2013-07-10

Analysis of Temporally Dependent Extremes for the Gumbel Distribution

Ji, Chaoqun

UNIVERSITY OF CALGARY

Analysis of Temporally Dependent Extremes for the Gumbel Distribution

by

Chaoqun Ji

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF MATHEMATICS AND STATISTICS

CALGARY, ALBERTA

June, 2013

Abstract

For modeling extremal behaviors, the Generalized extreme value distribution that originated from the well established Extreme Value Theory has been widely used. As a special case of such Generalized extreme value distribution, the Gumbel family is suitable for modeling maximum values from light-tailed distributions. A common assumption used in the central models of extreme values is the independence of extremes in most previous studies. However, short-term dependence among extremes might exist. In this thesis, we study a linear Gumbel distributed autoregressive model which was introduced by Toulemonde *et al*. (2010) to simulate dependent extremes that follow the Gumbel distribution. Our main goal is to investigate that if Gumbel distributed short-term maxima are weakly/moderately/strongly dependent, but this dependence is not recognized, what will happen to the resulting estimates of the Gumbel parameters. To reach this goal, simulations and a numerical example in environmental science are presented to quantify the above issue.

Acknowledgements

I would like to express my deepest appreciation to all those who provided me the possibility to complete my program and this thesis.

A special gratitude I give to my supervisor Dr. Gemai Chen for the useful commons, financial support and engagement through the learning process of this master thesis. Without his guidance and persistent help, this dissertation would not have been possible. A lot of thanks go to my co-supervisor, Dr. Hyang Mi Kim. I am extremely grateful and indebted to her for her expert, sincere and valuable guidance and encouragement extended to me.

I would like to thank my committee members Dr. Xuewen Lu and Dr. Jane Kang for taking some valuable time out of their schedules to read my thesis and for giving me some critical suggestions in my thesis research.

I would like to thank the financial support that I have received from the Department of Mathematics and Statistics at the University of Calgary during my graduate studies, which ensured my life and research in Calgary very enjoyable.

In additional, I would like to thank Director of Graduate Studies Dr. Renate Scheidler and Graduate Program Administrator Yanmei Fei from the Department of Mathematics and Statistics for their hardworking and help to schedule and reschedule my Master's oral exam.

Moreover, special thanks for my parents for their selfless love and support, without which I wouldn't have achieved what I have now.

Lastly, I want to thank all of those who supported me with my study.

Table of Contents

List of Tables

List of Figures and Illustrations

Chapter 1

Introduction

What on earth can we expect next? Can we do anything to protect ourselves from the extreme events that lie years, decades, or even centuries in the future? Known as Extreme Value Theory (EVT), it can give insights into extreme events. EVT is a curious and fascinating blend of a variety of theories and applications involving natural phenomena, financial market, insurance industry, economics and other disciplines. In particular, in the financial world, EVT is being used to assess the risk of natural disasters, and to ensure institutions have the financial reserves needed to cover the likely impact. The technique is being used to protect ship from the most extreme storms they may face. Also it has helped architects design a sea wall to sustain the shock of waves. Many real life extreme events require EVT to provide a firm theoretical foundation on which we can build statistical models to describe those extreme events.

The musings of EVT have been traced back to the early eighteenth century. Yet it was not until the 1920s that the idea of predicting the unexpected events first attracted serious attention. In 1928, Cambridge mathematician R.A. Fisher and his colleague L.H.C. Tippett launched what became known as EVT with a paper showing that extreme events do indeed follow their own special types of distribution. However, this obvious practical value was regarded with suspicion for many years until it was used to predict flood levels from past records with great success in the 1940s. From then on, EVT was gradually improved and the Generalized Extreme Value (GEV) distribution is used as an approximation to model the extreme values. Although it is widely accepted and has been spread into ever more fields now, implementation of EVT still faces many challenges. In this thesis, we will concentrate on the model called Gumbel distribution which is a special case of the GEV distribution.

1.1 Introduction to the GEV distribution

In probability theory and statistics, the GEV distributions arise as the limiting distributions for maximum or minimums (extreme values) of a sample of independent and identically distributed (iid) random variables, as the sample size increases. The GEV distribution are widely used in insurance, economics, risk management, environmental science, telecommunications, and many other industries to deal with extreme events which occur with very small probability. The class of GEV distributions essentially involves three families of extreme value distributions, namely types I, II and III or **Gumbel**, **Weibull**, and **Fréchet** families.

In some fields of applications, the GEV distribution is also known as the Fisher-Tippett extreme value distribution which was named from Sir Ronald Aylmer Fisher (1890-1962) and Leonard Henry Caleb Tippett (1902-1985) who recognized three principle functions outlined later. As the most common type, the Gumbel distribution is a particular case of the GEV distribution and it was named after a German mathematician Emil Gumbel (1891-1966) who pioneered the mathematical field of extreme value theory along with Leonard Tippet and Ronald Fisher.

The Gumbel distribution has two forms. One is based on the smallest extremes and the other is based on the largest extremes. We call these the minimum and maximum cases respectively, although this thesis will only focus on the maximum case.

The general formula for the probability density function of the Gumbel (minimum) distribution is

$$f(x) = \frac{1}{\sigma} e^{\frac{x-\mu}{\sigma}} e^{-e^{\frac{x-\mu}{\sigma}}}, x \in (-\infty, +\infty),$$

where μ is the location parameter and σ is the scale parameter. The case where $\mu = 0$ and $\sigma = 1$ is called the Standard Gumbel distribution. The density for the standard Gumbel distribution (minimum) reduces to

$$f(x) = e^{x} e^{-e^{x}}, x \in (-\infty, +\infty).$$

Figure 1.1 is the plot of the Gumbel probability density function (PDF) for the minimum case.

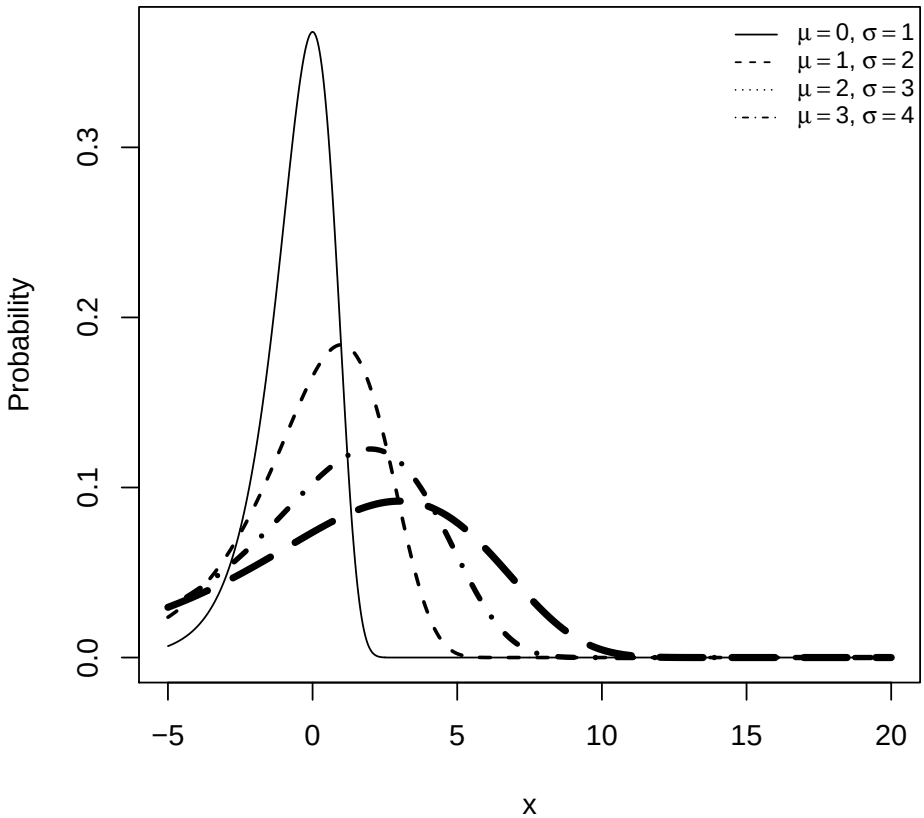

Figure 1.1: Gumbel distributionm PDF (Minimum).

The general formula for the PDF of the Gumbel (maximum) distribution is

$$f(x) = \frac{1}{\sigma} e^{-\frac{x-\mu}{\sigma}} e^{-e^{-\frac{x-\mu}{\sigma}}}, x \in (-\infty, +\infty),$$

where μ is the location parameter and σ is the scale parameter. The case where $\mu = 0$ and $\sigma = 1$ is called the Standard Gumbel distribution. The density for the standard Gumbel distribution (maximum) reduces to

$$f(x) = e^{-x} e^{-e^{-x}}, x \in (-\infty, +\infty).$$

The following Figure 1.2 is the plot of the Gumbel probability density function for the maximum case.

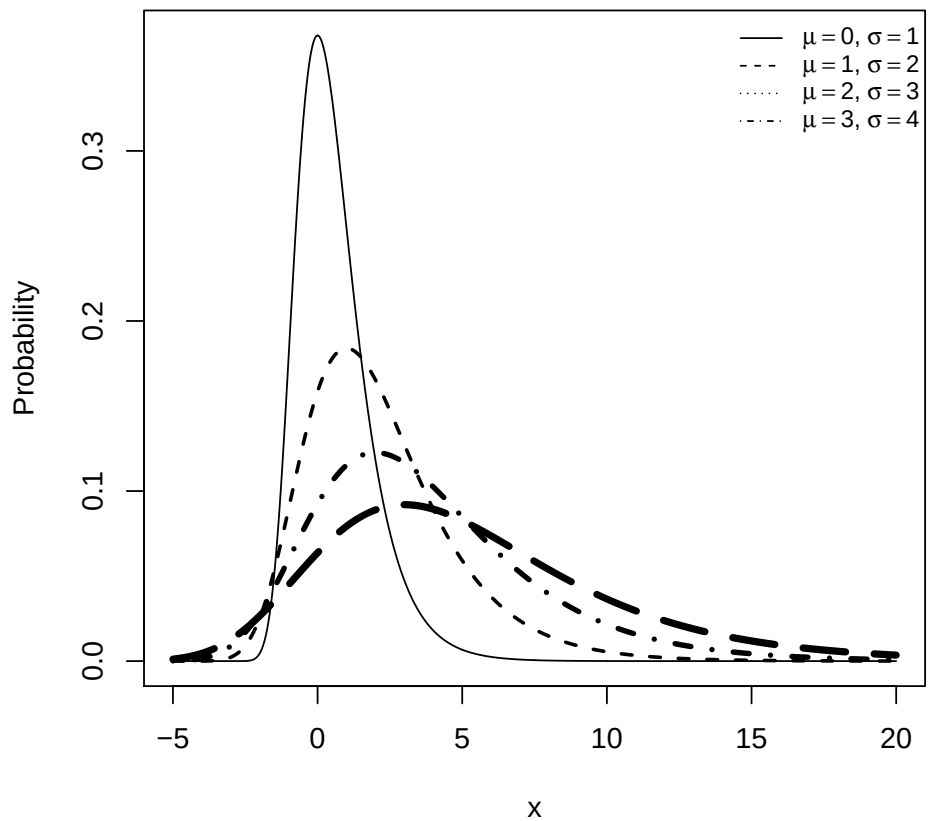

Figure 1.2: Gumbel distributionm PDF (Maximum).

