PREDICTIVE MODELLING OF EXTREME VALUES
IN UNBALANCED PANEL DATA

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE
IN
STATISTICS
UNIVERSITY OF REGINA

By
Xiaohua Liu

Regina, Saskatchewan
June 2019

© Copyright 2019: Xiaohua Liu
Xiaohua Liu, candidate for the degree of Master of Science in Statistics, has presented a thesis titled, *Predictive Modelling of Extreme Values in Unbalanced Panel Data*, in an oral examination held on June 11, 2019. The following committee members have found the thesis acceptable in form and content, and that the candidate demonstrated satisfactory knowledge of the subject material.

External Examiner: Dr. Samuel Gamtessa, Department of Economics

Supervisor: Dr. Taehan Bae, Department of Mathematics & Statistics

Committee Member: Dr. Andrei Volodin, Department of Mathematics & Statistics

Chair of Defense: Dr. Lisa Fan, Department of Computer Science
Abstract

This thesis aims at predictive modelling of a two-sided, heavy-tailed data under a mixture model setting. A robust regression method is used to fit the main body, while the peaks-over-threshold method is employed to select the tails or the extreme events. Based on the extreme value theory, the tails are modelled with Pareto or exponential distributions. For the estimation of the tail distributions, the Bayesian maximum a posterior estimation (MAP) with conjugate priors is used to smooth the maximum likelihood estimates (MLEs). With regard to each of the two tail, the MAP approach leads to two optimization problems for the estimation of tail parameters: the tail decay rate and the tail quantile level. This filter tuning process provides stability and efficiency in computation and prediction. Several constrained, non-convex optimization problems have been converted to unconstrained, convex problems by quadratic approximation and variable changes. Newton’s iteration method is employed to solve the optimization problems numerically.

This formulated methodology is applied to a large, multi-period, unbalanced data
set of daily returns of global stocks, containing nearly 120,000 records. The out-of-sample prediction results show the out-performance of the smoothed estimates over the regular MLEs.
Acknowledges

I appreciate all my supervisor Dr. Taehan Bae’s contributions of suggestions, ideas, time, patience and funding to make my master's experience productive. I would like to thank Dr. Andrei Volodin for being my supervisory committee member and thank the external examiner, Dr. Samuel Gamtessa, for the careful review of my thesis and the insightful questions and comments.

I gratefully acknowledge supports from Department of Mathematics and Statistics, Faculty of Graduate Studies and Research for contributing to my Master study. I was awarded UR Graduate Scholarship - 2018 Fall, Faculty of Graduate Studies and Research Master GTA Scholarship - 2018 Fall, UR Graduate Scholarship - 2019 Winter and Faculty of Graduate Studies and Research Master GTA Scholarship - 2019 Winter. I was also supported by Dr. Taehan Bae in summer 2019 to present my study at the 2019 Statistical Society of Canada (SSC) Annual Meeting in Calgary.

Finally, I give my special thanks to my family for their unconditional love, support and encouragement. Thank you all!
Contents

Abstract i

Acknowledgements iii

Contents iv

Table of Contents iv

List of Figures vii

List of Tables viii

1 Introduction 1

1.1 Backgrounds ............................................. 1

1.2 Purpose of Study and Contributions .................. 3

1.3 Thesis Structure ....................................... 4

2 Preliminaries 5
2.1 Heavy-Tailed Distribution .................................................. 6
  2.1.1 Distributions and Properties ...................................... 7
  2.1.2 Pareto Distribution .................................................... 7
2.2 Robust Regression ............................................................. 9
2.3 Extreme Value Theory ...................................................... 11
  2.3.1 Block Maximum Method .............................................. 11
  2.3.2 Peaks-Over-Threshold Method ..................................... 13
2.4 Bayesian Model .............................................................. 15
  2.4.1 Bayes' Theorem ...................................................... 15
  2.4.2 Maximum a Posterior Estimation .................................. 16
2.5 Optimizations ............................................................... 18
  2.5.1 Quadratic Approximation ............................................ 19
  2.5.2 Convex Optimization .................................................. 19
  2.5.3 Newton Algorithm .................................................... 20

3 Methodology ........................................................................ 21
  3.1 Problem Formulation ....................................................... 22
  3.2 Robust Regression .......................................................... 24
  3.3 Tail Models ................................................................. 26
    3.3.1 POT Method .......................................................... 26
    3.3.2 MLEs ................................................................. 30
List of Figures

3.1 Flowchart of estimation process ........................................ 22

3.2 Mean residual plot for stock return data ............................. 27

4.1 MLE vs. MAP estimates of $\theta_{t_i}^+$ .................................. 48
List of Tables

3.1 Indexes for prior construction ........................................ 34

4.1 Summary statistics of daily return dataset .......................... 45

4.2 Summary of robust regression .......................................... 47

4.3 Model fitting results - out of sample performance ............... 49

4.4 Model fitting results - in sample performance .................... 50
Chapter 1

Introduction

Aiming to describe unusual (rather than usual) events, extreme value theory provides a framework to estimate anticipated forces from historical data.

At first it looks as if this is an impossible task. Simply speaking, with just over 100 years of historical data, how can one estimate the probability of a once-in-10000 year event? Such thought is backed by the fact that the empirical distribution function carries all the information acquired, and going beyond its range is impossible. However extreme value theory and statistical modelling make it possible through (semi-)parametric methods, see de Haan (2006) for more details.

1.1 Backgrounds

Started from an abstract study of random phenomena, the study of extreme value theory can be tracked back to the early part of the 20th century. It has merged as
one of the most important statistical disciplines for the applied sciences over the last 70 years. This field was pioneered by Leonard Tippett (1902-1985) who worked to make cotton thread stronger. In his studies, he realized that the strength of a thread was controlled by the strength of its weakest fibres. With the help of R. A. Fisher (1890-1962), Tippett obtained three asymptotic limits describing the distributions of extremes assuming independent variables. Emil Julius Gumbel (1891-1966) reviewed this theory in his book *Statistics of Extremes*, including the Gumbel distribution that bore his name. Until the 1980s did the contours of the statistical theory take shape.

The first important theorem in extreme value theory is Fisher-Tippett-Gnedenko theorem (Fisher and Tippett (1928) and Gnedenko (1945)), also known as extreme value theorem, which is similar to the central limit theorem, except that the central limit theorem applies to the average of a sample from any distributions with finite variance, while the Fisher-Tippett-Gnedenko theorem only states that if the distribution of a normalized maximum converges, then the limit has to be one of a particular class of distributions (named as the generalized extreme value family). It relies on forming block maxima (or minima) as a preliminary step. In many practical situations it is customary and convenient to extract the annual maxima (or minima), generating an "Annual Maxima (or minima) Series". The Pickands-Balkema-de Haan theorem is often called the second theorem in extreme value theory (Balkema and de
Haan (1974), Pickands (1975)). It gives the asymptotic tail distribution of a random variable $X$, when the true distribution $F$ of $X$ is unknown. Unlike the first theorem, the interest here is extracting and analysing a group of peak values exceeding a certain threshold (or falling below a certain threshold). This method refers to as the Peaks-Over-Threshold (POT) method.

### 1.2 Purpose of Study and Contributions

This thesis has three main goals. First of all, we discuss the predictive modelling of the extreme values based on the POT framework within the classical extreme value theory. The next goal is to extend the methodology developed for one-tailed extreme value to a two-tailed one, and to provide reasonable and computationally efficient methods for multiple threshold selections. The third objective is to apply the developed methodology to a large stock return dataset.

The main contributions of this thesis to the application domain of extreme value analysis can be summarized as follows. Firstly, it proposes a computationally efficient methodology for modelling a two-sided heavy-tailed dataset. Next, regarding the several threshold selections, instead of heuristic choices based on the empirical quantiles, it uses a more theoretically sound method based on mean residuals plots. Additionally, it demonstrates that the prediction based on the smoothed MAP estimates can handle the common issue of missing and incomplete data, which is not
plausible under the MLEs.

1.3 Thesis Structure

The rest of this thesis organized with four chapters. Chapter 2 provides a brief literature review on heavy-tailed distribution, robust regression, extreme value theory, Bayesian model and optimization. Chapter 3 presents a methodology for predictive modelling of two-tailed extreme values. Based on the Pickands-Balkema-de Haan theorem, the POT method is used for the selection of tail values. A filter tuning process based on maximum a posterior (MAP) estimation which is used to smooth the maximum likelihood estimates (MLE) and to obtain reasonable and stable prediction results, is described. In Chapter 4, we apply the proposed methodology to a real stock return dataset. Finally, Chapter 5 gives some concluding remarks and future research.
Chapter 2

Preliminaries

In this chapter, a brief review of literature related to my research is presented. To begin with, some notions and examples of heavy-tailed distributions which play a major role in the analysis of extreme values, are given. Next the key concept of robust regression is explained for the purpose of fitting major trend and then selecting data potion for extreme value analysis. It follows a brief review of the classical extreme value theory which provides the theoretical foundation of the methods used in this research. We conclude this chapter with a review of Bayesian modelling approach and quadratic optimization problems which are involved in the process of smoothing parameter estimates.
2.1 Heavy-Tailed Distribution

Before introducing a formal definition of extreme values, let us start with a commonly used term "outlier". In statistics, an outlier is a data point that is distant from other observations. Such outliers can occur by chance as measurement errors, or the population in fact has a heavy-tailed distribution. Here we consider such outliers as extreme values generated from a heavy-tailed distribution.

Heavy-tailed distributions are essential tools for modelling extreme losses or risks in various areas including insurance, finance, environmental sciences and climatology. They also play a major role in the analysis of many stochastic systems. For example, heavy-tailed distributions are frequently used to accurately model inputs to computer and communication networks. They are also essential components of describing many risk processes. For example, they are effectively used in modelling epidemiological spread. There are substantial statistical evidences for their appropriateness in physics, geoscience and economics. See Foss (2013) for some applications.

