Topics in Statistical Modeling and Estimation of Extremes and Their Dependence

by

Kamal Hamidieh

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2008

Doctoral Committee:

Professor George Michailidis, Co-Chair
Assistant Professor Stilian A. Stoev, Co-Chair
Professor Robert W. Keener
Professor H. Nejat Seyhun

I dedicate this to my parents, sister, and my brother-in-law for providing me with free laptops.

ACKNOWLEDGEMENTS

This work could not have been completed without the guidance and patience of my advisors Professors Stilian Stoev and George Michailidis. I was fortunate to have Dr. Keener as a teacher and a committee member. I am grateful to Dr. Seyhun for his support and thoughtful advice. Dr. Amirdjanova provided me with much support and mentoring in my early days at the department. Dr. Gunderson has been a role model for me as an excellent instructor and communicator. I would also like to thank our department staff (in alphabetical order) Lu Ann Custer, Suleman Diwan, Candy Ellis, Meghan Genovese, Shannon Halbedel, Mary Ann King, and Amy Rundquist for doing a great job in making my graduate life as easy as possible.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER I

Introduction

The significant impact of the extreme and rare events such as floods, hurricanes, earthquakes, and stock market crashes has impelled many researchers to study the nature and effects of these events. Of interest to the researchers and indeed to the general public is how to use available historical data from these events such as the frequency of floods, the severity of earthquakes, and the magnitude of stock market crashes to make predictions about the size and frequency of the future extreme and rare events. From the statistical perspective, Extreme Value Theory (EVT) has been an important and successful set of tools developed for describing and modeling such events.

A particular area where the tools of EVT has become quite widespread is in the quantitative finance and risk management. The applications of these tools have grown tremendously due to the recognition by the financial community that extreme events and risks are more common than one would think based on the traditional statistical models. As pointed out by Mandelbrot and Hudson (2004)"Markets are very, very risky - more risky than the standard theories imagine". Recent events bear the concerns expressed: The stock market crash of 1987, the collapse of the Long Term Capital Management, and the massive losses at Baring PLS, Orange County,

Metallgesellschaft, and the sub-prime meltdown show that extreme losses do occur more frequently than anticipated. The interested reader in the aforementioned events should refer to Marthinsen (2008).

The main goal of this thesis is to use concepts and tools from EVT to model, make inference and develop prediction tools for the extremes of stationary data with a focus on the financial risk management applications. The background theory and the tools related to this work are presented within the next three chapters.

Chapter 2 presents a new estimator of the extremal index. The extremal index is the main parameter that describes and quantifies the clustering characteristics of the extreme values for many stationary time series. The extremal index has important applications in a number of areas, such as hydrology, telecommunications, finance, and environmental studies. In this chapter, an estimator for the extremal index based on the asymptotic scaling of block maxima and resampling is introduced. It is shown to be consistent and asymptotically normal for m-dependent time series. Further, a procedure for the automatic selection of its tuning parameter is discussed and different types of confidence intervals that prove useful in practice are suggested. The performance of the estimator is examined through simulations, which show its highly competitive behavior. Finally, the estimator is applied to two real data sets of daily Crude Oil prices and extreme temperatures.

Chapter 3 focuses on the point process modeling and estimation of the Value at Risk (VaR). A point process is a stochastic process, which describes events occurring at discrete random points of time. Our events of interest are related to the times when an asset sustains a large loss. VaR is defined as an extreme quantile of the loss distribution of an asset. It remains one of the most important and widely used extreme risk measures. The usual VaR is defined only in terms of the marginal

distribution of the asset returns. This definition ignores important dependence effects in the behavior of the returns. In this chapter, in addition to using the marginal distribution, we model the random times at which large losses of an asset occur. We then define a dynamic VaR measure, which predicts the quantiles of the loss distribution, conditionally on the present and past behavior of the returns, according to a certain point process model. The point process based estimation of the VaR is shown to be superior to the more widely accepted methods of unconditional or static VaR estimation. An analysis of real data set demonstrates the advantages of our methodology in practice.

In Chapter 4, new risk measures are proposed to incorporate the clustering of the large losses as often exhibited by the asset returns. This empirical fact has important consequences. For example, a risk manager is responsible for assessing the potential losses that an asset (or an entire portfolio of assets) can sustain. If clustering of extreme losses is present, it means that once the risk manager observes one loss, there is a good chance that he or she will see another big loss within a short period of time. In such short periods, the accumulation of the large losses can wipe out a significant portion of the value of an asset. The new risk measures we propose build on the conditional expected shortfall, which is just the expect loss of an asset given it has sustained a loss. They take into account the *duration* of the period of extreme losses. The risk measures are shown to be good predictors for the accumulation of extreme losses. Additionally, a new procedure for estimating the extremal index is proposed and verified through a simulation study. This procedure combines two established estimators of the extremal index to obtain better estimates of the extremal index and more accurately identify the independent clusters of the extreme values.

CHAPTER II

Estimation of the Extremal index Based on Scaling and Resampling

2.1 Introduction

Advances in computer technology have enabled the collection by research organizations and businesses of large time series data sets. These data sets are primarily characterized by the fine granularity (*high frequency*) of the time intervals at which the observations are collected; for example, Internet traffic is sampled at millisecond intervals, while stock trades at every second. Such time series data are characterized by the presence of long range dependence (the autocorrelation function decays at a polynomial rate) and the heavy tailed nature of the marginal distribution (see, e.g. Adler *et al.* (1998)). In many cases, another phenomenon can be observed, namely the presence of *clustering* of very large or very small values (*extremes*) of the data. For example, in Internet traces this is the result of bursty arrivals, while in data on the returns of a financial asset this is due to the arrival of an external market shock.

The daily log-returns of the spot price of West Texas Intermediate crude oil are shown for the period September 2006 – March 2007 in Figure 2.1. A pronounced temporal clustering of the extreme values can be seen, indicating the presence of local dependence in the extremes. Such a behavior is of interest to subject matter experts and it has important implications in practice, since it concerns large consecutive

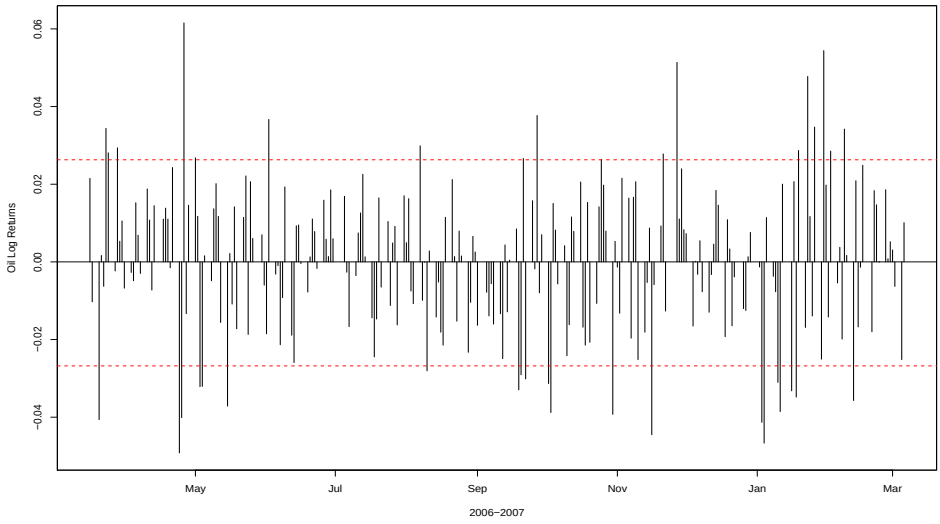

Figure 2.1: Observe the clustering of extremes, particularly evident in the extreme price drops or 'losses' (above the horizontal dotted line).

changes associated with large financial losses or gains. Therefore, quantifying the nature of the dependence structure as well as the duration of extreme events becomes an essential part of the understanding of these time series data.

The extremal index θ is the main parameter that describes and quantifies the clustering characteristics of the extreme values in many stationary time series. An informal interpretation of θ is given in Leadbetter *et al.* (1983), namely $\theta \approx$(mean cluster size)$^{-1}$. For example, for the crude oil log-returns, the extremal index is estimated to be around 0.56, which means that on the average, two large size losses or gains are recorded in relatively short time span. The modeling and analysis of rare events (extremes) has been an active area in probability and statistics (see e.g. Leadbetter *et al.* (1983), Resnick (1987), Embrechts *et al.* (1997), Beirlant *et al.* (2004)). In the context of extremes, the study and the estimation of the extremal index θ, formally defined next, plays an important role.

Let $X = \{X_k\}_{k\in\mathbb{Z}}$ be a strictly stationary time series. Let also

$$M_n := \max_{1\le k\le n} X_k \qquad \text{and} \qquad M_n^{iid} := \max_{1\le k\le n} \widetilde{X}_k,$$

where the $\widetilde{X}_k$'s are independent and identically distributed (iid) random variables with the same distribution as the X_k's. Formally, the time series X is said to have an *extremal index* θ, if for some norming sequences $c_n > 0$ and d_n, we have

$$\text{(2.1.1)}\quad \mathbb{P}\{c_n^{-1}(M_n^{iid} - d_n) \le x\} \xrightarrow{w} H(x) \quad \text{and} \quad \mathbb{P}\{c_n^{-1}(M_n - d_n) \le x\} \xrightarrow{w} H^\theta(x),$$

where $H(\cdot)$ is a non–degenerate extreme value distribution (see e.g. p. 417 in Embrechts *et al.* (1997)). An important mixing condition when dealing with the extremal index as follows:

Definition II.1 (Condition $D(u_n)$)**.** For any integers p,q, and n

$$1 \le i_1 < \ldots < i_p < j_1 < \ldots < j_q \le n$$

such that $j_1 - i_p \ge l$ we have

$$|\mathbb{P}\{\textstyle\bigvee_{i\in A_1 \bigcup A_2} X_i \le u_n\} - \mathbb{P}\{\bigvee_{i\in A_1} X_i \le u_n\}\mathbb{P}\{\bigvee_{i\in A_2} X_i \le u_n\}| \le \alpha_{n,l},$$

where $A_1 = \{i_1, \ldots, i_p\}, A_2 = \{j_1, \ldots, j_q\}$, $u_n = c_n x + d_n$, and $\alpha_{n,l} \to 0$ as $n \to \infty$ for some sequence $l = l_n = o(n)$.

The condition $D(u_n)$ is a form of an asymptotic independence condition but it is weaker than most traditional conditions since it only considers the events of the form $\{X \le x\}$. This condition gives the criteria to ensure asymptotic independence of extreme values of a stationary process when the groups of extreme values are separated both in the level *and* separation distance. If X is a strictly stationary sequence with $F(x)$ marginal distribution and $D(u_n)$, satisfying

$$\lim_{n\to\infty} n\bar{F}(u_n(\tau)) = \tau$$

for all $\tau > 0$, and if $\lim_{n\to\infty} \mathbb{P}\{M_n \leq u_n(\tau)\}$ exists for some $\tau > 0$, then

$$\lim_{n\to\infty} \mathbb{P}\{M_n \leq u_n(\tau)\} = e^{-\theta\tau},$$

where θ is the extremal index of X. Note if θ exists, it will not be a function of $u_n(\tau)$. We now directly state Theorem 3.7.1, page 66, from Leadbetter *et al.* (1983), which gives the necessary regularity conditions for the existence of the extremal index.

Theorem 3.7.1 - (Leadbetter *et al.* (1983))

Suppose $u_n(\tau)$ is defined for $\tau > 0$ and is such that $n(1 - F(u_n(\tau))) \to \tau$, and such that $D(u_n(\tau))$ holds for each such τ. Then there exists constants θ, θ', $0 \leq \theta \leq \theta' \leq 1$ such that $\limsup_{n\to\infty} \mathbb{P}\{M_n \leq u_n(\tau)\} = e^{-\theta\tau}$, $\liminf_{n\to\infty} \mathbb{P}\{M_n \leq u_n(\tau)\} = e^{-\theta'\tau}$. Hence if $\mathbb{P}\{M_n \leq u_n(\tau)\}$ converges for some $\tau > 0$, then $\theta = \theta'$, and $\mathbb{P}\{M_n \leq u_n(\tau)\} \to e^{-\theta\tau}$ for all $\tau > 0$.

Here, we focus on the non–degenerate case when the extremal index θ is positive. Observe that in this case the same normalization and centering sequences for the partial maxima M_n and M_n^{iid} above yield non–degenerate limit distributions. The extremal index takes values in the interval $[0, 1]$; a value close to 0 indicates a very strong short range extremal dependence, while a value close to 1 a rather weak dependence. In fact, for iid X_k's, by (2.1.1), we have $\theta = 1$. The extremal index, however, characterizes only the dependence of the extremes in the time series data and thus the data may still exhibit strong dependence, even though $\theta \approx 1$. The case of $\theta = 0$ is considered pathological.

Theoretical properties of the extremal index have been studied fairly extensively; (Leadbetter (1983), O'Brien (1987), Hsing *et al.* (1988), Leadbetter and Rootzén (1988), and references therein). The problem of estimating θ has also received some attention in the literature: Hsing (1993), Smith and Weissman (1994), Weissman and Novak (1998) and Ferro and Segers (2003). For example, Hsing (1993) studied

an estimator based on the idea that θ is proportional to the inverse of the mean cluster size of exceedances above a certain threshold. A key tuning parameter for this estimator is that the data are examined in blocks of size r. An alternative estimator proposed by O'Brien (1987) is based on the length of runs of values of the process falling below a prespecified threshold, given that an exceedance has occurred. The properties of these estimators are studied in Hsing (1993), Smith and Weissman (1994) and further refined in Weissman and Novak (1998). Recently, Ferro and Segers (2003) proposed another estimator which does not require the specification of a block size or a run length, and it exhibits a fairly robust performance in practice.

Applications of the extremal index in various scientific areas include its incorporation in calculations of the Value-at-Risk measure (Longin (2000) and Klüppelberg in Finkenstädt and Rootzén (2004)), in the study of the Nasdaq and S&P 500 indices (Galbraith and Zernov (2006)) and in the study of GARCH processes (Laurini (2004)). The estimation of the extremal index θ is an important practical problem with rapidly expanding areas of application to finance, insurance, hydrology and telecommunications, to name a few (for more details, see e.g. Embrechts *et al.* (1997) and Finkenstädt and Rootzén (2004)).

Most previous estimators of θ exploit its connection to the point process of exceedance. Here, we introduce a new method for estimating θ based on the asymptotic scaling properties of block–maxima and resampling. Specifically, let $X_1, \ldots, X_n$ be a data sample from a heavy–tailed time series (see (2.2.2)) with positive extremal index θ. The maximum values of the data calculated over blocks of size m, scale at a rate $m^{1/\alpha}$, where $\alpha > 0$ denotes the *tail index* of the marginal distribution of the data. Further, the normalized limit of the block maxima is proportional to $\theta^{1/\alpha}\sigma$, where $\sigma := c_X^{1/\alpha} > 0$ is an asymptotic scale coefficient of the X_k's (see (2.2.2)). Thus, by ex-

amining a sequence of growing, dyadic block sizes $m = 2^j,\ 1 \leq j \leq \lfloor \log_2 n \rfloor,\ j \in \mathbb{N}$, and subsequently estimating the mean of logarithms of block–maxima one obtains *estimating equations* involving both the tail index α and the parameter $\theta^{1/\alpha}\sigma$. In these equations, the scale σ and the extremal index θ are, however, coupled. Although, in principle θ can be calculated by solving an appropriate nonlinear equation, the resulting estimate proves to be too variable. Hence, we resort to *resampling*. Specifically, we consider either a *bootstrap* or a *random permutation* sample of the original data and *then* apply the previous methodology. The resampled data behaves, asymptotically, as an independent sequence with unit extremal index, that yields a second set of *estimating equations* of the tail index α and the parameter σ. By combining the resulting two estimating equations, one based on the original data and another based on the resampled data, we obtain a numerically stable estimate of θ.

The resulting estimators for θ are shown to be *consistent* and *asymptotically normal* for $m-$dependent sequences, while at the same time exhibiting good numerical performance in finite samples. An additional advantage of the resampling is that it provides a supplementary way of calculating confidence intervals for θ. Simulation studies show that the proposed estimator is a competitive alternative to existing ones. Further, it provides new insights at the important parameter θ from the perspective of resampling and can be successfully used to analyze small as well as large data sets in practice.

The remainder of the Chapter is organized as follows: Section 2.2 describes the proposed estimator. Its asymptotic properties are established in Section 2.3. Several methodological and implementation issues are discussed in Section 2.4, while Section 2.5 focuses on the evaluation of the estimator through an extensive simulation study. Two important data sets of daily Crude Oil prices and Wooster Extreme

Temperatures are examined in Section 2.6. The proofs and some auxiliary results are given in the Appendix.

2.2 The max–spectrum based estimator of θ

Let $X = \{X_k\}_{k\in\mathbb{Z}}$ be a positive ergodic strictly stationary sequence with *heavy tailed* marginals and *positive* extremal index $\theta > 0$. Specifically, assume that

$$\mathbb{P}\{X_k > x\} = 1 - F(x) \sim c_X x^{-\alpha}, \quad \text{as } x \to \infty \tag{2.2.2}$$

for some $\alpha > 0$ and $c_X > 0$, where $a_n \sim b_n$ means $a_n/b_n \to 1$, as $n \to \infty$. The parameter α is called the *tail index* of the distribution (see e.g. Resnick (2006)). Given a sample path $X_1, \ldots, X_n$, we define the dyadic block maxima as follows:

$$D(j,k) := \max_{1\le i\le 2^j} X_{2^j(k-1)+i} \equiv \bigvee_{i=1}^{2^j} X_{2^j(k-1)+i}\,, \tag{2.2.3}$$

where $j = 1, \ldots, \lfloor \log_2 n \rfloor$, $k = 1, \ldots, \lfloor n/2^j \rfloor$, and where $\lfloor\cdot\rfloor$ denotes the integer part function. For heavy–tailed X_k's, relation (2.1.1) holds with $H(x) = \exp\{-c_X x^{-\alpha}\}$, $x > 0$ and normalization constants $c_n := n^{1/\alpha}$ and $d_n := 0$. Therefore,

$$2^{-j/\alpha} D(j,k) \xrightarrow{D} \theta^{1/\alpha}\sigma Z^{1/\alpha}, \quad \text{as } j \to \infty. \tag{2.2.4}$$

where Z is a standard $1-$Fréchet random variable, i.e. $\mathbb{P}\{Z \le z\} = \exp(-z^{-1})$, $z > 0$, and where $\sigma := c_X^{1/\alpha}$ is the asymptotic scale coefficient of the X_k's. Due to the nature of the Fréchet extreme value distribution, the extremal index parameter θ appears in the scale coefficient of the limit distribution of the dependent maxima. This feature will play an important role in the estimation of θ discussed below.

Next, introduce the statistics

$$Y_j := \frac{1}{n_j}\sum_{k=1}^{n_j} \log_2(D(j,k)). \tag{2.2.5}$$

where $n_j = \lfloor n/2^j \rfloor$. The statistics $Y_j,\ j = 1 \ldots, \lfloor \log_2(n) \rfloor$ will be referred to as the *max–spectrum* of the data, and the j's as *scales*. By the assumed ergodicity and provided that the moments exist, for a fixed j, we get

$$Y_j \xrightarrow{a.s.} \mathbb{E}Y_j = j/\alpha + \mathbb{E}\log_2(2^{-j/\alpha}D(j,k)), \qquad \text{as } n \to \infty. \tag{2.2.6}$$

Assuming uniform integrability, Relation (2.2.4), on the other hand, implies that

$$\mathbb{E}Y_j \simeq j/\alpha + \log_2(\sigma) + \mathbb{E}\log_2(Z)/\alpha + \log_2(\theta)/\alpha, \qquad \text{as } j \to \infty, \tag{2.2.7}$$

where $a_n \simeq b_n$ means $a_n - b_n \to 0$, as $n \to \infty$. This indicates the existence of a *linear* relationship between the statistics Y_j and j up to an error term, which becomes negligible as n_j and j grow. The slope of a linear fit of Y_j versus j yields an estimator of $1/\alpha$ and thus α. Although our goal is to estimate θ, the estimation of the tail index α is an intermediate step and an integral part of our analysis.

Observe that on the other hand for iid data, we have $\theta = 1$ and thus (2.2.7) becomes:

$$\mathbb{E}Y_j^{iid} \simeq j/\alpha + \log_2(\sigma) + \mathbb{E}\log_2(Z)/\alpha, \tag{2.2.8}$$

where $\{Y_j^{iid}\}$ is the max–spectrum of an iid data set with the same distribution as the X_k's. Relations (2.2.7) and (2.2.8) suggest a method to obtain an estimate of θ. Namely, *resample* the data, for example, by randomly drawing (with or without replacement) a sample $X_1^*, \ldots, X_k^*$ of size $k = k(n)$ from the set $\{X_1, \ldots, X_n\}$. Intuitively, this destroys the dependence structure of the data, resulting in an approximately independent sample with the same marginal distribution as the original stationary sequence.

Let Y_j^* be as in (2.2.5) where now the $D(j,k)$'s are based on the resampled data $X_1^*, \ldots, X_k^*$. Since for an iid sequence we have $\theta = 1$, we expect the resampled

sequence to have $\theta \approx 1$, whereas α and σ will remain unchanged. Thus, Relation (2.2.7) becomes

$$\mathbb{E}[Y_j^*] \simeq j/\alpha + \log_2(\sigma) + \mathbb{E}[\log_2(Z)]/\alpha, \tag{2.2.9}$$

where the term $\log_2(\theta)/\alpha$ is no longer present since $\log_2(\theta \approx 1) \approx 0$.

Thus, in view of (2.2.7) and (2.2.8), we have

$$Y_j^* \approx j/\alpha + \log_2(\sigma) + \mathbb{E}\log_2(Z)/\alpha, \quad \text{and} \quad Y_j \approx j/\alpha + \log_2(\sigma) + \mathbb{E}\log_2(Z)/\alpha + \log_2(\theta)/\alpha.$$

Taking the difference between the last two estimating equations, replacing α by its estimate $\hat{\alpha}$ based on (2.2.7), and solving for θ we obtain the following estimator for the extremal index:

$$\hat{\theta}(j) = 2^{-\hat{\alpha}(j)(Y_j^* - Y_j)}. \tag{2.2.10}$$

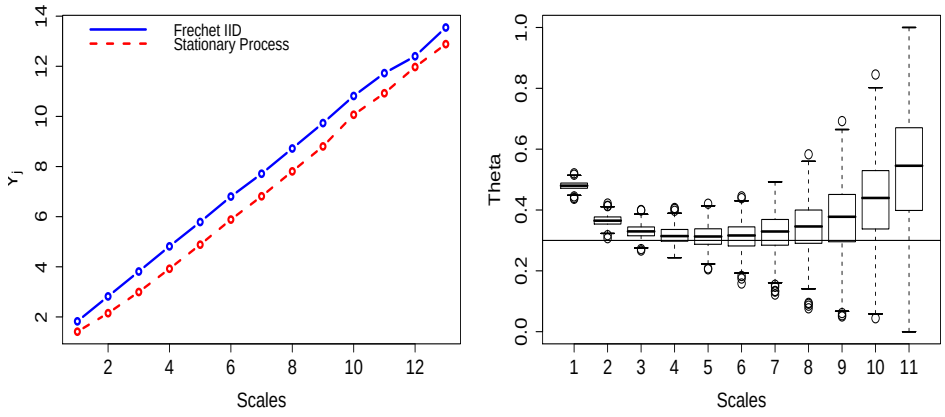

Figure 2.2: *Left panel:* The max–spectrum of $X_n = \max\{\frac{2}{3}X_{n-1}, \frac{1}{3}Z_n\}$, $\theta = 1/3$, with Z_i's iid standard 1−Fréchet (*solid line*) and the max–spectrum of iid copies of the X_k's (*broken line*). The two spectra are essentially linear with equal slopes. *Right panel:* boxplot of $\hat{\theta}(j)$'s obtained from different resampled versions of a single path of the process. The horizontal line is the theoretical value of $\theta = 1/3$.

Observe that for a single data set, one can obtain a large set of estimates $\hat{\theta}(j)$, based on different resampled versions of the data. Thus, resampling allows us to

gauge the variability of the estimates as well as the range of scales j where the asymptotics in (2.2.7) and (2.2.8) become applicable.

Figure 2.2 illustrates the main principle behind the proposed estimator. The left panel shows the combined max–spectra of a dependent sequence and an iid sample. The two max–spectra are parallel with equal slopes $\approx 1/\alpha$, since the marginal distributions behind the two spectra are the same. The difference is in the *intercept* and this is where the value of θ is derived from. The right panel shows boxplots of $\hat{\theta}(j)$ estimates obtained from 200 independent resampled versions of a single path of the process on the left. Observe that the medians of the $\hat{\theta}(j)$'s closely follow the true value $\theta = 1/3$ over a range of scales (for more details, see Section 2.4 below).

Remarks

1. The statistics Y_j's in (2.2.5) are not only dependent in j, but more importantly, they have different variances in j since they involve averages of $n_j \approx n/2^j$ terms. Thus, to reduce the variance in the regression estimators of α, it is essential to use a weighted or generalized least squares method (see e.g. Stoev *et al.* (2006), for more details).

2. The proposed resampling procedure avoids the problem of estimating the scale parameter $\sigma = c_X^{1/\alpha}$, however, an estimate of α is still needed. The algorithmic implementation of the estimators $\hat{\theta}(j)$ and other important practical issues are discussed below.

   The appropriate resampling sample size $k(n)$, from the perspective of asymptotics, is $o(\sqrt{n})$ (see, Section 2.3).

3. The estimate $\hat{\theta}(j)$ depends on the scale j, as indicated. An automatic procedure for the choice of j is presented in Section 2.4.

2.3 Theoretical properties

2.3.1 Description of the asymptotic regime

Let $X = \{X_k\}_{k\in\mathbb{Z}}$ be a strictly stationary time series with marginal c.d.f. F as in (2.2.2) and let also

$$M_n = \max_{1\le i\le n} X_i \equiv \bigvee_{i=1}^{n} X_i. \tag{2.3.11}$$

We shall assume that

$$F_n(x) := \mathbb{P}\{M_n \le n^{1/\alpha}x\} = \exp\{-c(n,x)x^{-\alpha}\}, \quad x \in \mathbb{R}, \tag{2.3.12}$$

for some function $c(n,x) > 0$, $n \in \mathbb{N}$, such that $F_n(x) \in (0,1)$.

As in (2.2.4), if the time series X has a positive *extremal index* $\theta \in (0,1]$, then

$$n^{-1/\alpha}M_n \xrightarrow{D} (\theta c_X)^{1/\alpha}Z^{1/\alpha}, \quad \text{as } n \to \infty, \tag{2.3.13}$$

where Z is a standard $1-$Fréchet variable: $\mathbb{P}\{Z \le x\} = e^{-x^{-1}},\ x > 0$.

Our asymptotic results rely on the moment behavior of $f(M_n/n^{1/\alpha})$, for certain deterministic functions f and involve some additional technical conditions, outlined below (for more details, see the Appendix).

Condition 1. *There exists $\beta > 0$ and $R \in \mathbb{R}$, such that*

(2.3.14)

$$|c(n,x) - \theta c_X| \le c_1(x)n^{-\beta}, \quad \textit{for all } x > 0, \qquad \textit{and} \qquad c_1(x) = \mathcal{O}(x^{-R}),\ x \downarrow 0,$$

where $\theta \in (0,1]$.

Condition 2. *$F_n(0) = 0$ and for all $x > 0$,*

$$c(n,x) \ge c_2 \min\{1, x^{\gamma}\}, \quad \textit{for some } \gamma \in (0,\alpha), \tag{2.3.15}$$

for all sufficiently large $n \in \mathbb{N}$, where $c_2 > 0$ does not depend on n.

Remarks

1. The conditions (2.3.14) and (2.3.15) are not very stringent. For example, let

$$X_k = \max\{Z_k, Z_{k-1}, \ldots, Z_{k-m+1}\}, \quad k \in \mathbb{Z}, \tag{2.3.16}$$

where the Z_k's are independent, standard $\alpha-$Fréchet. We then have

$$\mathbb{P}\{M_n \leq n^{1/\alpha}x\} \;=\; \mathbb{P}\{Z_{-m+1} \leq n^{1/\alpha}x, \cdots, Z_n \leq n^{1/\alpha}x\} = \exp\{-c(n,x)x^{-\alpha}\},$$

where the function $c(n,x) = (n+m-1)/n = 1 + \mathcal{O}(1/n)$ does not depend on x and $\beta = 1$, in this simple case.

Conditions 1 and 2 above hold for a more general class of moving maxima processes (see Proposition II.8).

2. Condition 1 and Relation (2.3.12) imply (2.3.13), that is, the extremal index of the time series X is precisely equal to θ in (2.3.14). Thus (2.3.14) quantifies further the *rate* of the convergence in (2.3.13).

Description of the asymptotic regime: To obtain the consistency of statistics based on the max–spectrum $Y = \{Y_j\}$, we focus on the range of scales $[j(n), \ell + j(n)]$, where $\ell \in \mathbb{N}$ is fixed and where $j(n) \to \infty$, as $n \to \infty$. We then define

$$\hat{\alpha}(j) := \Big(\sum_{i=0}^{\ell} w_i Y_{i+j(n)} \Big)^{-1}, \tag{2.3.17}$$

where the weights w_i's are fixed and such that $\sum_{i=0}^{\ell} w_i = 0$ and $\sum_{i=0}^{\ell} i w_i = 1$. The weights w_i's can be obtained, for example, either from GLS or WLS regression of $Y_{i+j(n)}$ versus i, for $0 \leq i \leq \ell$ (see Stoev *et al.* (2006), for more details).

The estimator $\hat{\theta}$ in (2.2.10) involves both the max–spectrum Y of the *dependent* data and the max–spectrum Y^* of the *resampled* data. Observe that

$$\hat{\theta}(j) = 2^{-\hat{\alpha}(j)(C^*(j) - C(j))}, \quad \text{where} \tag{2.3.18}$$

$$C^*(j) := Y_j^* - j/\alpha \quad \text{and} \quad C(j) := Y_j - j/\alpha, \tag{2.3.19}$$

since trivially $Y_j^* - Y_j = C^*(j) - C(j)$. We will establish the asymptotic normality of $\hat{\theta}(j)$ in three steps:

(Step 1.) We first establish rates of convergence for the quantities $\hat{\alpha}(j)$ and $C(j)$, which are based on the max–spectrum $\{Y_j\}$.

(Step 2.) We then show that the $C^*(j)$'s are asymptotically normal (under certain conditions) in two resampling schemes: *bootstrap* and *random permutations*.

(Step 3.) We finally combine the results from *Steps 1.* and *2.* above to establish the asymptotic normality of $\hat{\theta}(j)$.

2.3.2 Main results

We establish next the asymptotic normality of $\hat{\theta}(j)$ defined in (2.3.18), by following the three steps outlined in the previous section.

Step 1: The following result provides rates of convergence for $\hat{\alpha}(j)$ and $C(j)$, in the asymptotic regime described above.

Proposition II.2. *Let $X_1, \ldots, X_n$ be a sample from an $m-$dependent, strictly stationary time series $X = \{X_k\}_{k\in\mathbb{Z}}$, which satisfies Conditions 1 and 2 above.*

Then, for $\hat{\alpha}(j)$ and $C(j)$ in (2.3.17) and (2.3.19), we have, as $n \to \infty$

$$\hat{\alpha}(j) = \alpha + \mathcal{O}_P\left(\frac{1}{2^{j(n)\min\{1,\beta\}}}\right) + \mathcal{O}_P\left(\frac{2^{j(n)/2}}{n^{1/2}}\right), \quad \textit{and} \tag{2.3.20}$$

$$C(j) = C + \mathcal{O}_P\left(\frac{1}{2^{j(n)\min\{1,\beta\}}}\right) + \mathcal{O}_P\left(\frac{2^{j(n)/2}}{n^{1/2}}\right), \tag{2.3.21}$$

with $C = \log_2(\theta)/\alpha + \log_2(c_X)/\alpha + \mathbb{E}\log_2(Z)/\alpha$, where Z is a standard $1-$Fréchet variable.

The proof of this result is given the Appendix. Observe that Proposition II.2 is valid for an arbitrary stationary $m-$dependent time series which satisfies (2.3.14)

and (2.3.15). It is valid, in particular, for the simple process $\{X_k\}_{k\in\mathbb{Z}}$ in (2.3.16) and more generally for the moving maxima processes in (2.5.26) below (see Proposition II.8 in the Appendix).

Step 2: We now employ resampling to obtain an approximately independent data sample $X_1^*, \ldots, X_k^*$. Here, we consider two resampling schemes: *bootstrap* and *permutations* and in both cases, we obtain asymptotic normality results for the max–spectrum. The sample $X_1^* := X_{i_1},\ X_2^* := X_{i_2}, \ldots, X_k^* := X_{i_k}$ is a *bootstrap* sample from the data $X_1, \ldots, X_n$ if the indices $i_1, \ldots, i_k$ are drawn randomly and with replacement from the set $\{1, \ldots, n\}$. When these indices are drawn without replacement and $k \leq n$, we obtain a *permutation* sample. We need the following:

Lemma II.3. *Let $i_1, \ldots, i_k$ be collection randomly drawn indices either with replacement or without replacement from the set $\{1, \ldots, n\}$.. For any fixed $m \in \mathbb{N}$, we have*

$$\mathbb{P}\{\min_{1\leq j'<j''\leq k} |i_{j'} - i_{j''}| \geq m\} \geq 1 - mk^2/(n-k).$$

The proof is given in the Appendix. This result implies that for $k(n) = o(\sqrt{n}),\ n \to \infty$, the indices $\{i_j,\ 1 \leq j \leq k\}$ are spaced by at least $m-$lags away from each other, with probability asymptotically equal to 1, as $n \to \infty$. Therefore, if the data $X_1, \ldots, X_n$ come from an $m-$dependent time series, for the purposes of asymptotics in distribution, both the *bootstrap* and the *permutation* samples of size $k = o(\sqrt{n})$ become essentially independent, with high probability, as $n \to \infty$. This fact and Proposition 4.2 in Stoev *et al.* (2006), readily imply the following result.

Theorem II.4. *Let $X = \{X_i\}_{i\in\mathbb{Z}}$ be a strictly stationary $m-$dependent time series, which satisfies Conditions 1 and 2 above. Let $X_1^*, \ldots, X_k^*$ be either a bootstrap or a permutation sample from $X_1, \ldots, X_n$, where $k(n) \to \infty$ is such that $k(n) = o(n^{1/2})$,*

as $n \to \infty$, and let Y^* be its corresponding max–spectrum.

