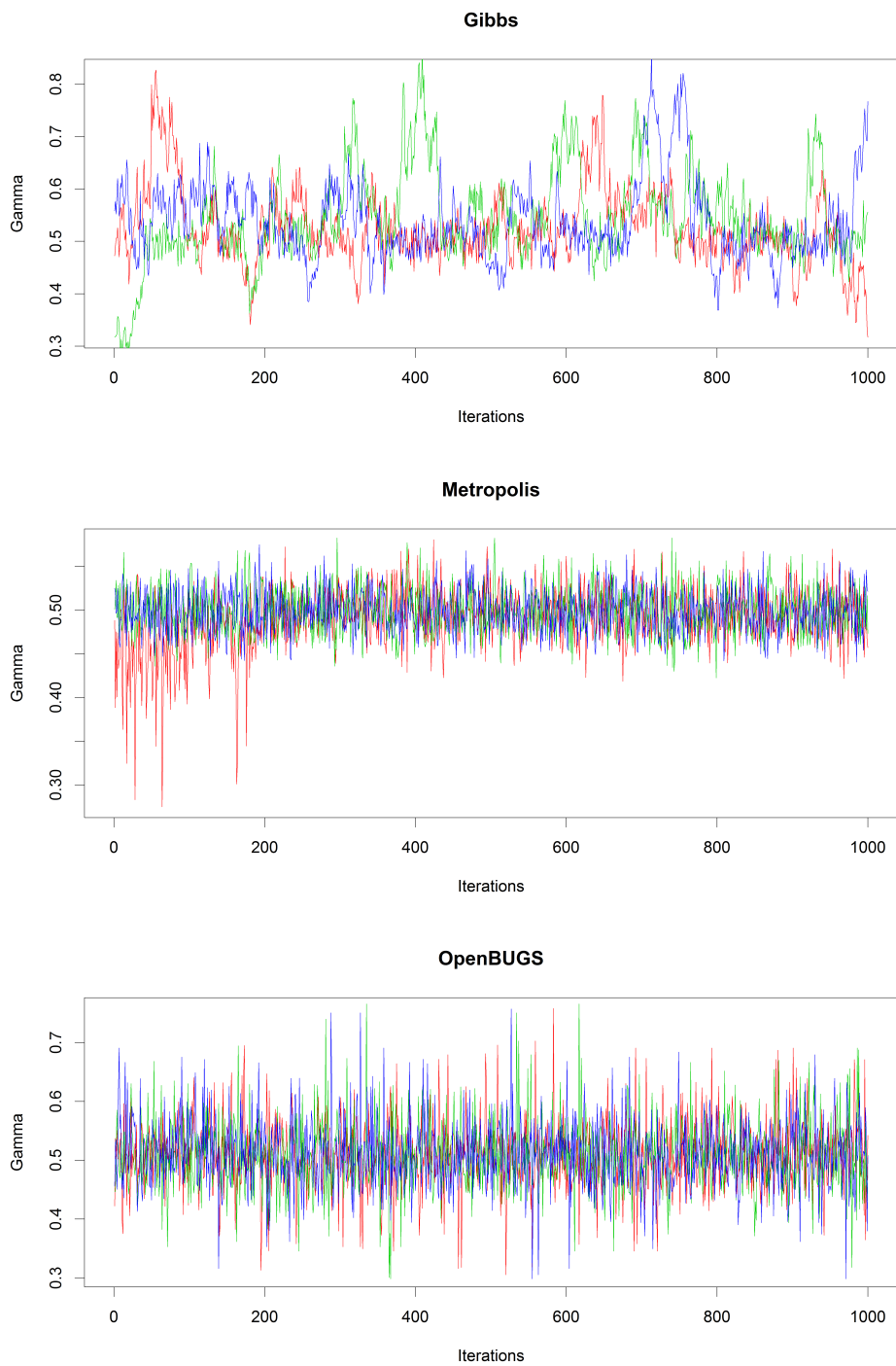


Figure 5.2: Traceplot of the EVI posterior simulations for 1000 iterations of 3 chains

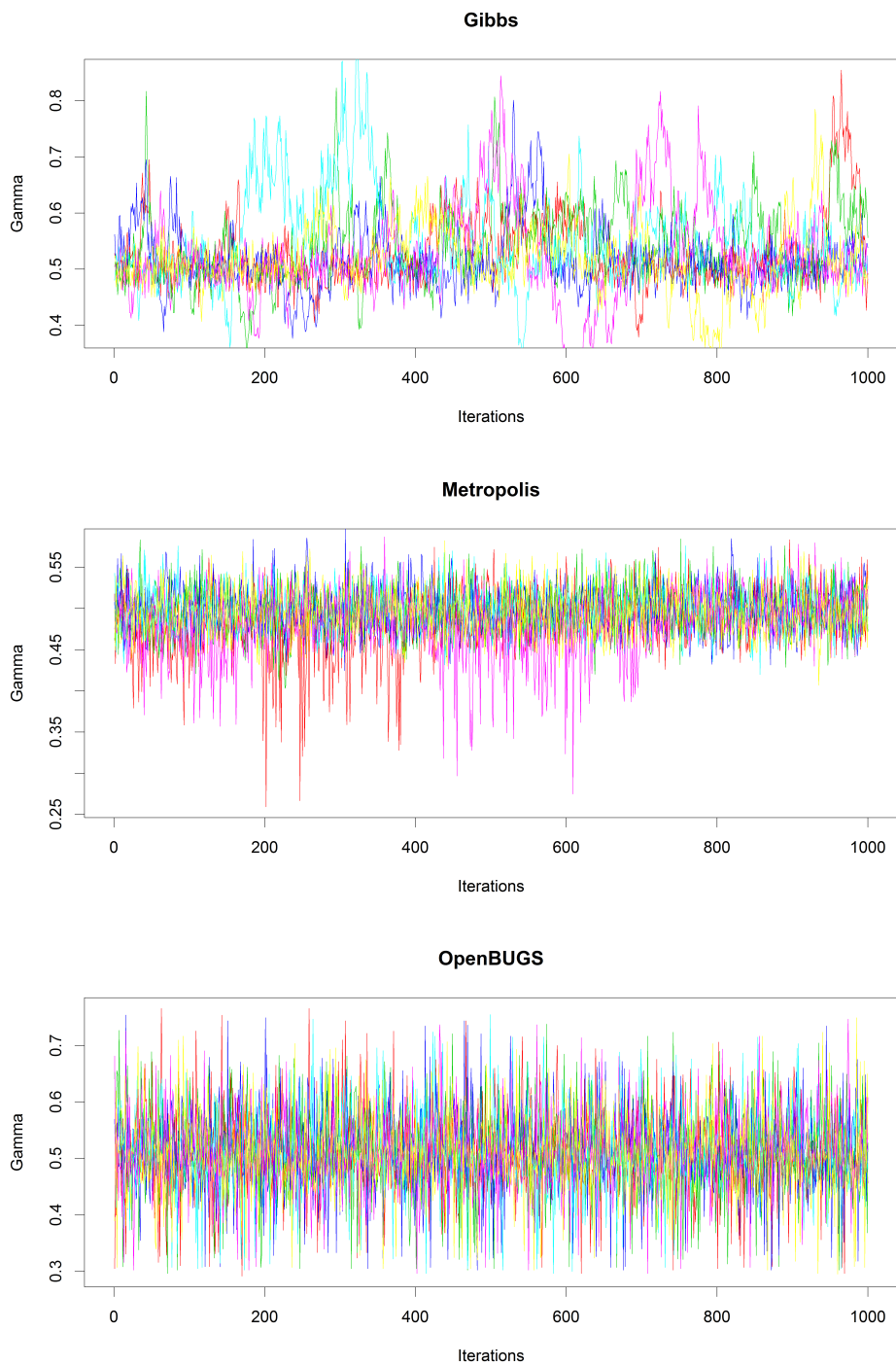


Figure 5.3: Traceplot of the EVI posterior simulations for 1000 iterations of 6 chains

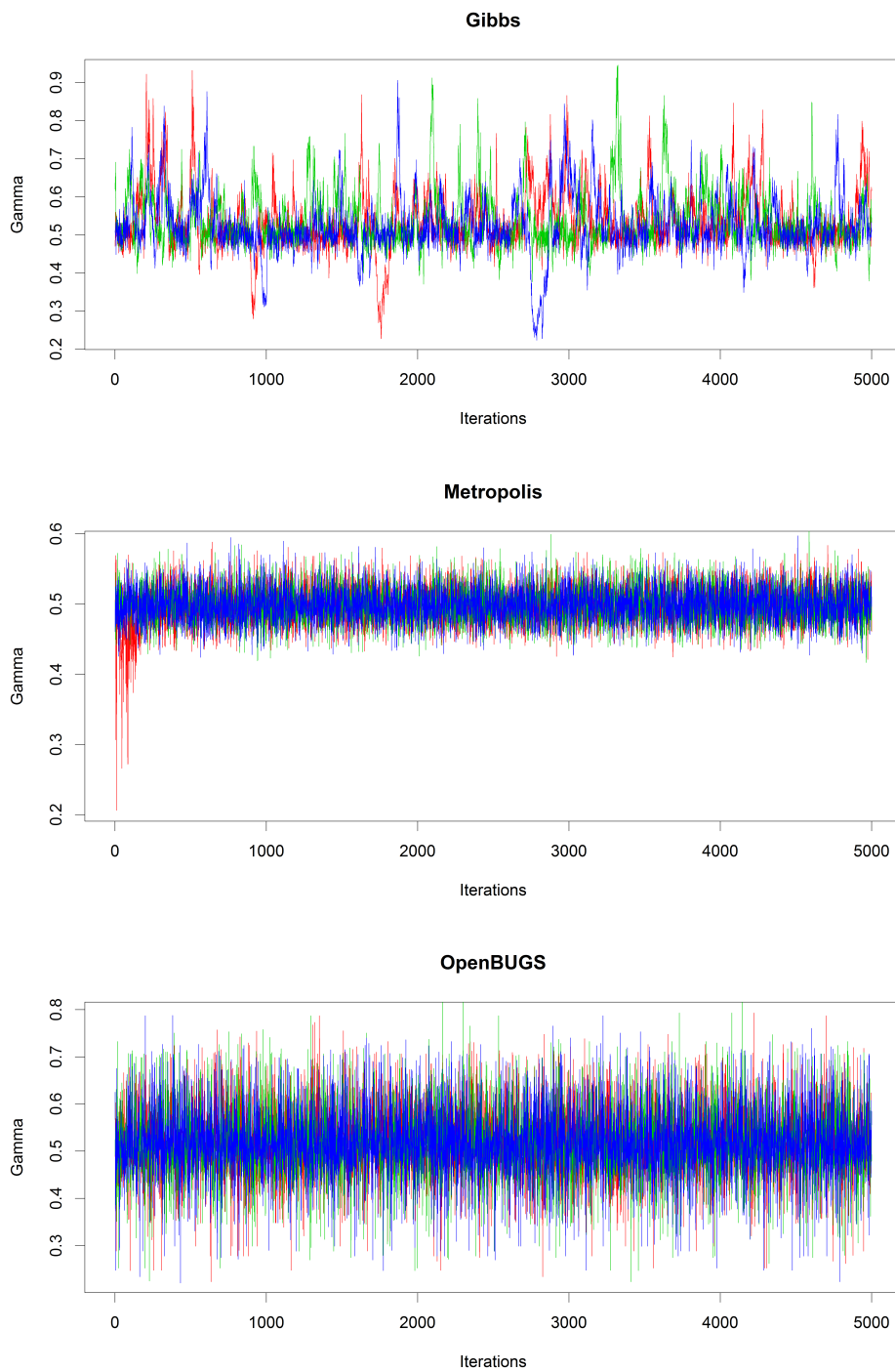


Figure 5.4: Traceplot of the EVI posterior simulations for 5000 iterations of 3 chains

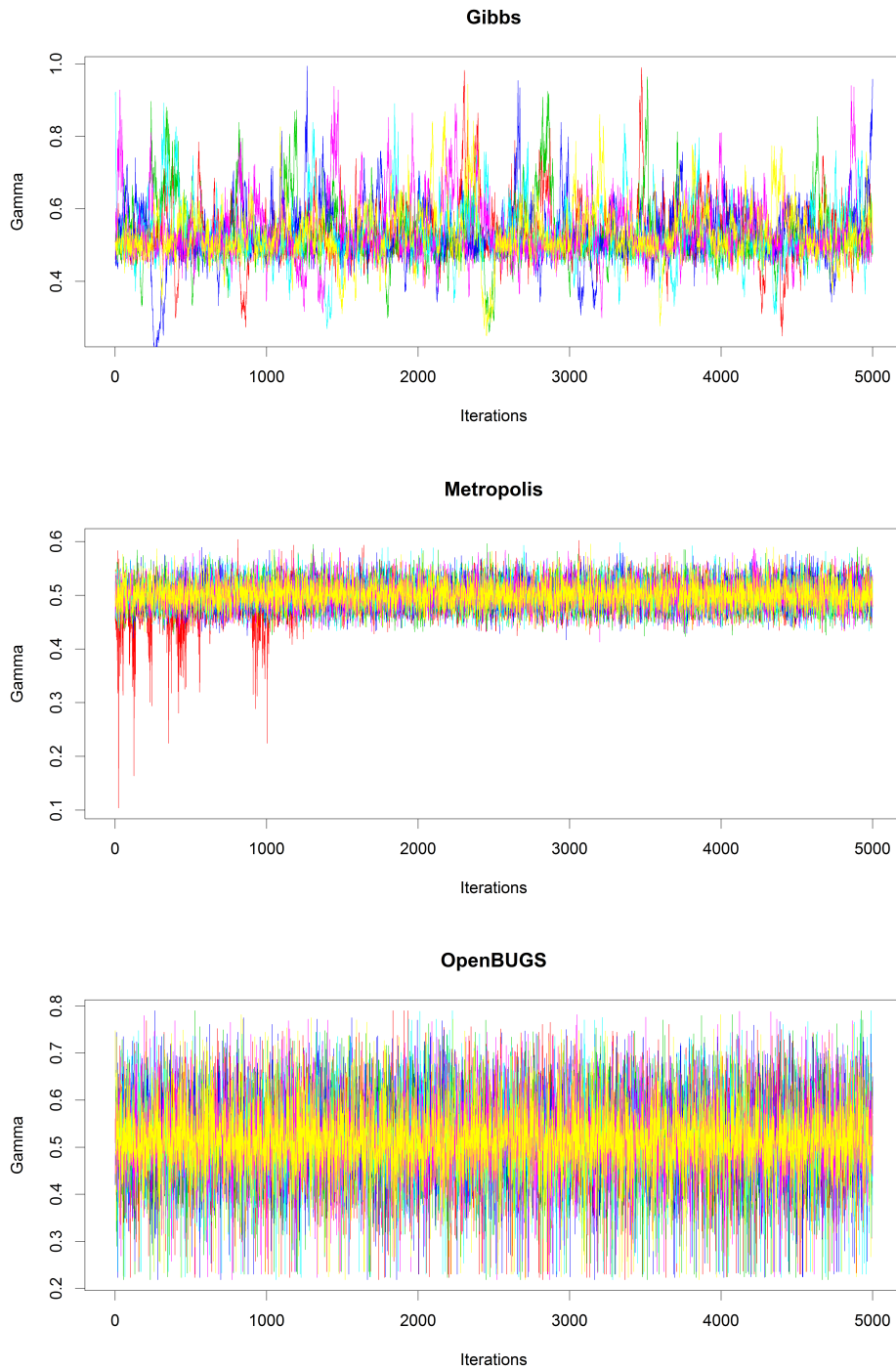


Figure 5.5: Traceplot of the EVI posterior simulations for 5000 iterations of 6 chains

As we can see from our convergence diagnostics that OpenBUGS seems to have satisfactory performance in terms of mixing and convergence. Moreover OpenBUGS is computationally efficient. Throughout this thesis we use OpenBUGS when making a Bayesian inference on parameters of the EPD.

5.3.6 OpenBUGS Efficient Sampling

BUGS does not generally perform Gibbs sampling, since it employs the most appropriate sampling scheme for each stochastic node, through a system known as BUGS ‘Expert system’. The system decides which of the MCMC algorithms it should use for a specific full conditional distribution, since the target densities of the parameters are not standard densities. In OpenBUGS however, the sampling scheme still has the Gibbs theme, since it traverses nodes in the graphical model and considers their full conditionals as target distributions, even if they cannot be precisely sampled (Lunn et al., 2012). Below is a sampling method hierarchy used by WinBUGS, also incorporated in OpenBUGS

Table 5.17: Sampling method hierarchy used by WinBUGS. Each method is only used if no previous method in the hierarchy is appropriate (Lunn et al., 2012).

Target full conditional	Sampling method
Discrete	Inversion of cumulative distribution function
Closed form (conjugate)	Direct sampling using standard algorithms
Log-concave	Derivative-free adaptive rejection sampling (Gilks and Wild, 1992)
Restricted range	Slice sampling (Neal, 2003)
Unrestricted range	Metropolis–Hastings (Hastings, 1970; Metropolis et al., 1953)

5.4 Asymptotic theory

A theoretical prerequisite for many estimators proposed in EVT literature, is the derivation of their asymptotic properties in order to assess performance. Within the POT framework, the excesses X_1, X_2, \dots, X_n over some threshold t are assumed to follow a parametric limiting distribution with some parameters (say θ). In making inference on the model’s parameters, the Maximum likelihood estimation method, assumes some asymptotic simplification in order to obtain analytic derivations of the parameters. This simplification neglects any uncertainty about the fixed parameter values since we cannot make any direct probability statements about these unknown parameters. In Bayesian statistics however, these parameters are treated as random variables having some probability distribution. Also unlike the MLE/Frequentist approach, which relies on asymptotic theory to

assess how well a model performs in the long run over an infinite number of hypothetical repetitions, Bernardo and Smith (2009) showed, that under suitable regularity conditions, as the sample size increases, the posterior distribution of the continuous parameter of interest $\boldsymbol{\theta}$ actually converges to a Normal distribution $N(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}_n, H(\hat{\boldsymbol{\theta}}_n))$ with precision matrix:

$$H(\hat{\boldsymbol{\theta}}_n) = \left(-\frac{\partial^2 \log p(x|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n} \quad (5.24)$$

$H(\hat{\boldsymbol{\theta}})$ is also known as the Hessian matrix which measures the local curvature of the log-likelihood function at its maximum, $\hat{\boldsymbol{\theta}}$ often called the observed information matrix. Similarly the large sample behavior of the ML estimate $\hat{\boldsymbol{\theta}}$ is also asymptotically Normal distributed by $N(\hat{\boldsymbol{\theta}}_n|\boldsymbol{\theta}, nI(\boldsymbol{\theta}))$, with precision matrix whose general element is:

$$(I(\boldsymbol{\theta}))_{ij} = \int p(x|\boldsymbol{\theta}) \left(-\frac{\partial^2 \log p(x|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right) dx \quad (5.25)$$

$I(\boldsymbol{\theta})$ is called the Fisher information matrix. Furthermore drawing from the work of Lindley (1958), Bernardo and Smith (2009) showed, by instituting the convergence of $H(\hat{\boldsymbol{\theta}})$ to $nI(\boldsymbol{\theta})$ as $n \rightarrow \infty$ and using the fact that the sampling distribution of $\hat{\boldsymbol{\theta}}$ becomes a location model for $\boldsymbol{\theta}$ for n large, that the reference posterior distribution for $\boldsymbol{\theta}$ and the asymptotic fiducial distribution of $\boldsymbol{\theta}$, based on the sampling distribution of $\hat{\boldsymbol{\theta}}_n$, are then asymptotically equivalent. In addition, the ML estimator of $\boldsymbol{\theta}$ and the asymptotic confidence intervals based on $\hat{\boldsymbol{\theta}}$ will be numerically identical to both the mode (or mean) and the HPD intervals, based on the applicable reference posterior distribution. Hence, very few differences exist between Bayesian and Frequentist inferential statements based on asymptotic properties. It is therefore theoretically unnecessary to derive asymptotics in Bayesian inference.

5.5 Simulation study: MLE vs Bayes

In this section, we perform a Monte Carlo simulation, by generating 1000 samples, each of size $n = 200$ and $n = 500$ from: the Fréchet, Burr and Student-t distribution with $\text{EVI} = 1$ and $\text{EVI} = 0.5$ respectively. For each distribution and each estimation method (Bayes and MLE), we compute Monte Carlo estimates of the EVI and the corresponding MSE by averaging out over the 1000 samples. For the Bayesian approach in OpenBUGS, we take 6000 draws of each parameter and discard the first 500.

5.5.1 Case 1: Estimating only two parameters

In this section, the second order parameter of the EPD ρ is estimated externally using the method discussed in Section 3.3.1 and fixed for every simulation. We do this for both the EPD Bayesian estimator and the EPD ML estimator.

Fréchet Distribution

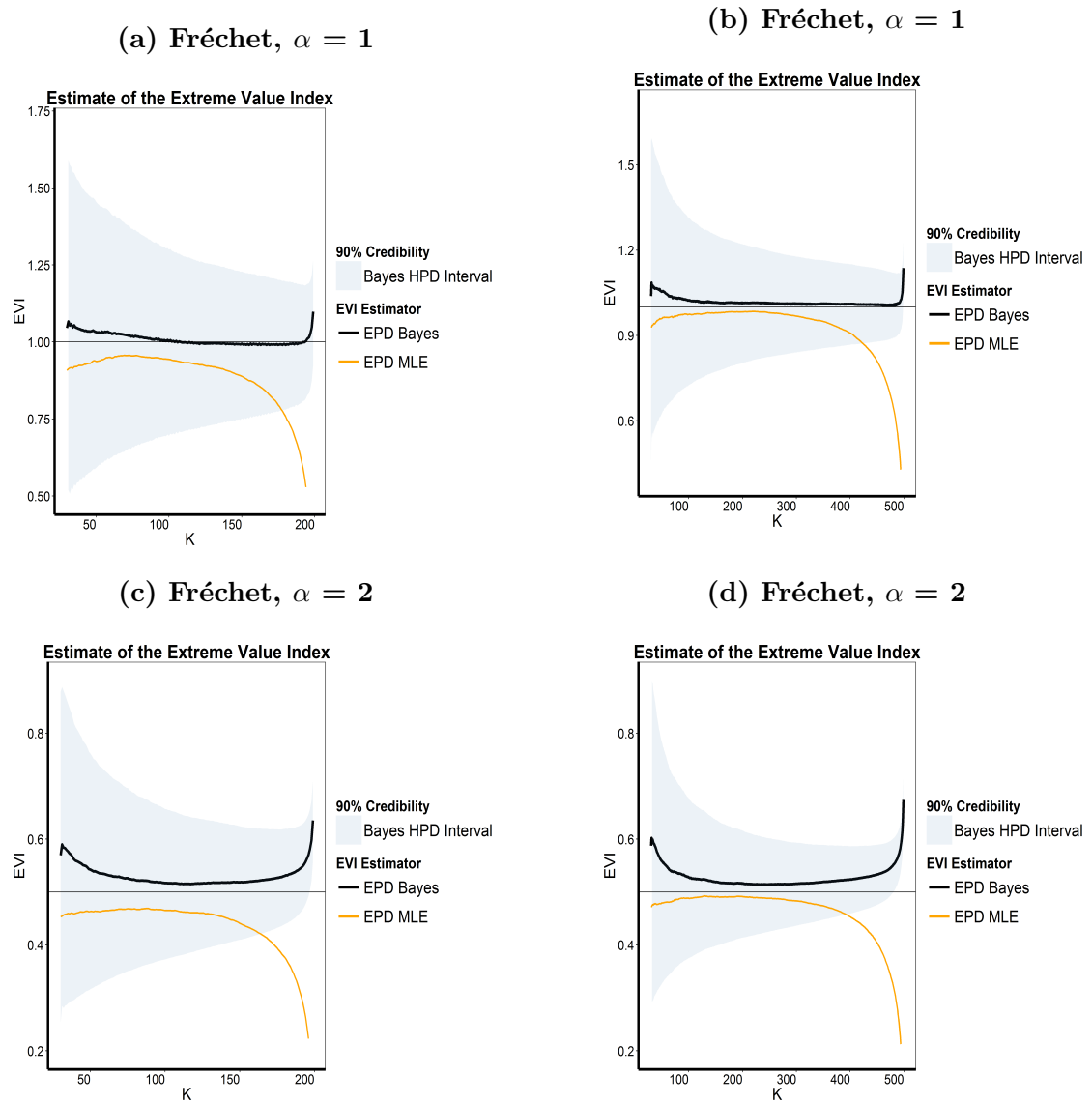


Figure 5.6: EVI estimates of a Fréchet sample of size $n=200$ (left), and size $n=500$ (right), and EVI of 1 (top) and 0.5 (bottom), 90 % HPD (shaded fill)

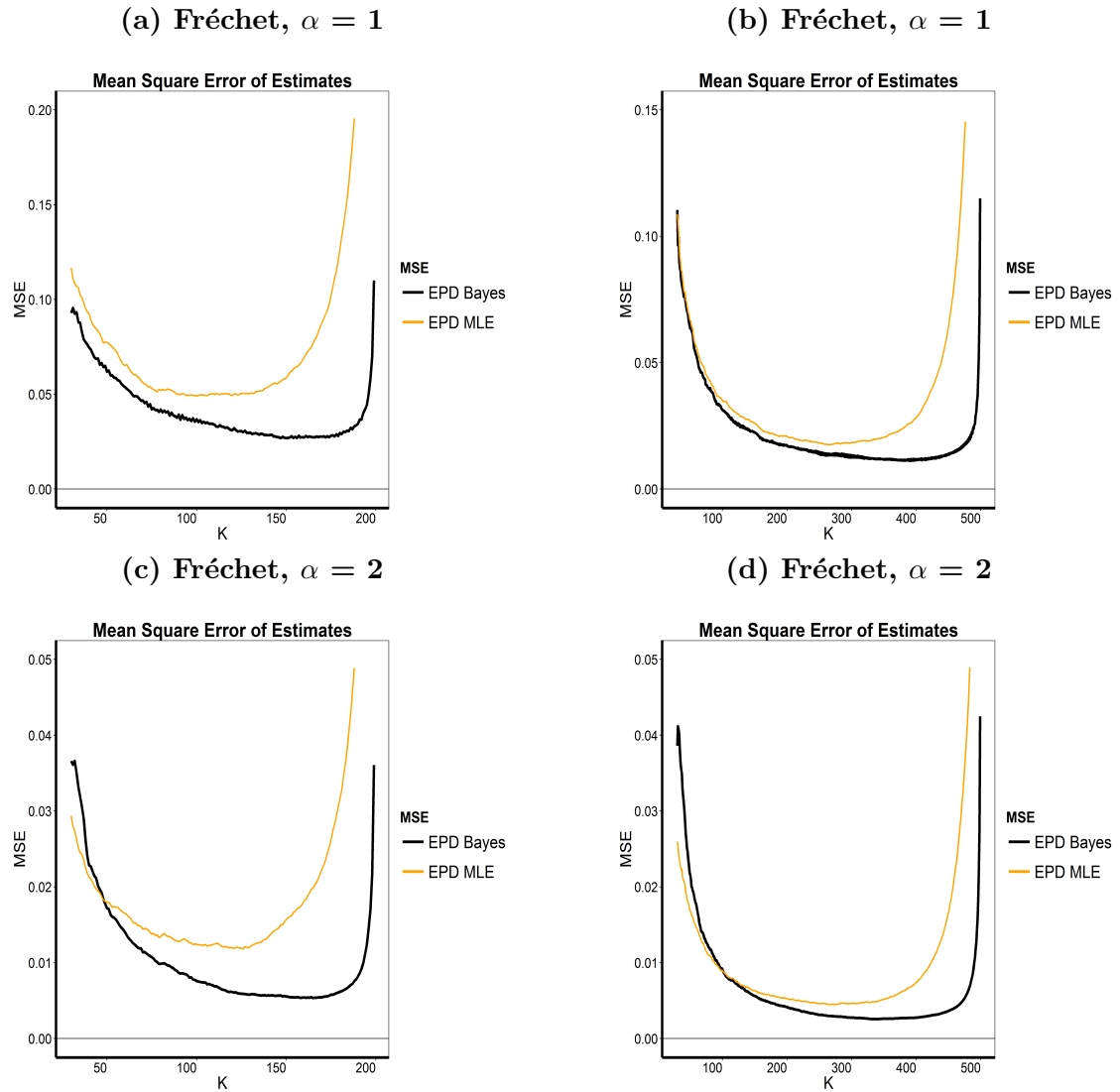


Figure 5.7: MSE of EVI estimates for a **Fréchet** sample of size $n=200$ (left), and size $n=500$ (right), and **EVI** of 1 (top) and 0.5 (bottom)

It is clear from Figure 5.6 and Figure 5.7 that the Bayesian estimates are performing much better than the ML estimates, for both smaller and larger samples. The asymptotic bias in Figure 5.6 and the MSE in Figure 5.7 for the Bayesian estimates is smaller than the ML estimates over a larger range of k values.

Burr Distribution

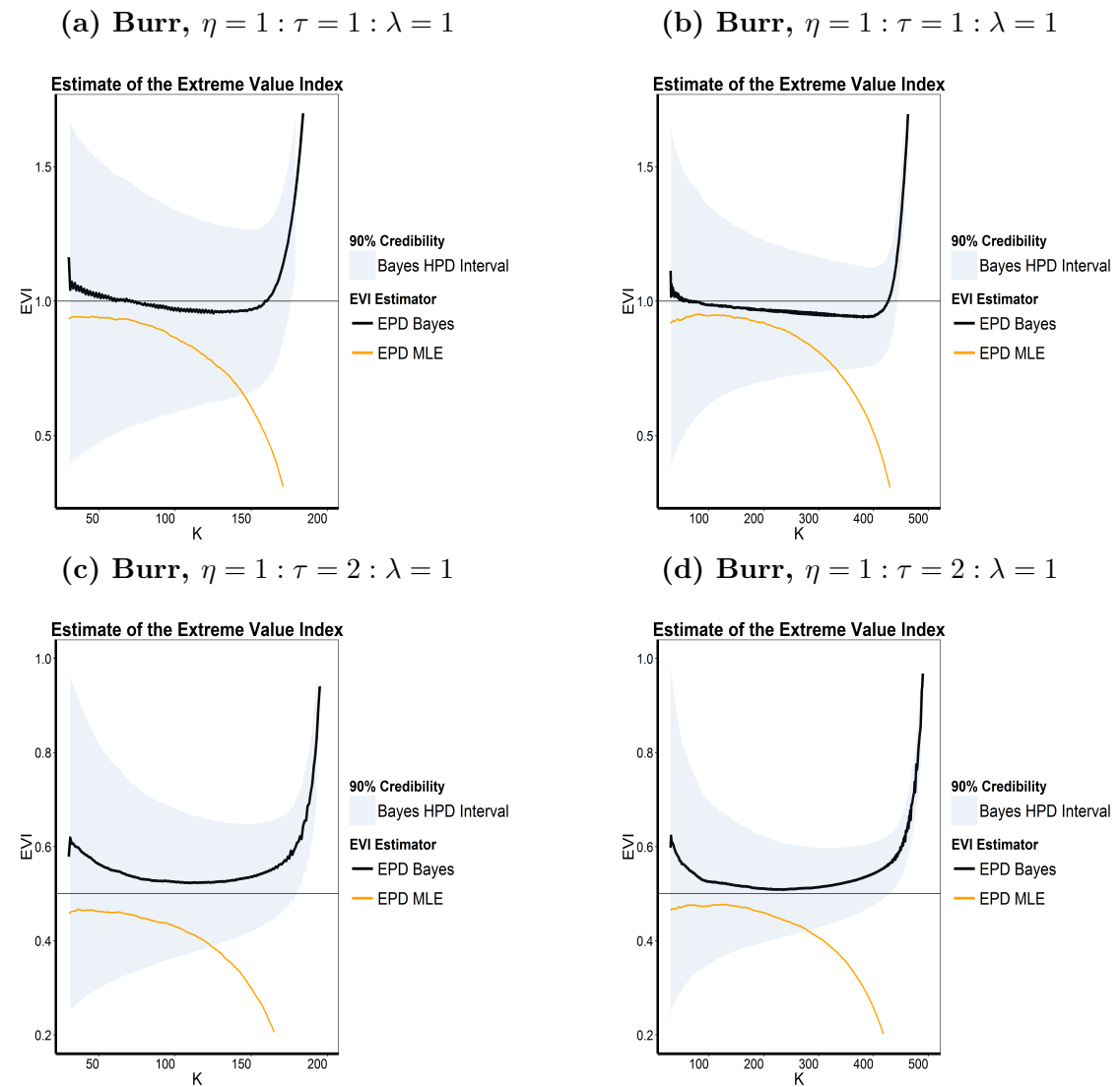


Figure 5.8: EVI estimates of a Burr sample of size $n=200$ (left), and size $n=500$ (right), and EVI of 1 (top) and 0.5 (bottom), 90 % HPD (shaded fill)

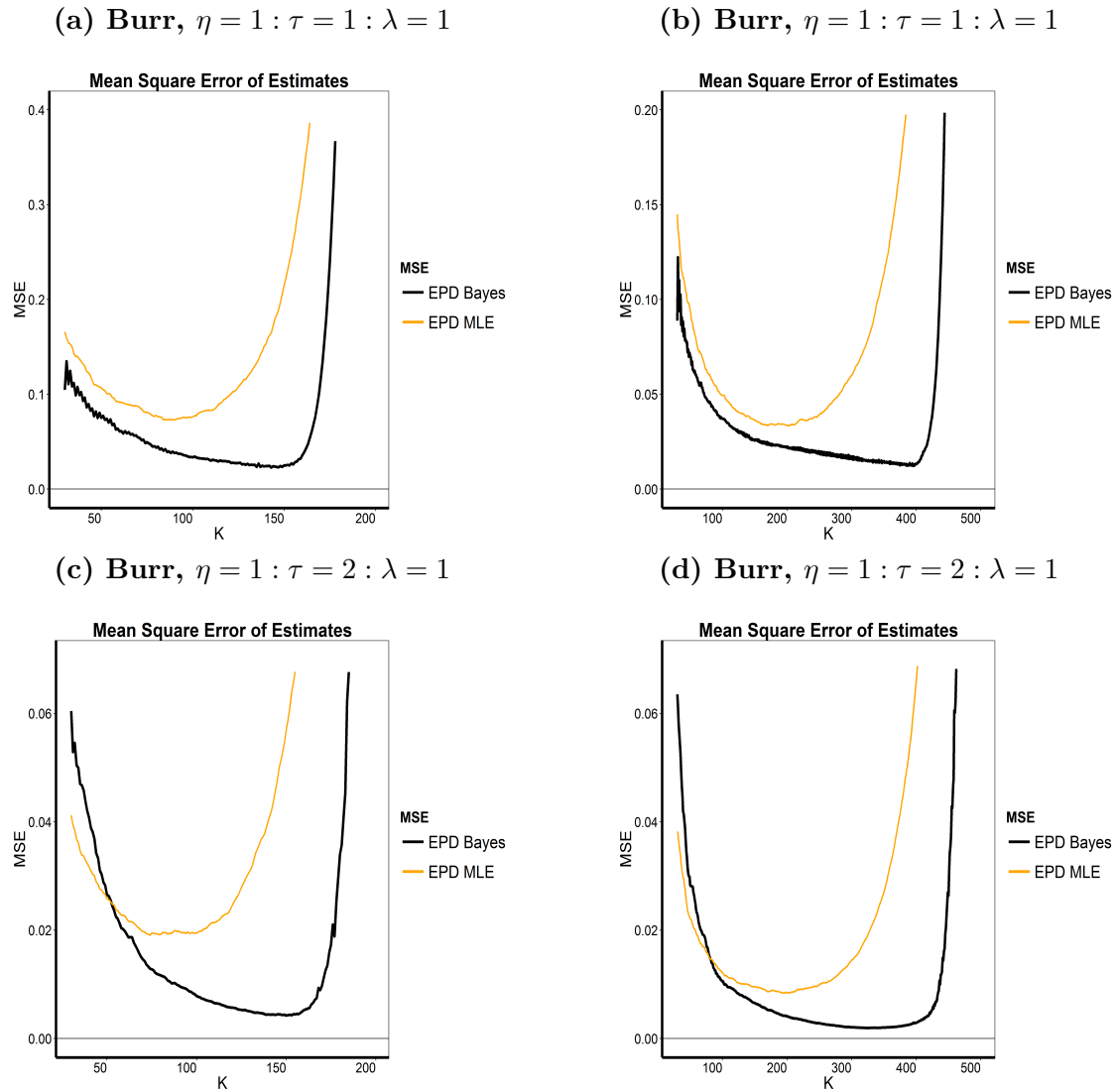


Figure 5.9: MSE of EVI estimates for a **Burr** sample of size $n=200$ (left), and size $n=500$ (right), and **EVI** of 1 (top) and 0.5 (bottom)

From Figure 5.8 and Figure 5.9 we can see that the Bayesian estimates are performing much better than the ML estimates in all cases. Note also that for the Bayesian estimate there is some indication of large metropolis jumps for cases where the $\text{EVI}=1$, the cycles are however smoothed out for larger samples. The Burr distribution is generally slightly heavier-tailed than the Fréchet, and as can be seen from Figure 5.9, the MSE of the Bayesian estimate performs even better for Burr samples than for Fréchet samples. The asymptotic bias in Figure 5.8 and the MSE in Figure 5.9 for the Bayesian estimates are smaller than the ML estimates over a larger range of k values.

Student-t Distribution

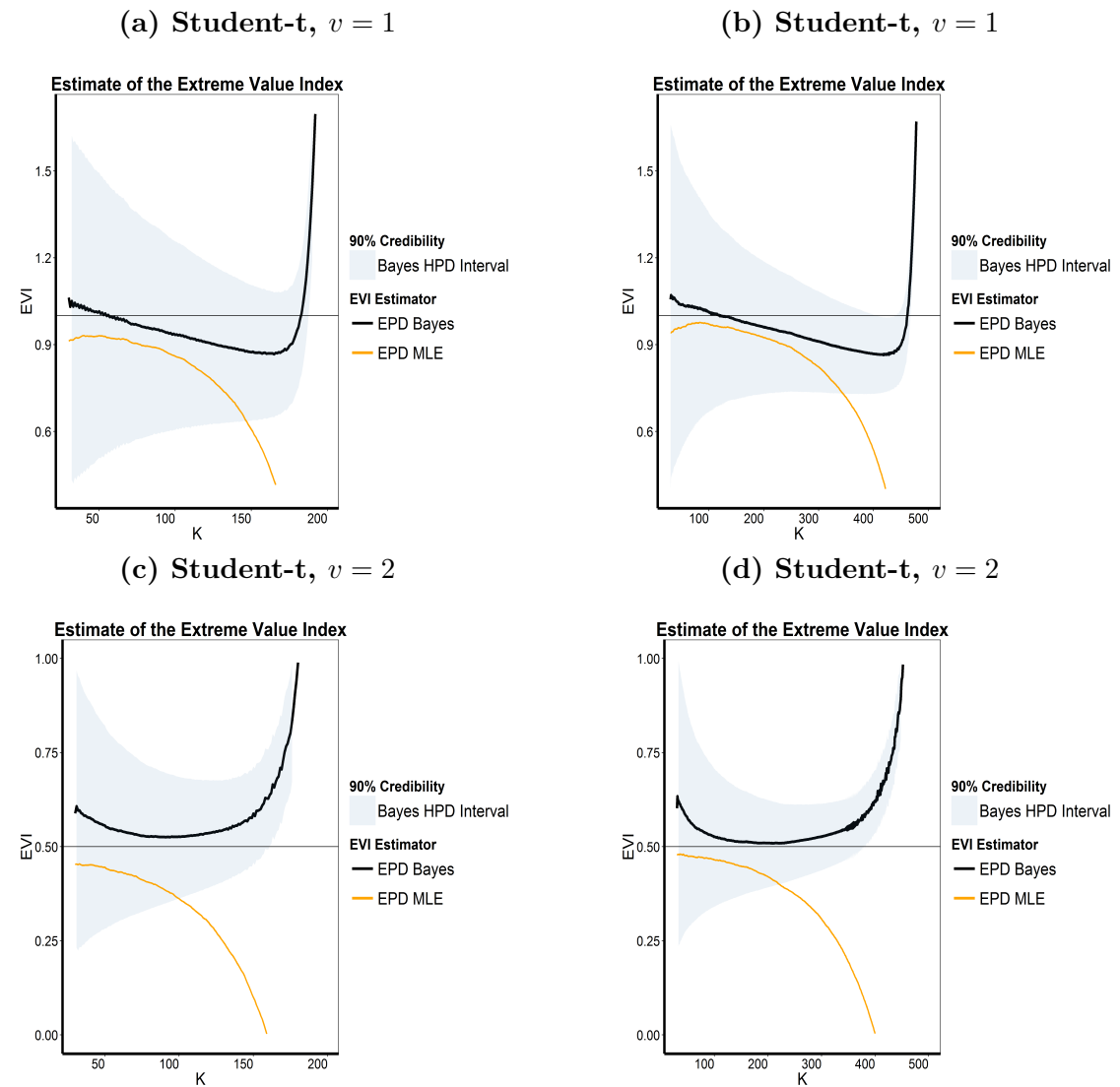


Figure 5.10: EVI estimates of a Student-t sample of size $n=200$ (left), and size $n=500$ (right), and EVI of 1 (top) and 0.5 (bottom), 90 % HPD (shaded fill)

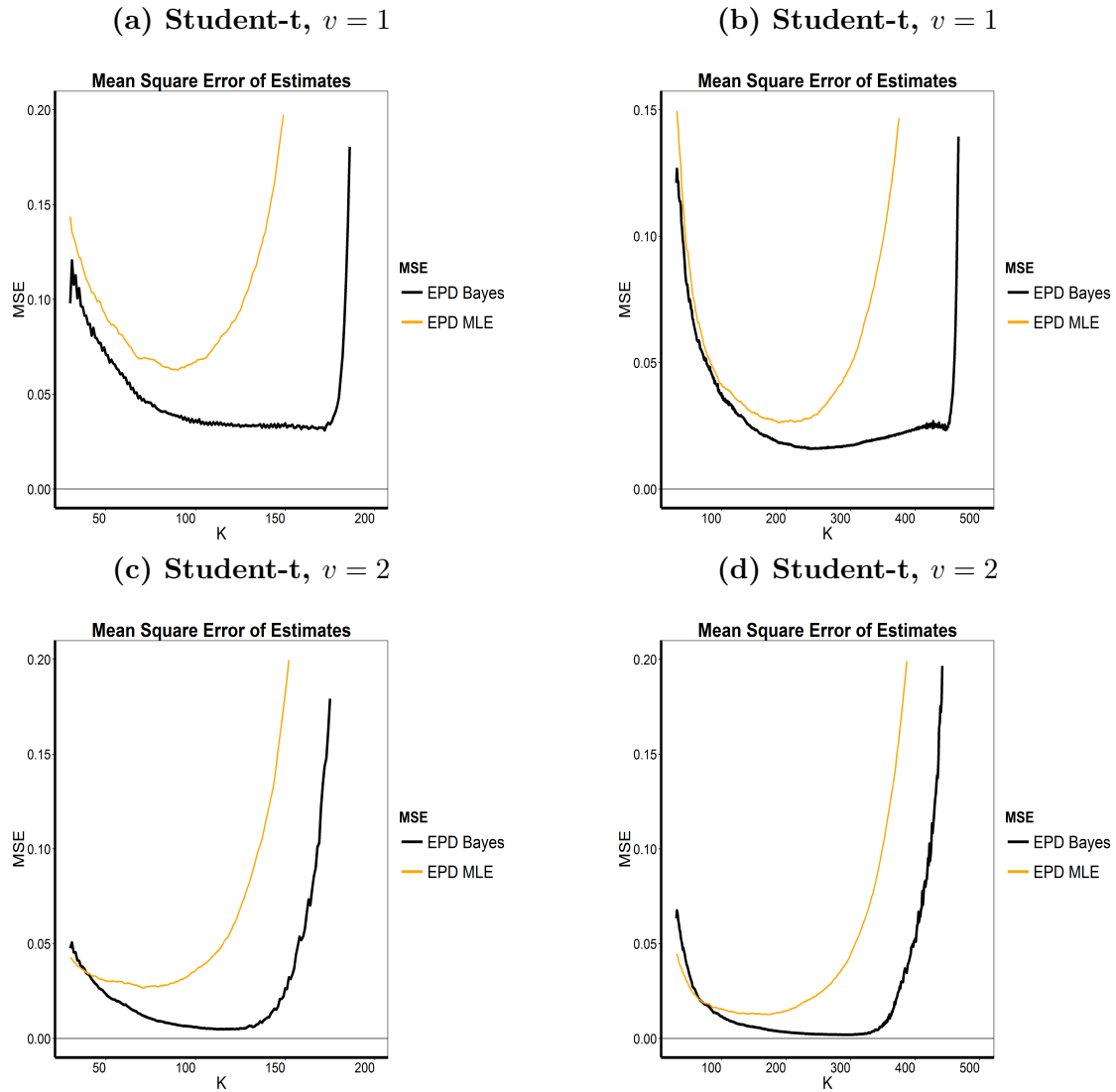


Figure 5.11: MSE of EVI estimates for a **Student-t** sample of size $n=200$ (left), and size $n=500$ (right), and **EVI** of 1 (top) and 0.5 (bottom)

From Figure 5.10 and Figure 7.29 we can see that the Bayesian estimates are out-performing those of the MLE, in all cases. In the case of the Student-t distribution, the out-performance is far more pronounced, this could be because the Student-t distribution is more heavier tailed than the Fréchet and the Burr distribution. The asymptotic bias in Figure 5.10 and the MSE in Figure 7.29 for the Bayesian estimates are smaller than the ML estimates over a larger range of k values.