The formula for the cumulative density function (CDF) of the standard Gumbel distribution (minimum) is

$$F(x) = 1 - e^{-e^{\frac{x-\mu}{\sigma}}}, x \in (-\infty, +\infty),$$

and the CDF for the standard maximum case is

$$F(x) = e^{-e^{-\frac{x-\mu}{\sigma}}}, x \in (-\infty, +\infty).$$

Figure 1.3 and Figure 1.4 show the plots of the Gumbel cumulative distribution functions for the minimum and maximum cases.

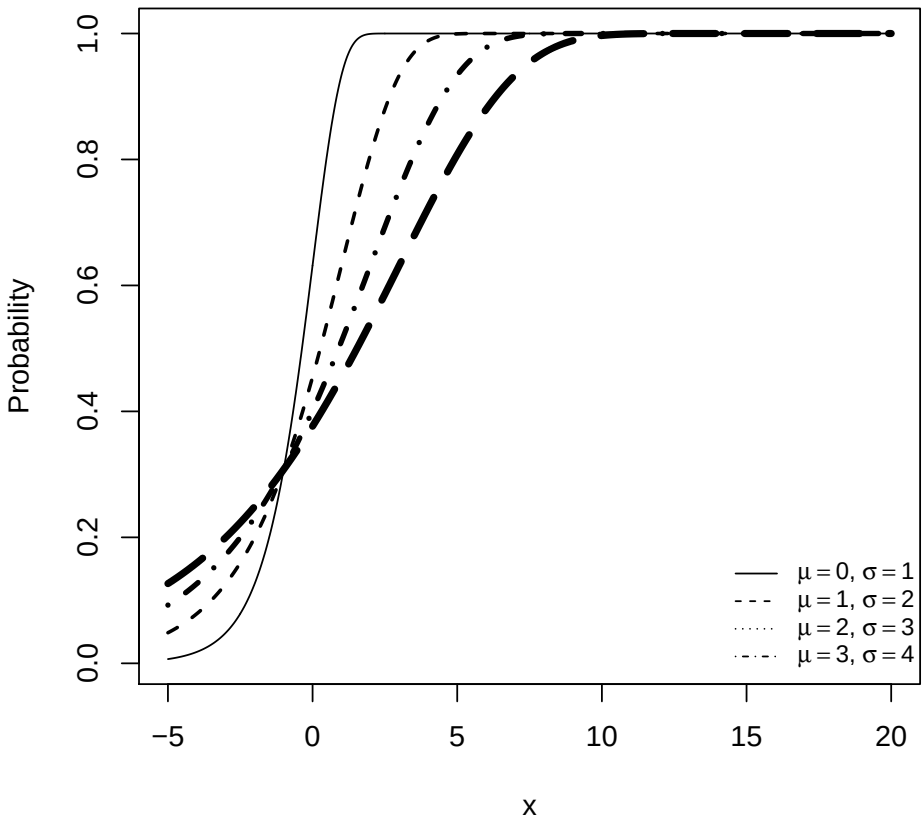

Figure 1.3: Gumbel distributionm CDF (Minimum).

1.2 Review of Previous Studies

Classical EVT dictates that correctly normalized maxima should follow (under various conditions) a GEV distribution. The key characteristic of the GEV is its stability for the max operator. Also, the GEV distribution is the limit distribution of normalized maxima of a sequence of independent and identically distributed random variables. When the target ran-

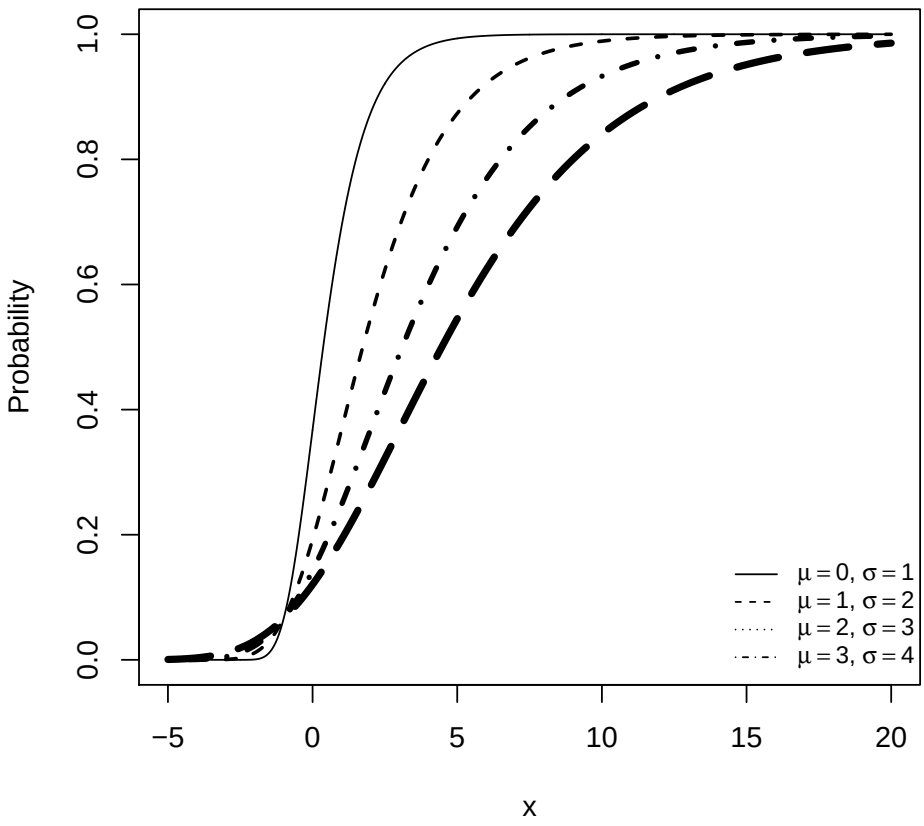

Figure 1.4: Gumbel distributionm CDF (Maximum).

dom variables are iid, a peaks over threshold analysis can be carried out, where only cluster maxima are used in the generalized Pareto distribution (GPD) fit and confidence interval evaluation (Davison and Smith, 1990). This approach can largely avoid dependence in exceedances. However, short term dependencies (e.g. day-to-day records in weather systems) might exist. When such dependence is present, one approach is to fit the GPD using all exceedances and then model the dependence between exceedances in a cluster (Smith *et al.*, 1997). Alternatively, one can fit the GPD using all exceedances and (falsely) assuming that they are independent, then account for the dependence in the confidence interval evaluation

by using appropriate block bootstrap methods (Davison and Hinkley, 1997). Davison and Ramesh (2000) admitted the difficulty to handle the failure of assuming the independence of extremes. Chavez-Demoulin and Davison (2005) argued that dependence between the extremes would not bias estimators of the parameters of a model built upon the independence assumption. However, no direct investigation on the effect of the failure of the independence assumption has ever been done, possibly because there was no easy way to simulate dependent extremes that follow GEV distribution or GPD. This situation changed due to the recent work such as Toulemonde *et al.* (2010), by which we can generate temporally dependent values that are Gumbel distributed.

1.3 Thesis Plan

As mentioned before, there is no direct investigation to show that failure of the independence assumption of extreme values may or may not affect the accuracy of the final modeling results using the Gumbel distribution. In particular, if Gumbel distributed maxima are weakly/moderately/strongly dependent, but this dependence is not recognized, what will happen to the resulting estimates of the Gumbel parameters? In this thesis we will answer this question. Our goal is to use the autoregressive (AR) processes (Toulemonde *et al*, 2010) to generate temporally dependent values under Gumbel distribution and compare the estimation results with previous works assuming independence of the extremes.

The thesis is organized as follows. In Chapter 2, we will introduce and discuss the maximum likelihood estimators of the parameters of the Gumbel distribution and then briefly explain how to extend linear AR models in such a way that they handle Gumbel maxima distribution. In Chapter 3, we will run simulations to study the bias and mean square error (MSE) of the maximum likelihood estimators of the Gumbel parameters by comparing the cases under the independence assumption with the dependence cases. Also the return levels of the two cases are considered. In Chapter 4, we will present the analysis of a real dataset

in environmental science using the methodology we introduced in Chapter 2. In Chapter 5, we will draw some conclusions and discuss some future study possibilities. The computer codes of our program are given in the Appendix.

Chapter 2

Methodology

The GEV distribution is widely used to model daily, weekly or yearly extreme values in many disciplines. One special case of such GEV distributions is the Gumbel family that corresponds to the modeling of maxima stemming from light-tailed distributions. In this chapter, we are going to focus on introducing the methodology of parameter estimation for the Gumbel distribution and the basic knowledge about the AR(1) Gumbel model. In the first section, asymptotic models for extreme value distribution are briefly introduced and basic concepts and properties of GEV distributions are provided. In the second section, we are going to discuss an efficient estimation algorithm based on the maximum likelihood method and give some simple model checking techniques. In the third section, we are going to introduce a way of extending linear autoregressive models to handle Gumbel distributed maxima in order to capture temporal dependencies and give an approach of the AR(1) model checking.

2.1 Asymptotic Models

2.1.1 Model Formulation

In this section the foundation of extreme value theory is used to build the model. The model focuses on the statistical behavior of the maximum, denoted by M_n, of a sequence of independent random variables $X_1, ..., X_n$,

$$M_n = max\left\{X_1, ..., X_n\right\},$$

where $X_1, ..., X_n$ have a common distribution function F. In applications, the X_i usually represent values of a process measured on a regular time-scale such as hourly measurements

of wind speed, daily average temperature or weekly rainfall depth and then M_n represents the maximum of the process over n time units of observation. If n is the number of observations in a year, then M_n corresponds to the annual maximum.