Let $j(k) \to \infty,\ n \to \infty$, be such that $k/2^{j(k)(1+2\beta)} + j(k)^2 2^{j(k)}/k \longrightarrow 0$, as $k \to \infty$.

Then, for $C^*(j)$ in (2.3.19), we have

$$\sqrt{k_j}(C^*(j) - C^*) \xrightarrow{D} \mathcal{N}(0, \sigma^2_{C^*}), \qquad as\ n \to \infty, \tag{2.3.22}$$

where $k_j = k(n)/2^{j(n)}$. Here $C^* := \log_2(c_X)/\alpha + \mathbb{E}\log_2(Z)/\alpha$, and $\sigma^2_{C^*} = \alpha^{-2}\mathrm{Var}(\log_2 Z)$, where Z is a standard $1-$Fréchet variable.

The proof is given in the Appendix.

Step 3: The following Theorem is the main result of the Section. It combines the results of Proposition II.2 and Theorem II.4 to establish the asymptotic normality of $\hat{\theta}(j)$.

Theorem II.5. *Assume the conditions of Theorem II.4 and let $\hat{\alpha}(j)$ be as in (2.3.17), where Y is the max–spectrum of the data $X_1, \ldots, X_n$. Let also $C(j)$ and $C^*(j)$ be as in (2.3.19), where Y^* is the max–spectrum of either a bootstrap or a permutation sample $X_1^*, \ldots, X_k^*$ of the data.*

Let $k(n) = o(\sqrt{n}),\ n \to \infty$ and $j(k) \to \infty,\ k \to \infty$, be such that

$$k/2^{j(k)(1+2\min\{1,\beta\})} + j(k)^2 2^{j(k)}/k \longrightarrow 0, \qquad as\ k \to \infty, \tag{2.3.23}$$

Then, for $\hat{\theta}(j)$ in (2.3.18), we have

$$\sqrt{k_j}(\hat{\theta}(j) - \theta) \xrightarrow{D} \mathcal{N}(0, \theta^2\pi^2/6), \qquad as\ n \to \infty,$$

where $k_j = k(n)/2^{j(n)}$.

The proof of this result is given in the Appendix. A few important remarks follow.

Remarks

1. Theorem II.5 applies, for example, to the class of moving maxima processes in (2.5.26), under mild assumptions on the innovations Z_k's (see Conditions $1'$ & $2'$ below). It holds, for example, for Pareto, mixtures of Pareto or Fréchet innovations.

2. Let $\delta \in (0, 2\min\{1,\beta\})$ be arbitrary and suppose that $k/2^{j(k)(1+2\min\{1,\beta\})} \sim k^{-\delta}$, $k \to \infty$. We then have $2^j \sim k^{(1+\delta)/(1+2\min\{1,\beta\})}$, $k \to \infty$ which, since $\delta < 2\min\{1,\beta\}$, implies that Relation (2.3.23) holds. This yields the rate $k_j \sim k^{(2\min\{1,\beta\}+\delta)/(1+2\min\{1,\beta\})}$ in Theorem II.5. Since $k = o(\sqrt{n})$ and since $\delta > 0$ can be taken arbitrarily small, we can achieve rates up to $n^{\frac{\min\{1,\beta\}}{(1+2\min\{1,\beta\})}}$. For example, if $\beta > 1/2$ the rate of $n^{1/4}$ is possible while the best possible rate is $o(n^{1/3})$.

2.4 Implementation issues

We present next an algorithmic implementation for the proposed estimator of θ and discuss its main features. We then propose a second algorithm for the automatic selection of scales.

In Theorem II.5, we only consider resampled sets from the data of size $k(n) = o(\sqrt{n})$. In practice, we found that the estimators of θ continue to work well even if one considers random permutations of the entire data sample of size $k(n) = n$. Using bootstrap instead of permutations results in estimates $\hat{\theta}(j)$ with larger variances and bias (for large j's). Thus, in the sequel, we focus on permutation based resampling and utilize the entire data set.

Algorithm 1. *(estimation of θ)*

1. Compute the Y_j's and the $\hat{\alpha}(j)$'s as in (2.2.5) and (2.3.17) based on the original data.

2. Randomly permute (i.e. shuffle) the data, N_{in} times and collect the N_{in} statistics Y_j^*.

3. Find the N_{in} differences of $Y_j^* - Y_j$ and compute the sample mean for the positive differences only: $\Delta(j) = \text{mean}\{Y_j^* - Y_j\}_+$.

4. Obtain the estimates of θ for each scale j: $\hat{\theta}(j) = \max\{2^{-\hat{\alpha}(j)\Delta(j)}, 1\}$.

5. Repeat steps 2, 3, and 4 N_{out} number of times and collect the $\hat{\theta}(j)$ values.

6. Produce a sequence of $\hat{\theta}(j)$ boxplots from the N_{out} available values, per each scale j.

7. Visually inspect the boxplots of $\hat{\theta}(j)$ and select a range of scales where the *medians* of the boxplots stabilize. Estimate θ by using the median values from this range of scales.

We now discuss the steps in the above algorithm.

Step 1: The estimate $\hat{\alpha}(j)$ is based on the range of scales $j, \ldots, j+\ell$, where $j + \ell = \lfloor \log_2(n) \rfloor - 1$ is chosen to be the second largest available scale in the data. In practice, we discard the highest scale since it involves an average of at most two block–maxima. We recommend using either generalized least squares with the asymptotic covariance matrix for the max–spectrum given in Stoev *et al.* (2006) or weighted least squares which account for the fact that $\text{Var}(Y_j) \propto 1/n_j \propto 2^j$. Both approaches are comparable and considerably better than ordinary least squares regression, which should not be used.

Steps 2 & 3: We introduce an *inner loop* with N_{in} iterations to reduce the variability of $Y_j^* - Y_j$. This considerably improves the variance of the θ estimates. On step 3, we average only the positive differences $Y_j^* - Y_j$ since by Relations (2.2.7) and

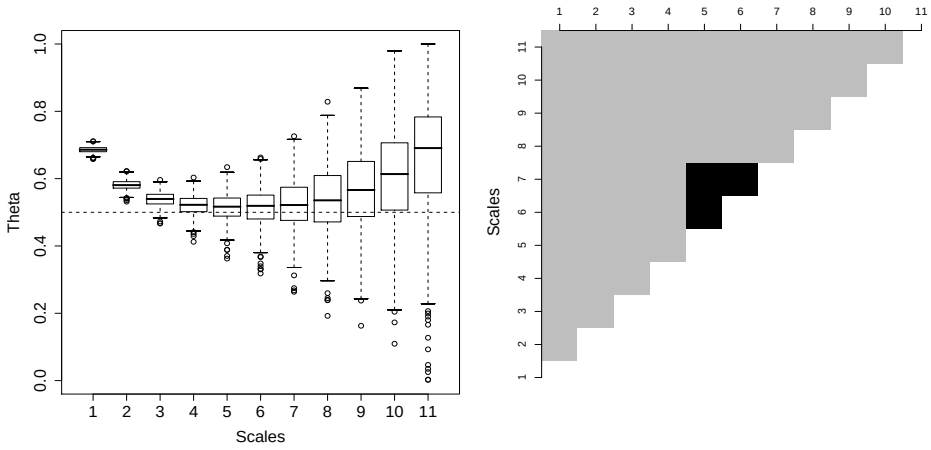

Figure 2.3: Estimation of θ for the process $X_n = \max\{\frac{1}{2}X_{n-1}, \frac{1}{2}Z_n\}$, $\theta = 1/2$, Z_i iid standard 1−Fréchet, with sample size of n=2^{13}, $N_{out} = 500$ and $N_{in} = 25$. *Left panel:* Boxplots of $\hat{\theta}(j)$'s with the last two scales omitted. *Right panel:* A "heat map" visualizing the Kruskal–Wallis test for the automatic selection of scales – black corresponds to $p-$values greater than 0.05.

(2.2.9), we have $\mathbb{E}Y_j^* \geq \mathbb{E}Y_j$. Our experiments indicate that replacing the "mean" by "median" in step 3 yields similar results.

Step 4: As in Ferro and Segers (2003), we take the minimum of the calculated estimate and 1 to ensure that $\hat{\theta}(j) \in [0, 1]$.

Step 5: This step yields a sample of N_{out} estimates of θ for each scale j. The choice of the parameters N_{out} and N_{in} is discussed in Section 2.5.

Step 6: In practice, the estimation of θ requires selecting the range of scales, where the best bias/variance trade–off is achieved. Estimating θ over the larger scales j (larger block sizes) involves lower bias, but leads to larger variance as the number of block–maxima is reduced. At lower scales j (smaller block sizes) the bias grows but the variance is reduced (see Figure 2.3). In general, reliable estimates of θ can be obtained from the middle range of scales. The choice of the scales j is addressed in the sequel.

Figure 2.3 (left panel) illustrates the above algorithm over a simulated process with

known extremal index $\theta = 1/2$. A stable range of scales 4 to 7 can be observed. In practice, we recommend taking the median of the sample of the pooled N_{out} estimates $\hat{\theta}(j)$ from each one of the scales j in the stable range. In this case we obtained a point estimate of 0.52. One can also obtain an empirical 95% confidence interval, based on 0.025–th and 0.975–th empirical quantiles of the pooled $\hat{\theta}(j)$ values to obtain $(0.40, 0.62)$. We discuss further the validity of such type of confidence intervals in Section 2.5.3.

Remark: In a simulation study where *independent* sequences of data are generated and computation time may be of a lesser concern, we recommend setting $N_{out} = 500$ and $N_{in} = 25$. Our experience shows that for the real date, in a finite sample situation, where resampling may not produce independent sets, setting $N_{out} = 200$ and $N_{in} = 1$ produce satisfactory results. Using $N_{in} = 1$ in such cases leads to slightly larger variance, leading to wider confidence intervals, but prevents missing the 'true value' due to the bias.

The selection of the stable range of scales j in Step 6 of the above algorithm is subjective. To *aid* one in the selection of the scales range, we recommend using the Kruskal–Wallis test to automatically determine a range of scales with approximately equal medians:

Algorithm 2. *(automatic selection of scales)*

1. For every given range $j_1 \leq j \leq j_2,\ j_1 < j_2$ of possible consecutive scales in the data, perform a Kruskal–Wallis test for equality of the medians, based on the samples of N_{out} values of $\hat{\theta}(j)$.

2. Consider the array of $p-$values: $p(j_1, j_2)$ from Step 1. Declare the medians over the range $[j_1, j_2]$ 'statistically different' if p is less than a prescribed significance threshold.

3. Produce a pooled estimate of θ based on the longest scale range where the medians are not statistically different.

4. If there are ties in Step 3, pick the range starting at the lowest scale. If all medians are statistically different, pick the middle scale and recommend visual inspection.

The proposed automatic scale selection procedure is evaluated in Section 2.5. One possible method to visualize the results of this analysis is to construct a heat map of the p-values for the Kruskal–Wallis tests – see the right panel of Figure 2.3. The axes correspond to scales j_1 and j_2 and the regions in black indicate ranges of scales $[j_1, j_2]$ with $p-$values greater than 0.05. This heat map shows that the medians over the scale range $[j_1, j_2] = [5, 7]$ are not statistically different at a level of 5%. A point estimate based on the pooled values from scales 5 to 7 is 0.52 with an empirical 95% confidence interval of (0.39, 0.63).

2.5 Performance evaluation

We provide next a brief description of two established estimators for θ, together with a number of processes for which θ can be computed explicitly. We then present a brief simulation study on the performance of these estimators.

2.5.1 Brief Description of Competing Estimators

The first estimator is based on the characterization of the extremal index given by O'Brien (1987). In this characterization, θ is expressed as the limiting probability that an exceedance is followed by a run of observations below a high threshold u_n:

$$\theta = \lim_{n\to\infty} \mathbb{P}\Big\{ \bigvee_{j=2}^{r_n} X_j \leq u_n | X_1 > u_n \Big\},$$

where $r_n = o(n)$ is the *length of runs* of values of the process falling below the threshold given that an exceedance has occurred. This characterization motivates the definition of the *runs* estimator for a fixed high threshold u and a specified runs length r as follows:

$$\hat{\theta}_{runs} = \frac{\sum_{j=1}^{n-r} \mathbf{I}(X_j \geq u \geq \bigvee_{i=j+1}^{j+r} X_i)}{\sum_{j=1}^{n-r} \mathbf{I}(X_j > u)}. \tag{2.5.24}$$

The runs estimator is asymptotically normal and consistent, under certain conditions (See Weissman and Novak (1998) and references therein for additional information.)

The second estimator is due to Ferro and Segers (2003). An interesting aspect of this estimator is that it does not require an auxiliary parameter (run length in the case of the runs estimator). However, one still has to choose the threshold. Using a point process approach, Ferro and Segers (2003) show that the inter–exceedance times - time differences between successive values above a threshold u_n - of the extreme values normalized by $\bar{F}(u_n)$ converge in distribution to a random variable T_θ with a mixture distribution. This is a mixture of a point–mass $1 - \theta$ at $t = 0$ and an exponential distribution with rate θ. Using a moment estimator, they first obtain:

$$\hat{\theta}_1 = \frac{2(\sum_{i=1}^{N-1} T_i)^2}{(N-1)(\sum_{i=1}^{N-1} T_i^2)}, \quad \text{and then consider} \quad \hat{\theta}_2 = \frac{2(\sum_{i=1}^{N-1}(T_i - 1))^2}{(N-1)(\sum_{i=1}^{N-1}(T_i - 1)(T_i - 2))}.$$

Here $\{T_i\}$ are the inter–exceedance times and N is the number of exceedances of a fixed high threshold u. The second estimator $\hat{\theta}_2$ is a bias corrected version allowing for zero inter–exceedance times.

The final form of the estimator ensures that it always lies between 0 and 1:

$$\hat{\theta}_{F/S} = \begin{cases} 1 \wedge \hat{\theta}_1 & \text{if } \max\{T_i : 1 \leq i < N-1\} \leq 2, \\ 1 \wedge \hat{\theta}_2 & \text{if } \max\{T_i : 1 \leq i < N-1\} > 2. \end{cases} \tag{2.5.25}$$

The Ferro–Segers estimator is consistent for m-dependent strictly stationary sequences.

2.5.2 Summary of numerical results

We start by presenting three types of processes with closed form expressions for the extremal index. These processes will be further used in to evaluate our estimators.

• The *max-autoregressive* (armax) process of order one is defined as:

$$X_n = \max\{bX_{n-1}, (1-b)Z_n\}, \quad \text{where} \quad 0 \leq b < 1,$$

and where $\{Z_n\}_{n\in\mathbb{Z}}$ is an iid sequence of standard $\alpha-$Fréchet random variables. For such processes $\theta = 1 - b^\alpha$ can take any value in the interval $(0, 1]$ (see e.g. Beirlant *et al.* (2004) for additional information).

• The *linear process* $\{Y_n\}$, $n \in \mathbb{Z}$ is defined as:

$$Y_n = \sum_{j\in\mathbb{Z}} \psi_j Z_{n-j},\ n \in \mathbb{Z}, \quad \text{where} \quad \sum_{j\in\mathbb{Z}} |\psi_j|^\delta < \infty, \quad \text{for some } 0 < \delta < \min\{1, \alpha\}.$$

Here $\{Z_n\}_{n\in\mathbb{Z}}$ is an iid sequence of heavy–tailed innovations with exponent $\alpha > 0$. When the Z_n's are symmetric, we have $\theta = \left(\psi_+^\alpha + \psi_-^\alpha\right) / \|\psi\|_\alpha^\alpha$, where $\psi_+ = \max_j(\psi_j \vee 0)$, $\psi_- = \max_j(-\psi_j \vee 0)$, and $\|\psi\|_\alpha^\alpha = \sum_{j\in\mathbb{Z}} |\psi_j|^\alpha$ (see, e.g. Corollary 5.5.3 in Embrechts *et al.* (1997)). We will use iid t-distributed innovations Z_n's where the degrees of freedom parameter is also equal to the tail index α.

• The *moving maxima process* $X = \{X_k\}_{k\in\mathbb{Z}}$ is defined as:

$$X_k := \max_{1\leq i\leq m} a_i Z_{k-i+1}, \quad k \in \mathbb{Z}, \tag{2.5.26}$$

with some coefficients $a_i > 0,\ i = 1, \ldots, m$, and $m \geq 1$, where the Z_k's are iid, positive heavy–tailed random variables with tail exponent α. The extremal index θ of X is: $\theta = \max_{1\leq i\leq m} a_i^\alpha / \sum_{i=1}^m a_i^\alpha$.

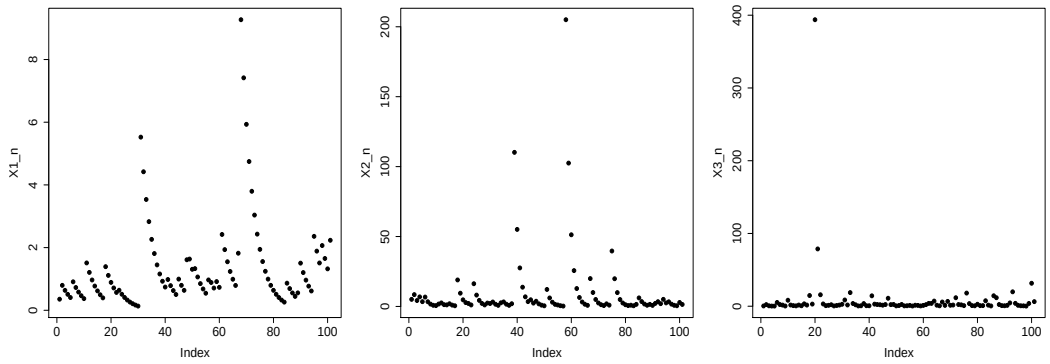

Figure 2.4: armax processes with $\theta = 0.2$, 0.5, and 0.8 and $\alpha = 1$. Note that the lower the extremal index, the higher the degree of clustering.

Simulation setup: In the interest of space, we present selected results for the processes under consideration that demonstrate best the behavior of the various estimators.

◦ $X_n = \max\{bX_{n-1}, (1-b)Z_n\}$, with Z_i iid standard 1−Fréchet.

◦ $Y_n = 0.50Z_n + 0.20Z_{n-1} + 0.10Z_{n-2}$, with Z_i iid t-distributed with α degrees of freedom.

◦ $W_n = \max\{0.80Z_n, 0.20Z_{n-1}, 0.40Z_{n-2}\}$, with Z_i iid Pareto with tail index α.

Figure 2.4 illustrates a few sample paths corresponding to the armax processes above.

• *Parameters:* For the armax processes, we fix the tail index at $\alpha = 1$ and vary the coefficient b to obtain a range of θ values. The coefficients of the linear and moving maxima processes are fixed (as indicated above), and the values of α for the Z_k's are varied to obtain a range of θ values. For all processes, other choices of parameters produced analogous results. For each type of process, 500 independent sample paths were generated of length $2^{13} = 8192$ for the armax and moving max processes and $2^{14} = 16384$ for the linear processes.

• *Estimators:* A range of thresholds corresponding to the largest 10% order statistics were selected for both the Ferro–Segers and the runs estimators. Further, for

the runs estimator, run lengths of 1, 5, and 9 were used.

• *Results:* For each generated sample path, the Ferro–Segers, the runs 1, 5, and 9 at each selected threshold were computed. The proposed *max–spectrum* based estimator was computed using both GLS and WLS and setting $N_{in} = 25$. The threshold (Ferro-Seggers and runs estimators) and the scale (proposed estimator) achieving the *best* Root-Mean-Square-Error (RMSE) is reported in Tables 2.1 – 2.3.

The results demonstrate that the proposed estimator exhibits a good overall performance in terms of RMSE and for many settings outperforms the Ferro-Segers estimator. The GLS and WLS variants produce similar results. The runs estimator performs exceptionally well for the armax process, *if* the "correct" run-length parameter is specified. However, it is quite sensitive to the type of process and to the choice of the run-length parameter. On the other hand, the other two estimators are fairly robust for linear processes and moving maxima, as shown next.

θ	α	GLS	WLS	F/S	$Runs-1$	$Runs-5$	$Runs-9$
0.10	1.00	0.0189	0.0197	0.0140	**0.0109**	0.0127	0.0137
0.20	1.00	0.0226	0.0256	0.0206	**0.0164**	0.0218	0.0247
0.30	1.00	0.0325	0.0291	0.0272	**0.0223**	0.0298	0.0343
0.40	1.00	0.0334	0.0290	0.0306	**0.0272**	0.0381	0.0440
0.50	1.00	0.0335	0.0308	0.0316	**0.0302**	0.0436	0.0520
0.60	1.00	0.0350	**0.0310**	0.0326	0.0316	0.0485	0.0569
0.70	1.00	0.0323	**0.0285**	0.0348	0.0327	0.0493	0.0584
0.80	1.00	0.0274	**0.0243**	0.0365	0.0323	0.0508	0.0638
0.90	1.00	0.0212	**0.0206**	0.0363	0.0284	0.0506	0.0621

Table 2.1: RMSE values for $X_n = \max\{bX_{n-1}, (1-b)Z_n\}$, with Z_i iid standard 1−Fréchet. The first column contains the θ values. The last 6 columns contain the best RMSE values for the max-spectrum estimates via GLS, WLS, and the competitors. The sample sizes were fixed at 2^{13}, with $N_{out} = 500$, and $N_{in} = 25$.

Figure 2.5 illustrates the estimators for 500 independent realizations of a linear process with $\theta = 0.625$. The boxplots for the WLS (GLS boxplots were very similar) method and the median of the estimates of the Ferro–Segers and the runs estimators per threshold are shown. The runs estimator is quite sensitive to the choice of the run–length and exhibits systematic bias. The Ferro–Segers and max–spectrum

θ	α	GLS	WLS	F/S	$Runs-1$	$Runs-5$	$Runs-9$
0.36	0.10	0.0226	0.0291	0.0172	**0.0100**	0.0155	0.0198
0.48	0.50	0.0262	0.0299	0.0204	**0.0181**	0.0322	0.0373
0.63	1.00	0.0328	0.0315	**0.0235**	0.0265	0.0441	0.0509
0.74	1.50	0.0226	**0.0203**	0.0404	0.0333	0.0509	0.0611
0.83	2.00	**0.0147**	0.0238	0.0598	0.0412	0.0576	0.0667
0.89	2.50	0.0032	0.0162	0.0007	0.0003	**0.0000**	0.0007
0.93	3.00	0.0013	0.0004	**0.0002**	0.0043	0.0043	0.0043

Table 2.2: RMSE values for $Y_n = 0.50Z_n + 0.20Z_{n-1} + 0.10Z_{n-2}$, with Z_i iid t-distributed. The first column contains the θ values. The tail index values are in the second column. The last 6 columns contain the best RMSE values for the max-spectrum estimates via GLS, WLS, and the competitors. The sample sizes were fixed at 2^{14}, with $N_{out} = 500$, and $N_{in} = 25$.

θ	α	GLS	WLS	F/S	$Runs-1$	$Runs-5$	$Runs-9$
0.36	0.10	0.0212	0.0287	0.0212	**0.0085**	0.0143	0.0181
0.45	0.50	**0.0244**	0.0311	0.0256	0.0557	0.0274	0.0334
0.57	1.00	**0.0315**	0.0325	0.0329	0.0867	0.0400	0.0474
0.68	1.50	0.0353	**0.0340**	0.0350	0.0844	0.0471	0.0560
0.76	2.00	0.0348	**0.0328**	0.0365	0.0606	0.0482	0.0571
0.83	2.50	**0.0320**	0.0323	0.0378	0.0324	0.0527	0.0625
0.88	3.00	0.0301	0.0297	0.0400	**0.0124**	0.0501	0.0594

Table 2.3: RMSE values for $W_n = \max\{0.80Z_n, 0.20Z_{n-1}, 0.40Z_{n-2}\}$, with Z_i iid Pareto. The first column contains the θ values. The tail index values are in the second column. The last 6 columns contain the best RMSE values for the max-spectrum estimates via GLS, WLS, and the competitors. The sample sizes were fixed at 2^{13}, with $N_{out} = 500$, and $N_{in} = 25$.

estimators are more robust and do not exhibit such strong bias, a fact observed in numerous other experimental settings.

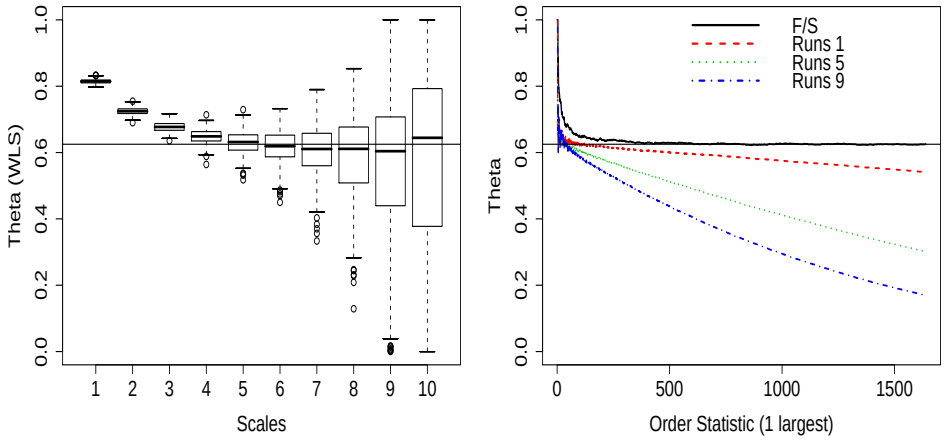

Figure 2.5: WLS simulation results for $Y_n = 0.50Z_n + 0.20Z_{n-1} + 0.10Z_{n-2}$, $\theta = 0.625$, Z_i iid t-distributed with $df = \alpha = 1.00$, and a sample size of 2^{14}, with $N_{out} = 500$ and $N_{in} = 25$. *Left panel:* Boxplots of max-spectrum $\hat{\theta}$. *Right panel:* $\hat{\theta}$ obtained form the runs and Ferro–Segers estimators. In both plots, the solid horizontal line corresponds to $\theta = 0.625$.

			Best Scale			Auto Selection		
Process	θ	α	*RMSE*	*Median*	*SD*	*RMSE*	*Median*	*SD*
AM	0.20	1.00	0.0252	0.22	0.0195	0.0439	0.22	0.0404
AM	0.50	1.00	0.0313	0.52	0.0268	0.0748	0.52	0.0713
AM	0.80	1.00	0.0257	0.81	0.0221	0.0717	0.81	0.0702
LP	0.48	0.50	0.0303	0.49	0.0301	0.0672	0.48	0.0670
LP	0.74	1.50	0.0200	0.76	0.0154	0.0635	0.74	0.0631
LP	0.89	2.50	0.0230	0.87	0.0090	0.0738	0.84	0.0620
MM	0.45	0.50	0.0324	0.47	0.0271	0.0513	0.47	0.0493
MM	0.68	1.50	0.0336	0.69	0.0276	0.0666	0.69	0.0638
MM	0.83	2.50	0.0337	0.85	0.0226	0.0700	0.84	0.0686

Table 2.4: Best RMSE values versus the RMSE from the automatic scale selection procedure.

Automatic selection of scales: We illustrate next the performance of the automatic selection procedure, introduced in Section 2.4. We use a subset of the armax, linear and moving maxima processes, described in the simulation setup above. As before, for each process, we generate 500 independent realizations, of length $2^{13} = 8192$ for the armax (AM) and moving maxima (MM) processes and $2^{14} = 16384$ for the linear processes (LP). We now use $N_{out} = 200$ and $N_{in} = 1$ and thus we obtain 200 *dependent* estimates of θ per scale, for each sample path. We apply the automatic selection procedure based on the Kruskal–Wallis test (at a level of 5%) for each set of 200 resampled θ estimates. We thus obtain a single θ estimate per simulated path.

This procedure is repeated for each independent realization and RMSE values are computed based on the obtained θ estimates from the automatic procedure. We report the best RMSE value (lowest RMSE value among scales), the median and the standard deviation of the estimates based on the automatic procedure and the same values corresponding to the scale at which the best RMSE value was obtained (as in Tables 2.1–2.3).

Table 2.4 indicates that the automatic selection procedure performs very well in terms of bias (as compared to the best–RMSE scale). The RMSE values for the automatic selection method are larger than the best-scale-RMSE values. This is due to the larger variance as seen from the reported standard deviations. Such a

behavior is to be expected since the automatic selection procedure does not involve any knowledge of the true value of θ. In practice, since θ is unknown, one cannot identify the best scale j and hence one cannot achieve the best–RMSE. In such a setting the automatic selection procedure appears to perform well, by producing estimates with low bias and paying a small price in higher variability.

2.5.3 Confidence Intervals

We discuss next two procedures to construct confidence intervals and investigate their coverage probabilities. The first one is based on asymptotic normality (Theorem II.5) and the second on resampling. We examine the coverage probabilities over the same set of processes as in Section 2.5.2.

• *Asymptotic confidence intervals:* For an appropriately chosen scale j in view of Theorem II.5, we propose the following asymptotic $100q\%$ confidence interval:

$$\hat{\theta}(j) \pm z_{(1-q)/2}\hat{\theta}(j)\pi\sqrt{1/6n_j}, \tag{2.5.27}$$

where $z_{(1-q)/2}$ is a $(1+q)/2-$th quantile of the standard normal distribution and n and $n_j = \lfloor n/2^j \rfloor$ are the total sample size and the number of block–maxima involved in the calculation of the Y_j statistic, respectively. Table 2.5 displays coverage probabilities for nominal levels .05 and .10 for scales j between 4 and 8, where the $\hat{\theta}(j)$ estimates typically stabilize. These results are based on 500 independent realizations for each process.

• *Resampling–based confidence intervals:* These are based on the estimated θ's from many shuffled (resampled) versions of a single sample path. The computed θ estimates are pooled across a range of scales with reasonable estimates, and then take the appropriate empirical quantiles:

$$(\hat{\theta}(j_1, j_2)_{(\frac{1-q}{2})},\ \hat{\theta}(j_1, j_2)_{(\frac{1+q}{2})}), \tag{2.5.28}$$

			90% - Scales					95%- Scales				
Process	θ	α	4	5	6	7	8	4	5	6	7	8
AM	0.20	1.00	36	72	82	85	89	48	84	90	93	96
AM	0.50	1.00	88	96	96	96	96	94	99	98	99	99
AM	0.80	1.00	99	99	99	99	98	100	100	100	100	99
LP	0.48	0.50	56	81	80	72	65	68	88	85	78	70
LP	0.74	1.50	94	90	88	84	79	98	95	93	89	83
LP	0.89	2.50	49	80	90	89	86	62	87	93	93	89
MM	0.45	0.50	68	95	99	99	99	82	98	100	100	100
MM	0.68	1.50	93	99	99	100	100	98	100	100	100	100
MM	0.83	2.50	99	99	100	100	99	100	100	100	100	99

Table 2.5: Coverage probabilities for a selected set of processes using equation (2.5.27).

			90% - Scales					95%- Scales				
Process	θ	α	4	5	6	7	8	4	5	6	7	8
AM	0.20	1.00	10	33	37	34	31	13	38	43	40	34
AM	0.50	1.00	34	58	62	61	61	40	66	69	67	68
AM	0.80	1.00	75	79	79	80	81	83	85	86	88	87
LP	0.48	0.50	31	61	58	56	53	36	69	66	64	60
LP	0.74	1.50	79	75	75	71	74	86	82	82	80	80
LP	0.89	2.50	20	57	75	82	83	28	65	82	90	90
MM	0.45	0.50	17	55	68	74	79	20	63	75	81	87
MM	0.68	1.50	31	67	78	81	83	37	79	86	88	90
MM	0.83	2.50	60	81	84	85	84	70	88	92	91	89

Table 2.6: Coverage probabilities for a selected set of processes using equation (2.5.28).

where $\hat{\theta}(j_1, j_2)_{(\tau)}$ represents the empirical $\tau-$th quantile of the pooled $\hat{\theta}(j)$ values across scales $j_1 \leq j \leq j_2$. The coverage probabilities based on (2.5.28) are reported in Table 2.6.

• *Empirical confidence intervals based on Normality:* To further assess the validity of the asymptotic results, we produced empirical confidence intervals based on 500 independent estimates of θ under the assumption that the $\hat{\theta}(j)$'s follow approximately a normal distribution. Namely, we estimate the standard error of the $\hat{\theta}(j)$ and compute $100q\%-$confidence intervals as follows:

$$\hat{\theta}(j) \pm z_{(1-q)/2}\text{S.E.}(\hat{\theta}(j)). \tag{2.5.29}$$

Table 2.7 indicates that the empirical confidence intervals based on independent samples and normality assumptions have relatively accurate coverages. This shows that the asymptotic approximation in Theorem II.5 is in fact applicable. This fact is further confirmed by the coverages reported in Table 2.5. The coverage probabilities

			90% - Scales					95%- Scales				
Process	θ	α	4	5	6	7	8	4	5	6	7	8
AM	0.20	1.00	62	88	92	89	87	62	88	92	89	87
AM	0.50	1.00	79	90	94	94	94	79	90	94	94	94
AM	0.80	1.00	91	94	95	94	98	91	94	95	94	98
LP	0.48	0.50	67	91	89	87	86	81	95	91	91	91
LP	0.74	1.50	89	90	89	90	89	94	95	97	95	95
LP	0.89	2.50	45	76	86	91	90	56	85	93	94	93
MM	0.45	0.50	52	82	85	89	89	66	90	92	95	94
MM	0.68	1.50	64	84	90	90	90	75	92	94	95	94
MM	0.83	2.50	79	89	90	88	97	87	93	95	93	98

Table 2.7: Coverage probabilities for a selected set of processes using equation (2.5.29).

therein are near their nominal levels but the confidence intervals based on equation (2.5.27) are generally wider. In a real data analysis one could only use equation (2.5.27) and not equation (2.5.29) since no independent realizations of the process are available.

As expected, the coverage probabilities in Table 2.6 are not as accurate as the ones obtained from the 500 independent values of θ and the resulting confidence intervals tend to undercover θ, on the average. Nevertheless, this method together with the asymptotic confidence intervals in (2.5.27) yields useful confidence lower bound in practice.

Tables 2.5 – 2.7 indicate coverage probabilities for the middle range of scales. Figure 2.6 illustrates further these coverages for confidence intervals based on (2.5.27) – (2.5.29). For lower scales, the coverage probabilities suffer substantially due to bias; however, as j increases all methods rapidly improve. The asymptotic confidence intervals based on (2.5.27) generally provide reasonable coverage results for scales j beyond 3 or 4.