5.5.2 Case 2: Estimating all three parameters

In this section, all the parameters of the EPD are estimated simultaneously using the ML method of estimation as discussed in Chapter 3 and the Bayesian method

of estimation.

Fréchet Distribution

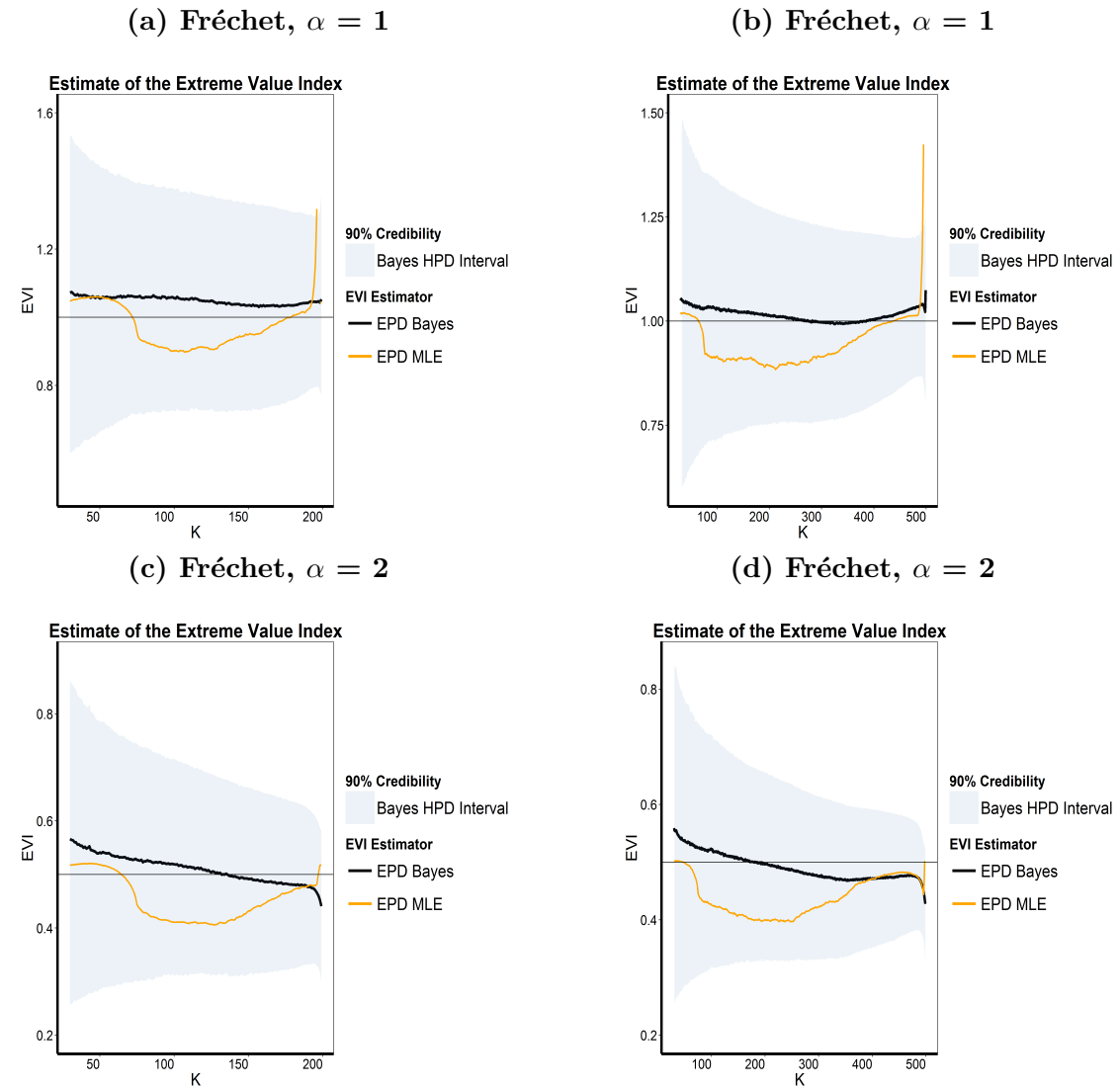


Figure 5.12: EVI estimates of a **Fréchet** sample of size $n=200$ (left), and size $n=500$ (right), and **EVI** of 1 (top) and 0.5 (bottom), 90 % HPD (shaded fill)

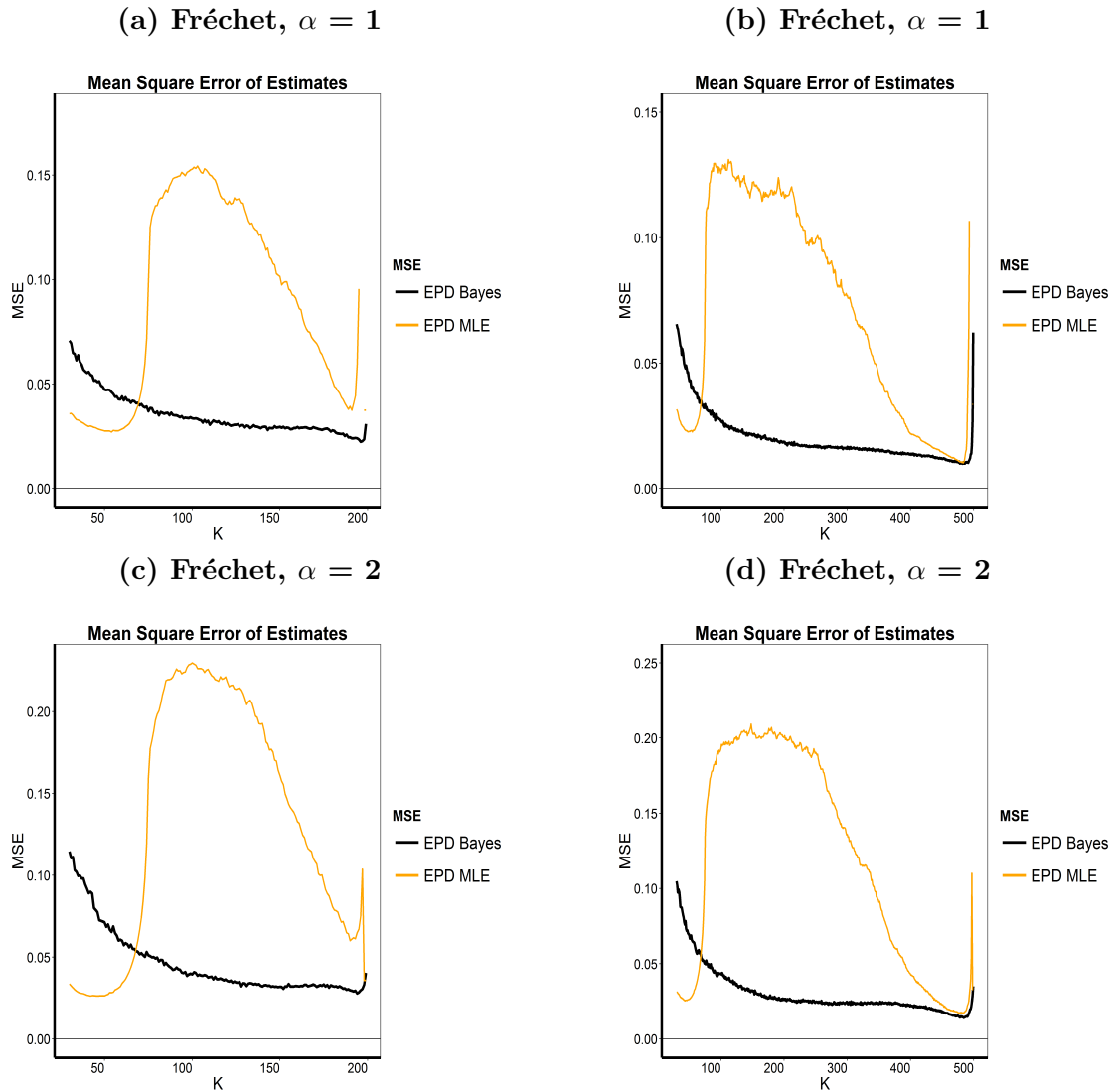


Figure 5.13: MSE of EVI estimates for a **Fréchet** sample of size $n=200$ (left), and size $n=500$ (right), and **EVI** of 1 (top) and 0.5 (bottom)

It is clear from Figure 5.12 and Figure 5.13 that the Bayesian estimates are performing much better than the ML estimates, for both smaller and larger samples. The asymptotic bias in Figure 5.12 and the MSE in Figure 5.13 for the Bayesian estimates are smaller than the ML estimates over a larger range of k values.

Burr Distribution

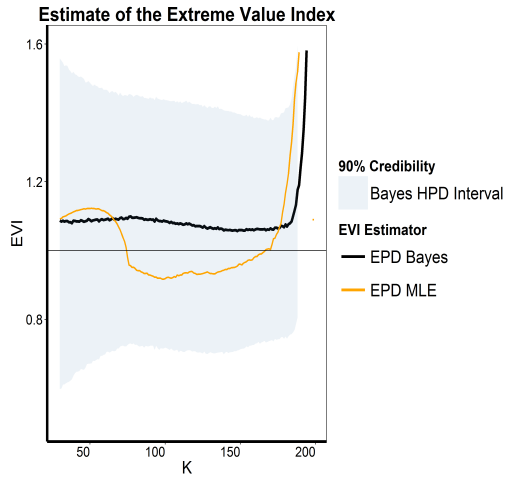
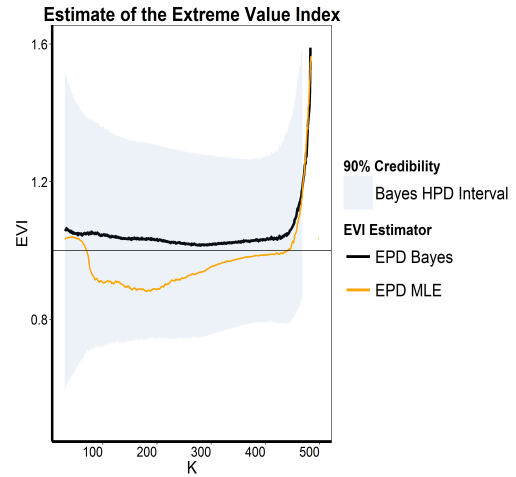
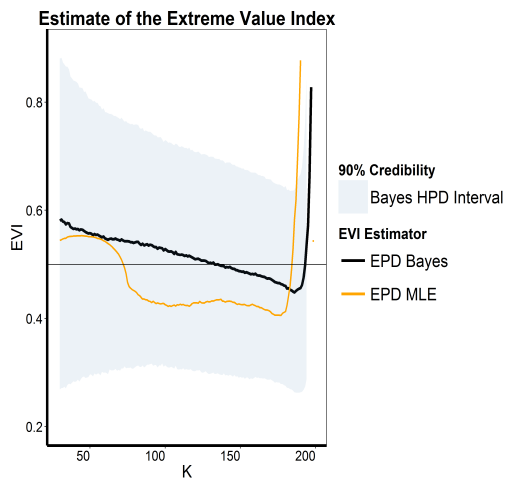
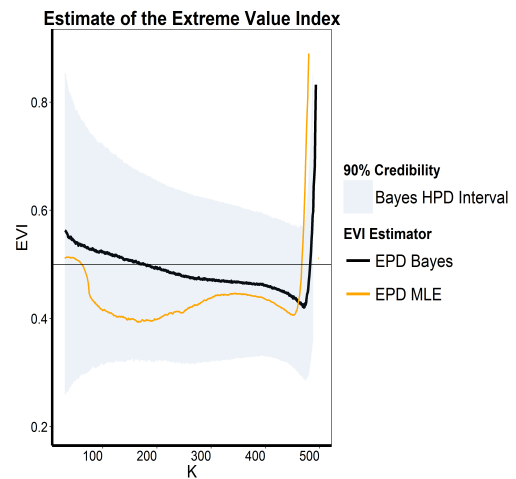
(a) Burr, $\eta = 1 : \tau = 1 : \lambda = 1$ (b) Burr, $\eta = 1 : \tau = 1 : \lambda = 1$ (c) Burr, $\eta = 1 : \tau = 2 : \lambda = 1$ (d) Burr, $\eta = 1 : \tau = 2 : \lambda = 1$ 

Figure 5.14: EVI estimates of a Burr sample of size $n=200$ (left), and size $n=500$ (right), and EVI of 1 (top) and 0.5 (bottom), 90 % HPD (shaded fill)

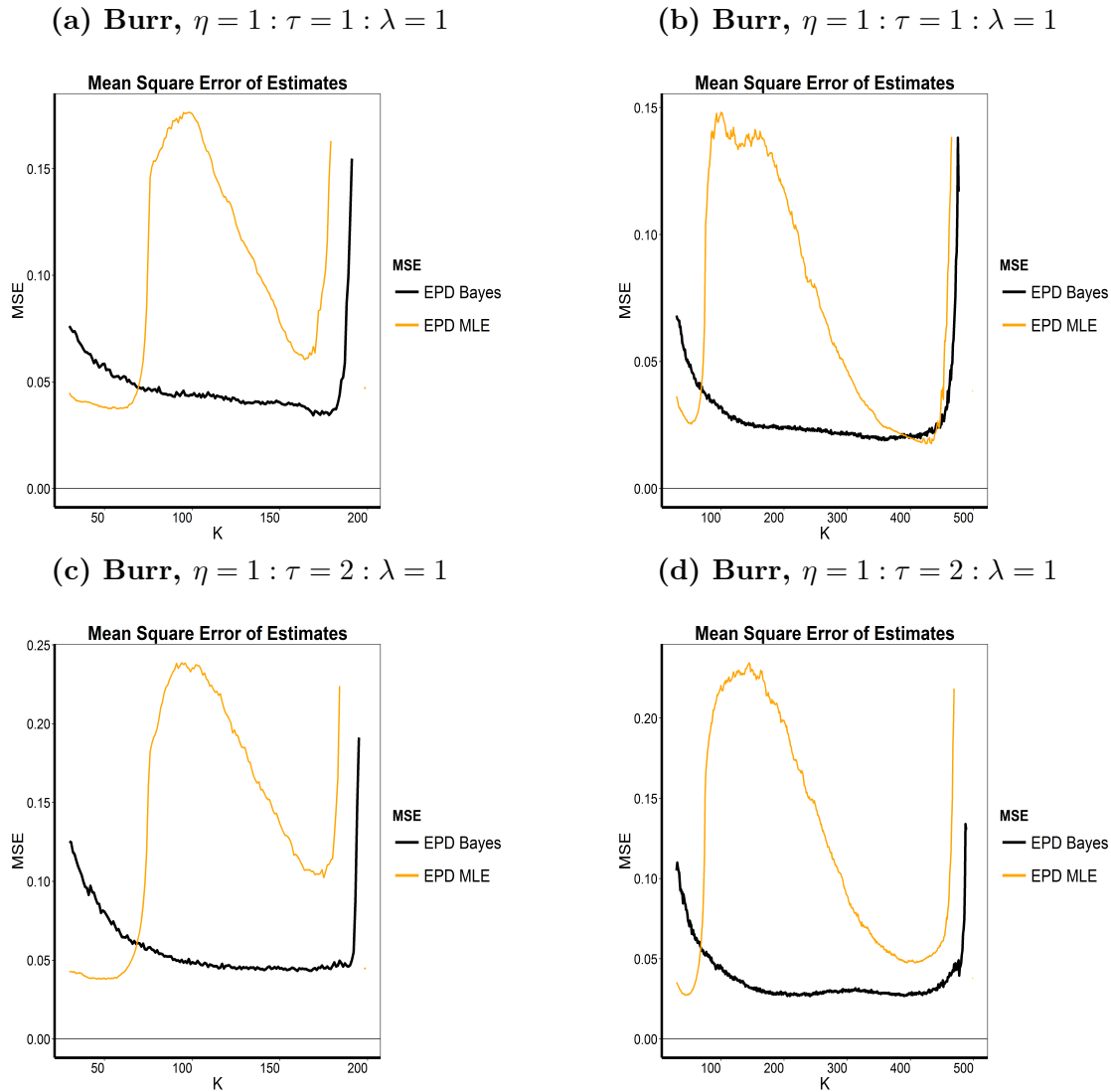


Figure 5.15: MSE of EVI estimates for a **Burr** sample of size $n=200$ (left), and size $n=500$ (right), and **EVI** of 1 (top) and 0.5 (bottom)

From Figure 5.14 and Figure 5.15 we can see that the Bayesian estimates are performing much better than the ML estimates in all cases. Note also for the Bayesian estimate there is some indication of large metropolis jumps for cases where the $\text{EVI}=1$, the cycles are however smoothed out for larger samples. The asymptotic bias in Figure 5.14 and the MSE in Figure 5.15 for the Bayesian estimates are smaller than the ML estimates over a larger range of k values.

Student-t Distribution

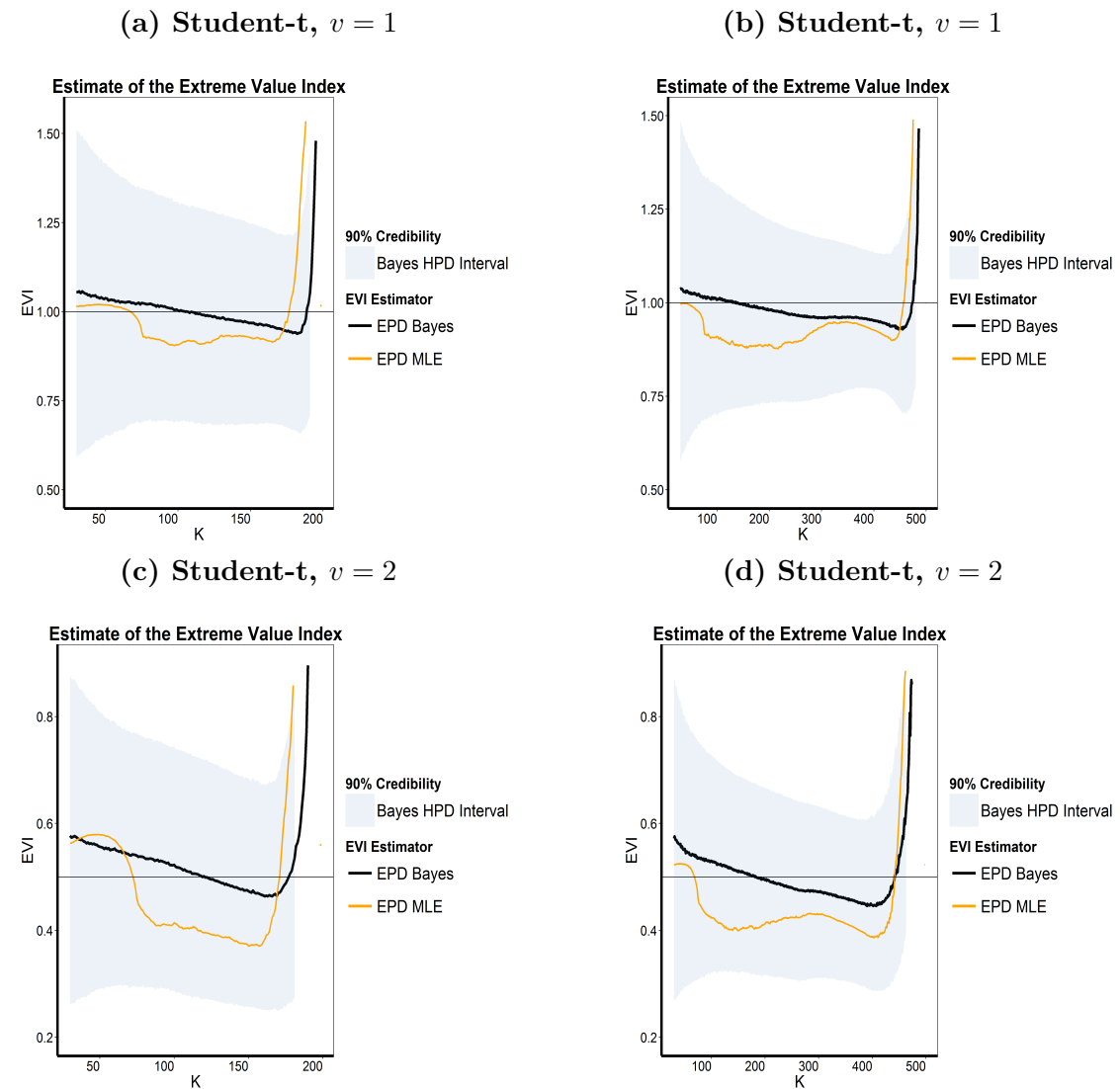


Figure 5.16: EVI estimates of a Student-t sample of size $n=200$ (left), and size $n=500$ (right), and EVI of 1 (top) and 0.5 (bottom), 90 % HPD (shaded fill)

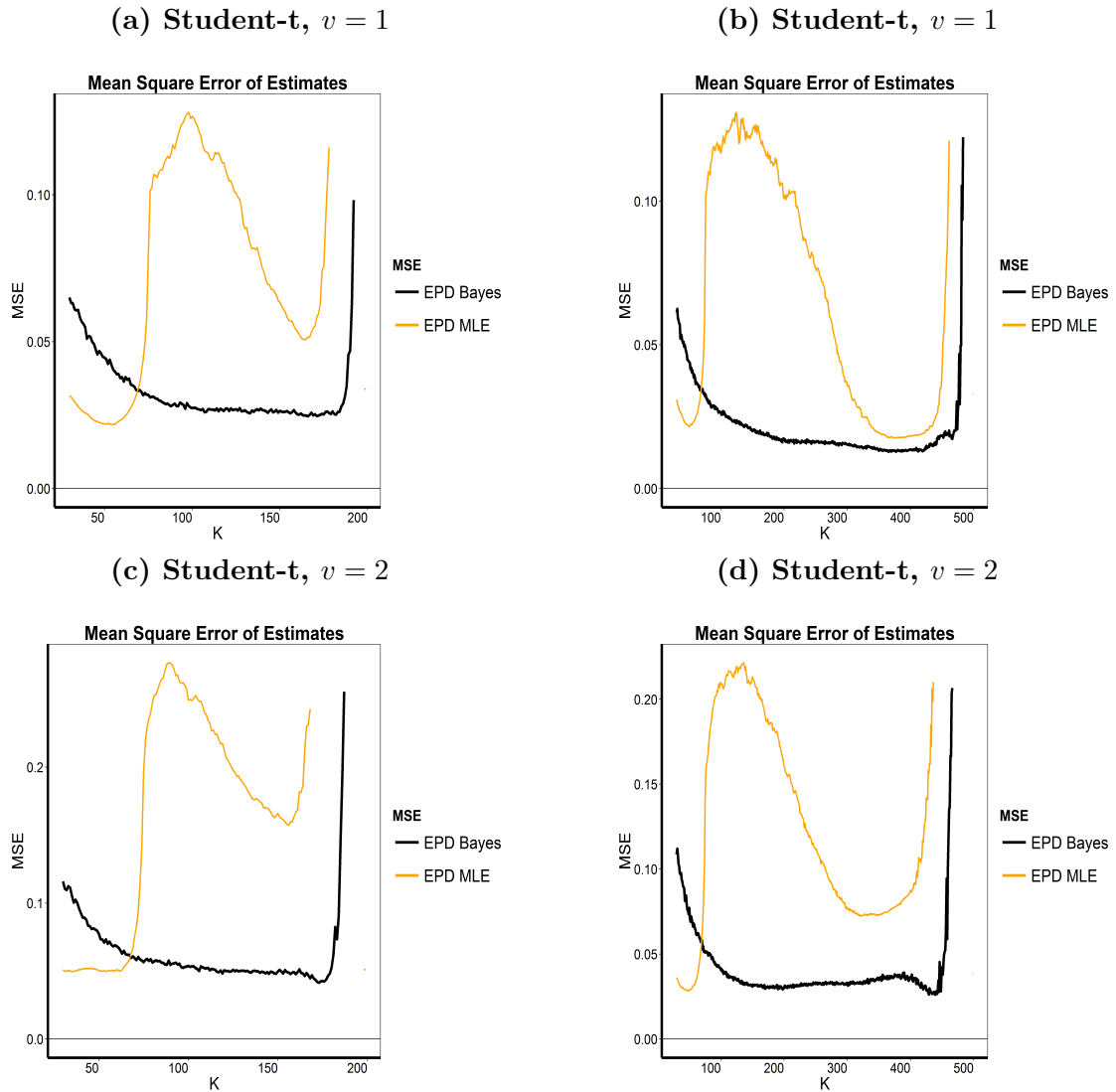


Figure 5.17: MSE of EVI estimates for a **Student-t** sample of size $n=200$ (left), and size $n=500$ (right), and **EVI** of 1 (top) and 0.5 (bottom)

From Figure 5.16 and Figure 5.17 we can see that the Bayesian estimates are outperforming those of the MLE, in all cases. The asymptotic bias in Figure 5.16 and the MSE in Figure 5.17 for the Bayesian estimates are smaller than the ML estimates over a larger range of k values.

5.6 Estimates of the Second Order Parameter

It is clear from the simulation study conducted in Section 5.5.2 that when the second order parameter ρ is estimated concurrently with other parameters for $k \in \{1, \dots, n - 1\}$, the ML estimator does not perform as good as the Bayesian

estimator. It is actually a strength of the Bayesian approach to fit models with a large number of parameters through the use of Monte Carlo Markov chain methods. This method is one of the most reliable methods to choose a suitable approximating distribution when sampling from complex Bayesian posterior distributions.

In the following section we produce a similar simulation study to the one conducted in Section 5.5.2 to further investigate how the Bayesian approach estimates the second order parameter τ . Table 5.18 shows the three distributions from which the samples are generated.

Table 5.18: Pareto type distributions with the second order parameter

Distribution	$1 - F(x)$	γ	ρ	τ
Fréchet (α)	$1 - \exp(-x^{-\alpha})$	$\frac{1}{\alpha}$	-1	$-\alpha$
Burr (η, τ, λ)	$\left(\frac{\eta}{\eta + x^\tau}\right)^\lambda$	$\frac{1}{\lambda\tau}$	$-\frac{1}{\lambda}$	$-\tau$
Student-t	$\int_x^\infty \frac{2\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} \left(1 + \frac{w^2}{n}\right)^{-\frac{n+1}{2}} dw$	$\frac{1}{n}$	$-\frac{2}{n}$	-2

From this table, since $\tau = \rho/\gamma$ by setting $\gamma = 1$ for the Fréchet we have $\tau = \rho = -1$, setting $\eta = 1, \tau = 1, \lambda = 1$ for the Burr(XII) distribution we have $\tau = \rho = -1$ and setting $v = 1$ for the Student-t distribution we have $\tau = \rho = -1$. We can therefore closely study the behaviour of the second order parameter ρ through estimates of τ .

We estimate τ using the following methods

1. $\tau = \hat{\rho}/\gamma$ where $\hat{\rho}$ is Fraga's estimate and γ is the true EVI,
2. Maximum likelihood parameter estimation method as mentioned in Chapter 4
3. Lastly Bayesian parameter estimation method as explained in this chapter.

Fréchet

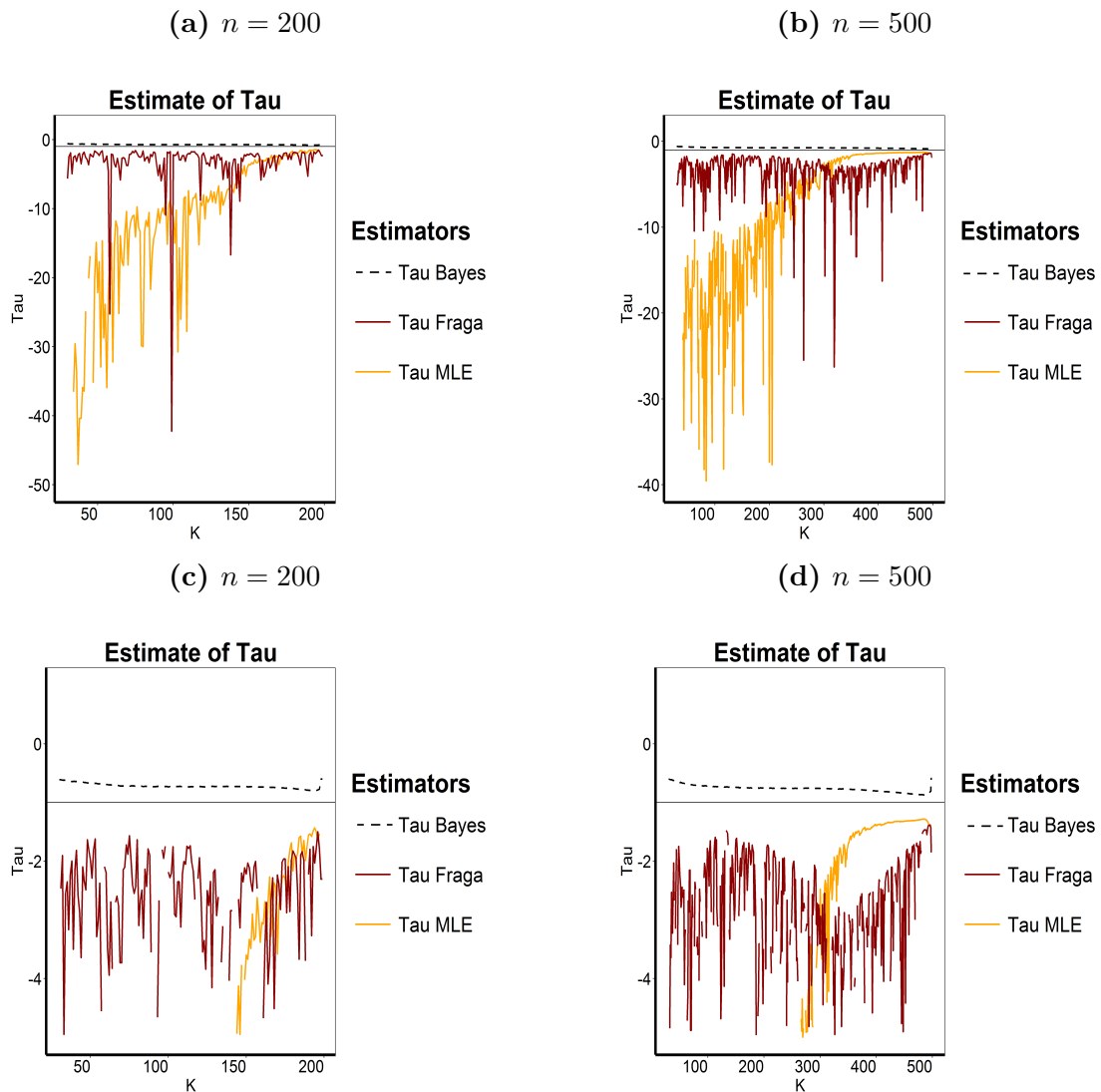


Figure 5.18: Simulated mean value of τ for a sample of sizes $n = 200$ (left) and $n = 500$ (right) from a Fréchet distribution (where true value $\tau = \rho = -1$), zoomed out view (top), zoomed in view (bottom)

As compared to γ estimates of the second order parameter ρ should be considered in a lower part of the sample. From the Figure 5.18 an interesting similarity between the ML and Fraga's estimator is apparent. The scale however of the ML estimator is wider than that of Fraga's estimate. Moreover, both the ML and Fraga's estimates become more stable and pronounced at the higher top order statistics, say $k_1 \rightarrow \infty$. The Bayesian estimator however is stable throughout.

Burr

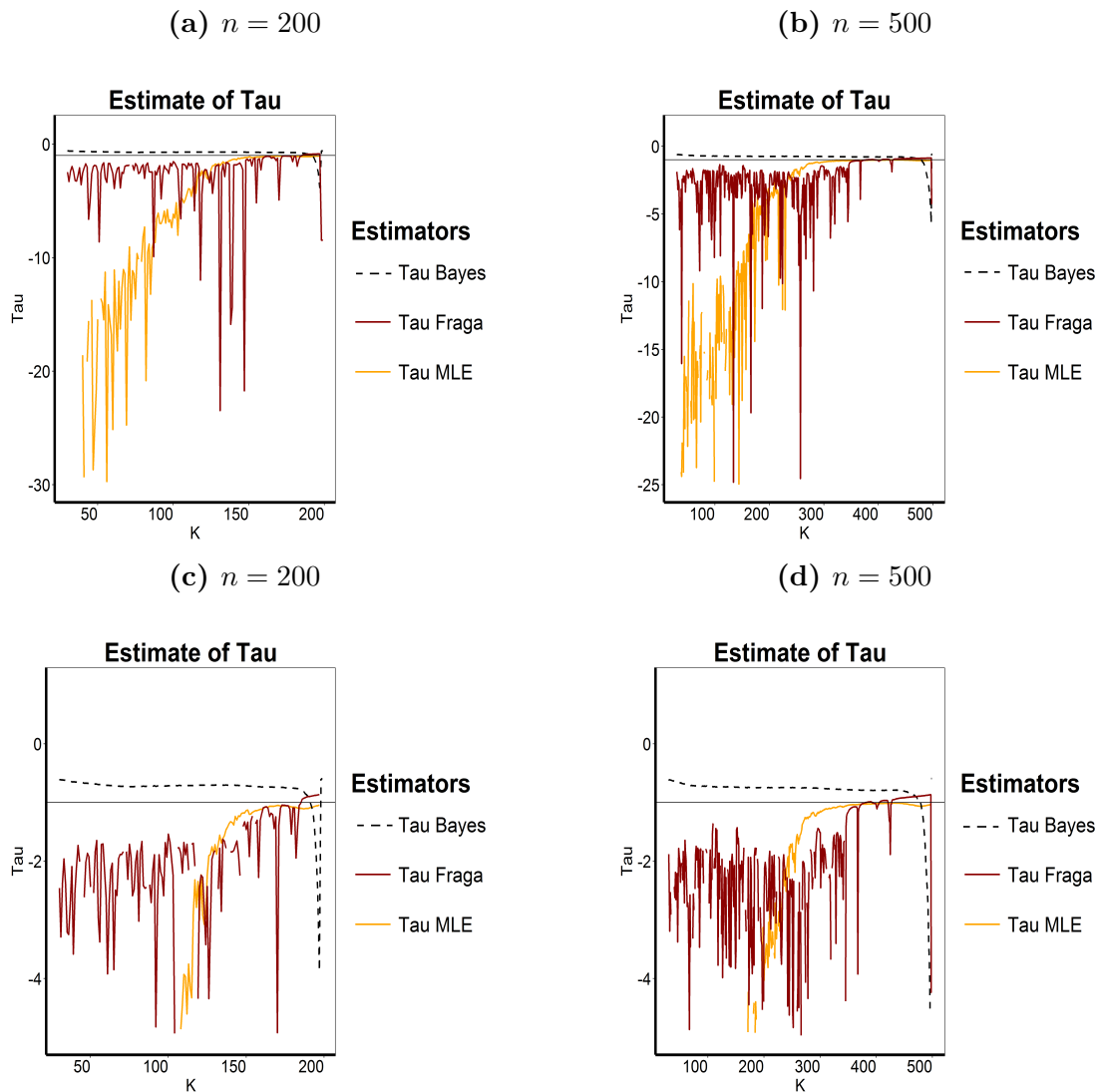


Figure 5.19: Simulated mean value of τ for a sample of sizes $n = 200$ (left) and $n = 500$ (right) from a Burr distribution (where true value $\tau = \rho = -1$), zoomed out view (top), zoomed in view (bottom)

The Burr distribution is generally more heavier tailed than the Fréchet distribution. From Figure 5.19 it appears the ML and Fraga's estimators become more pronounced and more accurate at higher top order statistics $k_1 \rightarrow \infty$. The Bayesian estimator is also very stable for all $k \in \{1, \dots, n - 1\}$ however it does not perform as good as the Fraga and ML method in the higher top order statistics.

Student-t

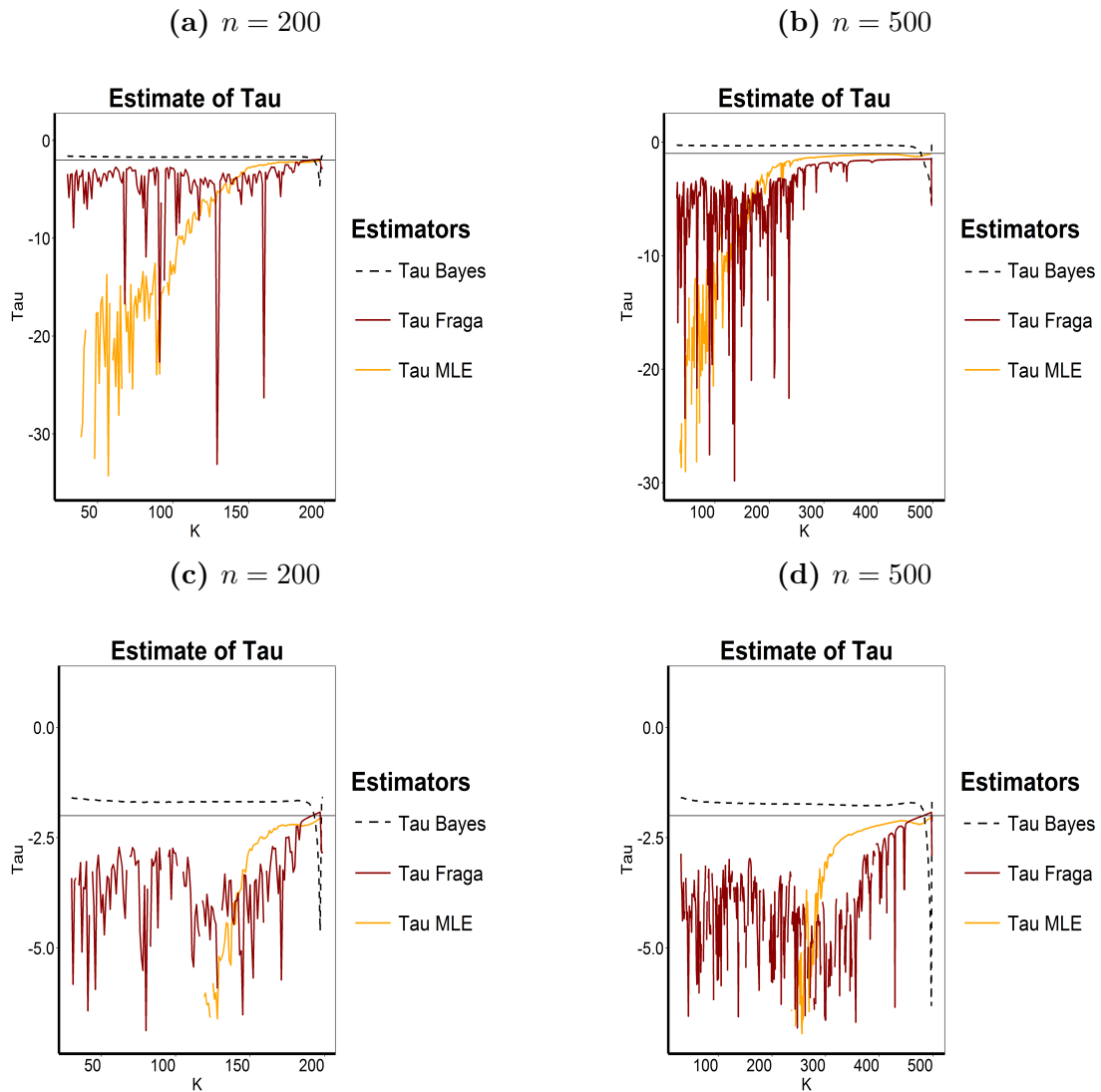


Figure 5.20: Simulated mean value of τ for a sample of sizes $n = 200$ (left) and $n = 500$ (right) from a Student-t distribution (where true value $\tau = \rho = -2$), zoomed out view (top), zoomed in view (bottom)

Lastly the Student-t distribution, also heavier tailed than the Fréchet distribution. Figure 5.20 presents results more similar to those in Figure 5.19. Fraga’s estimate is much more accurate for very high top order statistics $k \in \{1, \dots, n - 1\}$.