Some important examples include Pareto distributions (and other essential power-law distributions), lognormal distributions, and Weibull distributions (with shape parameter less than one).
2.1.1 Distributions and Properties

Foss (2013) defines heavy-tailed as:

A distribution $F$ on the real line $\mathbb{R}$, is said to be (right-) heavy-tailed if

$$\int_{-\infty}^{\infty} e^{\lambda x} dF(x) = \infty \text{ for all } \lambda > 0,$$

that is, if and only if $F$ fails to possess any positive exponential moment. Otherwise $F$ is said to be light-tailed. The distribution $F$ is heavy-tailed if and only if its tail function $\overline{F}(x) = 1 - F(x) = F(x, \infty), x \in \mathbb{R}$, fails to be bounded by any exponentially decreasing function.

A class of long-tailed distributions is an important subclass of heavy-tailed distribution. Formally, a distribution $F$ on $\mathbb{R}$ is said to be long-tailed if $F$ has the right-unbounded support and, for any fixed $y > 0$,

$$\frac{F(x+y)}{F(x)} \to 1 \text{ as } x \to \infty.$$

What makes some heavy-tailed distribution to be long-tailed is the degree of smoothness in the tail distribution function $\overline{F}(x)$.

2.1.2 Pareto Distribution

The Pareto distribution is the most widely used heavy-tailed distribution. Suppose $X$ is a random variable with a Pareto distribution, then the survival function
(also called the tail function, see more details in Dobson, 2008), is given by

$$\bar{F}(x) = \Pr(X > x) = \begin{cases} 
(x_m/x)^\alpha & x \geq x_m \\
1 & x < x_m,
\end{cases}$$

where \(x_m\) is the (necessarily positive) minimum possible value of \(X\), and \(\alpha\) is a positive parameter. Therefore, the Pareto distribution is characterized by a scale parameter \(x_m\) and a shape parameter \(\alpha\) which is known as the tail index. Specifically when this distribution is used to model the distribution of wealth, the parameter \(\alpha\) is called the Pareto index.

It is important to note that the Pareto distribution is related to the exponential distribution as follows. If \(X\) is Pareto-distributed with minimum value \(x_m\) and index \(\alpha\), then

$$Y = \log\left(\frac{X}{x_m}\right)$$

is exponentially distributed with rate parameter \(\alpha\). Equivalently, if \(Y\) is exponentially distributed with rate \(\alpha\), then \(x_m \exp(Y)\) is Pareto-distributed with minimum value \(x_m\) and index \(\alpha\). This can be shown using the standard change-of-variable
techniques:

$$\Pr(Y < y) = \Pr\left(\log\left(\frac{X}{x_m}\right) < y\right)$$

$$= \Pr(X < x_m e^y)$$

$$= 1 - \left(\frac{x_m}{x_m e^y}\right)^\alpha$$

$$= 1 - e^{-\alpha y}.$$  \hspace{1cm} (2.1)

The last expression (2.1) is the cumulative distribution function of an exponential distribution with rate $\alpha$. This relation will be used later in Chapter 4.

2.2 Robust Regression

The Central Limit Theorem is the most fundamental result in classical statistical theory.

**Central Limit Theory**: Let $\{X_1, \cdots, X_n\}$ be a random sample of size $n$, that is, a sequence of independent and identically distributed (i.i.d.) random variables drawn from a distribution with mean $\mu$ and finite variance $\sigma^2$. Suppose we are interested in the sample average

$$S_n := \frac{X_1 + \cdots + X_n}{n}$$

of these random variables. By the law of large numbers, the sample average converge in probability (and almost surely) to the expected value $\mu$ as $n \to \infty$.

The most widely used model assumption is that the observed data is drawn from
a normal (Gaussian) distribution. This assumption has been present in statistics for two centuries, and has been the framework for the classical methods in regression, analysis of variance and multivariate analysis. However, the classical statistical methods that rely on the normality assumption, fail to provide reliable results when applied to non-normal data. The robust approach to statistical modelling and data analysis aims at developing methods that produce reliable parameter estimates, associated statistical tests and confidence intervals, not only when the data follows a given distribution exactly, but also when this happens only approximately.

Regarding the identification of outliers, a traditional measure of the outlyingness of an observation $x_i$ with respect to a sample, is the ratio between its distance to the sample mean $\bar{x}$ and the sample standard deviation $s$:

$$t_i = \frac{x_i - \bar{x}}{s}.$$  

As discussed in Maronna (2006), observations with $|t_i| > 3$ are traditionally deemed as suspicious (the three-sigma rule), based on the fact that they would be very unlikely under the normality assumption. Removing these outlying observations and modelling the rest of the data points are the key steps of robust regression.

The purpose of using robust regression for modelling the majority data is to increase the reliability and accuracy of statistical modelling in regression settings. In fact robust regression is an essential tool when detecting extreme values, since they focus on the estimation of majority body by limiting the influence of outlier and thus,
keep the outlyingness of the extreme values.

2.3 Extreme Value Theory

Compared with classical statistical methods and robust regression which focus on the main body of the data distribution, extreme value theory is concerning about "meaningful outliers"; according to the Pareto approach, data points belong to the tail distribution can have a much higher impacts in reality.

In determining what extreme values are, there are two kinds of extreme value selection methods introduced in Coles (2001): Block Maximum (BM) and Peaks-Over-Threshold (POT). In the following, two limit theorems associated with these two methods are presented.

2.3.1 Block Maximum Method

In Block Maximum method, non-overlapping, equal-sized sets of observations are formed, and the maximum observation in each set is taken. The selected observations approximately follow a generalized extreme value distribution (GEV). The BM method is a natural way to determine extreme values but it can be a wasteful approach since only one observation is selected in each subset. This results in omission of some of the large observations in the set with multiple large values, but some of the small observations (for the subsets with all small values) will be contained.
In the following a formal definition of the GEV distribution is given. Let \( \{X_1, X_2, \ldots\} \) be a sequence of i.i.d. random variables having a distribution function \( F \) and \( M_n \) be the maximum value in this sequence, i.e., \( M_n = \max\{X_1, X_2, \ldots, X_n\} \).

**Theorem 1** (Extreme Value Theorem, Fisher and Tippett (1928), Gnedenko (1943))

Suppose there exist sequences of constants \( \{a_n > 0\} \) and \( \{b_n\} \) such that

\[
\Pr\left\{ \frac{M_n - b_n}{a_n} \leq z \right\} \to G(z) \quad \text{as} \quad n \to \infty,
\]

where \( G \) is a non-degenerate distribution function. Then \( G \) is a member of the GEV family

\[
G(z) = \exp\left\{-\left[1 + \xi \left(\frac{z - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}
\]

defined on \( \{z : 1 + \xi(z - \mu)/\sigma > 0\} \), where \( -\infty < \mu < \infty \), \( \sigma > 0 \) and \( -\infty < \xi < \infty \).

This theory indicates that, when \( M_n \) can be stabilized with suitable sequences \( \{a_n\} \) and \( \{b_n\} \), the corresponding normalized variable \( M_n^* \) has a limiting distribution that must be one of the three types of extreme value distributions. This means if we can find a set of sequences \( a_n \) and \( b_n \) to let \( M_n \) converge, then the limiting distribution must belongs to the GEV family:

\[
G(z) = \exp\left\{-\left[1 + \xi \left(\frac{z - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}.
\]

Depending on the tail index parameter \( \xi \), we have the following three special cases:
I. Gumbel as $\xi \to 0$

$$G(z) = \exp\left\{ -\exp\left[ -\left(\frac{z - b}{a}\right)^{-\frac{1}{\xi}}\right] \right\}, \ -\infty < z < \infty$$

II. Fréchet as $\xi > 0$

$$G(z) = \begin{cases} 
0 & z \leq b, \\
\exp\{-\left(\frac{z - b}{a}\right)^{-\xi}\} & z > b; 
\end{cases}$$

III. Weibull as $\xi < 0$

$$G(z) = \begin{cases} 
\exp\{-\left[ -\left(\frac{z - b}{a}\right)^{\xi}\right]\} & z < b; \\
1 & z \geq b. 
\end{cases}$$

Each family has a location parameter ($b$), scale parameter ($a$), and the Fréchet and the Weibull families have a shape parameter ($\xi$).

### 2.3.2 Peaks-Over-Threshold Method

The POT method is a popular and effective approach to define extreme value.

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables having the marginal distribution function $F$. Given that $X_i$ exceeds some high threshold $u$, the conditional probability of exceedance is

$$\Pr\{X > u + y|X > u\} = \frac{1 - F(u + y)}{1 - F(u)}, \ y > 0.$$
**Theorem 2** (Pickands-Balkema-de Haan Theorem, *Pickands* (1975), *Balkema and de Haan* (1974)) Suppose a sequence satisfies the conditions of Theorem 1, so that for large $n$,

\[ \Pr\{M_n \leq z\} \approx G(z). \]

Then for large enough $u$, the distribution function of $(X - u)$, conditional on $X > u$, is approximately

\[ \Pr\{X \leq u + y|X > u\} \approx H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi}, \]

defined on \( \{y : y > 0 \text{ and } (1 + \xi y/\tilde{\sigma}) > 0\} \), where \( \tilde{\sigma} = \sigma + \xi(u - \mu) \).

The function $H(y)$ is called the generalized Pareto distribution (GPD). The theorem implies that, if the block maxima has the approximate distribution $G$, then the threshold exceedances have a corresponding approximate distribution within the generalized Pareto family.

Specifically, if $\xi = 0$, $H(y)$ corresponds to an exponential distribution with parameter $1/\tilde{\sigma}$; if $\xi > 0$, $H(y)$ corresponds to the Pareto distribution. That is, the tail distribution has either exponential tail or Pareto tail. However, the logarithm of a Pareto random variable is exponentially distributed as already proven in the previous section. Based on this relationship, we can simply consider the conditional distribution as exponential distribution on either the original observations (for exponential tail) or the logarithm of the original observations (for Pareto tail).
2.4 Bayesian Model

Bayesian statistics is a statistical theory based on the Bayesian interpretation of probability where the probability expresses a degree of belief in an event. It describes the conditional probability of an event based on the data as well as the prior information or beliefs about the event or the conditions related to the event. Bayes’ theorem can directly assign a probability distribution that quantifies the belief to the parameter or a set of parameters. A Bayesian model has the identifying feature by the specification of the prior distributions for any unknown parameters.