Ferro and Segers (2003) compare the coverage probabilities of their estimator and the runs estimators. Their estimator is shown to be quite robust and the confidence intervals have relatively accurate coverage probabilities for thresholds corresponding to the 0.90th and up to the 0.99th quantiles. The runs perform very poorly, due to

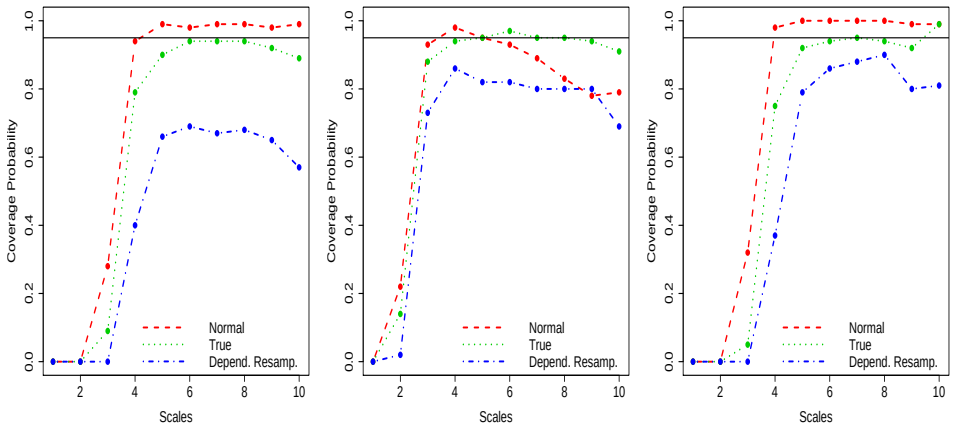

Figure 2.6: Coverage probabilities (for 95% confidence intervals) based on equations (2.5.27) designated as *normal*, (2.5.29) designated as *true*, and (2.5.28) designated as *dependent resampling* are shown. The left, center and right panels correspond to an armax, linear, and moving maxima processes, respectively (with corresponding θ values 0.50, 0.74, and 0.68). The solid line in all three plots corresponds to the nominal value of 0.95.

the large systematic bias of the runs estimator.

There is a heuristic correspondence between the scales j and the quantiles of the data. Namely, the block–maxima on scale j involve roughly the largest $1/2^j$ proportion of the data. Thus, the qth quantile corresponds to a scale of $j \approx -\log_2(1-q)$. Hence, the range of 0.90th to 0.99th quantile in the Ferro–Segers estimators corresponds to scales j between 3 and 7. This and Figure 2.6 suggest that the asymptotic confidence intervals perform relatively well and are comparable to the Ferro–Segers estimator over its corresponding range of thresholds.

2.6 Applications

Crude Oil Data: The daily log returns of West Texas Intermediate (WTI) crude oil prices from January 2, 1986 to March 6, 2007 (5342 observations) are analyzed and the extremal index estimated. Note that the daily log returns (referred as *returns* henceforth) are approximately equal to the daily percentage changes in the price.

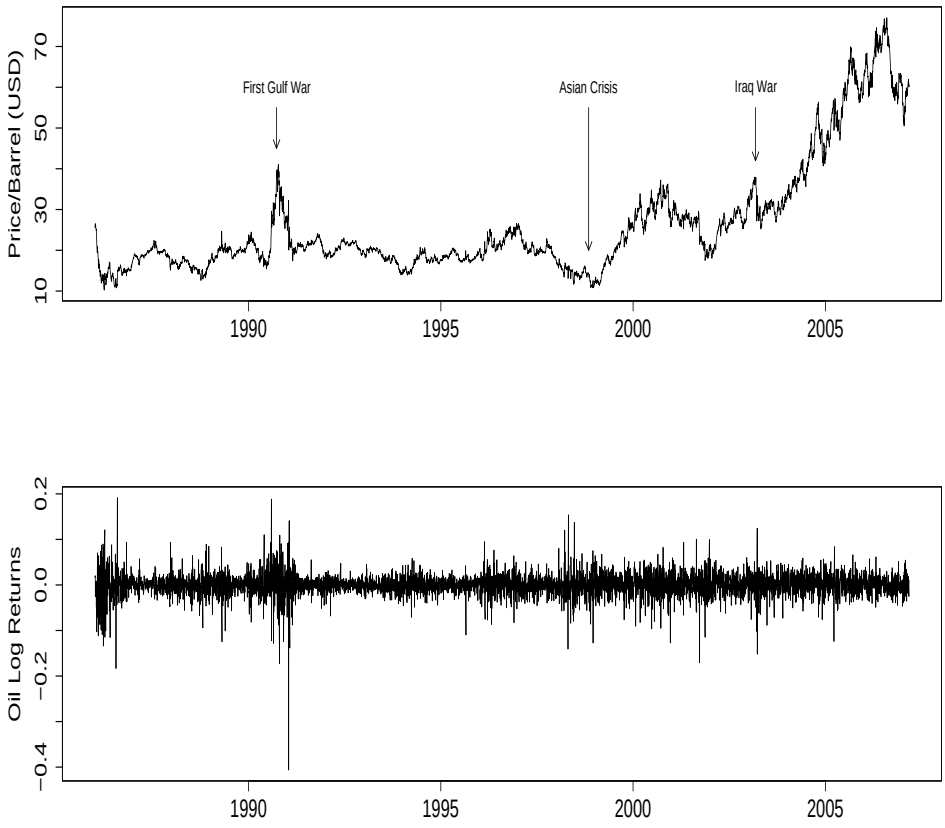

Figure 2.7: Top Plot: West Texas Intermediate (WTI) crude oil prices from January 2, 1986 to March 6, 2007. Bottom Plot: The daily log returns of oil prices for the same period.

WTI represents a benchmark against which all oil bound for the US is priced at and hence its market is deep and liquid. The data were obtained from Energy Information Administration (see http://www.eia.doe.gov/). For a useful reference on oil markets see Geman (2005).

Figure 2.7 shows a plot of the data and the corresponding returns. The return series appears to be approximately stationary, with the exception of a few instances, the result of events of major economic impact. In the top panel, the run up of the oil prices before the first Persian Gulf war can be seen, together with its subsequent rapid drop once it became apparent that the coalition forces would prevail. A similar

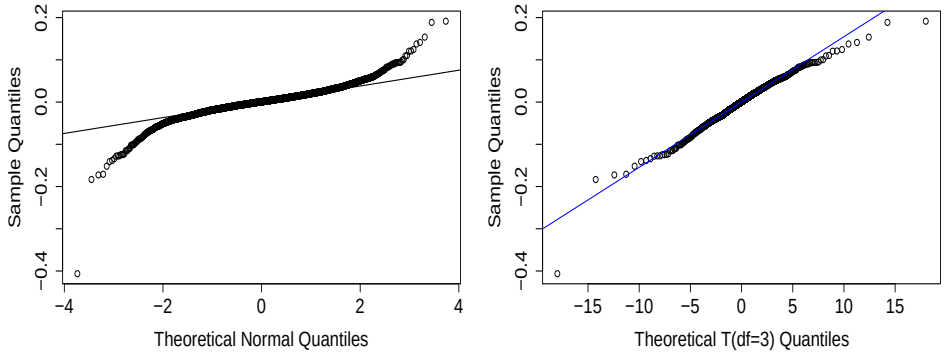

Figure 2.8: Q-Q plots of the log–returns of WTI crude oil prices versus normal (left panel) and t-distribution with df=3 (right panel).

pattern is observed at the onset of the recent Iraq war. The run up in oil prices over the course of the last three years is also evident in the plot.

We address next the question of whether the data exhibit heavy tailed behavior. In general, returns are assumed to be normally distributed. In Figure 2.8 q-q plots of the returns versus the normal distribution and a t-distribution with 3 degrees of freedom are shown. The first plot indicates a strong departure from normality in the tails, while the second one fits the data fairly well, suggesting that the tail index is about 3.

For the remainder, separate analyses are carried out for the positive (right tail of the distribution) and negative (left tail) returns. This is motivated by the empirical fact that positive and negative returns exhibit different behavior. The presence of heavy tailed marginals was confirmed by estimating the tail index using the max-spectrum and Hill estimators. Values of $\alpha \approx 2.8$ and 3 were obtained for the right and left tails, respectively.

We estimate next the *extremal index* θ of the returns using the max-spectrum, the runs 1, 5, 9 and the Ferro–Segers estimators. The results are shown in Figure 2.9.

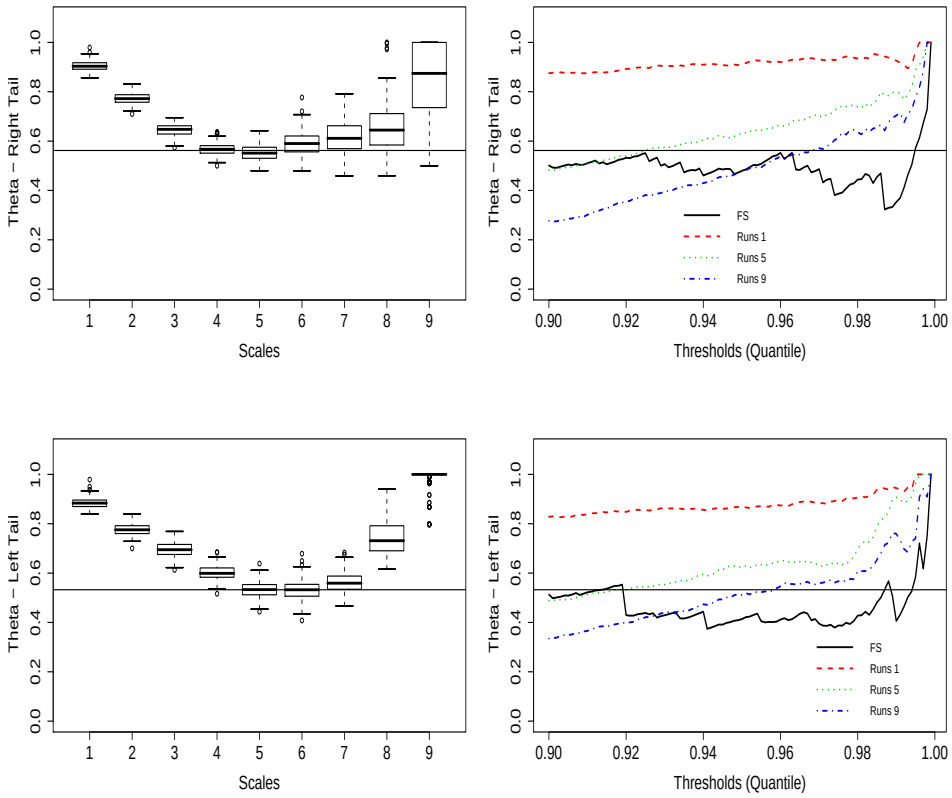

Figure 2.9: Top Row: Estimates of θ for the right tail. The left panel is the max spectrum estimates. The right panel is the Ferro–Segers and runs estimates. The solid horizontal line in both plots corresponds to the max spectrum point estimate of 0.56. Bottom Row: Estimates of θ for the left tail. The left panel is the max spectrum estimates. The right panel is the Ferro–Segers and runs estimates. The solid horizontal line in both plots corresponds to the max spectrum point estimate of 0.53.

The max-spectrum estimates of θ were obtained by setting $N_{out} = 200$ and $N_{in} = 1$ and using WLS. It can be seen that stable θ estimates for the right tail can be obtained at scales $j = 4$ to $j = 5$. Pooling these results yield a value for $\theta = 0.56$ with a 95% confidence interval of (0.50, 0.63) based on equation (2.5.28). It should be noted that the automatic selection procedure chooses scale $j = 5$, which gives comparable results. The 95% confidence interval obtained from (2.5.27) is (0.54, 0.58). The main reason that these confidence intervals are narrow is because they ignore the uncertainty regarding scale selection. For the left tail, we choose the

median value at scales $j = 5$ and to obtain a pooled estimate of 0.53 with a 95% confidence interval of (0.47, 0.61) using (2.5.28) and (0.51, 0.55) using (2.5.27).

A reasonably stable estimate obtained from the Ferro–Segers procedure is around 0.51 for the right tail and 0.42 for the left one. However, another choice for the left tail is 0.50, corresponding to the range of 0.90th to 0.92nd quantiles. The max-spectrum and Ferro–Segers estimates are to some extent in agreement for the right tail and possibly for the left tail as well, depending on the choice of a stable range for the Ferro–Segers estimate. On the other hand, the results of the runs-1 estimator are highly suspect. The results of the runs-1 indicate little or no clustering of extremes (as $\hat{\theta} \approx 1$). The fact that runs-1 fails to capture the clustering may be explained by the behavior of financial returns, where one extremely large positive return is commonly followed by a large negative return. Thus runs-1 often identifies clusters with a single extreme value, as in the case of independent data. Increasing the number of the run length parameter yields estimates more in agreement with the other two procedures. The results strongly suggest clustering of large losses and gains that can in turn have serious consequences in terms of risk exposure of portfolios that include WTI.

Wooster Temperature Data: We now examine the daily minimum temperature series for Wooster, OH, available through the `ismev` **R** package. This data set covers the years 1983 to 1987 and has served as a benchmark for several previous studies of the extremal index. We focus in our analysis only on the winter periods (December 21-March 21) and retain a total of 456 observations. The sign of the data values has been reversed, so that large values represent extremely low temperatures. The series is approximately stationary and heavy tailed over the chained winters. Coles (2001) using the runs estimator for various thresholds and runs lengths obtains estimates in the range of 0.16 to 0.42. Ferro and Segers (2003) using their method obtain

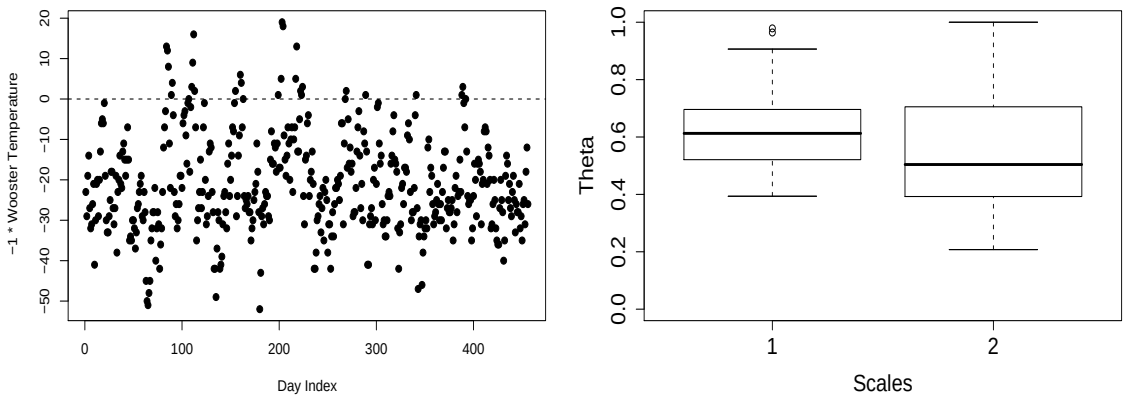

Figure 2.10: Left Panel: Wooster winter 1983-1987 daily minimum temperatures. Right Panel: Max spectrum estimates of θ. Note only 26 point, all point above the dashed line in the left panel figure, were used to estimating θ.

estimates in the range of 0.4 to 0.6 depending on the threshold, but settle on the final value of 0.60 for a stable point estimate. Laurini and Tawn (2003) using their proposed extremal index estimator obtain estimates ranging from 0.40 to 0.65 with a final estimate of 0.65. See also Smith *et al.* (1997) for a more detailed analysis of this data set.

For our analysis we set $N_{in} = 200$, and $N_{out} = 1$ and use WLS. The sample size of 26 - only 26 observations out of total sample size of 456 are positive - used for the max spectrum estimator is quite small in this case. With only 2 scales to examine, we report our results for each scale and the pooled version. Pooling the results from scales 1 and 2, we obtain a point estimate of 0.58, with a 95% confidence interval of (0.27,1) based on equation (2.5.28). The point estimate and the corresponding 95% confidence interval based on equation (2.5.28) for scale 1 is 0.61 and (0.42,0.88) correspondingly. For scale $j = 2$ the results are 0.50 and (0.26,1). This interval is fairly wide and hence its usefulness rather doubtful. The confidence intervals based on equation (2.5.27) are (0.61, 0.63) and (0.54, 0.58) for scales 1 and 2 respectively.

Our results are consistent with the results of previous studies.

2.7 Appendix

2.7.1 Rates of convergence for moment functionals of dependent maxima

Proposition II.6. *Suppose that $f : (0, \infty) \to \mathbb{R}$ is an absolutely continuous function on any compact interval $[a, b] \subset (0, \infty)$, and such that $f(x) = f(x_0) + \int_{x_0}^{x} f'(u)du$, $x > 0$ for some (any) $x_0 > 0$.*

Let for some $m \in \mathbb{R}$ and $\delta > 0$,

$$x^m|f(x)| + \text{esssup}_{0<y\leq x} y^m|f'(y)| \longrightarrow 0, \qquad \text{as } x \downarrow 0, \tag{2.7.30}$$

$$x^{-\alpha}|f(x)| + x^{1+\delta}\text{esssup}_{y\geq x} y^{-\alpha}|f'(y)| \longrightarrow 0, \qquad \text{as } x \to \infty. \tag{2.7.31}$$

Suppose also that the time series $X = \{X_n\}_{n\in\mathbb{Z}}$ satisfies Conditions 1 and 2, where $c_1(x)$ is such that:

$$\int_1^\infty c_1(x)x^{-\alpha}|f'(x)|dx < \infty. \tag{2.7.32}$$

Then, $E|f(M_n)| < \infty$, for all sufficiently large $n \in \mathbb{N}$, and for some $C_f > 0$, independent of n,

$$|\mathbb{E}f(M_n/n^{1/\alpha}) - \mathbb{E}f(Z)| \leq C_f n^{-\beta}, \tag{2.7.33}$$

where Z is an $\alpha-$Fréchet variable with scale coefficient $\sigma := c_X^{1/\alpha}$.

Proof: The proof is similar to the proof of Theorem 3.1 in Stoev *et al.* (2006). Indeed, as in the above reference, one can show that $\mathbb{E}|f(Z)| < \infty$ and $\mathbb{E}|f(M_n)| < \infty$, for all sufficiently large n. Further, by using the conditions (2.7.30) and (2.7.31) and integration by parts, we have that

$$\mathbb{E}f(M_n/n^{1/\alpha}) - \mathbb{E}f(Z) = \int_0^\infty (G(x) - F_n(x))f'(x)dx, \tag{2.7.34}$$

where $F_n(x) := \mathbb{P}\{M_n/n^{1/\alpha} \leq x\}$ and $G(x) = \mathbb{P}\{Z \leq x\}$. Since $F_n(x) = e^{-c(n,x)x^{-\alpha}}$, by the mean value theorem, we have

$$\begin{aligned} |G(x) - F_n(x)| &= |e^{-c_X x^{-\alpha}} - e^{-c(n,x)x^{-\alpha}}| \leq |c(n,x) - c_X| x^{-\alpha} e^{-\min\{\theta c_X, c(n,x)\}x^{-\alpha}} \\ &\leq n^{-\beta} c_1(x) x^{-\alpha} \Big(e^{-c_2 x^{-(\alpha-\gamma)}} + e^{-\theta c_X x^{-\alpha}} \Big), \end{aligned}$$

where in the last inequality, we used Relations (2.3.14) and (2.3.15).

Thus, by (2.7.34), we have that

$$\begin{aligned} |\mathbb{E}f(M_n/n^{1/\alpha}) - \mathbb{E}f(Z)| &\leq n^{-\beta} \int_0^\infty c_1(x) x^{-\alpha} |f'(x)| \Big(e^{-c_2 x^{-(\alpha-\gamma)}} + e^{-c_X x^{-\alpha}} \Big) dx \\ &=: n^{-\beta} \Big(\int_0^1 + \int_1^\infty \Big). \end{aligned} \tag{2.7.35}$$

The last integral is finite. Indeed, since the exponential terms above are bounded, Relation (2.7.32) implies that the integral "$\int_1^\infty$" is finite. On the other hand, conditions (2.3.14) and (2.7.30) imply that, $c_1(x)|f'(x)| = \mathcal{O}(x^{-R}),\ x \downarrow 0$, for some $R \in \mathbb{R}$. However, for all $p > 0$, we have $(e^{-c_2 x^{-(\alpha-\gamma)}} + e^{-c_X x^{-\alpha}}) = o(x^p),\ x \downarrow 0$, since $\alpha - \gamma > 0$. This implies that the integral in "$\int_0^1$" in (2.7.35) is also finite. This completes the proof of (2.7.33). $\square$

Proposition II.7. *Let $X = \{X_k\}_{k\in\mathbb{Z}}$ be a strictly stationary time series which satisfies Conditions 1 and 2 in Section 2.3.1 above. Suppose that $\int_1^\infty c_1(x) x^{-\alpha-1+\delta} dx < 0$, for some $\delta > 0$.*

Then, with $M_n := \max_{1\leq k\leq n} X_k$, we have $\mathbb{E}|\ln(M_n)|^p < \infty$, for all $p > 0$ and all sufficiently large $n \in \mathbb{N}$. Moreover, for any $p > 0$ and $k \in \mathbb{N}$, we have:

$$\Big| \mathbb{E}|\ln(M_n/n^{1/\alpha})|^p - \mathbb{E}|\ln(Z)|^p \Big| = \mathcal{O}(n^{-\beta}), \quad \textit{and} \quad \Big| \mathbb{E}(\ln(M_n/n^{1/\alpha}))^k - \mathbb{E}(\ln(Z))^k \Big| = \mathcal{O}(n^{-\beta}),$$

as $n \to \infty$, where Z is an $\alpha-$Fréchet random variable with scale coefficient $\theta^{1/\alpha} c_X^{1/\alpha}$.

Proof: It is enough to show that the functions $f(x) := |\ln(x)|^p$ and $f(x) := (\ln(x))^k,\ p > 0,\ k \in \mathbb{N}$ satisfy the conditions of Proposition II.6. In the first case, for example, $|f'(x)| = px^{-1}|\ln(x)|^{p-1},\ x > 0$. Therefore, the assumption $\int_1^\infty c_1(x)x^{-\alpha-1+\delta}dx < \infty$ implies (2.7.32), since $|\ln(x)|^{p-1} \leq \text{const}\, x^\delta$, for all $x \in [1,\infty)$. The conditions (2.7.30) and (2.7.31) are also fulfilled in this case, and hence Proposition II.6 yields the desired order of convergence. The functions $f(x) = (\ln(x))^k,\ k \in \mathbb{N}$ can be treated similarly. □

Note that Proposition II.7 readily implies:

$$\mathbb{E}(Y_j - j/\alpha) \equiv \mathbb{E}\log_2(D(j,k)/2^{j/\alpha}) = \mathbb{E}\log_2(\theta^{1/\alpha}c_X^{1/\alpha}Z_1) + \mathcal{O}(1/2^{j\beta}), \tag{2.7.36}$$

as $j \to \infty$, where Z_1 is a standard $\alpha-$Fréchet variable. This important fact is used in the proofs of the asymptotic results given below.

In the remainder of the section we illustrate that Conditions 1 and 2 in Section 2.3.1 apply to a general class of moving maxima processes.

Let $\{Z_n\}_{n\in\mathbb{N}}$ be a sequence of iid random variables with cumulative distribution function $P\{Z \leq z\} = F_Z(z)$. As in Stoev *et al.* (2006), for a function $c(z) > 0$ such that $c(z) \to \text{constant} > 0$, as $z \to \infty$, it is assumed that

$$F_Z(z) = \exp\{-c(z)z^{-\alpha}\}, \quad z > 0, \tag{2.7.37}$$

and two further conditions, analogous to Conditions 1 and 2, are imposed:

Condition 1′. *There exists $\beta' > 0$, such that*

$$|c(z) - c_Z| \leq Kz^{-\beta'}, \quad \textit{for all } z > 0, \tag{2.7.38}$$

where $c_Z > 0$ and $K \geq 0$.

Condition 2′. *$F_Z(0) = 0$ and for all $x > 0$,*

$$c(z) \geq c\min\{1, z^\gamma\}, \quad \textit{for some } \gamma \in (0,\alpha), \tag{2.7.39}$$

with $c > 0$.

Observe that (2.7.38) implies $c(z) \to c_Z,\ z \to \infty$, and in fact

$$\mathbb{P}\{Z > z\} = 1 - F_Z(z) \sim c_Z z^{-\alpha}, \quad \text{as } z \to \infty.$$

The following result shows that the moving maxima process X defined in (2.5.26) satisfies conditions Conditions 1 & 2.

Proposition II.8. *If the* Z_n*'s satisfy Conditions* $1'$ *and* $2'$*, then the process* $X = \{X_k\}_{k\in\mathbb{Z}}$ *in (2.5.26) satisfies Conditions 1 and 2 with* γ *as in (2.7.39),*

$$\beta = \min\{1, \beta'/\alpha\} \quad \textit{and} \quad c_1(x) := \text{const}(1 + x^{-\beta'}), \tag{2.7.40}$$

where β' *is as in (2.7.38). Moreover, the extremal index of* X *is*

$$\theta = \max_{1\le i\le m} a_i^\alpha / \sum_{i=1}^m a_i^\alpha.$$

Proof: We first derive the marginal distribution of the X_k's. By (2.7.37) and (2.5.26), we have

$$\mathbb{P}\{X_k \le x\} = P\{Z_k \le x/a_1, \ldots, Z_{k-m+1} \le x/a_m\} = \exp\{-\sum_{i=1}^m c(x/a_i) a_i^\alpha x^{-\alpha}\}.$$

Thus, in view of (2.7.38), $c(x/a_i) \to c_Z,\ x \to \infty$, and hence, as $x \to \infty$

$$\mathbb{P}\{X_k > x\} \sim \sigma_0^\alpha x^{-\alpha}, \quad \text{where } \sigma_0^\alpha := c_Z \sum_{i=1}^m a_i^\alpha. \tag{2.7.41}$$

We now focus on the maxima $M_n := \max_{1\le i\le n} X_i$. For $n > m$, and $x > 0$, we have that $F_n(x) := \mathbb{P}\{M_n/n^{1/\alpha} \le x\}$ equals

$$\begin{aligned} F_n(x) &= \mathbb{P}\{X_1 \le n^{1/\alpha}x, \ldots, X_n \le n^{1/\alpha}x\} \\ &= \mathbb{P}\Big\{ \bigvee_{j=2-m}^{0} g_{j,m} Z_j \le n^{1/\alpha}x,\ \bigvee_{j=1}^{n-m+1} a_{(1)} Z_j \le n^{1/\alpha}x,\ \bigvee_{j=0}^{m-2} h_j Z_{n-j} \le n^{1/\alpha}x \Big\} \end{aligned}$$

where

$$a_{(1)} := \bigvee_{k=1}^{m} a_k, \qquad g_{j,m} = \bigvee_{k=2-j}^{m} a_k, \qquad h_j = \bigvee_{k=1}^{1+j} a_k.$$

Therefore, by using the independence of the Z_j's and Relation (2.7.37), we get $F_n(x) = \exp\{-c(n,x)x^{-\alpha}\}$, $x > 0$, where

$$\begin{aligned} c(n,x) &= \frac{1}{n}\Big(\sum_{j=2-m}^{0} c(n^{1/\alpha}x/g_{j,m})g_{j,m}^{\alpha} + (n-m+1)a_{(1)}^{\alpha}c(n^{1/\alpha}x/a_{(1)}) \\ &\quad + \sum_{j=0}^{m-2} c(n^{1/\alpha}x/h_j)h_j^{\alpha}\Big). \end{aligned} \tag{2.7.42}$$

We will now show that Relation (2.3.14) holds with β and $c_1(\cdot)$ as in (2.7.40). Let $c_X := c_Z a_{(1)} = c_Z \max_{1\le i\le m} a_i^{\alpha}$. By (2.7.42), we have

$$\begin{aligned} |c(n,x) - c_X| &= |c(n,x) - c_Z a_{(1)}^{\alpha}| \\ &\le \frac{1}{n}\sum_{j=2-m}^{0} \Big|c(n^{1/\alpha}x/g_{j,m}) - c_Z\Big| g_{j,m}^{\alpha} + \\ &\quad \frac{(n-m+1)}{n}\Big|c(n^{1/\alpha}x/a_{(1)}) - c_Z\Big| a_{(1)}^{\alpha} \\ &\quad + \frac{1}{n}\sum_{j=0}^{m-2}\Big|c(n^{1/\alpha}x/h_j) - c_Z\Big| h_j^{\alpha} + \frac{C}{n} =: A_1 + A_2 + A_3 + \frac{C}{n}, \end{aligned} \tag{2.7.43}$$

where the constant C does not depend on x. In the last relation, we add and subtract the finite number of $2(m-1)$ terms of the type $g_{j,m}^{\alpha}c_Z$ and $h_j^{\alpha}c_Z$ and apply the triangle inequality.

Now, by applying Relation (2.7.38) to each one of the absolute value terms in A_1, we obtain

$$A_1 \le \frac{Ka_{(1)}^{\alpha}}{n}\sum_{j=2-m}^{0} n^{-\beta'/\alpha}x^{-\beta'}g_{j,m}^{\beta'} \le \frac{m-1}{n^{1+\beta'/\alpha}}Ka_{(1)}^{\alpha+\beta'}x^{-\beta'} = \frac{C_1}{n^{1+\beta'/\alpha}}x^{-\beta'}, \tag{2.7.44}$$

where the constant C_1 does not depend on n and x and where in the last inequalities we used that $g_{j,m} \le a_{(1)}$. One obtains a similar bound for the term A_3 in (2.7.43):

$$A_3 \le \frac{C_3}{n^{1+\beta'/\alpha}}x^{-\beta'}, \tag{2.7.45}$$

where the constant C_3 does not depend on n and x.

Now, for the term A_2 in (2.7.43), we also have by (2.7.38) that

$$A_2 \le \frac{n-m+1}{n} K a_{(1)}^{\alpha+\beta'} x^{-\beta'} n^{-\beta'/\alpha} \le \frac{C_2}{n^{\beta'/\alpha}} x^{-\beta'}, \tag{2.7.46}$$

where the constant C_2 does not depend on n and x.

By combining the bounds in (2.7.44) – (2.7.46), for the terms in (2.7.43), we obtain

$$|c(n,x) - c_X| \le \frac{(C_1+C_3)}{n^{1+\beta'/\alpha}} x^{-\beta'} + \frac{C_2}{n^{\beta'/\alpha}} x^{-\beta'} + \frac{C}{n},$$

which shows that (2.3.14) holds with $c_1(x) = \text{const}\,(1 + x^{-\beta'})$, where $\beta := \beta'/\alpha$.

We now show that (2.3.15) holds. Since (2.3.15) involves a lower bound, we can ignore the two positive sums in (2.7.42). Recall (2.7.39) and note that $c(n^{1/\alpha}x/a_{(1)}) \ge c_2' \min\{1, (n^{1/\alpha}x/a_{(1)})^\gamma\}$. Since, for sufficiently large n, $n^{1/\alpha} > a_{(1)}$, and $(n^{1/\alpha}x/a_{(1)})^\gamma \ge x^\gamma$, we obtain $c(n^{1/\alpha}x/a_{(1)}) \ge c_2' \min\{1, x^\gamma\}$. Therefore, by (2.7.42), since for all sufficiently large n, $(n-m+1)/n \ge 1/2$, we have $c(n,x) \ge c_2 \min\{1, x^\gamma\}$, where $c_2 = a_{(1)}^\alpha c_2'/2$. This implies (2.3.15) and completes the proof of the proposition. □

2.7.2 Proofs for Section 2.3.2

Proof of Proposition II.2: Recall that by (2.2.3),

$$D(j,k) := \bigvee_{i=1}^{2^j} X_{2^j(k-1)+i} \qquad \text{and introduce} \qquad \widetilde{D}(j,k) := \bigvee_{i=1}^{2^j-m} X_{2^j(k-1)+i}. \tag{2.7.47}$$

Observe that $\widetilde{D}(j,k),\ k = 1, \ldots, n_j$ ($n_j = \lfloor n/2^j \rfloor$) are independent in k since they are "separated by m" block–maxima of the $m-$dependent process X..

Recall also that by (2.2.5)

$$Y_j := \frac{1}{n_j} \sum_{k=1}^{n_j} \log_2 D(j,k) \qquad \text{and introduce the statistics} \qquad \widetilde{Y}_j := \frac{1}{n_j} \sum_{k=1}^{n_j} \log_2 \widetilde{D}(j,k).$$

We first establish Relation (2.3.20). Let

$$\hat{H} = \sum_{i=0}^{\ell} w_i Y_{i+j(n)}, \qquad \text{and} \qquad \widetilde{H} = \sum_{i=0}^{\ell} w_i \widetilde{Y}_{i+j(n)}, \tag{2.7.48}$$

so that $\hat{\alpha}(j)$ in (2.3.17) equals $1/\hat{H}$. The weights w_i's, the range ℓ and the quantity $j(n)$ are described in Section 2.3.1.

To prove that $\hat{\alpha}(j) - \alpha = \mathcal{O}_P(a_n),\ n \to \infty$, for some $a_n \to 0$, it suffices to show that $\mathbb{E}(\hat{H} - H)^2 = \mathcal{O}(a_n^2)$, where $H := 1/\alpha$.. Observe that by adding and subtracting the term $\widetilde{H}$, and by applying the inequality $(a+b)^2 \le 2a^2 + 2b^2$, we get

$$\begin{aligned} \mathbb{E}(\hat{H} - H)^2 \le 2\mathbb{E}(\hat{H} - \widetilde{H})^2 + 2\mathbb{E}(\widetilde{H} - H)^2 \quad &= \quad 2\mathrm{Var}(\hat{H} - \widetilde{H}) \\ &\quad +2(\mathbb{E}\hat{H} - \mathbb{E}\widetilde{H})^2 + 2\mathbb{E}(\widetilde{H} - H)^2 \\ &=: \quad 2A_1 + 2A_2 + 2A_3, \end{aligned} \tag{2.7.49}$$

where in the last relation we also used the fact that $\mathbb{E}\xi^2 = \mathrm{Var}(\xi) + (\mathbb{E}\xi)^2$.