Remarks

Through the simulation study conducted in Section 5.6 we are now able to see the behavior of the EPD Bayesian estimator in comparison to the ML and Fraga’s

estimator. It is now apprehensible why the Bayesian approach works better than the ML estimator in estimating the EVI. In the case of the ML estimator, the likelihood of the EPD is maximized and all three parameters are estimated concurrently and therefore the value of γ is dependent on the value of τ . And therefore if the value of τ is estimated for every $k \in \{1, \dots, n - 1\}$ then we can see from the simulation study conducted in Section 5.6 that the estimate of the second order parameter ρ or τ is unstable and thus distorts the trajectories of the sample paths for γ .

In the case of the Bayesian estimator though, we can refer to the graphical diagrams in Figure 5.1b. From this Figure we can see that γ is a parent to the node τ and therefore γ has much influence on τ . Because of this dependence structure the Bayesian estimator is bound to perform better than the ML estimate if we estimate parameters for different $k \in \{1, \dots, n - 1\}$.

In the next section we see the performance of our estimators through an application on a real dataset. In the next case study, the second order parameter ρ will be estimated externally using Fraga's estimator as given in Section 3.3.1. And fixed for all $k \in \{1, \dots, n - 1\}$.

5.7 Case-Study

In this section, we assess the performance of our estimators using a real dataset. An Extreme Value Analysis (EVA) can be very useful in helping insurance companies price the unlimited excess-loss layer above some operational priority R , and the most important estimated quantity in such an analysis is often the Extreme Value index, from which we can extrapolate high quantile estimates, tail probabilities and various other EVA quantities. We focus on the estimation of the EVI, and use our estimates to compute corresponding quantiles and small tail probabilities using both the Bayesian and MLE parameter estimates.

We use data presented in Beirlant et al. (2004) and subsequently studied by Vandewalle and Beirlant (2006), Beirlant et al. (2008) and Beirlant et al. (2009). The data is from a European re-insurer: Secura Belgian Re and contains 371 automobile claims exceeding €1,200,000, spanning over a period of 1988 to 2001 and gathered from several European insurance companies co-operating with Secura Belgian Re. This dataset is corrected amongst other things for inflation.

The EVI and high quantile estimation

From the second order condition (\mathcal{R}) and using the EPD as the distribution of X/u_n given $X > u_n$. Goegebeur et al. (2014) introduced for $\bar{F}(u_n) \rightarrow 0$:

$$U_0\left(\frac{1}{p_n}\right) := u_n \left(\frac{p_n}{\bar{F}(u_n)} \right)^{-\gamma} \left(1 - \delta(u_n) \left(1 - \left(\frac{p_n}{\bar{F}(u_n)} \right)^{-\rho} \right) \right) \quad (5.26)$$

as an approximation for $U(1/p_n)$. Furthermore let X_1, \dots, X_n be independent and identically distributed (i.i.d) random variables with a distribution function satisfying (\mathcal{R}) and denote by $X_{1,n} \leq \dots \leq X_{n,n}$ the corresponding order statistics. Taking the data adaptive threshold $u_n = X_{n-k,n}$, replacing F by the empirical distribution function in (Equation 5.26), and using the fact that $e^{-x} \approx 1 - x$ for $x \rightarrow 0$, Goegebeur et al. (2014) introduced the following extreme quantile estimator:

$$\hat{U}\left(\frac{1}{p_n}\right) := X_{n-k,n} \left(\frac{np_n}{k} \right)^{-\hat{\gamma}_n} \exp \left(-\hat{\delta}_n \left(1 - \left(\frac{np_n}{k} \right)^{-\hat{\rho}_n} \right) \right) \quad (5.27)$$

where $(\hat{\gamma}_n, \hat{\delta}_n)$ are estimates of (γ, δ) and $\hat{\rho}_n$ is a consistent estimator sequence for ρ . In this section we use the Bayesian parameter estimates as well as the ML estimates to estimate the $1 - 1/1000$ quantile.

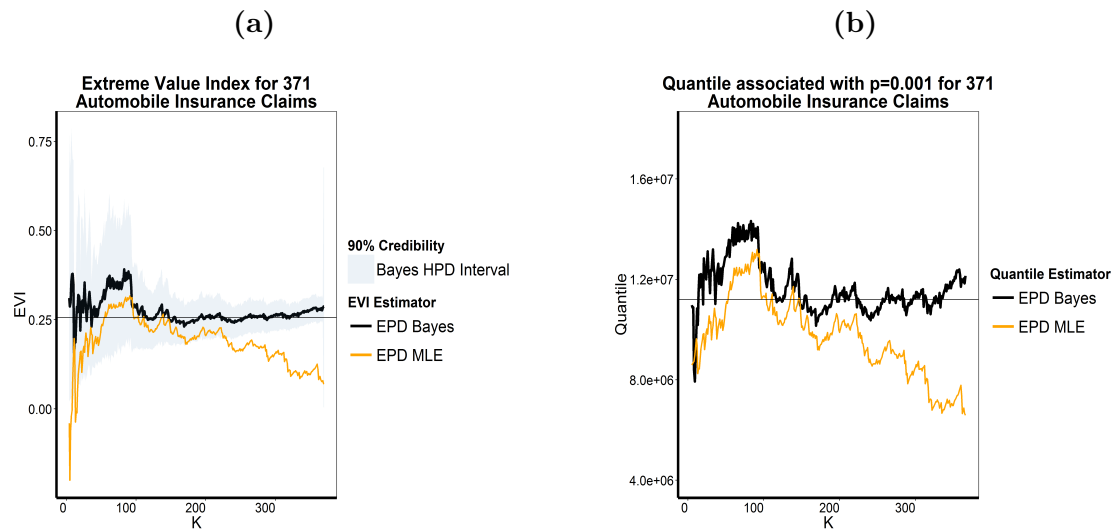


Figure 5.21: Estimates of the Extreme value index (a) and the quantile estimate associated with $p = 0.001$ (b) for the Secura Belgian Re data

From Figure 5.21 projected estimates of $\hat{\gamma}_{Bayes}$ and \hat{Q}_{Bayes} depicted as straight lines in Figure 5.21 (a) and (b) are the mean of the Bayesian estimates between $k = 100$ and $k = 350$, this is taken as the range for which the Bayesian estimates

are most stable. It is suggested in Beirlant et al. (2004) to model the complete distribution of this data by a two components mixture, namely: the Exponential and the Pareto, with a knot at roughly €2,600,000, corresponding to the order statistic $X_{n-k:n}$ with $k = 95$. Beirlant et al. (2009) however showed that this knot, although detected by the EPD, does not drastically distort tail parameter estimates. It is clear from Figure 5.21, that the Bayesian estimates of both the EVI and corresponding Quantile are more stable than those of the MLE. From $k = 100$, trajectories of our Bayesian estimates become relatively stable, oscillating around the points $\hat{\gamma}_{Bayes} \approx 0.25$ in Figure 5.21 (a) and $\hat{Q}_{Bayes} \approx \text{€}11,198,220$ in Figure 5.21 (b). Note also that the 90% HPD interval for the EVI narrows down for k large (as we would expect), and all Bayesian EVI estimates over k from $k = 5$ up to, and including $k = (n - 1)$ are within this credibility region.

Tail probability estimation

Given the order statistic $X_{1:n} \leq \dots \leq X_{n:n}$ of an independent sample from an unknown distribution function F , we are interested in an estimate of the tail probability $p_n = \bar{F}(x_n)$, where $x_n \rightarrow \infty$ and thus $p_n \rightarrow 0$ as $n \rightarrow \infty$. At a random threshold $t = X_{n-k:n}$ and given some parameter estimates of the EPD $(\hat{\gamma}_n, \hat{\delta}_n, \hat{\tau}_n)$, the tail probability estimate is constructed in Beirlant et al. (2009) as

$$\hat{p}_{k,n} = \hat{F}(x_n) = \frac{k}{n} \bar{G}_{\hat{\gamma}_n, \hat{\delta}_n, \hat{\tau}_n}(x_n / X_{n-k:n}) \quad (5.28)$$

An estimator of τ is taken as $\hat{\tau}_n = \hat{\rho}_n / \hat{\gamma}_n$, where $\hat{\rho}_n$ is a consistent estimator of the second order parameter ρ (Fraga Alves et al., 2003b).

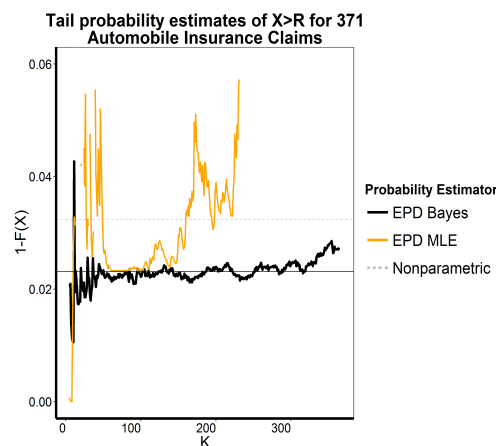


Figure 5.22: Estimates of the tail probabilities of X greater than an operational priority level $R = 5,000,000$ for the Secura Belgian Re data

The horizontal dashed line in Figure 5.22 is a nonparametric estimate of the exceedance probability $\frac{12}{371} = 3.24\%$, where 12 is the number of claims exceeding the operational priority R and is divided by the total number of claims. We can see again that trajectories of the Bayesian estimates are relatively more stable than those of the MLE and oscillate around the point $\hat{p}_{Bayes} \approx 2.31\%$. From this case study, we can conclude that a Bayesian approach to EVA does indeed yield more stability and assurance to some degree, of the accuracy of our estimates. We therefore regard the Bayesian approach as more desirable than the classical MLE approach.

5.8 Conclusion

We introduced an improved and efficient Bayesian approach to estimating the Extreme Value Index (EVI) for heavy-tailed data using the computationally efficient MCMC engine called OpenBUGS. This proposed approach, reduces the bias and improves stability when compared to the classical MLE. The basic result of consistency of EPD Bayesian estimators can be drawn from the theme of consistency of general Bayes estimators as given by Schwartz (1965) and Diaconis and Freedman (1986). As we get more data, the posterior distribution will tend to a ‘point mass’ at the true value of γ, δ and τ .

Through a simulation study we were also able to assess performance of our estimators using samples generated from three heavy-tailed distributions. We found, in general, that the Bayesian approach with vague prior distributions was better than the classical MLE in terms of bias reduction and stability. We further compared the results between the two approaches by conducting a case study using a real dataset and showed that our Bayesian estimates yield more stable approximations. We therefore recommend the use of Bayesian methods to estimate parameters of the EPD.

In the next Section, the EPD Bayesian estimator will be altered and adapted for right censored data.

Chapter 6

Estimation of the EVI under random censorship

6.1 Introduction

In extreme value theory we often assume a complete random univariate sample (X_1, \dots, X_n) to be independent and identically distributed by some distribution function F . The majority of available literature on estimating tail quantities assume the data is complete, non-censored and observable. In this chapter we assume the data is randomly right censored and assess some well known adapted semi-parametric EVI estimators for censored data developed by Einmahl et al. (2008). These estimators are then compared to the GPD censored likelihood developed by Beirlant et al. (2007) and the parametric EPD censored likelihood developed by Beirlant et al. (2016) in estimating the EVI for Pareto type distributions *i.e* in the estimation of a positive tail index γ_X . It goes beyond the scope of this thesis to investigate or derive censored extreme quantile estimators for the EPD.

The rest of this chapter is organized as follows. In Section 6.1.1 we look into some situations in practice in which random censoring can occur. In Section 6.2 we provide the general treatment of random right censoring in EVT and mention all estimators that will be compared in this chapter. In Section 6.3 we present results of a Monte Carlo simulation conducted to assess the behavior of the EVI estimators mentioned in Section 6.2. In addition to the Monte Carlo simulation conducted in Section 6.3, we illustrate the behavior of the same EVI estimators for censored cancer data in Section 6.4.

In Section 6.5 we extend the EPD censored likelihood approach of Beirlant et al. (2016) into a Bayesian setting, and construct a censored posterior. In Section 6.6 we present results of a Monte Carlo simulation conducted to compare the behavior of the MLE likelihood adapted estimator and the Bayesian censored posterior estimator. In Section 6.7 a conclusion is made.

6.1.1 Censoring in Practice

Incomplete data is a very common issue when analyzing survival and insurance data and sometimes physical phenomena data where extreme measurements might be incomplete due to damage to the measuring instrument. We start by illustrating possible situations where random right censoring can occur.

Reinsurance

Assume that for some insurance company, X_i is the loss from an i -th claim. In practice however, this claim will have a limit say, C_i , possibly induced by an insurance assessor or negotiator, who investigates the claim and negotiates a settlement figure below or equal to the actual claim. If the claim exceeds this limit i.e. $(X_i > C_i)$, $i = 1, \dots, n$ then the observed claim will be right-censored. We assume here that X and C are independent and under this set-up X is non-observable, we therefore only observe $(\mathbf{Z}, \boldsymbol{\omega})$ where $Z_i = \min(X_i, C_i)$ and:

$$\omega_i = \begin{cases} 1 & \text{if } X_i \leq C_i, \\ 0 & \text{if } X_i > C_i \end{cases} \quad (6.1)$$

Here, ω_i is called the indicator variable which determines whether or not the variable X , is censored. An observation is thus censored whenever $\omega_i = 0$. A reinsurer however is more interested in making inference on the right tail of the lifetime distribution of X , say F , since it will have to insure the insurance company over its policy limits. However, due to the presence of right-censoring, the right tail of the distribution F cannot be accurately estimated unless additional information is given about the incomplete observations, and the estimation procedures are adapted for the censoring.

Survival Data

Censoring is a common problem in clinical trials, we therefore provide an small application study in Section 6.4 where we estimate the extreme index of partially censored times to death of patients with Cancer of the tongue.

As a general treatment, assume that patients with some type of cancer are studied to investigate their time to death. Follow-up survival data is then taken on each patient at the end of the study. Some patient X_i could still be alive at the end of the study period or analysis and thus the event of interest namely: their time to death, has not occurred, the observation is therefore right-censored by the time of study C_i . However censoring may be induced by other factors such as death from competing risks (for example, the patient dies from a car accident), or the patient withdraws from the study entirely. Again we can only observe (Z_i, ω_i) defined as

above. Our main interest is to make inference on the right tail of the distribution function (d.f) of X .

6.2 Random right Censoring in EVT

For a more general treatment as given in Beirlant et al. (2007), let $X_i, i = 1, \dots, n$, be independent and identically distributed random variables with common continuous d.f F_X , and let $Y_i, i = 1, \dots, n$ be a another sequence of independent and identically distributed random variables with continuous d.f F_Y independent of F_X . We then observe $Z_i = X_i \wedge Y_i$ and $\omega_i = \mathbb{1}_{\{X_i \leq Y_i\}}, i = 1, \dots, n$. Let H be a d.f of Z_1 with $\tau_H = \inf\{x : H(x) = 1\}$ the support of H . Furthermore let $\bar{H}^1(z) = \mathbb{P}(Z > z, \omega = 1) = \mathbb{P}(z < X \leq Y)$. We assume that both F_X and F_Y are in the Fréchet domain i.e. $F_X \in \mathcal{D}(G_{\gamma_1})$ and $F_Y \in \mathcal{D}(G_{\gamma_2})$ with $\gamma_1 = \gamma_X$ the parameter of interest.

We emphasize the following assumptions:

1. X_1, X_2, \dots, X_n are all nonnegative and i.i.d. with distribution function F_X and density f_X .
2. Y_1, Y_2, \dots, Y_n are all nonnegative and i.i.d. with distribution function F_Y and density f_Y .
3. X and Y are independent.

Censored Fréchet Distribution

Let both X and Y be independent random variables with distribution Fréchet(γ_1) and Fréchet(γ_2) respectively, thus for all $x > 0$, $F_X(z) = 1 - \exp(-z^{-1/\gamma_1})$ and $F_Y(z) = 1 - \exp(-z^{-1/\gamma_2})$. Consequently:

$$\begin{aligned} F_Z(z) &= \mathbb{P}(\min(X, Y) \leq z) = 1 - \mathbb{P}(X > z)\mathbb{P}(Y > z) \\ &= 1 - \exp(-z^{-1/\gamma_1})\exp(-z^{-1/\gamma_2}) \\ &= 1 - \exp(-z^{-(\gamma_1+\gamma_2)/\gamma_1\gamma_2}) \end{aligned}$$

i.e. $Z \sim \text{Fréchet}(\gamma)$, where $\gamma = \gamma_1\gamma_2/(\gamma_1 + \gamma_2)$, and

$$f_Z(z) = -\exp(-z^{-1/\gamma_1})\exp(-z^{-1/\gamma_2}) \left(\frac{1}{\gamma_1} z^{-1/\gamma_1-1} + \frac{1}{\gamma_2} z^{-1/\gamma_2-1} \right)$$

We can now estimate

$$\begin{aligned}
 p \equiv p_Z &:= \mathbb{P}(X \leq Y | Z = z) = \mathbb{P}(\omega = 1 | Z = z) = \frac{f_X(z)(1 - F_Y(z))}{f_Z(z)} \\
 &= \frac{-\exp(-z^{-1/\gamma_1}) \frac{1}{\gamma_1} z^{-1/\gamma_1 - 1} \exp(-z^{-1/\gamma_2})}{-\exp(-z^{-1/\gamma_1}) \exp(-z^{-1/\gamma_2}) \left(\frac{1}{\gamma_1} z^{-1/\gamma_1 - 1} + \frac{1}{\gamma_2} z^{-1/\gamma_2 - 1} \right)} \\
 &= \frac{\frac{1}{\gamma_1} z^{-1/\gamma_1 - 1}}{\frac{1}{\gamma_1} z^{-1/\gamma_1 - 1} + \frac{1}{\gamma_2} z^{-1/\gamma_2 - 1}} \stackrel{z \rightarrow \infty}{=} \frac{\frac{1}{\gamma_1}}{\frac{1}{\gamma_1} + \frac{1}{\gamma_2}} \\
 &= \frac{\gamma_2}{\gamma_1 + \gamma_2}
 \end{aligned}$$

Therefore, the quotient between any estimator of $\gamma = \gamma_Z$ and an estimator of $p = p_Z$ will provide an estimator of the parameter of interest $\gamma_X = \gamma_1$. The general case for any extreme value distribution can then be written as:

$$F_X \in \mathcal{D}(G_{\gamma_1}), F_Y \in \mathcal{D}(G_{\gamma_2}) \rightarrow F_Z \in \mathcal{D}(G_{\gamma}), \text{ where } \gamma = \frac{\gamma_1 \gamma_2}{\gamma_1 + \gamma_2}.$$

We assume the following construction from Einmahl et al. (2008), where Z is a random variable with d.f H , in this case —Fréchet(γ) and let U be a random variable with d.f U uniform(0,1) and independent of Z . ω is then defined as:

$$\omega = \begin{cases} 1, & \text{if } U \leq p(Z), \\ 0, & \text{if } U > p(Z) \end{cases} \quad (6.2)$$

and,

$$\tilde{\omega} = \begin{cases} 1, & \text{if } U \leq p, \\ 0, & \text{if } U > p \end{cases} \quad (6.3)$$

Einmahl et al. (2008) showed that the (Z_i, ω_i) , $i = 1, \dots, n$, resulting from repeating this construction independently n times, have the same distribution as the initial pairs (Z_i, ω_i) , $i = 1, \dots, n$, where $Z = \min(X, Y)$ and $\omega = \mathbb{1}_{\{X \leq Y\}}$, for all $n \in \mathbb{N}$. \hat{p} can then also be written as:

$$\hat{p} = \frac{1}{k} \sum_{j=1}^k \mathbb{1}_{\{U_{[n-j+1, n]} \leq p(Z_{n-j+1, n})\}} \quad (6.4)$$

and, similarly

$$\tilde{p} := \frac{1}{k} \sum_{j=1}^k \tilde{\omega}_{[n-j+1, n]} = \frac{1}{k} \sum_{j=1}^k \mathbb{1}_{\{U_{[n-j+1, n]} \leq p\}}$$

Clearly showing that $\tilde{p} \stackrel{d}{=} \frac{1}{k} \sum_{j=1}^k \mathbb{1}_{\{U_j \leq p\}}$

More generally any function based on the $(k+1)$ top order statistics in the observed (Z_1, Z_2, \dots, Z_n) , used to estimate the EVI in complete samples, and generally denoted as $\hat{\gamma}_{k,n}(\mathbf{Z})$, will converge towards $\gamma_1\gamma_2/(\gamma_1 + \gamma_2)$ for intermediate k , i.e. whenever

$$k = k_n \rightarrow \infty \text{ and } \frac{k}{n} \rightarrow 0, \text{ as } n \rightarrow \infty$$

and

$$\frac{\hat{\gamma}_{k,n,Z}}{p} \xrightarrow{p} \gamma_1 \equiv \gamma_X, \text{ as } n \rightarrow \infty \quad (6.5)$$

6.2.1 Semi-Parametric Estimators

Recently in EVT, semi-parametric approaches have gained more popularity over parametric methods. This is due to the fact that semi-parametric estimation is drawn from a general framework, and therefore not based on probabilistic asymptotic results, as is the case in parametric approaches.

We consider two classical extreme value estimators, namely: the moment (Dekkers et al., 1989) and the Generalized Hill (Beirlant et al., 1996) to estimate the uncensored EVI of the complete sample $\mathbf{Z} = (Z_1, \dots, Z_n)$. For $j \geq 1$, $1 \leq k < n$, and with $(Z_{1,n}, \dots, Z_{n,n})$ denoting a set of ascending order statistics associated with the sample \mathbf{Z} , the moment estimator (M) is given by:

$$\hat{\gamma}_{k,n}^M(\mathbf{Z}) := M_{k,n}^{(1)} + \frac{1}{2} \{1 - (M_{k,n}^{(2)} / [M_{k,n}^{(1)}]^2 - 1)^{-1}\} \quad (6.6)$$

where,

$$M_{k,n}^{(2)} := \frac{1}{k} \sum_{i=1}^k [\log Z_{n-i+1:n} - \log Z_{n-k:n}]^2 \text{ and } M_{k,n}^{(1)} \text{ is the Hill (1975) estimator}$$

defined by:

$$H_k := \frac{1}{k} \sum_{i=1}^k \log Z_{n-i+1:n} - \log Z_{n-k:n} \quad (6.7)$$

Also for $k = 2, \dots, n-1$ the generalized Hill(GH) estimator is given by:

$$\hat{\gamma}_{k,n}^{GH}(\mathbf{Z}) := \frac{1}{k} \sum_{j=1}^k \log UH_{j,n} - \log UH_{k,n}, \text{ where } UH_{j,n} = Z_{n-j:n}H_j, 1 \leq j \leq k, \quad (6.8)$$

and H_k defined as in Equation 6.7. Whenever

$$k = k_n \rightarrow \infty \text{ and } \frac{k}{n}, \text{ as } n \rightarrow \infty$$

we can prove convergence in probability to $\gamma \equiv \gamma_Z$ in the domain of attraction where they are valid and under additional validity of second-order condition in Equation 2.23 asymptotic normality of both estimators can be guaranteed (see for instance Gomes and Neves, 2011).

Einmahl et al. (2008) showed that all estimators of a non-negative EVI can be adapted for censoring by simply dividing the original estimator based on the Z sample by the proportion of non-censored observations in the k largest Z 's, i.e.

$$\hat{\gamma}_{k,n}^{(c,M)}(\mathbf{X}) := \frac{\hat{\gamma}_{k,n}^M(\mathbf{Z})}{\hat{p}_k} \quad \text{and} \quad \hat{\gamma}_{k,n}^{(c,GH)}(\mathbf{X}) := \frac{\hat{\gamma}_{k,n}^{GH}(\mathbf{Z})}{\hat{p}_k} \quad (6.9)$$

with a simple, semi-parametric estimator of \hat{p} given by:

$$\hat{p} = \hat{p}_{k,n} := \frac{1}{k} \sum_{j=1}^k \omega_{[n-j+1]}, \quad (6.10)$$

where $\omega_{[n-j+1]}, 1 \leq j \leq n$, are the induced or concomitant order statistics associated with the ordered sample $Z_{n-j+1:n}, 1 \leq j \leq n$.

6.2.2 Parametric Estimators

Censoring adapted generalized Pareto distribution

The GPD, as defined in Section 2.3.1 is used to model the right tail of an underlying distribution function F , if, $F \in \mathcal{D}(G_\gamma)$ we say F belongs to the max-domain of attraction of G_γ . In the case of censoring the likelihood can be adjusted in the following way (see Beirlant et al., 2007):

$$\mathbb{P}(Z, \omega = 0) = \mathbb{P}(Z = Y | \omega = 0) \mathbb{P}(\omega = 0) = \mathbb{P}(\omega = 0) \quad (6.11)$$

$$= \mathbb{P}(X > Y) = \bar{F}(Y) \quad (6.12)$$

$\bar{F}(Y)$ is the right censored observation. Also for $\omega = 1$,

$$\mathbb{P}(Z, \omega = 1) = \mathbb{P}(Z = X | \omega = 1) \mathbb{P}(\omega = 1), \quad (6.13)$$

$$= \mathbb{P}(Z = X | X \leq Y) \mathbb{P}(X \leq Y) \quad (6.14)$$

$$= \left[\frac{f(Z)}{F(Y)} \right] [F(Y)] = f(Z) \quad (6.15)$$

where $\mathbb{P}(Z, \omega = 0)$ is the joint probability and is equivalent to $\mathbb{P}(Z \cap \omega = 0)$. These expressions can thus be combined into a single expression:

$$\mathbb{P}(Z, \omega) = [f(Z)]^\omega [\bar{F}(Z)]^{1-\omega}. \quad (6.16)$$

(see for example, Klein and Moeschberger, 2005)

The 2-parameter GPD cumulative distribution function is given by:

$$F_{\gamma, \sigma}(Z) = 1 - \left(1 + \frac{\gamma Z}{\sigma}\right)^{-1/\gamma} \quad (6.17)$$

(Matthys and Beirlant, 2003) where $Z \geq 0$ and $\gamma, \sigma > 0$. Consequently, the two-parameter density function of a GPD is given by:

$$f(Z) = \frac{1}{\sigma} \left(1 + \frac{\gamma Z}{\sigma}\right)^{-\frac{1}{\gamma}-1} \quad (6.18)$$

By setting $E_j = Z_j - t$, given $Z_j > t$ where $t \rightarrow \tau_F$ (the right end point of the distribution F) and denoting the number of absolute excesses over t by N_t , we use Equation 6.16 to derive the following GPD MLE estimator, adapted for censoring:

$$\hat{\gamma}_{ML,k}^{c,GPD} = \operatorname{argmax}_{\gamma_1} \prod_{j=1}^{N_t} [f_{\gamma_1, \sigma}(E_{n-j+1,n})]^{\omega_{n-j+1,n}} [\bar{F}_{\gamma_1, \sigma}(E_{n-j+1,n})]^{1-\omega_{n-j+1,n}} \quad (6.19)$$

For numerical stability we use logs:

$$\begin{aligned} \hat{\gamma}_{ML,k}^{c,GPD} = \operatorname{argmax}_{\gamma_1} &= \sum_{j=1}^k \omega_{n-j+1,n} \log \sigma + \left(\frac{1}{\gamma} + 1\right) \sum_{j=1}^k \omega_{n-j+1,n} \log E_{n-j+1} \\ &+ \left(\frac{1}{\gamma}\right) \sum_{j=1}^k (1 - \omega_{n-j+1,n}) \log E_{n-j+1,n} \end{aligned}$$

We solve this by taking the negative log-likelihood as the objective function and minimizing it over its sample parameter space $\gamma, \sigma > 0$.

Extended Pareto Distribution (EPD)

The distribution of the Peaks Over Thresholds can be approximated by the EPD with distribution function:

$$\bar{G}_{\gamma, \delta, \tau}(y) = \{y(1 + \delta - \delta y^\tau)\}^{-1/\gamma}, \quad y > 1 \quad (6.20)$$

where the parameter vector (γ, δ, τ) is in the range $\tau < 0 < \gamma$ and $\delta > \max(-1, 1/\tau)$, with $\delta = \delta = -D\gamma^{-1}C^{\rho/\gamma}t^{-\rho/\gamma}$ and $\tau = -\rho/\gamma$. The density function is defined by:

$$g_{\gamma, \delta, \tau}(y) = \frac{1}{\gamma} x^{-1/\gamma-1} \{1 + \delta(1 - x^\tau)\}^{-1/\gamma-1} [1 + \delta\{1 - (1 + \tau)x^\tau\}] \quad (6.21)$$

Using Equation 6.16, Bardoutsos et.al (2015) derived the following EPD MLE estimator adapted for censoring:

$$\hat{\gamma}_{ML,k}^{c,EPD} = \operatorname{argmax}_{\gamma_1} \prod_{j=1}^k \left[g_{\gamma_1, \delta, \tau} \left(\frac{Z_{n-j+1,n}}{Z_{n-k,n}} \right) \right]^{\omega_{n-j+1,n}} \left[\bar{G}_{\gamma_1, \delta, \tau} \left(\frac{Z_{n-j+1,n}}{Z_{n-k,n}} \right) \right]^{1-\omega_{n-j+1,n}} \quad (6.22)$$

We take the objective function to be the negative log-likelihood of the EPD and use the fact that $\tau = \rho/\gamma$. All optimization procedures are executed in **R** using the Limited Memory Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970), for a detailed description and implementation of this algorithm see Byrd et al. (1995). The algorithm used to estimate the second order parameter $\rho \leq 0$, controlling the speed of convergence of maximum values, as mentioned in Equation 2.23, is given in Fraga Alves et al. (2003b).

6.3 Monte Carlo Simulation

In this section we perform a simulation experiment to investigate the behavior of the parametric against the semi-parametric estimators for different values of k . We consider 3 cases, in the first case only 25% of the data is censored, in the 2nd case, 40% of the data is censored and in the 3rd case, 55% of the data is censored.

In all cases we perform Monte Carlo simulations, based on 10,000 runs, with samples of size $n = 1000$ from Z . We show the EVI estimates when all estimators are adapted for censoring, i.e. inferring on the right tail of X , and also the EVI estimates when the estimators are not adapted for censoring, thus making inference on the right tail of Z . We lastly show the Root Mean Square Error (RMSE) between the true estimators of γ_X and the censoring adapted estimators. The plots are obtained by averaging out over the 1,000 samples.

6.3.1 Case 1: 25% Censoring

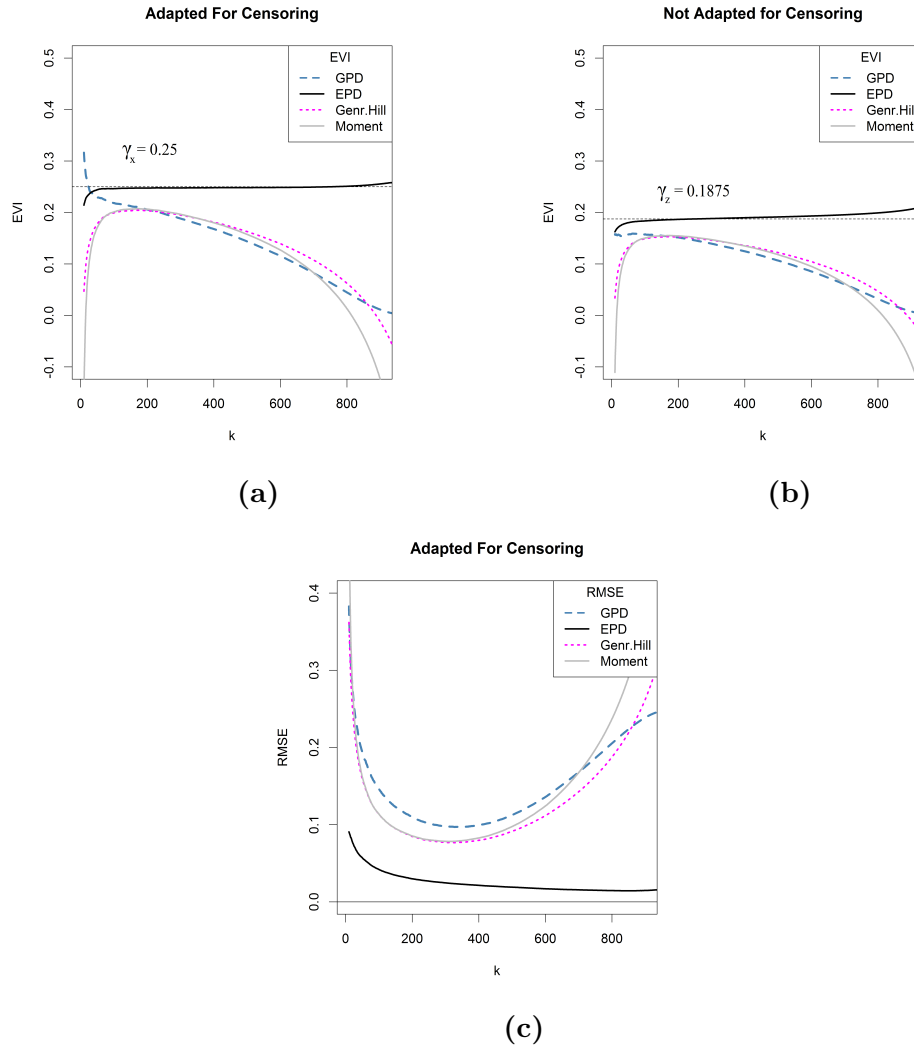


Figure 6.1: A $\text{Fréchet}(\gamma_X = 0.25)$ distribution censored by a $\text{Fréchet}(\gamma_Y = 0.75)$, thus $\gamma_Z = \frac{0.25 \cdot 0.75}{0.25 + 0.75} = 0.1875$ and proportion of non-censored data — $p = 75\%$ (25% censoring in the right tail); (a) EVI estimates adapted for censoring and (b) EVI estimate not adapted for censoring, (c) RMSE

From Figure 6.1, we can see that both the EPD estimators adapted and not adapted for censoring, are performing better than all other estimators in terms of bias and Root Mean Square Error (RMSE). Both GPD estimators seem to be performing poorly in comparison to the semi-parametric estimators. Looking at the RMSE in Figure 6.1c and EVI in Figure 6.1a, the Generalized Hill is performing better than the Moment estimator, and also better than the GPD when k is large.

6.3.2 Case 2: 40% Censoring

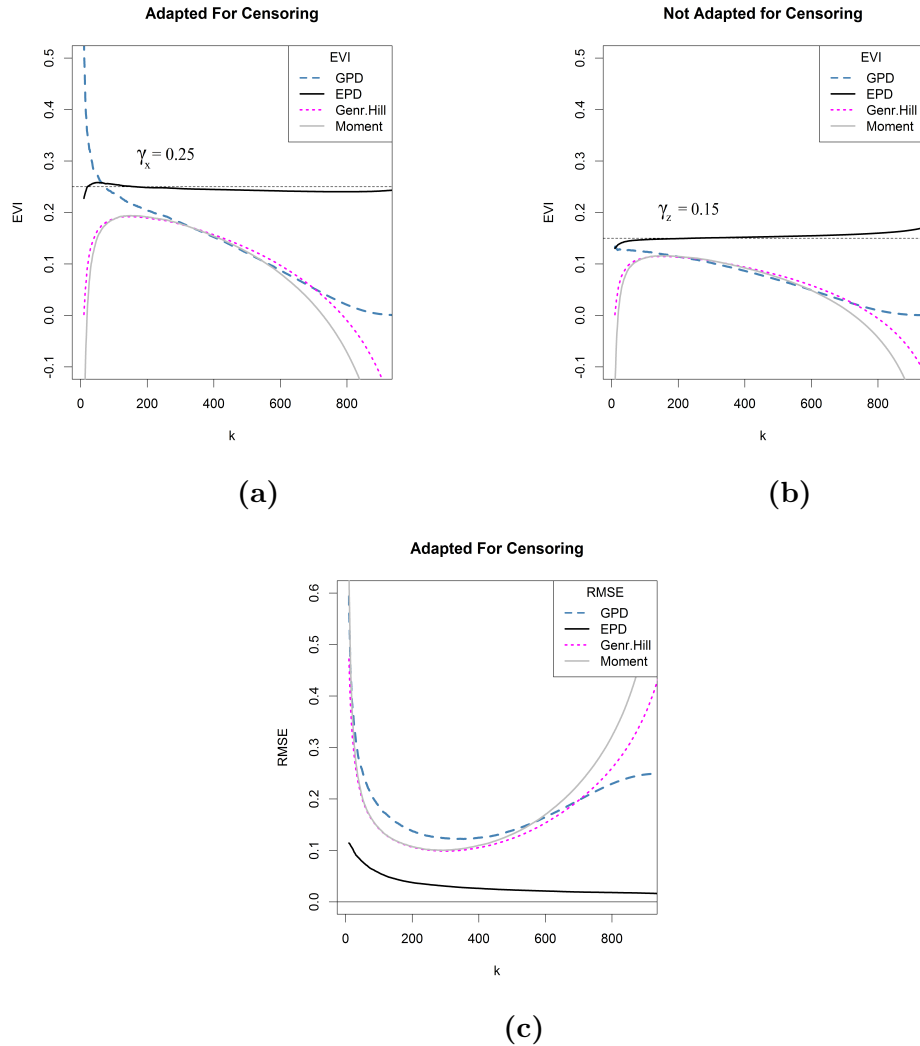


Figure 6.2: A $\text{Fréchet}(\gamma_X = 0.25)$ distribution censored by a $\text{Fréchet}(\gamma_Y = 0.375)$, thus $\gamma_Z = \frac{0.25 \cdot 0.375}{0.25 + 0.375} = 0.15$ and proportion of non-censored data $p = 60\%$ (40% Censoring in the right tail); (a) EVI estimates adapted for censoring and (b) EVI estimate not adapted for censoring, (c) RMSE

From Figure 6.2, we see that the EPD is still doing considerably better than all the other estimators. When a larger portion of the data is censored there is an improvement in bias of the GPD for smaller k . Also, the RMSE of the GPD coincides with that of the semi-parametric estimators for smaller k .

6.3.3 Case 3: 55% Censoring

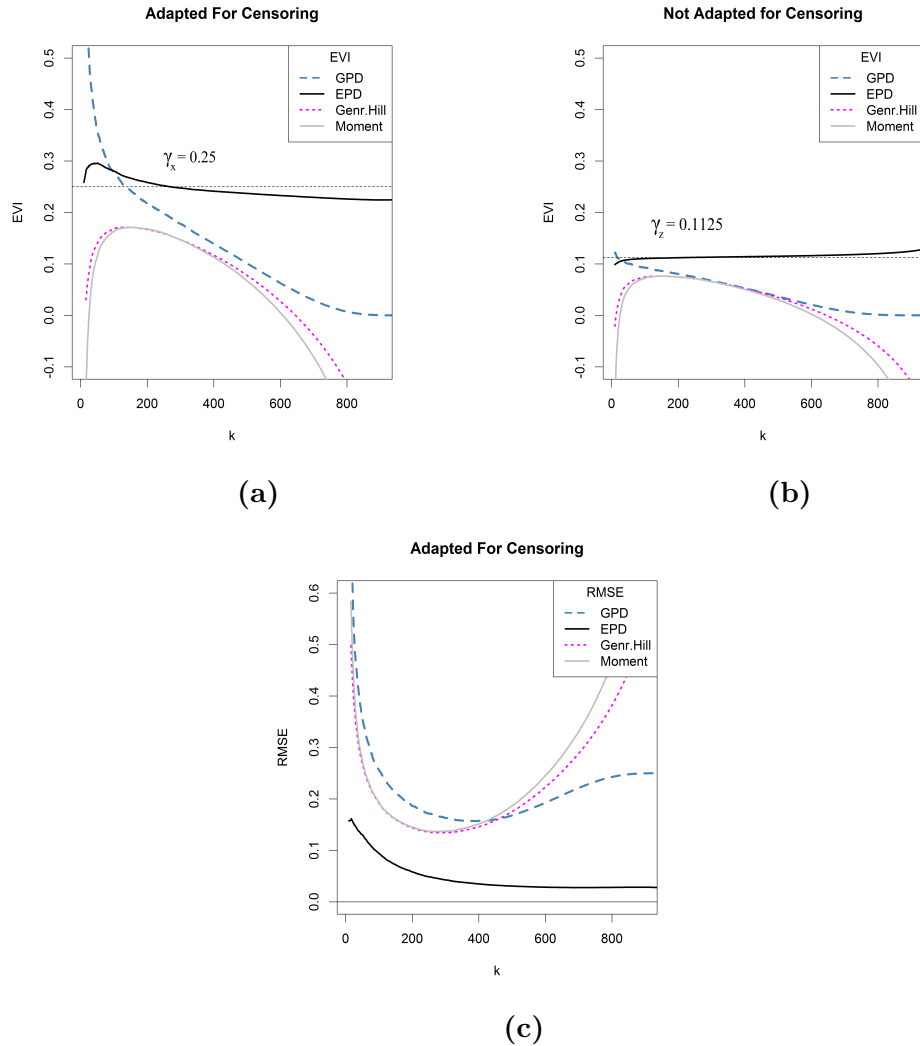


Figure 6.3: A Fréchet($\gamma_X = 0.25$) distribution censored by a Fréchet($\gamma_Y = 0.2046$), thus $\gamma_Z = \frac{0.25 \cdot 0.2046}{0.25 + 0.2046} = 0.1125$ and proportion of non-censored data — $p = 45\%$ (65% Censoring in the right tail); (a) EVI estimates adapted for censoring and (b) EVI estimate not adapted for censoring, (c) RMSE

Lastly in Figure 6.3, we see that for high proportion of censoring all estimates seem to perform relatively poorly, however, the EPD is still sufficiently stable given the amount of censoring. The GPD is performing better than other semi-parametric estimators in terms of the EVI and RMSE. Thus, the two censor-adapted parametric estimators are generally better than the two semi-parametric estimators at high levels of censoring. Most importantly, the EPD offers a more stable and bias reduced estimator of the EVI, under censorship. Note, if we censor X with the same distribution as X , we have 50% censorship in the right tail.