We use $F(x)$ to represent the distribution of $X_1, ..., X_n$, then the distribution of M_n can be derived exactly for all values of n by

$$\begin{aligned} P(M_n \leq x) &= P(X_1 \leq x, ..., X_n \leq x) \\ &= P(X_1 \leq x) \times ... \times P(X_n \leq x) \\ &= [F(x)]^n . \end{aligned} \tag{2.1}$$

However, the distribution function F is usually unknown in practice. One possible method is to use standard techniques to estimate F from observed data, and then to substitute this estimate into equation 2.1. Unfortunately, some very small discrepancies in the estimate of F can lead to substantial discrepancies for F^n in applications.

An alternative method is to accept that F is unknown and to look for approximate models for F^n, which can be estimated on the analysis of extreme data instead of the sequence of independent random variables. This idea is similar to the usual practice of approximating the distribution of sample means by the normal distribution, as justified by the central limit theorem. By this idea, we need to look at the behavior of F^n as $n \to \infty$. However, for any $x < x_+$, where x_+ is the upper end-point of F, $F^n(x) \to 0$ as $n \to \infty$, so that the distribution of M_n degenerates to a point mass on x_+. This difficulty is avoided by allowing a linear renormalization of the variable M_n :

$$M_n^* = \frac{M_n - \mu_n}{\sigma_n},$$

for sequences of constants $\{\mu_n\}$ and $\{\sigma_n > 0\}$. Appropriate choices of the $\{\mu_n\}$ and $\{\sigma_n\}$ stabilize the location and scale of M_n^* as n increases, avoiding the difficulties that arise with the variable M_n. Therefore, we look for the limit distributions for M_n^* instead of M_n with proper choices of $\{\mu_n\}$ and $\{\sigma_n\}$.

2.1.2 Extremal Types Theorem

The entire range of possible limit distributions for M_n^* is given by Theorem 1 (Coles, S. G., 2001), the extremal types theorem.

Theorem 1 *If there exist sequences of constants* $\{\mu_n\}$ *and* $\{\sigma_n > 0\}$ *such that*

$$Pr\left\{\left(M_n - \mu_n\right)/\sigma_n \leq x\right\} \rightarrow F(x),\ as\ n \rightarrow \infty,$$

where F *is a non-degenerate distribution function, then* F *belongs to one of the following families:*

$$\begin{aligned}
I : Fx &= exp\left\{-exp\left[\left(\frac{x-\mu}{\sigma}\right)\right]\right\}, -\infty < x < \infty; \\
II : Fx &= exp\left\{-\left(\frac{x-\mu^{\epsilon}}{\sigma}\right)^{-\epsilon}\right\},\ x > \mu; \\
III : Fx &= exp\left\{-\left[-\left(\frac{x-\mu^{\epsilon}}{\sigma}\right)^{\epsilon}\right]\right\},\ x < \mu,
\end{aligned}$$

for parameters $\mu, \sigma > 0$*, and in the cases of families II and III,* $\epsilon > 0$*.*

Theorem 1 states that the rescaled sample maxima $(M_n - \mu_n)/\sigma_n$ converges in distribution to a variable having a distribution within one of the three families labeled I, II and III. Collectively, these three families of distributions are termed the extreme value distributions, known as the **Gumbel**, **Fréchet**, and **Weibull** families, respectively. Each family has a location and a scale parameter, μ and σ respectively. In addition, the Fréchet and Weibull families have a shape parameter ϵ.

2.1.3 The Generalized Extreme Value Distribution

The three types of limits that arise in Theorem 1 have distinct forms of function. The distribution function of the three-parameter GEV distribution with parameters $\mu, \sigma > 0$,

and ϵ is

$$F(x) = exp\left\{-\left[1-\epsilon\left(\frac{x-\mu}{\sigma}\right)\right]^{1/\epsilon}\right\}, \tag{2.2}$$

where $1-\epsilon(x-\mu)/\sigma > 0$. The model has three parameters: a location parameter, μ; a scale parameter, σ; a shape parameter, ϵ. The Fréchet and Weibull families of extreme value distribution correspond respectively to the case $\epsilon > 0$ and $\epsilon < 0$ in the parameterization. The subset of the GEV family with $\epsilon = 0$ is interpreted as the limit of equation 2.2 as $\epsilon \to 0$, leading to the **Gumbel** family with the distribution function

$$F(x) = exp\left\{-exp\left[-\left(\frac{x-\mu}{\sigma}\right)\right]\right\}, -\infty < x < +\infty. \tag{2.3}$$

Although these three families have different requirement of ϵ, the data themselves determine the most appropriate type of tail behavior, and there is no need to make subjective a priori judgements about which individual extreme family to adopt.

For modeling the extremes of a series of independent observations $x_1, x_2, ...,$ data are blocked into sequences of observations of length n, for some large value of n, generating a series of block maxima, $M_{n,1}, ...M_{n,m}$, say, to which the GEV distribution can be fitted. Often the blocks are chosen to correspond to a time period of length one year, such as n is the number of observations in a year and the block maxima is the annual maxima. We will use the annual maxima in the following discussion.

By inverting equation 2.2, estimates of extreme quantiles of the annual maximum distribution are obtained:

$$x_p = \mu + \sigma\left\{1-\left[-log\left(1-p\right)\right]^{\epsilon}\right\}/\epsilon, \tag{2.4}$$

where $0 < p < 1$, $F(x_p) = 1-p$ and $\epsilon \neq 0$.

Similarly, for a Gumbel family with $\epsilon = 0$,

$$x_p = \mu + \sigma\left\{-log\left(1-p\right)\right\}. \tag{2.5}$$

By definition, x_p is the return level corresponding to return period $1/p$. The level x_p is expected to be exceeded on average once every $1/p$ years. More specifically, the annual maximum in any particular year is supposed to exceed x_p with probability p.

2.1.4 Asymptotic Models for Minima

Some applications require models for extremely small, rather than extremely large observations. For example, the lifetime of a system needs to be tested when the overall system breaks down if any of the individual components fails. In this case, the target is the working time of the weakest components of the system. To model the minima cases, we denote $\widetilde{M_n} = min\{X_1, ...X_n\}$, where the X_i is independent and identically distributed. Also analogous arguments which were applied to M_n still apply to $\widetilde{M_n}$, leading to to a similar limiting distribution of $\widetilde{M_n}$.

Letting $Y_i = -X_i$ for $i = 1, ..., n$, the change of sign means that small values of X_i correspond to large values of Y_i. So if $\widetilde{M_n} = min\{X_1, ...X_n\}$ and $M_n = max\{Y_1, ..., Y_n\}$, then $\widetilde{M_n} = -M_n$. Hence, for large n,

$$\begin{aligned}
P(\widetilde{M_n} \leq x) &= P(-M_n \leq x) \\
&= P(M_n \geq -x) \\
&= 1 - P(M_n \leq -x) \\
&\approx 1 - exp\left\{-\left[1 - \epsilon\left(\frac{-x-\mu}{\sigma}\right)\right]^{1/\epsilon}\right\} \\
&= 1 - exp\left\{-\left[1 - \epsilon\left(\frac{x-\widetilde{\mu}}{\sigma}\right)\right]^{1/\epsilon}\right\},
\end{aligned}$$

on $\{x : 1 - \epsilon(\widetilde{\mu} - x)/\sigma > 0\}$, where $\widetilde{\mu} = -\mu$. This distribution is the **GEV distribution for minima**, which was introduced by Coles, S. G. (2001).

2.2 Inference for the GEV Distribution

2.2.1 The Choice of Block Size

By Theorem 1, the GEV provides a model for the distribution of block maxima. Usually the preparatory work consists of blocking the data into blocks of equal length and fitting the GEV to the set of block maxima. For some particular dataset, the choice of block size can be critical. If the blocks are too small, the approximation by the limit model in Theorem 1 is likely to be poor, creating bias in estimation and extrapolation. Otherwise, large blocks generate few block maxima, leading to large estimation variance. In applications, the block size of one year is often to be adopted because only the annual maximum data may have been recorded in some situations such that the annual rainfall depth or annual highest sea-level. Many datasets contain seasonal factor which might lead to different distributions for block maxima. For example, daily temperatures are likely to vary according to season, violating the assumption that the x_i have a common distribution. If the data were blocked into block lengths of around 3 months, the maximum of the summer block is likely to be much greater than that of the winter block and then it would be likely to give inaccurate results.

2.2.2 Maximum Likelihood Estimation

If the data set $\{X_i\}$ are independent and identically distributed from a GEV distribution, then the minus of the log-likelihood function for a sample of n observations $\{x_1, ...x_n\}$ is

$$-L(\mu, \sigma, \epsilon) = nlog\sigma + (1-\epsilon)\sum_i^n y_i + \sum_i^n e^{-y_i}, \tag{2.7}$$

where $y_i = -\epsilon^{-1}log\left\{1 - \epsilon\left(x_i - \mu\right)/\sigma\right\}$ and provided that

$$1 - \epsilon(x_i - \mu)/\sigma > 0,\ i = 1, ..., n. \tag{2.8}$$

At parameter combinations for which equation 2.8 is violated, corresponding to a configuration for which at least one of the observed data point falls beyond an end-point of the distribution, the likelihood is zero and the log-likelihood equals $-\infty$.

The case $\epsilon = 0$ requires separate treatment using the Gumbel limit of the GEV distribution. This leads to the log-likelihood

$$-L(\mu, \sigma) = nlog\sigma + \sum_{i}^{n} z_i + \sum_{i}^{n} e^{-z_i}, \tag{2.9}$$

where $z_i = (x_i - \mu)/\sigma$.

Maximization of the log-likelihood function is equivalent to the minimization of the pair of equations 2.7 and 2.9 with respect to the parameter vector (μ, σ, ϵ), which gives to the maximum likelihood estimates with respect to the entire GEV family. There is no analytical solution, but for any given dataset the minimization is straightforward using standard numerical optimization algorithms. The maximum likelihood estimators of μ and σ are denoted by $\hat{\mu}$ and $\hat{\sigma}$.

By substitution of the maximum likelihood estimates of the GEV parameters into equation 2.4, the maximum likelihood estimate of x_p for $0 < p < 1$, the $1/p$ return level, is obtained as

$$\hat{x_p} = \hat{\mu} + \hat{\sigma} \left\{ 1 - \left[-log\left(1 - p\right) \right]^{\hat{\epsilon}} \right\} / \hat{\epsilon}, \tag{2.10}$$

where $\epsilon \neq 0$ and $F(x_p) = 1 - p$. Similarly, for the Gumbel family ($\epsilon = 0$),

$$\hat{x_p} = \hat{\mu} + \hat{\sigma} \left\{ -log\left(1 - p\right) \right\}. \tag{2.11}$$

2.2.3 Model Checking

Though it is impossible to check the validity of an extrapolation based on a GEV model, assessment can be made with reference to the observed data. This is not sufficient to justify extrapolation, but is a reasonable prerequisite. We will discuss the use of probability plots and quantile plots for model checking.