We will first show that $A_1 = o(1/n_j)$ in (2.7.49) is negligible. Indeed, by (2.7.48), we have

$$\hat{H} - \widetilde{H} = \sum_{i=0}^{\ell} w_i (Y_{i+j(n)} - \widetilde{Y}_{i+j(n)}), \tag{2.7.50}$$

and thus by using the inequality $\mathrm{Var}(\xi_0 + \cdots + \xi_\ell) \le (\ell+1)^2(\mathrm{Var}(\xi_0) + \cdots + \mathrm{Var}(\xi_\ell))$, we get $\mathrm{Var}(\hat{H} - \widetilde{H}) \le (1+\ell)^2 \sum_{i=0}^{\ell} w_i^2 \mathrm{Var}(Y_{i+j(n)} - \widetilde{Y}_{i+j(n)})$. Thus, by Lemma II.9 below, since ℓ is fixed,

$$\mathrm{Var}(\hat{H} - \widetilde{H}) \le \frac{\mathrm{const}}{n_j} \sum_{i=0}^{\ell} \mathrm{Var}\Big(\log_2 D(i + j(n), 1) - \log_2 \widetilde{D}(i + j(n), 1)\Big), \tag{2.7.51}$$

where $n_j = n/2^{j(n)}$. Lemmas II.10 and II.11, on the other hand, yield

$$\mathrm{Var}(\hat{H} - \widetilde{H}) = o(1/n_j), \qquad \text{as } n \to \infty. \tag{2.7.52}$$

Now, we focus on the term A_2 in (2.7.49). By (2.7.50), we have

$$\begin{aligned}\sqrt{A_2} &= \sum_{i=0}^{\ell} w_i(\mathbb{E}Y_{i+j(n)} - \mathbb{E}\widetilde{Y}_{i+j(n)}) = \sum_{i=0}^{\ell} w_i \log_2(D(i+j(n),1)/2^{(i+j(n))/\alpha}) \\ &\quad - \sum_{i=0}^{\ell} w_i \log_2(\widetilde{D}(i+j(n),1)/2^{(i+j(n))/\alpha}) \\ &= \sum_{i=0}^{\ell} w_i \mathbb{E}\log_2(Z) + o(1/2^{j(n)\beta}) \\ &\quad - \sum_{i=0}^{\ell} w_i \Big(\mathbb{E}\log_2(\widetilde{D}(i+j(n),1)/(2^{(i+j(n))} - m)^{1/\alpha}) \\ &\quad - \frac{1}{\alpha}\log_2((2^{i+j(n)} - m)/2^{i+j(n)}) \Big),\end{aligned}$$

where the last relation follows from (2.7.36) and where Z is an $\alpha-$Fréchet variable with scale coefficient $(\theta c_X)^{1/\alpha}$. Now, since $\widetilde{D}(i+j(n),1)/(2^{i+j(n)}-m)^{1/\alpha}$ is a properly normalized block–maximum (recall (2.7.47) above), by Relation (2.7.36), we further have that

$$\begin{aligned}\sqrt{A_2} &= \sum_{i=0}^{\ell} w_i \mathbb{E}\log_2(Z) - \sum_{i=0}^{\ell} w_i \mathbb{E}\log_2(Z) + o(1/2^{j(n)\beta}) + \mathcal{O}(\log_2(1 - m/2^{j(n)}) \\ &= o(1/2^{j(n)\beta}) + \mathcal{O}(1/2^{j(n)}),\end{aligned}$$

as $r \to \infty$, since $\log_2(1-x) = \mathcal{O}(x),\ x \to 0$. We thus have,

$$A_2 = \mathcal{O}(1/2^{j(n)\min\{1,\beta\}}), \quad \text{as } j(n) \to \infty. \tag{2.7.53}$$

Consider now the term A_3 in (2.7.49). As above, we have

$$\mathbb{E}(\widetilde{H} - H)^2 = \mathrm{Var}(\widetilde{H} - H) + (\mathbb{E}\widetilde{H} - H)^2 =: A_3' + A_3'',$$

and as in (2.7.51), we get

$$A_3' \le (\ell+1)^2 \sum_{i=0}^{\ell} w_i \mathrm{Var}(\widetilde{Y}_{i+j(n)}) = o(1/n_j) = o(2^{j(n)}/n), \quad \text{as } n_j \to \infty.$$

Also, as argued above, since $\sum_{i=0}^{\ell} w_i(i+j(n))/\alpha = 1/\alpha \equiv H$, we obtain

$$\mathbb{E}\widetilde{H} - H = \sum_{i=0}^{\ell} w_i(\mathbb{E}\log_2 \widetilde{D}(i+j(n),1) - (i+j(n))/\alpha) = \mathcal{O}(1/2^{j(n)\min\{1,\beta\}}),$$

as $j(n) \to \infty$ (see (2.7.53) above). By combining the bounds for terms A_1 , A_2 and A_3 in (2.7.52), (2.7.53) and the last two relations, we obtain

$$\hat{H} = H + \mathcal{O}_P(1/2^{j(n)\min\{1,\beta\}}) + \mathcal{O}_P(2^{j(n)/2}/n^{1/2}), \qquad \text{as } j(n),\ n/2^{j(n)} \to \infty.$$

This completes the proof (2.3.20).

The proof of (2.3.21) is simpler. By introducing the quantity $\widetilde{C}(j) := \widetilde{Y}_j - j/\alpha$, we have

$$C(j) - \widetilde{C}(j) = Y_j - \widetilde{Y}_j = \frac{1}{n_j}\sum_{k=1}^{n_j} \log_2(D(j,k)/\widetilde{D}(j,k)).$$

One can similarly show that $\mathrm{Var}(C(j) - \widetilde{C}(j))$ is of order $o(1/n_j)$, as $n \to \infty$. Thus, the order of $C(j) - C$ is dictated by the orders of the *bias* and *standard error* for the quantity $\widetilde{C}(j)$. These can be handled as the terms A_2 and A_3 in (2.7.49). □

The following three lemmas were used in the proof Proposition II.2.

Lemma II.9. *Under the conditions of Proposition II.2, for all $j > \log_2 m$, we have*

$$\mathrm{Var}(Y_j - \widetilde{Y}_j) \le \frac{3}{n_j}\mathrm{Var}(\log_2(D(j,1)/\widetilde{D}(j,1))).$$

Proof: For notational simplicity, let $\xi_k := \log_2(D(j,k)/\widetilde{D}(j,k)),\ k = 1,\ldots,n_j$. We have, by the stationarity of ξ_k in k, that

$$\mathrm{Var}(Y_j - \widetilde{Y}_j) = \frac{1}{n_j}\mathrm{Var}(\xi_1) + \frac{2}{n_j^2}\sum_{k=1}^{n_j-1}(n_j - k)\mathrm{Cov}(\xi_{k+1}, \xi_1).$$

Note that $\xi_{k+1} = \log_2(D(j,1+k)/\widetilde{D}(j,1+k))$ and $\xi_1 = \log_2(D(j,1)/\widetilde{D}(j,1))$ are independent if $k > 1$. Indeed, this follows from the fact that the process X is $m-$dependent, and since ξ_{k+1} and ξ_1 depend on blocks of the data separated by at

least $2^j > m$ lags. Therefore, only the lag–1 covariances in the above sum will be non–zero and hence

$$\mathrm{Var}(Y_j - \widetilde{Y}_j) \leq \frac{1}{n_j}\mathrm{Var}(\xi_1) + \frac{2}{n_j}\Big|\mathrm{Cov}(\xi_2, \xi_1)\Big| \leq \frac{3}{n_j}\mathrm{Var}(\xi_1),$$

since by the Cauchy–Schwartz inequality we have $|\mathrm{Cov}(\xi_2, \xi_1)| \leq \mathrm{Var}(\xi_2)^{1/2}\mathrm{Var}(\xi_1)^{1/2} = \mathrm{Var}(\xi_1)$. This completes the proof of the lemma. □

Lemma II.10. *For $D(j,k)$ and $\widetilde{D}(j,k)$, defined in (2.7.47) above, for any fixed k, we have $D(j,k)/\widetilde{D}(j,k) \xrightarrow{P} 1$, as $j \to \infty$.*

Proof: Let $\delta \in (0, 1/\alpha)$ be arbitrary and observe that

(2.7.54)

$$\mathbb{P}\{D(j,k)/\widetilde{D}(j,k) < 1\} = \mathbb{P}\{R > \widetilde{D}(j,k)\} \leq \mathbb{P}\{R > 2^{j\delta}\} + \mathbb{P}\{2^{j\delta} > \widetilde{D}(j,k)\},$$

where $R = \max_{1\leq i\leq m} X_{2^j(k-i)+1}$. Now, by stationarity,

$$\mathbb{P}\{R > 2^{j\delta}\} = \mathbb{P}\{\max_{1\leq i\leq m} X_i > 2^{j\delta}\} \to 0, \quad \text{as } j \to \infty.$$

On the other hand, Relation (2.3.14) implies that $2^{-j/\alpha}\widetilde{D}(j,k) \xrightarrow{d} Z$, as $n \to \infty$, where Z is a non–degenerate $\alpha-$Fréchet variable. Thus, since $\delta \in (0, 1/\alpha)$, we have that

$$\mathbb{P}\{2^{j\delta} > \widetilde{D}(j,k)\} \to 0, \qquad \text{as } j \to \infty.$$

The last two convergences and the inequality (2.7.54) imply that $\mathbb{P}\{D(j,k)/\widetilde{D}(j,k) < 1\} \to 0$, $j \to \infty$. Since trivially $\mathbb{P}\{D(j,k)/\widetilde{D}(j,k) > 1\} = 1$, we obtain $D(j,k)/\widetilde{D}(j,k)$ converges in distribution to the constant 1, as $j \to \infty$. This completes the proof since convergence in distribution to a constant implies convergence in probability. □

Lemma II.11. *The set of random variables $\left|\log_2\left(D(j,k)/\widetilde{D}(j,k)\right)\right|^p$, $j,k \in \mathbb{N}$ is uniformly integrable, for all $p > 0$, where $D(j,k)$ and $\widetilde{D}(j,k)$ are defined in (2.7.47).*

Proof: Let $q > p$ be arbitrary. By using the inequality $|x+y|^q \le 2^q(|x|^q + |y|^q)$, $x, y \in \mathbb{R}$, we get

$$\mathbb{E}\Big|\log_2 \frac{D(j,k)}{\widetilde{D}(j,k)}\Big|^q \le 2^q\mathbb{E}|\log_2(D(j,k)/2^{j/\alpha})|^q + 2^q\mathbb{E}|\log_2(\widetilde{D}(j,k)/2^{j/\alpha})|^q.$$

In view of Proposition II.7, applied to the block–maxima $D(j,k)$ and $\widetilde{D}(j,k)$, we obtain

$$\mathbb{E}|\log_2(D(j,k)/2^{j/\alpha})|^q = \mathbb{E}|\log_2(M_{2^j}/2^{j/\alpha})|^q \longrightarrow \text{const}, \quad \text{as } j \to \infty.$$

Thus the set $\{\mathbb{E}|\log_2(D(j,k)/2^{j/\alpha})|^q,\ j,k \in \mathbb{N}\}$ is bounded. We similarly have that the set $\{\mathbb{E}|\log_2(\widetilde{D}(j,k)/2^{j/\alpha})|^q,\ j,k \in \mathbb{N}\}$ is bounded since $\log_2(2^j - m) \sim j,\ j \to \infty$, for any fixed m.

We have thus shown that

$$\sup_{j,k\in\mathbb{N}} \mathbb{E}\Big|\log_2 \frac{D(j,k)}{\widetilde{D}(j,k)}\Big|^q < \infty,$$

for $q > p$, which yields the desired uniform integrability. □

Proof of Lemma II.3: Suppose that the indices $i_1, \ldots, i_k$ are drawn without replacement. Let $A_1 = \Omega$ and

$$A_j := \{\omega \in \Omega : |i_{j'}(\omega) - i_{j''}(\omega)| \ge m, \ \text{ for all } j' \ne j'',\ 1 \le j', j'' \le j\}, \tag{2.7.55}$$

for $j \ge 2$, that is, A_j is the event that the first j random indices are spaced further away from each other by at least m lags.

We need to show $\mathbb{P}(A_k) \ge 1 - mk^2/(n-k)$. Note that

$$\mathbb{P}(A_{j+1}) = \mathbb{P}(A_{j+1}|A_j)\mathbb{P}(A_j) \ge (1 - 2mj/(n-j))\mathbb{P}(A_j). \tag{2.7.56}$$

Indeed, the probability $\mathbb{P}(A_{j+1}|A_j)$ of choosing the index i_{j+1} to be within m lags from one of the chosen j indices $i_1, \ldots, i_j$ is at most $2mj/(n-j)$. Thus,

$$\mathbb{P}(A_k) = \prod_{j=1}^{k-1} \mathbb{P}(A_{j+1}|A_j)\mathbb{P}(A_1) \ge \prod_{j=1}^{k-1}(1 - 2mj/(n-j)).$$

Now, by the inequality $\prod_{j=1}^{k-1}(1-x_j) \geq 1-\sum_{j=1}^{k-1} x_j$, valid for all $x_j \in [0,1]$, we obtain

$$\mathbb{P}(A_k) \geq 1 - \sum_{j=1}^{k-1} 2mj/(n-j) \geq 1 - mk(k-1)/(n-k) > 1 - mk^2/(n-k). \tag{2.7.57}$$

The case when the indices are drawn with replacement is similar. □

Proof of Theorem II.4: Consider either a bootstrap or a permutation sample $X_l^* = X_{i_l},\ l = 1, \ldots, k$, where $i_1, \ldots, i_k$ are randomly chosen indices from $\{1, \ldots, n\}$, independently from the original data $X_1, \ldots, X_n$. In the case of bootstrap these indices are chosen with replacement and in the case of permutations – without replacement, respectively.

Let the event A_k be defined as in (2.7.55), which corresponds to the indices being spaced by at least $m-$lags away from each other. Thus, since the time series $X = \{X_i\}_{i\in\mathbb{Z}}$ is $m-$dependent,

$$(X_1^*, \cdots, X_k^*)1_{A_k} \stackrel{D}{=} (\widetilde{X}_1, \cdots, \widetilde{X}_k)1_{A_k},$$

where $\widetilde{X}_l,\ l = 1, \ldots, k$ are iid random variables with the same distribution as the X_n's which are *independent* from the event A_k. Observe that the event A_k is also independent from the time series X since it depends only on the random indices $i_1, \ldots, i_k$. Further, note that in the last relation, we have only *equality in distribution* and not equality almost surely.

Now, by Lemma II.3, we have $\mathbb{P}(A_n) \to 1$, as $n \to \infty$, since $k(n) = o(\sqrt{n})$. Thus, Lemma II.12 implies that any statistic based on the bootstrap or the randomly permuted sample will have the same limiting distribution as the corresponding statistic based on the iid sample $\{\widetilde{X}_l\}_{1\leq l\leq k}$.

Let $\widetilde{C}^*(j) = \widetilde{Y}_j - j/\alpha$ be defined as the quantity $C^*(j)$ in (2.3.19), but where now $\widetilde{Y}_j$ is the max–spectrum based on the iid data $\widetilde{X}_1, \ldots, \widetilde{X}_k$. Theorem 4.1 in Stoev

et al. (2006) implies that

$$\sqrt{k_j}(\widetilde{C}^*(j) - C) \xrightarrow{D} \mathcal{N}(0, \sigma^2_{C^*}), \quad \text{as } k \to \infty, \tag{2.7.58}$$

where $\sigma^2_{C^*}$ is as in Theorem II.4. As argued above, Lemma II.12 and Relation (2.7.58) imply (2.3.22), which completes the proof of the theorem. □

Lemma II.12. *Let X_n, X and Y_n be real random variables such that $X_n \xrightarrow{D} X$, as $n \to \infty$. Let also A_n and B_n be some events such that $Y_n 1_{B_n} \stackrel{D}{=} X_n 1_{A_n}$. If $\mathbb{P}(A_n) = \mathbb{P}(B_n) \to 1$, $n \to \infty$, then $Y_n \xrightarrow{D} X$, as $n \to \infty$.*

Proof: Let $f : \mathbb{R} \to \mathbb{R}$ be an arbitrary bounded and continuous function. Since $\mathbb{E}|f(Y_n)1_{B_n^c}| \leq \text{const}\mathbb{P}(B_n) = o(1)$, as $n \to \infty$, we have

$$\mathbb{E}f(Y_n) = \mathbb{E}f(Y_n)1_{B_n} + o(1) = \mathbb{E}f(X_n)1_{A_n} + o(1) = \mathbb{E}f(X_n) + o(1), \quad \text{as } n \to \infty.$$

This shows that $\lim_{n\to\infty} \mathbb{E}f(Y_n) = \lim_{n\to\infty} \mathbb{E}f(X_n)$, which completes the proof. □

Proof of Theorem II.5: Recall Relation (2.3.18) and observe that by Proposition II.2, we have

$$\hat{\alpha}(j) = \alpha + \mathcal{O}_P(b_n), \qquad \text{and} \qquad C(j) = C + \mathcal{O}_P(b_n),$$

as $n \to \infty$, where

$$b_n = 1/2^{j(k(n))\min\{1,\beta\}} + 2^{j(k(n))/2}/n^{1/2}. \tag{2.7.59}$$

Also, by Theorem II.4, we have $a_n^{-1}(C^*(j) - C^*) \xrightarrow{D} \mathcal{N}(0, \sigma^2_{C^*})$, as $n \to \infty$, where $a_n = 1/\sqrt{k_j} = 2^{j(k(n))/2}/k(n)^{1/2}$. Relation (2.3.23), implies that $b_n = o(a_n)$, $n \to \infty$. Indeed, since $k(n) = o(n)$, $n \to \infty$, we have $2^{j(k(n))/2}/n^{1/2} = o(2^{j(k(n))}/k(n)^{1/2}) \equiv o(a_n)$, as $n \to \infty$. This shows that the second term of b_n in (2.7.59) is negligible with respect to a_n. By Relation (2.3.23), we also have $k/2^{j(k)(1+2\min\{1,\beta\})} \to 0$, as $k \to \infty$,

or, equivalently $1/2^{j(k)\min\{1,\beta\}} = o(2^{j(k)/2}/k^{1/2})$, as $k \to \infty$. Hence, the first term of b_n in (2.7.59) is also of order $o(2^{j(k(n))/2}/k(n)^{1/2}) \equiv o(a_n)$, as $n \to \infty$.

Now, by using the fact that $b_n = o(a_n),\ n \to \infty$ and the 'Delta–method' (see e.g. Theorem 3.1 in van der Vaart (1998)), applied to the function $f(x, y, z) = 2^{x(y-z)}$ and $x_0 = \alpha$, $y_0 = C$ and $z_0 = C^*$ (see also (2.3.18)), we obtain

$$a_n^{-1}(\hat{\theta}(j) - \theta) \xrightarrow{D} \partial_z f(\alpha, C, C^*)\, Z \sim \mathcal{N}(0, \sigma_\theta^2), \qquad \text{as } n \to \infty.$$

Since $\partial_z f(x_0, y_0, z_0) = -\ln(2)\alpha\theta$, we obtain

$$\sigma_\theta^2 = \Big(\partial_z f(\alpha, C, C^*)\Big)^2 \sigma_{C^*}^2 = \ln(2)^2\theta^2 \mathrm{Var}(\log_2(Z)),$$

where Z is a $1-$Fréchet variable (see Theorem II.4).

It remains to show that $\sigma_\theta^2 = \theta^2\pi^2/6$, or equivalently, $\ln(2)^2\mathrm{Var}(\log_2(Z)) = \pi^2/6$. Since $\ln(2)\log_2(x) = \ln(x)$, $x > 0$, we have that

$$\ln(2)^2\mathrm{Var}(\log_2(Z)) = \mathrm{Var}(\ln(Z)) = \int_0^\infty \ln^2(x) de^{-x^{-1}} - \Big(\int_0^\infty \ln(x) de^{-x^{-1}}\Big)^2.$$

By making a change of variables, we obtain

$$\mathrm{Var}(\ln(Z)) = \int_0^\infty \ln^2(u) e^{-u} du - \Big(\int_0^\infty \ln(u) e^{-u} du\Big)^2 = \gamma^2 + \pi^2/6 - (-\gamma)^2 = \pi^2/6,$$

where γ denotes the Euler constant and where the last two integrals are given, for example, by Relations 4.331 (1) and 4.335 (1) on pages 573 and 574 in Gradshteyn and Ryzhik (1965). This completes the proof of the theorem. □

CHAPTER III

Joint Modeling of Extremes and their Clustering with Applications in Finance

3.1 Introduction

In this chapter, we propose a new method to estimate the Value at Risk (VaR) of an asset by explicitly taking into account the extremes of the asset returns and *their arrival intensities*. Quoting Tsay (2005), page 289, VaR's intuitive definition and interpretation follows:

> With probability $1 - q$, the potential loss encountered by the holder of the financial position over the time horizon T is less than or equal to VaR.

For example, suppose the VaR for an asset's daily percentage changes has been estimated to be 2.0% with $q = 0.99$. Suppose that the value of the asset is $\$1,000,000$. This means that a of loss of 2.0%, or $\$20,000$, of the asset value occurs every 100 ($100 = 1/(1 - 0.99)$) days on average. In the non-finance statistical literature, VaR is referred to as the return level.

VaR as a risk measure has been criticized by Artzner *et al.* (1999) for not being a coherent risk measure; One characteristic of a coherent risk measure is that the risk of holding two assets must be less than sum of the risks associated with each asset. VaR can violate this characteristic. However, it still remains a popular and

widely used extreme risk measure since it is easily communicable to the corporate decision makers. Furthermore, regulatory bodies such as the Bank for International Settlements (BIS) have adopted VaR for setting capital requirements. A recent article by Jorion (2007a) discusses the strengths and weaknesses of VaR.

Many risk managers mainly at financial institutions utilize VaR to monitor and control risk. For example, a violation of VaR by a trader's position may alert the risk manager to force the trader to reduce the risk by either hedging, pruning down the position, or even liquidating the position. VaR can be extended or re-applied for other types of risk such as credit risk and operational risk. The discussion and the applications of VaR in operational risk can be found in Panjer (2006). Bluhm *et al.* (2002) applies VaR to credit risk.

Jorion (2007b) provides an encyclopedic treatment of VaR. Other textbook level treatments are given by Dowd (2005), Christoffersen (2003), and Hull (2007). McNeil *et al.* (2005) focus more on the quantitative aspects of VaR, where as Crouhy *et al.* (2006) give a less quantitative and a broader overview of VaR. Dowd and Blake (2006) give an up-to-date account on the more recent advances in the estimation of VaR and other risk measures with emphasis on the insurance applications.

In this chapter, we propose a new method to estimate VaR. We estimate VaR, conditionally on the past and current behavior of the asset and show that this leads to substantial improvements over the benchmark methodology. Out method estimates conditional VaR by modeling the times of the occurrence and the magnitude of the extreme losses with a marked point process. A marked point process is a stochastic process such that some event of interest occurs at random times with additional random variables associated with the occurrence of the event. In our case, large losses occur at random times with random magnitudes. We will assume that the times of

the extreme losses above a fixed threshold are governed by a point process with stochastic intensity, which we estimate from the data. We incorporate the estimated intensity along with the classic Generalized Pareto Distribution (GPD) model for the tail of the loss distribution to estimate VaR. Intuitively, during high intensity periods, the large losses tend to cluster, and thus successive violations of VaR can occur. Such successive violations of VaR are not predicted by the classical theory, which models only the marginal distribution of the loss and ignores the temporal dependence. Our results show that incorporating the intensity into VaR estimation is able to account for this clustering. The stochastic intensity of the point process of the extreme losses is modeled in the general framework of the Autoregressive Conditional Duration (ACD) model first described in Engle and Russell (1998). To the best of our knowledge, with the exception of Focardi and Fabozzi (2005), this is the first time ACD models have been used outside the high frequency domain. Our approach is similar to Chaves-Demoulin *et al.* (2005), except we will estimate the stochastic intensity from the ACD model rather than assume a specific form for it.

The remainder of the chapter is organized as follows: In Section 3.2, we review the two key building components of our modeling strategy which are point processes and tail modeling. In Section 3.3, we derive explicit equations for predicting the next day VaR, based on the past data, and also the future 10 day cumulative excess quantiles. Section 3.4, contains some theoretical results regarding the stationarity of our model. In Section 3.5, we illustrate our methodology on real data sets; specifically, we describe the data and their key features, apply the proposed methodology, and compare the results to two benchmark methods of unconditional quantile estimation and GARCH. We conclude in Section 3.6, and comments on our future work.

3.2 Background Information

In this section, the two main theoretical underpinnings of our model are described in detail. We will only highlight the references most relevant to our work as extensive literature exists in both areas. The first subsection describes point processes, and the second subsection presents the modeling of the excesses over a high threshold via GPD.

3.2.1 Point Processes

A point process is simply a collection of random points in space and/or time. Our focus will be on point processes in time. Each point represents the time of the occurrence of a predefined event. A particularly general definition of a point process is given by Brémaud (1999).

Definition: A random point process on the positive half-line is a sequence $\{T_n\}_{n\geq 0}$ of nonnegative random variables such that, almost surely,

1. $T_0 \equiv 0$,
2. $0 < T_1 < T_2 < \ldots$,
3. $\lim_{n\to\infty} T_n = +\infty$.

The sequence $\{T_n\}_{n\geq 0}$ denotes the times of the events. With each random variable T_i, we can associate another random variable Y_i. The Y_i's are called marks and the two dimensional sequence $\{(T_i, Y_i)\}_{n\geq 0}$ is then called a *marked point process.* We are most often interested in the times between the occurrence of the events, i.e. the inter-arrival times. We will let the sequence $\{X_n\}_{n\geq 0}$, $X_i = T_i - T_{i-1}$, $X_0 = 0$, represent the inter-arrival times of the events of the process. Furthermore, a counting random variable, $N(s,t)$, is associated with the point process, which counts the number of

events in the time interval (s, t). By convention, we let $N(0, t) = N(t)$, when $s = 0$. We will let $\mathcal{H}_t$ represent the history of the process up to time t. Formally, we can define the history of the process as follows: Let $\mathcal{F}_n = \sigma\{(X_i, Y_i), 1 \leq i \leq n\}$, and let $N(t) = \sup\{n : \sum_{i=1}^{n} X_i \leq t\}$. Then $N(t)$ is a stopping time. Define $\mathcal{H}_t = \mathcal{F}_{N(t)}$, where $\mathcal{F}_\tau = \{A \in \mathcal{F}_n : A \cap \{\tau \leq n\} \in \mathcal{F}_n\}$ for a stopping time τ with respect to the filtration $\{\mathcal{F}_n\}_{n \geq 0}$.

Another important function associated with a point process is the intensity function, denoted as $\lambda(t)$. The intensity function $\lambda(t)$ is the instantaneous mean rate of the event occurrence. The units of $\lambda(t)$ are number of events per unit of time as we exclude the possibility of two events occurring simultaneously. The most well known point process is the Poisson process where $\lambda(t)$ is constant. Poisson processes generally provide the starting model for the analysis of most point processes. When $\lambda(t)$ is a deterministic function of time, the point process is called a non-homogeneous or non-stationary Poisson process. A process when $\lambda(t)$ itself is a stochastic process is called a doubly stochastic process or a Cox process. Generally the intensities of Cox processes are driven by some outside stochastic process, which is independent from the point process itself.

Processes such that $\lambda(t)$ is causally dependent on the point process itself are called the *self-exciting* point processes. In such processes, the marks are considered endogenous to the process itself. Detailed analysis of such processes from an applied perspective is in Snyder and Miller (1991). Karr (1991) and Daley and Vere-Jones (2003) provide a measure theoretic (and very dense) coverage. Engle and Russell (1998) give a good summary as well. The seminal results and applications are contained in Hawkes (1971b), Hawkes (1971a), Hawkes (1972), and Hawkes and Oakes (1974). Not surprisingly, self-exciting processes are sometimes referred to as *Hawkes*

processes. As discussed in Daley and Vere-Jones (2003), page 183, self-exciting processes come "closest to fulfilling, for point processes, the kind of role that the autoregressive model plays for conventional time series." This is one of the main advantages of the self-exciting point process modeling.

Our focus will be on the self-exciting marked point process. The intensity for the next infinitesimal time interval, $(t, t + dt)$, is determined conditionally on the values of the point process and its marks in the past time interval $(0, t)$. Furthermore, given the conditional specification of the intensity function, the statistical structure of the point process can be characterized completely. Namely, all the finite-dimensional distributions of the inter-arrival times and the associated marks can be explicitly handled. We now provide a non-mathematical heuristic definition of a self exciting marked point process. This definition is sufficient for our purposes.

- We initialize the process by $N(0) = 0$.
- Our processes will be *conditionally orderly* that is at any time $t \geq t_0$, for an infinitesimally small time interval $(t, t + dt)$, conditional on any events defined by the realization of the process on $(t_0, t]$, the probability of two or more events to occur in $(t, t + dt)$ will be infinitesimally small relative to the probability of one event to occur.
- We will assume that the last event of the process occurred at time t or before.

A formal definition is provided in Snyder and Miller (1991), pages 288–291. The resulting marked point process is called a self-exciting process. Its stochastic behavior is best understood in terms of its conditional intensity function, defined as follows:

$$\lambda(t|\mathcal{H}_t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}\{N(t, t + \Delta t) > 0|\mathcal{H}_t\}}{\Delta t}.$$

The conditional intensity $\lambda(t|\mathcal{H}_t)$ can be interpreted as the intensity of the point process immediately after t. For questions concerning the existence of the conditional intensity, we refer the reader to page 70 of Karr (1991). For $N(t) \geq 1$, we have the following equivalence:

$$\mathbb{P}\{N(t,t+1) = 0|\mathcal{H}_t\} = \exp\Big[-\int_t^{t+1} \lambda(y|\mathcal{H}_t)dy\Big]. \tag{3.2.1}$$

This is analogous to a formula for the non-homogeneous Poisson process, when the conditioning is ignored. We will make use of this quantity later on. See Snyder and Miller (1991), page 291, for the derivation of this result, or page 287 of Davison (2003) for a heuristic argument.

Example: The shot noise process used in the study of electrical current fluctuations is an example of a self-exciting point process. Formally, the shot noise process can be defined as

$$\lambda(t|N(t), T_1, \ldots, T_{N(t)}) = \sum_{i=1}^{N(t)} g(t - T_i).$$

In this case, the intensity function at time t becomes a linear combination of the difference between the current time, and the previous occurrence times through a function $g()$.

There are various ways to statistically model self-exciting processes. A popular one in the context of the analysis of high frequency financial data was the ACD models first put forth by Engle and Russell (1998). In this paper, the authors propose a model for the point process by explicitly describing the times between the buy/sell transactions of a stock. The original ACD model, which includes the marks, is defined as:

$$X_{N(t)+1} = \Psi_{N(t)+1}\varepsilon_{N(t)+1}, \tag{3.2.2}$$

$$\mathbb{E}[X_{N(t)+1}|\mathcal{H}_t] \equiv \Psi_{N(t)+1}, \tag{3.2.3}$$

where $\{\varepsilon_i\}$ are independent and identically distributed non-negative noise terms. In the context of Engle and Russell (1998), at time t, $X_{N(t)+1}$ represents the time till the next buy or sell transaction. The term $\Psi_{N(t)+1}$ is $\mathcal{H}_t$ measurable at time t since its value depends on the previous realized (historical) values of X_i and Y_i.

The model bears close resemblance to the GARCH models. Like the GARCH models, the ACD models have proven to be very useful in capturing the clustering effects of certain processes. The literature on the ACD modeling is extensive. Excellent up-to-date reviews of such models are given in Tsay (2005), Pacurar (2006), Bauwens and Hautsch (2006) and the references therein.

Various forms and parameterizations of the equations (3.2.2) and (3.2.3) are possible. Our interest is in the *log-ACD* model as first suggested by Bauwens and Giot (2000):

$$X_{N(t)+1} = \exp(\Psi_{N(t)+1})\varepsilon_{N(t)+1}, \tag{3.2.4}$$

$$\Psi_{N(t)+1} = \omega + \alpha\varepsilon_{N(t)} + \beta\Psi_{N(t)} + \eta Y_{N(t)}, \tag{3.2.5}$$

where ε_i's are iid exponential random variables with mean 1. The marks, Y_i's, are iid positive random variables independent from the error terms ε_i's. Note, we only included the last mark $Y_{N(t)}$ rather than all the past marks $Y_1, \ldots, Y_{N(t)}$.

The intensity of the log-ACD model for $t \geq T_{N(t)}$ is

$$\lambda(t|\mathcal{H}_t) = \exp(-\Psi_{N(t)+1}) = \exp\{-(\omega + \alpha\varepsilon_{N(t)} + \beta\Psi_{N(t)} + \eta Y_{N(t)})\}. \tag{3.2.6}$$

The random variable $X_{N(t)+1}$, in the context of our modeling, represents the time till the next event of large loss. At a fixed point of time t, conditionally on the realization of $X_{N(t)}$ and $Y_{N(t)}$, $X_{N(t)+1}$ is simply an exponential random variable

multiplied by a known constant $\Psi_{N(t)+1}$. This follows from the conditional specification of the model in equations (3.2.4) and (3.2.4). As mentioned in Daley and Vere-Jones (2003), page 231, the conditional intensity can be defined by the hazard function corresponding to $X_{N(t)+1}$. Using this fact, we now give a detailed derivation for the equation (3.2.6):

$$\begin{aligned}
\lambda(t|\mathcal{H}_t) &= \frac{d}{dx}\mathbb{P}\{X_{N(t)+1} \leq x|\mathcal{H}_t\}/\mathbb{P}\{X_{N(t)+1} > x|\mathcal{H}_t\} \\
&= \frac{d}{dx}\mathbb{P}\{\varepsilon_{N(t)+1} \leq xe^{-\Psi_{N(t)+1}}|\mathcal{H}_t\}/ \\
&\quad \mathbb{P}\{\varepsilon_{N(t)+1} > xe^{-\Psi_{N(t)+1}}|\mathcal{H}_t\} \\
&= \exp(-\Psi_{N(t)+1})\exp(-x\ \Psi_{N(t)+1})/\exp(-x\ \Psi_{N(t)+1}) \\
&= \exp(-\Psi_{N(t)+1}) = \exp\{-(\omega + \alpha\varepsilon_{N(t)} + \beta\Psi_{N(t)} + \eta Y_{N(t)})\}.
\end{aligned}$$

An advantage of this model is that the intensity is positive regardless of the signs of the coefficients ω, α, β, and η in equation (3.2.5). Note also that the intensity does not explicitly depend on the value of $N(t)$.