In this next section we apply the mentioned estimators on a real survival dataset to further assess their performance in reality.

6.4 Application to Cancer Survival Data

We consider a data set presented by Klein and Moeschberger (2005) and subsequently analyzed in the context of Extreme Value by Gomes and Neves (2011). This data is from a study conducted on the effects of ploidy on the prognosis of 80 male patients diagnosed with cancers of the tongue. $Z =$ time of death or study completion measured in weeks. We consider both the aneuploid and diploid tumors. Below is a Pareto QQ plot of the Z -sample.

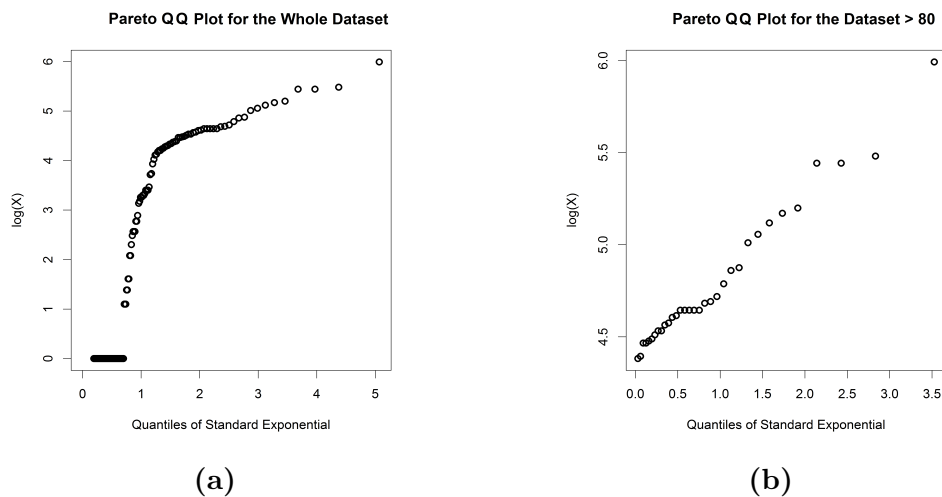


Figure 6.4: Pareto Quantile-Quantile plots for survival data related with cancer of the tongue

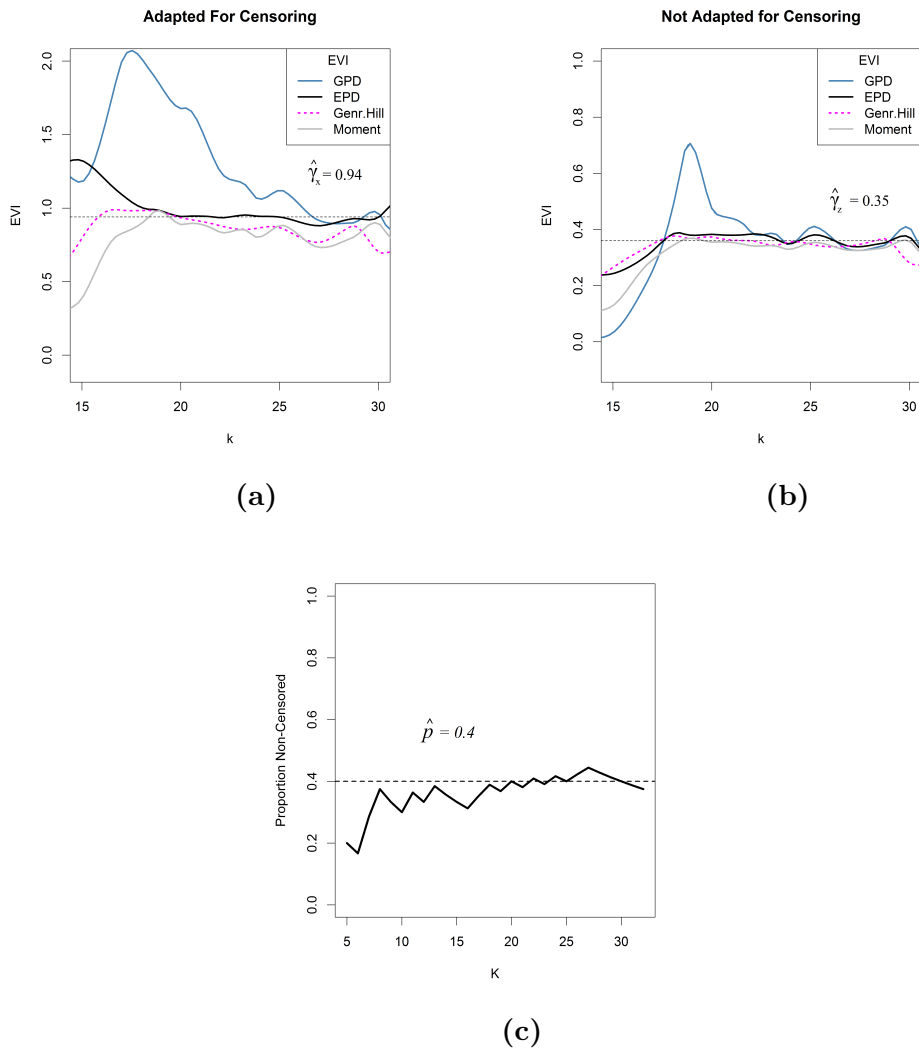


Figure 6.5: Estimates of the EVI for survival data related with cancer of the tongue

From Figure 6.4a we can clearly see that there is evidence of a right heavy tail and we therefore only consider values above 80. Next we estimate $p = \mathbb{P}(\omega = 1|Z = z)$ using Equation 6.10. We use p to choose the optimum k . p has a stable region over $20 \leq k \leq 30$. From this region we can choose an optimal k at $k = 25$, and thus $\hat{p}_{k,n} \approx 0.4$. Therefore approximately 60% of this Z sample is censored, at the point $k = 25$.

As depicted in Figure 6.5 the EPD estimate of the EVI at $k = 25$ is $\gamma_Z = 0.39$ while $\gamma_X = 0.94$. The Generalized Hill and the Moment Estimators also seem to be oscillating around this region. The GPD however becomes more erratic given this high percentage of censoring, just as we saw in the simulation experiment. Using censor adapting estimates of the EVI yields more accurate estimations, we are thus less likely to make adversely incorrect approximations of our quantities of interest.

We can see that if censoring is not taken into account, the EPD estimators leads to more conservative estimates of the EVI, remember in this case that, heavier tails imply larger tail values and thus far more optimistic estimates of survival times of patients with tongue cancer. Moreover, using parametric estimators is only as good as knowing which parametric form the data assumes. In this case, the EPD is a good fit for this Pareto type data, and hence more accurate in approximating estimates of the EVI.

In the next section we establish an alternative estimation procedure for the EPD under random right censorship, by altering the EPD posterior density.

6.5 An asymptotically unbiased Bayesian estimator for random right censored data

In this section, we extend the adapted EPD likelihood approach of Bardoutsos et.al (2015) to construct what Gatarek et al. (2013) refers to as a censored posterior distribution by replacing the likelihood in the posterior, by a censored likelihood. Here we use the same censored likelihood as the one given in Equation 6.22.

It goes beyond the scope of this thesis to analytically evaluate consistency of the censored posterior formulated here. Simulation methods are used instead, and we show that this extended approach can be considered as an alternative to the MLE likelihood adapted estimator as it also leads to much improved results in terms of the bias and MSE.

Quite contrary to what the name censored posterior suggests, it is not the likelihood for a censored data set, where all observations lying outside a particular area are censored. The censored posterior is defined as the product of the conditional densities of the censored observations given the past observations in the right tail of F (where we assume all past observations remain uncensored). Since we are interested in making inference about the right tail of F_X , considering also the fact that X and Y are independent, we do not have to adapt the prior for censoring i.e. we only need prior knowledge of the right tail of F_X . As given in Chapter 5, the Pareto MDI prior is again considered.

The censored EPD posterior is thus given by

$$p^{c,EPD}(\gamma, \delta | \mathbf{z}) = p(\gamma, \delta) \times p^{c,EPD}(\mathbf{z} | \gamma, \delta) \quad (6.23)$$

In the next section we conduct a Monte Carlo simulation to illustrate the behaviour of the censored Bayesian estimator and compare it to the censored ML estimator. The second order parameter ρ is estimated externally using Fraga Alves et al. (2003b)'s algorithm.

6.6 Monte Carlo Simulation: Bayes vs. MLE

We consider again three cases, in the first case only 25% of the data is censored, in the 2nd case, 40% of the data is censored and in the 3rd case, 55% of the data is censored. We generate samples from three heavy-tailed distributions, namely: the Fréchet, Burr and Student-t distribution. In each case we simulate 1000 samples of size $n = 500$ with $EVI = 1$ and $EVI = 0.25$ respectively, each distribution censored by a similar distribution. In the case of the Burr, we vary only the parameter τ . For instance to achieve 25% censoring in the right tail of a Burr distribution, we censor a Burr ($\eta = 1, \tau = 1, \lambda = 1$) with a Burr ($\eta = 1, \tau = 1/3, \lambda = 1$) resulting in samples from $Z \sim Burr(EVI = 0.75)$. The same applies to other distributions.

For each distribution and each estimation method, we compute Monte Carlo estimates of the EVI and the corresponding MSE by averaging out over the 1000 samples. For the Bayesian approach in OpenBUGS, we take 6000 draws of each parameter and discard the first 500.

We show the EVI estimates when all estimators are adapted for censoring, i.e. inferring on the right tail of X , and also the EVI estimates when the estimators are not adapted for censoring, thus making inference on the right tail of Z . We lastly show the Mean Square Error (MSE) between the true estimators of γ_X and the censoring adapted estimators. We included a 90 % HPD with a shaded fill. The second order parameter ρ is estimated using Fraga Alves et al. (2003b)'s method.

6.6.1 Fréchet

Case 1: 25% Censoring

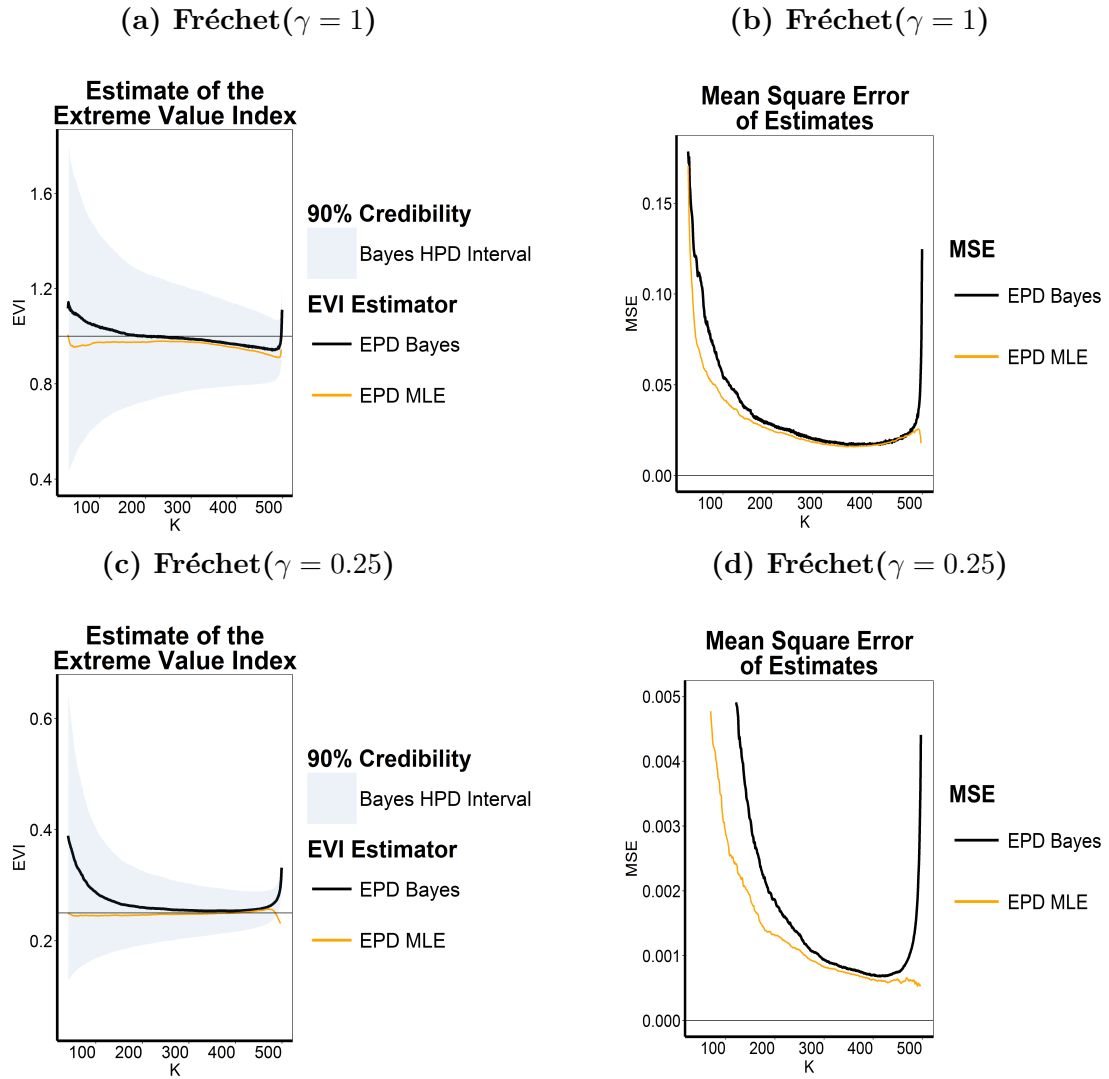


Figure 6.6: Estimates of γ_X (left) and **MSE**(right) for a **Fréchet**(γ_X) distribution censored by another **Fréchet**(γ_Y) at 25% censoring in the right-tail of X

From Figure 6.6 when subjected to 25% censorship in the right tail of the Fréchet distribution we can see that the ML estimator is performing slightly better than the Bayesian estimator in terms of both the bias and the MSE for both heavier tails ($\gamma_X = 1$) and less heavy tails ($\gamma_X = 0.25$).

Case 2: 40% Censoring

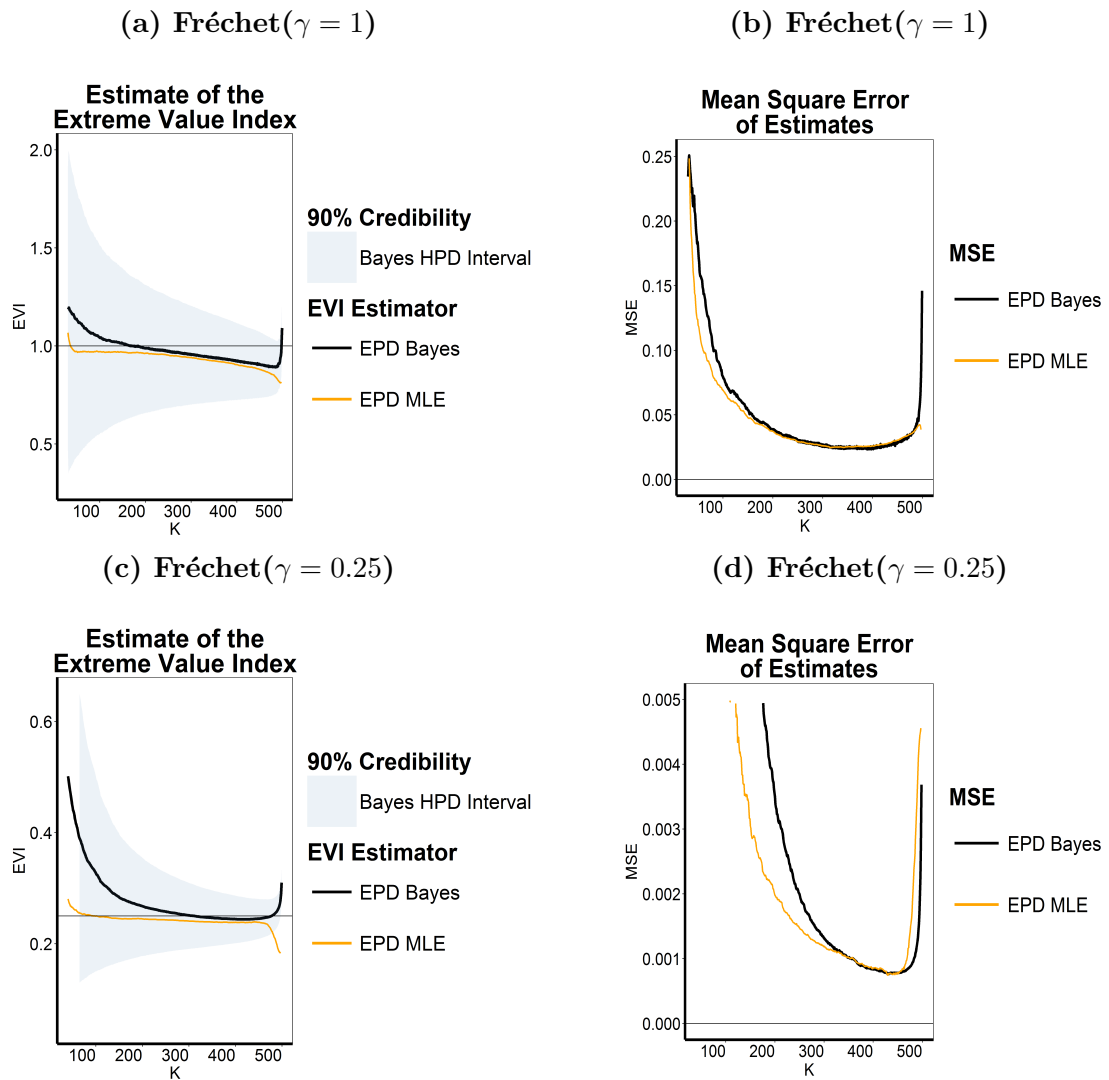


Figure 6.7: Estimates of γ_X (left) and **MSE**(right) for a **Fréchet**(γ_X) distribution censored by another **Fréchet**(γ_Y) at 40% censoring in the right-tail of X

From Figure 6.7 when subjected to 40% censorship in the right tail of the Fréchet distribution, the MSE of the Bayesian estimator now coincides with the ML estimator. And performance of the two estimators is ultimately similar for lower thresholds. The ML estimator still showing much better performance at high thresholds $k \rightarrow 0$.

From Figure 6.8 when subjected to 55% censorship in the right tail of the Fréchet distribution, the Bayesian estimator performs slightly better for heavier tails ($\gamma_X = 1$) in terms of both the bias and MSE. For less heavy tails however, the ML estimator seems to be doing better than the Bayesian estimator only at high thresholds,

while the Bayesian estimator performs better at lower thresholds.

Case 3: 55% Censoring

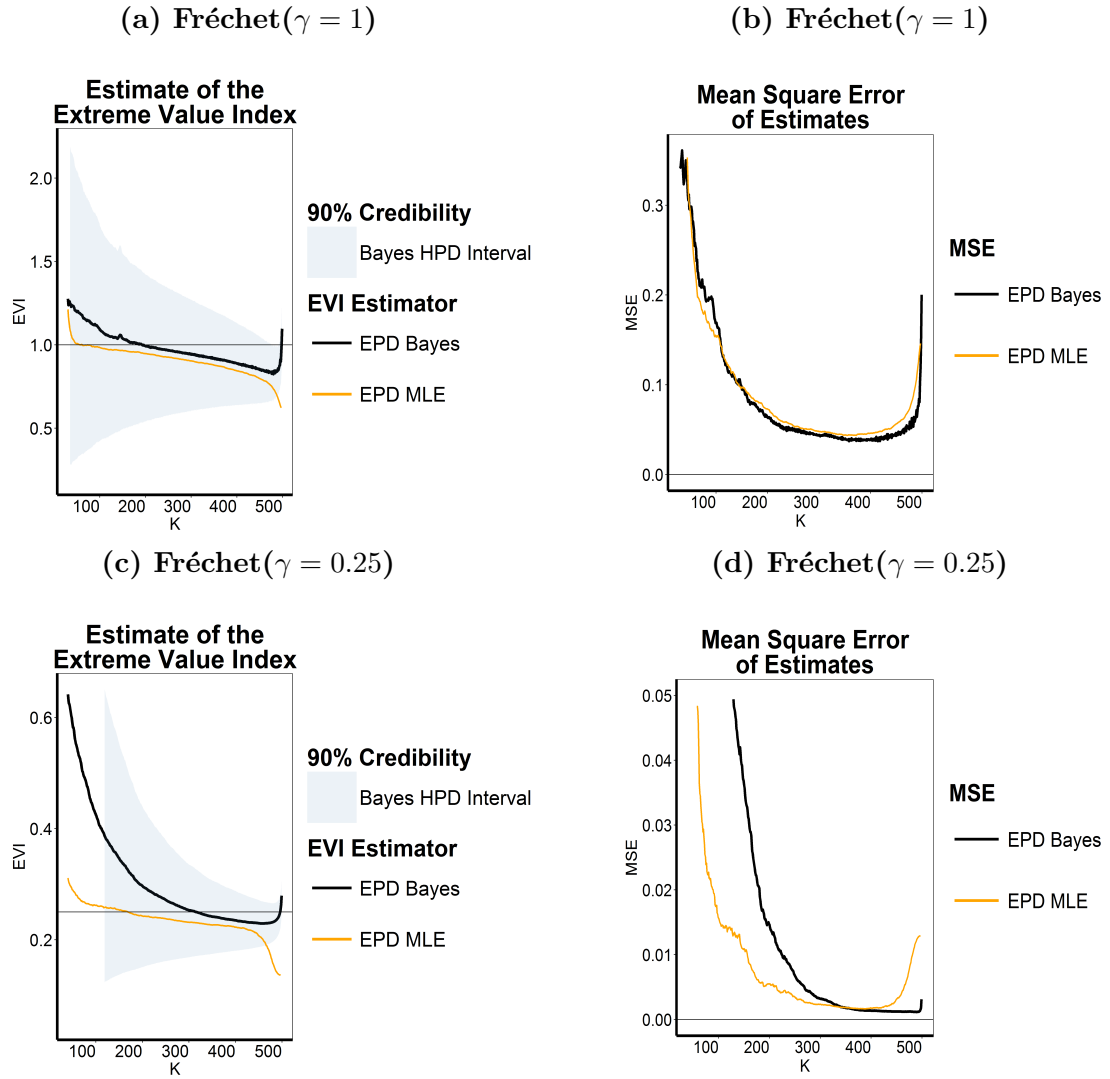


Figure 6.8: Estimates of γ_X (left) and **MSE**(right) for a **Fréchet**(γ_X) distribution censored by another **Fréchet**(γ_Y) at 55% censoring in the right-tail of X

6.6.2 Burr (XII)

Case 1: 25% Censoring

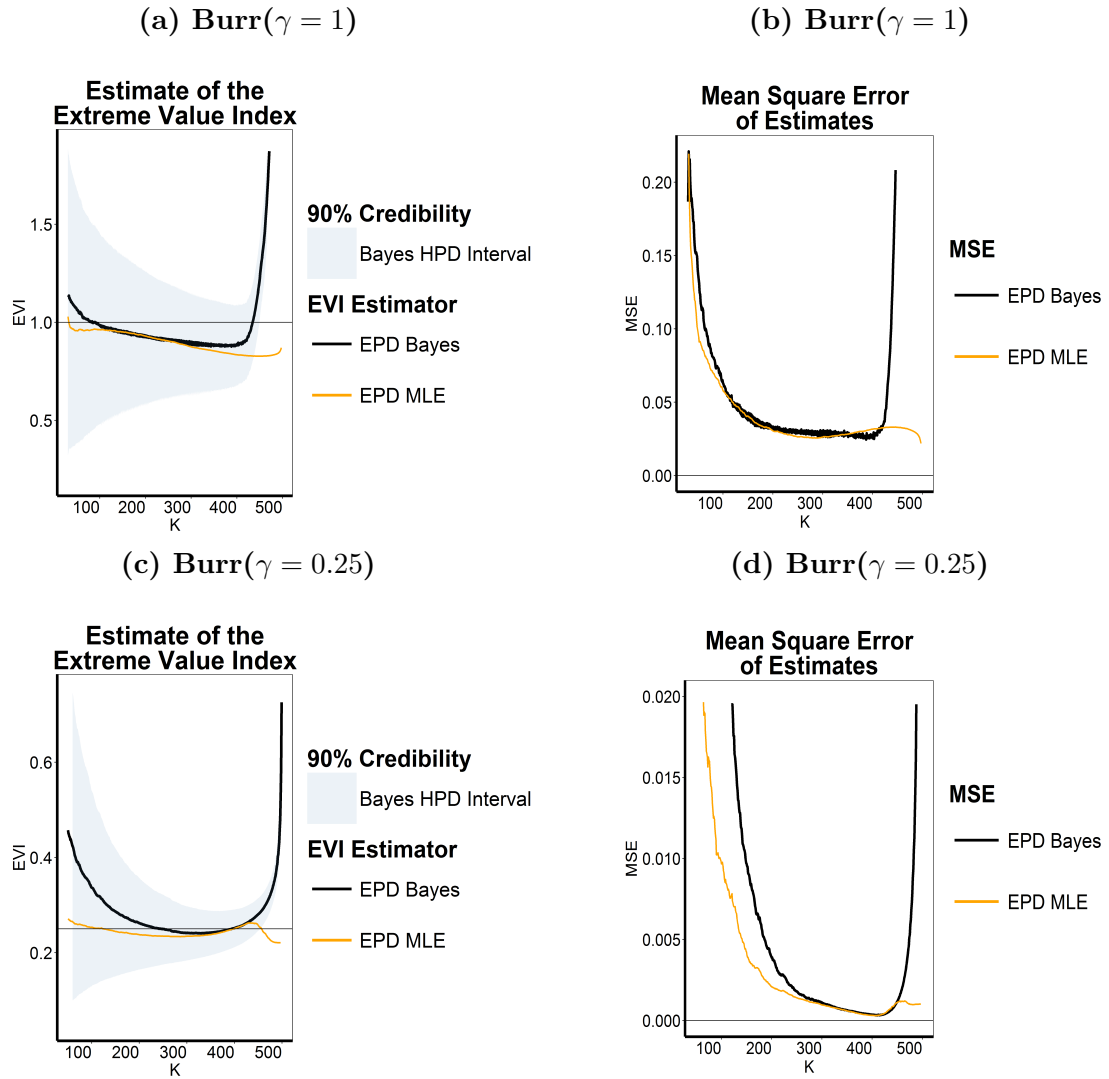


Figure 6.9: Estimates of γ_X (left) and **MSE**(right) for a **Burr**(γ_X) distribution censored by another **Burr**(γ_Y) at 25% censoring in the right-tail of X

From Figure 6.9 when subjected to 25% censoring in the right tail of a Burr distribution, both the Bayesian and ML Estimators offer a somewhat similar performance in estimating the true value of the EVI for both $\gamma_X = 1$ and $\gamma_X = 0.25$, The ML estimator however offers a better performance at high thresholds $k \rightarrow 0$.

Case 2: 40% Censoring

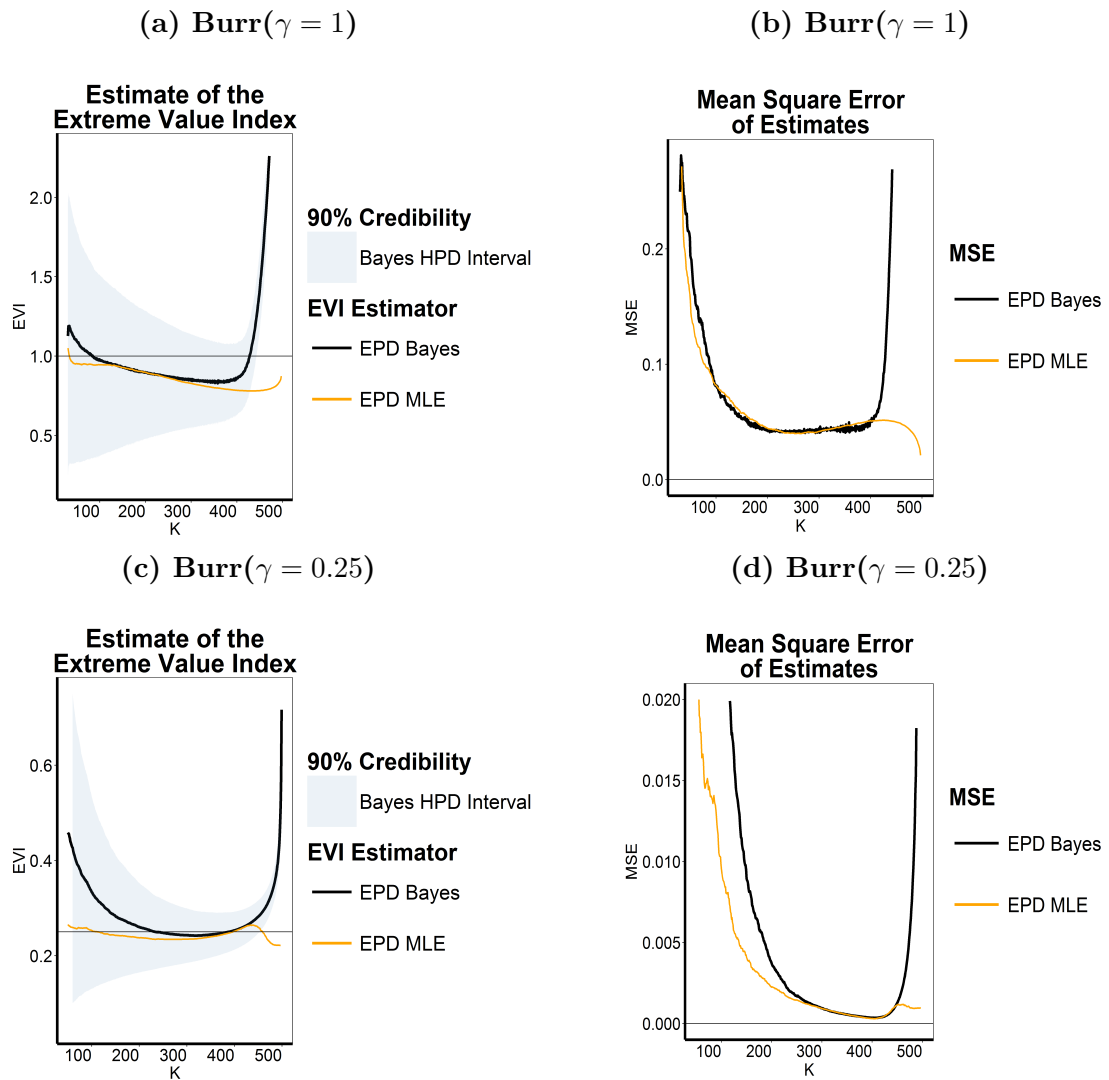


Figure 6.10: Estimates of γ_X (left) and **MSE** (right) for a $\text{Burr}(\gamma_X)$ distribution censored by another $\text{Burr}(\gamma_Y)$ at 40% censoring in the right-tail of X

From Figure 6.10 with 40% censoring in the right tail of the Burr distribution. The Bayesian and ML estimator offer a similar performance, however as we can see from Figure 6.9, for smaller EVI values, the ML estimator seems to be performing only slightly better than the Bayesian Estimators, particularly at high thresholds.

Case 3: 55% Censoring

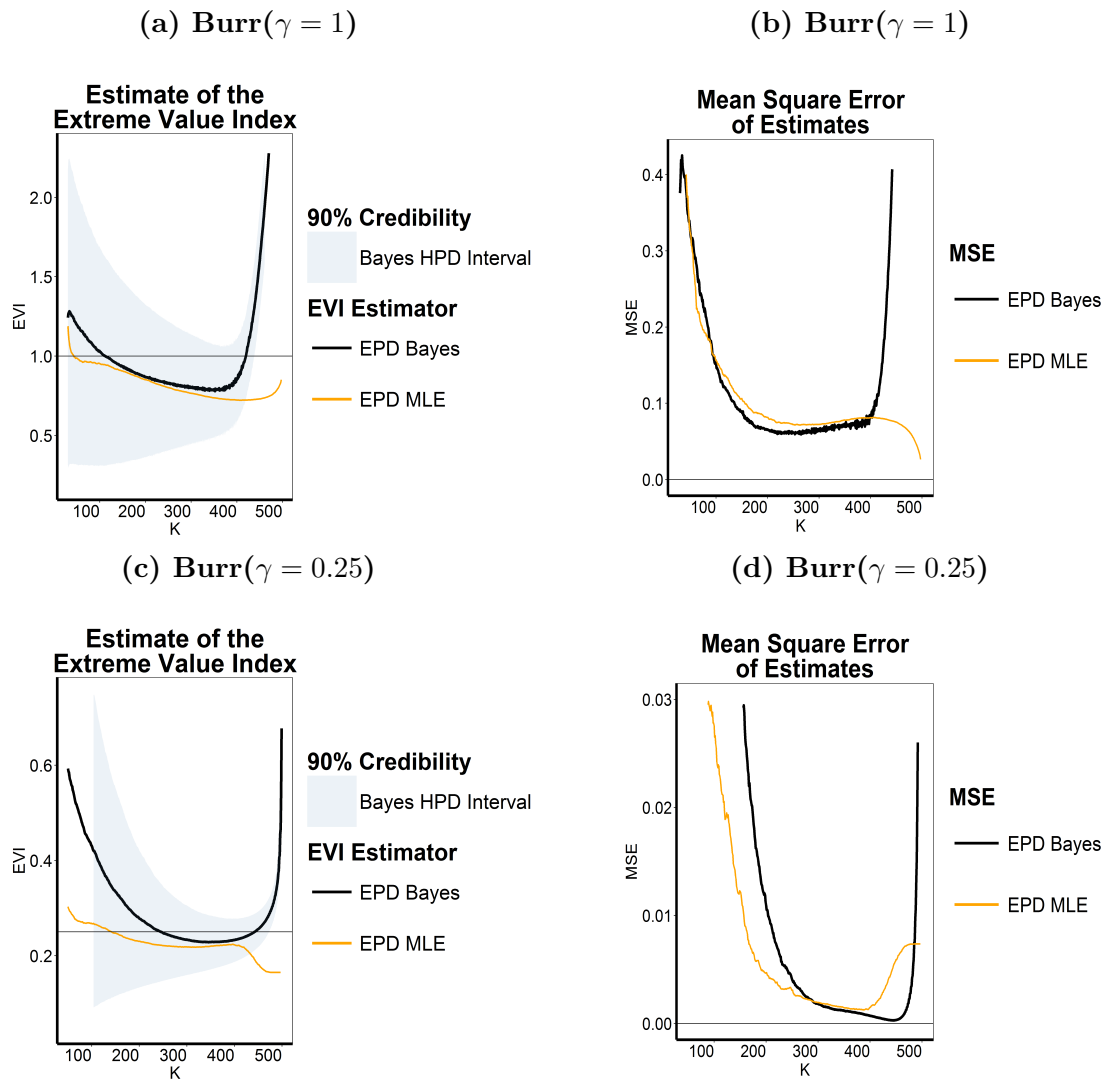


Figure 6.11: Estimates of γ_X (left) and **MSE** (right) for a **Burr**(γ_X) distribution censored by another **Burr**(γ_Y) at 55% censoring in the right-tail of X

From Figure 6.11 subjected to 55% censorship in the right tail of the Burr distribution, we can see that for heavier tails $\gamma_X = 1$, the Bayesian estimator is doing considerably better than the ML estimator in terms of the bias and MSE. For less heavy tails ($\gamma_X = 0.25$) the MSE of the two estimators coincides, with the Bayesian estimator ultimately better at lower thresholds.

6.6.3 Student-t

Case 1: 25% Censoring

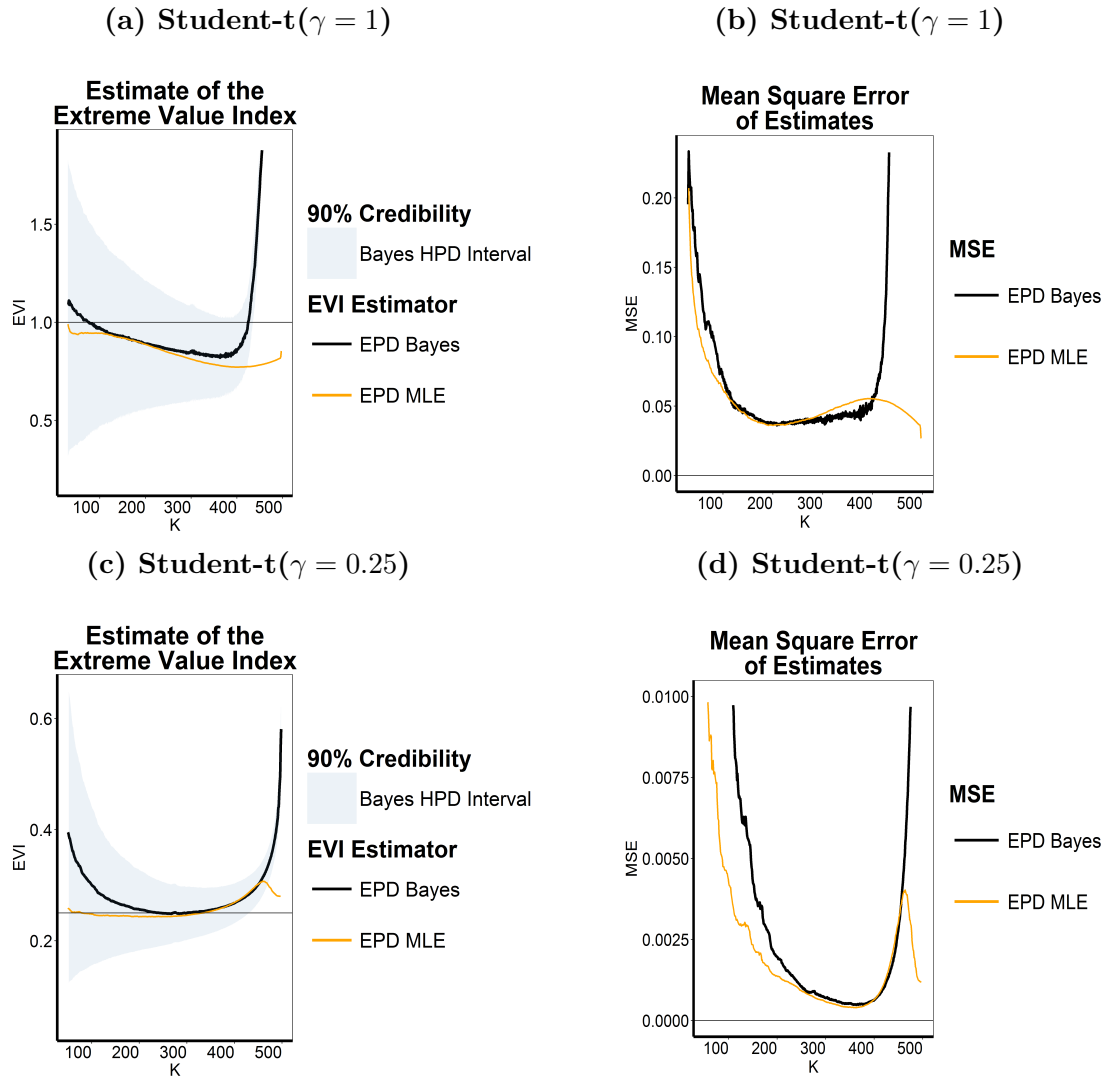


Figure 6.12: Estimates of γ_X (left) and MSE(right) for a **Student-t**(γ_X) distribution censored by another **Student-t**(γ_Y) at 25% censoring in the right-tail of X

Keeping in mind that the Student-t distribution is heavier tailed than both the Fréchet and the Burr distribution. From Figure 6.12 subjected to 25% censorship in the right tail of the Student-t distribution, we can see that both the ML and the Bayesian estimator coincide for both heavier tails ($\gamma_X = 1$) and less heavy tails ($\gamma_X = 0.25$). The ML estimator however still performing slightly better than the Bayesian estimator at high thresholds ($k \rightarrow 0$).