A probability plot is a comparison of the empirical and fitted distribution functions. With ordered block maximum data $x_1 \leq x_2 \leq ... \leq x_n$, the empirical distribution function evaluated at x_i is given by

$$\widetilde{F}(x_i) = i/(n+1).$$

By substitution of parameter estimates into equation 2.2, the corresponding model based estimates are

$$\widehat{F}(x_i) = exp\left\{-\left[1-\hat{\epsilon}\left(\frac{x_i-\hat{\mu}}{\hat{\sigma}}\right)\right]^{1/\hat{\epsilon}}\right\}.$$

Also for the Gumbel model with $\epsilon = 0$,

$$\widehat{F}(x_i) = exp\left\{-exp\left[-\left(\frac{x_i-\hat{\mu}}{\hat{\sigma}}\right)\right]\right\}.$$

If the distribution is working well,

$$\widehat{F}(x_i) \approx \widetilde{F}(x_i),$$

for each i, so a probability plot, consisting of the points

$$(\widehat{F}(x_i), \widetilde{F}(x_i)),\ i = 1, ...n,$$

should lie close to the unit diagonal. Any substantial departures from linearity are indicative of some failing in the GEV distribution.

A weakness of the probability plot for the GEV distribution is that both $\widehat{F}(x_i)$ and $\widetilde{F}(x_i)$ are bound to approach 1 as x_i increases, while it is usually the accuracy of the model for large values of x that is of the greatest concern. That is, the probability plot provides the least information in the region of most interest. This deficiency is avoided by the quantile plot, consisting of the points

$$\left(\widehat{F}^{-1}(\frac{i}{n+1}), x_i\right),\ i = 1, ..., n,$$

where from equation 2.4,

$$\widehat{F}^{-1}(\frac{i}{n+1}) = \hat{\mu} + \hat{\sigma}\left\{1-[-log\,(1-p)]^{\hat{\epsilon}}\right\}/\hat{\epsilon}.$$

Also for the Gumbel model with $\epsilon = 0$

$$\widehat{F}^{-1}(\frac{i}{n+1}) = \hat{\mu} + \hat{\sigma}\left\{-log\left(1-p\right)\right\}.$$

Departures from linearity in the quantile plot also indicate model failure.

2.3 Gumbel AR Models

2.3.1 Literature Review

The Gumbel distribution is usually used to describe independent and identically distributed samples. This is a common assumption used in the central models of extreme values. However, temporal dependencies may exist in some cases. For example, the highest temperature of today may affect the highest temperature of tomorrow. As mentioned in chapter 1, some statisticians have attempted some approaches to avoid dependence in exceedances (Smith *et al.*, 1997) or argued that dependence between the extreme values would not bias estimators (Chavez-Demoulin and Davison, 2005).

The key characteristic of the GEV is its stability for the max operator. The maximum of two independent and identically distributed GEV distributed random variables is still GEV distributed. But adding two GEV random variables does not generate a GEV distributed random variable. This explains why linear autoregressive processes are not generally used to describe maxima behavior and other methods have to be developed to combine linear AR processes for light tails and maxima. It is well known that correctly normalized maxima from light-tailed distributions belong to the Gumbel domain. This means that maxima can be expected to be adequately fitted by the Gumbel distribution defined by

$$F_{\mu,\sigma}(x) = exp\left\{-exp\left(-\frac{x-\mu}{\sigma}\right)\right\}, -\infty < x < \infty,$$

where μ and σ are the so-called location and scale parameters.

Toulemonde *et al.* (2010) presented a method using linear autoregressive processes to generate temporally dependent values that are Gumbel distributed. This method changes

the investigation of temporal dependence of extreme values. They took advantage of the stability of Gumbel random variables when added to the logarithm of a positive α-stable random variable. This can propose a linear Gumbel distributed AR model whose main theoretical properties are derived.

In this section, we are going to introduce the method to generate temporally dependent extreme values that are Gumbel distributed.

2.3.2 Studies of Gumbel AR Models

The building block of our model is an additive relationship between Gumbel and positive α-stable variables. Recall that a random variable S is said to be stable if for all non-negative real numbers c_1 and c_2, there exist a positive real number a and a real number b such that $c_1S_1 + c_2S_2$ is equal in distribution to $aS + b$ where S_1, S_2 are iid copies of S.

If X is Gumbel distributed with parameters μ and σ and is independent of S which represents a positive α-stable variable with $\alpha \in (0, 1)$ defined by its Laplace transform

$$E(exp(-\mu S)) = exp(-\mu^{\alpha}), \ for\ all\ \mu \geq 0, \tag{2.12}$$

then the sum $X + \sigma logS$ is also Gumbel distributed with parameters μ and σ/α. Such an additive property has been recently studied by Fougèes *et al.* (2009) in a mixture context. In time series analysis, the additive stability between Gumbel and positive α-stable random variables allows us to propose a simple linear AR model that can be summarized by the following proposition (Toulemonde *et al.*, 2010).

Theorem 2 *Let S_i be iid positive α-stable variables defined by equation 2.12 for any $i \in Z$, $Z = \{0, \pm1, \pm2, ...\}$. Let $\{X_i, i \in Z\}$ be a stochastic process defined by the recursive relationship*

$$X_i = \alpha X_{i-1} + \alpha\sigma logS_i, \tag{2.13}$$

where $\sigma > 0$ and $\alpha \in (0, 1)$, then X_i follows a Gumbel distribution with parameters $(0, \sigma)$.

An advantage of the equation 2.13 is that X_i follows a Gumbel distribution whose parameters are independent of α. The covariance between X_i and X_{i-h} is increasing with α.

To simplify the statement of our discussion, the Gumbel location parameter μ is set equal to zero. In practice, μ can be different from zero. It suffices to add μ to X_i in equation 2.13 to have a Gumbel(μ, σ).

Concerning the estimation of α, a least-square estimator can be introduced by writing

$$argmin_r \left\{ \sum_{i=1}^{n-1} \left([X_{i+1} - E(X_0)] - r\,[X_i - E(X_0)] \right)^2 \right\} = \frac{\sum_{i=1}^{n-1} (X_i - E(X_0))\,(X_{i+1} - E(X_0))}{\sum_{i=1}^{n-1} (X_i - E(X_0))^2}.$$

This is similar to the classical Yule-Walker equation for AR (1) models. It follows that given a sample $X_1, ..., X_n$ that satisfies equation 2.13, our estimator of α is simply

$$\tilde{\alpha} = \frac{1}{s^2 n} \sum_{i=1}^{n-1} \left(X_i - \overline{X}\right) \left(X_{i+1} - \overline{X}\right),$$

with standard error

$$s.e.(\tilde{\alpha}) = \sqrt{(1 - \tilde{\alpha}^2)/n},$$

where $\overline{X} = \sum_{i=1}^{n-1} X_i/n$ and $s^2 = \sum_{i=1}^{n} \left(X_n - \overline{X}\right)^2 /n$.

2.3.3 AR(1) Model Checking

When we use the Gumbel AR(1) model, it is important to check whether it is good or not to fit the extreme values. Section 2.2.3 introduced model checking for the Gumbel model with iid assumption. Here is a method of checking whether the dataset fits the AR(1) model.

First, graph the sample autocorrelation function (ACF) and partial autocorrelation function (PACF) for values of the lag h. Second, values of the ACF should decay rapidly as h increases, which indicate short-term dependency in the time series. Third, as a rough guide, the sample PACF of AR(1) model should lie between the plotted bounds $\pm 1.96/\sqrt{n}$ for lags $h > 1$. That means the sample PACF cuts off at lag 1.

Chapter 3

The Simulation Study

In this chapter, simulations will be run to study what happens to the maximum likelihood estimators $\hat{\mu}$ and $\hat{\sigma}$ of the Gumbel parameters μ and σ obtained under the assumption that the data $X_1, ..., X_n$ are iid Gumbel (μ, σ) when in fact the data $X_1, ..., X_n$ are following the Gumbel AR(1) model with series dependence.

3.1 Studies of the Estimators

To study the behavior of our estimators $\widehat{\mu}$ and $\widehat{\sigma}$, we generated 10,000 samples from the Gumbel AR (1) model defined in equation 2.13 with different values of n (sample size) and α. Simulations were performed for $n = 10, 20, 50, 100, 200, 500$ and $\alpha \in \{0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5,\ 0.6,\ 0.7,\ 0.8,\ 0.9\}$, and without loss of generality, we set $\mu = 0$ and $\sigma = 1$. For each combination of n and α, 10,000 dependent samples were generated from the Gumbel AR(1) model. As α increases, the dependence is enhanced. Also, we generated a comparison dataset with $\alpha = 0$ which means that the 10,000 samples are iid.

For each sample, we estimate $\mu = 0$ and $\sigma = 1$ using the maximum likelihood estimators $\widehat{\mu}$ and $\widehat{\sigma}$ under the independence assumption.

Our simulation results are summarized in Tables 3.1 to 3.4 and Figure 3.1 to 3.4, which give the bias and mean squared error (MSE) of the two estimators $\widehat{\mu}$ and $\widehat{\sigma}$. The bias of an estimator is defined to be

$$Bias(\widehat{\theta}) = E(\widehat{\theta} - \theta),$$

and the MSE of an estimator is defined as

$$MSE(\widehat{\theta}) = E\left((\widehat{\theta} - \theta)^2\right),$$

where θ is the true value of a parameter and $\widehat{\theta}$ is an estimator of θ. If $Bias(\widehat{\theta}) = 0$, $\widehat{\theta}$ is an unbiased estimator of θ, and small MSE values indicate better estimation accuracy. In our simulation study, we report sample bias and sample MSE for each combination of n and α, where

$$Sample\ Bias(\hat{\theta}) = \frac{1}{10000}\sum_{i=1}^{10000}(\hat{\theta}_i - \theta),$$

and

$$Sample\ MSE(\hat{\theta}) = \frac{1}{10000}\sum_{i=1}^{10000}(\hat{\theta}_i - \theta)^2,$$

with $\hat{\theta}_i$ being the estimate of θ using the ith sample.