3.2.2 Tail Modeling

Our presentation in this section draws mainly from Leadbetter *et al.* (1983), Embrechts *et al.* (1997), Coles (2001) and Richard Smith's discussion in Finkenstädt and Rootzén (2004). Let $\{Z, Z_1, \ldots, Z_n\}$ be a sequence of iid random variables such that for suitably chosen normalizing constants $a_n > 0$, and b_n, as $n \longrightarrow \infty$,

$$\mathbb{P}\{(\max(Z_1, Z_2, \ldots, Z_n) - b_n)/a_n \leq z\} \xrightarrow{D} G(z),$$

where $G(z)$ is a non-degenerate generalized extreme value (GEV) distribution:

$$G(z) = \exp\Big\{-\Big(1 + \xi\frac{z-\mu}{\psi}\Big)_+^{-1/\xi}\Big\}. \tag{3.2.7}$$

The parameters μ, $\psi > 0$, and ξ of the limiting distribution G are the location, scale and shape parameters, respectively. Now, consider the distribution of an excess

of Z over a high threshold u, namely $Z-u$, where we have $Z-u>0$. Then Pickands-Balkema-De Haan theorem - a statement of this theorem can be found in McNeil *et al.* (2005), page 277 - implies that conditionally on the event $\{Z>u\}$, the distribution of $Z-u$ has approximately the Generalized Pareto Distribution:

$$\mathbb{P}\{Z-u<z|Z>u\}\approx H(z)=1-\left(1+\xi\frac{z}{\sigma}\right)_{+}^{-1/\xi}. \tag{3.2.8}$$

The support of the distribution H is defined on

$$\{z:z>0,1+(\xi z)/\sigma>0\},$$

and $\sigma>0$ is defined as:

$$\sigma=\psi+\xi(u-\mu).$$

The original result is due to Pickands (1975). The final form of $H(z)$ depends on the sign of ξ:

1. $\xi<0$: the distribution $H(z)$ has a finite upper end point of $-\sigma/\xi$.

2. $\xi>0$: the distribution $H(z)$ is heavy tailed and more precisely:

$$1-H(z)=\bar{H}(z)\sim\left(\xi\frac{z}{\sigma}\right)^{-1/\xi},\quad z\to+\infty.$$

3. $\xi=0$: the distribution $H(z)$ reduces to the exponential distribution $H(z)=1-\exp(-z/\sigma)$.

The important point here is that all distributions satisfying (3.2.7), will have GPD excess distribution beyond an appropriately chosen threshold.

Example: Let Z follow a Cauchy distribution with CDF,

$$F(z)=1/2+\arctan(z)/\pi,\quad -\infty<z<+\infty.$$

Note that as $z \to +\infty$, $\arctan(z) \sim \pi/2 - 1/z$ so $F(z) \sim 1 - /(\pi z)$. Then with $b_n = 0$ and $a_n = n/\pi$, we obtain

$$\lim_{n\to\infty} \mathbb{P}\{(\max(Z_1, Z_2, \ldots, Z_n) - b_n)/a_n \leq z\} = \lim_{n\to\infty} (1 - 1/zn)^n = \exp\{-(1+z-1)^{-1}\}.$$

The distribution in the right hand side of the last expression is GEV with $\xi = 1$, and $\mu = \psi = 1$. Now, using the approximation in equation (3.2.8) for the CDF again we have:

$$\mathbb{P}\{Z - u < z | Z > u\} = \frac{F(z+u) - F(u)}{1 - F(u)} \approx \frac{1 - \pi/(u+z) - 1 + \pi/u}{1 - (1 - \pi/u)} = 1 - (1 + z/u)^{-1}.$$

The distribution on the right hand side is GPD with $\xi = 1$, and $\sigma = u$.

Equation (3.2.8) can be used to obtain an extreme quantile estimate from F. Define first the quantile of a distribution F as follows:

$$z_q = \inf\{z_q \in \mathbb{R} : F(z_q) \geq q\}, \quad q \in (0, 1).$$

Pick a threshold u such that excesses above u are approximately GPD. A typical threshold corresponds to the 0.90 quantile of the data. Various methods can be used to choose the thresholds. For a detailed description on the choice of threshold refer to Embrechts *et al.* (1997), Chapter 6. Conditioning on the event $\{Z > u\}$, we need to solve for z_q:

$$1 - q = \mathbb{P}\{Z > z_q\} = \mathbb{P}\{Z - u > z_q - u | Z > u\}\mathbb{P}\{Z > u\}.$$

The first term on the right hand side of the above equation is then approximated by the GPD:

$$\mathbb{P}\{Z - u > z_q - u | Z > u\} = \Big(1 + \xi \frac{z_q - u}{\sigma}\Big)_+^{-1/\xi}. \tag{3.2.9}$$

$\mathbb{P}\{Z > u\}$ is just the proportion of values over the threshold u so if u is set at the 0.90 quantile of the data, then $\mathbb{P}\{Z > u\} = 0.10$. We let $\mathbb{P}\{Z > u\} = 1 - q_u$ and

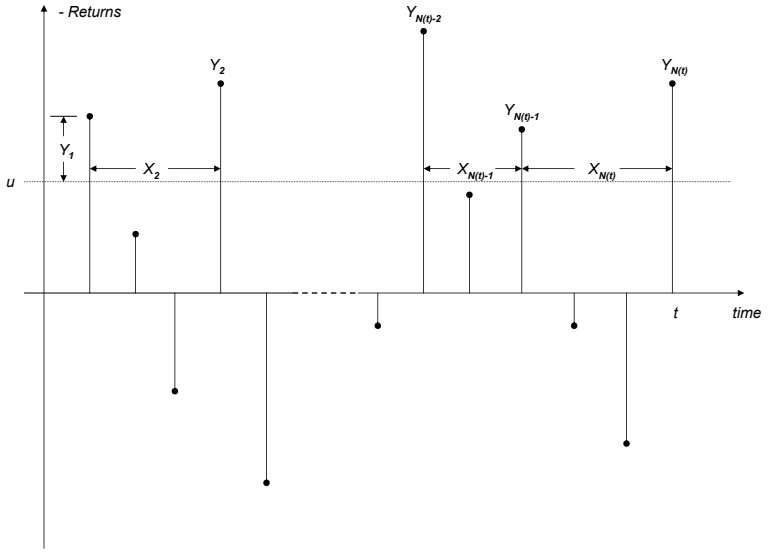

Figure 3.1: The point process of the extremes: our focus will be on the X_i and on the excesses $Y_i = Z_i - u$ over the fixed threshold u.

thus,

$$1 - q = \Big(1 + \xi \frac{z_q - u}{\sigma}\Big)_+^{-1/\xi} (1 - q_u).$$

Solving the above equation for z_q we obtain the quantile estimator:

$$z_q = u + \frac{\sigma}{\xi}\Big[\Big(\frac{1 - q_u}{1 - q}\Big)^{\xi} - 1\Big]. \tag{3.2.10}$$

The above equation is only valid for $z_q > u$ and $q < q_u$, regardless of the sign of ξ.

3.3 Methodology

3.3.1 Estimating the 1-Day VaR

Our goal in this section is to obtain an expression and its estimate for the quantile of the one day ahead predictive distribution of the returns, conditionally on the past and current data. We shall refer to this quantile as the *1 day VaR* since it predicts the value of VaR for the next day, given the past and current conditions in the

market, as captured by our model. More precisely, the 1 day VaR is a quantile of the following conditional CDF:

$$\mathbb{P}\{Z_{t+1} \leq z|\mathcal{H}_t\} = F_{Z_{t+1}|\mathcal{H}_t}(z).$$

Thus the qth quantile of this distribution is:

$$z_q^t = \inf\{z \in \mathbb{R} : F_{Z_{t+1}|\mathcal{H}_t}(z) \geq q\}, \tag{3.3.11}$$

which is a solution to

$$\mathbb{P}\{Z_{t+1} > z_q^t|\mathcal{H}_t\} = 1 - q. \tag{3.3.12}$$

We will take equation (3.3.11) to be the formal definition of 1 day VaR:

$$\text{1 day VaR} = z_q^t.$$

We now give the precise definitions of $N(t)$, X_i, and Y_i in the context of our model defined in equations (3.2.4) and (3.2.5).

Let $\{Z_n\}_{n\geq 1}$ denote the *negative* daily log-returns of a financial asset. Our event of interest is when the loss exceeds a pre-specified high threshold u: $\{Z_i \geq u\}$. We will assume that this is the event of a large loss from the perspective of a long holder of the asset. The counting variable of our point process, $N(t)$, counts the number of large losses up to time t. The sequence $\{X_n\}_{n\geq 1}$ will represent the interexceedance (inter-arrival) times between the large losses, which can be dependent on the previous excesses $Y_i = Z_i - u$. We will assume that the excesses $\{Y_n\}_{n\geq 1}$ are iid. This assumption in general is not met for the real data. We could have, for example, taken this dependence into account by time series modeling of the excesses. However, based on our own experimentation, additional modeling of the dependence of the excesses does not lead to significantly better estimation of z_q^t; our results in subsection 3.5.3 are quite satisfactory.

From (3.3.12), and $z_q^t > u$, we have:

$$\begin{aligned}\mathbb{P}\{Z_{t+1} > z_q^t|\mathcal{H}_t\} &= \mathbb{P}\{Z_{t+1} - u > z_q^t - u|\mathcal{H}_t\} \\ &= \mathbb{P}\{Z_{t+1} - u > z_q^t - u|Z_{t+1} > u, \mathcal{H}_t\} \times \\ &\quad \mathbb{P}\{Z_{t+1} > u|\mathcal{H}_t\} \qquad (3.3.13)\end{aligned}$$

Using relationship (3.2.1) and noting that

$$\{Z_{t+1} \leq u|\mathcal{H}_t\} = \{No\ exceedance\ from\ t\ to\ t+1|\mathcal{H}_t\},$$

the second term on the right hand of (3.3.13) is

$$\begin{aligned}\mathbb{P}\{Z_{t+1} > u|\mathcal{H}_t\} &= 1 - \mathbb{P}\{N(t, t+1) = 0|\mathcal{H}_t\} \\ &= 1 - \exp\Big[-\int_t^{t+1} \lambda(y|\mathcal{H}_t)dy\Big].\end{aligned}$$

The first term in the right hand side of equation (3.3.13) can be approximated via equation (3.2.9). Thus equation (3.3.13) becomes

$$\mathbb{P}\{Z_{t+1} > z_q^t|\mathcal{H}_t\} = \Big(1 + \xi\frac{z_q^t - u}{\sigma}\Big)_+^{-1/\xi}\Big(1 - \exp\Big[-\int_t^{t+1} \lambda(y|\mathcal{H}_t)dy\Big]\Big).$$

Simplifying the notation $\lambda(y|\mathcal{H}_t) = \lambda$ we obtain a solution for (3.3.12) as follows:

$$z_q^t = u + \frac{\sigma}{\xi}\Big[\Big(\frac{1 - \exp(-\lambda)}{1 - q}\Big)^{\xi} - 1\Big]. \qquad (3.3.14)$$

Note, regardless of the sign of ξ, the above estimate is valid for $z_q^t > u$, and $\lambda \geq -\log(q)$.

At time t, we have a realized marked point process: $N(t) = n$, and $\{(X_i = x_i, Y_i = y_i), i = 1, \ldots, n\}$ as shown in Figure 3.1. The future marks Y_i and the interexceedance times X_i are conditionally independent, given the past. Therefore, by conditioning, the joint likelihood of the data can be expressed as

$$f_{X_1,Y_1}(x_1, y_1)\prod_{i=2}^{n} f_{X_i,Y_i \mid X_{i-1},Y_{i-1}}(x_i, y_i \mid x_{i-1}, y_{i-1}) =$$

$$\underbrace{f_{X_1}(x_1)\prod_{i=2}^{n} f_{X_i \mid X_{i-1},Y_{i-1}}(x_i \mid x_{i-1}, y_{i-1})}_{I}\underbrace{\prod_{j=1}^{n} f_{Y_j}(y_j)}_{II}.$$

We describe next our methodology for estimating the parameters in the model by essentially maximizing the above likelihood function. *We focus on component II first.* Assuming that the threshold u is set sufficiently high so that the marks Y_i are iid GPD, the parameters σ and ξ can be estimated via maximum likelihood (MLE). The log-likelihood is

$$\log(\prod_{j=1}^{n} f_{Y_j}(y_j)) = n\log\sigma + (1+1/\xi)\sum_{i=1}^{n}\log(1+\xi y_i/\sigma). \tag{3.3.15}$$

The MLE of the GPD parameter is of non-regular nature as the support of the distribution depends on the parameters. Smith (1985) investigated the MLE properties of the GPD parameter estimation and concluded that the MLE based estimates still maintain the usual asymptotic MLE properties as long as $\xi > -1/2$.

The residuals defined below provide a useful diagnostic tool for assessing the GPD assumption and the goodness-of-fit of our model:

$$\widehat{W_i} = (\hat{\xi}^{-1})\log(1+\hat{\xi}Y_i/\hat{\sigma}). \tag{3.3.16}$$

By simple variable transformation, one can verify that $\widehat{W_i}$ form an iid sequence of exponential random variables with mean 1 if Y_i are iid GPD with parameters $\widehat{\xi}$ and $\widehat{\sigma}$. In practice, the independence assumption can be checked via an ACF plot of the residuals. Additionally, a Q-Q plot of the $\widehat{W_i}$ residuals versus the standard exponential random variable should look linear if our GPD assumption of the excesses is met.

An important remark regarding a potential pitfall in the GPD estimation of ξ needs to be made here. Often times, GPD MLE underestimates ξ. One culprit

is the presence of extremal dependence, i.e. extremal index $\theta < 1$. The effective sample size n of the data is reduced to θn for the purposes of studying the tail and marginal distribution of the data. Another potential reason is the slow rate of convergence of the conditional excess distribution to the limiting GPD. Table 1 in Degen *et al.* (2007) shows the rate of convergence of the excess to GPD as a function of the threshold level for various distributions. The rates can be quite slow. For example if the returns are log-normal - a common assumption made in the finance literature - the rate is $O(\frac{1}{\log(u)})$. However, poor estimates of ξ do not lead to poor quantile estimates based on (3.3.22) and (3.3.12). This fact was observed, tested via simulations, and reported in McNeil and Frey (2000).

Now, we focus on maximizing component I. We will assume the set up in equations (3.2.4) and (3.2.5). The choice of the exponential distribution is motivated by the results in Hsing *et al.* (1988) showing that for increasing thresholds, under general conditions, the exceedance point process converges to a cluster Poisson process with exponentially distributed random times between the clusters of exceedances. Furthermore, our own extensive experimentations with different error distributions such as the Weibull and Generalized Gamma, and also with more complicated functional forms of ψ_i did not lead to better results.

For suitably chosen initial value ψ_0, the log-likelihood of the component I is

(3.3.17)
$$\log\Big(f_{X_1}(x_1)\prod_{i=2}^{n} f_{X_i \mid X_{i-1},Y_{i-1}}(x_i \mid x_{i-1}, y_{i-1})\Big) = l(\omega,\alpha,\beta,\eta) = -\sum_{i=1}^{n}\Big(\frac{x_i}{e^{\psi_i}} + \psi_i\Big).$$

Component I residuals for the model diagnosis and goodness of fit purposes are obtained as follows:

(3.3.18)
$$\widehat{\varepsilon_i} = X_i \ \exp\{-\widehat{\psi_i}\}.$$

The above residuals form a sequence of approximately iid exponential random variables when the model fits. Similar to the $\widehat{W_i}$, the residuals $\hat{\varepsilon}_i$ are examined via ACF and Q-Q plots. The values of the $\widehat{\psi_i}$'s can be computed from the past values of the data and the estimated parameters.

Components I and II can be optimized separately to estimate our parameters. The maximization can be performed using any standard statistical packages. In our work, we used the **R**'s "optim" function. This function is the general purpose optimization tool in **R**. We used the default method of optimization which is Nelder-Mead. We chose the Nelder-Mead method since it performed with the minimum amount of difficulty in our experimentations among various optimization routines. The standard errors for both components are computed numerically by asking the optim function to return differentiated Hessian matrices. The standard errors for component II could also be obtained by plugging the estimates into the expected Fisher information matrix as given by Embrechts *et al.* (1997), Page 357. The code for the optim usage can be found in Section 3.7.

The outputs from the components I and II maximization are our parameter estimates: $\hat{\omega}$, $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\eta}$. Now that we have obtained the estimates for the model parameters, we can proceed in estimating the 1 day VaR.

The first step is to obtain an MLE estimate of the intensity $\lambda(y|\mathcal{H}_t)$, for the time interval t to $t+1$ by using (3.2.6). *We will consider including the excess in three forms, leading to three different intensities.* They are as follows:

$$\lambda_{linear}(t|\mathcal{H}_t) = \exp\{-(\omega + \alpha\varepsilon_{N(t)} + \beta\Psi_t + \eta(Y_{N(t)}))\} \tag{3.3.19}$$

$$\lambda_{log}(t|\mathcal{H}_t) = (Y_{N(t)})^{\eta}\exp\{-(\omega + \alpha\varepsilon_{N(t)} + \beta\Psi_{N(t)})\} \tag{3.3.20}$$

$$\lambda_{plain}(t|\mathcal{H}_t) = \exp\{-(\omega + \alpha\varepsilon_{N(t)} + \beta\Psi_{N(t)})\} \tag{3.3.21}$$

We will name these *linear*, *log*, and *plain* versions. Unlike the first two versions, the last version implies that the intensity is not affected by an excess. In the plain version, we are actually excluding the marks. As a byproduct of our investigation, we will be able to infer on how large excesses can affect the time to the next large loss or gain based on the statistical significance of η. The MLE estimated equivalents of the intensity functions are:

$$\begin{aligned}
\widehat{\lambda_{linear}}(t|N(t)=n, X_{N(t)}=x_n, Y_{N(t)}=y_n) &= \exp\{-(\hat{\omega}+\hat{\alpha}\widehat{\varepsilon_n}+\hat{\beta}\widehat{\Psi_n}+\hat{\eta}(y_n))\},\\
\widehat{\lambda_{log}}(t|N(t)=n, X_{N(t)}=x_n, Y_{N(t)}=y_n) &= (y_n)^{\hat{\eta}}\exp\{-(\hat{\omega}+\hat{\alpha}\widehat{\varepsilon_n}+\hat{\beta}\widehat{\Psi_n})\},\\
\widehat{\lambda_{plain}}(t|N(t)=n, X_{N(t)}=x_n) &= \exp\{-(\hat{\omega}+\hat{\alpha}\widehat{\varepsilon_n}+\hat{\beta}\widehat{\Psi_n})\},
\end{aligned}$$

where $\widehat{\varepsilon_n} = x_n\ \exp\{-\widehat{\psi_n}\}$. The values x_n and y_n are just the last observed interexceedance time and the excess value respectively. From now on, we will suppress the version subscripts for our intensities.

In the second step, an estimate of 1 day VaR, $\hat{z_q^t}$, is obtained by directly substituting our parameter and intensity estimates into the equation (3.3.14):

$$\hat{z_q^t} = u + \frac{\hat{\sigma}}{\hat{\xi}}\Big[\Big(\frac{1-\exp(-\hat{\lambda})}{1-q}\Big)^{\hat{\xi}} - 1\Big]. \tag{3.3.22}$$

We will have three versions of $\hat{z_q^t}$ based on the three versions of the intensities defined in equations (3.3.19)-(3.3.19).

3.3.2 Estimating the T-Day Excess VaR

We may also be interested in estimating the cumulative T-day excess conditional quantile, with T representing the future time horizon over which the excesses will accumulate. For the T day ahead cumulative excess loss distribution of the returns,

$$\mathbb{P}\left\{\sum_{i=1}^{T} Y_{t+i} \le y|\mathcal{H}_t\right\} = F_{\sum_{i=1}^{T} Y_{t+i}|\mathcal{H}_t}(y),$$

we define its corresponding conditional quantile as follows:

$$y_q^T = \inf\{y \in \mathbb{R} : F_{\sum_{i=1}^T Y_{t+i}|\mathcal{H}_t}(y) \geq q\}, \tag{3.3.23}$$

which is a solution to

$$\mathbb{P}\left\{\sum_{i=1}^{T} Y_{t+i} > y_q^T|\mathcal{H}_t\right\} = 1 - q.$$

Let y_q^T be the T-day cumulative excess loss VaR. Closed form solutions for y_q^T can not be obtained. In subsection 3.5.4, we will outline a simulation procedure to estimate y_q^T.

3.4 Theoretical results

In this section, we provide sufficient conditions for the strict stationarity of our model. We state a key theorem from Brandt (1986) which we use to establish the result. *The notation used in the theorem is as it appears in the paper and does not correspond to the notation we have been using.*

Theorem III.1. *Brandt, 1986*

If the sequence $\Psi = \{(A_n, B_n)\}$ *is stationary and ergodic and one of the conditions*

$$-\infty \leq \mathbb{E}[\log(A_0)] < 0 \qquad \textit{and} \qquad \mathbb{E}[(\log(B_0))^+] < \infty,$$

or

$$P(A_0 = 0) > 0$$

is satisfied, then

$$y_n(\Psi) = \sum_{j=0}^{\infty}(\prod_{i=n-j}^{n-1} A_i)B_{n-j-1}, \quad n \in \mathbb{Z},$$

is the only proper stationary solution of

$$Y_{n+1} = A_n Y_n + B_n, \quad n \in \mathbb{Z}$$

for a given Ψ. *Furthermore the solution converges absolutely almost surely.*

This theorem is applicable in a number of situations, including time series analysis of non-linear processes such as GARCH and ARCH processes. The following result gives the conditions under which the log-ACD model is strictly stationary.

Theorem III.2. *Define the Log-ACD model as follows:*

$$X_i = \exp(\Psi_i)\varepsilon_i,$$

$$\Psi_i = \omega + \alpha\varepsilon_{i-1} + \beta\Psi_{i-1} + \eta f(Y_{i-1}),$$

where $\{\varepsilon_i\}_{i\geq 0}$ *and* $\{Y_i\}_{i\geq 0}$ *are iid almost surely positive, and* $f : \mathbb{R}^+ \mapsto \mathbb{R}^+$. *The sequence* $\{X_i\}_{i\geq 1}$ *is stationary if*

$$0 \leq \beta < 1, \qquad \textit{and} \qquad \mathbb{E}[f(Y_i)] < \infty.$$

Furthermore, when $\{Y_i\}_{i\geq 0}$ *are GPD with parameters* ξ *and* σ, $\mathbb{E}[f(Y_i)] < \infty$ *if* $f : y \mapsto y$ *and* $\xi < 1$ *or* $f : y \mapsto \log(y)$.

Proof: We will use the Brandt (1986) result in a straight forward manner. Apply the Brandt theorem to Ψ:

$$\underbrace{\Psi_i}_{Y_{n+1}} = \underbrace{\beta}_{A_n}\underbrace{\Psi_{i-1}}_{Y_n} + \underbrace{\omega + \alpha\varepsilon_{i-1} + \eta f(Y_{i-1})}_{B_n}.$$

The sequence $\{(A_n, B_n)\} = \{(\beta, \omega + \alpha\varepsilon_{i-1} + \eta f(Y_{i-1}))\}$ is a two dimensional iid sequence and hence stationary and ergodic.

We need to check the conditions:

$$-\infty \leq \beta < 0, \qquad \text{and} \qquad \mathbb{E}[\log(\omega + \alpha\varepsilon_{i-1} + \eta f(Y_i))] < \infty.$$

Note $-\infty \leq \log(\beta) < 0$ since $0 \leq \beta < 1$. Using Jensen's inequality we have,

$$\mathbb{E}[\log(\omega+\alpha\varepsilon_{i-1}+\eta f(Y_i))] \leq \log(\mathbb{E}[\omega+\alpha\varepsilon_{i-1}+\eta f(Y_i)]) = \log(\omega+\alpha+\eta\mathbb{E}[f(Y_i)]) < \infty,$$

since $\mathbb{E}[f(Y_i)] < \infty$. The stationarity of X_i follows from the stationarity of Ψ_i.

Now for $Y_i \sim GPD_{\xi,\sigma}$, and $f(Y_i) = Y_i$, $\mathbb{E}[Y_i] < \infty$ if and only if $\xi < 1$ as shown in Embrechts *et al.* (1997), page 165. If $f(Y_i) = \log(Y_i)$, and $\xi > 0$, then $\mathbb{E}[\log(Y_i)] < \infty$, since,

$$\mathbb{E}[\log(Y_i)] \;=\; \mathbb{E}[\log(\xi Y_i/\sigma)] - \log(\xi/\sigma) \;\leq\; \xi\mathbb{E}[(\xi^{-1})\log(1+\xi Y_i/\sigma)] \;=\; \xi$$

as $(\xi^{-1})\log(1+\xi Y_i/\sigma)$ is just an exponential random variable with rate 1, from (3.3.16). If $\xi \leq 0$, all the moments of Y_i exist so again a simple application of the Jensen's inequality would yield $\mathbb{E}[\log(Y_i)] < \infty$. $\square$

Item	Data	Start Date	End Date	Length
1	S&P 500	01-01-1960	10-18-2007	12,031
2	WTI Oil	01-02-1986	10-10-2007	5,495
3	IBM	01-02-1980	10-12-2007	7,012
4	AT	07-19-1984	10-12-2007	5,862
5	JNJ	01-02-1980	10-12-2007	7,012
6	JPM	12-30-1983	10-12-2007	6,001
7	10 Year Bond	01-02-1962	10-15-2007	11,435
8	USD/GBP	01-02-1990	10-17-2007	4,478

Table 3.1: Summary information of the data.

3.5 Illustration of Methodology on Data Sets

3.5.1 Data Description

Table 3.1 gives a summary of the data we will use for our analysis. The frequency for all data is daily. All data were obtained from Yahoo! Finance (http://finance.yahoo.com/) with the exception of the 10 year US Treasury bond and the currency rates (both downloaded from http://www.federalreserve.gov) and the oil data (downloaded from http://www.eia.doe.gov/). For our analysis, we used the differenced log returns of the data. For the stocks, only dividend adjusted data were used.

We focus on the key *empirical* features of the data, which are related to our estimation of VaR: the tail behavior and the clustering of the extreme values. The period from 01-02-1986 to 10-10-2007 is chosen for a total of 5,468 trading days of S&P 500, Oil, and IBM to illustrate these features. We initially focus on the losses, but deal with the gains later on as well.

To show how large data values cluster, we estimate the extremal index using the Ferro-Segers estimator for a moving block of size 1000 (about 4 years of trading data) for each sequence of data and present the results in Figure 3.2. As seen in Figure 3.2, all data sets exhibit clustering of extreme values, i.e. $\theta < 1$. This could lead to successive violations of one's VaR estimates. The patterns for the IBM and S&P 500

data sets are similar as one would expect. The pattern for the Oil data set seems to be the opposite of the IBM and S&P 500 patterns with more clustering in the early periods of the series and less in the later periods.

Using the same blocks, we estimate the shape parameter ξ via GPD modeling and MLE. Based on Figure 3.3, we can see that Oil and IBM exhibit heavy tails, i.e. $\xi > 0$. S&P500 at times exhibits "short" tails, i.e. $\xi < 0$. The results seem to go against the conventional wisdom that the tails are *always* either heavy or normally distributed. However, recall our discussion that the ξ are generally underestimated when using the GPD modeling, and estimation via MLE.

The values of θ and ξ change over time, which suggests the presence of non-stationarity. Note that we have not included confidence intervals in our analysis. Our purpose up to this point, has been descriptive and exploratory.

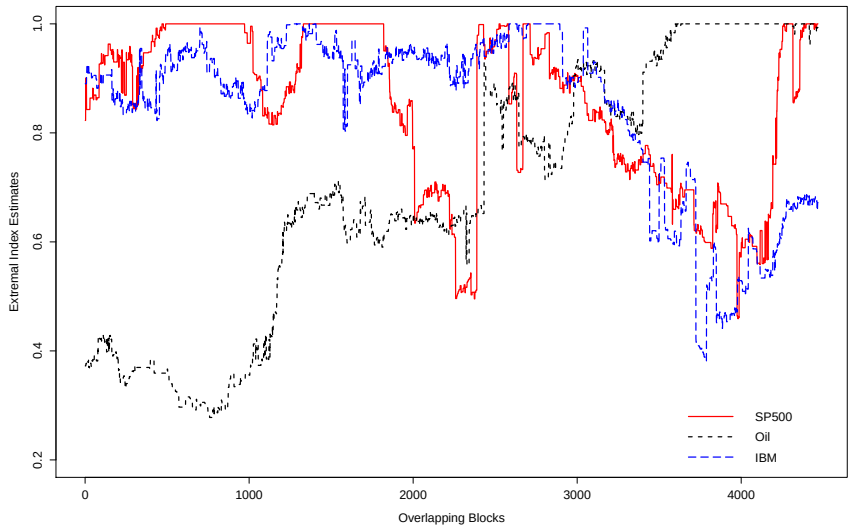

Figure 3.2: θ estimates for the three assets: S&P 500, IBM, and WTI Crude Oil.

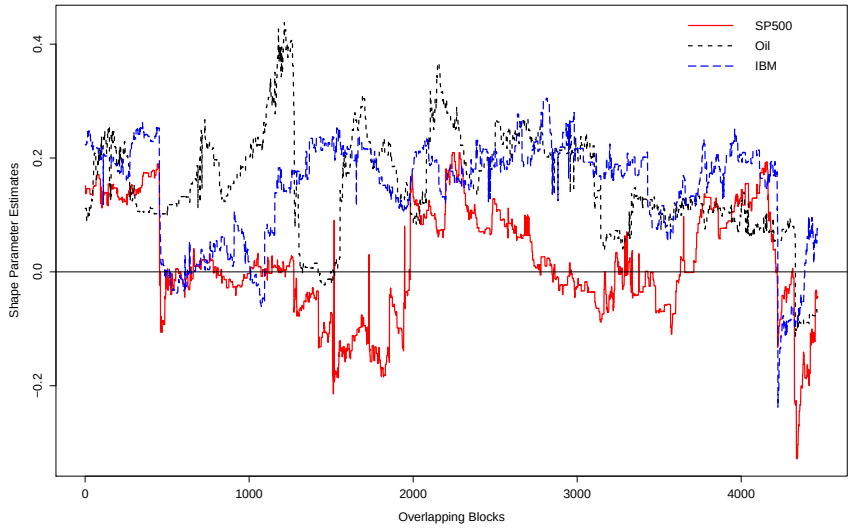

Figure 3.3: ξ estimates for the three assets: S&P 500, IBM, and WTI Crude Oil.

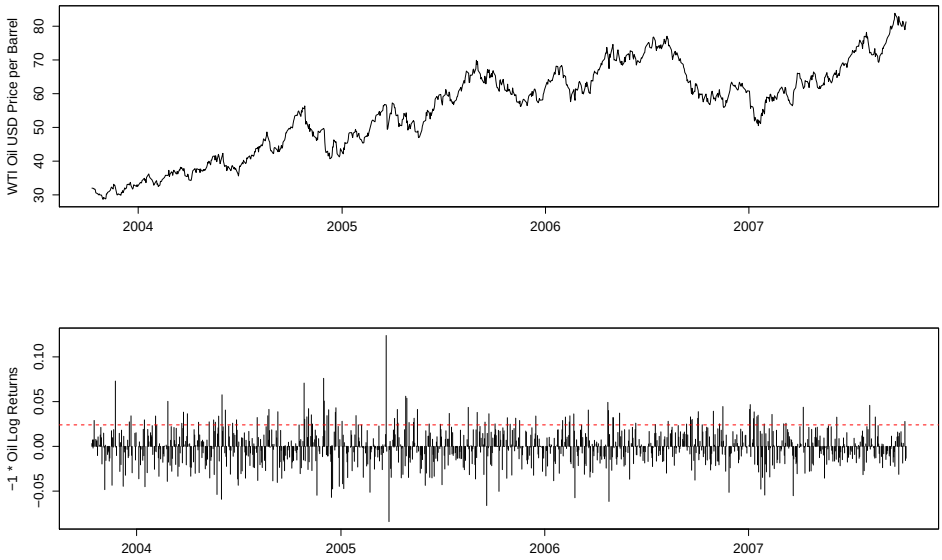

Figure 3.4: Top Panel: WTI oil price per barrel , Bottom Panel: Negative log returns of Oil. The dotted line corresponds to the 0.90 quantile of the data for this period.

3.5.2 Example 1 Day VaR Estimation

In this section, we adopt the view of a practitioner interested in using the proposed methodology. We illustrate our method for estimating the 1 day VaR at a fixed level of $q = 0.99$, and we give a detailed analysis including all the relevant diagnostics. Whenever possible, we provide interpretations of the estimated values. We choose the last 1000 days of oil data from October 2003 to October 2007. See Figure 3.4. Our method is most effective when high degree of clustering is present. However, our estimate of the extremal index is about 0.98 for this period which indicates minimal clustering of the extreme losses.

The first step is to estimate the GPD shape and scale parameters, ξ and σ, in equation (3.2.8). The top left panel of Figure 3.5 shows the successive MLE estimates of the shape parameter against higher thresholds. There seems to be a stable region just above the line $\xi = 0$. Our estimate of ξ at the threshold value of

Version	$\widehat{\omega}$	$\widehat{\alpha}$	$\widehat{\beta}$	$\widehat{\eta}$	$\widehat{\mathbb{E}[\varepsilon]}$	$\widehat{\mathbb{E}[\varepsilon^2]}$	$\widehat{\lambda}$	$\widehat{z}^t_{q=0.99}$
Plain	1.704(1.057)	0.271(0.130)	0.130(0.458)	-	1.012	1.899	0.046	0.040
Linear	0.667(0.366)	0.172(0.162)	0.703(0.195)	-14.135(5.504)	1.017	1.852	0.055	0.042
Log	0.027(0.898)	0.222(0.204)	0.547(0.356)	0.155(0.084)	1.014	1.879	0.053	0.042

Table 3.2: Parameter estimates for our three version of Oil Returns from October 2003 to October 2007. Quantities in parentheses are standard errors.

0.023, corresponding to the 0.90 quantile of the data, is 0.062 with a standard error of 0.086, giving us a 95% confidence interval of (-0.106, 0.230). Note the units of the threshold of 0.023 is in percent points so this value corresponds to a loss of 2.3% in the asset value. The confidence interval contains the value of zero, thus indicating that the assumption of the returns being normally or log-normally distributed seems to be valid. The estimated σ at various quantile levels (not shown here) also showed a stable region around the threshold value of 0.023. The corresponding point estimate, standard error and the 95% confidence interval are 0.011, 0.0015, (0.008, 0.014) respectively. The Q-Q plot of the excesses versus the GPD quantiles in the top left panel of Figure 3.5 indicates that we have a good fit.