Case 2: 40% Censoring

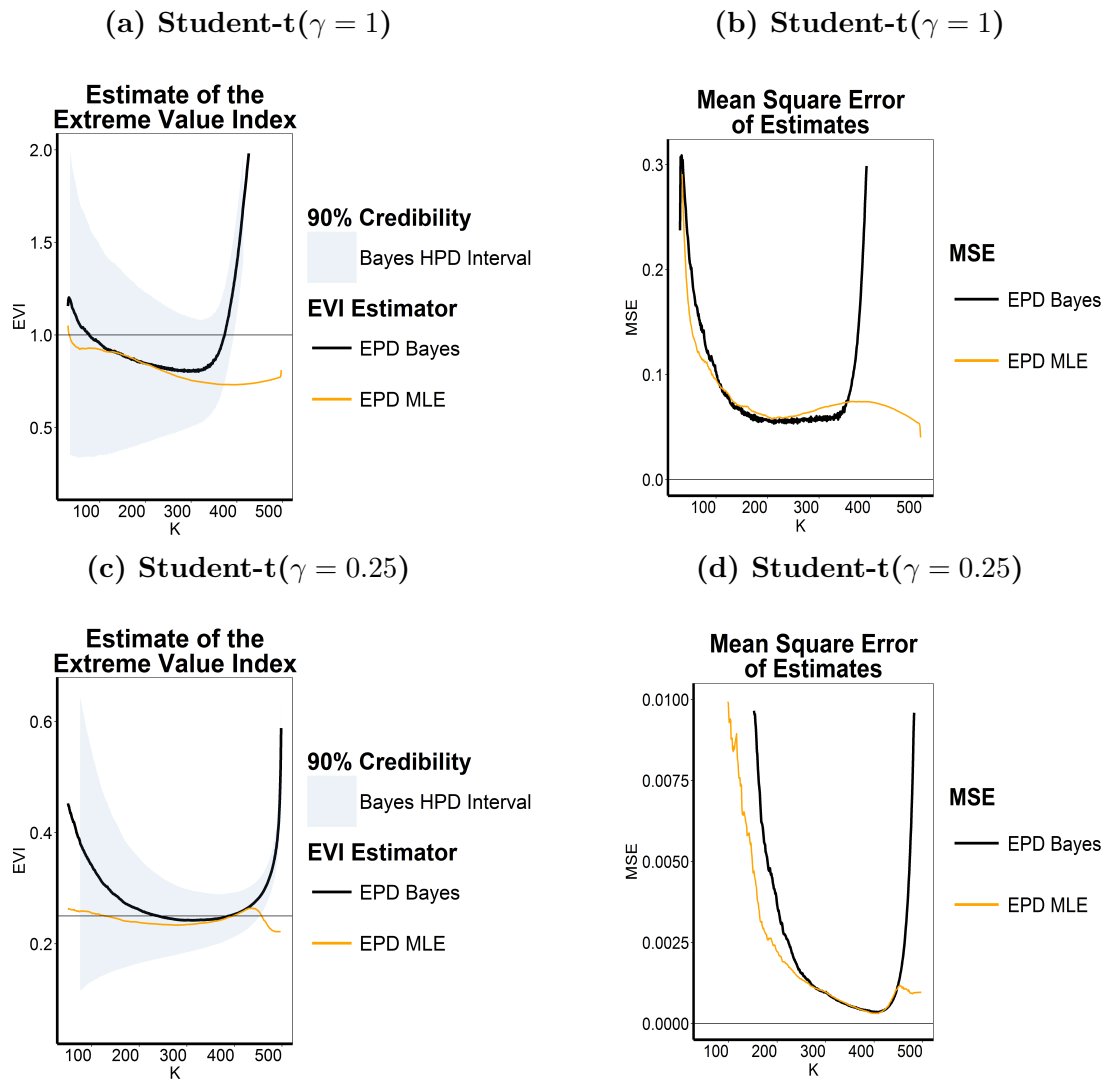


Figure 6.13: Estimates of γ_X (left) and **MSE**(right) for a **Student-t**(γ_X) distribution censored by another **Student-t**(γ_Y) at 40% censoring in the right-tail of X

From Figure 6.13 subjected to 40% censorship in the right tail of the Student-t distribution, the Bayesian estimator seems to be performing slightly better than the ML estimator, both in terms of bias and MSE. The ML estimator is still performing considerably better for heavier tails at high thresholds ($k \rightarrow 0$)

Case 3: 55% Censoring

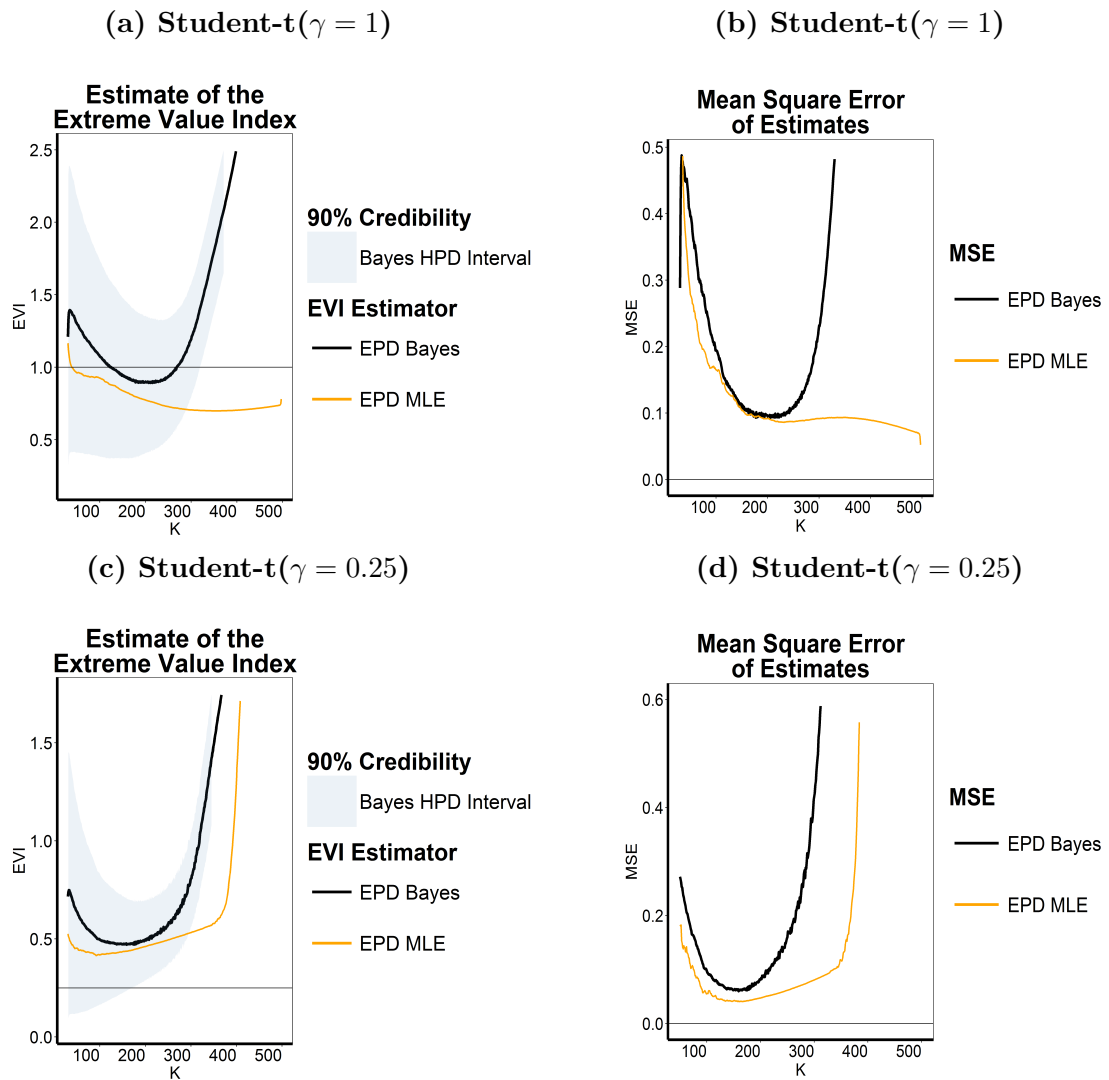


Figure 6.14: Estimates of γ_X (left) and **MSE**(right) for a **Student- $t(\gamma_X)$** distribution censored by another **Student- $t(\gamma_Y)$** at 55% censoring in the right-tail of X

From Figure 6.14 subjected to 55% censorship in the right tail of the Student- t distribution, the Bayesian estimator is performing much better than the ML estimator particularly for heavier tails ($\gamma_X = 1$). For less heavy tails, the MSE of both estimators coincide.

6.7 Conclusion

We investigated the EVI estimators under random right censorship. Through some numerical studies using the Limited Memory Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm, we were able to implement the parametric EPD censored likelihood estimator of Beirlant et al. (2016).

Parametric estimators generally enjoy a considerable advantage of being invariant to location shifts in the data, as opposed to semi-parametric estimation which can be very sensitive to location shifts in the data. However, an important requisite to employing parametric approaches, is that we know the parametric form of the data for which we wish to make inference about the tail. The parametric form can sometimes be difficult to recognize. In order to assess and illustrate the behavior of the chosen parametric and semi-parametric estimators of the EVI, we conducted a simulation study.

Judging from the EPD’s predominating overall performance in estimating the EVI under random right censorship, we make a concluding recommendation of using the EPD to model randomly censored **Pareto-Type** data. In Section 6.5 we constructed a censored posterior density function for the EPD, and conducted a simulation study in Section 6.6 to assess the performance of this estimator in comparison to the ML estimator. From this simulation study it was seen that the ML estimator generally performs better than the Bayesian estimator for higher thresholds as $k \rightarrow 0$, while the Bayesian estimator performs better at lower thresholds.

Also it was seen that the Bayesian estimator performs better than the ML estimator for heavier tailed distributions, while the ML estimator performs better for less heavier tailed distributions. Lastly it was apparent the more the data is censored the better the Bayesian estimator is at estimating the true EVI than the ML estimator.

We therefore recommend using the Bayesian estimator under the following circumstances:

- The data is very heavy tailed ($\gamma > 0.5$).
- The data points are few, and you thus wish to use lower thresholds.
- The data is overly censored in the right tail (more than 50% censored).
- Lastly and more importantly, when expert opinion can be elicited and incorporated into a prior distribution.

In the next chapter we look again at the estimators mentioned in previous chapters and assess their performance at contaminated data. And also investigate the newly developed robust method of estimating EPD parameters –The Minimum Density Power Divergence Estimator (MDPDE).

Chapter 7

Estimation of the EVI for contaminated data

7.1 Introduction

In this chapter we investigate the performance of the different estimation methods mentioned in previous chapters when applied to contaminated data. Throughout this study, we make an explicit assumption about the data being independent and identically distributed, thus making these observations mathematically convenient to work with. In practice however, these assumptions are not precisely true and reality is a much more complex system. The previous chapter discussed one such complexity where data is randomly censored. When modelling randomly censored excesses with the EPD, the effects of censoring can be overcome by adapting the model's likelihood.

All estimation procedures discussed in previous chapters are however not robust against deviations of the true distribution from the assumed parametric model. As shown in Section 5.3.4, for the Bayesian estimator the choice of the prior can be an undesirable source of ambiguity. It can be seen from Section 5.3.4 that the Bayesian estimator is not robust against misspecification of the prior and as will be seen in this chapter, the Bayesian, ML, and Hill estimators are not robust against model contamination. An additional method of estimation also investigated in this chapter is the Minimum Density Power Divergence Estimator (MDPDE), which was shown by Dierckx et al. (2013) to be robust against outliers.

This chapter is organized as follows: In Section 7.2 we briefly define the MDPDE. In Section 7.2.1 we provide in detail, the estimation procedures using the MDPDE. In Section 7.3 we conduct an exhaustive simulation study, to assess the performance of these estimates. A case study is carried out in Section 7.4 to further assess the behaviour of the MDPDE in comparison to other mentioned estimators when data is both contaminated and uncontaminated. Some concluding remarks

are made in Section 7.5

7.2 Minimum Density Power Divergence Estimator

This section discusses the Minimum Density Power Divergence Estimator (MDPDE) as a method for estimating parameters of the EPD model. Dierckx et al. (2013) introduced the MDPDE in an effort to construct a robust and asymptotically unbiased estimator of the EVI. The idea of the density power divergence for developing a robust estimation criterion was first introduced by Basu et al. (1998). In this chapter we will assess the performance of the MDPDE and compare it to other estimators mentioned in previous chapters, namely; the ML, Bayesian and Hill estimator. Using all four estimators of the EVI, we will further estimate high quantiles associated with some probability p as defined by Goegebeur et al. (2014). We also consider the Weissman (1978) estimator given as

$$\hat{U}\left(\frac{1}{p_n}\right) = X_{n-k,n} \left(\frac{k+1}{(n+1)p} \right)^{H_{k,n}} \quad (7.1)$$

where $H_{k,n}$ is the Hill (1975) estimator.

The density power divergence between the density functions f and g is given by Dierckx et al. (2013) as

$$\Delta_\alpha(f, g) := \begin{cases} \int_{\mathbb{R}} [g^{1+\alpha}(y) - (1 + \frac{1}{\alpha})g^\alpha(y)f(y) + \frac{1}{\alpha}f^{1+\alpha}(y)]dy, & \alpha > 0, \\ \int_{\mathbb{R}} \log \frac{f(y)}{g(y)} f(y)dy, & \alpha = 0 \end{cases} \quad (7.2)$$

Assume that the density function g depends on a parameter vector θ , and let Y_1, \dots, Y_n be independent and identically distributed random variables according to the density function f . The MDPDE is the point $\hat{\theta}$ that minimizes the empirical density power divergence

$$\hat{\Delta}_\alpha(\theta) := \int_{\mathbb{R}} g^{1+\alpha}(y)dy - \left(1 + \frac{1}{\alpha}\right) \frac{1}{n} \sum_{i=1}^n g^\alpha(Y_i) \quad (7.3)$$

for $\alpha > 0$, and

$$\hat{\Delta}_0(\theta) := -\frac{1}{n} \sum_{i=1}^n \log g(Y_i) \quad (7.4)$$

for $\alpha = 0$, the latter corresponds with the negative log-likelihood function. The parameter α controls the trade-off between robustness and efficiency of the MDPDE: The estimator becomes less robust against outliers and more efficient as $\alpha \rightarrow 0$.

7.2.1 Estimation procedure

Let the sample X_1, \dots, X_n be independent and identically distributed from a distribution function satisfying (\mathcal{R}) , and denote the ascending order statistics by $X_{1,n} \leq \dots \leq X_{n,n}$. Dierckx et al. (2013) showed that the MDPDE for (γ, δ) satisfies the system of equations:

$$0 = \int_1^\infty g^\alpha(y) \frac{\partial g(y)}{\partial \gamma} dy - \frac{1}{k} \sum_{j=1}^n g^{\alpha-1}(Y_j) \frac{\partial g(Y_j)}{\partial \gamma} \quad (7.5)$$

and

$$0 = \int_1^\infty g^\alpha(y) \frac{\partial g(y)}{\partial \delta} dy - \frac{1}{k} \sum_{j=1}^n g^{\alpha-1}(Y_j) \frac{\partial g(Y_j)}{\partial \delta} \quad (7.6)$$

where $g(y)$ is the density function of the EPD, given by:

$$g(y) = \frac{1}{\gamma} y^{-1/\gamma-1} \{1 + \delta(1 - y^\tau)\}^{-1/\gamma-1} [1 + \delta\{1 - (1 + \tau)y^\tau\}] \quad (7.7)$$

Both these equations depend also on the unknown parameter τ . As noted by Dierckx et al. (2013), estimating the second order parameter ρ externally leads to bias-corrected estimators of γ with a smaller asymptotic variance compared to a joint estimation of (γ, δ, τ) . In this section, we adopt the parameterization $\tau = \rho/\gamma$ and use an external estimator to estimate the second order parameter ρ for all estimation procedures.

Assuming that τ is known, Dierckx et al. (2013) stated and proved the asymptotic normality of the sequence of consistent solutions of Equations 7.5 and 7.6, and further showed that by replacing ρ by an external consistent estimator leads to the same limiting distribution as in the case where the true ρ is known. For brevity, proofs of asymptotic properties of the MDPDE are omitted in this study (see Dierckx et al., 2013, for more details).

7.3 Simulation Study

In this section we study the finite sample behavior of all estimators mentioned in previous chapters and compare them to the MDPDE. To investigate the optimum trade off between robustness and efficiency of the MDPDE, we consider three cases of the MDPDE $\gamma_{n,\alpha}$ where $\alpha = 0.1$, $\alpha = 0.5$ and $\alpha = 1$. We also consider estimates of high quantiles associated with $p = 1/500$ and the corresponding Mean Square Error (MSE): $(\hat{U}_*(1/p_n)/U(1/p_n) - 1)^2$, where $\hat{U}_*(1/p_n)$ denotes any of the considered estimators of $U(1/p_n)$ as a function of k , the true quantile is indicated with a straight line.

We denote $\hat{\gamma}_n^{Bayes}$ as the Bayesian EVI estimator and $\hat{U}_*^{Bayes}(1/p_n)$ as the Bayesian Quantile estimator. We denote in a similar manner the ML and the Hill estimators.

In Section 7.3.1 1,000 Samples of size 200 and 500 are generated from three different Pareto type distributions given in Table 2.2, namely: the Fréchet, Burr (XII) and Student-t distribution all with $\gamma = 0.5$. In Section 7.3.2, 1,000 contaminated samples of size 500 are generated from the three previously mentioned distributions with $\gamma = 0.5$. The plots are obtained by averaging out over the 1,000 samples.

7.3.1 Uncontaminated Distributions

To assess the behavior of estimators at uncontaminated data we simulate 1,000 samples of sizes 200 and 500 respectively from three Pareto type distributions mentioned above with $\text{EVI} = 0.5$. Table 7.1 shows the different settings of the simulation study at uncontaminated data.

Table 7.1: Simulation setting for uncontaminated data

Distribution F	Quantity	Figure
Fréchet ($\alpha = 2$)	$\hat{\gamma}$ and $\hat{U}(1 - 1/500)$	7.1
	$\hat{\gamma}^{MSE}$ and $\hat{U}^{MSE}(1 - 1/500)$	7.2
Burr (XII) ($\eta = 1, \tau = 2, \lambda = 1$)	$\hat{\gamma}$ and $\hat{U}(1 - 1/500)$	7.3
	$\hat{\gamma}^{MSE}$ and $\hat{U}^{MSE}(1 - 1/500)$	7.4
Student-t ($df = 2$)	$\hat{\gamma}$ and $\hat{U}(1 - 1/500)$	7.5
	$\hat{\gamma}^{MSE}$ and $\hat{U}^{MSE}(1 - 1/500)$	7.6

Fréchet

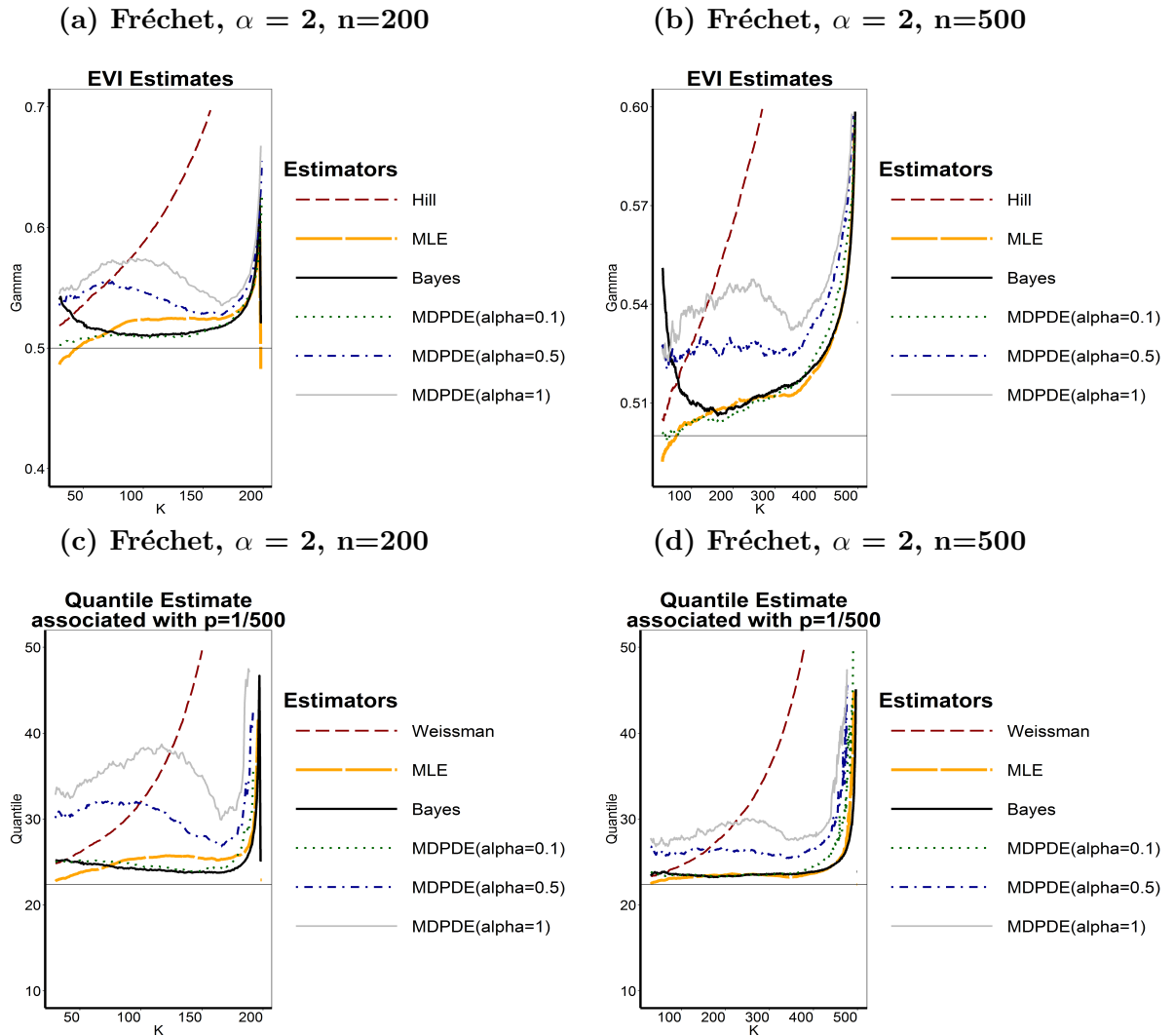


Figure 7.1: Estimates of the **EVI** (top) and **U(1-1/500)** (bottom) of a **Fréchet** sample with $\text{EVI} = 0.5$.

From Figure 7.1 we can see that for uncontaminated samples $\hat{\gamma}_n^{Bayes}$, $\hat{\gamma}_n^{MLE}$ and $\hat{\gamma}_{0.1,n}$ are performing considerably better than other estimators in terms of bias for both smaller and larger samples. The same is true in the case of quantile estimates, however for smaller samples $\hat{U}_*^{Bayes}(1/p_n)$ seems to be performing slightly better than all considered estimators.

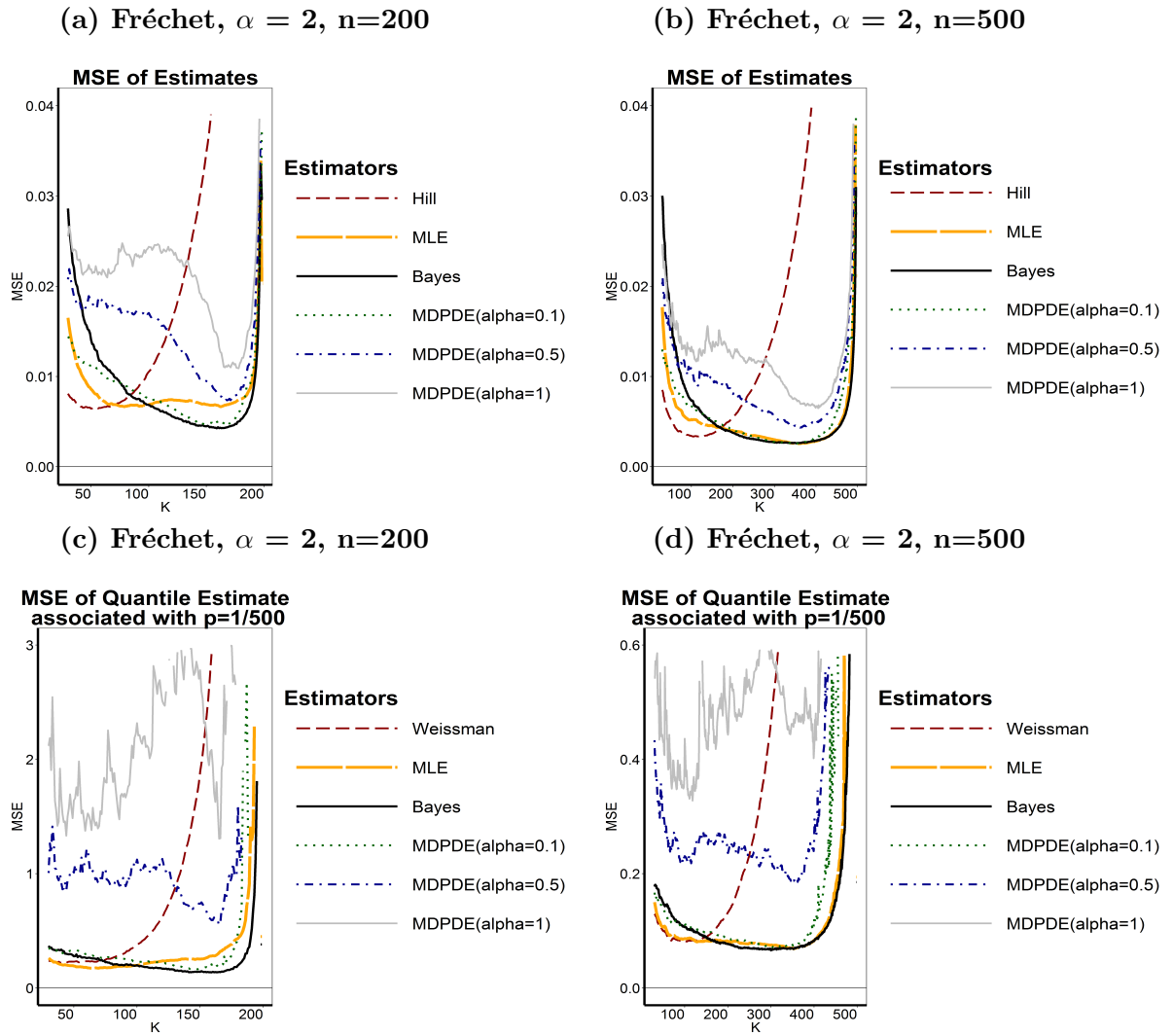


Figure 7.2: MSE of EVI estimates (top) and $U(1-1/500)$ (bottom) of a Fréchet sample with $EVI = 0.5$.

From Figure 7.2 it is more apparent that $\hat{\gamma}_n^{Bayes}$ and $\hat{U}_*^{Bayes}(1/p_n)$ seem to outperform other considered estimators in terms of the MSE for lower thresholds $k \rightarrow \infty$. It is also clear that $\hat{\gamma}_{1,n}$ and $\hat{\gamma}_{0.5,n}$ are under performing.

Burr

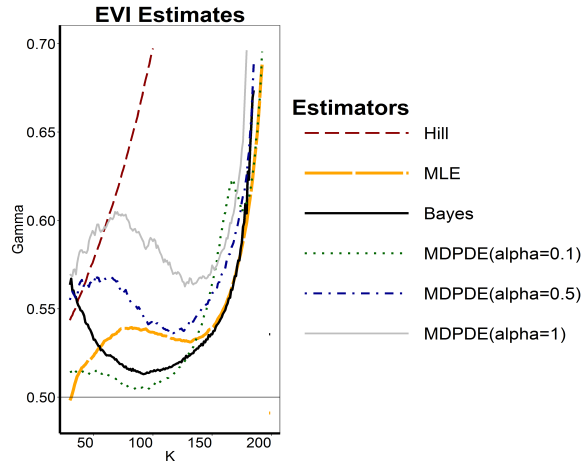
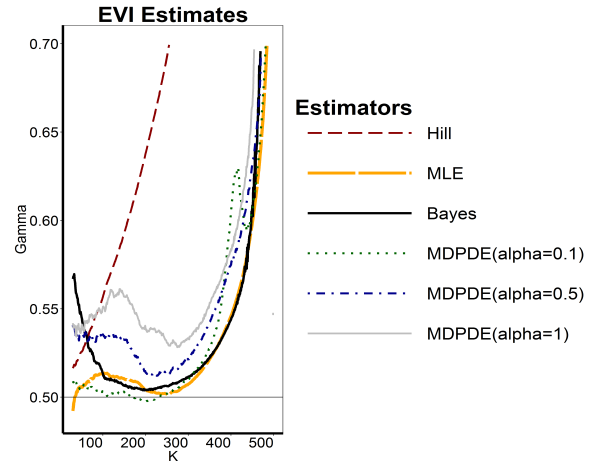
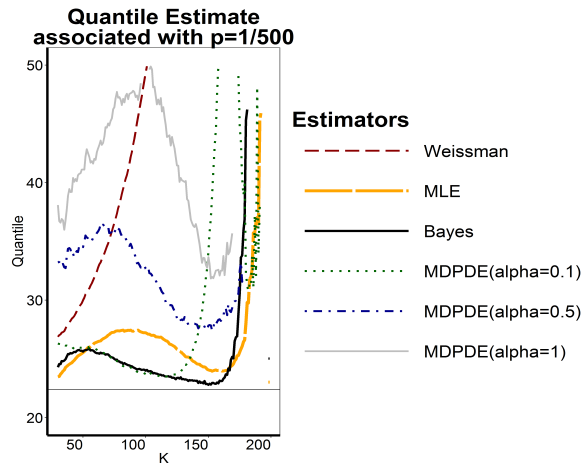
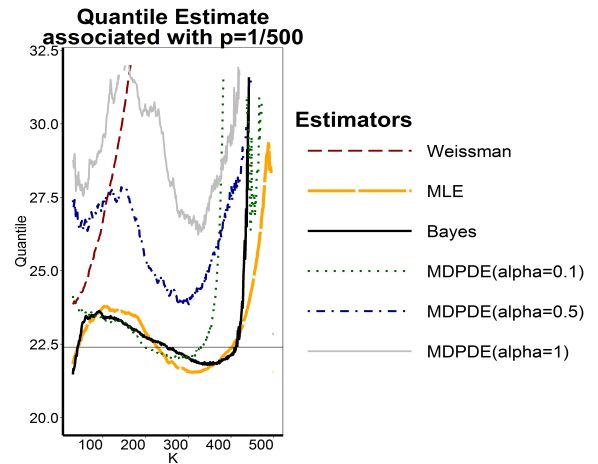
(a) Burr (XII), $\eta = 1, \tau = 2, \lambda = 1$, $n=200$ (b) Burr (XII), $\eta = 1, \tau = 2, \lambda = 1$, $n=500$ (c) Burr (XII), $\eta = 1, \tau = 2, \lambda = 1$, $n=200$ (d) Burr (XII), $\eta = 1, \tau = 2, \lambda = 1$, $n=500$ 

Figure 7.3: Estimates of the **EVI** (top) and **U(1-1/500)** (bottom) of a **Burr (XII)** sample with $\text{EVI} = 0.5$.

Firstly we know that Burr samples are generally heavier tailed than Fréchet samples, and from Figure 7.3 we can see that $\hat{\gamma}_{0.1,n}$ seems to be performing much better than all considered estimators for both smaller and larger samples. $\hat{U}_*^{Bayes}(1/p_n)$ and $\hat{U}_*^{MLE}(1/p_n)$ are also performing reasonable well. Performance of $\hat{U}_*^{Bayes}(1/p_n)$ can be accredited to the efficient estimation of δ in approximating the posterior.

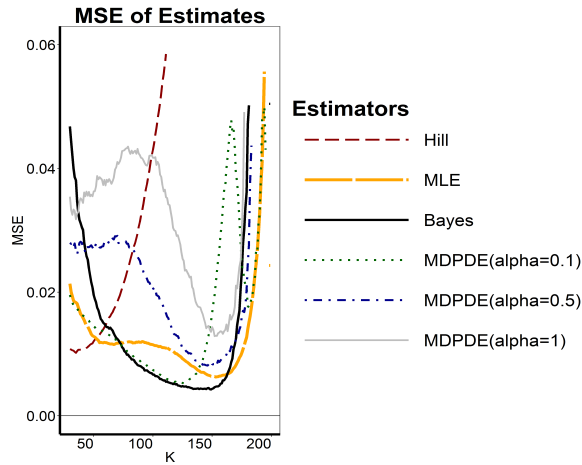
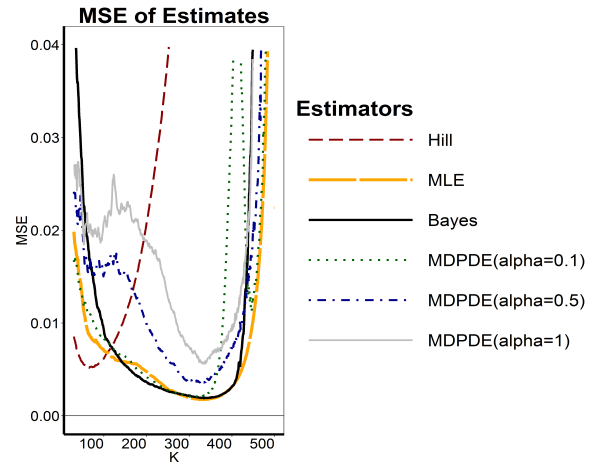
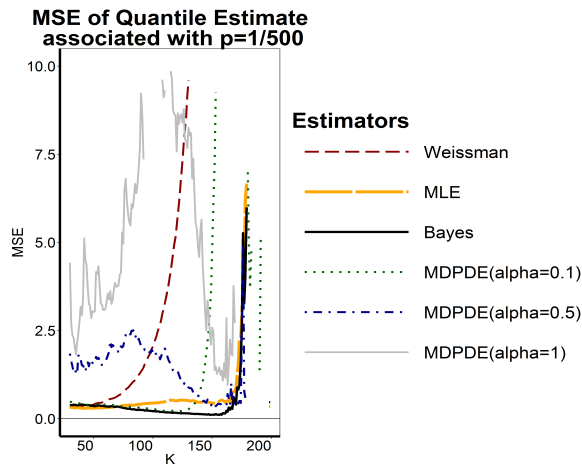
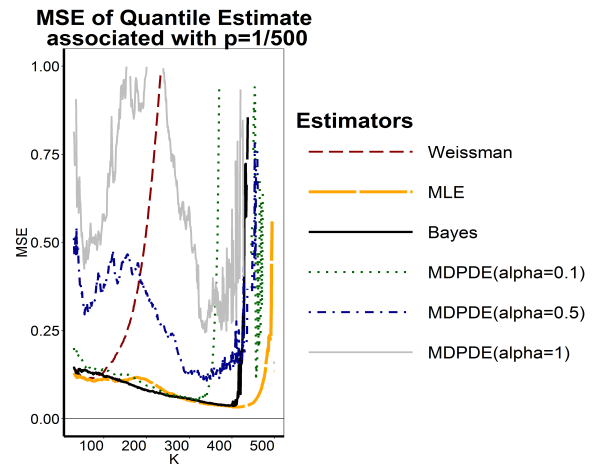
(a) Burr (XII), $\eta = 1, \tau = 2, \lambda = 1$, $n=200$ (b) Burr (XII), $\eta = 1, \tau = 2, \lambda = 1$, $n=500$ (c) Burr (XII), $\eta = 1, \tau = 2, \lambda = 1$, $n=200$ (d) Burr (XII), $\eta = 1, \tau = 2, \lambda = 1$, $n=500$ 

Figure 7.4: MSE of EVI estimates (top) and $U(1-1/500)$ (bottom) of a Burr (XII) sample with $EVI = 0.5$.

From Figure 7.4 we can see that in terms of the MSE, $\hat{\gamma}_n^{Bayes}$ and $\hat{U}_*^{Bayes}(1/p_n)$ as well as $\hat{\gamma}_{0.1,n}$ and $\hat{U}_{0.1,n}^{MDPDE}(1/p_n)$ are performing considerably better than all considered estimators. We can also spot a decent performance from $\hat{\gamma}_n^{MLE}$ particularly in larger samples.

Student-t

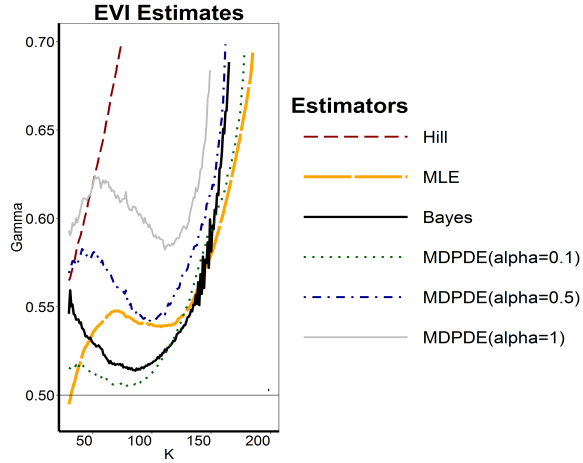
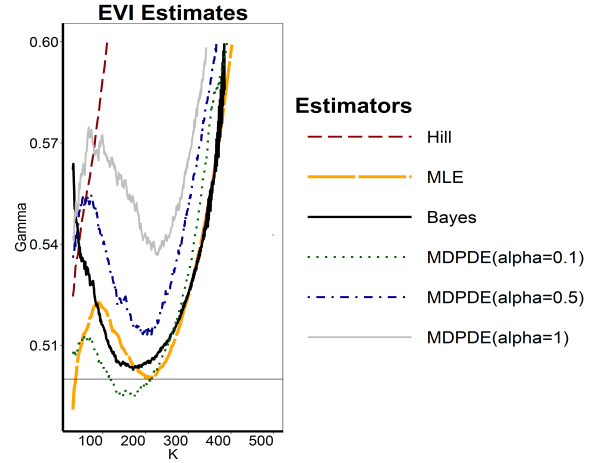
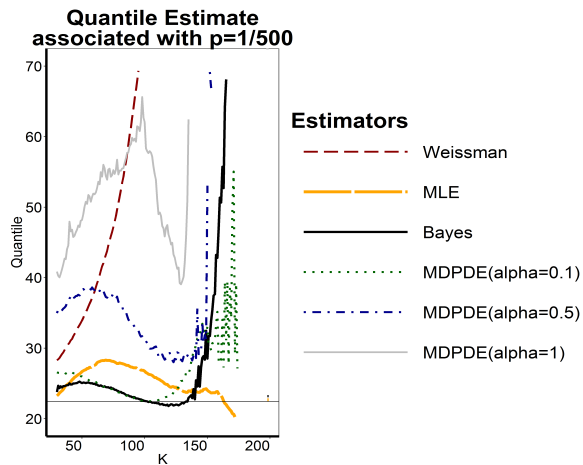
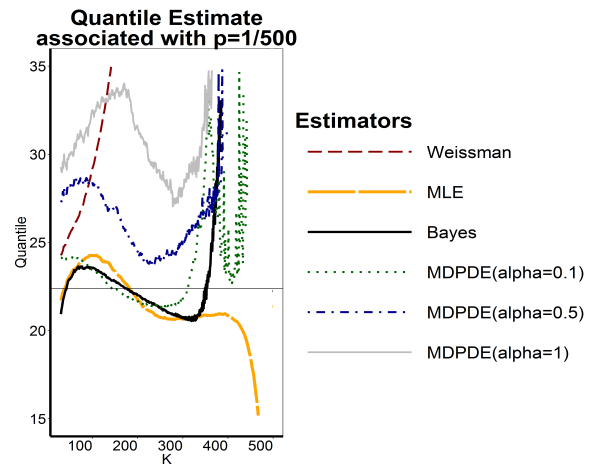
(a) Student-t, $df = 2$, $n=200$ (b) Student-t, $df = 2$, $n=500$ (c) Student-t, $df = 2$, $n=200$ (d) Student-t, $df = 2$, $n=500$ 

Figure 7.5: Estimates of the **EVI** (top) and **U(1-1/500)** (bottom) of a **Student-t** sample with **EVI = 0.5**.

From Figure 7.5, $\hat{\gamma}_{0.1,n}$ is performing considerably better than all considered estimators for both smaller and larger samples. However $\hat{U}_*^{Bayes}(1/p_n)$ seems to outperform $\hat{U}_{0.1,n}^{MDPDE}$ for smaller samples.

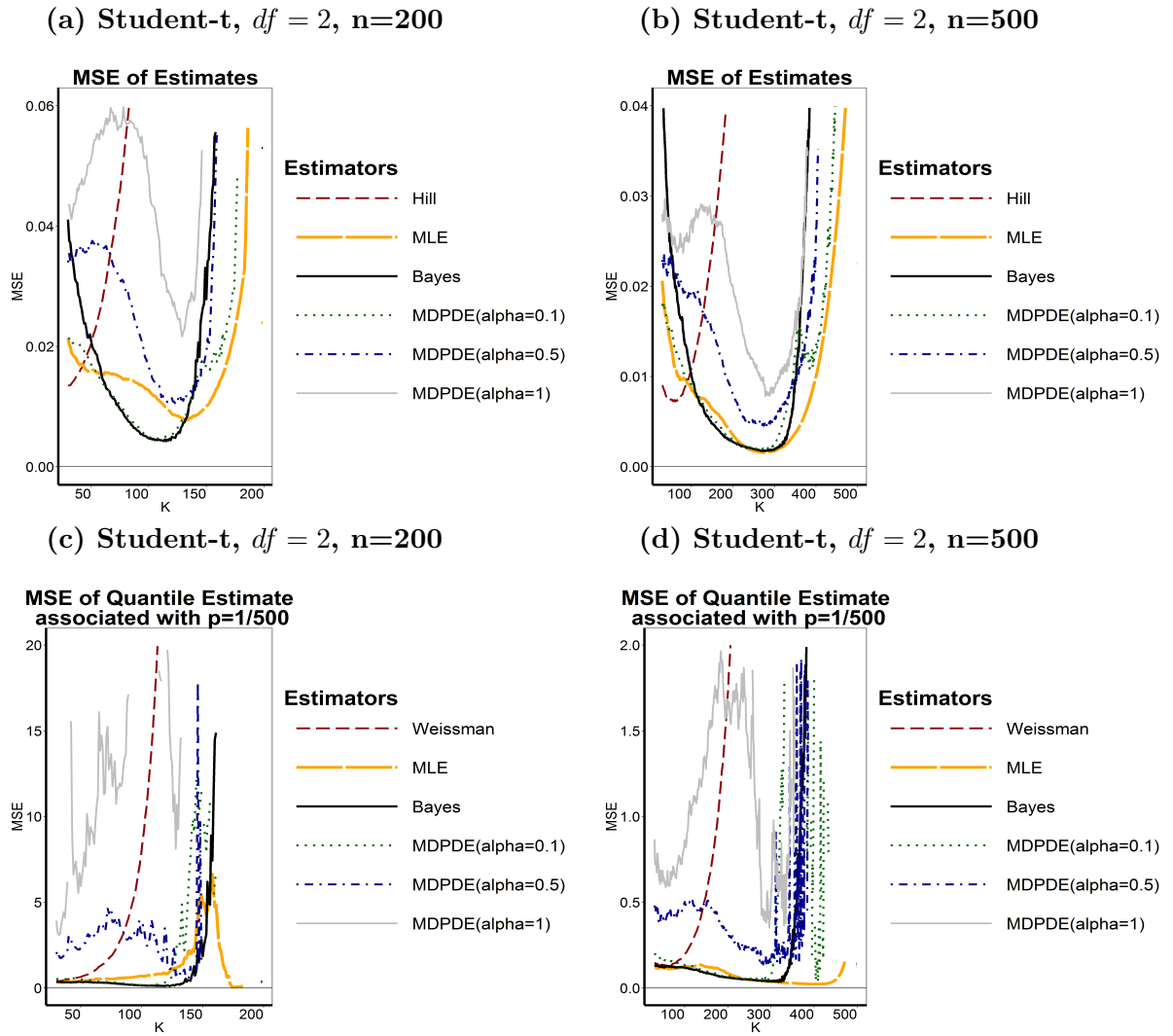


Figure 7.6: MSE of EVI estimates (top) and $U(1-1/500)$ (bottom) of a **Student-t** sample with $EVI = 0.5$.