Table 3.1: Biases of $\widehat{\mu}$.

n	$\alpha = 0$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$
10	0.0430	0.0485	0.0505	0.0615	0.0817	0.0912	0.1141	0.1539	0.2037	0.3091
20	0.0209	0.0216	0.0267	0.0323	0.0419	0.0467	0.0602	0.0889	0.1357	0.2143
50	0.0086	0.0084	0.0100	0.0161	0.0174	0.0224	0.0257	0.0387	0.0591	0.1024
100	0.0033	0.0055	0.0047	0.0069	0.0089	0.0112	0.0132	0.0169	0.0278	0.0514
200	0.0008	0.0034	0.0020	0.0032	0.0017	0.0057	0.0096	0.0098	0.0123	0.0352
500	0.0012	-0.0003	0.0008	0.0022	0.0033	0.0022	0.0036	0.0035	0.0049	0.0134

Table 3.1 and Figure 3.1 present the biases of the estimator $\widehat{\mu}$. When $\alpha = 0$ the generated Gumbel samples are independent and $\widehat{\mu}$ and $\widehat{\sigma}$ should perform well. This is indeed the case. With the increase of α, the biases of $\widehat{\mu}$ also increase for a fixed sample size. On the other hand, at a fixed α value, the biases of $\hat{\mu}$ decrease as the sample size increases. In particular, for large α values ($\alpha \geq 0.7$ or strong dependence) and small sample sizes ($n \leq 50$), the biases of $\hat{\mu}$ are not negligible.

The biases of $\widehat{\sigma}$ shown in Table 3.2 and Figure 3.2 are mostly negative under estimation. The biases are relatively high for small samples and/or large α values. Figure 3.2 shows clearly that biases of $\widehat{\sigma}$ is in positive correlation with α, but the influence becomes weak with the increase of the sample size. Compared with the biases of $\hat{\mu}$, large α values have larger influence on the biases of $\hat{\sigma}$.

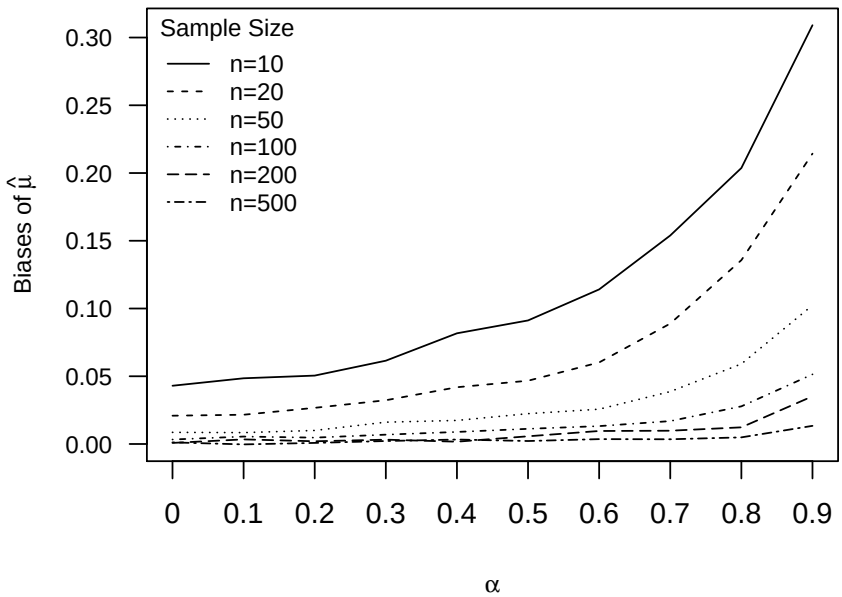

Figure 3.1: Biases of $\widehat{\mu}$ with $\alpha \in (0, 1)$.

Table 3.2: Biases of $\widehat{\sigma}$.

n	$\alpha = 0$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$
10	-0.0737	-0.0803	-0.0994	-0.1257	0.1387	-0.1665	-0.2094	-0.2651	-0.3561	-0.5128
20	-0.0383	-0.0437	-0.0511	-0.0617	-0.0751	-0.0888	-0.1103	-0.1505	-0.2201	-0.3554
50	-0.0144	-0.0172	-0.0197	-0.0263	-0.0300	-0.0363	-0.0487	-0.0671	-0.0985	-0.1954
100	-0.0078	-0.0091	-0.0112	-0.0142	-0.0163	-0.0188	-0.0258	-0.0367	-0.0539	-0.1083
200	-0.0042	-0.0033	-0.0055	-0.0070	-0.0074	-0.0096	-0.0124	-0.0185	-0.0279	-0.0559
500	-0.0011	-0.0018	-0.0019	-0.0020	-0.0030	-0.0041	-0.0057	-0.0074	-0.0108	-0.0257

Since $\sigma = 1$ and $\mu = 0$ are used to generate data from the Gumbel distribution, so the estimates $\widehat{\mu} = 0.3091$ and $\widehat{\sigma} = 1 - 0.5128 = 0.4872$ with $n = 10$ and $\alpha = 0.9$ are a bit too large to be acceptable. But even though α is still 0.9, the estimates $\widehat{\mu} = 0.0134$ and $\widehat{\sigma} = 1 - 0.0257 = 0.9743$ with $n = 500$ are acceptable when the sample size is large.

From Tables 3.3 and 3.4 and Figures 3.3 and 3.4, it is clear that the MSEs in the first column with $\alpha = 0$ are fairly small. The biggest MSE values generally occur when α is large and close to 1. Increasing the sample size generally reduce the MSEs for all different values of α.

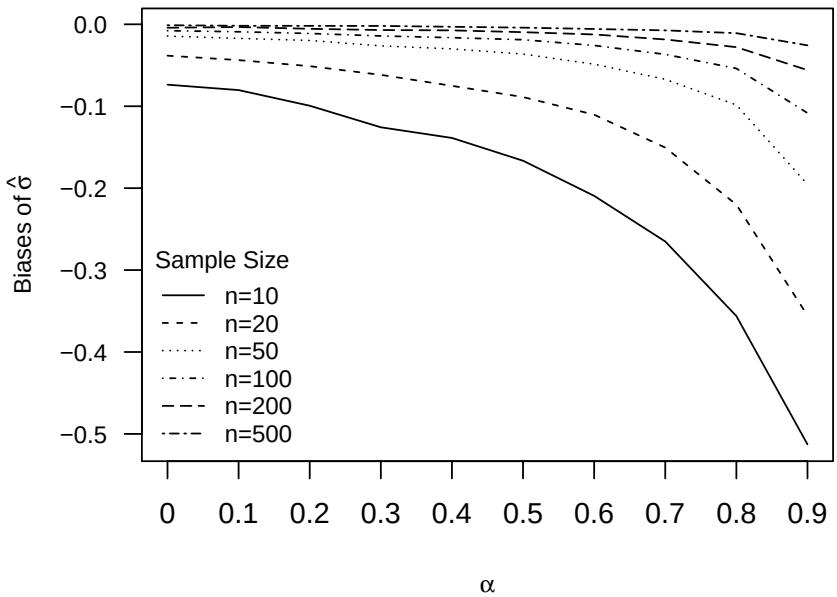

Figure 3.2: Biases of $\widehat{\sigma}$ with $\alpha \in (0, 1)$.

Table 3.3: Mean Squared Errors of $\widehat{\mu}$.

n	$\alpha = 0$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$
10	0.1174	0.1399	0.1700	0.2125	0.2613	0.3291	0.4047	0.5311	0.7469	1.0744
20	0.0565	0.0712	0.0851	0.1061	0.1356	0.1688	0.2195	0.3065	0.4558	0.7527
50	0.0225	0.0278	0.0347	0.0437	0.0536	0.0701	0.0917	0.1276	0.1930	0.3955
100	0.0114	0.0137	0.0171	0.0209	0.0275	0.0351	0.0462	0.0673	0.1038	0.2138
200	0.0056	0.0069	0.0086	0.0107	0.0135	0.0169	0.0229	0.0323	0.0521	0.1076
500	0.0023	0.0028	0.0035	0.0044	0.0055	0.0069	0.0091	0.0133	0.0213	0.0438

Tables 3.1 to 3.4 and Figures 3.1 to 3.4 demonstrate that for small sample sizes ($n \leq 50$), ignoring the dependence between the Gumbel extreme values leads to increased biases and MSEs for $\widehat{\mu}$ and $\widehat{\sigma}$. When the dependence is strong ($\alpha \geq 0.7$), the increased biases and MSEs become serious and can provide misleading results when fitting data to the Gumbel model.

3.2 Studies of the Return Level

The return value is defined as a value that is expected to be equaled or exceeded on average once every interval of time $1/p$ (with a probability of p). We run simulations as in section

Table 3.4: Mean Squared Errors of $\widehat{\sigma}$.

n	$\alpha = 0$	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$
10	0.0650	0.0692	0.0724	0.0816	0.0929	0.1093	0.1336	0.1654	0.2263	0.3589
20	0.0323	0.0331	0.0353	0.0391	0.0469	0.0544	0.0696	0.0944	0.1360	0.2258
50	0.0123	0.0125	0.0139	0.0160	0.0189	0.0230	0.0287	0.0401	0.0608	0.1161
100	0.0061	0.0063	0.0070	0.0079	0.0094	0.0120	0.0150	0.0210	0.0320	0.0638
200	0.0031	0.0032	0.0035	0.0040	0.0047	0.0058	0.0075	0.0104	0.0166	0.0328
500	0.0012	0.0013	0.0014	0.0015	0.0019	0.0024	0.0030	0.0043	0.0064	0.0137

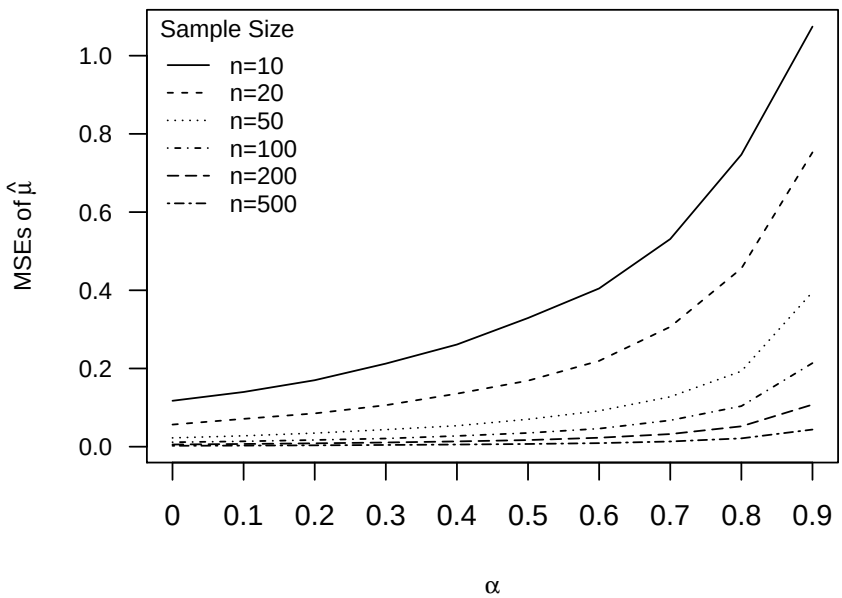

Figure 3.3: MSEs of $\widehat{\mu}$ with $\alpha \in (0, 1)$.