2.4.1 Bayes’ Theorem

In probability theory and statistics, Bayes’ theorem (alternatively Bayes’ law or Bayes’ rule) describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

As in Hogg (2005), Bayes’ theorem is stated mathematically as follows:

\[ \Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)}, \quad (2.2) \]

where \( A \) and \( B \) are events and \( \Pr(B) \neq 0 \). \( \Pr(A \mid B) \) (or \( \Pr(B \mid A) \)) is a conditional probability of the occurrence of event \( A \) (or \( B \)) given that \( B \) (or \( A \)) is true. \( \Pr(A) \) and \( \Pr(B) \) are the probabilities of observing \( A \) and \( B \), respectively. These are called the marginal probabilities.
In (2.2):

- Pr(A) is the prior probability which is the initial degree of belief in A.
- Pr(B | A) is the conditional probability or likelihood that measures the degree of belief in B, given that the proposition A is true.
- Pr(A | B) is the posterior probability which is the probability of A after taking into account B for and against A.

For different choices of the priors, as long as we collect enough data, the posteriors will converge to the true one.

In a parametrized form, the prior is often assumed to come from the conjugate prior. If the posterior distributions are in the same probability distribution family as the prior probability distribution, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function. The usefulness of a conjugate prior is that the corresponding posterior distribution will be in the same family, and the expressions may be given in a closed form.

### 2.4.2 Maximum a Posterior Estimation

For parameter estimation, Bayes’ theorem provides a hint: the posterior distribution \( \pi_{\theta | x}(\theta | x) \) which is the conditional probability distribution of the parameters
given the observed data can be computed as

$$\pi_{\Theta|X}(\theta | x) = \frac{f_{X|\Theta}(x | \theta)\pi(\theta)}{\int f_{X|\Theta}(x | \theta)\pi(\theta) d\theta},$$

while the predictive distribution which is the conditional probability distribution of a new observation $y$ given that data $x$ can be computed as

$$f_{Y|X}(y | x) = \int f_{Y|\Theta}(y | \theta)\pi_{\Theta|X}(\theta | x)d\theta,$$

where $f_{Y|\Theta}(y | \theta)$ is the probability density function of the new observation, given the parameter value. $\pi(\theta)$ is the prior distribution which is a probability distribution over the space of possible parameter values. See Klugman (2008) for more details about posterior distribution.

The maximum a posterior (MAP) estimation can be used to obtain a point estimate of an observed quantity on the basis of empirical data. Compared with the maximum likelihood estimation, the MAP employs an augmented optimization objective over the quantity one wants to estimate. This augmented component incorporates a prior probability which quantifies the additional information available through prior knowledge of related events. The MAP estimation can be seen as a regularization of MLE.

- Log-likelihood function

**MLE** \[ \log[f_{X|\Theta}(x | \theta)] \]

**MAP** \[ \log[\pi_{\Theta|X}(\theta | x)] = \log[f_{X|\Theta}(x | \theta)\pi(\theta)] + c = \log[f_{X|\Theta}(x | \theta)] + \log[\pi(\theta)] + c \]
- Estimators

\[
\begin{align*}
\text{MLE} & \quad \hat{\theta}_{\text{MLE}} = \arg \max \log f_{X|\theta}(x \mid \theta) \\
\text{MAP} & \quad \hat{\theta}_{\text{MAP}} = \arg \max \{ \log f_{X|\theta}(x \mid \theta) + \log \pi(\theta) \} 
\end{align*}
\]

where \( \hat{\theta}_{\text{MAP}} \) is equivalent to the posterior mode. The objective function

\[
\log f_{X|\theta}(x \mid \theta) + \log \pi(\theta)
\]

is also called a penalized log-likelihood function (Klugman, 2008).

### 2.5 Optimizations

From the previous section, a few models for the tail distribution are introduced. The estimation of model parameters based on data necessarily involves optimization problems. Specifically, the MLE (or the MAP) involves the maximization of the log-likelihood (or the penalized log-likelihood) function.

Here, we consider the convex optimization, a special class of mathematical optimization problems, which includes the least-squares and the linear programming problems. It is well known that these two optimization problems arise in a variety of applications and several efficient numerical solutions are available.
2.5.1 Quadratic Approximation

Taylor’s formula provides an approximation of a function in the neighbourhood of a point by a polynomial:

\[ f(x) = f(x_0) + f'(x_0)(x - x_0) + \ldots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + r_n(x_0; x), \]

where \( r_n(x_0; x) \) is the remainder term. Therefore, the quadratic approximation of a function \( f(x) \) is given by

\[ f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2. \]

2.5.2 Convex Optimization

A convex optimization problem (with affine constraints) is of the form

\[
\begin{align*}
\text{minimize} & \quad f_0(x) \\
\text{subject to} & \quad f_i(x) \leq 0, i = 1, \ldots, m \\
& \quad a_i^T x = b_i, i = 1, \ldots, p
\end{align*}
\]

where \( f_0, \ldots, f_m \) are convex functions; see Boyd (2004) for more details about convex optimization.

The convex problem has three additional requirements. Firstly, the objective function must be convex. Secondly, the inequality constraint functions must be convex. Finally, the equality constraint functions \( h_i(x) = a_i^T x - b_i \) must be affine.
In particular when the objective function is (convex) quadratic and the constraint functions are affine, the convex optimization problem is called a quadratic problem.

2.5.3 Newton Algorithm

In order to solve a quadratic optimization problem, we introduce Newton’s method which is a numerical, iterative method for finding successively better approximations to the roots (or zeroes) of a real-valued function. The algorithm is summarized as follows:

Given a starting point \( x \in \text{dom}(f) \) and tolerance \( \epsilon > 0 \), repeat the following steps:

1. Compute the Newton step and decrement:
   
   \[
   \Delta x_{nt} := -\nabla^2 f(x)^{-1} \nabla f(x);
   \]
   
   \[
   \lambda^2 := \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x).
   \]

2. Stopping criterion: quit if \( \lambda^2/2 \leq \epsilon \).

3. Line search: Choose step size \( t \) by backtracking line search.

4. Update: \( x := x + t \Delta x_{nt} \).
Chapter 3

Methodology

In this chapter, we present our approaches on predictive modelling of the extreme values. For the majority data, a robust regression method is used to fit the model. With the residuals from the robust regression, the POT method is employed to select the tails observations (the extreme events). To estimate the tail distributions, the MAP estimation with conjugate priors is used to smooth the MLEs. With regard to each of the two tail, this MAP problem leads to two problems for tail parameters (the tail decay rate and the tail quantile level) which will be solved by Newton’s iteration method.

Figure 3.1 is a flowchart that describes our statistical modelling process:
3.1 Problem Formulation

Consider a dataset with multiple time periods and multiple individuals in a population. We assume that the data distribution features heavy-tailedness on both left and right tails.

The collection $D$ describes the data:

$$D = \{([x_{tia}, y_{tia}]_{a=1}^{M_{ii}})_{i=1}^{N} \}_{t=1}^{T},$$

(3.3)

where for each time period $t$ and individual $i$, the sequential pairs
\{(x_{ti1}, y_{ti1}), \ldots, (x_{tiM_i}, y_{tiM_i})\} \text{ is an i.i.d. sample of an underlying probability distribution. Here } y_{ti} \text{ is the response variable and } x_{ti} \text{ is a vector of covariates.}

In the following, we describe the models and methods to be applied to the data separated into several subsets in terms of different time periods and different individuals. From the expression (3.3), there are } T \times N \text{ subsets and } M_{ti} \text{ observations in each subset.}

As mentioned before, in order to capture the key characteristics of the data distribution, we use a mixture model where a robust regression method is used to fit the main body, while the POT method is employed to model the tails.

Specifically, the mixture model contains a normal distribution which explains the main body, and two tail distributions which are fitted to the left and right tails, respectively:

\[ y_{ti} = x_{ti}^T \beta_{ti} + z_{ti}^{(3)} v_{ti}^{(n)} + z_{ti}^{(1)} v_{ti}^{(e+)} + z_{ti}^{(2)} v_{ti}^{(e-)}, \quad (3.4) \]

where

\[ v_{ti}^{(n)} \sim N(0, \sigma_{ti}^2), \]
\[ v_{ti}^{(e+)} \sim p(\theta_{ti}^+ | x_{ti}), \]
\[ v_{ti}^{(e-)} \sim p(\theta_{ti}^- | x_{ti}), \]

\[ z_{ti}^{(k)} \sim \text{Trinomial}(q_{ti}^{(k)}), k = 1, 2, 3. \]

Marginally } z_{ti}^{(1)} \text{ and } z_{ti}^{(2)} \text{ follow Bernoulli distributions with success probabilities
\(q_{ti}^{(1)}\) and \(q_{ti}^{(2)}\), respectively:

\[
\begin{align*}
z_{ti}^{(1)} & \sim \text{Bernoulli}(q_{ti}^{(1)}), \\
z_{ti}^{(2)} & \sim \text{Bernoulli}(q_{ti}^{(2)}),
\end{align*}
\]

We will write the exceedance probabilities \(q_{ti}^{(1)} = q_{ti}^{+}\) and \(q_{ti}^{(2)} = q_{ti}^{-}\) throughout the rest of this thesis.

Note that our model generalizes the mixture model considered in Shenoy and Gorinevsky (2015) in which only the right tail is assumed heavy.

According to extreme value theory, \(p(\theta_{ti}^{\pm} | x_{ti})\) for tails can be modelled either by exponential distribution or Pareto distribution. As mentioned previously, the logarithm of Pareto random variable has an exponentially distribution, and thus, the same analysis can be applied with a log-transformation to the original dataset. The following methodology will consider these two possibilities in a unified fashion.