From Figure 7.6 $\hat{\gamma}^{Bayes}$ and $\hat{U}_*^{Bayes}(1/p_n)$ as well as $\hat{\gamma}_{0.1,n}$ and $\hat{U}_{0.1,n}^{MDPDE}$ are performing considerably better than all considered estimators in terms of the MSE.

Remarks

In the case of uncontaminated samples, the MDPD estimator performs better for α small, in other words by making the MDPDE more efficient. The Bayesian estimator also performs well for uncontaminated samples. We also note that the Bayesian quantile estimator performs better in most cases than other considered quantile estimators.

In the next section we investigate the behavior of the mentioned estimators at contaminated data.

7.3.2 Contaminated Distributions

With the finite sample behavior of all mentioned estimators, when the data is contaminated, we consider the same list of distributions but with different measures of contamination. The contamination distribution \tilde{F} is chosen to be a Pareto distribution, and is given by $\tilde{F}(x) = 1 - \left(\frac{x}{x_c}\right)^{-\beta}$, for $x \geq x_c$, and $\beta = 0.5$. Therefore the true distribution of the data is given by

$$F_\epsilon(x) = (1 - \epsilon)F(x) + \epsilon\tilde{F}_\epsilon \quad (7.8)$$

where F is the Fréchet, Burr (XII) or Student-t distribution and ϵ is the amount of contamination. Two cases of x_c are shown where x_c is 1.2 times the 99.99% quantile of the uncontaminated distribution F and where x_c is 2 times the 99.99% quantile of the uncontaminated distribution F . Three values of contamination are chosen as 0.01, 0.05 and 0.1. Table 7.2 shows the different settings of the simulation study at contaminated data.

Table 7.2: Simulation setting for contaminated data

Distribution F	Quantity	ϵ	Figure
Fréchet ($\alpha = 2$)	$\hat{\gamma}$ and $\hat{U}(1 - 1/500)$	0.01	7.7
		0.05	7.9
		0.1	7.11
	$\hat{\gamma}^{MSE}$ and $\hat{U}^{MSE}(1 - 1/500)$	0.01	7.8
		0.05	7.10
		0.1	7.12
Burr (XII) ($\eta = 1, \tau = 2, \lambda = 1$)	$\hat{\gamma}$ and $\hat{U}(1 - 1/500)$	0.01	7.13
		0.05	7.15
		0.1	7.17
	$\hat{\gamma}^{MSE}$ and $\hat{U}^{MSE}(1 - 1/500)$	0.01	7.14
		0.05	7.16
		0.1	7.18
Student-t ($df = 2$)	$\hat{\gamma}$ and $\hat{U}(1 - 1/500)$	0.01	7.19
		0.05	7.21
		0.1	7.23
	$\hat{\gamma}^{MSE}$ and $\hat{U}^{MSE}(1 - 1/500)$	0.01	7.20
		0.05	7.22
		0.1	7.24

Fréchet Distribution

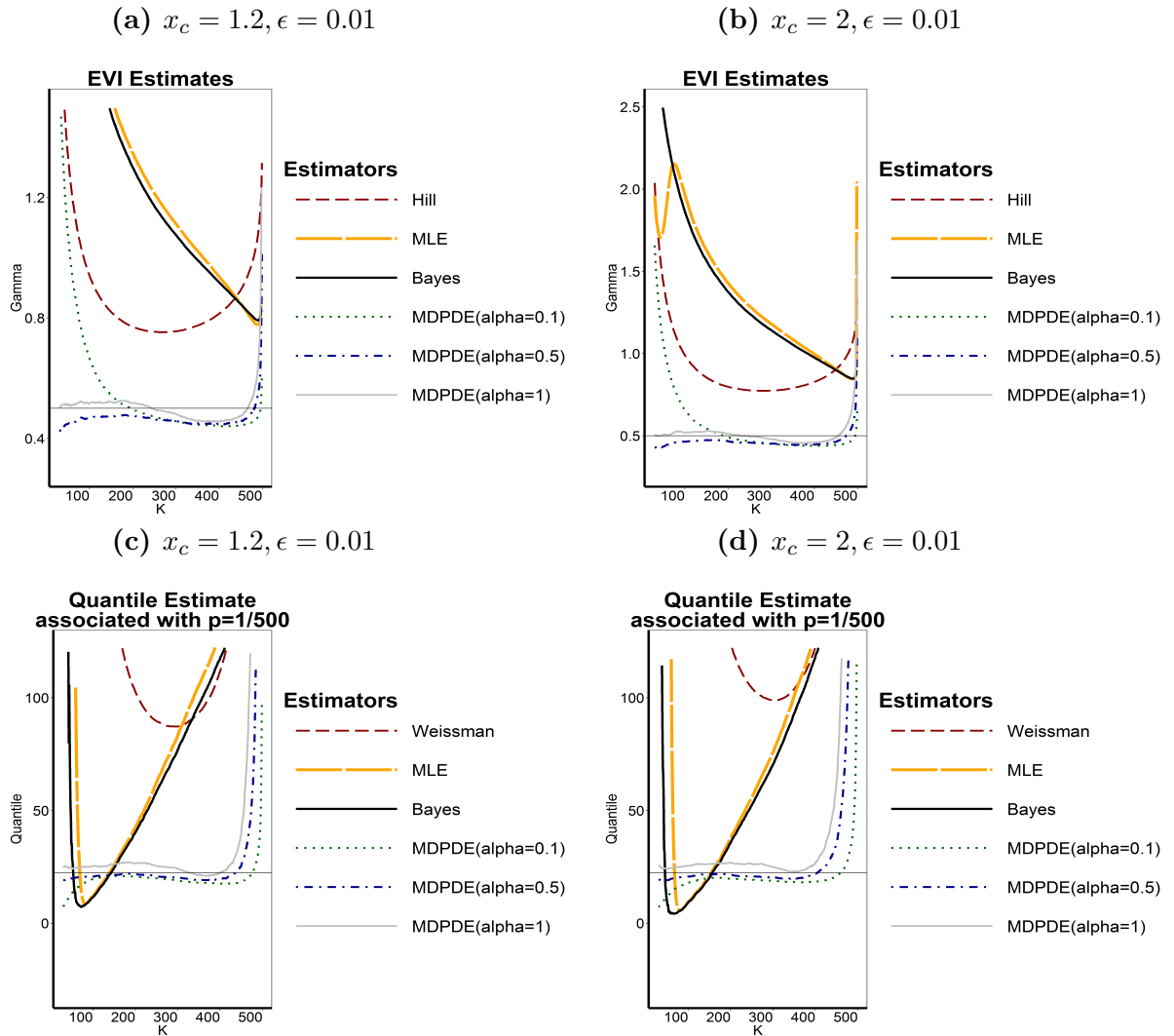


Figure 7.7: Estimates of the **EVI** (top) and **U(1-1/500)** (bottom) of a **Fréchet** sample with $\text{EVI} = 0.5$.

From Figure 7.7 we can see that the MDPD estimators are performing much better than the Bayes, ML and Hill estimators. $\hat{\gamma}_{1,n}$ seems to be performing better than other MDPD estimators. The choice of $x_c = 1.2$ and $x_c = 2$ does not seem to have much influence on the estimators.

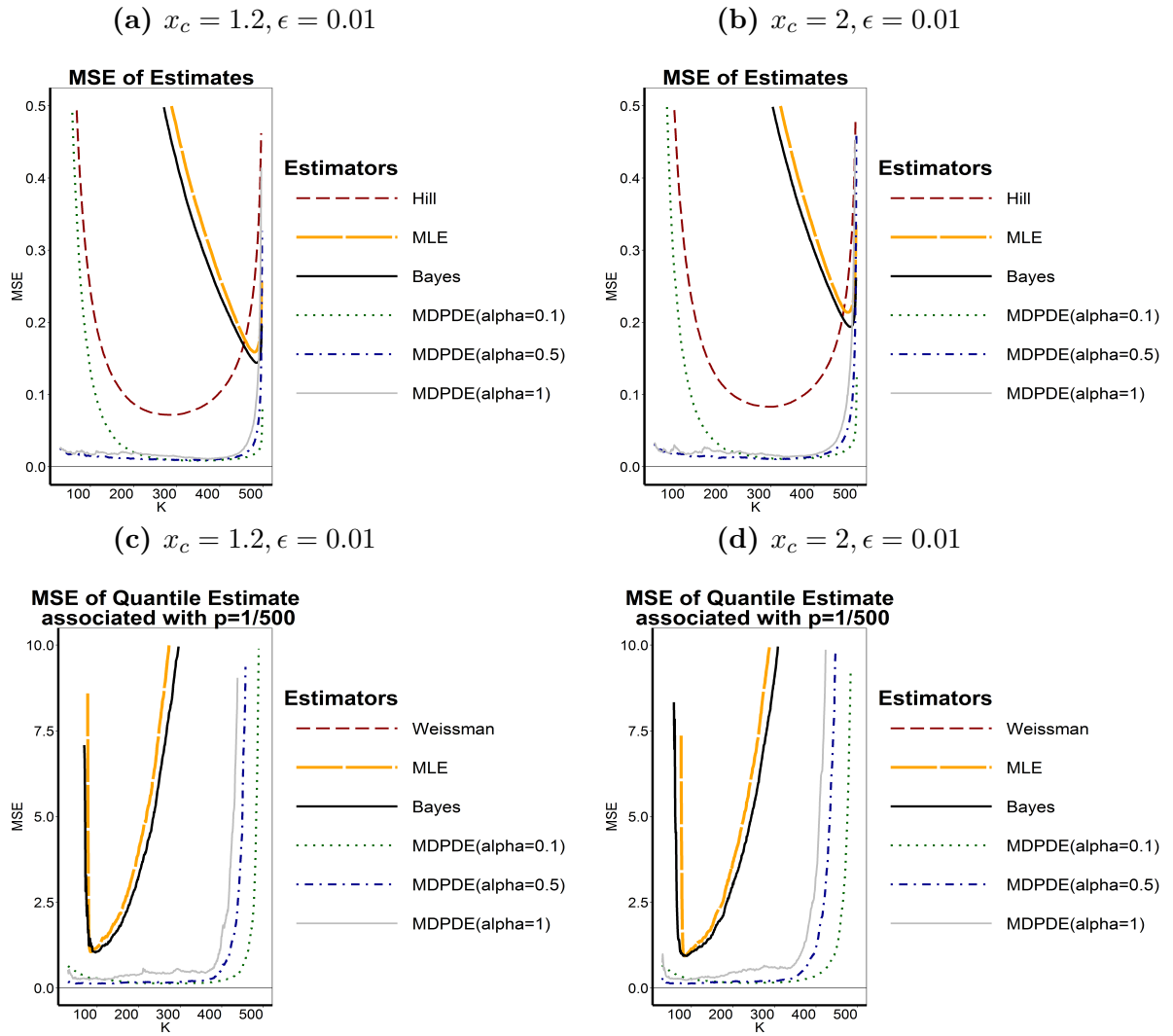


Figure 7.8: MSE of EVI estimates (top) and $U(1-1/500)$ (bottom) of a Fréchet sample with $EVI = 0.5$.

From Figure 7.8 $\hat{\gamma}_{0.5,n}$ and $\hat{U}_{0.5,n}^{MDPDE}(1/p_n)$ seem to be outperforming the non-robust estimators $\hat{\gamma}_n^{Hill}$, $\hat{\gamma}_n^{Bayes}$ and $\hat{\gamma}_n^{MLE}$ as well as their corresponding quantile estimators in terms of MSE.

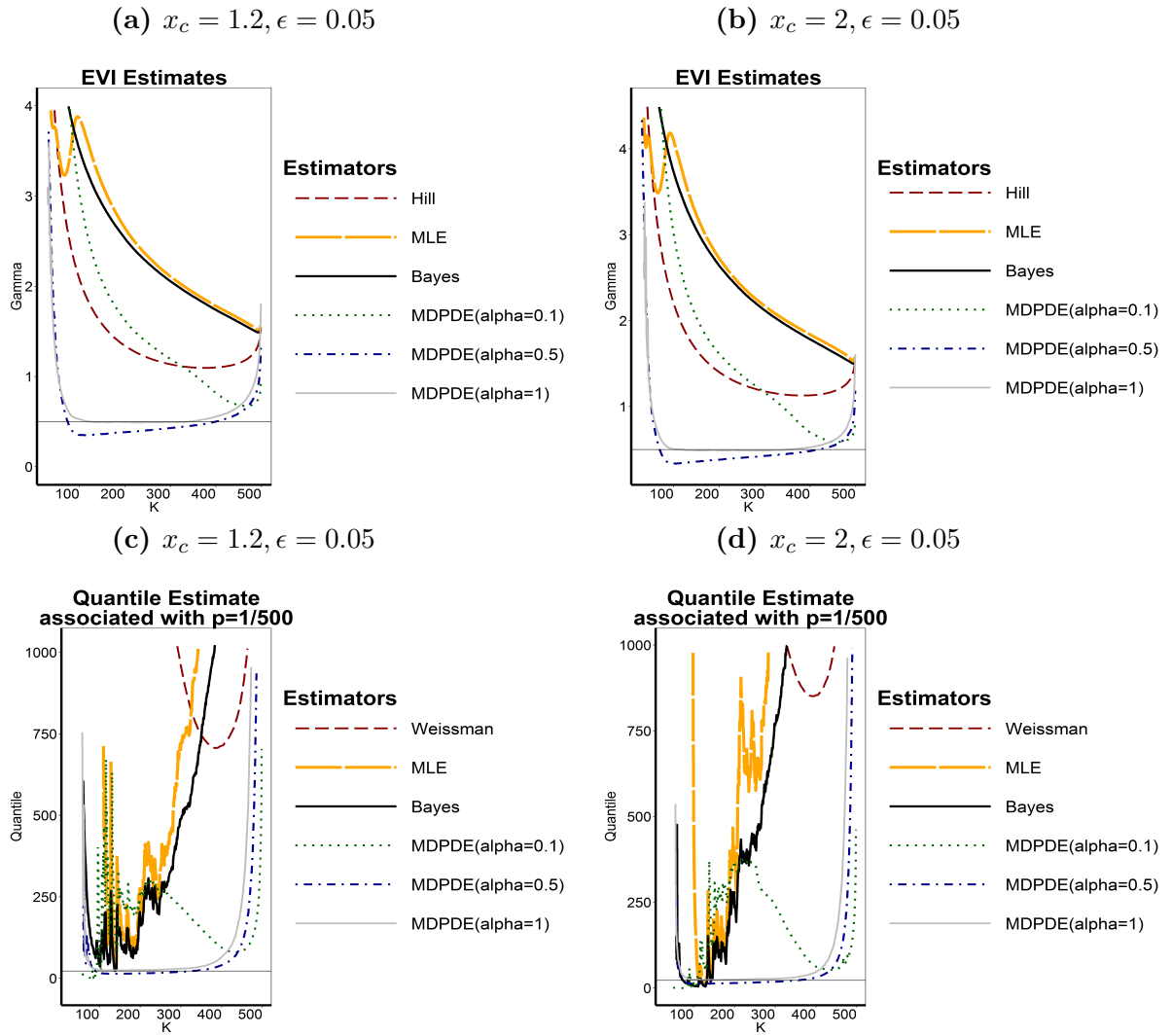


Figure 7.9: Estimates of the **EVI** (top) and **U(1-1/500)** (bottom) of a **Fréchet** sample with **EVI = 0.5**.

From Figure 7.9 $\hat{\gamma}_{1,n}$ and $\hat{U}_{1,n}^{MDPDE}(1/p_n)$ seem to be outperforming all other considered robust and non-robust MDPD estimators. This is due to the fact that we have more contamination in the right tail of the distribution F and an MDPD estimator with $\alpha = 1$ is more robust against such contamination than an MDPE with $\alpha < 1$.

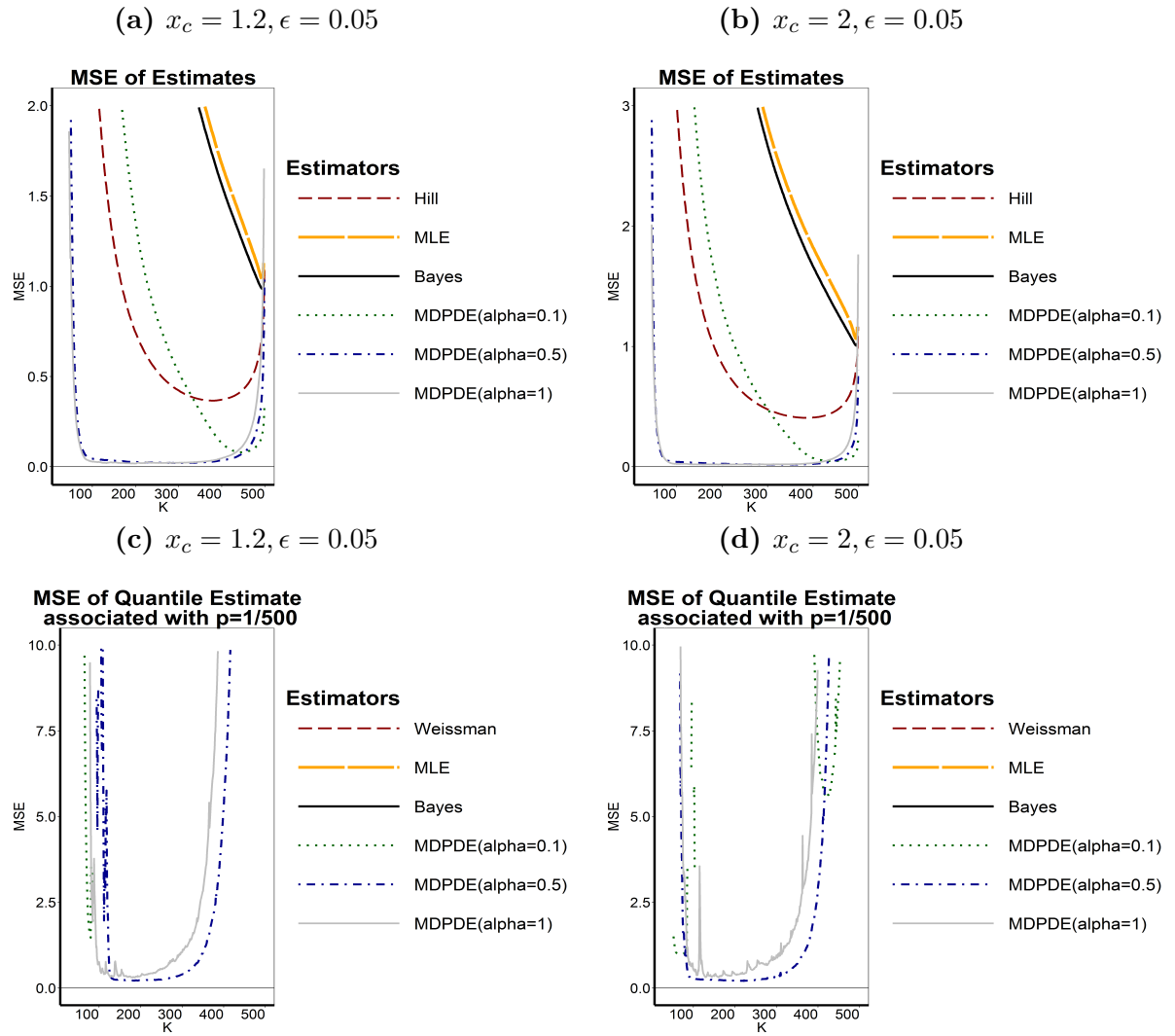


Figure 7.10: MSE of EVI estimates (top) and $U(1-1/500)$ (bottom) of a Fréchet sample with $EVI = 0.5$.

From Figure 7.10 $\hat{\gamma}_{0.5,n}$ and $\hat{U}_{0.5,n}^{MDPDE}(1/p_n)$ tends to compete with $\hat{\gamma}_{1,n}$ and $\hat{U}_{1,n}^{MDPDE}(1/p_n)$ in terms of MSE.

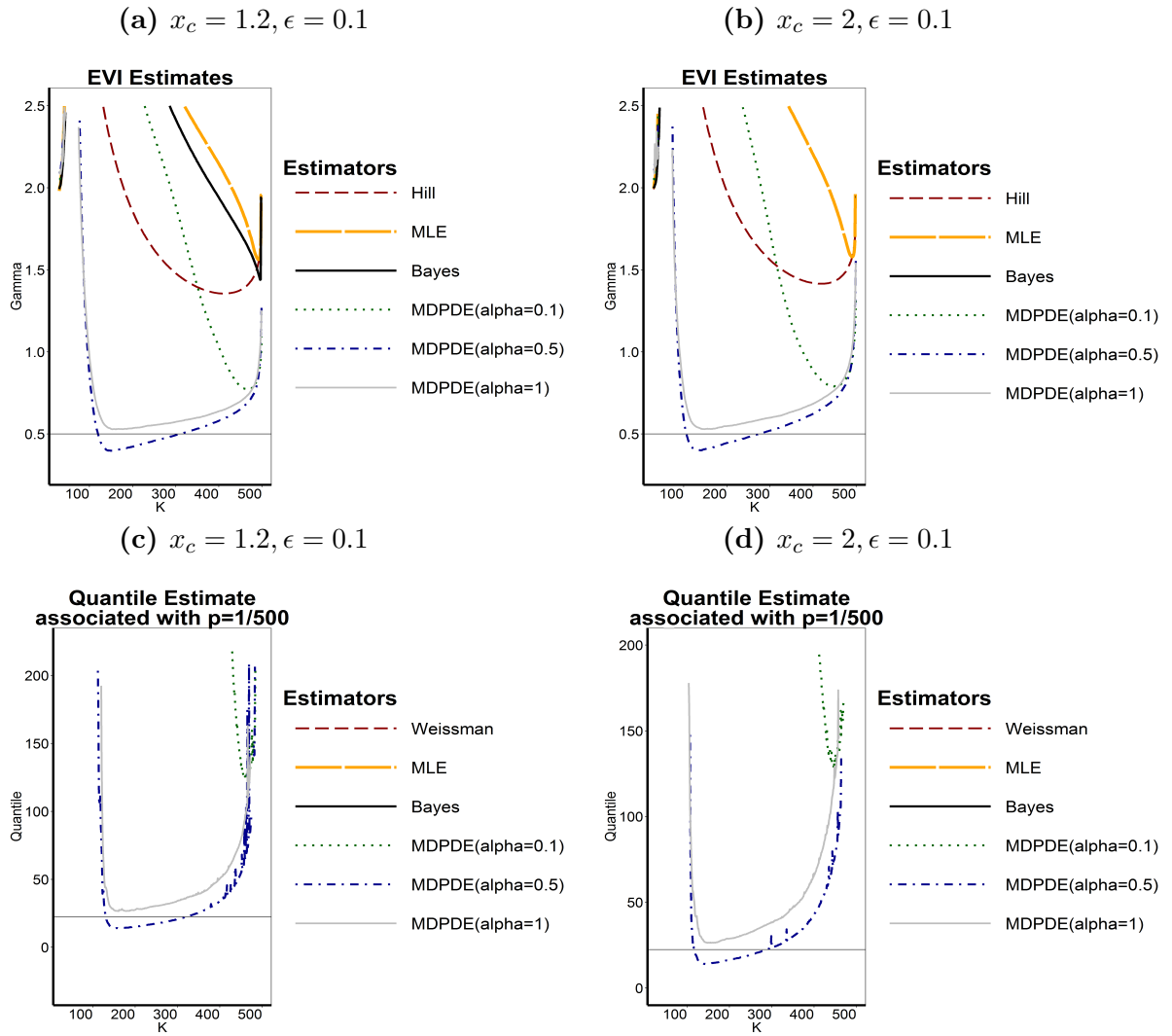


Figure 7.11: Estimates of the **EVI** (top) and **$U(1-1/500)$** (bottom) of a **Fréchet** sample with **$EVI = 0.5$** .

From Figure 7.11 $\hat{\gamma}_{0.5,n}$ and $\hat{U}_{0.5,n}^{MDPDE}(1/p_n)$ seem to be performing better than all other considered robust and non-robust estimators. Keeping in mind that there is now more contamination in the right tail of F ($\epsilon = 0.1$). It is also clear that the non-robust estimators considered under perform more when the amount of contamination increases.

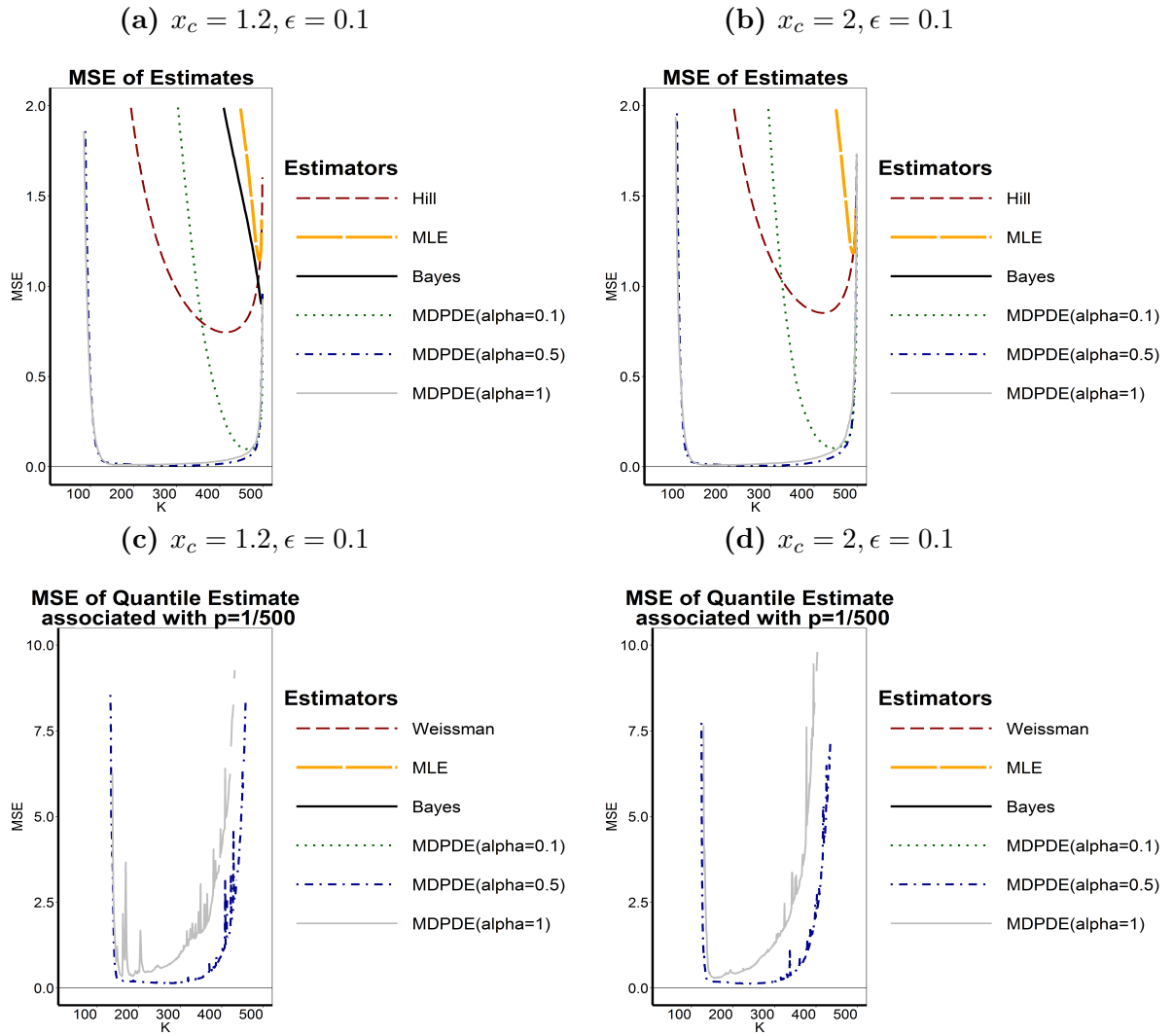


Figure 7.12: MSE of EVI estimates (top) and $U(1-1/500)$ (bottom) of a Fréchet sample with $EVI = 0.5$.

Figure 7.12 confirms the good performance of $\hat{\gamma}_{0.5,n}$ and $\hat{U}_{0.5,n}^{MDPDE}(1/p_n)$ in terms of both the bias and MSE in comparison to other considered estimators. It is also worth noting that $\hat{\gamma}_n^{Hill}$ and $\hat{U}_*^{Wiseman}(1/p_n)$ are performing much better than all considered non-robust estimators $\hat{\gamma}_n^{MLE}$, $\hat{\gamma}_n^{Bayes}$, $\hat{U}_*^{Bayes}(1/p_n)$ and $\hat{U}_*^{MLE}(1/p_n)$.

Burr Distribution

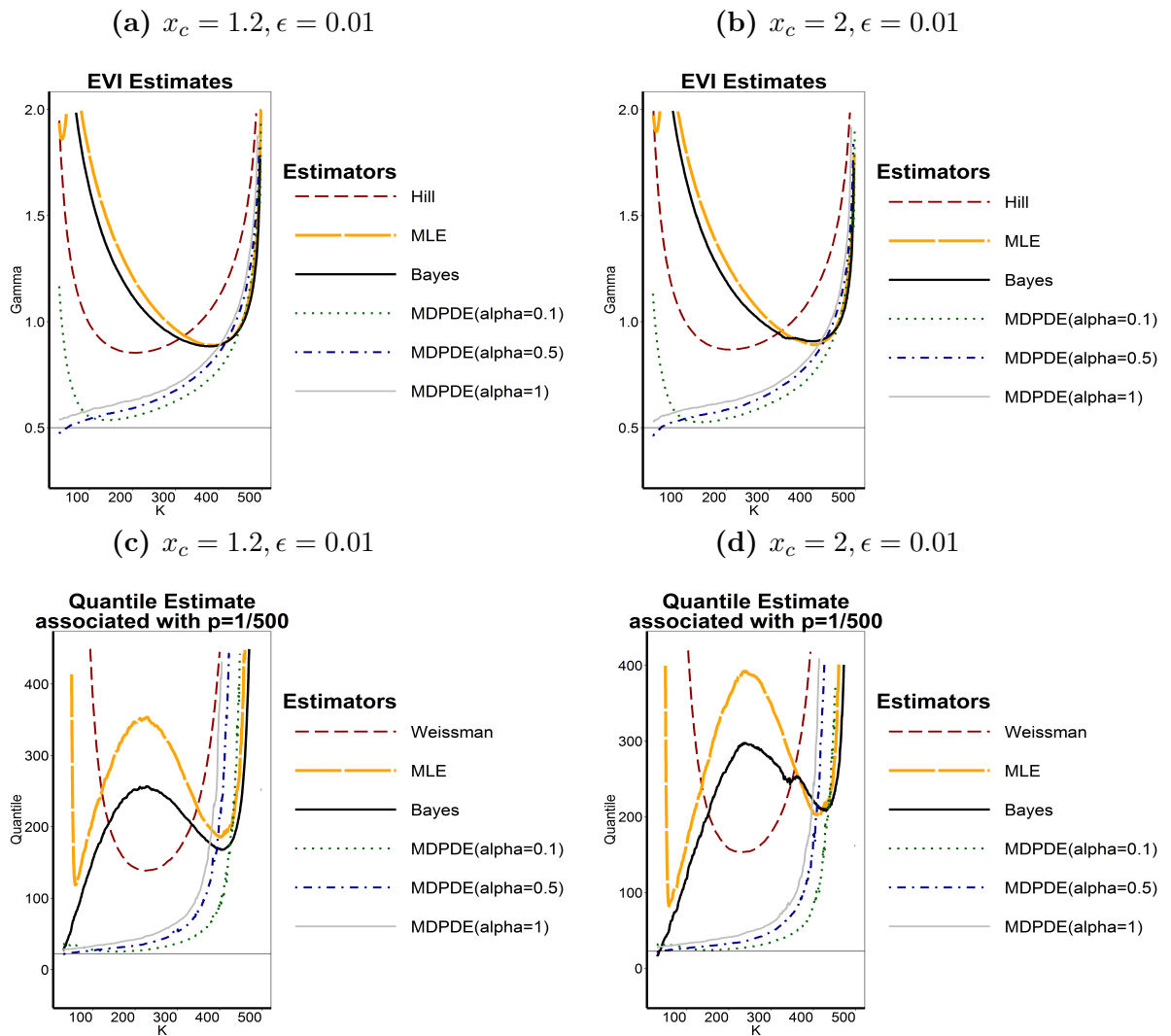


Figure 7.13: Estimates of the **EVI** (top) and **U(1-1/500)** (bottom) of a **Burr (XII)** sample with **EVI = 0.5**.

From Figure 7.13 we can see that for the Burr distribution $\hat{\gamma}_{0.1,n}$ and $\hat{U}_{0.1,n}^{MDPDE}(1/p_n)$ are performing better than all other considered robust and non-robust estimators. The $x_c = 1.2$ and $x_c = 2$ still does not have much influence on the estimators.

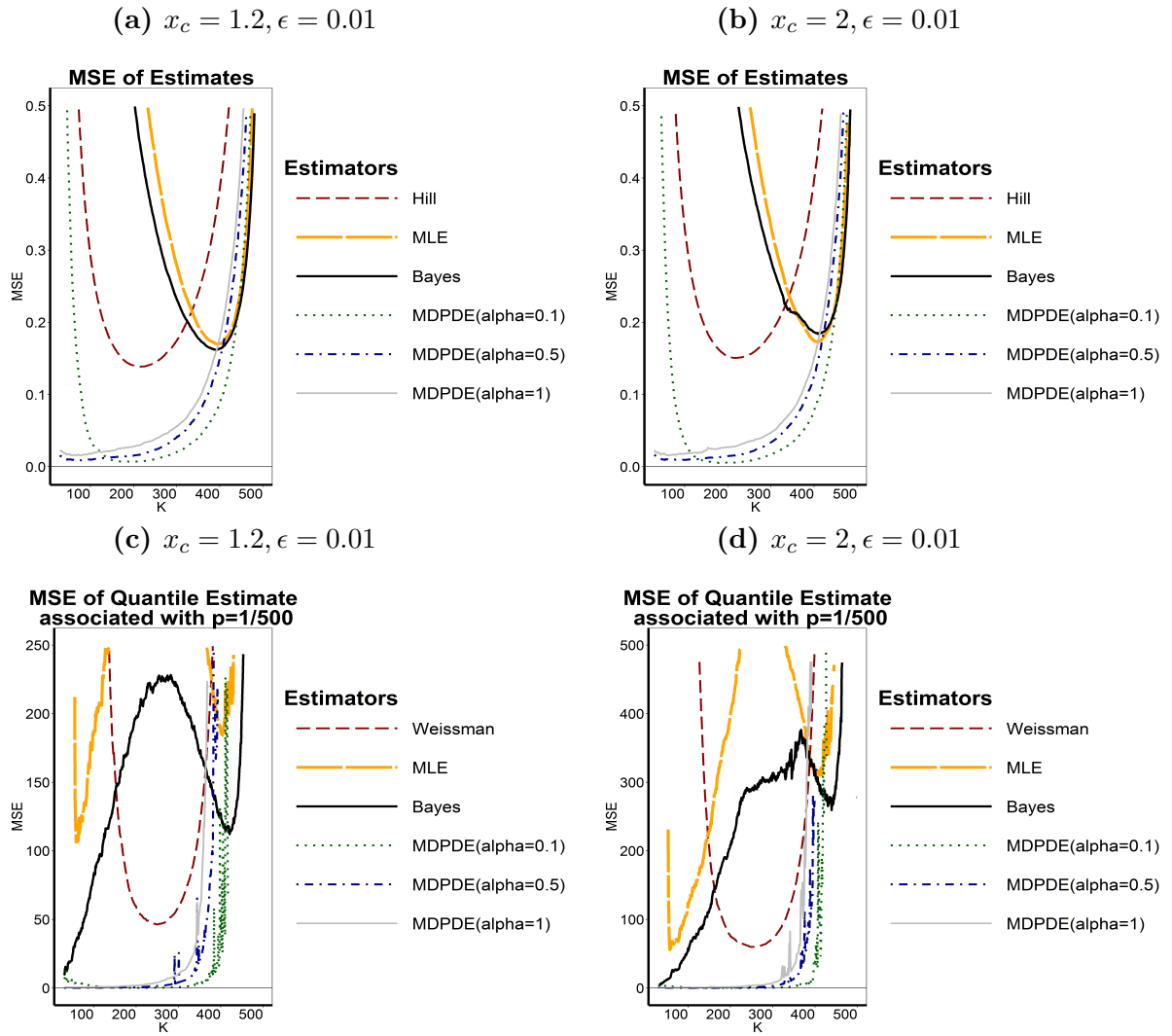


Figure 7.14: MSE of EVI estimates (top) and $U(1-1/500)$ (bottom) of a Burr (XII) sample with $EVI = 0.5$.

Figure 7.14 confirms the good performance of $\hat{\gamma}_{0.1,n}$ and $\hat{U}_{0.1,n}^{MDPDE}(1/p_n)$ in comparison to all other considered robust and non-robust estimators.

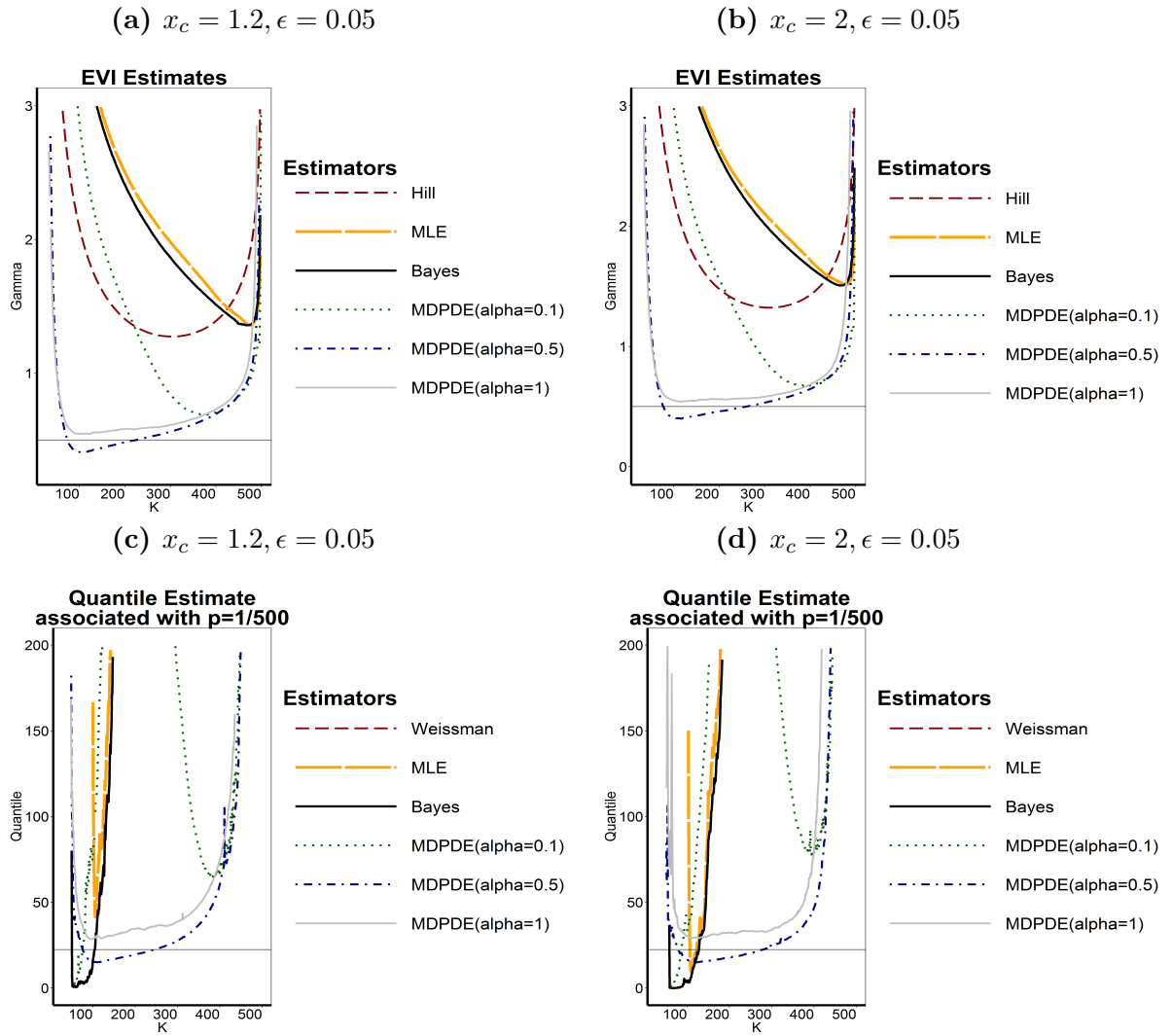


Figure 7.15: Estimates of the **EVI** (top) and **U(1-1/500)** (bottom) of a **Burr (XII)** sample with $\text{EVI} = 0.5$.

From Figure 7.15 $\hat{\gamma}_{0.5,n}$ and $\hat{U}_{0.5,n}^{MDPDE}(1/p_n)$ are once again performing better than all other considered robust and non-robust estimators. This is due to the increased amount of contamination in the right tail of F ($\epsilon = 0.05$).