3.1 and report, for each combination of n, α and p , the average return levels, where for each simulated sample, are estimated according to

$$\widehat{x_p} = \widehat{\mu} - \widehat{\sigma} log \left\{ -log \left(1 - p \right) \right\}.$$

We choose $p \in \{0.01, 0.02, 0.05, 0.1, 0.2\}$, then x_p is the return level associated with the return period $1/p \in \{100, 50, 20, 10, 5\}$ time periods.

Figure 3.5 shows plots of return levels for different α, p and sample size $n = 10, 20, 50,$ 100, 200, 500. We see that for each combination of sample size n and p, the return levels decrease with increasing α values, especially when the sample size is small. Moreover, the

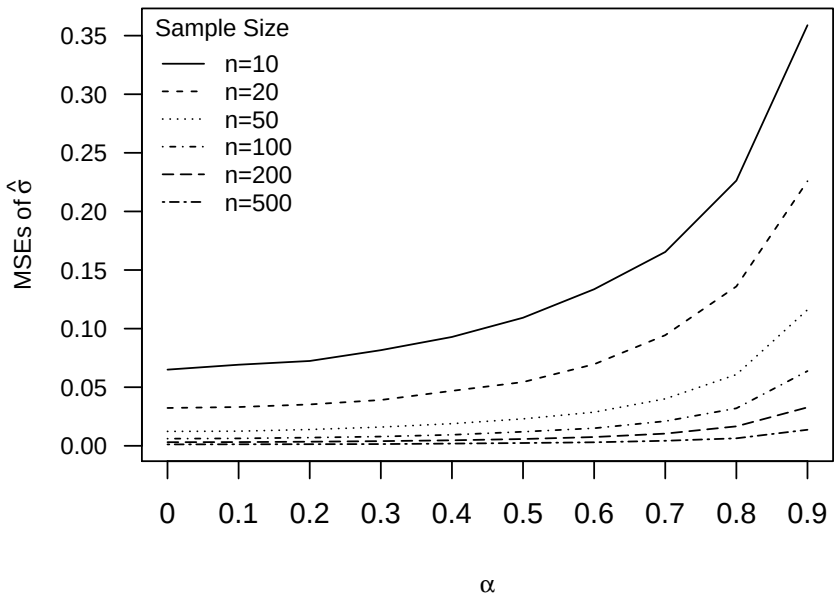

Figure 3.4: MSEs of $\widehat{\sigma}$ with $\alpha \in (0, 1)$.

speed of decreasing is slowing down with the increasing of p. For example, if one year is the time period length, 100 years return level will be affected more seriously with the increase of α than 10 years return level when the sample size is fixed. Increasing the sample size apparently reduces the decrease of return levels for different values of α and p, but for $\alpha \geq 0.8$ and $p \leq 0.02$, the influence of the dependence between the Gumbel extreme values on the return level estimates is still noticeable even when the sample size $n = 100$. From the figures, we can conclude that when n is larger than 100, the influence will become quite weak.

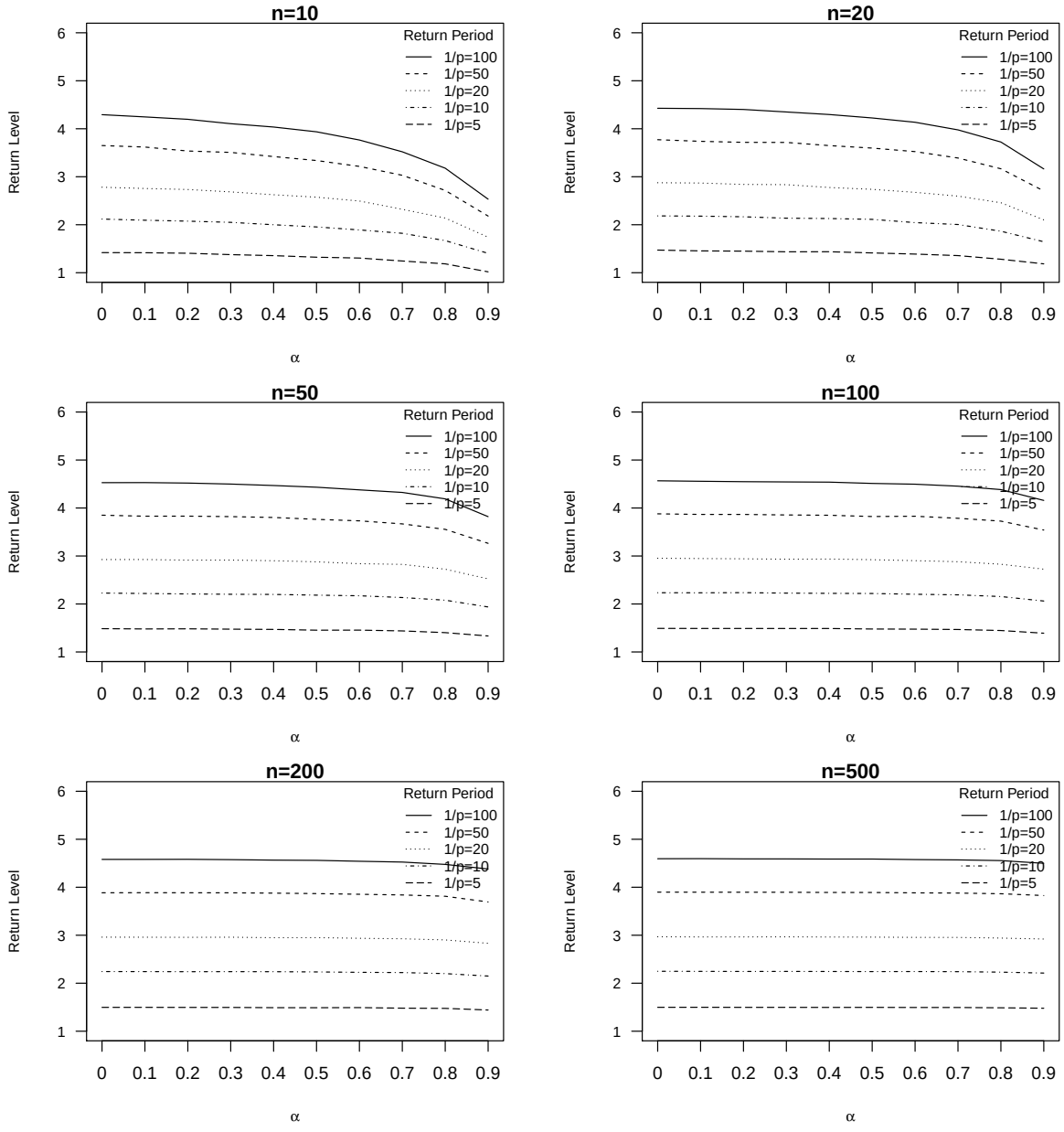

Figure 3.5: Return levels.

Chapter 4

An Example

In this chapter, we use an example to illustrate how to identify the temporal structure among the extreme values in the Gumbel AR(1) model.

This example uses an environmental dataset to illustrate the connection between light-tailed maxima and the Gumbel distribution. The dataset is maximum daily wind speeds (m/s) in Calgary downtown in 2012. Wind storms can generate severe damage to infrastructure and therefore lead to large economic losses. That is why the prediction of maximum wind speeds is essential in many areas of industry. In this example, identifying the temporal structure of the daily fastest wind speed measurements is of primary interest for atmospheric science because this can help to predict future maxima of wind speeds at a specific location.

A scatter plot of the daily maximum wind speeds is displayed in Figure 4.1. It seems to show a linear decrease of maximum wind speeds. To study the dependency between the maximum wind speeds, a scatter plot of x_{i+1} vs x_i is given in Figure 4.2, where a positive dependence is observed.

To see more about this dependence, the sample ACF and sample PACF are plotted in Figure 4.3. We see from Figure 4.3 that the sample ACF decays exponentially and the sample PACF (mostly) cuts off at lag 1, so an autoregressive model of order 1 is suggested. Since we are dealing with maximum wind speeds, we can try to model the data with an AR(1) Gumbel model.

Using the method in Toulemonde *et al.*(2010), for the maximum wind speeds data, the least-square estimator of α is $\tilde{\alpha} = 0.3931$ with a standard error of $s.e.(\tilde{\alpha}) = 0.0481$. This further confirms that the dependence does exist among daily maxima wind speeds in Calgary downtown in 2012.

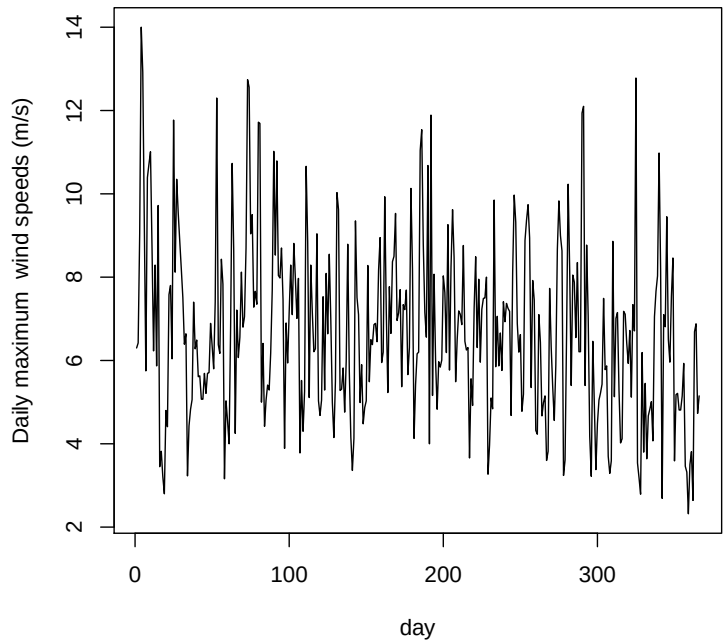

Figure 4.1: Plot of daily maximum wind speeds.