### 3.2 Robust Regression

In the presence of extreme values, a robust estimation method should be used for the estimation of the body. While there are many robust methods available in the literature (e.g. Maronna, 2006), we follow the practical approach used in Shenoy and Gorinevsky (2014, 2015) for a robust estimation of the body part and specifically, the model to be estimated is \(y_{tia} \sim N(x_{tia}^{T} \beta_{tia}, \sigma_{tia}^{2})\). We apply the maximum likelihood
estimation (MLE) iteratively to get the parameter estimates:

\[
\hat{\beta}_{ti} = \arg \min_{\beta_{ti}} \| y_{C_{ti}} - X_{C_{ti}} \beta_{ti} \|_2^2
\]

\[
= (X_{C_{ti}}^T X_{C_{ti}})^{-1} X_{C_{ti}}^T y_{C_{ti}},
\]

\[
\hat{\sigma}_{ti} = \text{card}(C_{ti})^{-1/2} \| y_{C_{ti}} - X_{C_{ti}} \hat{\beta}_{ti} \|_2,
\]

where \( X_{C_{ti}} \) is the design matrix whose \( a \)th row is \( x_{tia}^T \) and \( \text{card}(\cdot) \) is the cardinality (size) of the data set. \( C_{ti} \) is the set of indexes of the points that belong to the distribution body for time period \( t \) and population \( i \), for a threshold parameter \( A_{ti} \) defined as

\[
C_{ti} = \{ a \mid A_{ti} > |v_{tia}| \},
\]

where the residuals (Weisberg, 2015)

\[
v_{tia} = y_{tia} - \hat{y}_{tia},
\]

\[
\hat{y}_{tia} = x_{tia}^T \hat{\beta}_{ti}.
\]

As a start, we set \( C_{ti} = D \) (the whole dataset), which is equivalent to \( A_{ti} = \infty \). For the threshold \( A_{ti} \), we take a practical approach such as \( A_{ti} = 3\hat{\sigma}_{ti} \). With \( A_{ti} \), do the iterations until \( C_{ti} \) converges.

The converged \( C_{ti} \) includes all the data with small (we define "small" as smaller than \( 3\hat{\sigma} \)) residuals. The resulting robust estimates \( \hat{\beta}_{ti} \) and \( \hat{\sigma}_{ti} \) are used in the next step to get the residuals and to deal with the tail distributions.
3.3 Tail Models

With the parameters estimated from the robust regression step, we obtain residuals for data points in each subset.

For the modelling of tail distributions, we first choose the tail data points by the peaks-over-threshold method. Then the maximum likelihood estimation method is applied to fit the parameters in either exponential or Pareto distribution.

3.3.1 POT Method

In practice, the empirical quantiles are often used to determine the threshold. In Chapter 4, we use a method based on mean residual plot to determine the threshold in an objective fashion. This method is based on the fact that, for a large enough $u$, the mean excess $E(X - u \mid X > u)$ is linear function of $u$ (Cole, 2001).

If $Y$ has a generalized Pareto distribution with parameters $\sigma$ and $\xi$, then

$$E(Y) = \frac{\sigma}{1 - \xi},$$

provided $\xi < 1$. Based on Theorem 2, suppose the generalized Pareto distribution is valid as a model for the excesses of a threshold $u_0$ generated by a series $X_1, \ldots, X_n$, of which an arbitrary term is denoted by $X$, then it should equally be valid for all thresholds $u > u_0$, subject to the appropriate change of scale parameter to $\sigma_u$. Hence,
for $u > u_0$, 

$$E(X - u \mid X > u) = \frac{\sigma_u}{1 - \xi} = \frac{\sigma_{u_0} + \xi u}{1 - \xi} \quad (3.5)$$

According to (3.5), the sample mean of the threshold excesses of $u$ are expected to change linearly with $u$, at levels of $u$ for which the generalized Pareto model is appropriate.

Figure 3.2: Mean residual plot for stock return data

Figure 3.2 shows the mean residual plot for stock AAP in 2011-2012 subset. The graph appears to curve from $u = 0$ to $u \approx 0.5$, beyond which is approximately linear until $u \approx 1$, whereupon it decays sharply. It is tempting to conclude that there is no
stability until \( u = 0.5 \), after which there is approximately linearity. See more details in Chapter 4.

Next, we deal with positive residuals \( v_{tia}^+ \) and negative residuals \( v_{tia}^- \) separately to model the two extreme tails.

Define exceedances \( e_{ti}^\pm \) for time period \( t \) and population \( i \) as

\[
e^+_ti = \{ v_{tia}^+ - \Omega^+_ti \}_{a=1}^M,
\]
\[
e^-ti = \{ v_{tia}^- - \Omega^-ti \}_{a=1}^M,
\]

where \( \Omega^\pm ti \) are two thresholds. For the initial choice of the threshold \( \Omega^\pm ti \) or more clearly, considering the quantile level \( q^\pm ti \), the standard practice is to adopt it as low as possible subject to select enough data points for estimation.

We also define \( T^\pm ti \) to be the set of indexes of the data points in the tails. According to the data structure of our problem, the POT exceedances are the set of positive exceedances for \( T^+_ti \) and negative exceedances for \( T^-ti \):

\[
T^+_ti = \{ a \mid e^+_ti \geq 0 \},
\]
\[
T^-ti = \{ a \mid e^-ti \leq 0 \}.
\]

The probability of exceedances are denoted by

\[
q^+_ti = P(e^+_ti \geq 0),
\]
\[
q^-ti = P(e^-ti \leq 0).
\]

We are interested in estimating the risk (probability) of extreme residuals \( v_{tia}^\pm \)
which correspond to large exceedances \( e_{tia}^+ \). By Bayes’ theorem, the risks are expressed as:

\[
R_{ti}^+(u^+) = P(e_{ti}^+ \geq u^+)
\]
\[
= P(e_{ti}^+ \geq u^+ | e_{ti}^+ \geq 0) P(e_{ti}^+ \geq 0),
\]
\[
R_{ti}^-(u^-) = P(e_{ti}^- \leq u^-)
\]
\[
= P(e_{ti}^- \leq u^- | e_{ti}^- \leq 0) P(e_{ti}^- \leq 0). \tag{3.6}
\]

As mentioned before, the conditional tail will be approximated by Pareto distribution or exponential distribution, while Pareto tail can be switched into exponential tail by means of logarithm transformation of the residuals. Therefore, we simply consider the first multiplier in (3.6) follows the exponential distribution.

\[
e_{ti}^+ | (e_{ti}^+ \geq 0, \theta_{ti}^+) \sim \text{Exp} (\theta_{ti}^+),
\]
\[
e_{ti}^- | (e_{ti}^- \leq 0, \theta_{ti}^-) \sim \text{Exp} (\theta_{ti}^-). \tag{3.7}
\]

From (3.6), we consider that each data point is an independent realization of the mixture model, i.e., with probability \( q_{ti}^+ \), the point belongs to the right tail (or \( q_{ti}^- \) to the left tail). More precisely, the number of the right tail (or left tail) points \( n_{ti}^+ \) (or \( n_{ti}^- \)) follows the Binomial distribution:

\[
n_{ti}^+ | q_{ti}^+ \sim \text{Bin}(M_{ti}, q_{ti}^+),
\]
\[
n_{ti}^- | q_{ti}^- \sim \text{Bin}(M_{ti}, q_{ti}^-). \tag{3.8}
\]
3.3.2 MLEs

To find the MLEs for the parameters in (3.7) and (3.8), for each subset \((t, i)\), the likelihood functions are:

\[
L(\theta^+_{ti} | e^+_ti) = (\theta^+_{ti})^{n^+_{ti}} \exp(-\sum_a n^+_{ti} e_{tica} \theta^+_{ti}) = (\theta^+_{ti})^{n^+_{ti}} \exp(-n^+_{ti} \bar{e}^+_ti \theta^+_{ti}),
\]

\[
L(\theta^-_{ti} | e^-_ti) = (\theta^-_{ti})^{n^-_{ti}} \exp(+\sum_a n^-_{ti} e_{tica} \theta^-_{ti}) = (\theta^-_{ti})^{n^-_{ti}} \exp(+n^-_{ti} \bar{e}^-_ti \theta^-_{ti}),
\]

\[
L(q^+_{ti}) = (q^+_{ti})^{n^+_{ti}} (1-q^+_{ti})^{M_{ti}-n^+_{ti}},
\]

\[
L(q^-_{ti}) = (q^-_{ti})^{n^-_{ti}} (1-q^-_{ti})^{M_{ti}-n^-_{ti}},
\]

where we denote:

\[
\bar{e}^+_ti = \text{mean}_{a \in T^+_ti} \{e_{tica}\}, \quad \bar{e}^-_ti = \text{mean}_{a \in T^-_ti} \{e_{tica}\},
\]

\[
n^+_{ti} = \text{card}(T^+_ti), \quad n^-_{ti} = \text{card}(T^-_ti).
\]

Regarding \(\theta^+_{ti}\), by maximizing the log-likelihood function, we obtain the MLE as follows:

\[
L(\theta^+_{ti} | e^+_ti) = (\theta^+_{ti})^{n^+_{ti}} \exp(-n^+_{ti} \bar{e}^+_ti \theta^+_{ti}),
\]

\[
l(\theta^+_{ti} | e^+_ti) = \log[(\theta^+_{ti})^{n^+_{ti}} \exp(-n^+_{ti} \bar{e}^+_ti \theta^+_{ti})],
\]

\[
= n^+_{ti} \log(\theta^+_{ti}) - n^+_{ti} \bar{e}^+_ti \theta^+_{ti},
\]

\[
\frac{\partial l}{\partial \theta^+_{ti}} = \frac{n^+_{ti}}{\theta^+_{ti}} - n^+_{ti} \bar{e}^+_ti = 0,
\]

\[
\Rightarrow \hat{\theta}^+_{ti} = 1/\bar{e}^+_ti.
\]