Note that the selection of the appropriate threshold for quantile estimation remains an open problem. Prominent researchers have dubbed this the "Achilles heel" of the extreme value theory. For a discussion of the ways to choose the appropriate threshold see Beirlant *et al.* (2004) Sections 4.7 and 5.8.

The ACF and the Q-Q plot of the W_i residuals, as defined in equation (3.3.16), are displayed in the bottom portion of Figure 3.5. A slightly significant peak is seen at the lag of 9 but otherwise there does not appear to be any dependence based on the ACF plot. The Q-Q plot indicates that the W_i residuals are exponential, as expected, when the excesses are iid GPD. The first and the second sample moments of the W_i are 1.000 and 2.017 respectively, which are very close to the theoretical moment values of 1 and 2 for an exponential random variable with mean 1.

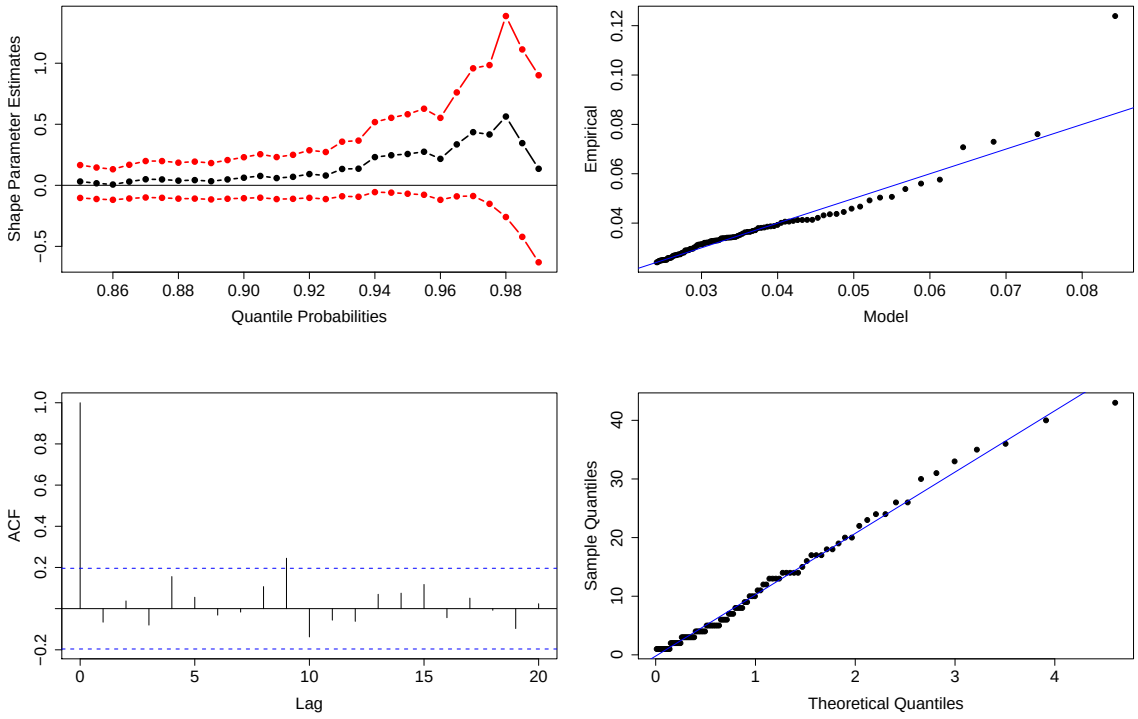

Figure 3.5: Top Left Panel: ξ estimates for the Oil, Top Right Panel: Q-Q plot of the Oil excess versus GPD. Bottom Left Panel: ACF of W_i residuals, Bottom Right Panel: QQ-Plot of W_i residuals versus Exponential with mean 1.

Next, we estimate the parameters related to the intensities based on the three versions suggested in Section 3.3. The intensity estimates are obtained via equations (3.3.19) to (3.3.21) and the corresponding residuals are computed using equation (3.3.18). Table 3.2 displays our parameter estimates. Figure 3.6 displays the diagnostics for the ACD models. We also estimate the conditional quantile estimates using equation (3.3.22). Devoting more attention to the estimated intensities and quantiles, we summarize the results from Table 3.2 and Figure 3.6 next.

The estimated intensities give the instantaneous probability of an exceedance. For example, the linear version predicts that there is a $\widehat{\lambda} = 0.055$ instantaneous probability that a value over a threshold of $u = 0.023$ will occur. This instantaneous

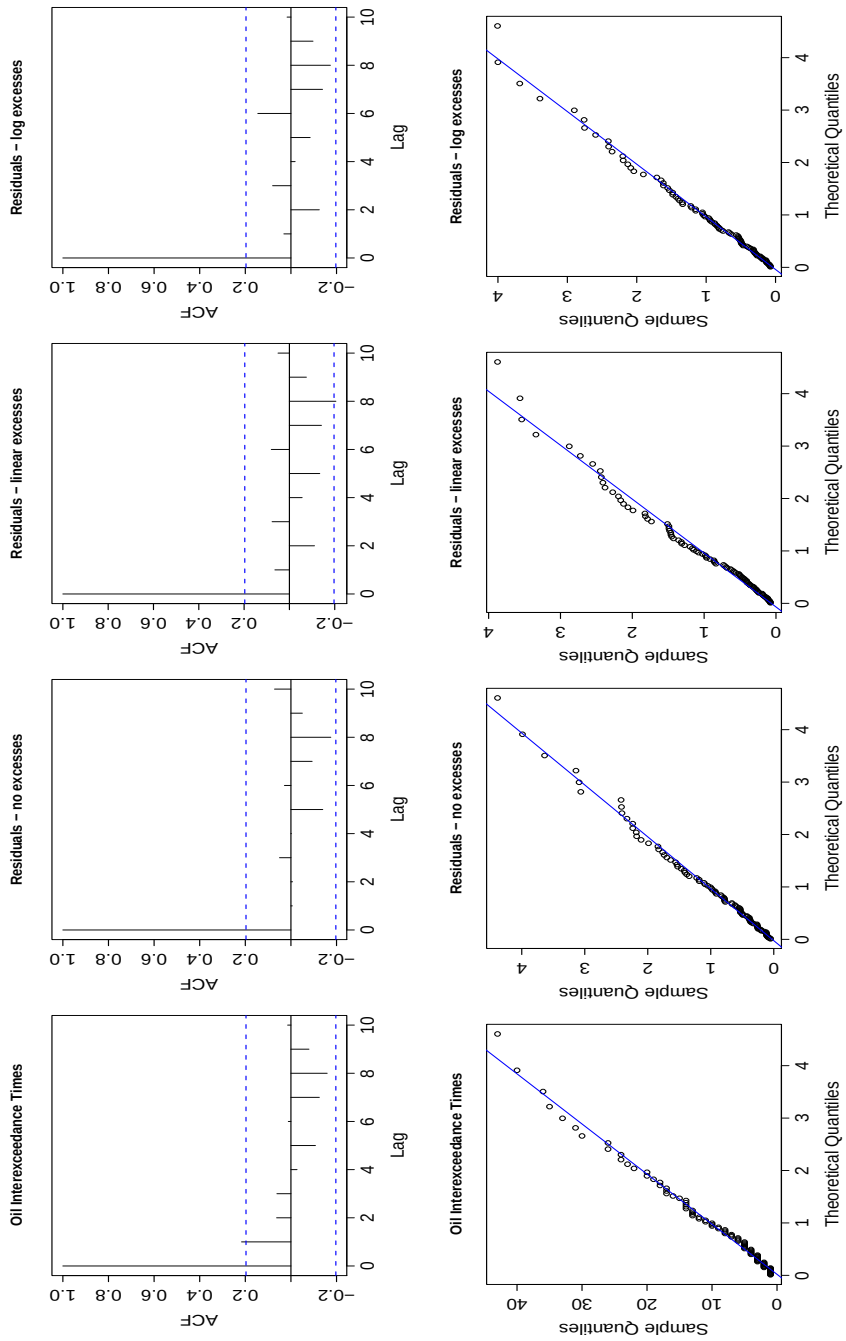

Figure 3.6: Top, Left to Right: ACF of the Oil interexceedance times, and our model residuals. Bottom, left to Right: Q-Q plot of Oil, and our model residuals versus exponential distribution.

probability can be taken as an *estimate* for the next time period since we assume no information arrives from now to the next time period. Typically $\lambda = 0.1$, but in this case a lower probability is predicted. This is partly due to the fact that the last interexceedance time (not shown) is 36 days, that is we are not in a state where extreme values are arriving intensely. All three versions give very similar estimates for the one day ahead 0.99 VaR as seen under the column $\widehat{z}^t_{q=0.99}$.

Not all estimated parameters are statistically significant at 5%, i.e. the 95% would contain zero. However, η is statistically significant in the linear version and very close to being significant in the log version. The entire 95% confidence interval for η in the linear version is below zero, suggesting only negative values for η. Based on this period, we have statistical evidence that large losses increase the intensity of the process and thus increase the probability of observing a large loss in the next period. Using equation (3.3.19), and the estimated values of 2.582 and 2.723 for $\widehat{\varepsilon}_{100}$ and $\widehat{\psi}_{100}$ respectively, for an observed 2% in the excess loss over the threshold of 2.3%, that is for a total loss of 4.3%, the next period intensity and thus the instantaneous probability of observing another large loss is 0.064:

$$\widehat{\lambda_{linear}} = \exp\{-(0.667 + 0.172(2.582) + 0.703(2.723) - 14.135(0.02))\} = 0.064.$$

Another interpretation is that an increase of Δ units in the excesses increases λ_{linear} by $100 \times \exp(-14.135\Delta)\%$.

The diagnostics show that the plain version seems to capture the dependence of the X_i by removing the dependence in the residuals well. The Q-Q plots of our models also indicate that the slight clustering effect, as seen in the lower part of the Q-Q plot of the oil interexceedance times, has been removed. However, in this case some of the apparent clustering is due to the discreteness of the interexceedance times of the oil data. The moments of the residuals match well with the moments of

an exponential distribution although there appears to be some under-dispersion, as seen from the values of $\widehat{\mathbb{E}[\varepsilon^2]}$ in Table 3.2.

3.5.3 Backtesting the 1 Day VaR

To validate our modeling procedure and assess its accuracy in predicting the large quantiles, we use the backtesting procedure as described in McNeil and Frey (2000). We perform our predictions and backtesting on both the gains and the losses of the returns. We now briefly describe the procedure. Additional general information on backtesting can be found in Jorion (2007b), Hull (2007), Dowd (2005), McNeil *et al.* (2005), and Christoffersen (2003).

For each time series of the returns $\{z_1, \ldots, z_n\}$, a moving window of 1000 days - approximately four years of trading data - in a set $\{z_{t-999}, \ldots, z_t\}$, $t \in \{1000, 1001, \ldots, n-2, n-1\}$ is considered. The data from each window is used to estimate our parameters and predict the 1 day VaR based on equation (3.3.22) and each intensity version in equations (3.3.19) to (3.3.21). The estimated VaR, $\hat{z_q}^t$, for each quantile level $q \in \{0.95, 0.99, 0.995\}$, is compared with the realized value z_{t+1}. We say a *violation* has occurred, when $\hat{z_q}^t < z_{t+1}$.

Under the null hypothesis that our model is accurately predicting the quantiles, the number of violations is a binomial random variable with the expected value equal to $(n - 1000)(1 - q)$ and the success probability of $1 - q$. Therefore, a two sided binomial test can be used to assess whether our predictions are overestimating or underestimating the conditional quantiles. Larger p-values will support the null hypothesis and thus support the method being used. From a practical perspective, overestimating VaR is detrimental to the risk taker as it may trigger overly conservative capital reserve requirements and thus limit the profit potential. On the other hand, underestimating VaR means underestimating risk, which can lead

to large financial losses. Thus, accurate estimates of VaR, which neither over- nor under-estimate the loss are desirable.

We compare our method to the unconditional EVT, based on equation (3.2.10) and a GARCH model. The competing GARCH model is given by:

$$\sigma_t^2 = \alpha_0 + \alpha_1 \sigma_{t-1}^2 + \alpha_2 z_{t-1}^2,$$

where $z_t|\mathcal{F}_{t-1} \sim \mathcal{N}(0, \sigma_t^2)$, and $\mathcal{F}_{t-1}$ represents the information set available in $\{z_{t-999}, \ldots, z_t\}$. This model is similar to the RiskMetrics^{TM} as described in Tsay (2005), page 290. The difference between our model and RiskMetrics is that the latter model assumes $\alpha_0 = 0$ and $\alpha_1 + \alpha_2 = 1$. Our model is quite common for GARCH process modeling of the returns. Likewise at each moving block, we estimate the GARCH parameters, and obtain the conditional quantile based on

$$\hat{z_q}^t = \hat{\sigma}_t \Phi^{-1}(q),$$

where $\Phi()$ denotes the cdf of a standard normal random variable.

Tables 3.3 to 3.4 report the expected number of violations, the observed number of violations, and the two sided binomial test p-values for our three models, based on modeling the intensities and excesses, the unconditional EVT, and GARCH process with normal innovations. We choose 5% as our level of significance.

At the 0.95 quantile prediction, the GARCH based method fails 6 out of 16 times, the unconditional EVT fails 5 out of 16 times. Among our methods, the linear and the plain versions fail only once each and the log version fails 2 out of 16 times. The major downfall of the unconditional EVT at this level is due to its inability to adjust for the clustering of the extremes.

At the 0.99 quantile prediction, the GARCH method fails 15 out of 16 times. The unconditional EVT fails 4 out of 16 times. Except for the plain version for the bonds,

	SP500	Oil	USD/GBP	Bond	ATT	IBM	JNJ	JPM
0.95 Quantile								
Expected	551.5	224.7	173.85	521.7	243.05	300.55	300.55	250
GARCH	562(0.65)	187(0.01)	168(0.65)	531(0.68)	275(0.04)	281(0.25)	303(0.89)	255(0.75)
Uncond. EVT	629(0.00)	231(0.67)	169 0.71)	617(0.00)	258(0.33)	315(0.39)	272(0.09)	276(0.09)
Plain	602(0.03)	218(0.65)	178(0.75)	581(0.01)	257(0.36)	295(0.74)	283(0.30)	255(0.75)
Linear	576(0.28)	225(0.98)	179(0.69)	585(0.00)	275(0.04)	308(0.66)	295(0.74)	251(0.95)
Log	590(0.09)	242(0.24)	181(0.58)	583(0.01)	269(0.09)	316(0.36)	287(0.42)	261(0.48)
0.99 Quantile								
Expected	110.3	44.94	34.77	104.34	48.61	60.11	60.11	50
GARCH	140(0.00)	52(0.29)	56(0.00)	196(0.00)	76(0.00)	88(0.00)	86(0.00)	69(0.01)
Uncond. EVT	141(0.00)	47(0.76)	39(0.47)	160(0.00)	53(0.53)	66(0.45)	50(0.19)	62(0.09)
Plain	131(0.05)	42(0.66)	35(0.97)	135(0.00)	52(0.63)	64(0.61)	57(0.69)	59(0.20)
Linear	115(0.65)	44(0.89)	38(0.58)	140(0.00)	53(0.53)	61(0.91)	56(0.59)	56(0.39)
Log	121(0.31)	48(0.65)	37(0.70)	149(0.00)	52(0.63)	62(0.81)	58(0.78)	61(0.12)
0.995 Quantile								
Expected	55.15	22.47	17.385	52.17	24.305	30.055	30.055	25
GARCH	78(0.00)	33(0.03)	30(0.00)	133(0.00)	44(0.00)	61 0.00)	57(0.00)	47(0.00)
Uncond. EVT	75(0.01)	27(0.34)	24(0.11)	93(0.00)	32(0.12)	35(0.37)	28(0.71)	32(0.16)
Plain	66(0.14)	20(0.60)	24(0.11)	79(0.00)	32(0.12)	36(0.28)	32(0.72)	30(0.32)
Linear	61(0.43)	21(0.76)	23(0.18)	76(0.01)	32(0.12)	32(0.72)	30(0.99)	32(0.16)
Log	57(0.80)	26(0.46)	24(0.11)	81(0.00)	31(0.17)	33(0.59)	32(0.72)	34(0.07)

Table 3.3: The gains theoretical expected number of violations, observed number of violations, and the corresponding p-values in parenthesis for our models, unconditional EVT and GARCH with normal errors.

none of our methods fail.

At the 0.995 quantile prediction, the GARCH method always fails. The GARCH method, particularly at the higher quantile levels, fails since the normal innovations are not able to account for the size of the losses. The unconditional EVT fails 4 times. Except for the bonds losses, our method never fail. The unconditional EVT seems to be competitive with ours at the highest quantile. This is due to the fact that there is less clustering of extremes at highest quantile level.

We have solid evidence in favor of our method. The linear version seems to have a slight edge overall but there is no clear winner among our three versions. All methods seem to have trouble with the bonds data. Even though we have treated the loss and the gains sides separately, it is clear from the table that most often the losses and the gains fail at the same time for the same method. Note since we were applying all the methods to the same data sets, we can point out that our p-values were higher on average.

	SP500	Oil	USD/GBP	Bond	ATT	IBM	JNJ	JPM
0.95 Quantile								
Expected	551.5	224.7	173.85	521.7	243.05	300.55	300.55	250
GARCH	534(0.45)	198(0.07)	167(0.59)	485(0.10)	211(0.04)	246(0.00)	253(0.01)	216(0.03)
Uncond. EVT	608(0.01)	244(0.19)	144(0.02)	620(0.00)	264(0.17)	307(0.70)	273(0.10)	265(0.33)
Plain	591(0.08)	231(0.67)	158(0.22)	551(0.19)	241(0.89)	290(0.53)	276(0.15)	253(0.85)
Linear	577(0.27)	222(0.85)	168(0.65)	559(0.09)	257(0.36)	297(0.83)	271(0.08)	255(0.75)
Log	573(0.35)	232(0.62)	167(0.59)	574(0.02)	245(0.90)	308(0.66)	29(0.66)	263(0.40)
0.99 Quantile								
Expected	110.3	44.94	34.77	104.34	48.61	60.11	60.11	50
GARCH	178(0.00)	66(0.00)	51(0.01)	166(0.00)	68(0.01)	94(0.00)	87(0.00)	73(0.00)
Uncond. EVT	129(0.07)	48(0.65)	27(0.19)	135(0.00)	51(0.73)	67(0.37)	55(0.51)	51(0.89)
Plain	120(0.35)	48(0.65)	28(0.25)	128(0.02)	45(0.60)	61(0.91)	56(0.59)	44(0.39)
linear	110(0.98)	47(0.76)	28(0.25)	118(0.18)	45(0.60)	57(0.69)	56(0.59)	43(0.32)
Log	114(0.72)	50(0.45)	33(0.76)	123(0.07)	49(0.96)	62(0.81)	58(0.78)	45(0.48)
0.995 Quantile								
Expected	55.15	22.47	17.385	52.17	24.305	30.055	30.055	25
GARCH	112(0.00)	47(0.00)	32(0.00)	116(0.00)	39(0.00)	64(0.00)	60(0.00)	48(0.00)
Uncond. EVT	74(0.01)	29(0.17)	11(0.13)	81(0.00)	26(0.73)	35(0.37)	25(0.36)	24(0.84)
Plain	69(0.06)	27(0.34)	12(0.20)	63(0.13)	24(0.95)	33(0.59)	30(0.99)	24(0.84)
Linear	68(0.08)	26(0.46)	12(0.20)	62(0.17)	25(0.89)	32(0.72)	29(0.85)	26(0.84)
Log	66(0.14)	24(0.75)	12(0.20)	63(0.13)	24(0.95)	33(0.59)	31(0.86)	26(0.84)

Table 3.4: The losses theoretical expected number of violations, observed number of violations, and the corresponding p-values in parenthesis for our models, unconditional EVT and GARCH with normal errors.

3.5.4 Backtesting the 10 Day Cumulative Excess Losses and Gains

We outline a simulation algorithm to estimate y_q^T in equation (3.3.23) with $T = 10$ using the previously estimated model parameters for the 1 day VaR.

1. Let L be the number of excess losses that will occur in the time duration $T = 10$. Randomly draw L $\hat{\varepsilon}_i$ Log-ACD residuals (as defined in (3.3.18)) with replacement such that $L = 0$ if $\hat{\varepsilon}_1 > 10$ and else
$$L = \min\Big(\min\{n \in \mathbb{N} : \sum_{i=1}^{n+1} \hat{\varepsilon}_i > 10\}, 10\Big).$$
2. Randomly draw L Y_i excesses with replacement to obtain the cumulative sum
$$S_j = \sum_{i=1}^{L} Y_i.$$
3. Repeat steps 1 and 2 1000 times and collect the S_j terms.
4. By equation (3.2.10), obtain $q = 0.95$ and $q = 0.99$ quantile estimates for $VaR^{T=10}$ from the empirical distribution of the S_j.

5. Move the data window by 1 day and repeat steps 1 to 4.

By directly resampling from the Log-ACD model residuals, we mimic the current market conditions; for example, during high intensity periods with large losses, the residual values are small and the excesses are large. We also tried generating the number of excesses via the estimated intensity function, and generating the excesses from the estimated GPD model, but the resampling results were better on average. Once the estimated values for y_q^T at $q = 0.95$ and $q = 0.99$ are obtained, we can compare them to the realized cumulative excesses, and similar to the 1 day analysis, we collect the number of violations. However, we will not be able to perform the binomial test as we did for the 1 day violations since the time periods overlap, which introduces dependence. In Tables 3.5 to 3.6, we merely report the number of violations along with the expect number of violations. The "Uncond. EVT" row is obtained by setting $L = 1$ since we would expect the returns to exceed the $q = 0.90$ quantile in 1 day out of 10 days. We simply use the unconditional extreme quantile equation (3.2.10) with q_u set to 0.90. In other words, we are eliminating the effect of our conditional estimation.

From the tables we can see that unconditional EVT performs very poorly compared to our conditional methods. At the 0.95 quantile prediction, for both the gains and losses, our methods seem to perform well except for the S&P 500 data. The linear and the plain versions have a slight edge. Our results for the 0.99 quantile prediction are not satisfactory. As a matter of fact a discernable pattern seems to be that the number of violations is twice the expected ones on average. However, our method still outperforms the unconditional method considerably. This may be contributed to the downward bias in estimating ξ when summing random number of the excesses.

	SP500	Oil	USD/GBP	Bond	ATT	IBM	JNJ	JPM
0.95 Quantile								
Expected	551	224	173	521	243	300	300	250
Plain	672	187	172	539	299	335	309	319
Linear	614	175	179	567	293	330	316	353
Log	645	202	170	624	308	357	304	326
Uncond. EVT	1230	340	334	1423	517	474	436	494
0.99 Quantile								
Expected	110	45	35	104	49	60	60	50
Plain	232	74	25	238	99	139	101	126
Linear	208	70	34	226	114	149	105	122
Log	206	84	27	232	113	151	104	120
Uncond. EVT	708	165	203	970	313	255	213	296

Table 3.5: The gains theoretical cumulative excess expected number of violations and observed number of violations based on simulations.

	SP500	Oil	USD/GBP	Bond	ATT	IBM	JNJ	JPM
0.95 Quantile								
Expected	551	224	173	521	243	300	300	250
Plain	624	268	137	543	280	300	281	239
Linear	603	255	145	556	278	334	289	233
Log	607	260	160	590	317	326	307	230
Uncond. EVT	1243	385	277	1219	559	505	483	455
0.99 Quantile								
Expected	110	45	35	104	49	60	60	50
Plain	261	97	57	245	62	120	67	97
Linear	210	100	61	240	70	123	79	88
Log	249	103	55	256	80	111	92	94
Uncond. EVT	713	222	165	789	295	232	211	264

Table 3.6: The gains theoretical cumulative excess expected number of violations and observed number of violations based on simulations.

3.6 Final Remarks

The results show that our proposed method is accurately predicting the 1 day VaR. Specifically, the practitioner would find our practical example in subsection 3.5.2 with its guidelines and recommendations useful for estimating the 1 day VaR. Furthermore, the investigation of the estimated intensities could be of great interest since it would answer the question "given the current market conditions, what is the probability of experiencing a large loss tomorrow?" This could be a useful risk measure in its own right. We have additional results from analyzing the estimated intensities collected from our backtesting. We are working to extend the results to the expected shortfall and other risk measures. We would like to investigate, which form of the three proposed intensities is best to use since at this point there are no clear winners. The inclusion of the excesses in the intensity brings new insight, since it allows us to test the hypothesis of whether large losses increase the conditional probability of observing another large loss. Our results for the 10 day excess VaR were not satisfactory at $q = 0.99$ level. A possible reason is that the ξ in the second round of GPD estimation is underestimated. We plan to investigate the reasons for such unsatisfactory performance and possibly suggest new ways to obtain estimates for the cumulative losses and gains quantiles. Lastly, we are working to create prediction intervals for our estimates in two ways: delta method, and simulation. Since all of our unknown parameters and their standard errors are estimated via MLE, we can apply the delta method directly to the equation (3.3.22). However, due to the small sample sizes we are dealing with, the results of the delta method need to be taken with caution. Another alternative is to take a simulation based approach similar to the one taken in subsection 3.5.4. The coverage probabilities would could aid us in

deciding which method, simulation or delta method, to use.

3.7 Appendix

The following code is a part of the main function for backtesting our data which includes the use of the optim function.

```
( This part of the code for the main function is omitted.)
start.index = i
end.index   = i + window.size - 1
temp.data   = x[start.index:end.index]
q.level[i] = quantile(temp.data, prob = prob.threshold)
durations  = get.interarrival.times(temp.data, threshold = q.level[i])
excesses   = temp.data[which(temp.data > q.level[i])] - q.level[i]
acd.mle.1  = optim(c(initial.omega.1, initial.alpha.1, initial.beta.1,
               initial.eta.1),
               negloglikelihood.elog2acd11.with.excesses,
               my.data = durations, my.excesses = excesses,
               hessian = T,
               control=list(maxit=3000)  )
acd.mle.2  = optim(c(initial.omega.2, initial.alpha.2, initial.beta.2),
               negloglikelihood.elog2acd11,
               my.data = durations,
               hessian = T,
               control=list(maxit=3000)  )
acd.mle.3  = optim(c(initial.omega.3, initial.alpha.3, initial.beta.3,
               initial.eta.3),
```

```
                negloglikelihood.elog2acd11.with.excesses.log.form,
                my.data = durations, my.excesses = excesses,
                hessian = T,
                control=list(maxit=3000)  )
( This part of the code for the main function is omitted.)
```

CHAPTER IV

Incorporating Clustering of Large Losses into Risk Measures

4.1 Introduction

The classical risk measures such as VaR and Expected Shortfall (ES) are defined for fixed periods of time such as 1 day to 10 days. VaR was defined in the previous chapter. ES answers the question, given that an asset has sustained a large loss, how much loss can be expected, on average? Note that such measures are used as predictions about the future losses. These measures as defined, are *unconditional*, in a sense that the dependence in the data is ignored, and only the marginal distribution of the losses is considered. However, it has been shown for example in McNeil and Frey (2000), Chaves-Demoulin *et al.* (2005), Christoffersen and Gonçalves (2005), and Chapter 2 of this thesis that if one introduces conditioning in estimating the risk, that is taking into account the dependence of the losses in addition to the marginal distribution of the losses, one obtains better predictions for the future losses.

These methods, in large part, work better because they are able to account for the *temporal clustering* of large losses that many financial and tradable assets exhibit. As mentioned in Chapter 1, accumulation of large losses due to the temporal clustering could have grave consequences if provisions for handling such cases have not been made. The drawback in employing such methods is that the estimation procedure

is more complicated, and the results may be difficult to interpret or communicate to the non-technical decision makers. This brings us to the main question that this chapter is attempting to answer:

Can a single statistic incorporate the clustering of large losses into a risk measure?

To answer this question, we must be able to *decluster*, that is to identify the beginning and the ending periods of the clusters of large losses. Therefore, declustering becomes an important step in our investigation. A parameter that is directly related to declustering is the extremal index. Recall that an extremal index close to 1 implies that the average cluster size is near 1, and an extremal index close to 0 implies that average cluster size is much large than 1. Recall also that an extremal index of say 0.5 does not imply that exactly 2 large values happen consecutively. The correct interpretation is that within *a short period of time* - within a cluster - 2 large losses are *expected* to occur. An up-to-date reference on declustering is Chapter 10 of Beirlant *et al.* (2004).

In this chapter, new risk measures are defined and an estimation procedure is proposed and tested to answer the above question. The proposed procedure first identifies the clusters of extremes, which may be viewed as approximately independent. The isolated clusters of extremes correspond to the *extreme states* of the asset. Once the extreme states are identified, various statistics can be defined. For example, one may be interested in the cumulative loss during such states. Then, a straightforward approach in estimating the *extreme state cumulative loss* is to sum the losses in each cluster and then average them. This risk measure would estimate the mean cumulative loss an asset can sustain during an extreme state. Note that this new measure takes into account the temporal accumulation of the large losses. A second

contribution in this chapter is the discovery that combining of two estimators of the extremal index leads to a better estimator.

The following remark is in order: The extreme states correspond to the periods when the asset is incurring large losses in a short duration of time. This is to be differentiated from high volatility periods when an asset's return could fluctuate in the loss and gain directions. In practice, the extreme state and high-volatility periods may coincide, overlap, or they may be completely unrelated.

4.2 Introduction to Declustering

The key task in our estimation of the risk measures is accurate declustering: this corresponds to identifying the approximately independent clusters of the large losses. Declustering of extremes and the estimation of the extremal index, θ, are inherently related problem as summarized by the following relationship:

$$\theta \approx \frac{Number\ of\ Independent\ Clusters\ with\ at\ Least\ One\ Exceedance}{Total\ Number\ of\ Exceedances}, \tag{4.2.1}$$

where exceedances are just the data values exceeding a pre-specified threshold u. The important takeaway from the relationship above is that an accurate estimation of θ will lead to better declustering results and vice versa. Declustering can be accomplished in two ways:

1. Decluster the data first by specifying auxiliary parameters, and without estimating θ first. Then use relationship (4.2.1) to obtain an estimate of θ. The runs approach as defined in equation (2.5.24) uses this method.

2. Estimate θ first and then using asymptotic theory justifications, decluster the data. Ferro and Segers (2003) propose this approach.

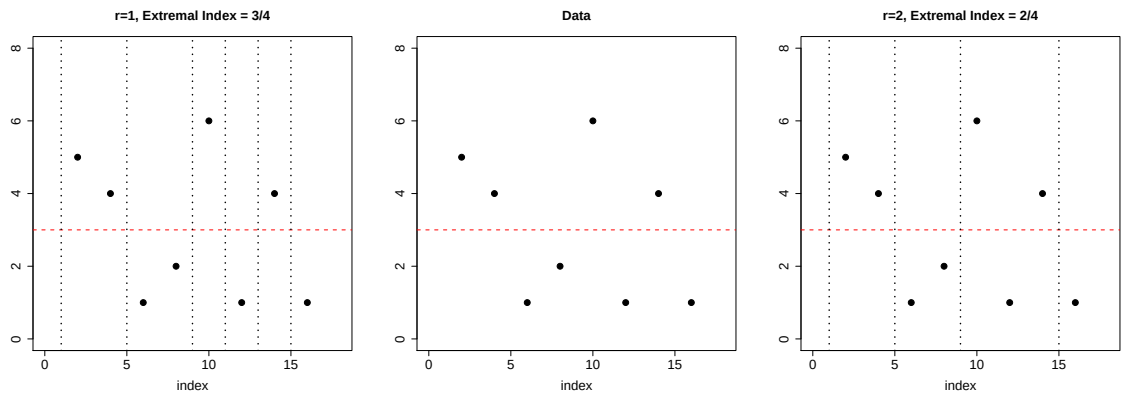

Figure 4.1: Illustration of the Runs Method. The left picture displays the separated clusters by dashed vertical lines when $r = 1$ and thus $\hat{\theta} = 3/4$. The middle picture is just the plot of the data. The right picture displays the declustering when $r = 2$ and thus $\hat{\theta} = 1/2$.

Our previous discussion of the Ferro-Segers and the runs estimators in subsection 2.5.1 focused on the estimation of θ and not on the method of declustering. In the next two subsections, we elaborate on the declustering methods. In the last subsection, we propose a new method to combine the two estimators to achieve better declustering and estimation of θ.

4.2.1 Runs Based Declustering

In the runs method, the run-length auxiliary parameter, r, is used to identify the independent clusters. Recall from Section 2.5.1 that this same parameter was defined to obtain an estimate of θ. A cluster begins when a large value exceeds a prespecified threshold u, and continues until r consecutive values fall below u. Figure 4.1 illustrates how the runs method selects clusters. The center plot is a display of the data. The dashed horizontal line in all three plots depicts the fixed threshold level. The left-most plot shows the case when $r = 1$. Three independent clusters are identified. Starting with the first exceedance, a cluster is terminated as soon as one value falls below the threshold; therefore, $\hat{\theta} = (\#\text{clusters})/\#(\text{exceedances}) = 3/4$.

The right-most plot shows the case when $r = 2$. Now, two independent clusters are identified. Starting with the first exceedance, a cluster is terminated as soon as two values fall below the threshold; therefore, $\hat{\theta} = 2/4$. This simple example also illustrates that the choice of r can substantially affect both the delcustering and estimation of θ.

4.2.2 Ferro-Segers Based Declustering

This declustering method is due to Ferro and Segers (2003). We describe it briefly next. The declustering involves the estimation of θ, first, as discussed in Section 2.5.1. The method proposes an automatic way to select r. Recall from Section 2.5.1 that the interexceedance times consist of two groups: one group corresponding to the inter-cluster times, and the other one corresponding to the intra-cluster times. Based on the asymptotic theory, it is postulated, that the $1 - \theta$ proportion of the smallest interexceedance times belong to the intra-cluster times, and the rest belong to the inter-cluster times. Therefore, given m sorted interexceedance times, we can take the $(\lfloor m\theta \rfloor + 1)$th interexceedance time as the smallest interexceedance time that separates the clusters. Now, declustering can proceed as in the previous subsection 4.2.1 with $r = \lfloor m\theta \rfloor + 1$.

4.2.3 Combing estimators

Generally the runs method of θ estimation produces the lowest rmse values *once the correct value of r is chosen.* Here by "correct value of r", we mean the one that leads to the smallest rmse - root mean square error - value. Results from Chapter 2 of this thesis shows that the runs method beats Ferro-Segers often with the correct choice of r. Even Ferro and Segers (2003) in their own simulations show that the runs method gives the lowest rmse values especially at higher threshold levels. The

main advantage of the Ferro-Segers estimator is that it does not require the selection of the additional parameter r. Our experience shows that the Ferro-Segers estimator tends to identify r well. Therefore, we propose the following four step procedure to estimate θ and decluster:

1. Choose a high threshold u. We recommend choosing u corresponding to at least the 0.90 quantile of the data.
2. Estimate the runs parameter r via Ferro-Segers, as described in Section 2.5.1.
3. Decluster the data, as discussed in subsection 4.2.1 using the value of r from step 2.
4. Using the runs parameter r from step 2, estimate θ via the runs method, as described in Section 2.5.1.