By the conclusion of chapter 3, sample size is an important aspect of the estimation. The daily maxima of wind speeds dataset comes from a sample of size $n = 366$ which is far larger than 100 and $\tilde{\alpha} = 0.3931$ which is less than 0.7, so the dependence among the maximum wind speeds may not affect the accuracy of the modeling results if the Gumbel distribution under the independence assumption is used. In this case, the previous method to obtain the maximum likelihood estimates can still be used. The maximization of the Gumbel log-likelihood for this dataset leads to the estimates

$$(\widehat{\mu}, \widehat{\sigma}) = (5.6944, 1.8263) \; and \; (s.e.(\widehat{\mu}), s.e.(\widehat{\sigma})) = (0.1008, 0.0730),$$

for which the log-likelihood is 788.32. The 100-day return level is estimated to be $\widehat{x}_{0.01} = 14.10$. This value implies that, the speed of 14.10 m/s is expected to be exceeded on average once every 100 days. The corresponding estimate for the 500-day return level is

Table 4.1: Daily maximum wind speeds (m/s) recorded at a specific location in Calgary downtown, 2012.

6.30	6.42	14.00	12.92	9.27	5.75	10.39	10.71	11.01	8.69	6.23	8.29	5.87
9.72	3.45	3.21	2.80	4.80	4.41	7.58	7.80	6.04	11.77	8.12	10.35	9.54
8.86	8.20	6.39	6.64	3.23	4.43	4.83	5.06	7.40	6.28	6.49	5.61	5.63
7.83	3.16	4.55	4.00	6.81	10.73	8.65	4.25	7.21	6.07	6.56	8.12	6.80
7.05	8.70	12.56	9.04	9.50	7.28	7.66	7.35	11.72	11.69	5.00	6.41	4.42
5.05	5.41	6.18	7.69	11.02	8.53	10.79	8.04	7.98	8.70	7.46	3.89	6.90
5.94	7.17	7.10	8.81	7.98	7.01	7.97	3.78	5.52	4.30	5.26	10.66	8.85
5.11	8.29	6.21	6.28	9.04	5.03	4.68	5.05	7.53	5.29	8.09	6.64	8.55
6.98	4.99	5.50	10.03	9.62	5.28	5.31	5.82	4.76	5.99	8.79	5.74	4.16
3.36	4.13	7.51	7.10	4.99	5.90	4.48	4.87	5.03	8.28	5.49	6.50	6.38
6.86	6.89	8.17	8.95	5.95	6.20	9.93	7.18	5.23	7.77	6.65	8.37	8.50
9.53	6.96	7.70	5.37	7.34	7.22	7.69	5.66	6.31	10.13	8.20	4.13	5.42
6.16	6.20	11.54	8.79	7.16	6.56	10.68	4.00	11.89	5.16	8.07	5.98	4.83
5.97	5.84	8.03	7.63	6.19	9.26	5.77	8.24	9.62	8.47	5.49	6.50	7.19
7.11	6.86	6.46	6.25	6.31	3.66	5.56	4.92	7.26	8.49	6.31	7.95	5.96
7.23	7.49	8.00	3.27	3.97	5.10	4.84	9.85	5.85	7.06	5.87	6.66	5.76
7.41	6.93	7.25	7.18	4.68	8.11	9.97	9.32	6.89	6.20	6.62	4.78	5.19
8.88	9.37	8.94	5.35	7.92	7.45	4.32	4.23	7.10	6.43	4.67	4.99	5.15
3.60	3.82	6.32	5.35	4.56	5.89	8.67	9.83	9.00	8.63	3.24	3.61	6.67
10.23	8.14	8.05	7.88	6.55	8.35	6.21	6.21	11.94	12.10	5.40	8.77	7.07
4.18	3.22	4.92	3.38	4.49	5.05	5.22	5.43	7.49	5.78	5.87	3.70	3.29
3.59	8.86	7.01	7.15	5.27	4.02	4.12	7.18	7.11	6.42	5.93	6.72	5.12
7.34	6.71	3.54	3.18	2.79	6.19	3.80	5.45	3.64	4.67	4.83	5.01	4.07
7.05	7.66	10.98	8.17	2.69	7.10	6.81	9.45	6.51	5.96	7.54	8.46	3.59
5.18	5.21	4.81	5.15	5.93	3.46	3.32	2.32	3.50	3.81	2.64	6.69	6.88
4.73	5.15											

$\widehat{x}_{0.002} = 17.04$. Also we use the model checking approach introduced in section 2.2.3 to test whether the dataset fits the Gumbel model.

Two diagnostic plots for assessing the accuracy of the Gumbel model fitted to the data are shown in Figures 4.4. Both plots show a linear relationship, which indicates that a Gumbel fit seems to be reasonable.

Therefore, although there is a short-term dependence among the maximum wind speeds in this example, the maximum likelihood estimators $\hat{\mu}$ and $\hat{\sigma}$ of the Gumbel parameters μ and σ obtained under the independence assumption can still be used because of the large

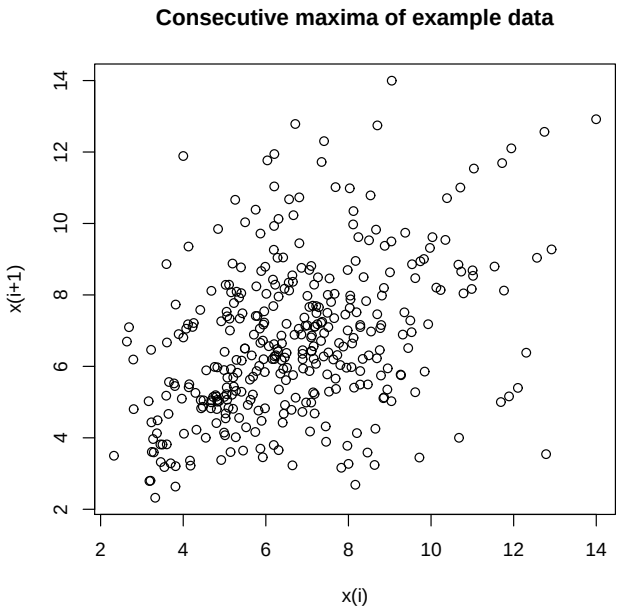

Figure 4.2: Scatter plot of the consecutive daily maxima of the wind speeds.

sample size.

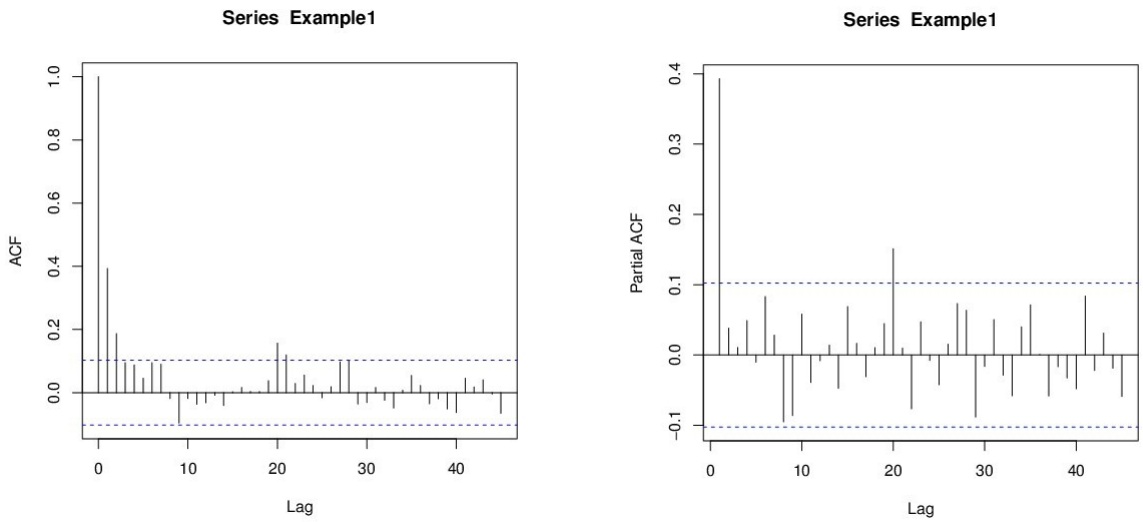

Figure 4.3: The sample ACF and PACF of daily maximum wind speed series.

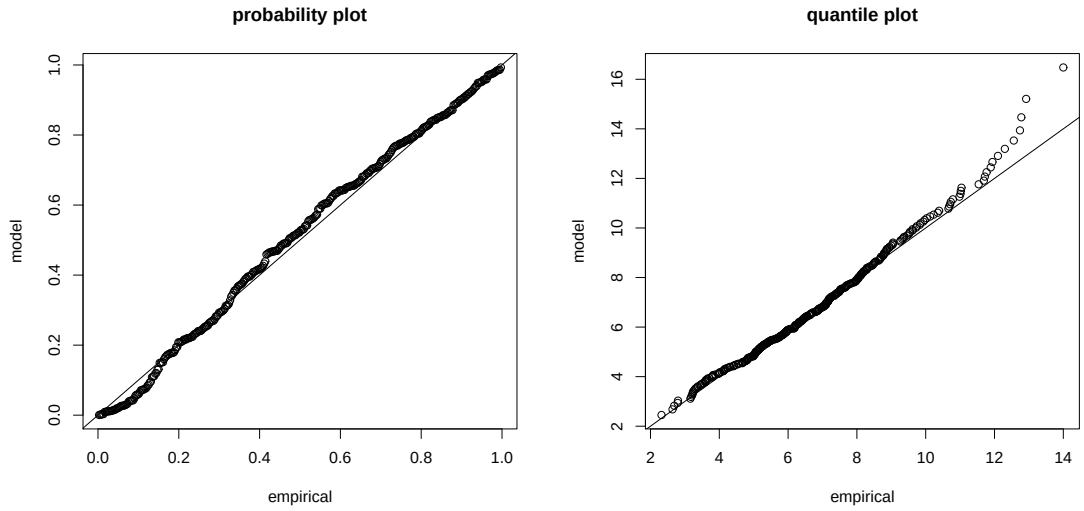

Figure 4.4: Diagnostic Check: probability plot and quantile plot.

Chapter 5

Conclusion and Future Work

Extreme Value Theory has been widely used in the past 50 years in hydrology, meteorology and insurance, especially in the cases where one has the most to lose or to win: stockmarket crash, earthquake, and so on. Some models based on extreme value techniques, such as threshold and multivariate extreme models are attractive to many researchers. Among these popular models, the GEV distribution has been widely and successfully used in fitting maximal data in many areas. In general, extreme values in the GEV models are assumed to be independent. However, the failure of assuming independence of extremes may or may not bias estimators of the parameters of a model. Although some arguments and studies have been done to figure out this problem, the situation was changed until the work of Toulemonde *et al.*(2010).