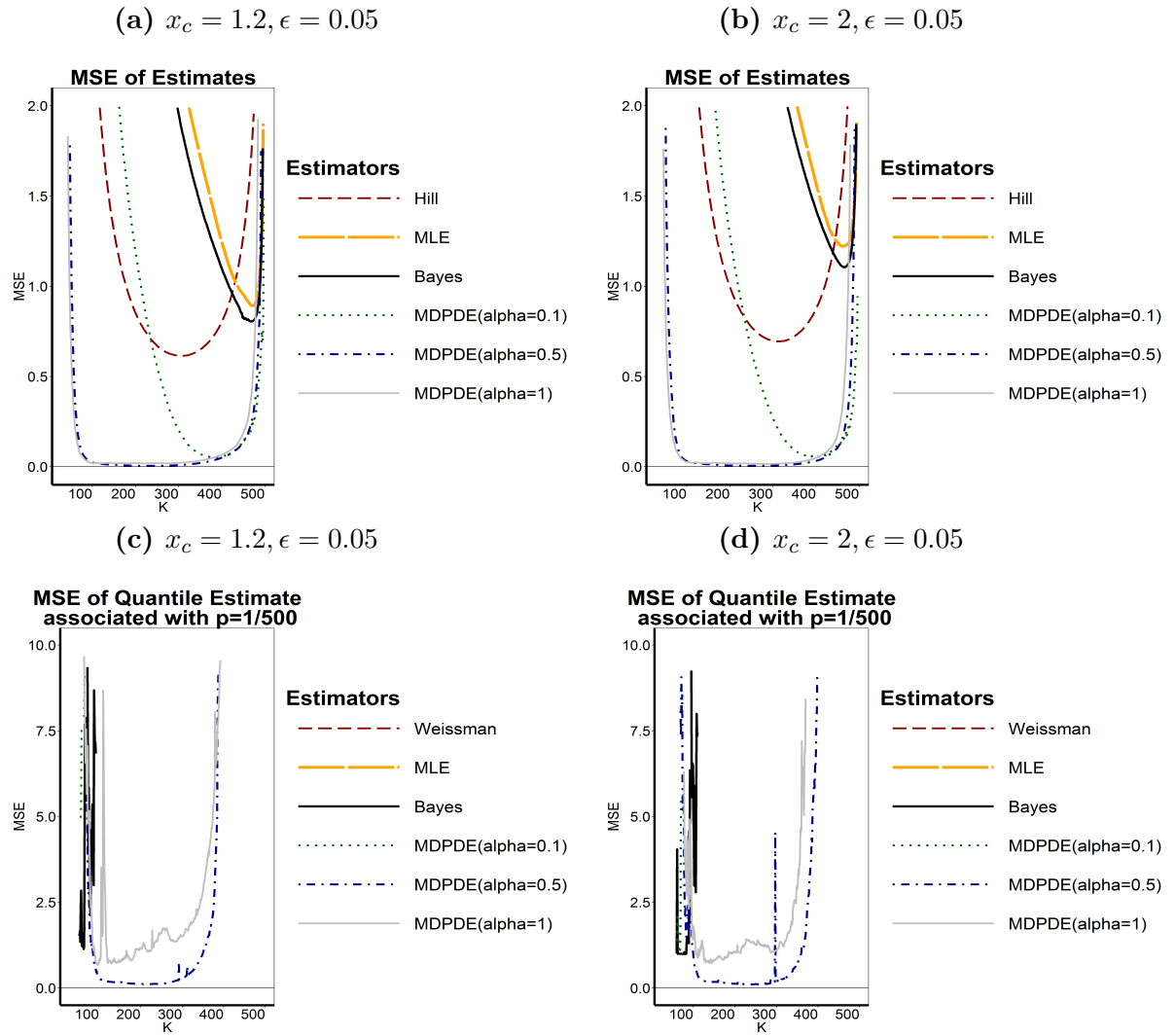


Figure 7.16: MSE of EVI estimates (top) and $U(1-1/500)$ (bottom) of a Burr (XII) sample with $EVI = 0.5$.

Figure 7.16 confirms the good performance of $\hat{\gamma}_{0.5,n}$ especially $\hat{U}_{0.5,n}^{MDPDE}(1/p_n)$. At this point all considered non-robust estimators are performing poorly.

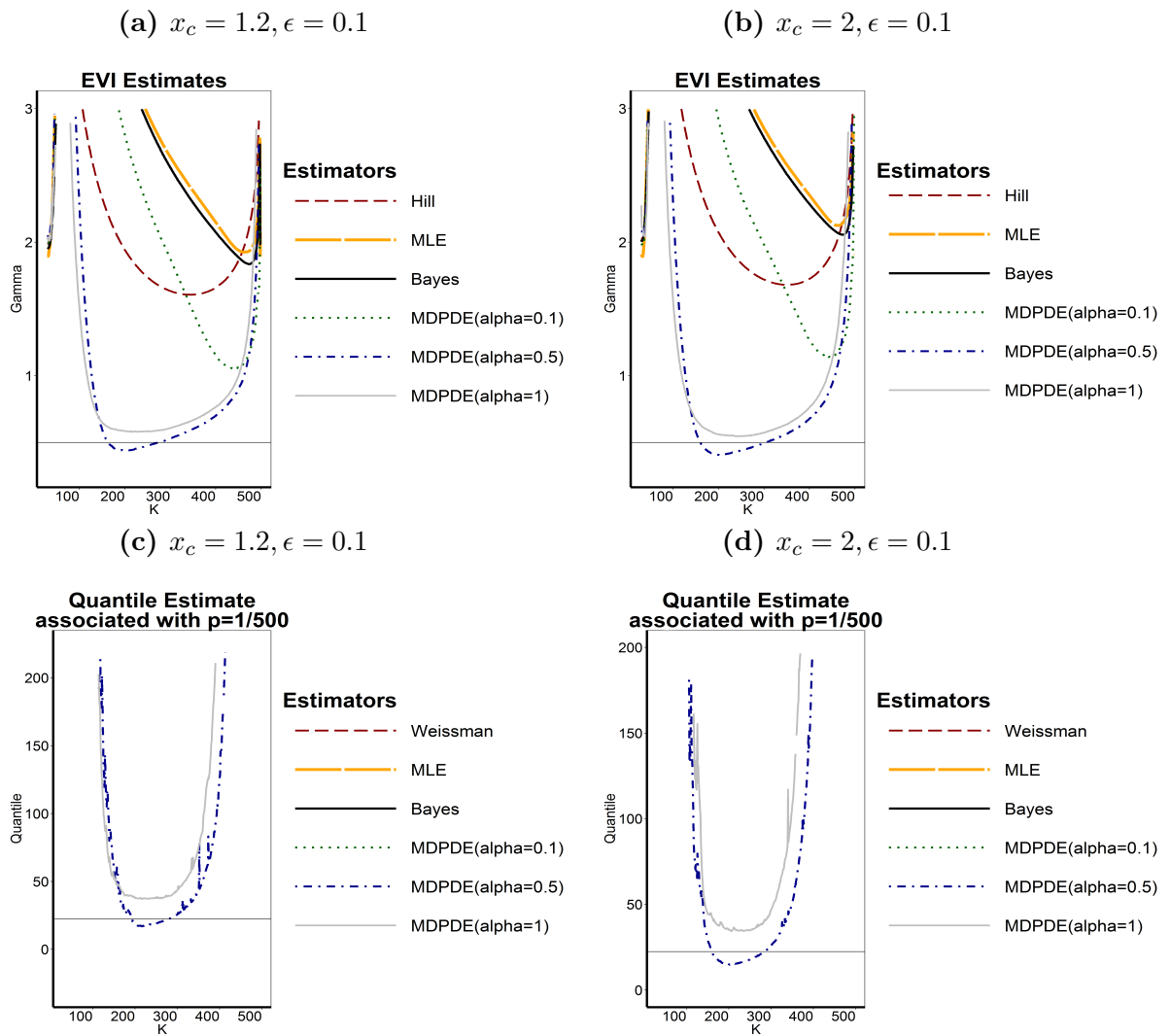


Figure 7.17: Estimates of the **EVI** (top) and **U(1-1/500)** (bottom) of a **Burr (XII)** sample with **EVI = 0.5**.

By further increasing the amount of contamination to $\epsilon = 0.1$ of the Burr samples, it can be seen from Figure 7.17 that $\hat{\gamma}_{0.5,n}$ and $\hat{U}_{0.5,n}^{MDPDE}(1/p_n)$ continue to outperform all considered robust and non-robust estimators.

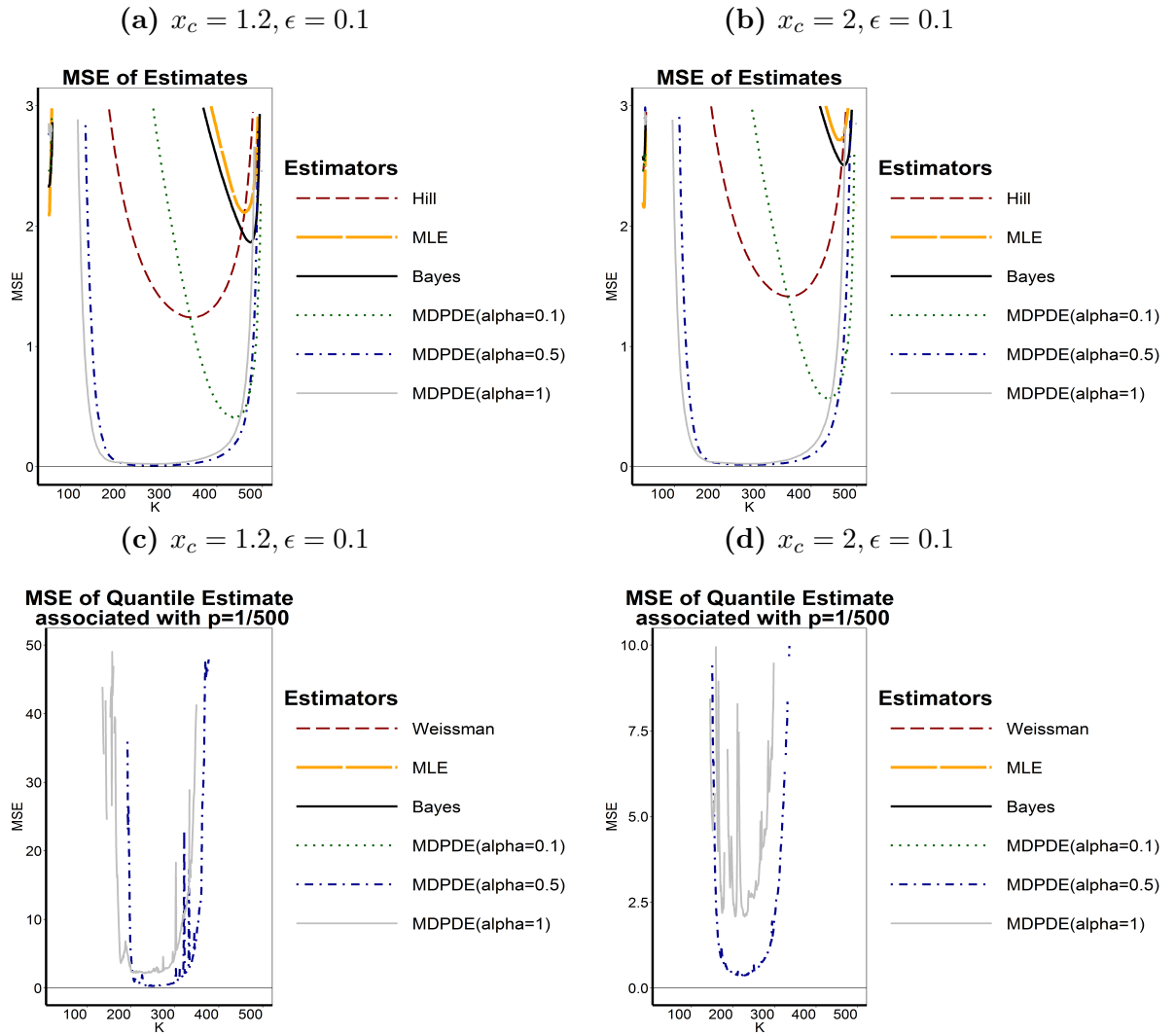


Figure 7.18: MSE of EVI estimates (top) and $U(1-1/500)$ (bottom) of a Burr (XII) sample with $EVI = 0.5$.

Figure 7.18 confirms the good performance of $\hat{\gamma}_{0.5,n}$ and $\hat{U}_{0.5,n}^{MDPDE}(1/p_n)$ in terms of the MSE and bias.

Student-t Distribution

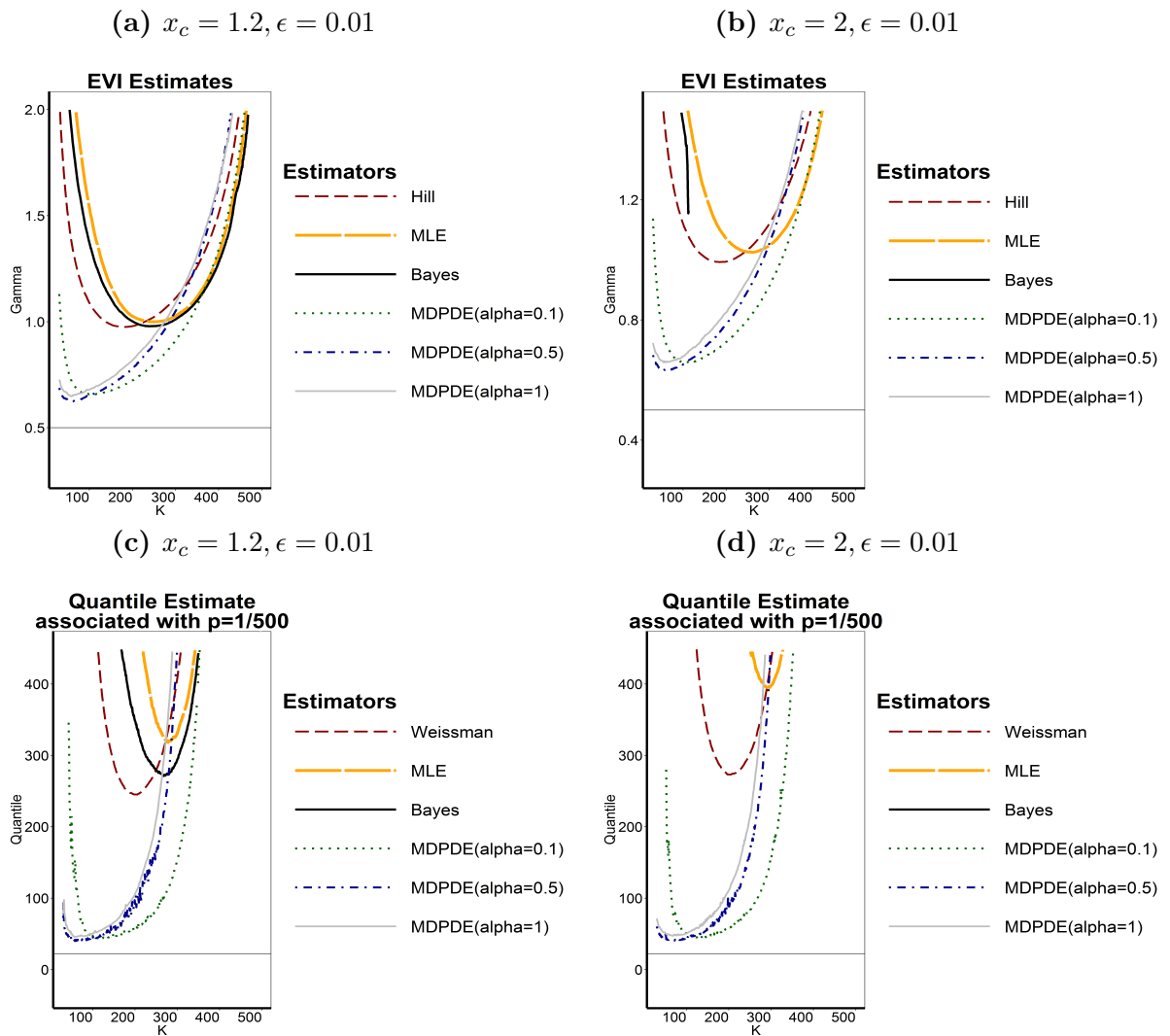


Figure 7.19: Estimates of the **EVI** (top) and **U(1-1/500)** (bottom) of a **Student-t** sample with **EVI = 0.5**.

From Figure 7.19 $\hat{\gamma}_{0.1,n}$ and $\hat{U}_{0.1,n}^{MDPDE}(1/p_n)$ appears to be competitive. This is due to the fact that there is a small amount of contamination in the right tail of F .

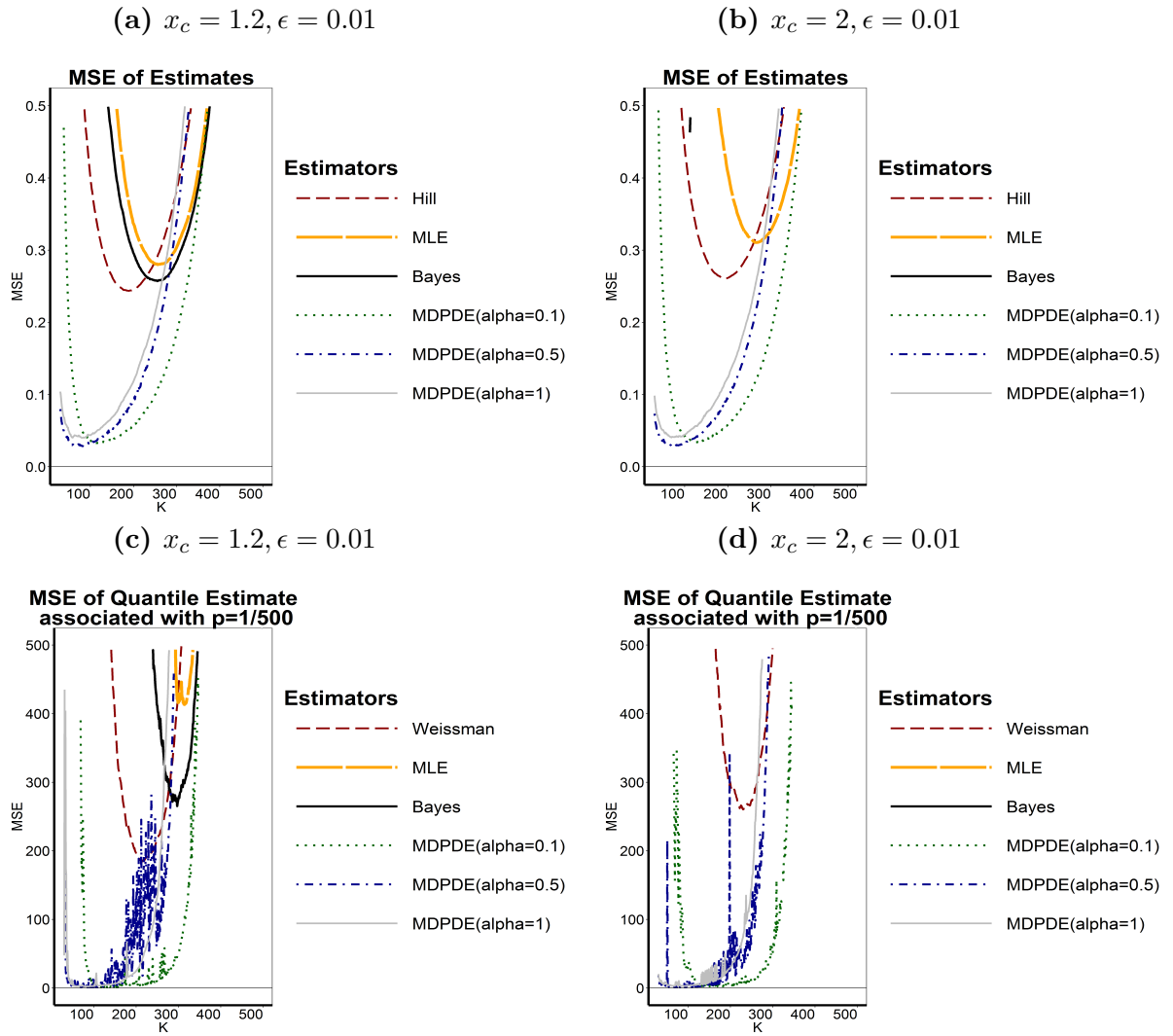


Figure 7.20: MSE of EVI estimates (top) and $U(1-1/500)$ (bottom) of a **Student-t** sample with $EVI = 0.5$.

Figure 7.20 confirms the good performance of $\hat{\gamma}_{0.1,n}$ and $\hat{U}_{0.1,n}^{MDPDE}(1/p_n)$ in terms of the MSE. We also note the explosive behavior of $\hat{U}_{0.1,n}^{MDPDE}(1/p_n)$.

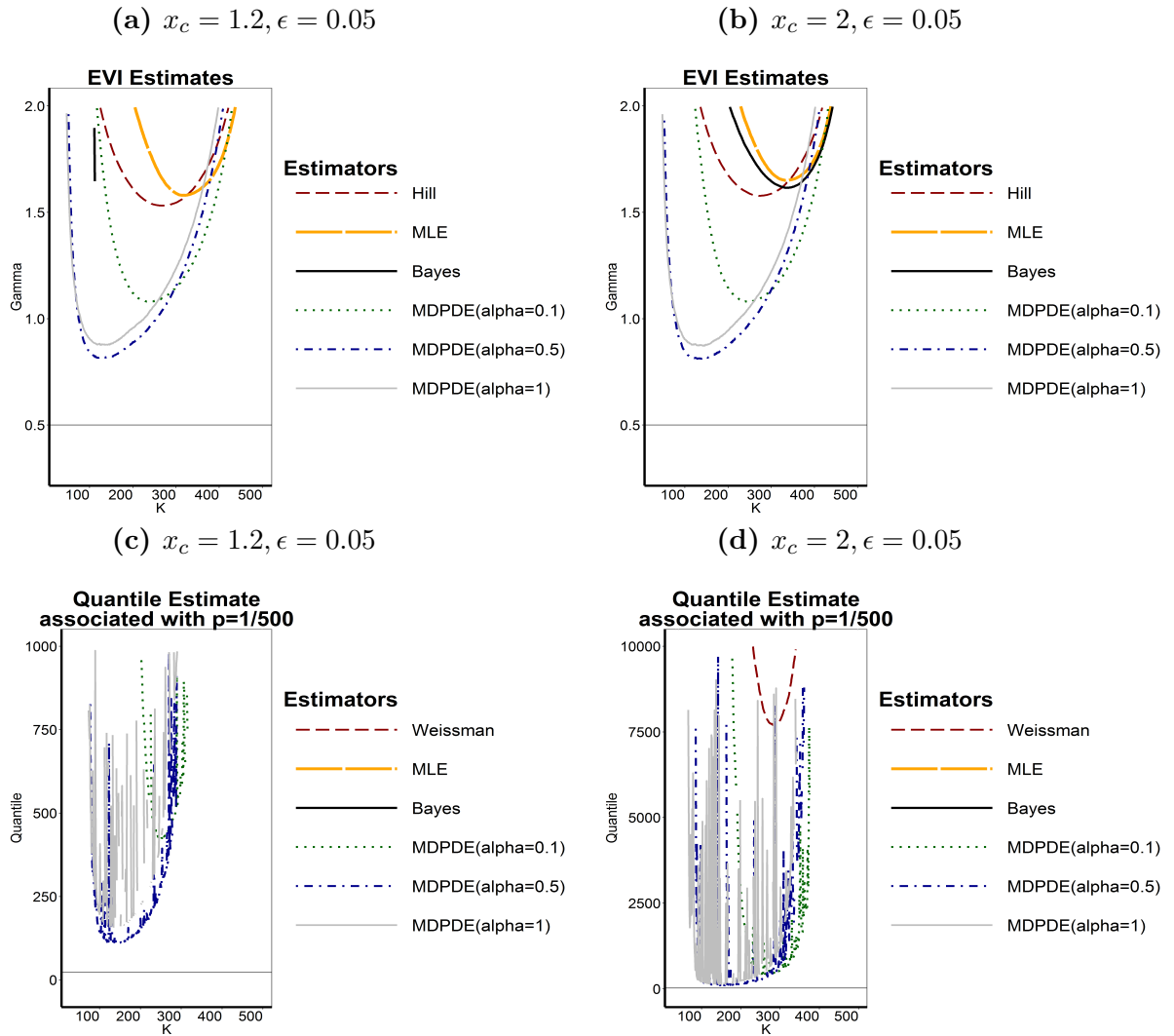


Figure 7.21: Estimates of the **EVI** (top) and **U(1-1/500)** (bottom) of a **Student-t** sample with **EVI = 0.5**.

From Figure 7.21 we can see that all estimators perform poorly in terms of bias at this amount of contamination ($\epsilon = 0.05$), however $\hat{\gamma}_{0.5,n}$ and $\hat{U}_{0.5,n}^{MDPDE}(1/p_n)$ maintain a reasonably good performance also noting the volatile behavior of $\hat{U}_{1,n}^{MDPDE}(1/p_n)$.

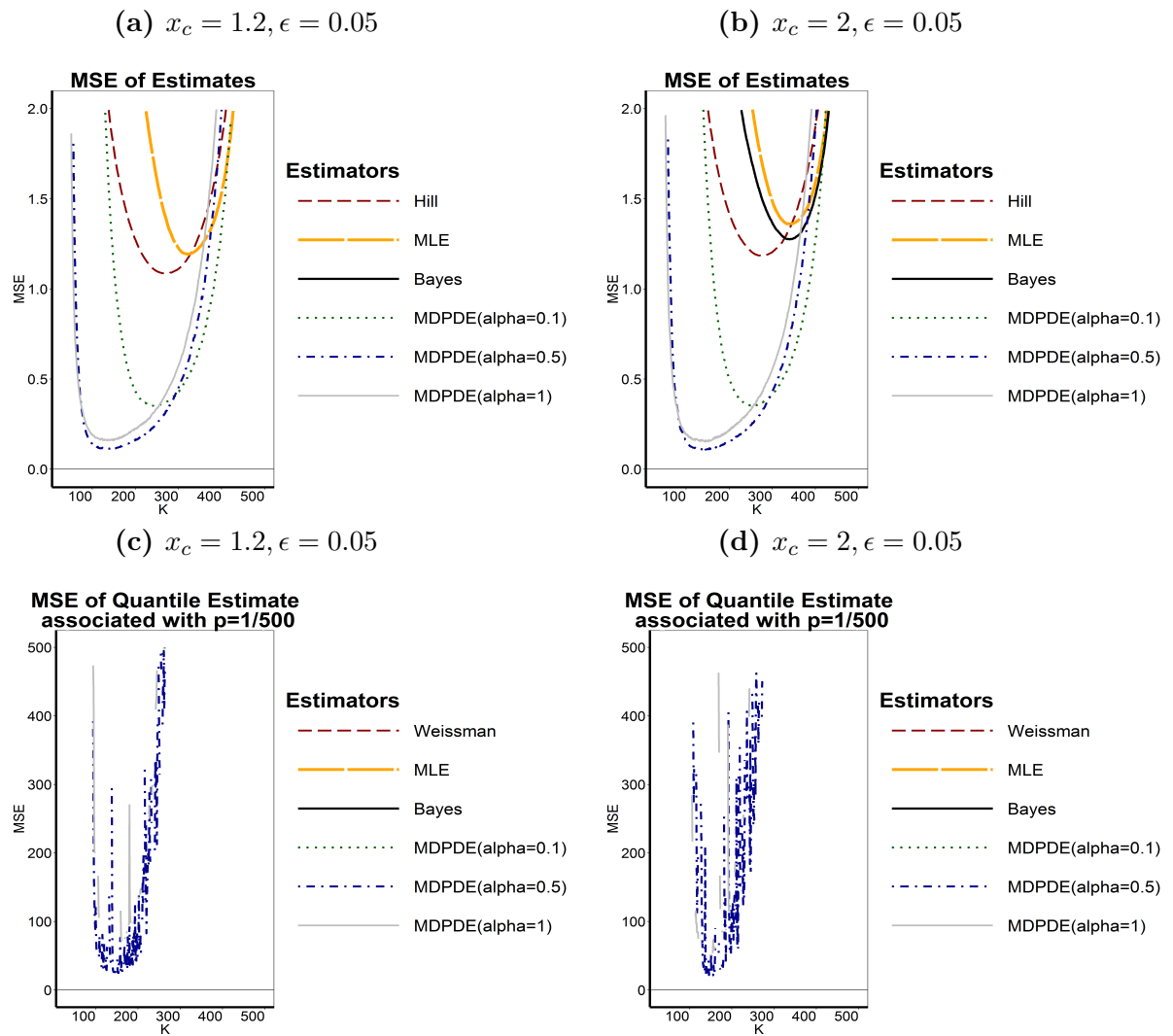


Figure 7.22: MSE of EVI estimates (top) and $U(1-1/500)$ (bottom) of a **Student-t** sample with $EVI = 0.5$.

Figure 7.22 confirms the reasonable out-performance of $\hat{\gamma}_{0.5,n}$ and $\hat{U}_{0.5,n}^{MDPDE}(1/p_n)$ in comparison to all considered robust and non-robust estimators.

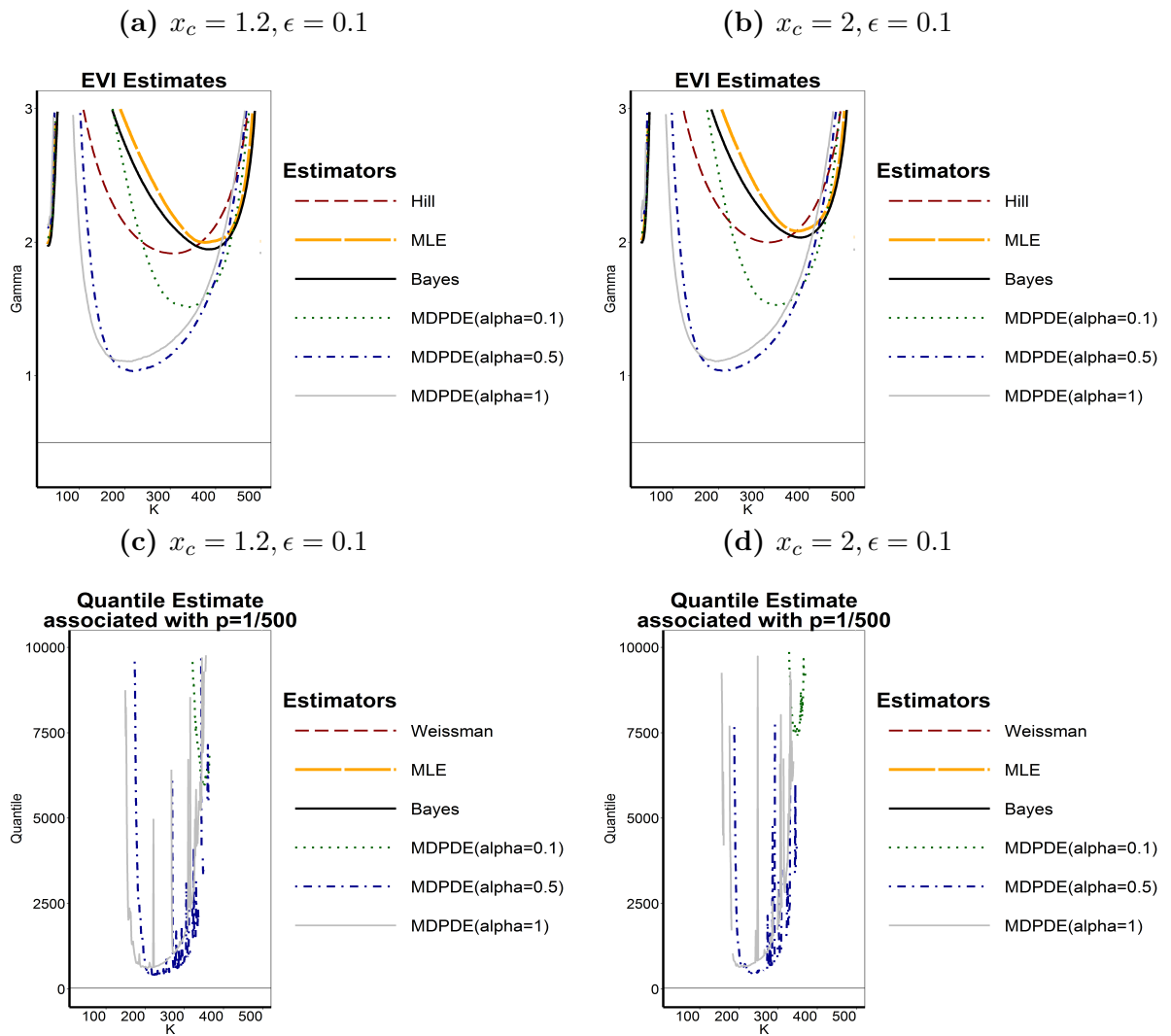


Figure 7.23: Estimates of the EVI (top) and $U(1-1/500)$ (bottom) of a Student- t sample with $EVI = 0.5$.

From Figure 7.23 we can see that, despite the reasonable performance of $\hat{\gamma}_{0.5,n}$ and $\hat{U}_{0.5,n}^{MDPDE}(1/p_n)$, all considered estimators are very volatile at this amount of contamination ($\epsilon = 0.1$) in the right tail of F .

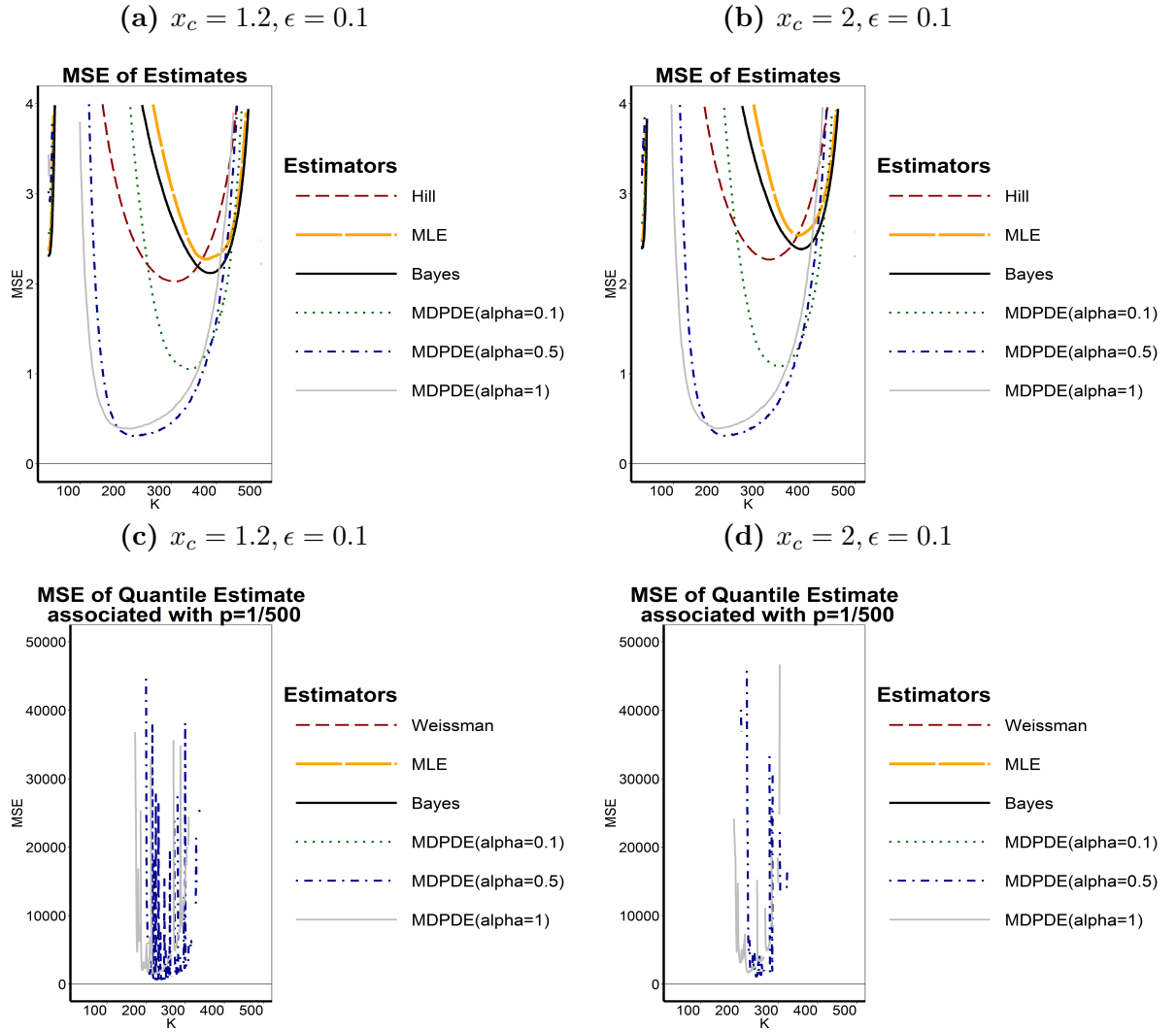


Figure 7.24: MSE of EVI estimates (top) and $U(1-1/500)$ (bottom) of a **Student-t** sample with $EVI = 0.5$.

Lastly in Figure 7.24 we can see that $\hat{\gamma}_{0.5,n}$ and $\hat{U}_{0.5,n}^{MDPDE}(1/p_n)$ maintain a reasonable performance. At this stage all considered robust and non-robust estimators are performing poorly in terms of both bias and MSE.

Remarks

In case of contaminated samples, the estimators $\hat{\gamma}_{0.5,n}$ and $\hat{U}_{0.5,n}^{MDPDE}(1/p_n)$ clearly outperform all the considered non-robust estimators both in terms of bias and MSE. $\hat{\gamma}_{0.5,n}$ and $\hat{U}_{0.5,n}^{MDPDE}(1/p_n)$ show much more stable sample paths.

In conclusion, $\hat{\gamma}_n^{Bayes}$ is in general a good alternative to estimating the EVI at uncontaminated data as well as $\hat{\gamma}_n^{MLE}$ and $\hat{\gamma}_{\alpha,n}$ where α is close to zero as well as

the corresponding quantile estimators. When contamination is present $\hat{\gamma}_{0.5,n}$ and $\hat{U}_{0.5,n}^{MDPDE}(1/p_n)$ seem to be less biased compared to all other MDPD estimators and all considered non-robust estimators. We also noted that $\hat{\gamma}_n^{Hill}$ and $\hat{U}_*^{Wiseman}(1/p_n)$ performs better when contamination is present than all other considered non-robust estimators $\hat{\gamma}_n^{Bayes}$, $\hat{U}_*^{Bayes}(1/p_n)$, $\hat{\gamma}_n^{MLE}$ and $\hat{U}_*^{MLE}(1/p_n)$.

In the next section we further assess the performance of all previously considered estimators with an application to a real dataset.

7.4 Case Study

In this section all previously discussed estimators are applied to a real dataset and assessed in a real world setting. We use soil data studied by Goegebeur et al. (2005), Beirlant et al. (2004) and further studied by Vandewalle et al. (2007). This data is from the Condruz region in Belgium, gathered by the Unit of Geopedology at Gembloux Agricultural University (see Goegebeur et al., 2005, for more information). In agriculture, soil analysis is the basis of fertilizer and amendment recommendations in the context of managing soil fertility and crop performance.

The development of precision farming has drastically increased the demand for soil data and laboratories are burdened with large data sets, which inevitably causes concern about outliers and their apparent influence on the quality of the information. Robust estimation procedures therefore play a crucial role in the database management of soil data in order to provide high quality information.

The Condruz database contains calcium content and pH level measurements of 19,516 soil samples originating from different cities in the Condruz, a geographical region in the southern part of Belgium. Beirlant et al. (2004) saw the need to describe the tail of the calcium distribution in terms of the covariate pH. Figure 7.25 shows a scatter plot of calcium (Ca) versus pH. From this plot it is clear that these two variables are positively associated and extreme calcium measurements tend to occur more often at the higher pH levels.

In this section we focus on estimation of the EVI and high Quantiles conditional on the $\text{pH} = \{6, 6.5, 7.1, 7.2, 7.3 \text{ and } 7.4\}$ covariate. Figure 7.26 shows the Box plots of Ca measurements at different pH levels. It can be seen from these plots that the data is sometimes more than 1.5 times the interquartile range away from the box which could imply existence of suspicious observations. The box plots for $\text{pH}=6.5$ and 7.1 seem to only have one suspicious observation while box plot $\text{pH}=7.2$ and 7.3 have more observations that can be flagged as suspicious.

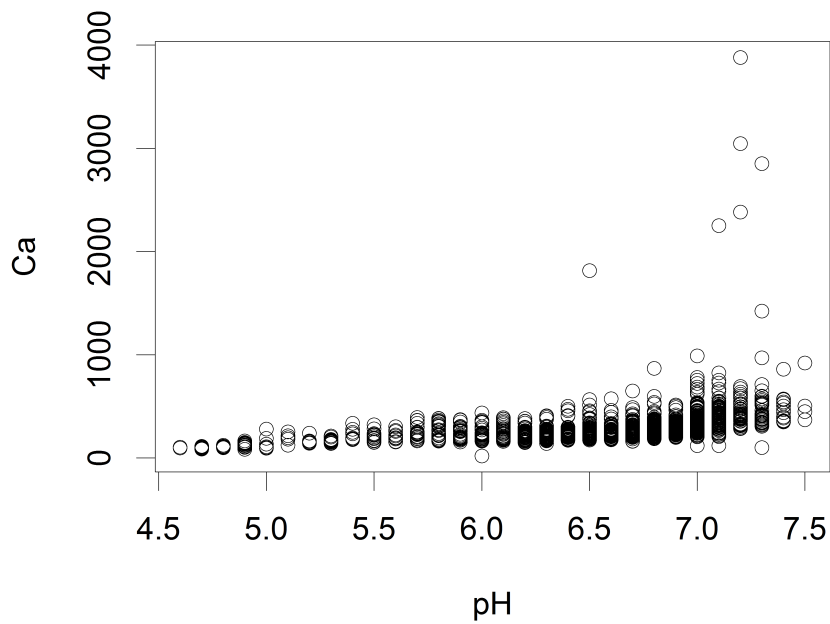


Figure 7.25: Plot of Ca versus pH

We further investigate large observations that do not follow the ultimate linearity of the Pareto quantile plot in Figure 7.27, and regard them as outliers.

From Figure 7.26 and Figure 7.27 we can see that there exists some outliers that could be considered as suspicious. For instance in Figure 7.27a we can see from its ultimate linearity of the Pareto QQ that the data is Pareto, however there is an observation that is much lower than the others. Also with Figure 7.27b, Figure 7.27c, Figure 7.27d and Figure 7.27f we get abnormally high observations far from the linear pattern of the Pareto QQ. It is for this reason that we expect a robust estimator of the EVI, to outperform all other estimators and not be greatly influenced by the suspicious observations.