In Section 4.4.1, we conduct a simulation study to verify our estimation procedure.

4.3 New Risk Measures

In this section, the new risk measures are defined and motivated first. The estimation procedure is then presented.

4.3.1 Definitions

Let $\{X_1, \ldots, X_i, \ldots, X_n\}$, $\mathbb{E}[X_i^2] < \infty$, be a realization of a strictly stationary process with $\theta \in (0, 1]$. The X_i will represent the *negative* daily return series of an asset. For a fixed large threshold u such as $P(X_i < u) = q$ with $q \geq 0.90$, and $r \in \{0\} \cup \mathbb{N}$, define the extreme state beginning at time period i as follows:

$$A_{(i,r)} = \{X_i > u\} \cap \{X_{i+L(\theta)-1} > u\} \cap \{X_{i+L(\theta)} \leq u, \ldots, X_{i+L(\theta)+r} \leq u\}, \tag{4.3.2}$$

where $L(\theta)$, taking values in $\mathbb{N}$, is a random variable denoting the duration of the extreme state in days such that no r consecutive values in the set $\{X_i, \ldots, X_{i+L(\theta)-1}\}$

are less than or equal to u. The smaller the θ, the larger the $L(\theta)$ will be and vice versa. From now on we will let $L(\theta) = L$.

An extreme state always begins when a return value at time i exceeds a high threshold u, that is when a large loss occurs, and it ends when another large loss occurs at the end of L days, followed by r days of "calm", when all returns are below u. When $L = 1$, the extreme state consists of a single large loss. Our measures of risk will be only affected by the values in $\{X_i, \ldots, X_{i+L-1}\}$ but we need to include the event $\{X_{i+L} \leq u, \ldots, X_{i+L+r} \leq u\}$ to demarcate the end of an extreme state in a precise manner. We will treat r as an auxiliary parameter and u as a non-random quantity. The reason we do not define an extreme state simply as a cluster of *consecutive* large losses is that financial data could exhibit a pattern in which large losses are inter-dispersed with gains. This motivates the following definitions for new measures of risk:

Extreme State Cumulative Return:

$$M_1(u, r) = \mathbb{E}_{X,L}\Big[\sum_{j=i}^{L+i-1} X_j \;\Big|\; A_{(i,r)} \Big]. \tag{4.3.3}$$

Extreme State Cumulative Loss:

$$M_2(u, r) = \mathbb{E}_{X,L}\Big[\sum_{j=i}^{L+i-1} X_j \mathbf{I}\{X_j > 0\} \;\Big|\; A_{(i,r)} \Big]. \tag{4.3.4}$$

Extreme State Cumulative Excess Loss:

$$M_3(u, r) = \mathbb{E}_{X,L}\Big[\sum_{j=i}^{L+i-1} X_j \mathbf{I}\{X_j > u\} \;\Big|\; A_{(i,r)} \Big]. \tag{4.3.5}$$

Extreme State Adjusted Expected Shortfall:

$$M_4(u, \theta) = \mathbb{E}_X[\, X_i \mid X_i > u \,]/\theta. \tag{4.3.6}$$

The notation $\mathbb{E}_{X,L}$ emphasizes that the expectation is taken with respect to X and L. However, for ease of notation, we will drop the subscripts from now on. Similarly,

$M_i(u, r)$ emphasizes that the risk measure is a function of u and r, but again we will drop the u and r from the notation henceforth.

Typically, the expected shortfall is defined as

$$\mathbb{E}[\ X \mid X > \text{VaR}\ ].$$

However, we define our risk measures relative to a fixed threshold u rather than VaR for the following reasons:

1. The estimation of VaR would introduce additional complexity and further uncertainty into our risk measures. Estimation of extreme quantiles such as VaR using the empirical distribution of the data leads to poor results as shown in Novak and Beirlant (2006). We believe choosing u as a fixed quantity overcomes this issue *without reducing the practical usefulness of the risk measure.* Otherwise, we would have to resort to semi-parametric estimation methods borrowed from the extreme value theory, for example in McNeil and Frey (2000) or extrapolation methods as suggested by Inui and Kijima (2005).

2. Theoretically, the extremal index, and thus the declustering results should be independent of the chosen threshold. However, as pointed out by Ancona-Navarrete and Tawn (2000), most existing estimators actually estimate $\theta(u)$, the extremal index, as a function of the threshold, rather than θ and conclude that the established methods do well only when $\theta(u) \approx \theta$. By defining our risk measures using u, the degree of clustering and the large losses are both defined relative to the same threshold. This keeps our risk measures consistent in terms of our definition. Estimating θ at VaR , generally leads to $\hat{\theta} \approx 1$ since VaR typically corresponds to very high quantile of the data with very few data points. It is even possible that the estimated VaR is not even an observed value

of the data.

We will refer to $\mathbb{E}[\ X \mid X > u\ ]$ as the *adjusted* expected shortfall.

M_1 measures the mean total (negative) return during an extreme state. We use the word "return" rather than loss since this is just a straight forward mean of the sums of the returns for each extreme state. It is possible that during an extreme state for several gains to occur, which offset the losses. Thus, M_1 can take negative values or values close to 0. For each extreme state, M_2 takes into account only the losses whereas M_3 measures the total excess losses. We always have that $M_2 \geq M_1$ and $M_2 \geq M_3$. The measure M_4 is simply an adjustment made to the usual expected shortfall, with u in place of VaR, to account for the clustering of large values; the idea is to incorporate the factor $1/\theta$ in order to account for the expected size of a cluster. When no clustering is present, theoretically $\theta = 1$ and with $L = 1$, then $M_1 = M_2 = M_3 = M_4 = \mathbb{E}[\ X_i \mid X_i > u\ ]$. This would for example be the case for iid data, when on average large losses occur in an isolated manner, rather than in clusters. Finally, when $\mathbb{E}[L] = 1/\theta$, and the extreme state consist only of losses, M_3 reduces to M_4. Therefore, M_4 can be viewed as a proxy and a more convenient way to estimate M_3, since only an estimate of θ and the adjusted expected shortfall without declustering is needed.

Perhaps the most important risk measure from the perspective of the risk manager is M_2. If clustering of large losses is present, then a large value of M_2 tells the risk manager that the assets can experience large losses in short durations of time on average. This clustering of large losses could lead to the triggering of margin calls, automatic sell offs, or severe capital withdrawals. As mentioned before, since losses can be inter-dispersed with gains, the risk manager can take this into account in M_1. A large discrepancy between M_1 and M_2 would tell the risk manager that the

extreme states consists of both large losses and large gains. When $M_1 \approx M_2$ then the extreme states consist mostly of large losses. M_3 and M_4 may be of interest when excess losses trigger a transaction such as insurance or re-insurance payments made over a pre-specified level.

4.3.2 Estimation

In this section, we give a step by step procedure for estimating the risk measures and provide relevant comments. Our assumption is that the estimated risk measures for an observed period of time give reasonable estimates for the next *unobserved extreme state time period*. This is a reasonable assumption for stationary data. Later in Section 4.4.2, we elaborate on how to check the validity of the estimation procedure.

1. Choose a reasonably large window size, n, for example $n = 1000$. We will use this value for our data analysis.

2. Choose a threshold u relevant to the asset. We recommend setting u to at least $q = 0.90$ quantile of the data. This would give $\lfloor n(1-q) \rfloor$ large losses or 100 when $q = 0.90$ and $n = 1000$. For $q = 0.90$, typical values of u for stocks and stock indices can be chosen around 1.5%, 1% for treasury bills and bonds, and about 2-3% for crude oil and natural gas.

3. Estimate the extremal index as outlined in subsection 4.2.3. If $\hat{\theta} = 1$, then you can assume $M_1 = M_2 = M_3 = M_4 = \mathbb{E}[\ X_i \mid X_i > u\ ]$. At this point there is no use in going any further. One can just resort to estimating the adjusted expected shortfall by any established methods.

4. When $\hat{\theta} < 1$, decluster the data as outlined in subsection 4.2.3, obtaining m clusters. Note $1 \leq m \leq 99$ if u is set to the 0.90 quantile of the data and

$n = 1000$. You will have m sets of approximately independent clusters. Denote the cluster sizes as $\{L_1, \ldots, L_m\}$. We will represent the corresponding data within the clusters as $X_{i,j}$ where the second subscript j represents the cluster number, $j \in \{1, \ldots, m\}$, and the first subscript i, $1 \leq i \leq L_j$, represents the element number within the cluster j.

5. Estimate the risk measures by just the empirical versions of the equations (4.3.3) to (4.3.6) as follows:

$$\widehat{M_1} = \sum_{j=1}^{m} \Big(\sum_{i=1}^{L_j} X_{i,j} \Big) / m. \tag{4.3.7}$$

$$\widehat{M_2} = \sum_{j=1}^{m} \Big(\sum_{i=1}^{L_j} X_{i,j} \mathbf{I}\{X_{i,j} > 0\} \Big) / m. \tag{4.3.8}$$

$$\widehat{M_3} = \sum_{j=1}^{m} \Big(\sum_{i=1}^{L_j} X_{i,j} \mathbf{I}\{X_{i,j} > u\} \Big) / m. \tag{4.3.9}$$

$$\widehat{M_4} = \frac{\sum_{k=1}^{n} X_k \mathbf{I}\{X_k > u\}}{\lfloor n(1-q) \rfloor \hat{\theta}}. \tag{4.3.10}$$

That is sum the values within each cluster, and then take the sample mean of the resulting cluster statistics. For later reference, we define the following quantities as well:

$$S_j^{M_1} = \sum_{i=1}^{L_j} X_{i,j}. \tag{4.3.11}$$

$$S_j^{M_2} = \sum_{i=1}^{L_j} X_{i,j} \mathbf{I}\{X_{i,j} > 0\}. \tag{4.3.12}$$

$$S_j^{M_3} = \sum_{i=1}^{L_j} X_{i,j} \mathbf{I}\{X_{i,j} > u\}. \tag{4.3.13}$$

Note that our estimation is fully non-parametric.

4.4 Simulations

In this section, we conduct two sets of simulation experiments. The first corresponds to checking that our 4 step combined runs and Ferro-Segers estimation gives better results than just the Ferro-Segers when estimating θ. The second set of simulations is done to validate that our estimated risk measures correctly predict the next unobserved value.

We will use three linear processes for our simulations. The noise terms $\{Z_i\}_{i\in\mathbb{Z}}$ for all the processes will have t-distribution with 4 degrees of freedom. Recall that the tail index, α, for a t-distribution is equal to its degrees of freedom so that $P(Z_i > z) \sim cz^{-4}$, as $z \to \infty$, and c is a constant. Given $(a_0, a_1, a_2, a_3, a_4) = (0.9, 0.8, 0.7, 0.6, 0.5)$, the three processes are:

$$X_n(1) = \sum_{i=0}^{4} a_i Z_{n-2i} \; - \; \sum_{i=0}^{4} a_i Z_{n-2i-1}, \tag{4.4.14}$$

$$X_n(2) = \sum_{i=0}^{4} a_i Z_{n-2i}, \tag{4.4.15}$$

$$X_n(3) = Z_n. \tag{4.4.16}$$

The theoretical values of the extremal index computed via Corollary 5.5.3 in Embrechts *et al.* (1997) for both $X_n(1)$ and $X_n(2)$ is 0.4380132. The qualitative extremal behavior of $X_n(1)$ is quite different than $X_n(2)$ even though they both have the same extremal index. The extremes of $X_n(1)$ tend to whipsaw, whereas the extremes of $X_n(2)$ occur in successions. $X_n(3)$ is just an iid process and its extremal index is 1. See Figure 4.2 for the sample plots.

Note, we do not assume or model the asset returns with simple linear processes. However, some important features of the asset returns can be replicated with the

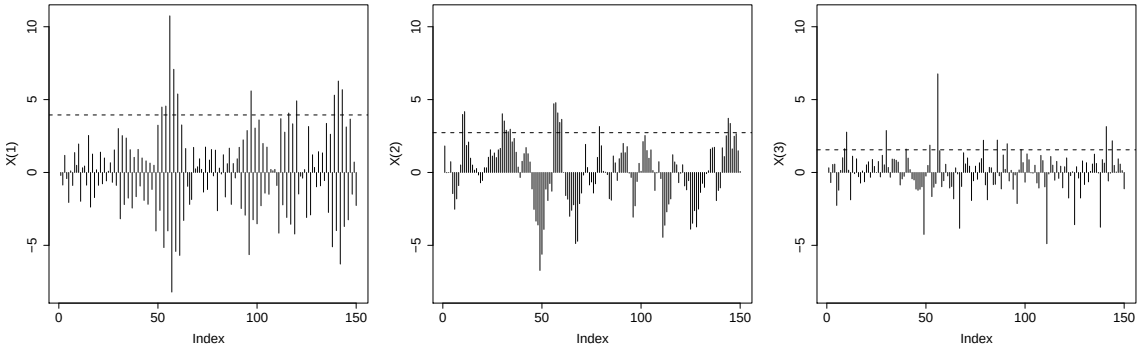

Figure 4.2: From left to right, sample paths of the $X_n(1)$, $X_n(2)$, and $X_n(3)$ processes as defined in (4.4.14) to (4.4.16). The dashed lines in each plot correspond to the 0.90 quantile of the data.

above models, namely the clustering of the extremes, the tendency to have large losses followed by large gains and the heavy tail behavior of the returns.

4.4.1 Combined Estimator RMSE Simulations

For our simulation study, we generate 1000 sets of values of $X_n(1)$, $X_n(2)$, and $X_n(3)$, with sample sizes of $n = 1000$ and 10000. At each run, the extremal index is estimated via the Ferro-Segers estimator as described in Section 2.5.1, and the proposed combined estimator as described in Section 4.2.3 with the thresholds set to the qth quantile of the data with $q \in \{0.85, 0.86, \ldots, 0.98, 0.99\}$. The rmse at each threshold level is estimated for each process. The results are summarized in Figure 4.3. The top row is for $n = 1000$ and the bottom row is for $n = 10000$. The combined estimator clearly beats the Ferro-Segers estimator for $X_n(1)$. In case of the $X_n(2)$ process, the combined estimator outperforms the Ferro-Segers for the quantiles above $q > 0.90$. Since the thresholds are generally set at large quantiles, and we are concerned about clustering of large losses at larger quantiles, this is good news too. For the $X_n(3)$ process, the Ferro-Segers estimator marginally outperforms

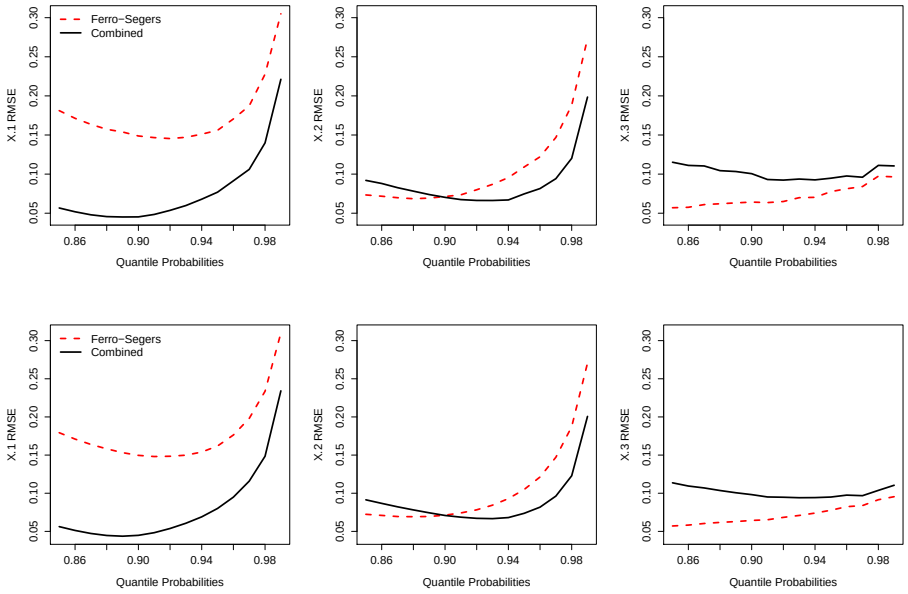

Figure 4.3: RMSE values for the combined method estimation of θ. Top row is for $n = 1000$, bottom row is for $n = 10000$.

the combined estimator. This can be attributed to the fact that for iid data, we do not need the additional sophisticated declustering step, which adds to the variance of r and therefore leads to a small bias in estimating the extremal index. The combined estimator tends to underestimate θ. However, the gap is small and becomes even smaller for $n = 10000$. This would lead to the overestimates of the risk measures. Although this is not a desirable effect, it would have been even more problematic, if the estimation underestimated risk.

4.4.2 Cluster Prediction Simulations

In this section, we carry out a simulation study to validate our proposed risk measures by examining their predictive ability. In the risk management literature, this process of observing the predictive performance of a risk measure is called *back-testing*. The main idea is to generate a long sequence of a desired process, estimate

the risk measures in a small subset of the sequence and compare the result with the next out-of-sample value. We will still use the same processes in equations (4.4.14) to (4.4.16) represented by $X(k)$, $k \in \{1, 2, 3\}$. Next, we describe the simulation and backtesting procedures.

1. Generate N data points from $X(k)$, where N is the length of the long sequence of the process.

2. Choose $n \ll N$, where n is the size of the small subset of the sequence or the *window size*.

3. Set $c = 1$. The final value of $c = C$ will be the eventual number of predictions we will make. This quantity is random and unknown in the beginning.

4. For $\{X_c(k), \ldots, X_{n+c-1}(k)\}$, set u to the qth quantile of the data, estimate θ and decluster by the combined method, as described in Section 4.2.3. The outputs from this step are $\hat{\theta}$, r, m set of clusters, and $\widehat{S_j^{M_1}}$, $\widehat{S_j^{M_2}}$, $\widehat{S_j^{M_3}}$, $j \in \{1, \ldots, m\}$.

5. Estimate M_1 to M_4 as defined in (4.3.3) to (4.3.3) via equations (4.3.7) to (4.3.10) to obtain $\widehat{M_1}$, $\widehat{M_2}$, $\widehat{M_3}$, and $\widehat{M_4}$. We will also refer to these quantities as *in-sample* values.

6. Recalling the definition of the extreme state in equation (4.3.2), obtain the one cluster ahead risk measures by rolling forward through the data starting from $n + c$ until the next extreme state is identified. Use r from Step (4) to identify the cluster ahead. Compute the cluster ahead risk measures to obtain M_1^+, M_2^+, $M_3^+ = M_4^+$, where the superscript $+$ represents the one cluster ahead. We will also refer to these quantities as *out-of-sample* values. Let the position of the termination of the one cluster ahead be l.

7. Obtain the following *standardized differences* at each step c:

$$\widehat{D_c^{M_1}} = \frac{M_1^+ - \widehat{M_1}}{SD(\widehat{S_j^{M_1}})}, \tag{4.4.17}$$

$$\widehat{D_c^{M_2}} = \frac{M_2^+ - \widehat{M_2}}{SD(\widehat{S_j^{M_2}})}, \tag{4.4.18}$$

$$\widehat{D_c^{M_3}} = \frac{M_3^+ - \widehat{M_3}}{SD(\widehat{S_j^{M_3}})}, \tag{4.4.19}$$

$$\widehat{D_c^{M_4}} = \frac{\hat{\theta}(M_4^+ - \widehat{M_4})}{SD(\sum_{j=1}^{n} X_j \mathbf{I}\{X_j > u\}/\lfloor n(1-q)\rfloor)}, \tag{4.4.20}$$

and the following *empirical proportions*:

$$\widehat{P_c^{M_1}} = \frac{\#\{M_1^+ > \widehat{M_1}\}}{m}, \tag{4.4.21}$$

$$\widehat{P_c^{M_2}} = \frac{\#\{M_2^+ > \widehat{M_2}\}}{m}, \tag{4.4.22}$$

$$\widehat{P_c^{M_3}} = \frac{\#\{M_3^+ > \widehat{M_3}\}}{m}, \tag{4.4.23}$$

where for example, $\#\{M_1^+ > \widehat{M_1}\}$ counts the number of clusters such that

$$M_1^+ > S_j^{M_1},$$

for $j \in \{1, \ldots, m\}$.

8. Set $c = l - n + 1$ so the window is $\{X_{l-n+1}(k), \ldots, X_l(k)\}$. Therefore, we move the window of the data up to the end of the one cluster ahead.

9. Increment c by 1, go back to step (4) and repeat until the end of the data sequence is reached. We will now have C standardized differences and empirical proportions.

To test our results we conduct two tests:

1. Under the null hypothesis that risk measures are being correctly estimated, the mean of the standardized differences is zero. That is

$$H_0 : \mathbb{E}[\widehat{D_c^{M_i}}] = 0, \qquad H_1 : \mathbb{E}[\widehat{D_c^{M_i}}] > 0, \tag{4.4.24}$$

for all $i \in \{1, 2, 3, 4\}$. We choose the alternative to be one sided to the right, since the underestimation of the risk is of most concern here. However, the differences tend to be severely skewed and a simple t-test is not applicable. Therefore, we use a bootstrap test as described on page 224 of Efron and Tibshirani (1993) with no assumptions about the distribution of the standardized differences. Even though the tails of the standardized differences may be heavy, using the bootstrap is justified as long as $\mathbb{E}[(D_c^{M_i})^2] < \infty$, that is the second moment of the standardized differences exist. This fact is proved in Mason and Shao (2001). See Section 4.7 for the **R** code used in bootstrapping the mean.

2. Under the null hypothesis that the one cluster ahead risk measures come from the same distribution as the past m clusters, the sequence of the empirical proportions $\widehat{P_c^{M_i}}$ will follow a uniform distribution. That is, for $\{1 \leq c \leq C\}$, we expect to have

$$\widehat{P_c^{M_i}} \overset{d}{\approx} \text{iid uniform}(0, 1),$$

for all $i \in \{1, 2, 3\}$. This is simply due to the fact that for any continuous random variable $Y = 1 - F_Z(Z)$,

$$\mathbb{P}\{Y > y\} = \mathbb{P}\{Z < F_Z^{-1}(1 - y)\} = F_Z(F_Z^{-1}(1 - y)) = 1 - y,$$

so Y is a uniform $(0, 1)$ random variable. Empirical proportions are just the empirical versions of $\mathbb{P}\{Y > y\}$. A histogram of the empirical proportions

	$X(1)$		$X(2)$		$X(3)$	
	$n = 1000$	$n = 10000$	$n = 1000$	$n = 10000$	$n = 1000$	$n = 10000$
M_1	0.5240	0.4370	0.1870	0.4390	0.5110	0.4920
M_2	0.3100	0.4380	0.1910	0.4340	0.5000	0.5080
M_3	0.3010	0.4330	0.1960	0.3920	0.5040	0.5200
M_4	0.2790	0.3870	0.2040	0.3950	0.5250	0.5390

Table 4.1: P-values for testing the hypothesis that the standardized differences are zero.

should yield a uniform distribution under the null hypothesis. One could even perform a goodness of fit test but we will not pursue this here.

In testing our results, we are making the assumption that the clusters are approximately independent and this independence becomes more pronounced as the separation between the cluster grows. For more details on the theoretical backing of this assumption refer to Hsing *et al.* (1988).

We run our simulations with $n = 1000$ and $n = 10000$ for $N = 20000$, and $N = 200000$ respectively. The threshold value u is always set to the 0.90 quantile of the data. This means that when $\theta < 1$, $\max(m) = 99$ for $n = 1000$, and $\max(m) = 999$, when $n = 10000$. The main function written in **R** is in Section 4.7. Table 4.1 contains the p-values for testing the hypothesis in equation (4.4.24). All p-values are greater than the most common significance levels of 0.01 to 0.10. Furthermore, with the exception of two cases, the p-values improve for larger sample sizes.

It is also worthwhile to examine the standardized differences. We pick two sets: Figures 4.4 and 4.5 show the histogram, boxplots, and time plots of $\widehat{D_c^{M_1}}$ for $X(1)$ and $\widehat{D_c^{M_3}}$ for $X(3)$ with $n = 1000$ and $n = 10000$. Although two sets were picked, the results for the rest of the standardized differences are quite similar. The obvious feature of the standardized differences is their right skewness. The standardized differences $\widehat{D_c^{M_1}}$ for $X(1)$ tend to have a slightly more symmetrical distribution since this process could produce large gains and losses within the same cluster. The standardized differences $\widehat{D_c^{M_3}}$ for $X(3)$ are *very* right skewed since the iid process

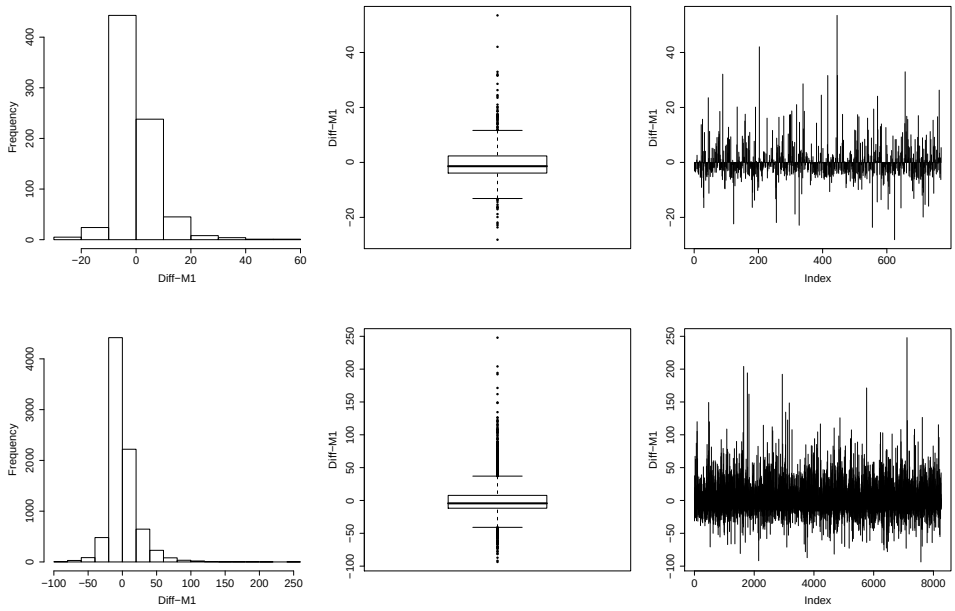

Figure 4.4: Plots of $\widehat{D_c^{M_1}}$ for $X(1)$. Top row is for $n = 1000$, bottom row is for $n = 10000$.

on the average produces only large losses within clusters even though they rarely happen.

The histograms of the empirical proportions for all the risk measures and the processes are shown in Figures 4.6 - 4.8. In all the cases, the empirical proportions appear uniform and the fit to uniform seems to improve for $n = 10000$. There is a slight peak or "bump" in the smaller values of the histograms in Figure 4.8 for the $X(3)$ process. This indicates that in slightly higher percentages of the times, the out-of-sample values were less than the in-sample values.

Finally, we examine the extremal index estimates, as shown in Figure 4.9. As in the previous figures, the top row is for $n = 1000$, and the bottom row is for $n = 10000$. The dashed lines in the plots indicate the theoretical θ values. Our combined estimator seems to be doing a good job. The variability and bias seem to be substantially reduced for the $X(1)$ process with the large sample size. A slight

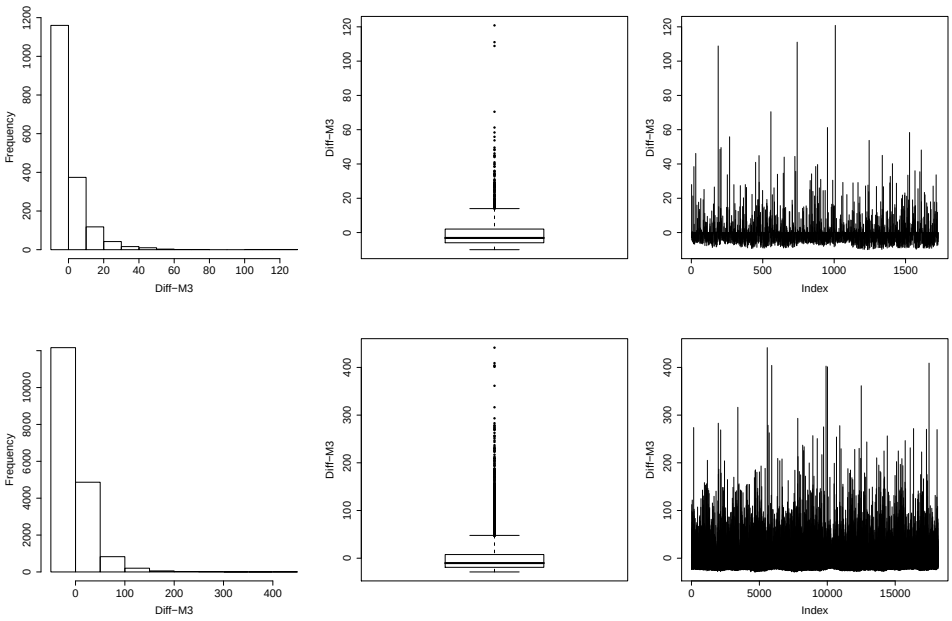

Figure 4.5: Plots of $\widehat{D_c^{M_3}}$ for $X(3)$. Top row is for $n = 1000$, bottom row is for $n = 10000$.

bit of bias is present for the $X(2)$ process. Although the variance appears reduced for the $X(3)$ process, $\hat{\theta}$ values seem to oscillate between 1 and 0.9.

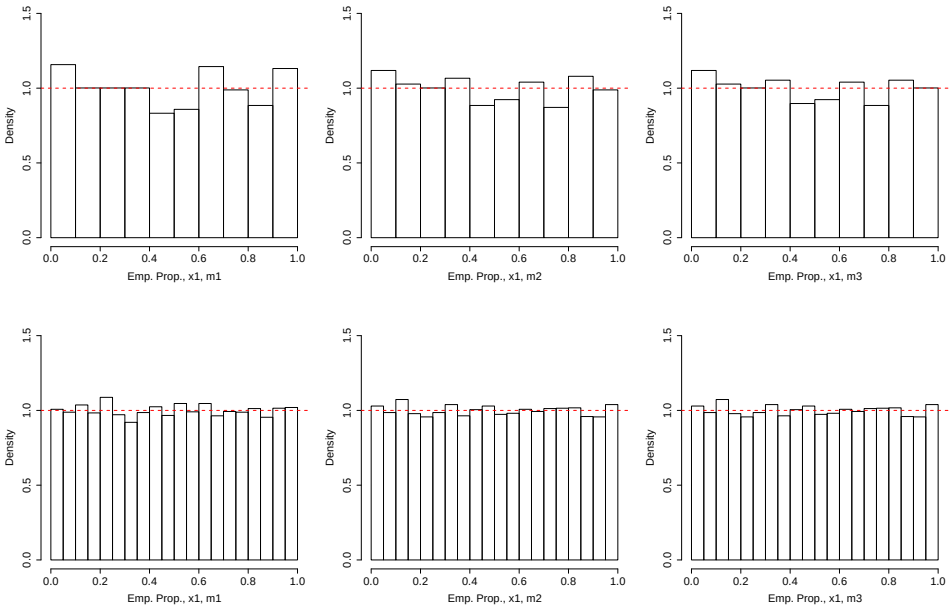

Figure 4.6: Histogram of the empirical proportions for the $X(1)$ process. Top row is for $n = 1000$, bottom row is for $n = 10000$.

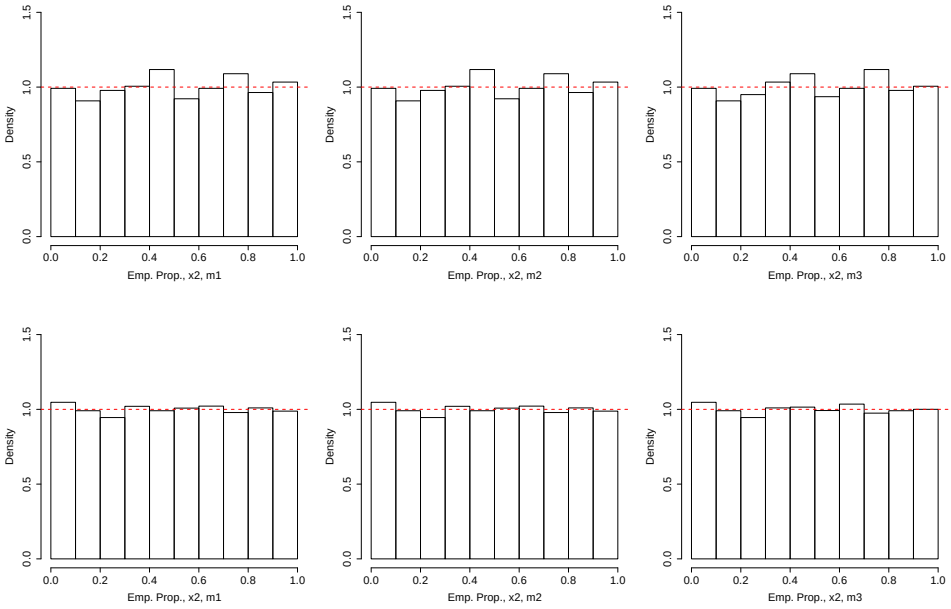

Figure 4.7: Histogram of the empirical proportions for the $X(2)$ process. Top row is for $n = 1000$, bottom row is for $n = 10000$.