Similarly, maximizing the log-likelihoods for the four tail parameter distributions
yields (repeat for each subset \((t, i)\)):

\[
\begin{align*}
\text{maximize} & \quad \sum_{t=1}^{T} \sum_{i=1}^{N} n_{ti}^+ (\log\theta_{ti}^+ - \theta_{ti}^+ \bar{e}_{ti}^+), \\
\text{maximize} & \quad \sum_{t=1}^{T} \sum_{i=1}^{N} n_{ti}^- (\log\theta_{ti}^- + \theta_{ti}^- \bar{e}_{ti}^-), \\
\text{maximize} & \quad \sum_{t=1}^{T} \sum_{i=1}^{N} [n_{ti}^+ \log q_{ti}^+ + (M_{ti} - n_{ti}^+)\log(1 - q_{ti}^+)], \\
\text{maximize} & \quad \sum_{t=1}^{T} \sum_{i=1}^{N} [n_{ti}^- \log q_{ti}^- + (M_{ti} - n_{ti}^-)\log(1 - q_{ti}^-)].
\end{align*}
\] (3.9)

The estimates can be obtained by differentiating each term with respect to \(\theta_{ti}^+\), \(\theta_{ti}^-\), \(q_{ti}^+\) or \(q_{ti}^-\) and finding the zeros. The resulting estimates are:

**Tail decay rate \(\theta\)**

\[
\begin{align*}
\hat{\theta}_{ti}^+ & = \frac{1}{\bar{e}_{ti}^+}, \\
\hat{\theta}_{ti}^- & = -\frac{1}{\bar{e}_{ti}^-}.
\end{align*}
\] (3.10)

**Tail quantile level \(q\)**

\[
\begin{align*}
\hat{q}_{ti}^+ & = \frac{n_{ti}^+}{M_{ti}}, \\
\hat{q}_{ti}^- & = \frac{n_{ti}^-}{M_{ti}}.
\end{align*}
\] (3.11)

### 3.4 Filter Tuning Process

This section describes the method of Bayesian maximum a posterior estimation (MAP) on the tail parts of the dataset.
Firstly, we make the following non-linear variable changes (see more details in Martin, 2014). Define

\[
\begin{align*}
    r_{ti}^+ &= \psi(\theta_{ti}^+) = \log(\theta_{ti}^+), \\
    r_{ti}^- &= \psi(\theta_{ti}^-) = \log(\theta_{ti}^-), \\
    w_{ti}^+ &= \phi(q_{ti}^+) = \log(-\log(q_{ti}^+)), \\
    w_{ti}^- &= \phi(q_{ti}^-) = \log(-\log(q_{ti}^-)),
\end{align*}
\]

(3.12)

where \(\psi(\cdot) = \log(\cdot)\) or \(\phi(\cdot) = \log(-\log(\cdot))\) are variable changes which make the constrained parameters \(\theta_{ti}^\pm \in (0, +\infty)\) and \(q_{ti}^\pm \in (0, 1)\) to be unconstrained.

Then optimization problems (3.9) can be compactly rewritten as

\[
\begin{align*}
    \text{maximize} & \quad \sum_{t=1}^{T} \sum_{i=1}^{N} n_{ti}^+ \Psi(r_{ti}^+; \bar{e}_{ti}^+), \\
    \text{maximize} & \quad \sum_{t=1}^{T} \sum_{i=1}^{N} n_{ti}^- \Psi(r_{ti}^-; -\bar{e}_{ti}^-), \\
    \text{maximize} & \quad \sum_{t=1}^{T} \sum_{i=1}^{N} n_{ti}^+ \Phi(w_{ti}^+; M_{ti}/n_{ti}^+ - 1), \\
    \text{maximize} & \quad \sum_{t=1}^{T} \sum_{i=1}^{N} n_{ti}^- \Phi(w_{ti}^-; M_{ti}/n_{ti}^- - 1),
\end{align*}
\]

(3.13)

where

\[
\begin{align*}
    \Psi(x; \gamma) &= x - \gamma e^x, \\
    \Phi(x; \gamma) &= -e^x + \gamma \log(1 - \exp(-e^x)).
\end{align*}
\]

(3.14)
3.4.1 The Priors

In order to construct the priors for the MAP estimates, we will make assumptions both to the prior structures and to the prior distributions.

Since the points from the right and the left tails are assumed independent to each other, $\theta^+$ and $\theta^-$, $q^+$ and $q^-$ are supposed to be independent to each other, respectively. With the following prior structures, we are able to construct optimization problems for the four parameters for the MAP estimation.

Based on some empirical evidences, we consider the following prior structure which relates the tail rate parameter $\theta$ and tail quantile parameter $q$ at two indexes sets $\{t_k, i_k\}$ and $\{t'_k, i'_k\}$ for the left and the right tail, respectively. Firstly, the parameters for a particular individual $i$ in two adjacent sequential time period $t$ and $t'$ are close. Secondly, the parameters in a particular time period $t$ for an individual $i$ and the average parameter of all individuals $i'$ are close (see Chapter 4 for more details).

Based on this prior belief, we give a new index named $k$ for the pair $\{t_k, i_k\}$ and $\{t'_k, i'_k\}$. Table 3.1 summaries the indexes and their ranges.
Here $i_0$ stands for the average parameter of all individuals in a particular time period. Therefore, for this new index $k$, $k = 1, 2, \ldots, N_p$, the total number is $N_p = N \cdot (T - 1) + T \cdot N = 2TN - N$.

For notational simplicity, we write $\theta^\pm \equiv \{\theta^\pm_{ti} \}$ and $q^\pm \equiv \{q^\pm_{ti} \}$, and consider that two sets of $N_p$ priors $\{(\Delta \psi)_k(\theta^+), (\Delta \phi)_k(q^+)\}_{k=1}^{N_p}$ and $\{(\Delta \psi)_k(\theta^-), (\Delta \phi)_k(q^-)\}_{k=1}^{N_p}$ have the form:

\[
(\Delta r)_k^+ = r_{t_k,i_k}^+ - r_{t_k,i_k'}^+ = (\Delta \psi)_k(\theta^+) = \psi(\theta_{t_k,i_k}^+) - \psi(\theta_{t_k,i_k'}^+) \sim \psi(\chi_{k}^+),
\]

\[
(\Delta w)_k^+ = w_{t_k,i_k}^+ - w_{t_k,i_k'}^+ = (\Delta \phi)_k(q^+) = \phi(q_{t_k,i_k}^+) - \phi(q_{t_k,i_k'}^+) \sim \phi(\xi_{k}^+),
\]

\[
(\Delta r)_k^- = r_{t_k,i_k}^- - r_{t_k,i_k'}^- = (\Delta \psi)_k(\theta^-) = \psi(\theta_{t_k,i_k}^-) - \psi(\theta_{t_k,i_k'}^-) \sim \psi(\chi_{k}^-),
\]

\[
(\Delta w)_k^- = w_{t_k,i_k}^- - w_{t_k,i_k'}^- = (\Delta \phi)_k(q^-) = \phi(q_{t_k,i_k}^-) - \phi(q_{t_k,i_k'}^-) \sim \phi(\xi_{k}^-),
\]

where $(\Delta \cdot)_k(\cdot)$ stands for the random variables in the prior distribution for the parameters $\theta^\pm$ and $q^\pm$ with variable change $\psi(\cdot)$ or $\phi(\cdot)$. $\psi(\cdot)$ and $\phi(\cdot)$ are given by (3.12).

About the prior distributions for $\chi_{k}^\pm$ and $\xi_{k}^\pm$, we have the following setting.

For an algebraic convenience, we consider conjugate priors which make the posterior distributions have the same algebraic forms as the prior distributions (generally
with different parameter values). Specifically, we postulate

\[ \chi_k^+ \sim \text{Gamma}(\alpha_k^+ + 1, \alpha_k^+), \]
\[ \chi_k^- \sim \text{Gamma}(\alpha_k^- + 1, \alpha_k^-), \]
\[ \zeta_k^+ \sim \text{Beta}(\eta_k^+ + 1, \eta_k^+ (e - 1) + 1), \]
\[ \zeta_k^- \sim \text{Beta}(\eta_k^- + 1, \eta_k^- (e - 1) + 1), \]

(3.16)

where \( \alpha_k^\pm \) and \( \eta_k^\pm \) are tuning parameters which control the degree of smoothness for the pairs \( \{t_k, i_k\} \) and \( \{t'_k, i'_k\} \) in Table 3.1.

The following gives some details about the parameter settings in (3.15). Considering \( \chi_k^+ \sim \text{Gamma}(\alpha, \beta) \) (or \( \zeta_k^+ \sim \text{Beta}(\alpha, \beta) \) (see Klugman, 2008 for more details), we make the mode of \( \psi(\chi_k^+) \) (or \( \phi(\zeta_k^+) \)) equals to 0, equivalent to the mode of \( \chi_k^+ \) equals to 1 (or \( \phi(\zeta_k^+) = 1/e \)). Note that for Gamma(\( \alpha, \beta \)), the mode is \((\alpha - 1)/\beta \) (or for Beta(\( \alpha, \beta \)), the mode is \((\alpha - 1)/(\alpha + \beta - 2) \)). Therefore, by setting \((\alpha - 1)/\beta = 1 \) (or \((\alpha - 1)/(\alpha + \beta - 2) = 1/e \)), we have \( \alpha = \alpha_k^+ + 1, \) \( \beta = \alpha_k^+ \) (or \( \alpha = \eta_k^+ + 1, \beta = \eta_k^+ (e - 1) + 1 \)). As a result, the mode of \( \psi(\chi_k^-) \) (or \( \phi(\zeta_k^-) \)) equals to zero. The intuition of these parameter settings is that, in the absence of POT exceedance data (3.7) and (3.8), the tail parameters \( \theta^\pm \) and \( q^\pm \) for time period \( t_k \) and individual \( i_k \) are exactly equal to those for the time period \( i'_k \) and individual \( i'_k \).