In this thesis, we use the Gumbel AR(1) model introduced by Toulemonde *et al.* (2010) to generate temporally dependent extreme values that are Gumbel distributed and investigate maximum likelihood estimators of the parameters under the Gumbel distribution. In the simulation part, biases and MSEs of the maximum likelihood estimates $\hat{\mu}$ and $\hat{\sigma}$ of the Gumbel parameters μ and σ are compared under multiple combinations of sample size n and parameter α. The return levels associated with the return period $1/p$ are also shown in chapter 3. We observe biases and MSEs in the estimation of location and scale parameters, μ and σ, of Gumbel models to conclude that ignoring the dependence between the Gumbel extreme values can yield increased biases and MSEs for $\hat{\mu}$ and $\hat{\sigma}$ when the sample size is small ($n \leq 50$). Moreover, large α ($\alpha \geq 0.7$), which means the dependence is strong, leads to increased biases and MSEs for $\hat{\mu}$ and $\hat{\sigma}$. The sample size is also an important aspect. When the sample size is large enough, the influence of dependence among the extreme values will

become weak. In the example, the sample size is large (more than 100) and α are small (less than 0.7), that is why we can estimate the Gumbel parameters μ and σ using the maximum likelihood method which is obtained under the independence assumption. Therefore, the dependence between the extreme values may not affect the accuracy of the final modeling results using the Gumbel distribution when the sample size is large enough and α is small.

Some useful R programs for generating temporally dependent values that are Gumbel distributed and estimating parameters with the maximum likelihood method is given in the Appendix.

For future research, the following issues are worth to be worked on:

- Propose a more general AR model like a GEV AR(1) model and figure out the way of estimating three parameters of the GEV distribution using maximum likelihood method.

- Extend the Gumbel AR(1) model to Gumbel AR(p) models for $p \geq 2$.

Appendix A

Useful R code

The function *rlaptrans* is introduced by Ridout, M.s. (2009). This function implementing the inversion method to generate random variables from the distribution, using a modified Newton-Raphson algorithm, with values of the distribution and density functions obtained by numerical transform inversion.

```
rlaptrans <- function(n, ltpdf, ..., tol=1e-7, x0=1, xinc=2,
                      m=11, L=1, A=19, nburn=38)
#==============================================================
{
    #-------------------------------------------------------
    # Function for generating a random sample of size n from a
    # distribution, given the Laplace transform of its p.d.f.
    #-------------------------------------------------------
    maxiter = 500
      # ----------------------------------------------------
      # Derived quantities that need only be calculated once,
      # including the binomial coefficients
      # ----------------------------------------------------
    nterms = nburn + m*L
    seqbtL = seq(nburn,nterms,L)
    y = pi * (1i) * seq(1:nterms) / L
    expy = exp(y)
    A2L = 0.5 * A / L
```

```
    expxt = exp(A2L) / L
    coef = choose(m,c(0:m)) / 2^m
# ---------------------------------------------------
     # Generate sorted uniform random numbers. xrand will
     # store the corresponding x values
     # ---------------------------------------------------
    u = sort(runif(n), method="qu")
    xrand = u
     #-------------------------------------------------------------
     # Begin by finding an x-value that can act as an upper bound
     # throughout. This will be stored in upplim. Its value is
     # based on the maximum value in u. We also use the first
     # value calculated (along with its pdf and cdf) as a starting
     # value for finding the solution to F(x) = u_min. (This is
     # used only once, so doesn't need to be a good starting value
     #-------------------------------------------------------------
    t = x0/xinc
    cdf = 0
    kount0 = 0
    set1st = FALSE
    while (kount0 < maxiter & cdf < u[n]) {
        t = xinc * t
        kount0 = kount0 + 1
        x = A2L / t
        z = x + y/t
        ltx = ltpdf(x, ...)
```

```
    ltzexpy = ltpdf(z, ...) * expy
    par.sum = 0.5*Re(ltx) + cumsum( Re(ltzexpy) )
    par.sum2 = 0.5*Re(ltx/x) + cumsum( Re(ltzexpy/z) )
    pdf = expxt * sum(coef * par.sum[seqbtL]) / t
    cdf = expxt * sum(coef * par.sum2[seqbtL]) / t
    if (!set1st & cdf > u[1]) {
        cdf1 = cdf
        pdf1 = pdf
        t1 = t
        set1st = TRUE
    }
}
if (kount0 >= maxiter) {
   stop('Cannot locate upper quantile')
}
upplim = t
  #--------------------------------
  # Now use modified Newton-Raphson
  #--------------------------------
lower = 0
t = t1
cdf = cdf1
pdf = pdf1
kount = numeric(n)
maxiter = 1000
for (j in 1:n) {
```

```
  #-------------------------------
  # Initial bracketing of solution
  #-------------------------------
upper = upplim
kount[j] = 0
while (kount[j] < maxiter & abs(u[j]-cdf) > tol) {
    kount[j] = kount[j] + 1
     #-----------------------------------------------
     # Update t. Try Newton-Raphson approach. If this
     # goes outside the bounds, use midpoint instead
     #-----------------------------------------------
    t = t - (cdf-u[j])/pdf
    if (t < lower | t > upper) {
       t = 0.5 * (lower + upper)
    }
   #-----------------------------------------------------
   # Calculate the cdf and pdf at the updated value of t
   #-----------------------------------------------------
    x = A2L / t
    z = x + y/t
    ltx = ltpdf(x, ...)
    ltzexpy = ltpdf(z, ...) * expy
    par.sum = 0.5*Re(ltx) + cumsum( Re(ltzexpy) )
    par.sum2 = 0.5*Re(ltx/x) + cumsum( Re(ltzexpy/z) )
    pdf = expxt * sum(coef * par.sum[seqbtL]) / t
    cdf = expxt * sum(coef * par.sum2[seqbtL]) / t
```

```
            #------------------
            # Update the bounds
            #------------------
          if (cdf <= u[j]) {
              lower = t}
            else {
              upper = t}
      }
      if (kount[j] >= maxiter) {
         warning('Desired accuracy not achieved for F(x)=u')
      }
      xrand[j] = t
      lower = t
   }
   if (n > 1) {
      rsample <- sample(xrand) }
    else {
      rsample <- xrand}
   rsample
}
```

Function *gumbelar1* is used to generate temporally dependent values that are Gumbel distributed through function *ralptrans*.

```
gumbelar1 <- function(n, alpha, burn=200){
 #n is sample size and alpha is between 0 and 1.
   ltpdf <- function(u,alpha) {
            exp(-u^alpha)
```

```
                                }
   e <- rlaptrans(n+burn,ltpdf, alpha=alpha)
   x <- rep(0,n+burn)
   for(i in 1:(n+burn-1)) {
           x[i+1] <- alpha*x[i] + alpha*log(e[i+1])
                          }
   x[-c(1:burn)]
}
```

The *gumbelmle* function returns the maximum likelihood estimates by optimizing the negative log likelihood function.

```
gumbelmle <- function(x){
    fngumbel <- function(theta,x){
 #Define likelihood function.
       mu <- theta[1]
       sig <- theta[2]
       u <- (x-mu)/sig
       length(x)*log(sig) + sum(u) + sum(exp(-u))
                                  }
    xbar <- mean(x)
    s <- sqrt(var(x))
    n <- length(x)
    gumbelini <- c(xbar - 0.57721*sqrt(6)*s/pi, sqrt(6)*s/pi)
    out <- optim(gumbelini, fngumbel, hessian = TRUE,
             control=list(maxit=1000), x=x)
    mle <- out$par
    mle
```

```
}
```

The *gumbelbe* function can return biases and MSEs of maximum likelihood estimators.

```
gumbelbe <- function(n, alpha, nrep=1000){
# Estimate mu=0 and sigma=1 in a Gumbel model
  when the data are Gumbel AR(1)
# bias and mean square error are computed
        z <- matrix(0,nrep,2)
        for(i in 1:nrep) {
            z[i,] <- gumbelmle(gumbelar1(n, alpha))
                        }
  bias <- apply(z,2,mean) - c(0, 1)
  mse <- c(mean(z[,1]^2),mean((z[,2]-1)^2))
  list(bias, mse)
}
```

The function *returnlevel* is used to obtain returnlevels of predictor.

```
returnlevel<- function(n,alpha,p,nrep){
 #1/p is return period.
     z <- rep(0,nrep)
     for (i in 1:nrep){
        est <- gumbelmle(gumbelar1(n,alpha,burn=200))
        z[i] <- est[1] - est[2]*log(-log(1-p))
                      }
     returnlevel <- mean(z)
}
```

Bibliography

[1] Andrews, B., Calder, M. and Davis, RA. (2008), Maximum likelihood estimation for α-stable autoregressive processes. *Annals of Statistics 2009*, 37, 1946-1982.

[2] Brockwell, P.J. and Davis, R.A. (2002), *Introduction to Time Series and Forecasting, Second Edition*. New York: Springer.

[3] Chavez-Demoulin, V. and Davison, A. C. (2005), Generalized additive modelling of sample extremes. *Applied Statistics*, C, 54, 207-222.

[4] Coles, S. G. (2001), *An Introduction to Statistical Modelling of Extreme Values.* London: Springer.

[5] Coles, S., Pericchi and L.R., Sisson, S. (2003), A fully probabilistic approach to extreme rainfall modeling. *Journal of Hydrology*, 273, 35-50.

[6] Davison, A. C. and Hinkley, D. V. (1997), *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

[7] Davison, A. C. and Ramesh, N. I. (2000), Local likelihood smoothing of sample extremes. *Jour- nal of Royal Statistical Society*, B, 42, 191-208.

[8] Davison, A. C. and Smith, R. L. (1990), Models for exceedances over high thresholds (with discussion). *Journal of Royal Statistical Society*, B, 52, 393-442.

[9] Fougères A-L., Nolan ,J. P. and Rootzén, H. (2009), Models for dependent extremes using stable mixtures. *Scandinavian Journal of Statistics*, 36, 42-59.

[10] Matthews, R. (2005), Extreme Value Theory. *25 big idears*, 104-108.

[11] Ridout, M.S. (2009) Generating random numbers from a distribution specified by its Laplace transform. *Statistics and Computing*, 19 (4), 439-450.

[12] Smith, R. L., Tawn, J. A. and Coles, S. G. (1997), Markov chain models for threshold ex- ceedances. *Biometrika*, 84, 249-268.

[13] Toulemonde, G., Guilloub, A. and Naveauc, P. (2012), Particle filtering for Gumbel-distributed daily maxima of methane and nitrous oxide. *Wiley Online Library*, available at `http://www.wileyonlinelibrary.com`.

[14] Toulemonde, G., Guillou, A., Naveau, P., Vrac, M. and Chevallier, F. (2010), Autoregressive models for maxima and their applications to CH4 and N2O. *Environmetrics*, 21, 189-207.