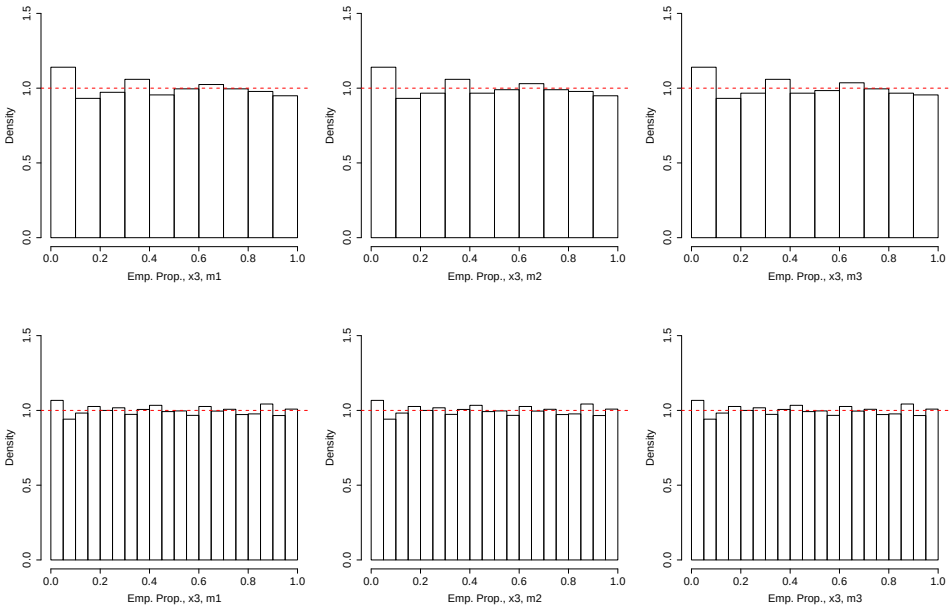

Figure 4.8: Histogram of the empirical proportions for the $X(3)$ process. Top row is for $n = 1000$, bottom row is for $n = 10000$.

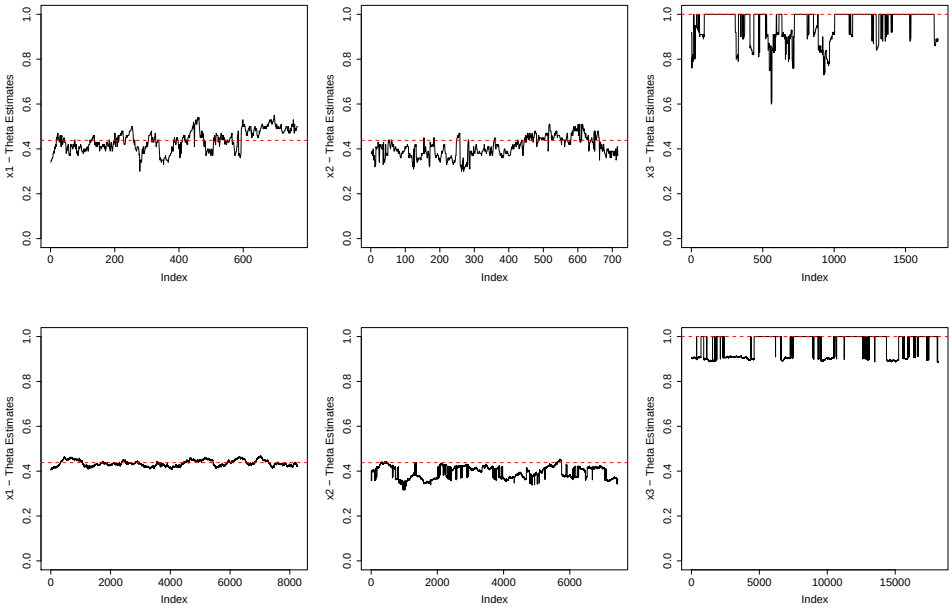

Figure 4.9: Extremal Index estimates for $X(1)$, $X(2)$ and $X(3)$. Top row is for $n = 1000$, bottom row is for $n = 10000$.

	M_1		M_2		M_3		M_4	
	Gains	*Losses*	*Gains*	*Losses*	*Gains*	*Losses*	*Gains*	*Losses*
S&P 500	0.1320	0.0010	0.1110	0.0150	0.0980	0.0100	0.1550	0.0440
WTI Oil	0.7900	0.0180	0.5530	0.0320	0.5110	0.0340	0.3570	0.0180
IBM	0.0000	0.0040	0.0020	0.0010	0.0030	0.0010	0.0380	0.0030
ATT	0.1940	0.2260	0.2040	0.2090	0.1780	0.1800	0.0460	0.1770
JNJ	0.3410	0.1380	0.2900	0.1970	0.2820	0.1840	0.5540	0.1470
JPM	0.0100	0.0260	0.0030	0.0180	0.0050	0.0180	0.0250	0.0270
10 *Year Bond*	0.0020	0.0810	0.0010	0.0210	0.0030	0.0220	0.0020	0.0190
USD/GBP	0.6650	0.9490	0.6930	0.9350	0.6530	0.9050	0.5160	0.9130

Table 4.2: P-values for testing the null hypothesis of zero mean for the standardized differences.

4.5 Application of the New Risk Measures to Financial Data

In this section, we use the same data as described in Table 3.1 to estimate our defined risk measures, and backtest them by reporting the p-values and examining the empirical proportions. We follow the same estimation and backtesting procedure as we did for the simulated data with $n = 1000$ and u set to the 0.90 quantile of the data. We conduct our analysis for both the gains and the losses. Table 4.2 contains the p-values for testing the null hypothesis of zero mean for the standardized differences, as stated in equation (4.4.24) for all our risk measures.

Taking the level of significance to be 5%, we have good results in 32 out of 64 cases, i.e. the p-value is greater than 0.05. The success of our backtesting seems to depend on the asset and sometimes on the gains or the losses. The p-values for the S&P 500 and Oil gains are good but the losses indicate that the risk measures are being underestimated. ATT, JNJ, and USD/GBP give excellent results while IBM, JPM, and the 10 Year Bonds returns give poor results. The worst offender seems to be IBM.

To see the effects of the clustering, we examine Figure 4.10 of the estimated M_i, $i = \{1, 2, 3, 4\}$, measures for IBM and USD/GBP along with the concurrent estimated 1 day *adjusted expected shortfall* (ES). In both cases, the estimated risk

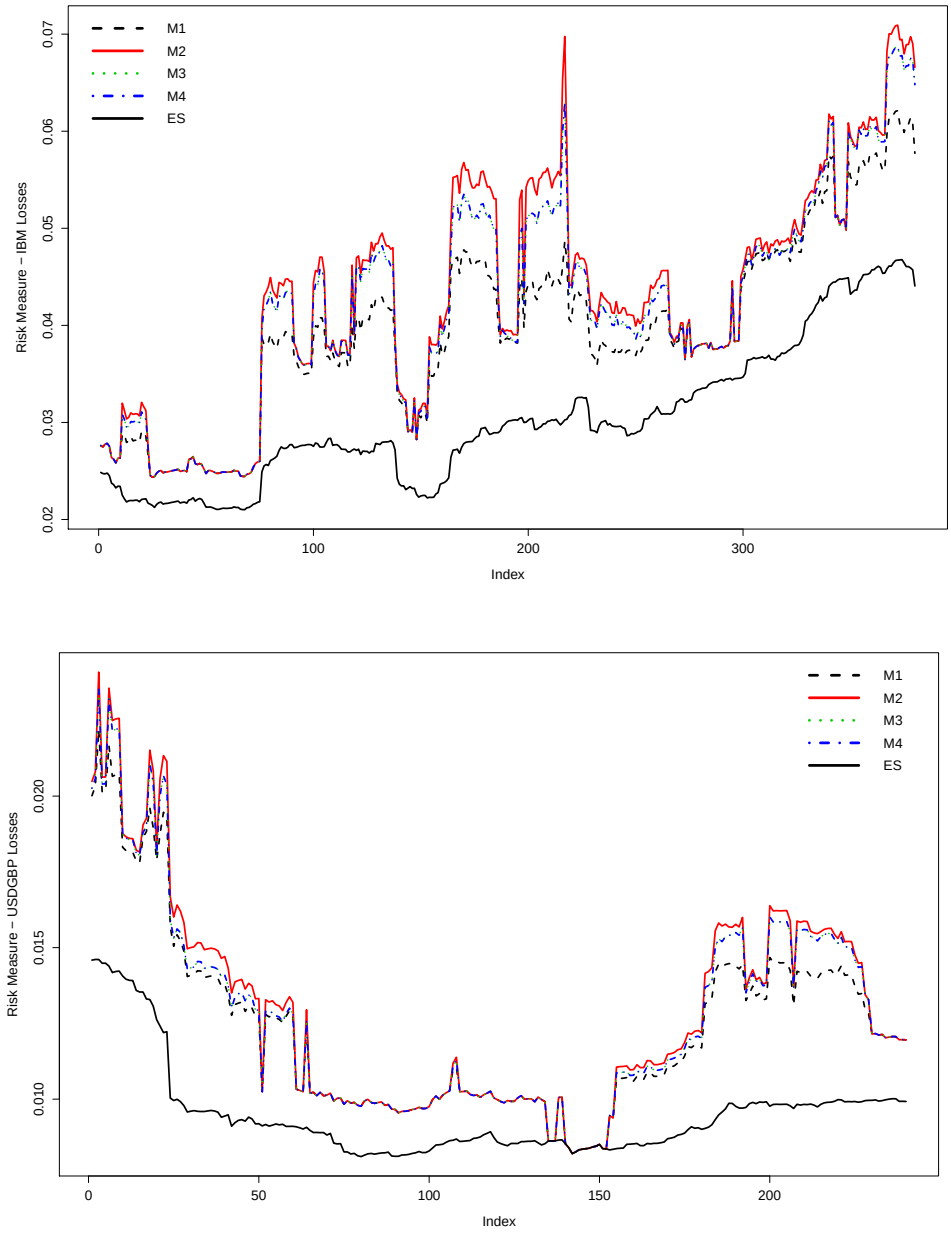

Figure 4.10: The top plot the estimated IBM M_i losses. The bottom plot is the estimated USD/GBP M_i losses. "ES" is just the estimated 1 day expected shortfall for the same period of estimation of the M_i measures. Thus ES ignores clustering.

measures are significantly greater than the 1 day adjusted expected shortfall with the differences being more in the case of IBM. If no clustering were present, then all the risk measures would be approximately equal to the 1 day adjusted expected shortfall.

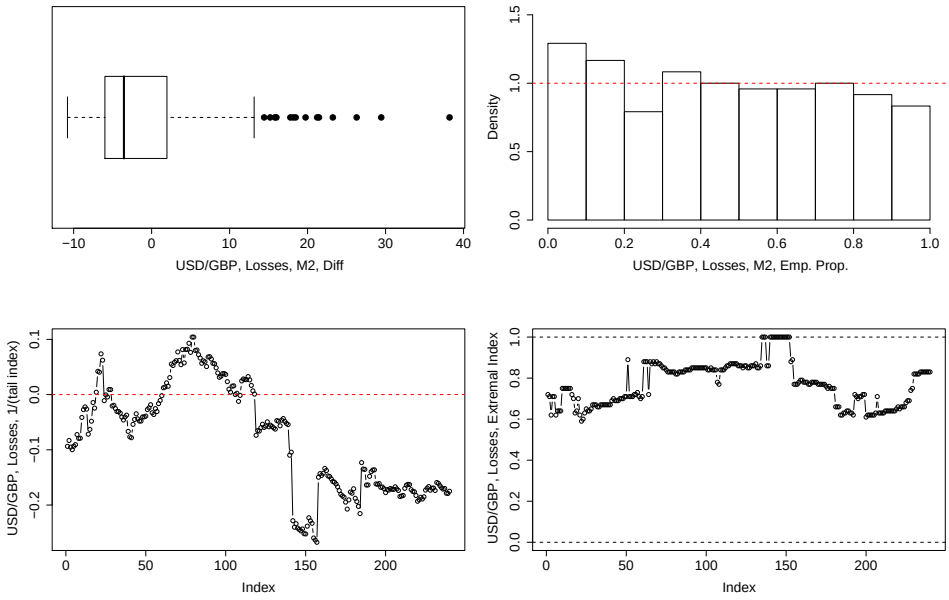

Figure 4.11: USD/GBP losses diagnostics plots for M_2 with ξ and θ estimates of the returns.

In the IBM case, M_4 seems to track M_3 well, supporting our assertion that M_4 is a proxy for M_3. For a significant number of time periods, there are discrepancies between M_2 and M_1 which indicates that large losses and possibly large gains mingle together in the extreme states. In the USD/GBP case, the risk measures tend to be approximately equal except for later time periods for M_1 and M_2. It appears that a few gains occur in the extreme states. There is even a time from about the 140th to 150th periods when all the risk measures collapse into the 1 day adjusted expected shortfall. Similar to the IBM case, M_4 tracks M_3 well.

We will now focus on the IBM and USD/GBP loss sides and their corresponding M_2 risk measures for further investigations of the results. We start with the USD/GBP diagnostics in Figure 4.11. The boxplot shows a number of outliers for $\widehat{D_c^{M_2}}$. The plot certainly shows that the distribution is right skewed. The empirical proportions look approximately uniform except there is a bump in the lower part of the histogram. This indicates that a slightly higher percentages of the out-of-sample

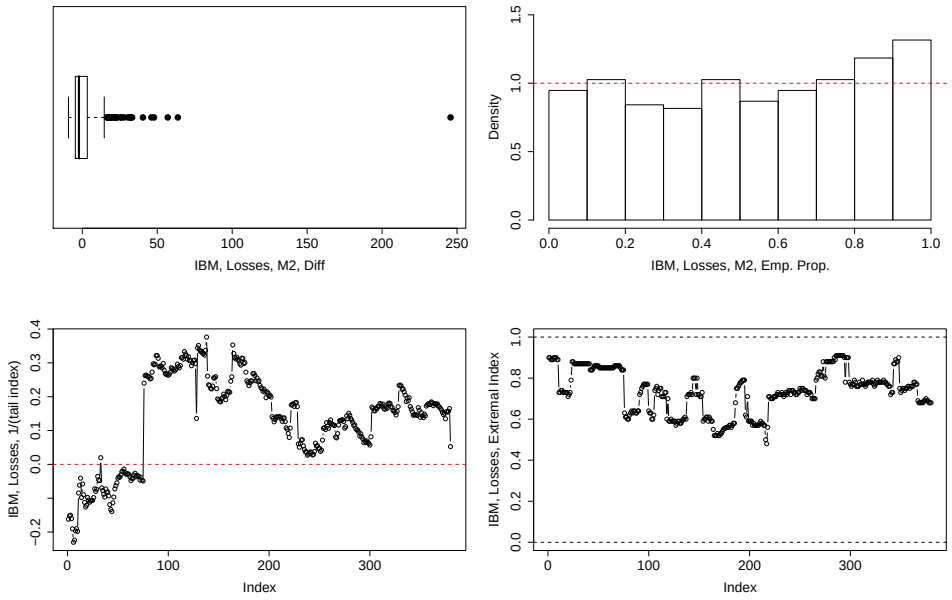

Figure 4.12: IBM losses diagnostics plots for M_2 with ξ and θ estimates of the returns.

values were less that the in-sample estimated risk measures. The GPD-MLE estimates of ξ for the negative *returns* of USD/GBP imply light tails but recall from Chapter 3 that GPD-MLE underestimates ξ in the presence of dependence. The extremal index estimates are based on the combined estimator. Extreme loss clustering is present except from about the 140th to the 150th time periods, when the extremal index is estimated to be 1. We also observed this in the lower plot of Figure 4.10, where all the risk measures are the same in the absence of clustering. Note both ξ and θ estimates show stability, in a sense that the estimated values tend not to change for long time periods.

The IBM diagnostics tell a different story as seen in Figure 4.12. First note the extreme outliers present in the boxplot of $\widehat{D_c^{M_2}}$. The distribution of $\widehat{D_c^{M_2}}$ is very skewed. Furthermore, the GPD-MLE estimates of ξ of IBM negative returns at certain periods come dangerously close to the value of 0.50, and may well be 0.5 due to the underestimation by GPD-MLE. When $\xi = 0.50$, the second moments do not

exist. This would violate our assumption that the second moments for the return series exist. Our assumption that standardized differences possess a second moment is also in jeopardy. A Hill plot of ξ for $\widehat{D_c^{M_2}}$ is shown in Figure 4.13. It appears that the $\widehat{D_c^{M_2}}$ have infinite second moment but the first moment exists. Further details on the estimation of $1/\xi$ via Hill plots are in Hill (1975). Since the first moment appears to exist, our *hypothesis* of testing for the mean of zero is still valid but our bootstrap procedure of testing for this hypothesis appears not to be valid. We would like to argue that the accuracy of our risk measures in predicting the out-of-sample values is not correctly reflected by the low p-values for the IBM case, since the bootstrap procedure is not valid. Furthermore, even though there is a slight peak in the histogram of the empirical proportions near 1, the bump is slight; had our risk measure systematically underestimated the out-of-sample values, this bump would have appeared much larger. As a simple first attempt to correct for the presence of extreme skewness, leaving the 5 largest $\widehat{D_c^{M_2}}$ out of the 380 values, and redoing the bootstrap test increases the p-value to 0.12 from 0.001. This is a drastic change. Admittedly, this is too ad hoc of a procedure to follow but we plan to investigate ways to correct our bootstrap procedure for our purposes. Our investigation of the other risk measures for IBM, and the rest of the data sets showed a general trend: the heavier the standardized differences, the worse the p-values. However, the empirical proportions did not indicate that a high proportion of the times, the risk measures were being underestimated.

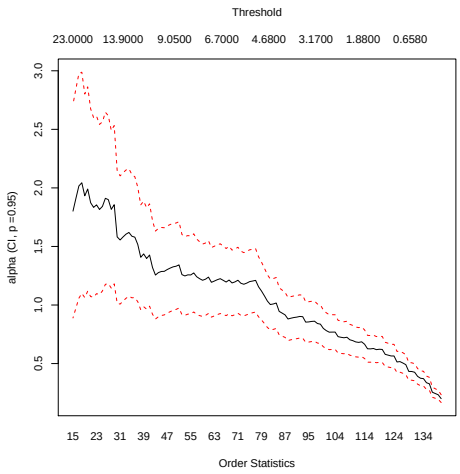

Figure 4.13: IBM Hill plot estimates of $\alpha = 1/\xi$.

4.6 Final Remarks

We have statistical evidence that the risk measures we have defined capture well the clustering effect in the extreme losses. To the best of our knowledge, the effect of clustering has never been taken into account explicitly, in the available literature, when measuring risk. We view temporal dependence and clustering as an integral part to measuring and predicting risk. Furthermore, our combined estimator is better able to estimate the extremal index in terms of rmse. In a practical scenario, we recommend estimating θ first. If $\theta \approx 1$, then there is no need to estimate any of our risk measures. Otherwise, estimate M_1, and M_2 next. A large difference between these two measures indicates that the extreme states consist of both large losses and gains. In this case, M_2 can be taken as an average estimate of the worst case scenario, and M_1 would provide an estimate of the average loss in an extreme state. If $M_1 \approx M_2$, then M_2 should be taken as the risk measure. If excess losses above a certain threshold are of concern, then M_3 would be the risk measure of choice. Recall also that M_4 provides an estimate of M_3 with the advantage that no

declustering needs to be done.

There are a number issues that need further attention and work. We discuss them next.

Perhaps the foremost issue is that of confidence intervals for our defined risk measures. We have not addressed how to create confidence intervals for our estimated risk measures. Confidence intervals for risk measures have not been fully studied in the available literature, and only recently researchers have begun to look into it. The latest explicit focus on the confidence intervals is in Christoffersen and Gonçalves (2005). Investigation of the asymptotic normality and consistency of our estimators will be helpful in this matter. Another approach would involve the bootstrap.

The next pressing issue is that of the bootstrapping test for the mean zero null hypothesis. If the data have heavy tails such that the moments of order greater than or equal to 2 do not exist, then the normalized bootstrapped distribution of the sample mean will not converge to the limit of the sample mean when the bootstrap resample size is equal to the sample size. These issues were investigated and reported in Athreya (1987) and Knight (1989). A fix, at the expense of efficiency and change of the test statistic, is suggested by setting the bootstrap resample size $l = l(n)$ to satisfy $l = l(n) \to \infty$, with $l/n \to 0$ as $n \to \infty$. DasGupta (2008) suggests setting $l = 2\sqrt{n}$, but we do not know the origins for this rule of thumb. We plan to investigate this area since testing for the null hypothesis of mean zero is of critical importance to the validation of our work.

Finally, in our work we have taken the liberty to define a new expected shortfall as the conditional expectation of a loss given that a return has exceeded a non-random quantity u. Certainly the additional step of estimating the VaR introduces more complexity and uncertainty for our procedure. We plan to investigate incorporating

the estimation of the VaR with our own defined risk measures. This, we believe, is merely a technical detail and all our conclusions should remain the same.

4.7 Appendix

The following code is the main function for backtesting.

```
obtain.risk.measure.differences.5 = function(x, u.prob = 0.90,
use.default.ei = T, ei.default = 1,
window.size=1000, end.cut.off = 250, gpd.u.prob = (u.prob -0.05))
{
  n                  =  length(x)
( This part of the code for the main function is omitted.)
  i = 0
  repeat
  {
    if ( (final.index + end.cut.off) >= n) break
    i = i + 1
    temp.data  =  x[start.index : final.index]
    u        =  quantile(temp.data,   prob = u.prob, names=FALSE)
    z        =  (temp.data > u)
    if (use.default.ei)
    {
      dc.obj   =  decluster.intervals(z, ei.default)
      r.par[i] =  dc.obj$par
      ei[i]    =  ei.default
    }
    if (!use.default.ei)
```

```
{
  ei[i] = exi.intervals.modified(z)
  if ( ei[i] == 1 )
  {
    dc.obj   =  decluster.intervals(z, 1)
    r.par[i] =  dc.obj$par
  }  else
  {
    dc.obj   =  decluster.intervals(z, ei[i])
    r.par[i] =  dc.obj$par
    ei[i]    =  exi(temp.data, u = u, r = r.par[i])
  }
}
index.matrix  =  create.exst.matrix(dc.obj)
no.of.clusters = dim(index.matrix)[1]
no.clus[i]     = no.of.clusters
m.obj = estimate.m.measures.2(temp.data, index.matrix, u = u )
m1[i]    = m.obj$m1
m2[i]    = m.obj$m2
m3[i]    = m.obj$m3
m1.sd[i]    = m.obj$m1.sd
m2.sd[i]    = m.obj$m2.sd
m3.sd[i]    = m.obj$m3.sd
gpd.threshold = quantile(temp.data, prob = gpd.u.prob,
names=FALSE)
```

```
gpd.mle.obj    =  gpd.fit(temp.data, gpd.threshold, show=F)

xi[i]          =  gpd.mle.obj$mle[2]

sigma[i]       =  gpd.mle.obj$mle[1]

one.day.gpd.ES  =  ( u / (1 - xi[i]) ) *

( 1 + (sigma[i] -

xi[i] * gpd.threshold) / u )

m31[i] = one.day.gpd.ES / ei[i]

one.day.observed.ES     =

mean(temp.data[which(temp.data > u)])

m32[i] = one.day.observed.ES / ei[i]

m32.sd[i]     =  sd(temp.data[which(temp.data > u)])

m31.sd[i]     =  m32.sd[i]

if ( sum(1*(x[(final.index + 1):(final.index + end.cut.off)]

> u))== 0)

{ break }

m.obj.plus = obtain.m.plus.measures.2(x[(final.index + 1):

(final.index + end.cut.off)], r = r.par[i], u = u)

diff.m1[i]      =  (m.obj.plus$m1.plus - m1[i])   /

( m1.sd[i]  / sqrt(no.of.clusters)  )

diff.m2[i]      =  (m.obj.plus$m2.plus - m2[i])   /

( m2.sd[i]  / sqrt(no.of.clusters)  )

diff.m3[i]      =  (m.obj.plus$m3.plus - m3[i])   /

( m3.sd[i]  / sqrt(no.of.clusters)  )

diff.m31[i]     =  ei[i] * (m.obj.plus$m3.plus - m31[i])

  /  ( m31.sd[i] / sqrt(length(z[z==T])) )
```

```
  diff.m32[i]     =  ei[i] * (m.obj.plus$m3.plus - m32[i])
    /  ( m32.sd[i] / sqrt(length(z[z==T])) )
  start.index        =  start.index + m.obj.plus$skip
  final.index        =  final.index + m.obj.plus$skip
  pval.m1[i]   =   sum(1*(m.obj.plus$m1.plus > m.obj$m1val))
   / no.of.clusters
  pval.m2[i]   =   sum(1*(m.obj.plus$m2.plus > m.obj$m2val))
   / no.of.clusters
  pval.m3[i]   =   sum(1*(m.obj.plus$m3.plus > m.obj$m3val))
   / no.of.clusters
  pval.m31[i]  =   sum(1*(m.obj.plus$m3.plus >
  (temp.data[which(temp.data > u)] / ei[i]) )) /
  length(temp.data[which(temp.data > u)])
  pval.m32[i]  =   sum(1*(m.obj.plus$m3.plus >
  (temp.data[which(temp.data > u)] / ei[i]) )) /
  length(temp.data[which(temp.data > u)])
  m1.plus[i]      =  m.obj.plus$m1.plus
  m2.plus[i]      =  m.obj.plus$m2.plus
  m3.plus[i]      =  m.obj.plus$m3.plus
  m4.plus[i]      =  m.obj.plus$m3.plus
  raw.diff.m1[i]  =  (m.obj.plus$m1.plus - m1[i])
  raw.diff.m2[i]  =  (m.obj.plus$m2.plus - m2[i])
  raw.diff.m3[i]  =  (m.obj.plus$m3.plus - m3[i])
  raw.diff.m4[i]  =  (m.obj.plus$m3.plus - m32[i])
}
```

```
( This part of the code for the main function is omitted.)
}
```

The following code is the main function for bootstrapping the mean.

```
perform.bootstrap = function(y, n=1000)
{
  y.mean  = mean(y)
  y.tilde = y - y.mean
  t.stat  = numeric(n)
  for (i in 1:n)
  {
    temp      = sample(y.tilde, replace=T)
    t.stat[i] = (mean(temp))/(sd(temp)/sqrt(length(temp)))
  }
  y.obs = mean(y)/(sd(y)/sqrt(length(temp)))
  p.value = length(t.stat[t.stat > y.obs])/n
  out = list(y.obs = y.obs, t.stat=t.stat, p.value = p.value)
  out
}
```

Bibliography

Adler, R., Feldman, R., and Taqqu, M. S., editors (1998). *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*. Birkhäuser, Boston.

Ancona-Navarrete, M. A. and Tawn, J. (2000). A comparison of methods for estimating the extremal index. *Extremes*, **3**(1), 5–38.

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, **9**, 203–228.

Athreya, K. B. (1987). Bootstrap of the mean in the infinite variance case. *Ann. Statist.*, **15**(2), 724–731.

Bauwens, L. and Giot, P. (2000). The logarithmic acd model: an application to the bid-ask quote process of three nyse stocks. *Annales dEconomie et de Statistique*, **60**, 117–149.

Bauwens, L. and Hautsch, N. (2006). Modelling financial high frequency data using point processes. *Available at SSRN: http://ssrn.com/abstract=945107*.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. (2004). *Statistics of Extremes*. Wiley.

Bluhm, C., Overbeck, L., and Wagner, C. (2002). *An Introduction to Credit Risk Modeling*. Chapman and Hall, New York.

Brandt, A. (1986). The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients. *Advances in Applied Probability*, **18**(1), 211–220.

Brémaud, P. (1999). *Markov Chains: Gibbs fields, Monte Carlo simulation, and queues*. Spinger.

Chaves-Demoulin, V., Davison, A. C., and McNeil, A. (2005). Estimating value-at-risk: a point process approach. *Quantitative Finance*, **5**(2), 227–234.

Christoffersen, P. and Gonçalves, S. (2005). Estimation risk in financial risk management. *The Journal of Risk*, **7**(3).

Christoffersen, P. F. (2003). *Elements of Financial Risk Management*. Academic Press.

Coles, S. G. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag.

Crouhy, M., Galai, D., and Mark, R. (2006). *The Essentials of Risk Management*. McGraw-Hill.

Daley, D. J. and Vere-Jones, D. (2003). *An introduction to the theory of point processes. Vol. I*. Probability and its Applications (New York). Springer-Verlag, New York, second edition. Elementary theory and methods.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. Springer.

Davison, A. C. (2003). *Statistical models*, volume 11 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

Degen, M., Embrechts, P., and Lambrigger, D. (2007). The quantitative modeling of operational risk: Between g-and-h and evt. *Astin Bulletin*, **37**(2), 265–291.

Dowd, K. (2005). *Measuring Market Risk*. Wiley.

Dowd, K. and Blake, D. (2006). After var: The theory, estimation, and insurance applications of quanitle-based risk measures. *The Journal of Risk and Insurance*, **73**(2), 193–229.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Embrechts, P., Kluppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events*. Springer-Verlag.

Engle, R. F. and Russell, J. R. (1998). Autoregressive conditional duration: a new model for irregulary spaced transaction data. *Econometrica*, **66**, 1127–1162.

Ferro, C. A. T. and Segers, J. (2003). Inference for clusters of extreme values. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **65**(2), 545–556.

Finkenstädt, B. and Rootzén, H., editors (2004). *Extreme Values in Finance, Telecommunications, and the Environment, Volume 99 of Monographs on Statistics and Applied Probability*. Chapman and Hall / CRC.

Focardi, S. and Fabozzi, F. (2005). An autoregressive conditional duration model of credit-risk contagion. *Journal of Risk Finance*, **6**, 208–225.

Galbraith, J. W. and Zernov, S. (2006). Extreme dependence in the nasdaq and s&p 500 composite indexes. *Preprint. http://www.mcgill.ca/files/economics/extremedependencein.pdf*.

Geman, H. (2005). *Commodities and Commodity Derivatives : Modelling and Pricing for Agriculturals, Metals and Energy*. Wiley.

Gradshteyn, I. S. and Ryzhik, I. M. (1965). *Table of integrals, series, and products*. Fourth edition prepared by Ju. V. Geronimus and M. Ju. Ceĭtlin. Translated from the Russian by Scripta Technica, Inc. Translation edited by Alan Jeffrey. Academic Press, New York.

Hawkes, A. G. (1971a). Point spectra of some mutually exciting point processes. *J. Roy. Statist. Soc. Ser. B*, **33**, 438–443.

Hawkes, A. G. (1971b). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, **58**, 83–90.

Hawkes, A. G. (1972). Spectra of some mutually exciting point processes with associated variables. In *Stochastic point processes: statistical analysis, theory, and applications (Conf., IBM Res. Center, Yorktown Heights, N.Y., 1971)*, pages 261–271. Wiley-Interscience, New York.

Hawkes, A. G. and Oakes, D. (1974). A cluster process representation of a self-exciting process. *J. Appl. Probability*, **11**, 493–503.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.*, **3**(5), 1163–1174.

Hsing, T. (1993). Extremal index estimation for a weakly dependent stationary sequence. *The Annals of Statistics*, **21**, 2043–2071.

Hsing, T., Husler, J., and Leadbetter, M. (1988). On the exceedance point process for a stationary sequence. *Probability Theory and Related Fields*, **78**, 97–112.

Hull, J. C. (2007). *Risk Management and Financial Institutions*. Pearson Prentice Hall.

Inui, K. and Kijima, M. (2005). On the significane of expected shortfall as a coherent risk measure. *Journal of Banking and Finance*, **29**, 853–864.

Jorion, P. (2007a). Risk management for hedge funds with position information: both managers and investors could be satisfied. *Journal of Portfolio Management*, **34**(1), 127–136.

Jorion, P. (2007b). *Value at Risk: The New Benchmark for Managing Financial Risk*. McGraw-Hill Companies, 3 edition.

Karr, A. F. (1991). *Point Processes and Their Statistical Inference*. Marcel Dekker, 2 edition.

Knight, K. (1989). On the bootstrap of the sample mean in the infinite variance case. *Ann. Statist.*, **17**(3), 1168–1175.

Laurini, F. (2004). Clusters of extreme observations and extremal index estimate in garch processes. *Studies in Nonlinear Dynamics and Econometrics*, **8**(4).

Laurini, F. and Tawn, J. (2003). New estimators for the extremal index and other cluster characteristics. *Exrtemes*, **6**, 189–211.

Leadbetter, M. (1983). Extremes and local dependence in stationary sequence. *Z. Wahrsch. Verw. Gebiete*, **65**, 291–306.

Leadbetter, M. R. and Rootzén, H. (1988). Extremal theory for stochastic processes. *Ann. Probab.*, **16**(2), 431–478.

Leadbetter, M. R., Lindgren, G., and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer-Verlag.

Longin, F. (2000). From value at risk to stress testing: the extreme value approach. *Journal of Banking and Finance*, **24**, 1097–1130.

Mandelbrot, B. and Hudson, R. L. (2004). *The Misbehavior of Markets*. Basic Books.

Marthinsen, J. E. (2008). *Risk Takers, Uses and Abuses of Financial Derivatives*. Pearson Addison Wesley, 2 edition.

Mason, D. M. and Shao, Q.-M. (2001). Bootstrapping the Student t-statistic. *Ann. Probab.*, **29**(4), 1435–1450.

McNeil, A. and Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach. *Journal of Empirical Finance*, **7**, 271–300.

McNeil, A. J., Frey, R., and Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques, and Tools*. Pearson Addison Wesley.

Novak, S. Y. and Beirlant, J. (2006). The magnitude of a market crash can be predicted. *Journal of Banking and Finance*, **30**, 453–462.

O'Brien, G. (1987). Extreme values for stationary and markov sequences. *The Annals of Probability*, **15**, 281–291.

Pacurar, M. (2006). Autoregressive conditional duration (acd) models in finance: A survey of the theoretical and empirical literature. *Available at SSRN: http://ssrn.com/abstract=933120*.

Panjer, H. H. (2006). *Operational Risk : Modeling Analytics*. Wiley-Interscience.

Pickands, J. (1975). Statistical inference using the extreme order statistics. *The Annals of Statistics*, **3**, 119–131.

Resnick, S. (2006). *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer, 1 edition.

Resnick, S. I. (1987). *Extreme Values, Regular Variation and Point Processes*. Springer-Verlag.

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, **72**, 67–90.

Smith, R. L. and Weissman, I. (1994). Estimating the extremal index. *Journal of the Royal Statistical Society, Series B*, **56**, 515–528.

Smith, R. L., Tawn, J. A., and Coles, S. G. (1997). Markov chain models for threshold exceedances. *Biometrika*, **84**, 249–268.

Snyder, D. L. and Miller, M. I. (1991). *Random Point Processes in Time and Space*. Springer, 2 edition.

Stoev, S. A., Michailidis, G., and Taqqu, M. S. (2006). Estimating heavy-tailed exponents through max self-similarity. Technical Report 445.

Tsay, R. (2005). *Analysis of Financial Time Series*. John Wiley & Sons, Hoboken, New Jersey, U.S.A., 2 edition.

van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

Weissman, I. and Novak, S. Y. (1998). On blocks and runs estimators of the extremal index. *J. Statist. Plann. Inference*, **66**(2), 281–288.