For the following, the derivation of the log-prior density for each parameter is given. Regarding the log-prior for \( \theta^+ \), let \( y \) be a random variable from \( \text{Gamma}(\alpha_k^+ + \)
1, θ_k). Then by (3.15) and (3.16), we have

\[(\Delta r)_k^+ = r_{t_k,i_k}^+ - r_{t_k,i_k}'^+ = (\Delta \psi)_k(\theta^+) = \psi(\theta_{t_k,i_k}^+) - \psi(\theta_{t_k,i_k}') - \psi(\chi_k^+),\]

\[\chi_k^+ \sim \text{Gamma}(\alpha_k^+ + 1, \alpha_k^+),\]

\[(\Delta r)_k^+ = \log(y).\]

Thus, the prior distribution is

\[\frac{\alpha_k^+}{\Gamma(\alpha_k^+ + 1)}y^{\alpha_k^+}e^{-\alpha_k^+y},\]

and the log prior is written as (up to constant)

\[c + \alpha_k^+ \log(y) - \alpha_k^+y = c + \alpha_k^+(\Delta r)_k^+ - \alpha_k^+e^{(\Delta r)_k^+} = c + \alpha_k^+[(\Delta r)_k^+ - e^{(\Delta r)_k^+}] = c + \alpha_k^+\Psi((\Delta \psi)_k(\theta^+); 1).\]

Similarly, regarding the log-prior for \(q_k^+\), let \(y\) be a random variable from \(\text{Beta}(\eta_k^+ + 1, \eta_k^+).\) Then by (3.15) and (3.16), we have:

\[(\Delta w)_k^+ = w_{t_k,i_k}^+ - w_{t_k,i_k}'^+ = (\Delta \phi)_k(q^+) = \phi(q_{t_k,i_k}^+) - \phi(q_{t_k,i_k}') - \phi(\xi_k^+),\]

\[\xi_k^+ \sim \text{Beta}(\eta_k^+ + 1, \eta_k^+(e - 1) + 1),\]

\[(\Delta w)_k^+ = \log(-\log(y)).\]

Thus, the prior distribution is

\[\frac{y^{\eta_k^+}(1 - y)^{\eta_k^+(e - 1)}}{B(\eta_k^+ + 1, \eta_k^+(e - 1) + 1)}\]
and the log-prior is written as (up to constant)

\[
\eta_k^+ \log(y) - \eta_k^+(e - 1)\log(1 - y) + c
\]

\[
= \eta_k^+(-e^{(\Delta w)^+_k}) + \eta_k^+(e - 1)\log(1 - \exp(-e^{(\Delta w)^+_k})) + c
\]

\[
= \eta_k^+ \Phi((\Delta \phi)_k(q^+); e - 1) + c.
\]

For summary, the log-prior for each parameter is:

\[
\alpha_k^+ \Psi((\Delta \psi)_k(\theta^+); 1);
\]

\[
\alpha_k^- \Psi((\Delta \psi)_k(\theta^-); 1);
\]

\[
\eta_k^+ \Phi((\Delta \phi)_k(q^+); e - 1);
\]

\[
\eta_k^- \Phi((\Delta \phi)_k(q^-); e - 1);
\]

(3.17)

### 3.4.2 MAP Problems

Recalling Section 2.4.2, the log-likelihood function (3.13) plus the log-priors derived above in (3.17) become the objective functions for MAP estimators:

\[
\text{maximize} \sum_{t=1}^{T} \sum_{i=1}^{N} n_{ti}^+ \Psi(r_{ti}^+; \bar{e}_{ti}^+) + \sum_{k=1}^{N_p} \alpha_k^+ \Psi((\Delta r)_k^+; 1),
\]

\[
\text{maximize} \sum_{t=1}^{T} \sum_{i=1}^{N} n_{ti}^- \Psi(r_{ti}^-; -\bar{e}_{ti}^-) + \sum_{k=1}^{N_p} \alpha_k^- \Psi((\Delta r)_k^-; 1),
\]

\[
\text{maximize} \sum_{t=1}^{T} \sum_{i=1}^{N} n_{ti}^+ \Phi(w_{ti}^+; M_{ti}/n_{ti}^+ - 1) + \sum_{k=1}^{N_p} \eta_k^+ \Phi((\Delta w)_k^+; e - 1),
\]

\[
\text{maximize} \sum_{t=1}^{T} \sum_{i=1}^{N} n_{ti}^- \Phi(w_{ti}^-; M_{ti}/n_{ti}^- - 1) + \sum_{k=1}^{N_p} \eta_k^- \Phi((\Delta w)_k^-; e - 1).
\]

(3.18)
where \((\Delta r)^\pm_k = r^\pm_{t_k i_k} - r^\pm_{t'_{k} i'_{k}}\) and \((\Delta \omega)^\pm_k = \omega^\pm_{t_k i_k} - \omega^\pm_{t'_{k} i'_{k}}\). The functions \(\psi(\cdot), \phi(\cdot), \Psi(\cdot)\) and \(\Phi(\cdot)\) are defined in (3.12) and (3.14).

These optimizations are non-convex problems. For computational efficiency, we consider the following steps. Note that all the objective functions in (3.18) have the following form:

\[
L(\xi) = \sum_{t=1}^{T} \sum_{i=1}^{N} n_{ti} f(\xi_{ti}; \rho_{ti}) + \sum_{k=1}^{N_p} \lambda_k f((\Delta \xi_k); \gamma),
\]  
(3.19)

where \(\xi_{ti}\) are the decision variables and \((\Delta \xi_k) = \xi_{t_k i_k} - \xi_{t'_{k} i'_{k}}\). The parameters \(\rho_{ti}, \lambda_k, \gamma\), and the function \(f(\cdot)\) can be related back to problems (3.18) (see details later in (3.24) and (3.25)).

By writing \(\xi_{ti} = \xi_{ti}^* + \epsilon_{ti}\), Taylor’s expansion (Rudin, 1976) at \(\xi_{ti}^*\) for the likelihood function \(L(\xi)\) above gives:

\[
L(\xi) = \sum_{t=1}^{T} \sum_{i=1}^{N} n_{ti} \left( f(\xi_{ti}^*; \rho_{ti}) + f'(\xi_{ti}^*; \rho_{ti}) \epsilon_{ti} + f''(\xi_{ti}^*; \rho_{ti}) \frac{\epsilon_{ti}^2}{2} + \ldots \right) \\
+ \sum_{k=1}^{N_p} \lambda_k \left( f((\Delta \xi_k^*); \gamma) + f'(\Delta \xi_k^*; \gamma) (\Delta \epsilon_k) + f''(\Delta \xi_k^*; \gamma) \frac{(\Delta \epsilon_k)^2}{2} + \ldots \right) \\
\approx \sum_{t=1}^{T} \sum_{i=1}^{N} n_{ti} \left( f'(\xi_{ti}^*; \rho_{ti}) \epsilon_{ti} + f''(\xi_{ti}^*; \rho_{ti}) \frac{\epsilon_{ti}^2}{2} \right) \\
+ \sum_{k=1}^{N_p} \lambda_k \left( f'(\Delta \xi_k^*; \gamma) (\Delta \epsilon_k) + f''(\Delta \xi_k^*; \gamma) \frac{(\Delta \epsilon_k)^2}{2} \right) + \text{constant} \\
\approx \sum_{t=1}^{T} \sum_{i=1}^{N} n_{t,i} f''(\xi_{ti}^*; \rho_{ti}) (\frac{f'(\xi_{ti}^*; \rho_{ti})}{f''(\xi_{ti}^*; \rho_{ti})} \cdot \epsilon_{ti} + \frac{1}{2} \cdot \epsilon_{ti}^2) \\
+ \sum_{k=1}^{N_p} \lambda_k f''(\Delta \xi_k^*; \gamma) \cdot (\frac{f'(\Delta \xi_k^*; \gamma)}{f''(\Delta \xi_k^*; \gamma)} \cdot (\Delta \epsilon_k) + \frac{1}{2} (\Delta \epsilon_k)^2) + \text{constant} \\
= \left( \frac{1}{2} F_0 \epsilon - d_0 \right)^T B_0 F_0 \epsilon + \left( -\Delta \epsilon - d_1 \right)^T B_1 \Delta \epsilon + \text{constant},
\]  
(3.20)
where \( \Delta \xi_k^* \) and \( \Delta \epsilon_k \) are defined as \( (\Delta \xi_k^*) = \xi_{tki_k}^* - \xi_{tk'i_k}^* \) and \( (\Delta \epsilon_k) = \epsilon_{ti_ki_k} - \epsilon_{ti_k'i_k} \). In (3.20), \( F_0 \) is the identity matrix. We assign

\[
\Delta \epsilon = F_1 \epsilon,
\]

where \( F_1 \) is the sparse difference matrix that maps the vectorized matrix \( \epsilon \) into the vector \( \Delta \epsilon \) in accordance with the prior structure. Specifically, \( F_1 \) is defined as

\[
F_1 = \left[ \begin{array}{c} I_N \otimes DT \\ I_{T \times N} - \frac{1}{N} I_T \otimes E_{N \times N} \end{array} \right]
\]

where \( I \) is the identity matrix, \( D \) is a bi-diagonal first difference matrix with -1 on the main diagonal (except the last, zero, entry) and 1 above the main diagonal, \( E \) is a square matrix of 1.