Boxplots

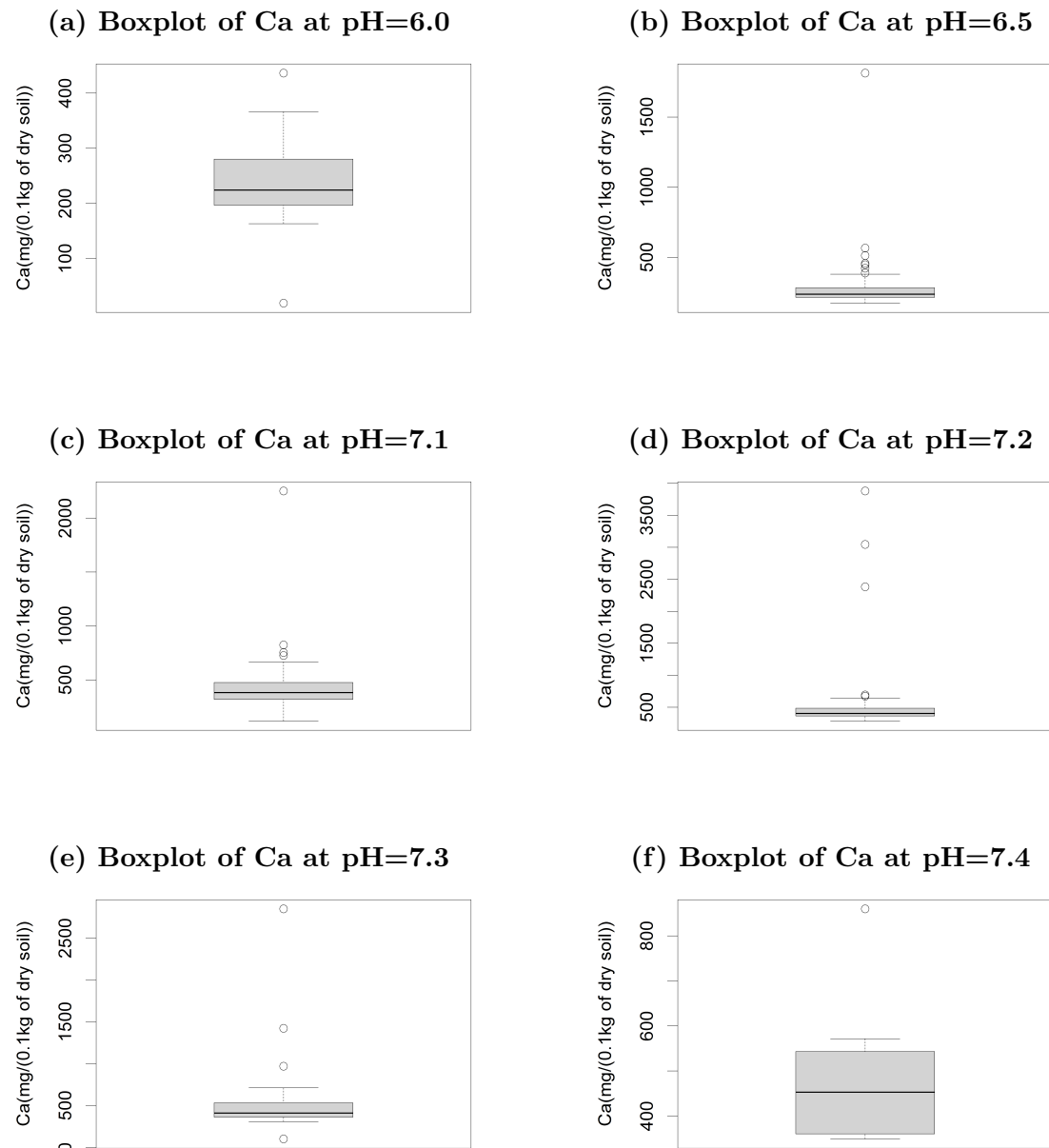
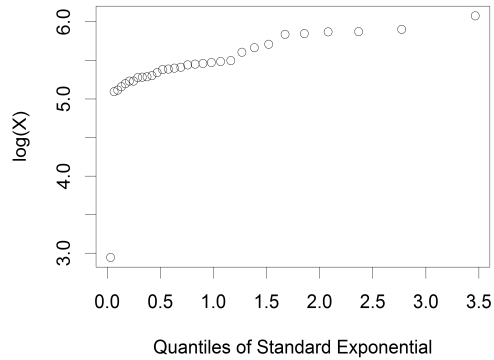


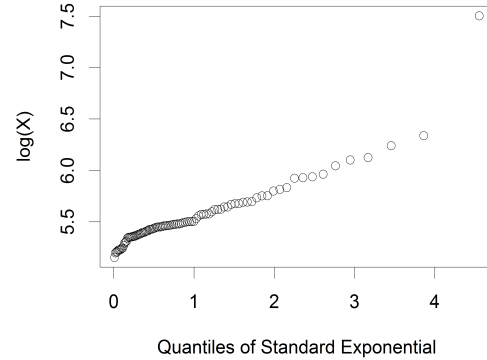
Figure 7.26: Boxplots of the Ca measurements at different pH levels

Pareto quantile Plots

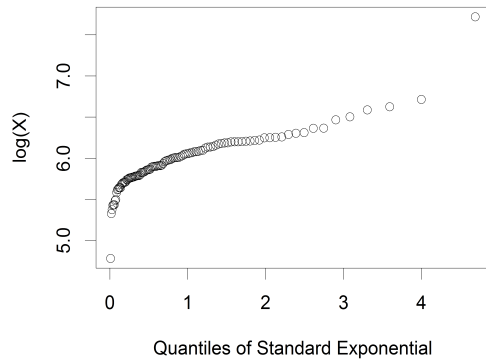
(a) Ca measurement at pH=6.0



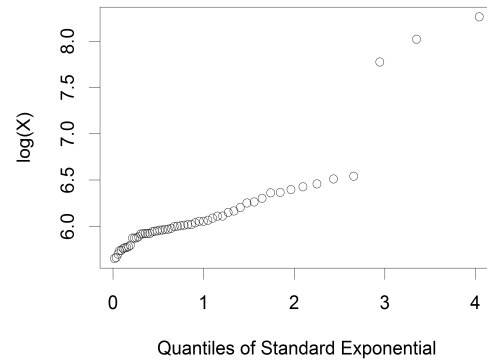
(b) Ca measurement at pH=6.5



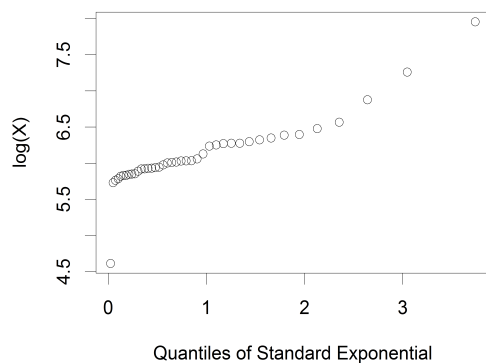
(c) Ca measurement at pH=7.1



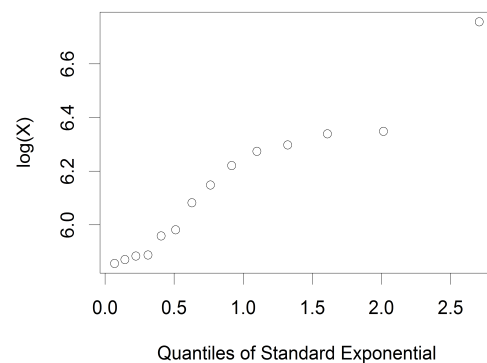
(d) Ca measurement at pH=7.2



(e) Ca measurement at pH=7.3



(f) Ca measurement at pH=7.4

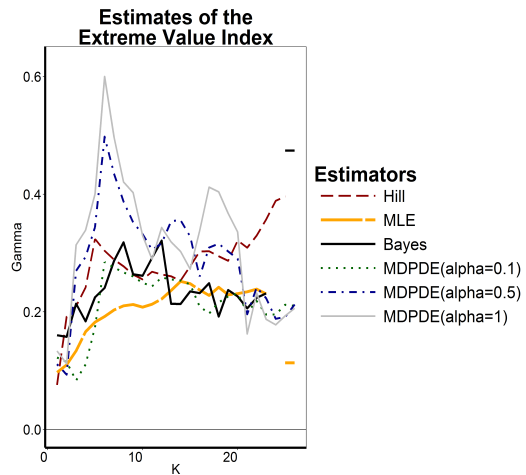
**Figure 7.27:** Pareto quantile plots of the Ca measurements at different pH levels

In Section 7.4.1 we estimate the EVI as well as the corresponding quantile associated with ($p = 1/500$), using similar estimation procedures as in Section 7.3, for

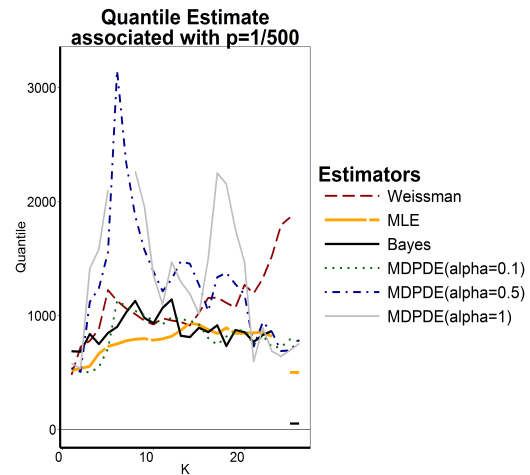
each covariate pH level.

7.4.1 EVI and High Quantile estimates

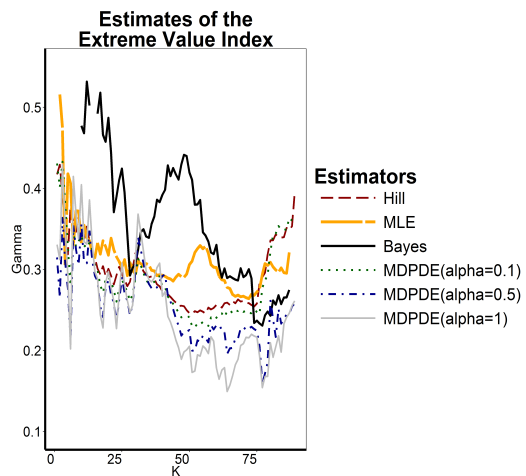
(a) Ca measurement at pH=6.0



(b) Ca measurement at pH=6.0



(c) Ca measurement at pH=6.5



(d) Ca measurement at pH=6.5

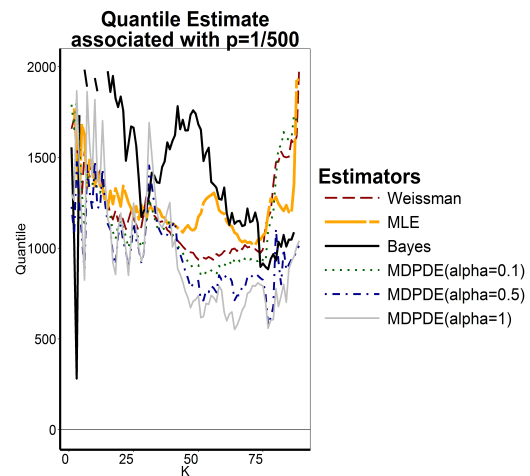
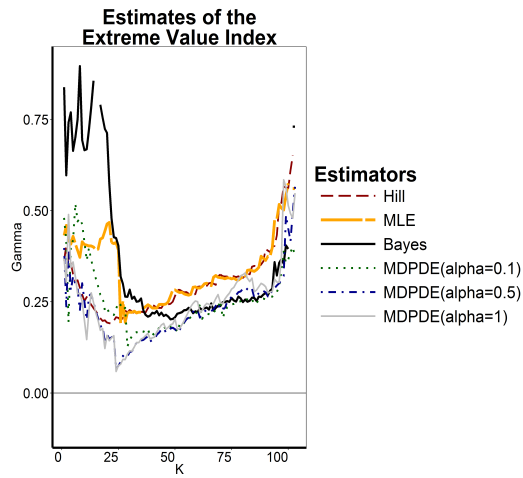
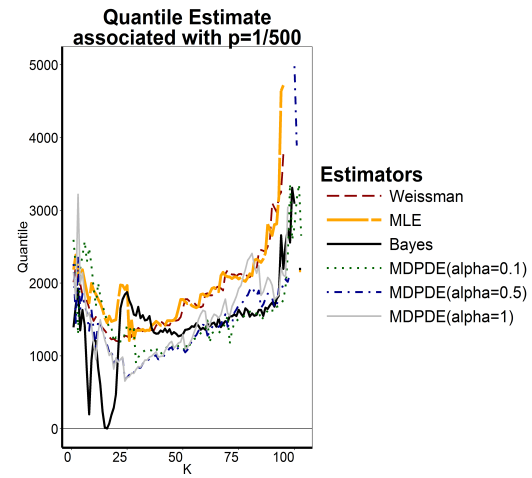


Figure 7.28: EVI (left) and corresponding quantiles (right) of the Ca measurements as a function of k at different pH levels

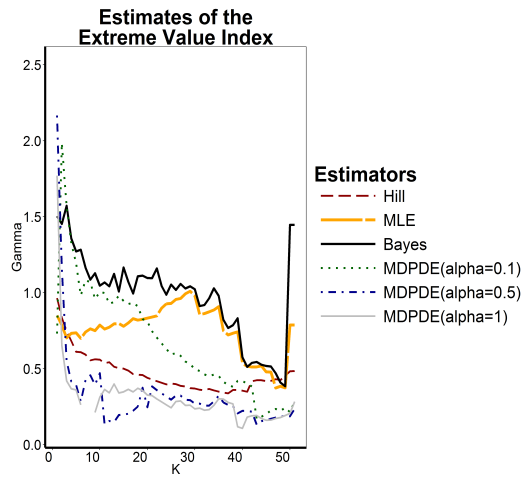
(a) Ca measurement at pH=7.1



(b) Ca measurement at pH=7.1



(c) Ca measurement at pH=7.2



(d) Ca measurement at pH=7.2

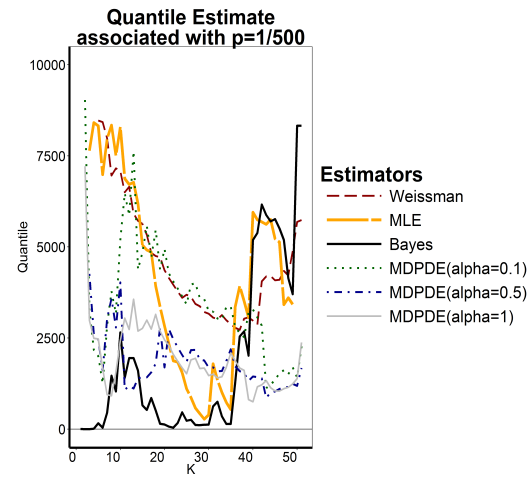
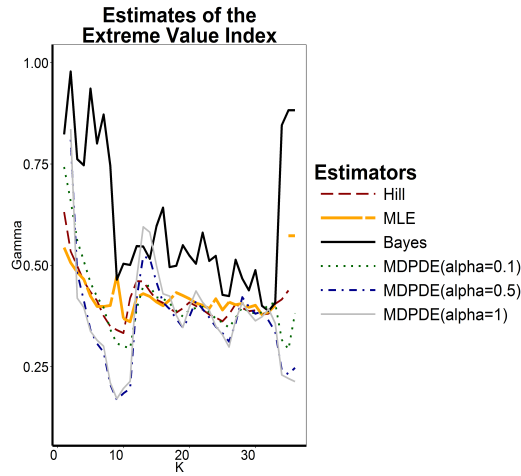
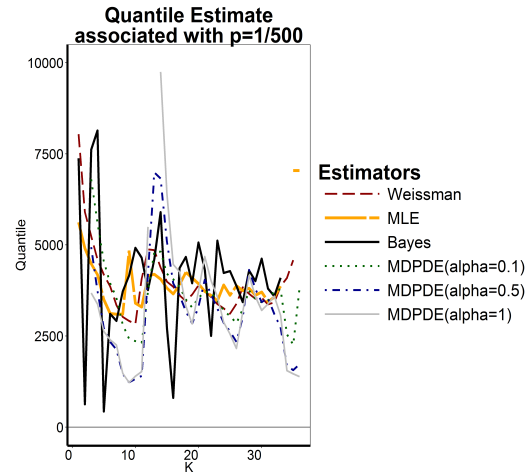


Figure 7.29: EVI (left) and corresponding quantiles (right) of the Ca measurements as a function of k at different pH levels

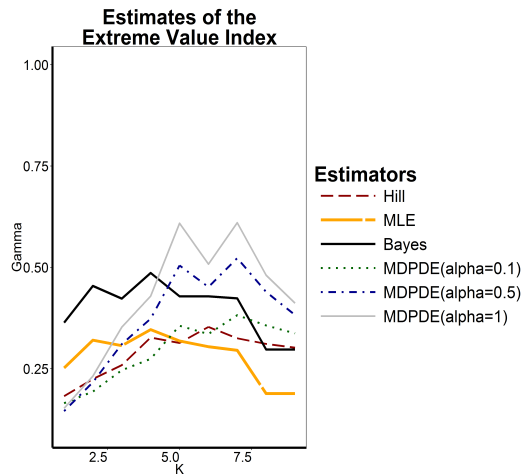
(a) Ca measurement at pH=7.3



(b) Ca measurement at pH=7.3



(c) Ca measurement at pH=7.4



(d) Ca measurement at pH=7.4

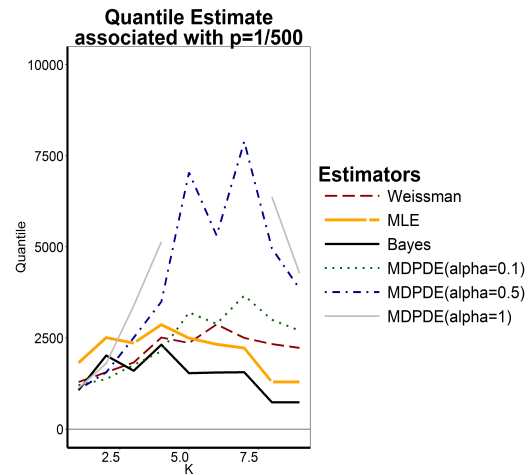


Figure 7.30: EVI (left) and corresponding quantiles (right) of the Ca measurements as a function of k at different pH levels

Remarks

We can see from Figure 7.26 and Figure 7.27 that at pH level 6.0, there exists one abnormally low observation, therefore when we estimate the EVI and the corresponding quantiles at lower thresholds, this observation will be included. Hence, as we can see from Figure 7.28a and Figure 7.28b the MPDE estimators are performing considerably better at lower thresholds. It seems as if $\hat{\gamma}_{0.1,n}$, $\hat{\gamma}_{0.5,n}$ and $\hat{\gamma}_{1,n}$ converge to some value around **0.2** also $\hat{U}_{0.1,n}^{MDPDE}(1/p_n)$, $\hat{U}_{0.5,n}^{MDPDE}(1/p_n)$ and $\hat{U}_{1,n}^{MDPDE}(1/p_n)$ converge to a value around **Ca=800** for lower thresholds. One suspiciously low observation is included at $k \approx n - 1$ and it is for this reason that

estimates of $\hat{\gamma}_n^{Bayes}$ and $\hat{\gamma}_n^{MLE}$ become very explosive.

From Figure 7.27b at a pH level of 6.5 there is evidence of an abnormally high observation, this observation will be included throughout the entire range of thresholds. We can see however from Figure 7.28c and Figure 7.28d that the MDPD estimators give more conservative estimates of the EVI and corresponding quantiles. At lower thresholds $\hat{\gamma}_{0.5,n}$ oscillates around **0.21** and around **0.32** at higher thresholds. Also $\hat{U}_{0.5,n}^{MDPDE}(1/p_n)$ oscillates around **Ca=800** at lower thresholds and around **Ca=1300** at higher thresholds, both estimates are below the suspicious observation.

We can see from Figure 7.27c at pH level 7.1 we have one abnormally high and one abnormally low observation. As explained previously, an abnormally low observation is negated by the fact that we consider only observations above a threshold. From Figure 7.29a and Figure 7.29b it is interesting to note that at high thresholds, the suspicious observation seems to throw off $\hat{\gamma}_n^{Bayes}$ more than any other estimator, this is because the Bayesian estimator is very sensitive to contamination in the data. It can be seen here that the MDPD estimators tend to be more conservative than the other non-robust estimators.

From Figure 7.27d at pH level 7.2 we have 3 abnormally high observations. The amount of contamination at this covariate level can be considered quite high. However in Figure 7.29c and Figure 7.29d $\hat{\gamma}_{0.1,n}$, $\hat{\gamma}_{0.5,n}$ and $\hat{\gamma}_{1,n}$ are not very sensitive to this contamination of data. In fact $\hat{\gamma}_{0.5,n}$ and $\hat{\gamma}_{1,n}$ seem to have much stable paths as compared to other estimators. $\hat{U}_*^{Bayes}(1/p_n)$ seems however to have a much more conservative estimate of the corresponding quantiles. This does not seem very reliable as it sometimes reaches zero. $\hat{U}_{0.5,n}^{MDPDE}(1/p_n)$ and $\hat{U}_{1,n}^{MDPDE}(1/p_n)$ seem to have more stable paths and is thus more reliable.

From Figure 7.27e at pH level 7.3 the data is reasonable Pareto despite the altered path of the Pareto QQ. In such a case, if contamination exists it should be of a very small amount. It can be seen from Figure 7.30a and Figure 7.30b that all considered robust and non-robust estimators pretty much oscillate around a common value. However $\hat{\gamma}_n^{Bayes}$ seems to be more sensitive to this small amount of contamination also $\hat{U}_*^{Bayes}(1/p_n)$ is very volatile when compared to other considered estimators.

From Figure 7.27f there are only a few data points at pH level 7.4 thus making it difficult to make a reliable inference. However there is an abnormally high observation, and linearity of the Pareto QQ can only be affirmed in the midsection of the Pareto QQ. It is clear from Figure 7.30c and Figure 7.30d that at high thresholds, $\hat{\gamma}_n^{Bayes}$ and $\hat{\gamma}_n^{MLE}$ are more sensitive.

7.5 Conclusion

We gave a brief definition of the MDPDE, as well as detailed estimation procedures of the MDPDE. We assessed the performance of the MDPDE and compared it to other estimators considered in this thesis through an exhaustive simulation study. It was clear that when the data is contaminated the MDPDE estimators give better performance than the non-robust estimators more so in the case where the parameter α controlling the trade off between efficiency and robustness is $\alpha = 0.5$. A case study was also carried out using a real dataset, to further illustrate the performance of the MDPD estimators.

We conclude this chapter by recommending the use of the MDPDE estimator when the heavy-tailed data seems suspiciously contaminated. Also we corroborated Dierckx et al. (2013) and Goegebeur et al. (2014)'s recommendation in setting $\alpha = 0.5$ as it seems more appropriate.

Chapter 8

Conclusion

The main aim of this thesis was to investigate the second-order refined Peaks Over Threshold (POT) model called the Extended Pareto Distribution (EPD) and to also look into currently existing methods of parameter estimation of the EPD.

We started with a brief overview of EVT including some literature on historical developments in EVT along with literature on Bayesian methods in EVT, tail estimation, tail estimation under contamination and tail estimation under right censoring.

The EPD was defined and its generalization was shown along with some properties of the Pareto Distribution (PD) and the Generalized Pareto Distribution (GPD). We investigated the effectiveness of the EPD in modelling heavy-tailed distributions and compared it to the GPD in terms of the bias, mean squared error and variance of the EVI. The ML estimation procedure was also discussed in Section 3.3. We verified the EPD to be a better POT model to consider than the GPD in terms of bias reduction.

An improved Maximum Likelihood (ML) estimation procedure was proposed and through a simulation study this ML procedure was compared with the one given in Beirlant et al. (2009). The improved ML estimation procedure seems to have more stable sample paths than the previously used procedure. However this new ML estimation procedure does seem to increase the variance by a tenuous amount.

A Bayesian EPD parameter estimator is also proposed with a detailed estimation procedure. An exhaustive Monte Carlo simulation study was conducted in order to assess the behavior of the Bayesian EPD estimator where the second order parameter is fixed (the first case) and where the second order parameter is estimated concurrently with other parameters (the second case). A case-study was also conducted in using a real dataset to further illustrate the performance of the Bayesian EPD estimator. From the simulation study, it was clear that the Bayesian EPD estimator was a good alternative to the ML method of estimation. In fact in cases where $\gamma > 0.5$, the Bayesian EPD estimator seems to be better than the ML

method. Through a simulation study we further studied how the Bayesian approach estimated the second order parameter τ . The study revealed that although the Bayesian estimation procedure estimated τ in a more stable manner, than the ML and Fraga's consistent estimator (Fraga Alves et al., 2003b), the Bayesian approach seems to have a static margin of error and it therefore does not converge very well to the true value of τ .

We investigated the performance of the EPD under random right censorship. A general treatment of random right censoring was provided and a simulation study was also conducted to assess the behavior of the EVI estimators adapted for censoring. In addition to the Monte Carlo simulation study, a case study on real data was also carried out to further illustrate the performance of the EPD for right censored data. We extended EPD likelihood-adapted approach of Beirlant et al. (2016) into a Bayesian setting, and constructed a censored posterior through which we approximate the posterior of the right tail of the distribution of interest. An additional simulation study was conducted to assess the censored Bayesian estimator in comparison to the censored ML estimator. We verified that when the EPD model is adapted for censoring it results in bias reduced estimates of the true EVI. We found the Bayesian approach to be a good alternative to the ML method more especially at higher levels of censoring and when $\gamma > 0.5$.

We investigated the performance of the EPD when the data is contaminated. A short definition of the minimum density power divergence estimator was given and estimation procedures of the MDPDE were shown.

An exhaustive simulation study was conducted in order to assess the performance of the estimates and a brief case study was carried out to further assess the behavior of the MDPDE in comparison to other mentioned estimators when data is both contaminated and uncontaminated. We ascertained that when the data is uncontaminated, the MDPDE is more reliable when α is close to zero, where α controls the trade off between efficiency and robustness. We further corroborate that when data is contaminated the MDPDE is more reliable when $\alpha = 0.5$.

8.1 Further research possibilities

- The Bayesian approach seems to over estimate the second order parameter by a small margin, one could further investigate how to improve the Bayesian estimation of this parameter which could lead to much improved estimates of the EVI;
- The MDI prior used in this study is assumed to be that of a Pareto. Despite its plausibility in simulation studies we could improve our posterior approximation by deriving a more accurate prior from the EPD model.

References

- Alves, M. F., Gomes, M. I., de Haan, L. and Neves, C. (2007), ‘A note on second order conditions in extreme value theory: linking general and heavy tail conditions’, *REVSTAT Statistical Journal* **5**(3), 285–304.
- Balkema, A. A. and De Haan, L. (1974), ‘Residual life time at great age’, *The Annals of probability* pp. 792–804.
- Basu, A. (1964), ‘Estimates of reliability for some distributions useful in life testing’, *Technometrics* **6**(2), 215–219.
- Basu, A., Harris, I. R., Hjort, N. L. and Jones, M. C. (1998), ‘Robust and efficient estimation by minimising a density power divergence’, *Biometrika Trust* **85**(3), 549–559.
- Beirlant, J., Bardoutsos, A., de Wet, T. and Gijbels, I. (2016), ‘Bias reduced tail estimation for censored pareto type distributions’, *Statistics & Probability Letters* **109**, 78–88.
- Beirlant, J., Figueiredo, F., Gomes, M. I. and Vandewalle, B. (2008), ‘Improved reduced-bias tail index and quantile estimators’, *Journal of Statistical Planning and Inference* **138**(6), 1851–1870.
- Beirlant, J., Guillou, A., Dierckx, G. and Fils-Villetard, A. (2007), ‘Estimation of the extreme value index and extreme quantiles under random censoring’, *Extremes* **10**(3), 151–174.
- Beirlant, J., Joossens, E. and Segers, J. (2009), ‘Second-order refined peaks-over-threshold modelling for heavy-tailed distributions’, *Journal of Statistical Planning and Inference* **139**(8), 2800–2815.
- Beirlant, J., Vynckier, P. and Teugels, J. L. (1996), ‘Tail index estimation, pareto quantile plots regression diagnostics’, *Journal of the American statistical Association* **91**(436), 1659–1667.
- Beirlant, J., Y. Goegebeur, Teugels, J. and Segers, J. (2004), *Statistics of Extremes: Theory and Applications*, Wiley Series in Probability and Statistics, John Wiley & Sons, Ltd.

- Bélisle, C. J. (1992), ‘Convergence theorems for a class of simulated annealing algorithms on rd’, *Journal of Applied Probability* pp. 885–895.
- Bernardo, J. M. and Smith, A. F. (2009), *Bayesian theory*, Vol. 405, John Wiley & Sons.
- Berning, T. L. (2010), Improved estimation procedures for a positive extreme value index, PhD thesis, Stellenbosch University.
- Bhattacharya, S. K. (1967), ‘Bayesian approach to life testing and reliability estimation’, *Journal of the American Statistical Association* **62**(317), 48–62.
- Bingham, N. H., Goldie, C. M. and Teugels, J. L. (1987), ‘Regular variation (encyclopedia of mathematics and its applications)’.
- Brazauskas, V. and Serfling, R. (2000), ‘Robust and efficient estimation of the tail index of a single-parameter pareto distribution’, *North American Actuarial Journal* **4**(4), 12–27.
- Brent, R. P. (2013), *Algorithms for minimization without derivatives*, Courier Corporation.
- Broyden, C. (1970), ‘The convergence of a class of double-rank minimization algorithms 1. general considerations’, *IMA Journal of applied Mathematics* **6**(1), 76–90.
- Byrd, R. H., Lu, P., Nocedal, J. and Zhu, C. (1994), ‘A limited memory algorithm for bound constrained optimization’, *Technical Report NAM-08. NORTHWESTERN UNIVERSITY*.
- Byrd, R. H., Lu, P., Nocedal, J. and Zhu, C. (1995), ‘A limited memory algorithm for bound constrained optimization’, *SIAM Journal on Scientific Computing* **16**(5), 1190–1208.
- Coles, S., Bawa, J., Trenner, L. and Dorazio, P. (2001), *An introduction to statistical modeling of extreme values*, Vol. 208, Springer.
- Coles, S. G. and Powell, E. A. (1996), ‘Bayesian methods in extreme value modelling: a review and new developments’, *International Statistical Review/Revue Internationale de Statistique* pp. 119–136.
- Coles, S. and Tawn, J. (1996), ‘Modelling extremes: a bayesian approach’, *Applied Statistics* **45**, 463–478.
- Davison, A. and Smith, R. L. (1990), ‘Model for exceedances over high thresholds’, *J.R. Statist. Soc.* **B**(52), 237–254.
- de Haan, L. (1970), ‘On regular variation and its application to the weak convergence of sample extremes’, *Mathematical Centre Tract* **32**.

- de Haan, L. (1984), Slow variation and characterization of domains of attraction, in ‘Statistical extremes and applications’, Springer, pp. 31–48.
- de Haan, L. and Stadtmüller, U. (1996), ‘Generalized regular variation of second order’, *Journal of the Australian Mathematical Society (Series A)* **61**(03), 381–395.
- De Zea Bermudez, P. and Kotz, S. (2010), ‘Parameter estimation of the generalised pareto distribution’, *Journal of Statistical Planning and Inference* **140**, 1353–1373.
- Dekkers, A. L., Einmahl, J. H. and De Haan, L. (1989), ‘A moment estimator for the index of an extreme-value distribution’, *The Annals of Statistics* pp. 1833–1855.
- Dell’Aquila, R. and Embrechts, P. (2006), ‘Extremes and robustness: a contradiction?’, *Financial Markets and Portfolio Management* **20**(1), 103–118.
- Diaconis, P. and Freedman, D. (1986), ‘On the consistency of bayes estimates’, *The Annals of Statistics* pp. 1–26.
- Dierckx, G., Goegebeur, Y. and Guillou, A. (2013), ‘An asymptotically unbiased minimum density power divergence estimator for the pareto-tail index’, *Journal of Multivariate Analysis* **121**, 70–86.
- Drees, H. (1996), ‘Refined pickands estimators with bias correction’, *Communications in Statistics - Theory and Methods* **25**(4), 837–851.
- Dupuis, D. (1998), ‘Exceedances over high thresholds: A guide to threshold selection’, *Extremes* **1**(3), 251–261.
- Dupuis, D. J. and Victoria-Feser, M.-P. (2006), ‘A robust prediction error criterion for pareto modelling of upper tails’, *Canadian Journal of Statistics* **34**(4), 639–658.
- Einmahl, J. H., Fils-Villetard, A., Guillou, A. et al. (2008), ‘Statistics of extremes under random censoring’, *Bernoulli* **14**(1), 207–227.
- Engelund, S. and Rackwitz, R. (1992), ‘On predictive distribution functions for the three asymptotic extreme value distributions’, *Structural Safety* **11**(3), 255–258.
- Fisher, R. and Tippett, L. (1928), ‘Limiting forms of the frequency distribution of the largest and smallest member of a sample’, *Proceedings of the Cambridge Philosophical Society* **24**, 180–190.
- Fletcher, R. (1970), ‘A new approach to variable metric algorithms’, *The Computer Journal* **13**(3), 317–322.
- Fletcher, R. and Reeves, C. M. (1964), ‘Function minimization by conjugate gradients’, *The computer journal* **7**(2), 149–154.

- Fraga Alves, M., Gomes, M. and de Haan, L. (2003b), ‘A new class of semi-parametric estimators of the second order parameter’, *Portugaliae Mathematica* **60**, 193–214.
- Fridman, M. and Harris, L. (1998), ‘A maximum likelihood approach for non-gaussian stochastic volatility models’, *Journal of Business & Economic Statistics* **16**(3), 284–291.
- Galambos, J. and Kotz, S. (1978), *Preliminaries and basic results*, Springer.
- Gatarek, L. T., Hoogerheide, L. F., Hooning, K. and Van Dijk, H. K. (2013), ‘Censored posterior and predictive likelihood in bayesian left-tail prediction for accurate value at risk estimation’, *Tinbergen Institute Discussion Paper 13-060/III*.
- Gelfand, A. E. (2000), ‘Gibbs sampling’, *Journal of the American Statistical Association* **95**(452), 1300–1304.
- Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2014), *Bayesian data analysis*, Vol. 2, Taylor & Francis.
- Gilks, W. R. and Wild, P. (1992), ‘Adaptive rejection sampling for gibbs sampling’, *Applied Statistics* pp. 337–348.
- Gnedenko, B. (1943), ‘On the limit distribution of the maximum term of a random series’, *Annals of Mathematics* **44**(3), 423–453.
- Goegebeur, Y., Guillou, A. and Verster, A. (2014), ‘Robust and asymptotically unbiased estimation of extreme quantiles for heavy tailed distributions’, *Statistics & Probability Letters* **87**, 108–114.
- Goegebeur, Y., Planchon, V., Beirlant, J. and Oger, R. (2005), ‘Quality assessment of pedometer data using extreme value methodology’, *Journal of applied sciences* **5**, 1092–1102.
- Goldfarb, D. (1970), ‘A family of variable metric updates derived by variational means’, *Mathematics of Computation* **24**(109), 23–26.
- Gomes, I. M. and Neves, M. M. (2011), ‘Estimation of the extreme value index for randomly censored data’, *Biometrical Letters* **48**(1), 1–22.
- Gomes, M. I., Rodrigues, L. H., Pereira, H. and Pestana, D. (2010), ‘Tail index and second-order parameters’ semi-parametric estimation based on the log-excesses’, *Journal of Statistical Computation and Simulation* **80**(6), 653–666.
- Gumbel, E. J. (2004), *Statistics of Extremes*, Courier Dover Publications.
- Hall, P. (1982), ‘On some simple estimates of an exponent of regular variation’, *J.R. Statist. Soc. B* **44**(1), 37–42.

- Hall, P. and Welsh, A. (1985), ‘Adaptative estimates of parameters of regular variation’, *Ann. Statist.* **13**, 331–341.
- Hastings, W. K. (1970), ‘Monte carlo sampling methods using markov chains and their applications’, *Biometrika* **57**(1), 97–109.
- Hill, B. (1975), ‘A simple general approach to inference about the tail of a distribution’, *annals of statistics* **3**(5), 1163–1974.
- Holla, M. (1966), ‘Bayesian estimates of the reliability function’, *Australian Journal of Statistics* **8**(1), 32–35.
- Hubert, M., Dierckx, G. and Vanpaemel, D. (2013), ‘Detecting influential data points for the hill estimator in pareto-type distributions’, *Computational Statistics & Data Analysis* **65**, 13–28.
- Jeffreys, H. (1961), ‘Theory of probability’, *Oxford University Press* .
- Juárez, S. F. and Schucany, W. R. (2004), ‘Robust and efficient estimation for the generalized pareto distribution’, *Extremes* **7**(3), 237–251.
- Karamata, J. (1930), ‘Sur un mode de croissance reguliere des fonctions, mathematica (cluj) 4 (1930), 38-53’, *Karamata384Mathematica (Cluj)* .
- Kass, R. E., Carlin, B. P., Gelman, A. and Neal, R. M. (1998), ‘Markov chain monte carlo in practice: a roundtable discussion’, *The American Statistician* **52**(2), 93–100.
- Kim, M. and Lee, S. (2008), ‘Estimation of a tail index based on minimum density power divergence’, *Journal of multivariate Analysis* **99**(10), 2453–2471.
- Klein, J. and Moeschberger, M. (2005), *Datasets for Survival Analysis – Techniques for Censored and Truncated Data*, Springer.
- Leugrands, S. (1987), ‘Linear models, random censoring and synthetic data’, *Biometrika* **74**(2), 301–309.
- Lindley, D. V. (1958), ‘Fiducial distributions and bayes’ theorem’, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 102–107.
- Loynes, R. (1986), ‘Statistical extremes and applications-j. tiago de oliveira.’, *Metrika* **33**, 110–110.
- Lunn, D., Jackson, C., Best, N., Thomas, A. and Spiegelhalter, D. (2012), *The BUGS book: A practical introduction to Bayesian analysis*, CRC press.
- Matthys, G. and Beirlant, J. (2003), ‘Estimating the extreme value index and high quantiles with exponential regression models’, *Statistica Sinica* **13**, 853–880.

- McNeil, A. and Saladin, T. (1997), ‘The peaks over thresholds method for estimating high quantiles of loss distributions’, *Departement Mathematik ETH Zentrum*.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953), ‘Equation of state calculations by fast computing machines’, *The journal of chemical physics* **21**(6), 1087–1092.
- Neal, R. M. (2003), ‘Slice sampling’, *Annals of statistics* pp. 705–741.
- Nelder, J. A. and Mead, R. (1965), ‘A simplex method for function minimization’, *The computer journal* **7**(4), 308–313.
- Peng, L. and Welsh, A. (2001), ‘Robust estimation of the generalized pareto distribution’, *Extremes* **4**(1), 53–65.
- Pericchi, L. and Smith, A. (1992), ‘Exact and approximate posterior moments for a normal location parameter’, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 793–804.
- Pfaff, B., McNeil, A. and Stephenson, A. (2012), *Extreme Values in R*.
URL: <http://cran.r-project.org/web/packages/evir/evir.pdf>
- Pickands III, J. (1994), Bayes quantile estimation and threshold selection for the generalized pareto family, in ‘Extreme Value Theory and Applications’, Springer, pp. 123–138.
- Pickands, J. (1975), ‘Statistical inference using extreme order statistics’, *The Annals of statistics* **3**(1), 119–131.
- R Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: <http://www.R-project.org/>
- Ramsey, F. P. (1931), ‘Truth and probability (1926)’, *The foundations of mathematics and other logical essays* pp. 156–198.
- Reiss, R.-D. and Thomas, M. (1999), ‘A new class of bayesian estimators in paretian excess-of-loss reinsurance’, *Astin Bulletin* **29**(02), 339–349.
- Reiss, R.-D., Thomas, M. and Reiss, R. (2001), *Statistical analysis of extreme values*, Springer.
- Schwartz, L. (1965), ‘On bayes procedures’, *Probability Theory and Related Fields* **4**(1), 10–26.
- Shanno, D. (1970), ‘Conditioning of quasi-newton methods for function minimization’, *Mathematics of Computation* **24**(111), 647–656.

- Sinha, S. K. and Sloan, J. (1988), ‘Bayes estimation of the parameters and reliability function of the 3-parameter weibull distribution’, *Reliability, IEEE Transactions on* **37**(4), 364–369.
- Smith, R. L. (1987), ‘Estimating tails of probability distributions’, *The Annals of Statistics* **15**, 1174–1207.
- Spiegelhalter, D. J., Thomas, A., Best, N. G. and Gilks, W. R. (1995), ‘Bugs: Bayesian inference using gibbs sampling, version 0.50’, *MRC Biostatistics Unit, Cambridge* .
- Stanton, R. (1987), ‘The work of lhc tippet’, *Ars Textrina* **7**, 179–185.
- Sturtz, S., Ligges, U. and Gelman, A. (2010), ‘R2openbugs: a package for running openbugs from r’.
- Taleb, N. N. (2010), *The black swan:: The impact of the highly improbable fragility*, Vol. 2, Random House.
- Vandewalle, B. and Beirlant, J. (2006), ‘On univariate extreme value statistics and the estimation of reinsurance premiums’, *Insurance: Mathematics and Economics* **38**(3), 441–459.
- Vandewalle, B., Beirlant, J., Christmann, A. and Hubert, M. (2007), ‘A robust estimator for the tail index of pareto-type distributions’, *Computational Statistics & Data Analysis* **51**(12), 6252–6268.
- Weissman, I. (1978), ‘Estimation of parameters and large quantiles based on the k largest observations’, *Journal of the American Statistical Association* **73**(364), 812–815.
- Worms, J. and Worms, R. (2014), ‘New estimators of the extreme value index under random right censoring, for heavy-tailed distributions’, *Extremes* **17**(2), 337–358.
- Zellner, A. (1997), *Bayesian Analysis in Econometrics and Statistics*, The Zellner View and Papers, Edward Elgar Publ. Co.