Other matrices in (3.20) are defined as

\[
\epsilon = \text{vec}(\epsilon_{ti})_{t=1,i=1}^{T,N},
\]

\[
B_0 = \text{diag}(\text{vec}(n_{ti} f''(\xi_{ti}^*; \rho_{ti}))_{t=1,i=1}^{T,N}),
\]

\[
B_1 = \text{diag}(\text{vec}(\lambda_{ti} f''((\Delta \xi_k^*); \gamma))_{k=1}^{N_p}),
\]

\[
d_0 = \text{vec} \left\{ -\frac{f'(\xi_{ti}^*; \rho_{ti})}{f''(\xi_{ti}^*; \rho_{ti})} \right\}_{t=1,i=1}^{T,N},
\]

\[
d_1 = \text{vec} \left\{ -\frac{f'((\Delta \xi_k^*); \gamma)}{f''((\Delta \xi_k^*); \gamma)} \right\}_{k=1}^{N_p}.
\]

By multiplying a minus sign to function \( f(\cdot) \), the four maximization problems in
become the minimization problems of the following forms:

\[
\text{minimize } \left( \frac{1}{2} F_0 \epsilon - d_0 \right)^T B_0 \left( F_0 \epsilon \right) + \left( \frac{1}{2} F_1 \epsilon - d_1 \right)^T B_1 \left( F_1 \epsilon \right),
\]

\[
\iff \frac{\partial}{\partial \epsilon} \left\{ \left( \frac{1}{2} F_0 \epsilon - d_0 \right)^T B_0 \left( F_0 \epsilon \right) + \left( \frac{1}{2} F_1 \epsilon - d_1 \right)^T B_1 \left( F_1 \epsilon \right) \right\} = 0,
\]

\[
\iff \frac{\partial}{\partial \epsilon} \left\{ (F_0 \epsilon - d_0)^T B_0 (F_0 \epsilon - d_0) + (F_1 \epsilon - d_1)^T B_1 (F_1 \epsilon - d_1) \right\} = 0,
\]

\[
\iff \text{minimize } (F_0 \epsilon - d_0)^T B_0 (F_0 \epsilon - d_0) + (F_1 \epsilon - d_1)^T B_1 (F_1 \epsilon - d_1). \tag{3.23}
\]

Now we consider the parameters \( \rho_{ti}, \lambda_k, \gamma \), and the function \( f(\cdot), f'(\cdot), f''(\cdot) \) in (3.19), respectively.

For \( \theta^\pm \),

\[
f(x; \nu) = -\Psi(x; \nu) = ve^x - x,
\]

\[
\rho_{ti} = \pm \hat{e}_{ti}^\pm, \quad \gamma = 1, \quad \lambda_k = \alpha_k^\pm,
\]

\[
\Psi'(x; \nu) = 1 - ve^x,
\]

\[
\Psi''(x; \nu) = -ve^x, \tag{3.24}
\]

For \( q^\pm \),

\[
f(x; \nu) = -\Phi(x; \nu) = e^x - \nu \log[1 - \exp(-e^x)],
\]

\[
\rho_{ti} = M_{ti}/n_{ti}^\pm - 1, \quad \gamma = e - 1, \quad \lambda_k = \eta_k^\pm,
\]

\[
\Phi'(x; \nu) = e^x \left[ \frac{\nu}{\exp(e^x) - 1} \right],
\]

\[
\Phi''(x; \nu) = e^x \left[ \nu \frac{\exp(e^x)(1 - e^x) - 1}{[\exp(e^x) - 1]^2} - 1 \right], \tag{3.25}
\]
Note that the problem (3.23) is a unconstrained convex optimization problem which can be solved explicitly as:

\[
\epsilon = H^{-1} \sum_{m=0}^{1} F_m^T B_m d_m,
\]

\[
H = \sum_{m=0}^{1} F_m^T B_m F_m.
\] (3.26)

With (3.26) we can find the decision variables \( \xi \), as long as we have Taylor’s expansion center \( \xi^* \). However, the accurate value of \( \xi^* \) can not be found directly. In the following part, we will use MLEs solved in (3.10) and (3.11) as the expansion center and update \( \xi \) with iteration as \( \xi \leftarrow \xi^* + \epsilon \) where \( \epsilon \) is getting from (3.26).

### 3.4.3 Optimization Steps

First, for empty subsets (i.e., \( M_{ti} = 0 \)), the MLEs are \( \log(\theta) = \infty \) and \( \log(-\log(q)) = \infty \). In this case, we use the average values for the expansion centers.

For non-empty subsets, we take the following steps:

(i) Set \( \xi_{ti}^* \) using the MLEs \( \theta_{ti}^\pm \) or \( q_{ti}^\pm \) and \( (\Delta \xi_k^*) \) are computed based on the prior assumptions. Then \( \epsilon \) is solved from (3.26) and \( \Delta \epsilon = F_1 \epsilon \).

(ii) Update the estimates using the approximation solutions in (3.26)

\[
\xi_{ti}^{(i+1)} \leftarrow \xi_{ti}^{(i)} + \epsilon^{(i)}
\]

\[
(\Delta \xi_k)^{(i+1)} \leftarrow (\Delta \xi_k)^{(i)} + (\Delta \epsilon_k)^{(i)}.
\]

(iii) The iteration continues until convergence.
The tail decay rate $\theta_{ti}^\pm$ and the tail quantile level $q_{ti}^\pm$ are recovered by the relations:

$$\theta_{ti}^\pm = \exp(\xi_{ti}^\pm) \text{ and } q_{ti}^\pm = \exp(-\exp(\xi_{ti}^\pm)),$$

respectively.

### 3.5 Prediction Accuracy Measure

To measure the predictive performance of the fitted model, we consider the Value-at-Risk (VaR). Formally, the VaR (or quantile) $\pi_p$ at level $p$ is defined as:

$$\Pr(x < \pi_p) = p,$$

where $x$ is a random variable (see Chinhamu, 2007; Shiryaev, 1995 for more details). Specifically, we construct the probability distribution of $y_{tia}$ using the parameter estimates and simulations.

$$y_{tia} = x_{tia}^T \beta_{ti} + z_{(n)}_{ti} v_{(n)}_{ti} + z_{(1)}_{ti} v_{(e^+)}_{ti} + z_{(2)}_{ti} v_{(e^-)}_{ti}, \quad (3.27)$$

where

$$v_{(n)}_{ti} \sim N(0, \sigma^2),$$

$$v_{(e^\pm)}_{ti} \sim \text{Exp}(\theta_{ti}^\pm),$$

$$z_{(k)}_{ti} \sim \text{Trinomial}(q_{ti}^{(k)}), \quad k = 1, 2, 3 \quad (3.28)$$

Then, for a set of prescribed $p$ values, the VaRs are determined and the proportions of the observations in the testing dataset below the corresponding VaRs, i.e.,
the coverage probabilities, are computed. The closer the coverage probability to the corresponding $p$ is, the better the model performs.
In this chapter, we apply the proposed methodology to a stock return dataset. This dataset is retrieved and compiled from various sources including Yahoo Finance and S & P Indexes. In the filter tuning process, the selection of the prior structure is based on some empirical evidences.

4.1 Dataset

The dataset used in this thesis is composed of daily returns of 47 stocks spanning from 2008 to 2018 (see Appendix for more details). For financial time series, there is a rich structure of temporal dependence (Reiss, 2007; Chavez-Demoulin, 2014). Since the raw data points of observed stock prices or related index prices are strongly non-stationary, we transform the observations as follows.
Daily stock return:

\[ y_{tj} = \log \left( \frac{p_{tj}}{p_{tj-1}} \right), \]

where \( p_{tj} \) is the stock price at time \( t_j \).

Daily Index return for \( k \), where \( k = 1, 2, 3, 4 \):

\[ x_{k:tj} = \log \left( \frac{p_{k:tj}}{p_{k:tj-1}} \right), \]

where Index \( k=1 \) is S&P global index (common for all stocks); Index 2 is country index (common for stocks from the same country); Index 3 is stock sector index (common for stocks from the same industry sector); Index 4 is company size index (common for stocks from the same size level).

This transformed data series is close to stationary. For the consistency of models for multiple time periods, the stock return data series have been standardized in each subset. Descriptive average statistics for this dataset is given in Table 4.1. It shows that the distribution of stock returns is asymmetric and fat-tailed.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td>6</td>
<td>47</td>
<td>430</td>
<td>120,000</td>
<td>-0.19</td>
</tr>
</tbody>
</table>

Table 4.1: Summary statistics of daily return dataset

Note here that, though each subset size is mostly large, it is possible that for some
subset \(t, i\), \(M_{ti} = 0\), that is, there is no observation for the given individual \(i\) at the
given time period \(t\) (or \(n_{ti} = 0\), that no extreme value in this subset). The MAP filter
tuning process considered in this thesis is able to deal with this situation.

4.2 Robust Regression, Threshold and Tuning Parameter Se-
lection

As described in Section 3.2, we first estimate the main body of the mixture model
\(3.4\). The robust regression requires the selection of the threshold \(A_{ti}\) for each sub-
set. Here we use the so-called three-sigma rule which states that, for many reason-
abley bell-shaped unimodal distributions, almost all of the observations lies within
three standard deviations from the mean. This empirical rule is often used in prac-
tice to isolate the majority body.

Table 4.2 gives the averages of the robust estimates of \(\beta\) and \(\sigma\) in \(3.4\). This aver-
age result may suggest that the Global Index (of which the average of coefficients is
negative) might be dropped due to the possible multicollinearity issue with the other
three indexes. However the robust regression results are significantly different by
time and individual, and thus, we include all four covariates for the sake of practical
simplicity.
Table 4.2: Summary of robust regression

<table>
<thead>
<tr>
<th>Estimates</th>
<th>Intercept</th>
<th>Global Index</th>
<th>Country Index</th>
<th>Sector Index</th>
<th>Size Index</th>
<th>Sigma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>-0.02</td>
<td>-29.48</td>
<td>30.86</td>
<td>40.15</td>
<td>13.95</td>
<td>0.53</td>
</tr>
</tbody>
</table>

The tail models are then fitted on the extreme tail statistics (i.e., residuals from the robust regression) over appropriate thresholds, $\Omega_{ti}^{\pm}$, which are chosen by the mean residual plots as explained in Chapter 3.

In the MAP filter tuning process, we need to construct a set of priors for the posterior maximum optimization problems (see more details in Shenoy, 2014; 2015). From empirical evidences, it is reasonable to believe that the parameters for a single stock in two sequential time period are close, and the parameters for an individual stock and the average of all stocks at the same period are close.

The prior distribution assumptions given in Section 3.4.1 involve four tuning parameters: $\alpha_k^{\pm}$ and $\eta_k^{\pm}$ for Gamma and Beta distributions, respectively. The larger the tuning parameters are, the smoother the MAP estimators will be. We choose the tuning parameters which are large enough to make the parameter estimates smooth but not too large to make them flat.

Figures 4.1 illustrate the MLEs and smoothed MAP estimates of $\theta_{ti}^+$ for five stocks.
in each time periods. In the figures, the MLEs are pretty jaggy while the MAP estimators are well-smoothed under sufficiently large values of the tuning parameters $\alpha$ and $\eta$.

![Figure 4.1: MLE vs. MAP estimates of $\theta_{ii}^+$](image)

To choose the optimal tuning parameters, we conduct dozens of training trials with different parameter values on this stock dataset, and the ones which provide the best out-of-sample testing performance are selected. A rolling prediction (training on three sequential time periods) also supports our choices in the following:

$$
\alpha_k^\pm = \begin{cases} 
5 \times 10^5, & k = 1,\ldots,(T-1)N \\
5 \times 10^4, & k = (T-1)N + 1,\ldots(2T-1)N,
\end{cases}
$$

$$
\eta_k^\pm = \begin{cases} 
1000, & k = 1,\ldots,(T-1)N \\
100, & k = (T-1)N + 1,\ldots(2T-1)N.
\end{cases}
$$

With these tuning parameter values, the MLEs are well smoothed. In the following
we present the out-of-sample testing results by MLEs and MAP estimates, respectively.

### 4.3 Out-of-Sample Testing

We use the smoothed estimates obtained based on the data up to 2016 (i.e., $T = 1, \ldots, 5$) to perform out-of-sample predictions for 2017-2018 time period ($T = 6$). Table 4.3 gives the prediction accuracy measures (the coverage probabilities) under Exponential tail and Pareto tail, respectively.

<table>
<thead>
<tr>
<th>Pr</th>
<th>5 %</th>
<th>10 %</th>
<th>20 %</th>
<th>50 %</th>
<th>80 %</th>
<th>90 %</th>
<th>95 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE (Exp)</td>
<td>9.721 %</td>
<td>15.415 %</td>
<td>25.515 %</td>
<td>50.233 %</td>
<td>74.563 %</td>
<td>84.296 %</td>
<td>89.873 %</td>
</tr>
<tr>
<td>MLE (Pareto)</td>
<td>9.739 %</td>
<td>15.410 %</td>
<td>25.487 %</td>
<td>50.185 %</td>
<td>74.576 %</td>
<td>84.296 %</td>
<td>89.873 %</td>
</tr>
<tr>
<td>MAP (Exp)</td>
<td>8.368 %</td>
<td>13.960 %</td>
<td>24.077 %</td>
<td>50.194 %</td>
<td>75.876 %</td>
<td>85.587 %</td>
<td>91.187 %</td>
</tr>
<tr>
<td>MAP (Pareto)</td>
<td>8.388 %</td>
<td>13.989 %</td>
<td>24.127 %</td>
<td>50.236 %</td>
<td>75.902 %</td>
<td>85.581 %</td>
<td>91.181 %</td>
</tr>
</tbody>
</table>

Table 4.3: Model fitting results - out of sample performance

From the result, we can see that the MAP provides a consistently better performance than the MLEs under both Exponential and Pareto tail models. Specifically, for the majority body (as $Pr = 50\%$), MLEs and MAP estimators both predict well.
For extreme values (as Pr = 5% or Pr = 95%), however, MAP estimates show advantages over MLEs in the out-of-sample prediction. The two tail models, Exponential and Pareto models, provide similar results in while the extreme tails are underestimated.

### 4.4 In-Sample Testing

To check the goodness of fits of the models and the appropriateness of the quadratic approximation of the penalized likelihood functions, we also conduct an in-sample-testing with the tuning parameters $\alpha_k^\pm$ and $\eta_k^\pm$ selected from multiple out-of-sample tests.

The in-sample prediction performance measures under Exponential tails and Pareto tails, respectively, are given in Table 4.4.

<table>
<thead>
<tr>
<th>Pr</th>
<th>5 %</th>
<th>10 %</th>
<th>20 %</th>
<th>50 %</th>
<th>80 %</th>
<th>90 %</th>
<th>95 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE (Exp)</td>
<td>6.553 %</td>
<td>11.500 %</td>
<td>21.482 %</td>
<td>50.109 %</td>
<td>78.396 %</td>
<td>88.131 %</td>
<td>93.093 %</td>
</tr>
<tr>
<td>MLE (Pareto)</td>
<td>6.494 %</td>
<td>11.367 %</td>
<td>21.390 %</td>
<td>50.230 %</td>
<td>83.782 %</td>
<td>93.282 %</td>
<td>97.450 %</td>
</tr>
<tr>
<td>MAP (Exp)</td>
<td>6.541 %</td>
<td>11.457 %</td>
<td>21.500 %</td>
<td>50.123 %</td>
<td>78.413 %</td>
<td>88.187 %</td>
<td>93.062 %</td>
</tr>
<tr>
<td>MAP (Pareto)</td>
<td>6.499 %</td>
<td>11.372 %</td>
<td>21.430 %</td>
<td>50.221 %</td>
<td>83.719 %</td>
<td>93.239 %</td>
<td>97.515 %</td>
</tr>
</tbody>
</table>

Table 4.4: Model fitting results - in sample performance
From Table 4.4, one can see that the results under MLE and MAP are very close, which implies that the filter tuning process does not introduce much bias in prediction. Also note that the Pareto tail model overestimates the right extreme tail while both Exponential and Pareto tail models slightly underestimate the left tail. This is in line with the empirical evidences that the density curve of stock returns have fatter tails and a higher peak at the mean than the normal distribution, which is also called leptokurtosis. This in-sample result substantiates the well-known asymmetry of the return distribution, i.e., the lower (left) tail is heavier than the upper (right) tail.
Chapter 5

Summary and Future Research

5.1 Summary

In this thesis, we have provided a formulated methodology for the prediction of heavy-tailed data with some covariates. For outlying observations in a large dataset, we consider them as meaningful extreme values rather than abnormal outliers as treated in classical statistics. Therefore several methods available in extreme value analysis have been employed to explicitly model them. First a robust regression method is used to fit the main body and to identify the extreme values (residuals) in a robust way. The Peaks-Over-Threshold (POT) method is applied to the residuals to select the extreme values. Based on the extreme value theory, these selected extreme values (or tail observations) follow a Pareto distribution or an Exponential distribution. For the estimation of these tail distributions, the Bayesian maximum a posterior (MAP) estimation is used to smooth the maximum likelihood estimates.
(MLEs). This filter tuning process provides stability and efficiency in both computation and prediction. The proposed methodology has been demonstrated with a stock return dataset on which the out-of-sample prediction results show the advantages of MAP estimation over the regular MLEs.

5.2 Future Research

Each subset in the dataset that has been used in this thesis has about 430 data points. Even though the size of the data is not small, it is still challenging to fit extreme tails since the extreme POT thresholds will end up with only a few observations (e.g., only 21 obs under 5% threshold). This results in unreliable performances for highly extreme tails in both in-sample and out-of-sample predictions regardless of the estimation methods. With mean residual plots, we are able to find theoretically better thresholds if the linearity appears. However, it is not guaranteed and needs to be individualized for each subset and each tail. I will study a more stable and efficient way to choose the thresholds for the POT method.

Recall that there are four stock index returns (the covariate vector) that were considered as given information. However, these stock indexes are stochastic in nature and thus, it is desirable to model them appropriately. The incorporation of randomness to indexes will result in a normal mixture model for the body part of the data distribution. This also provides a cross sectional dependence between the individuals
in the same characteristics (e.g., industry sector, country and size). For my future research, I will study a more general modelling (random factor models) and prediction framework which can explain both temporal and cross-sectional dependence in a big data situation.

In the filter tuning process, the MLEs are smoothed undiscriminated which may cause over-smoothing for some subsets. In the future study, we consider customizing the smoothing process (i.e., by different sectors or using different tuning parameters based on the specific subset).

Based on the extreme value theory, the tail distribution follows either a Pareto distribution or an exponential distribution. In this thesis both distributions were attempted to provide two possible results. It will be much more efficient and accurate if one of the distributions can be decided for each subset by using an appropriate model selection method. I will pursue an approach in this direction.
Bibliography


Appendix

Data Resources

1. Basic Stock Return data: Yahoo Finance (https://ca.finance.yahoo.com/)


   (For independence, considering Dow Jones indices first.)
   *Dow Jones Health Care Titans 30 Index, Ticker: DJTHCA;
   *Dow Jones U.S. Basic Materials Index, Ticker: DJUSBM;
   *Dow Jones U.S. Financial Services Index (USD), Ticker: DJUSFV;
   *Dow Jones U.S. Industrials Index, Ticker: DJUSIN;
   *S&P 500 Energy, Ticker: SPN;
   *Dow Jones U.S. Technology Index, Ticker: DJUSTC
   *Dow Jones U.S. Utilities Index, Ticker: DJUSUT
   *S&P Composite 1500 Real Estate;
   *Dow Jones U.S. Consumer Services Index, Ticker: DJUSCY as Consumer Defensive
4. Country Indices

Currency in USD (https://ca.finance.yahoo.com/)

*Asian (South Korea): S&P Asia 50, Ticker: SPA50

*Europe (Luxembourg and Switzerland): S&P Europe 350, Ticker: SPEURO

*Canada: S&P/TSX 60, Ticker: SPTSX60

5. Company Size Indices (www.spindices.com)

*Large: S&P Global LargeCap (US Dollar)

*Mid: S&P Global MidCap (US Dollar)

*Small: S&P Global SmallCap, Ticker: SBERGLU

*Micro: Dow Jones Select Micro-Cap Index, Ticker: DJSM

Note:

Company Size Categorized by

*Mega-Cap (greater than USD 100 billion) (Index includes this into Large-Cap)

*Large-Cap (between USD 10 billion to USD 100 billion)

*Mid-Cap (between USD billion and USD 10 billion)

*Small-Cap (between USD 300 million to USD 2 billion)

*Micro-Cap (between USD 50 million to USD 300 million)

*Nano-Cap (less than USD 50 million)